Dynamic Programming in Distributional Reinforcement Learning

Elie Odin, Arthur Charpentier

Dynamic Programming in Distributional Reinforcement Learning

Odin Elie[*1] and **Charpentier Arthur**[2]

[1]École Normale Supérieure de Rennes, elie.odin@ens-rennes.fr
[2]Université du Québec à Montréal, Département de Mathématiques,
charpentier.arthur@uqam.ca

September 2020

Abstract

The classic approach to reinforcement learning is limited in that it only predicts the expected return. The specialized literature has long tried to remedy this problem by studying risk-sensitive models, but the distributional approach will not emerge until 2017. Since the seminal article M. G. Bellemare, Dabney, and Munos 2017 and the state-of-the-art performance of the `C51` algorithm in the ATARI 2600 suite of benchmark tasks (M. G. Bellemare, Naddaf, et al. 2013), research has focused on understanding the behaviour of distributional algorithms. In this paper we place Bellemare's original results in distributional dynamic programming in parallel with the classic results.

One of the foundations of unsupervised learning is interaction with the environment. This involves observing how the environment reacts to certain actions and then using this information to achieve a specific goal. Let's take the example of a chess game. From one player's perspective, the environment consists of the chessboard, the pieces in play and the opponent. For each move made, the opposing player reacts and moves a piece in turn. The experience accumulated by the player over several games allows him to get back to situations that he recognizes as advantageous by specific combinations of moves, and conversely to prevent certain mistakes. Let's mention the following key idea : if the player, regardless of the state of the game, has the possibility to determine the chances of winning offered by each of the actions he can take, then he can easily reach his goal whenever possible.

In more abstract terms, let us consider an agent which interacts with an environment by executing successive actions. After each interaction, the environment give back a reward that depends on the state it was in before the action and the state that followed. The reward formalize the notion of goal and the agent's objective is to maximize the sum of rewards accumulated during the game by choosing the appropriate action for each state. This sum will be called the *return*. It should be noted that we cannot be satisfied with simply choosing the actions that maximize immediate reward. Indeed, in many situations, significant gains can only be obtained after moving through intermediate states and

[*]This work was carried out during an internship at UQAM.

after performing specific combinations of actions. For example, in chess, it may be more profitable to capture a queen in three moves than to capture a pawn in one move.

Reinforcement learning studies the methods by which such an agent can achieve this objective using the experience gained through the interaction. Throughout this paper we will solve this problem using the dynamic programming approach. Initiated by Bellman in 1957 (Bellman 1957), this approach has the advantage to provide encouraging theoretical results with a minimal mathematical background. On the other hand, learning algorithms using dynamic programming require a model of the environment. This hypothesis allows many simplifications because it is no longer necessary to interact with a real environment to recognize the effect of an action. It is moreover possible to study all outcomes from a given state without having to bring the environment into this state. Using the previous example, a model of the chess game allows the agent to try several possible moves from a particular configuration, which is not possible in real game conditions. From a practical point of view, it is not always possible to obtain a model of the environment and the method can become prohibitively time-consuming when the number of states in the environment becomes very large. For a more complete introduction to ideas behind reinforcement learning, see Sutton and Barto 2018 and C. J. C. H. Watkins 1989.

In section 2, we will solve the reinforcement learning problem with the classic approach, where only the expectation of the return is considered. In section 3, we present a variation of the previous approach where the agent consider the full distribution of the return, which removes the limitations caused by the mere consideration of expectation. Finally, we discuss in section 4 practical applications of these results and briefly review recent advances in distributional reinforcement learning.

1 Introduction and setting

The interaction of the agent with the environment is a discrete time process where each step is described as follows: the environment is in a particular state so the agent chooses an action according to this state; then, the environment reacts to this action and moves to another state; finally, a reward is given. Each step occurs at time $t \in \mathbb{N}$ and the next step at time $t+1$. The first step is interpreted as the first time the agent interacts with the environment and occurs at time $t = 0$.

The transition from one state to another is allowed to be random but will be assumed *Markovian*, i.e. the state at time $t+1$ is a random variable which depends only on the state at time t and on the action taken by the agent, not on past event. This hypothesis should be viewed more as a restriction on the states than a restriction on the process. Indeed, in a realistic environment, if a past event produces an effect at time t, then all the information leading to this effect must be contained in the state at time $t-1$. A sufficiently complete model of the environment must therefore satisfy this hypothesis. On the other hand, we have at first glance no restriction to pose regarding the dependence in time of the actions taken by the agent. Note that this freedom will later be restricted because it will be proved to have no impact on learning.

We emphasize that the environment and the actions constitute two separate random sources. The agent's behaviour evolves during the learning process, whereas the response of the environment remains the same.

Standard theoretical framework to model such a process are *Markov decision process* (MDP). The notations and vocabulary we will use are inspired by Sutton and Barto 2018. We let $\mathcal{S}$ denote the set of all possible states of the environment, $\mathcal{A}$ denote the set of all possible actions and $\mathcal{R} \subset \mathbb{R}$ is the set of rewards. It will be convenient to use a probability kernel to describe the transition from one state to another. We will therefore make the standard assumption that these three sets are countable. Note

that in reinforcement learning, $\mathcal{R}$ is sometimes allowed to be continuous (for example see Morimura et al. 2012). However, the study of more general models is rather related to dynamic programming and is outside the scope of this document (Denardo 1967, D. P. Bertsekas 2005, D. Bertsekas and Shreve 1996). To complete this environment, we must specify which actions are accessible from a state $s \in \mathcal{S}$. This is important when some actions are linked to a specific state and have no sense otherwise. We denote the set of all accessible actions from state s by $\mathcal{A}(s)$.

As mentioned above, it is always possible to construct the set of states in such a way that each state s contains all the informations which could affect the future behaviour of the process. The state transition will therefore be assumed Markovian. This means that if the agent takes an action a at a state s, then the probability that the next state will be s' depends only on the current state s and on the action a but not on past actions or events. Thus, completely describing the behaviour of the environment amounts to introduce the probability kernel $P : \mathcal{R} \times \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ which maps a quadruplet (r, s', s, a) to the probability $P(r, s'; s, a)$ to reach a state s' and get the reward r starting from state s and taking action a. The semicolon separates the random contribution from the conditional contribution. Note that for all pairs (s, a) in $\mathcal{S} \times \mathcal{A}$, we have,

$$\sum_{r,s' \in \mathcal{R} \times \mathcal{S}} P(r, s'; s, a) = 1.$$

When we are only concerned with the state following s and a, we introduce,

$$P(s'; s, a) := \sum_{r \in \mathcal{R}} P(r, s'; s, a), \tag{1.1}$$

the probability to reach s' starting from state s and taking action a.

At each time $t \in \mathbb{N}$, the current state, the action taken by the agent and the reward received after the transition are modelled by random variables S_t, A_t and R_{t+1} from a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and with values respectively in $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$. We say that a sequence of random variables $(S_0,\ A_0,\ S_1,\ R_1,\ A_1,\ S_2,\ \ldots)$ satisfies the Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P)$ with initial state S_0 and first action A_0 if for all time steps $t \geq 1$,

$$\begin{aligned} &\mathbb{P}(R_t = r,\ S_t = s' \mid S_{t-1} = s,\ A_{t-1} = a,\ S_{t-2} = s_2,\ A_{t-2} = a_2,\ \ldots) \\ &= \mathbb{P}(R_t = r,\ S_t = s' \mid S_{t-1} = s,\ A_{t-1} = a) \\ &= P(r, s'; s, a). \end{aligned} \tag{1.2}$$

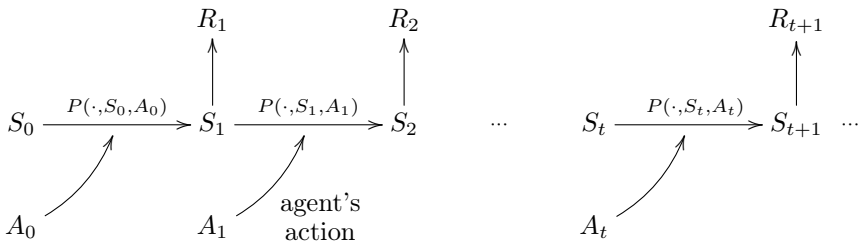
This definition clearly expresses the Markovian side of the process. However, it differs from a Markov chain by the external random source provided by A_t. From (1.2), one can derive these other formulas,

$$\begin{aligned} &\mathbb{P}(S_t = s' \mid S_{t-1} = s,\ A_{t-1} = a,\ S_{t-2} = s_2,\ A_{t-2} = a_2,\ \ldots) \\ &= \mathbb{P}(S_t = s' \mid S_{t-1} = s,\ A_{t-1} = a) \\ &= P(s'; s, a). \end{aligned} \tag{1.3}$$

and using the fact that $\mathbb{P}(A|B \cap C) = \mathbb{P}(A \cap B|C) / \mathbb{P}(B|C)$,

$$\begin{aligned} &\mathbb{P}(R_t = r \mid S_{t-1} = s,\ A_{t-1} = a,\ S_t = s',\ S_{t-2} = s_2,\ A_{t-2} = a_2,\ \ldots) \\ &= \mathbb{P}(R_t = r \mid S_{t-1} = s,\ A_{t-1} = a,\ S_t = s') \\ &= P(r, s'; s, a) / P(s'; s, a). \end{aligned} \tag{1.4}$$

The process is summarized in the following diagram,



In the following, we will often work with probability distributions rather than directly with random variables. We introduce some notations for this purpose.

Notation. For some random state S and random action A, $P(\cdot; S, A)$ will be the law of any random variable S' succeeding to S and A according to the transition kernel $P(\cdot;\cdot,\cdot)$. This distribution is well-defined and, according to the law of total probability, for all s' in $\mathcal{S}$,

$$P(\cdot; S, A)(\{s'\}) \;=\; \sum_{s,a \in \mathcal{S}\times\mathcal{A}} P(s'; s, a)\, \mathbb{P}(S = s, A = a). \tag{1.5}$$

In addition, we denote by $R(S, A)$ any reward variable succeeding to S and A with respect to the transition kernel $P(\cdot;\cdot,\cdot)$. Then, $\mathcal{L}\left(R_{(S,A)}\right)$ is the probability distribution of $R(S, A)$ and for all $r \in \mathcal{R}$,

$$\mathcal{L}\left(R_{(S,A)}\right)(\{r\}) \;=\; \mathbb{P}\left(R(S, A) = r\right) \;=\; \sum_{s,a,s' \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} P(r, s'; s, a)\, \mathbb{P}(S = s, A = a). \tag{1.6}$$

Since our model allows $R(S, A)$ to be dependent on S_1, it will sometimes be useful to denote by,

$$\mathcal{L}\left(R_{(s_1;S,A)}\right) := \mathcal{L}\left(R(S, A) \mid S_1 = s_1\right), \tag{1.7}$$

the conditional distribution of $R(S, A)$ given $S_1 = s_1$ and by $R(s_1; S, A) \sim \mathcal{L}\left(R_{(s_1,S,A)}\right)$ any random variable following this distribution.

Remark 1.1. By writing $\mathbb{P}(S = s, A = a) = \mathbb{P}(A = a \mid S = s)\cdot\mathbb{P}(S = s)$, we see that the distribution of any random variable S' and R succeeding to S and A with respect to the transition kernel P is entirely determined by the distribution of the current state S and the conditional distribution $\mathcal{L}(A|S)$ of A with respect to S. Thus, if an initial distribution $\mathcal{L}(S_0)$ is given and if all conditional distributions $\mathcal{L}(A_t|S_t)$, $t \geq 0$ are specified, then the process is fully determined in distribution. This precisely means that, for each random variable involved in the process, the probability for this variable to lie in a particular set is determined. However, the relationships between these variables are not : there can be a lot of different joint distributions for this process. We emphasize that taking action only with respect to the current state is sufficient to determine the expectation of all future rewards and, by linearity of expectation, sufficient to determine the *expected return*. Therefore, in the classic approach, the agent doesn't need to take into account past actions or states to maximise the expected return because only actions chosen with the knowledge of the current state have an impact on future rewards. However, this is not true in the distributional setting since considering the full distribution of return leads to considering dependences between rewards. (For more details about how the time dependence of actions impact the process, see D. P. Bertsekas 2005 vol. II)

As we said earlier, the goal of reinforcement learning is to determine an agent that maximize the sum of rewards by choosing the best action at each time step. The laws that determine the behaviour of our agent are specified in a *policy*. More precisely, for each state $s \in \mathcal{S}$ and for each time step $t \in \mathbb{N}$, a policy specifies the actions to choose when the state at time t is s. In some cases, a lot of actions can be taken at a particular state s and it would be arbitrary to impose one of them. We can then choose the action randomly following a probability distribution depending on state s ; the policy is said to be *stochastic*.

The random action A_t could be either dependent on the previous variables, we say that it is history-dependent, or independent. More precisely, if A_t is conditionally independent to the history given S_t, then the policy is referred to as *Markovian*. In the classic approach where we try to maximise the expected return, this distinction is irrelevant since we are unable to differentiate between an agent following an history dependent policy and a memoryless agent following a Markovian policy (see remarks 1.1 and 1.3) ; we will thus left the history dependence of A_t unspecified. In the distributional setting, the dependence in history of A_t has an impact on relations between rewards and could affect the distribution of the return. For this reason, in this setting we will always assume that policies are Markovian although this could lead to a loss of performance.

Formally, a policy π is a sequence of functions $(\pi_0, \pi_1, \dots, \pi_t, \dots)$ such that, for all $t \in \mathbb{N}$, π_t is a map from $\mathcal{S}$ to $\mathscr{P}(\mathcal{A})$ the space of probability distributions over $(\mathcal{A}, \mathcal{P}(\mathcal{A}))$. The distribution $\pi_t(s)$ models the manner in which the agent chooses the action a_t when $S_t = s$. We will denote the probability of choosing the action a at time t when $S_t = s$ by,

$$\pi_t(a|s) \;:=\; \mathbb{P}(A_t = a | S_t = s) \;=\; \pi_t(s)(\{a\}).$$

Note that in order for a policy to be strictly consistent with our model, we must have $\pi_t(a|s) \neq 0$ if and only if a is accessible from s, that is $a \in \mathcal{A}(s)$. A policy satisfying this condition is said to be *admissible*.

Remark 1.2. In practice, we will not restrict ourself to admissible policies. In fact, it is possible to construct the kernel transition P in such a manner that it takes into account the restriction to $\mathcal{A}(s)$ regardless of the policy. For example, we can left the environment unchanged and give a null reward after each non-accessible action. That is, we impose $P(0, s; s, a) = 1$ when $a \notin \mathcal{A}(s)$.

The set of all (admissible) policies is denoted by Π. A subclass of policies will prove to be very important later, that of policies which do not change over time. Theses policies are said to be *stationary* and are written $\pi = (\pi_0, \pi_0, \dots)$. The set of stationary policies is denoted by $\overline{\Pi}$. In this case, we will omit the mention of t in the notation $\pi_t(a|s)$.

We can now complete the previous Markov Decision Process (MDP). A sequence of random variables $(S_0, A_0, S_1, R_1, A_1, S_2, \dots)$ satisfies the MDP $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \pi)$, $\pi \in \Pi$, if it satisfies the MDP $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P)$ and if for all $a, s \in \mathcal{A} \times \mathcal{S}$,

$$\mathbb{P}(A_t = a \mid S_t = s) = \pi_t(a|s), \quad \text{for all } t \in \mathbb{N}. \tag{1.8}$$

Note that under a Markovian stationary policy, the sequence $(S_t \times A_t)_{t \in \mathbb{N}}$ becomes a well-defined Markov-chain with transition kernel $P(s', a'; s, a) = P(s'; s, a) \cdot \pi(a'|s')$.

As mentioned before, we will often work with the probability distribution of A_t. Then, $\pi_t(\cdot|S)$ will denote the law of any random variable A_t following the policy π at time t from a random state S_t,

$$\pi_t(\cdot|S)(\{a\}) = \sum_{s \in \mathcal{S}} \pi_t(a|s)\, \mathbb{P}(S_t = s). \tag{1.9}$$

From the beginning of this document, we implicitly assumed that the process, which is in discrete-time, has an infinite number of steps. This is an *infinite horizon* process. We will not treat the case of finite-horizon processes because they can be interpreted as a special case of infinite-horizon processes (see Sutton and Barto 2018 for details).

Since our goal is to find a policy which maximize the return, that is the cumulative reward, we must discuss the definition of this infinite summation. The most straightforward way to do this is to assume that rewards at any time are bounded above and below uniformly over Ω by a common constant, and then to assign multiplicative weights to each reward such that the sum of these weights over time is finite. As is usual, we prefer to give more importance to immediate performance so we will give more weight to rewards immediately following the action taken by the agent by introducing a discount factor $\gamma < 1$ and define the *return* by

$$Z = \sum_{k=0}^{\infty} \gamma^k R_{k+1}. \tag{1.10}$$

Here the weights strictly decrease over time and their sum is $1/(1-\gamma)$. Thanks to the Markovian properties of the environment, an agent is not required to consider past events to be optimal. Each time step can be considered as the initial one and it is enough for the agent to maximize the return calculated from time t. We thus define the *return after time* t by,

$$Z_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{1.11}$$

In the same way, we introduce $Z_\pi(S, A)$ the return of an MDP sequence $(\mathcal{A}, \mathcal{S}, \mathcal{R}, P, \pi)$ with initial pair (S, A). If π is a *stationary* policy, then according to the Markov property, Z_t the return after time t and $Z_\pi(S_t, A_t)$ the return from the initial pair (S_t, A_t) have the same distribution,

$$Z_t \overset{d}{=} Z_\pi(S_t, A_t). \tag{1.12}$$

Remark 1.3. $Z_\pi(S, A)$ is a random variable which is only partly determined by the MDP quintuplet $(\mathcal{A}, \mathcal{S}, \mathcal{R}, P, \pi)$ and that depends strongly on the underlying MDP sequence. Fortunately, its expectation does not. Indeed, $Z_\pi(S, A)$ is bounded and, by remark 1.1 the sequence of distributions $\mathcal{L}(R_t)$ is completely determined, so by the dominated convergence theorem the expectation of $Z_\pi(S, A)$ is equal to the infinite sum of discounted expectations $\gamma^t \mathbb{E}\, R_t$. This reasoning no longer work in the distributional setting. In fact, R_t is not necessarily independent of R_{t+1} so the distribution $\mathfrak{Z}_\pi(S, A) := \mathcal{L}(Z_\pi(S, A))$ is not fully determined. For the distributional Reinforcement Learning perspective, we need more informations, namely the joint law of all random rewards. It is sufficient for this to precise the dependence of A_t with all preceding variables. Then, for two MDP sequences $(\mathcal{A}, \mathcal{S}, \mathcal{R}, P, \pi)$ with reward variables R_t and R'_t respectively, both sequences $\left(\sum_{k=0}^{t} \gamma^k R_{k+1}\right)_t$ and $\left(\sum_{k=0}^{t} \gamma^k R'_{k+1}\right)_t$ converge almost surely to $Z_\pi(S, A)$ and $Z'_\pi(S, A)$ so the two sequences of probability distributions converge weakly to $\mathfrak{Z}$ and $\mathfrak{Z}'$ respectively. By uniqueness of the weak limit of sequences of measures (see Billingsley 2013), we have $\mathfrak{Z} = \mathfrak{Z}'$.

2 Classic approach : value function

In this chapter, we will focus on two reinforcement learning problems and solve them with the classic approach of dynamic programming, that is, by only considering the expected return. The first problem,

called *prediction problem* or sometimes *policy evaluation* consist in determining the return we can expect by following a particular policy. That is, we determine how good the policy is. The second problem consist in identifying the best policies among all existing ones. This is the *control problem*.

From then on, we will assume that rewards at any steps are bounded above and below by a common constant. As a consequence, remark 1.3 applies. We also make the additional assumption that the space of actions is finite ; other spaces are countable.

Improving an agent following a policy π requires knowing if it is profitable or not to be in a state s, choosing an action a and then following this policy. The classic approach for determining the effectiveness of a policy and answer the prediction problem is to use the *value function*,

$$Q_\pi(s,a) := \mathbb{E}[Z_\pi(s,a)], \tag{2.1}$$

which maps each state-action pair (s,a) to the expected return of a MDP sequence with initial state $S_0 = s$, initial action $A_0 = a$ and which follows the policy π from this point. Obviously, the value function lies in $\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$, the set of bounded functions from $\mathcal{S}\times\mathcal{A}$ to $\mathbb{R}$. The evaluation of this function is apparently computationally costly. However, when the policy is stationary, it is possible to rewrite Q_π as the unique solution of a fixed point problem. Let $\pi \in \overline{\Pi}$ be a *stationary* policy. Then,

$$\begin{aligned}
Q_\pi(s,a) &= \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k R_{k+1} \;\middle|\; S_0=s, A_0=a\right] \\
\text{(Markov property)} \quad &= \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k R(S_{k+1},A_{k+1}) \;\middle|\; S_0=s, A_0=a\right] \\
&= \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[\sum_{s',a'\in\mathcal{S}\times\mathcal{A}} Z_\pi(s',a')\,\mathbb{1}_{\{S_1=s',A_1=a'\}} \;\middle|\; S_0=s, A_0=a\right] \\
&= \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[\sum_{s',a'\in\mathcal{S}\times\mathcal{A}} Q_\pi(s',a')\,\mathbb{1}_{\{S_1=s',A_1=a'\}} \;\middle|\; S_0=s, A_0=a\right] \\
&= \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[Q_\pi(\cdot,\cdot)\circ(S_1\times A_1) \;\middle|\; S_0=s, A_0=a\right] \\
&= \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[Q_\pi(S_1,A_1) \;\middle|\; S_0=s, A_0=a\right].
\end{aligned} \tag{2.2}$$

with $A_t \sim \pi(\cdot|S_t)$ and $S_{t+1} \sim P(\cdot;S_t,A_t)$. This recursive expression for Q_π was first stated by Bellman in Bellman 1957 and is known as the *Bellman equation*. This key result in dynamic programming is the basis for the characterisation of value functions by fixed point equations.

From now on, we will omit the conditional part in the expectations since the MDP sequence of random variables $(S_0, A_0, S_1, R_1, A_1, S_2, \ldots)$ implicitly admits $S_0 = s$ as initial state and $A_0 = a$ as initial action. We will therefore simply write,

$$Q_\pi(s,a) = \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[Q_\pi(S_1,A_1)\right]. \tag{2.3}$$

To complete our problem, we introduce the *Bellman operator* $\mathcal{T}^\pi$ from $\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$ to itself such that for all $s,a \in \mathcal{S}\times\mathcal{A}$ and $Q \in \mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$,

$$\mathcal{T}^\pi Q(s,a) = \mathbb{E}[R(s,a)] + \gamma\,\mathbb{E}\left[Q(S_1,A_1)\right], \tag{2.4}$$

where A_1 follows policy π. Hence, the Bellman operator $\mathcal{T}^\pi$ is a map between value functions and admits Q_π as fixed point. As in Denardo 1967, we will use the fact that the value function space

$\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$ is complete under the uniform distance d_∞,

$$d_\infty(u,v) \;=\; \sup_{s,a\in\mathcal{S}\times\mathcal{A}} |u(s,a)-v(s,a)|, \quad u,v\in\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R}).$$

It remains to prove that the Bellman operator is a contraction under this distance.

Theorem 2.1. *Let $\pi\in\overline{\Pi}$ be a stationary policy. Then the Bellman operator $\mathcal{T}^\pi$ is a γ-contraction over $(\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R}),\ d_\infty)$. That is, for all $u,v\in\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$,*

$$d_\infty(\mathcal{T}^\pi u,\mathcal{T}^\pi v) \;\leq\; \gamma\cdot d_\infty(u,v).$$

Proof. Let u and v be elements of $\mathcal{B}(\mathcal{S}\times\mathcal{A},\mathbb{R})$. Then,

$$d_\infty(\mathcal{T}^\pi u,\mathcal{T}^\pi v) \;=\; \sup_{s,a\in\mathcal{S}\times\mathcal{A}} \gamma\cdot\left|\mathbb{E}\left[u(S_1,A_1)\right]-\mathbb{E}\left[v(S_1,A_1)\right]\right|,$$

with $S_1\sim P(\cdot|s,a)$ and $A_1\sim\pi(\cdot|S_1)$. In addition, the object of the supremum in the right-hand side of the above equation satisfies,

$$\left|\,\mathbb{E}\left[u(S_1,A_1)\right]-\mathbb{E}\left[v(S_1,A_1)\right]\,\right| \;\leq\; \mathbb{E}\left[|u(S_1,A_1)-v(S_1,A_1)|\right] \;\leq\; \sup_{s,a\in\mathcal{S}\times\mathcal{A}} |u(s,a)-v(s,a)|.$$

As a result,

$$d_\infty(\mathcal{T}^\pi u,\mathcal{T}^\pi v) \;\leq\; \gamma \sup_{s,a\in\mathcal{S}\times\mathcal{A}} |u(s,a)-v(s,a)| \;=\; \gamma\cdot d_\infty(u,v),$$

which ends the proof. □

Now we can apply the Banach fixed point theorem and conclude with,

Corollary 2.2. *If π is a stationary policy, then Q_π is the unique bounded solution of the Bellman equation. Moreover, the recursive application of $\mathcal{T}^\pi$ over an arbitrary bounded function Q induces a sequence $(Q,\mathcal{T}^\pi Q,\dots,(\mathcal{T}^\pi)^n Q,\dots)$ which converges exponentially quickly to Q_π.*

At this point, a natural question arises. How to find a policy that maximise the value function from some starting point ? We could first solve the following subsidiary issue : What is the maximum return we can hope by starting from some states s, then taking action a and following a policy π ? For this purpose, we introduce the notion of optimal value function.

Definition 2.1. A value function will be said to be *optimal* if it is greater at any points to any other value functions. More precisely the *optimal value function*, denoted Q^*, is the point-wise supremum of all value functions over all admissible policies,

$$Q^*(s,a)=\sup_{\pi\in\Pi} Q_\pi(s,a). \tag{2.5}$$

In the same way, a policy π whose value function is Q^* is an *optimal policy*. It is not clear that there exists an optimal policy (see Denardo 1967) and if it exists, there could be a lot of them. The set of all optimal policies, which could be empty, will be denoted by Π^*. In the following paragraphs, we prove that under our assumptions there indeed exists a stationary deterministic optimal policy and it has an explicit formula depending on Q^*. First, let us try to intuitively draw this formula. We first use the principle of optimality stated by Bellman.

Principle of optimality (Bellman 1957). *An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

This principle justify the research for a *stationary* optimal policy. From a starting configuration (s, a) we receive a random reward r_1 and pass to a random state s'. Then the best strategy is *a priori* to choose the action a' that maximizes the optimal value function, that is, to choose a' in $\arg\max_{a\in\mathcal{A}} Q^*(s', a)$. The policy which, at each steps, chooses a such action a' is denoted π^*. Thus,

$$\pi^*(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s, a), \tag{2.6}$$

and the Bellman operator associated with this policy $\mathcal{T}^{\pi^*}$ satisfies,

$$\mathcal{T}^{\pi^*} Q^*(s, a) = \mathbb{E}[R(s, a)] + \gamma\, \mathbb{E}\left[\max_{a'\in\mathcal{A}} Q^*(S_1, a')\right], \tag{2.7}$$

where $\max_{a'\in\mathcal{A}} Q^*(S_1, a')$ is the function which maps a state $s \in \mathcal{S}$ to $\max_{a\in\mathcal{A}} Q^*(s, a)$ composed with S_1. This maximum formulation only applies for Q^* and is not correct for other value functions because π^* takes after each state the best decision with the knowledge of Q^*. We say that π^* is *greedy* with the knowledge of Q^*. More generally, we can define greedy policies for any value function Q.

Definition 2.2. A greedy policy for a value function Q maximises $Q(s, \cdot)$ for each $s \in \mathcal{S}$. The set of greedy policies for Q is

$$\mathcal{G}_Q := \left\{\pi \in \Pi : \forall s \in \mathcal{S}, \sum_{a\in\mathcal{A}} \pi(a|s) Q(s, a) = \max_{a\in\mathcal{A}} Q(s, a)\right\} = \left\{\pi \in \Pi : \operatorname{supp} \pi(s) \subset \arg\max_{a\in\mathcal{A}} Q(s, a)\right\}.$$

We can then complete (2.7) by introducing an operator $\mathcal{T}^*$ that mimics a greedy update for all value functions Q, i.e.

$$\mathcal{T}^* Q(s, a) = \mathbb{E}[R(s, a)] + \gamma\, \mathbb{E}\left[\max_{a'\in\mathcal{A}} Q(S_1, a')\right]. \tag{2.8}$$

This is the *Bellman optimality operator*. In the following we prove that the optimal value function Q^* is a fixed point of $\mathcal{T}^*$. In other words, Q^* is the only bounded function that satisfies the *optimality equation*,

$$Q^*(s, a) = \mathbb{E}[R(s, a)] + \gamma\, \mathbb{E}\left[\max_{a'\in\mathcal{A}} Q^*(S_1, a')\right], \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{2.9}$$

Then, we prove that any greedy policy with respect to Q^* is optimal. As a result, π^* is an optimal stationary policy.

Proposition 2.3. *The optimal value function Q^* satisfies the optimality equation (2.9).*

The following proof is from Ross 1983 and recopied to make this document self-sufficient.

Proof. Let us prove that the left hand-side of (2.9) is everywhere inferior to the right hand-side. Let $s, a \in \mathcal{S} \times \mathcal{A}$ and let $\pi = (\pi_1, \pi_2, \ldots)$ be an admissible policy. We denote $\pi^+ = (\pi_2, \pi_3, \ldots)$ the

one-step-shifted policy, that is, the policy starting at time $t = 1$ instead of $t = 0$. We have,

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}\left[Q_{\pi^+}(S_1, A_1)\right] \\ &\leq \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}\left[Q^*(S_1, A_1)\right] \\ &\leq \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(S_1, a')\right]. \end{aligned}$$

This inequality is true for all policies so taking the supremum in the left hand-side $Q_\pi(s,a)$ and we have the desired result. For the reversed inequality, we remark that for all $\varepsilon > 0$ and all $s \in \mathcal{S}$, there exists an ε-optimal policy $\pi^{\varepsilon,s}$ such that $Q_{\pi^{\varepsilon,s}}(s, \pi^*(s)) \geq Q^*(s, \pi^*(s)) - \varepsilon$. We then constrain the sequence of random actions such that the first one is greedy according to Q^*, that is, $A_1 = \pi^*(S_1)$ and such that the sequence follows an ε-optimal policy thereafter, namely, if $S_1 = s'$, then the sequence $A_2, A_3, \ldots$ will follow $\pi^{\varepsilon,s'}$ conditionally with $S_1 = s'$. This sequence of actions is dependent on the past event S_1 but only the induced policy π^ε impacts the expected return. In fact, we have for all $t > 1$,

$$\pi_t^\varepsilon(a_t|s_t) := \mathbb{P}(A_t = a_t \mid S_t = s_t) = \sum_{s_1 \in \mathcal{S}} \pi_{t-1}^{\varepsilon,s_1}(a_t|s_t) \cdot \mathbb{P}(S_1 = s_1 \mid S_t = s_t).$$

If s and a are fixed, then $\mathcal{L}(S_1)$ is also fixed and π^ε is well-defined. By writing Q_{π^ε} we obtain,

$$\begin{aligned} Q_{\pi^\varepsilon}(s,a) &= \mathbb{E}[R(s,a)] + \gamma \sum_{s' \in \mathcal{S}} Q_{\pi^\varepsilon}(s', \pi^*(s'))\, \mathbb{P}(S_1 = s') \\ &\geq \mathbb{E}[R(s,a)] + \gamma \sum_{s' \in \mathcal{S}} \left(Q^*(s', \pi^*(s')) - \varepsilon\right) \mathbb{P}(S_1 = s') \\ &= \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(S_1, a')\right] - \varepsilon. \end{aligned}$$

And because $Q^*(s,a) \geq Q_{\pi^\varepsilon}(s,a)$ for all $s, a \in \mathcal{S} \times \mathcal{A}$ we have,

$$Q^*(s,a) \;\geq\; \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(S_1, a')\right] - \varepsilon.$$

This result is true for all $\varepsilon > 0$ so we have the other inequality. □

Proposition 2.4. *The Bellman optimality operator $\mathcal{T}^*$ is a γ-contraction and admits Q^* as fixed point. In particular, the optimal value function Q^* is the unique bounded solution of the optimal equation (2.9) and the recursive application of $\mathcal{T}^*$ to any bounded function Q produce a sequence that converges exponentially quickly to Q^*.*

Proof. It only remains to prove that $\mathcal{T}^*$ is a contraction for the d_∞ distance. The rest follows directly from the Banach fixed point theorem. Let u and v be bounded functions of $\mathcal{B}(\mathcal{S} \times \mathcal{A}, \mathbb{R})$. Then,

$$\begin{aligned} d_\infty(\mathcal{T}^*u, \mathcal{T}^*v) &= \sup_{s,a \in \mathcal{S} \times \mathcal{A}} \gamma \left| \mathbb{E}\left[\max_{a' \in \mathcal{A}} u(S_1, a')\right] - \mathbb{E}\left[\max_{a' \in \mathcal{A}} v(S_1, a')\right] \right| \\ &\leq \sup_{s,a \in \mathcal{S} \times \mathcal{A}} \gamma\, \mathbb{E}\left[\left| \max_{a' \in \mathcal{A}} u(S_1, a') - \max_{a' \in \mathcal{A}} v(S_1, a') \right|\right] \\ &\leq \sup_{s \in \mathcal{S}} \gamma \left| \max_{a \in \mathcal{A}} u(s,a) - \max_{a \in \mathcal{A}} v(s,a) \right| \;\leq\; \gamma \sup_{s,a \in \mathcal{S} \times \mathcal{A}} \left| u(s,a) - v(s,a) \right|. \end{aligned}$$

□

Proposition 2.5. *Any greedy policy π with respect to Q^* is optimal in the sense that $Q_\pi = Q^*$.*

Proof. Because Q^* satisfies the optimality equation, one can precise equation (2.7). For all $s, a \in \mathcal{S} \times \mathcal{A}$, we have,

$$\mathcal{T}^\pi Q^*(s,a) = \mathbb{E}[R(s,a)] + \gamma \, \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(S_1, a')\right] = Q^*(s,a).$$

Thus Q^* is a fixed point of $\mathcal{T}^\pi$ and according to corollary (2.2), we have $Q_\pi = Q^*$. □

From the previous results we can deduce an algorithm which, starting from any value function, successively applies the Bellman operator to it. We thus have a sequence that converges exponentially quickly to Q^*. When the iterates get close enough to each other, we could consider that the optimal value function is almost reached and use the last iteration to calculate an almost optimal policy.

Algorithm 1: Classic dynamic programming

Parameters: $\varepsilon > 0$ a small threshold determining accuracy of estimation.
Input : Q a value function in $\mathcal{B}(\mathcal{S} \times \mathcal{A}, \mathbb{R})$.
$\Delta \leftarrow \varepsilon + 1$
while $\Delta > \varepsilon$ **do**
  for $s, a \in \mathcal{S} \times \mathcal{A}$ **do**
    $Q'(s,a) \leftarrow \mathcal{T}^* Q(s,a) = \sum_{s', r \in \mathcal{S} \times \mathcal{R}} P(r, s'; s, a) \left(r + \gamma \max_{a' \in \mathcal{A}} Q(s', a')\right)$
  end
  $\Delta \leftarrow d_\infty(Q, Q') = \max_{s,a \in \mathcal{S} \times \mathcal{A}} |Q(s,a) - Q'(s,a)|$
  $Q \leftarrow Q'$
end
Output an almost optimal stationary policy π such that for all $s \in \mathcal{S}$,
$\pi(s)$ is in $\arg\max_{a \in \mathcal{A}} Q(s,a)$.

Note that, due to the multiple sweeps over the product space $\mathcal{S} \times \mathcal{A}$, this algorithm can only be used when the state space is finite and particularly small. For example, it cannot be applied to a chess game where the number of states is combinatorially large although finite. We discuss in the last section the extensions that overcome this problem.

3 Distributional perspective

In the previous chapter we answered the prediction and the control problems by considering the expected return through a value function. This approach is limiting from the agent's perspective and insufficient to get a deeper understanding of the return, for example variance or multimodality. Algorithms with value function based agents are thereby not risk sensitive. The distributional Reinforcement Learning counteract this difficulty by considering the full distribution of the return. In dynamic programming, the distributional approach is almost as old as the theory itself (see Jaquette 1973, Sobel 1982, White 1988), but these efforts have focused more on probabilistic criterias about the return than on the distribution itself. In reinforcement learning, the distributional setting has been used for specific purposes, for example to model parametric uncertainty (Dearden, Friedman, and Russell 1998) or to design risk-sensitive algorithms (Morimura et al. 2010, Morimura et al. 2012). It will take until 2017 for the prediction and control problems to be resolved in a purely distributional

way in M. G. Bellemare, Dabney, and Munos 2017. This article has laid the foundations for one of the most promising fields of reinforcement learning. Following their approach, this section is devoted to the extention of classic results introduced earlier to the distributional setting.

As in the previous section, all rewards function will be bounded by a common constant C. For simplicity, we will assume that the space of states $\mathcal{S}$ is finite, instead of countable. This is the standard theoretical framework for many recent studies (Rowland et al. 2018, M. G. Bellemare, Roux, et al. 2019), although the results we will present in dynamic programming can be extended to the countable case (M. G. Bellemare, Dabney, and Munos 2017). Moreover, all policies π will be assumed to be Markovian, i.e. for all $t \in \mathbb{N}$,

$$\mathbb{P}(A_t = a \mid S_t = s,\ R_t = r,\ S_{t-1},\ \ldots) \ = \ \pi_t(a|s). \tag{3.1}$$

This implies in particular that, for a MDP sequence $(\mathcal{A}, \mathcal{S}, \mathcal{R}, P, \pi)$ with initial pair (S, A), the distribution of the return is well-defined (see remark 1.3).

To deal with the prediction problem, we proceed similarly as in the precedent section : we first define a distributional Bellman operator associated with the distributional return ; then, we will prove that it is a contraction in an appropriate metric and conclude that the distributional return is the unique fixed point of the Bellman operator. Let us start by defining the distributional equivalent of the value function Q_π.

Notation. We will be led to see the *return* both as a distribution and as a random variable. To make the notation less cluttered, we will use the Fraktur letter $\mathfrak{Z}$ when the return is a distribution and the regular font Z when it is a random variable. Thus, when the random variable $Z_\pi(s,a)$ is defined, we have $\mathfrak{Z}_\pi(s,a) := \mathcal{L}(Z_\pi(s,a))$, and conversely when the distribution $\mathfrak{Z}_\pi(s,a)$ is defined, $Z_\pi(s,a)$ is a random variable with distribution $\mathfrak{Z}_\pi(s,a)$.

We define the *value distribution* as a function from $\mathcal{S} \times \mathcal{A}$ to the space of probability distributions,

$$\begin{aligned} \mathfrak{Z}_\pi : \quad \mathcal{S} \times \mathcal{A} \ &\to \ \mathscr{P}(\mathbb{R}) \\ (s,a) \ &\mapsto \ \mathfrak{Z}_\pi(s,a), \end{aligned}$$

where $\mathscr{P}(\mathbb{R})$ is the space of all probability distributions over $\mathbb{R}$. Its equivalent in terms of random variables is

$$\begin{aligned} Z_\pi : \quad \mathcal{S} \times \mathcal{A} \ &\to \ \mathcal{F}(\Omega, \mathbb{R}) \\ (s,a) \ &\mapsto \ Z_\pi(s,a). \end{aligned}$$

The space of all value distributions is denoted $\mathcal{Z}$. Obviously, the expectation of the value distribution is equal to the value function, $Q_\pi(s,a) = \mathbb{E}\, Z_\pi(s,a)$ for all $s, a \in \mathcal{S} \times \mathcal{A}$.

If the policy is stationary, we can, as in (2.2), write $\mathfrak{Z}_\pi$ in a recursive form.

Proposition 3.1 (distributional Bellman equation)**.** *Let $\pi \in \overline{\Pi}$ be a stationary policy. Then we have the distributional Bellman equation,*

$$\mathfrak{Z}_\pi(s,a) \ = \sum_{s_1, a_1 \in \mathcal{S} \times \mathcal{A}} \left[\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z_\pi(s_1,a_1)\big)\right] \pi(a_1|s_1) \cdot P(s_1;s,a), \tag{3.2}$$

where $\star$ is the convolution between probability measures.

The distributional Bellman equation exposed in M. G. Bellemare, Dabney, and Munos 2017 appear to be more straightforward but, as Morimura et al. 2010 and Morimura et al. 2012, we prefer to detail all independent contributions.

Proof. According to the weak Markov property, for all $s_1 \in \mathcal{S}$ and for all sequences of sets $(B_1, B_2, B_3, \dots) \in \mathcal{P}(\mathbb{R})^{\mathbb{N}}$ we have,

$$\mathbb{P}(R_2 \in B_2, R_3 \in B_3, \dots \mid R_1 \in B_1, S_1 = s_1) \;=\; \mathbb{P}(R_2 \in B_2, R_3 \in B_3, \dots \mid S_1 = s_1).$$

Consequently, R_1 and $\sum_{k=2}^{\infty} R_k$ are conditionally independent given S_1. Then,

$$\begin{aligned}
\mathfrak{Z}_\pi(s,a) \;&= \mathcal{L}\left(R_1 + \gamma \sum_{k=0}^{\infty} \gamma^k R_{k+2}\right) \\
&= \sum_{s_1 \in \mathcal{S}} \left[\mathcal{L}\left(R(s,a) \mid S_1 = s_1\right) \star \mathcal{L}\left(\gamma \sum_{k=0}^{\infty} \gamma^k R_{k+2} \;\middle|\; S_1 = s_1\right)\right] P(s_1; s, a) \\
&= \sum_{s_1 \in \mathcal{S}} \left[\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z_\pi(s_1, A_1)\big)\right] P(s_1; s, a),
\end{aligned}$$

where $\star$ is the convolution between probability measures and where A_1 is taken conditionally to s_1, $A_1 \sim \pi(\cdot|s_1)$. Moreover, π is Markovian so A_1 and R_1 are conditionally independent given S_1. Thus,

$$\mathfrak{Z}_\pi(s,a) \;= \sum_{s_1, a_1 \in \mathcal{S}\times\mathcal{A}} \left[\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z_\pi(s_1, a_1)\big)\right] \pi(a_1|s_1) \cdot P(s_1; s, a).$$

□

In the following, we will see $\mathfrak{Z}_\pi$ as a vector of $\mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$. We can therefore write the vectorial expectation,

$$\mathbb{E}\, Z = {}^t\big(\mathbb{E}\, Z(s,a)\big)_{s,a \in S \times \mathcal{A}}.$$

We then define the *distributional Bellman operator* $\mathfrak{T}^\pi$ between value distributions as a function from $\mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ to itself,

$$\begin{aligned}
\mathfrak{T}^\pi : \quad \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \;&\to\; \mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \\
\mathfrak{Z} \;&\mapsto\; \mathfrak{T}^\pi \mathfrak{Z},
\end{aligned}$$

such that for all $s, a \in \mathcal{S} \times \mathcal{A}$,

$$\mathfrak{T}^\pi \mathfrak{Z}(s,a) = \sum_{s_1, a_1 \in \mathcal{S}\times\mathcal{A}} \left[\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z(s_1, a_1)\big)\right] \pi(a_1|s_1) \cdot P(s_1; s, a). \tag{3.3}$$

The fraktur font is used to prevent confusion with the classic Bellman operator. The link between $\mathcal{T}^\pi$ and $\mathfrak{T}^\pi$ is derived as follows ; by linearity of the expectation, we get,

$$\begin{aligned}
\mathbb{E}\, \mathfrak{T}^\pi Z(s,a) \;&=\; \mathbb{E}\left[\sum_{s_1 \in \mathcal{S}} P(s_1; s, a)\left[R(s_1; s, a) + \gamma Z(s_1, A_1)\right]\right] \\
&=\; \mathbb{E}\left[R(S_1; s, a) + \gamma Z(S_1, A_1)\right] \\
&=\; \mathbb{E}\, R(s,a) + \gamma\, \mathbb{E}\, Q(S_1, A_1) \;=\; \mathcal{T}^\pi Q(s,a).
\end{aligned}$$

where $Q = \mathbb{E}\, Z$. Hence, in a more concise way,

$$\mathbb{E}\, \mathfrak{T}^\pi Z = \mathcal{T}^\pi Q. \tag{3.4}$$

It now remains to define a metric over $\mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ for which $\mathfrak{T}^\pi$ will be a contraction. Several statistical distances can already be eliminated. Indeed, M. G. Bellemare, Dabney, and Munos 2017 emphasize that $\mathcal{T}^\pi$ is neither a contraction in total variation distance nor in Kullback-Leibler divergence or Kolmogorov distance. In fact, these divergences ignore the geometry of the distributions. Moreover, the KL-divergence is only defined for distributions whose support of one is included in the support of the other. Yet, in the discrete case, value distributions and their Bellman update often have disjoint supports. In the next section we introduce a family of distances between distributions which behaves well with respect to the Bellman operator.

3.1 The ℓ_p family of distances between distributions

From now on, $\mathscr{P}_p(\mathbb{R})$ will denote the collection of probability distributions with finite p^{th} moment, that is, for all real valued random variables X, we have $\mathcal{L}(X) \in \mathscr{P}_p(\mathbb{R})$ if and only if $\mathbb{E}\,|X^p| < +\infty$.

To counteract the disjoint-support issues and the lack of sensitivity to the geometry of certain statistical distances, M. G. Bellemare, Dabney, and Munos 2017 used a maximal form of the Wasserstein metric and solve both prediction and control problems. Recall that for two probability distributions μ and ν in $\mathscr{P}_p(\mathbb{R})$, the p-Wasserstein distance between μ and ν is defined as

$$w_p(\mu,\nu) := \inf_{\substack{X\sim\mu\\ Y\sim\nu}} \|X - Y\|_p .$$

The infimum is attained by the quantile transform $F_\mu^{-1}(U)$ and $F_\nu^{-1}(U)$ of a random variable U uniformly distributed on $[0,1]$. Hence,

$$w_p(\mu,\nu) = \left\|F_\mu^{-1}(U) - F_\nu^{-1}(U)\right\|_p = \left\|F_\mu^{-1} - F_\nu^{-1}\right\|_p ,$$

where F_μ^{-1} and F_ν^{-1} are the quantile functions of μ and ν. The p-Wasserstein metric is scale sensitive, sum invariant (see definitions below) and makes complete the space $\mathscr{P}_p(\mathbb{R})$. It is thus adapted to our problem.

Another family of metrics can be used in distributional reinforcement learning: the ℓ_p family of metrics. Let $\mu,\nu \in \mathscr{P}_1(\mathbb{R})$ be two probability distributions over $\mathbb{R}$. For $1 \le p \le \infty$, the ℓ_p distance between μ and ν is defined as the L^p distance between cumulative distribution functions F_μ and F_ν,

$$\ell_p(\mu,\nu) := \|F_\mu - F_\nu\|_p = \Big[\int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|^p\, dx\Big]^{\frac{1}{p}}. \tag{3.5}$$

Since L^p distances are metrics over L^p spaces, the ℓ_p distance is a metric over probability distributions. As the p-Wasserstein metric, the ℓ_p metric is both scale sensitive and sum invariant and makes the space $\mathscr{P}_1(\mathbb{R})$ complete. It should be noted that the p-Wasserstein and the ℓ_p metrics coincide when $p = 1$ and are disjoint otherwise. When $p = 2$, the squared ℓ_2 metric coincide with the Cramér distance which, unlike Wasserstein, has unbiased sample gradients (M. Bellemare et al. 2017).

If we limit ourselves strictly to the theoretical analysis that follows, we can choose indifferently one or the other of these two family of metrics. We will use ℓ_p metrics just because some intermediate results are automatically verified. In practice, where algorithms learn from samples, the impact of

each of these metrics on learning performances is still unclear (see section 4 for further details). Let us now define and prove the properties of ℓ_p metrics previously stated.

Let $\mathbf{d}$ be a *divergence*, that is an application which maps each pair of probability distributions to a non-negative number and such that, for all $\mu, \nu \in \mathscr{P}(\mathbb{R})$, $\mathbf{d}(\mu, \nu) = 0$ if and only if $\mu = \nu$. For two random variables X and Y with distributions $X \sim \mu$ and $Y \sim \nu$, we write $\mathbf{d}(X, Y) := \mathbf{d}(\mu, \nu)$. We say that $\mathbf{d}$ is *scale sensitive* of order $\beta > 0$ if it has the property (**S**),

$$\textbf{(S)} \qquad \mathbf{d}(\gamma X, \gamma Y) \;\leq\; |\gamma|^{\beta} \mathbf{d}(X, Y), \quad \forall \gamma \in \mathbb{R}^*.$$

Likewise, if $Z \sim \eta$ is a random variable independent of X and Y, then $\mathbf{d}$ is said to be *sum invariant* if it has the property (**I**),

$$\textbf{(I)} \qquad \mathbf{d}(X + Z, Y + Z) \;\leq\; \mathbf{d}(X, Y),$$

or identically,

$$\mathbf{d}(\mu \star \eta, \nu \star \eta) \;\leq\; \mathbf{d}(\mu, \nu).$$

Proposition 3.2. *The ℓ_p metric is both scale sensitive of order $1/p$ and sum invariant. That is, it has both properties (**S**) and (**I**).*

The proof of the sum invariance is taken from M. Bellemare et al. 2017.

Proof. Let $\gamma \in \mathbb{R}^*$. Then,

$$\ell_p(\gamma X, \gamma Y) \;=\; \left\| F_\mu\left(\frac{\cdot}{\gamma}\right) - F_\nu\left(\frac{\cdot}{\gamma}\right) \right\|_p \;=\; \left(\int_{\mathbb{R}} \left| F_\mu\left(\frac{x}{\gamma}\right) - F_\nu\left(\frac{x}{\gamma}\right) \right|^p dx \right)^{\frac{1}{p}}.$$

Use the change of variable $u = x/\gamma$ in the integral and see,

$$\ell_p(\gamma X, \gamma Y) \;=\; \left(\gamma \int_{\mathbb{R}} |F_\mu(u) - F_\nu(u)|^p \, du \right)^{\frac{1}{p}} \;\leq\; |\gamma|^{\frac{1}{p}} \, \ell_p(\mu, \nu).$$

To prove the sum invariance, we use the dual form of the ℓ_p metric, which is then viewed as an Integral Probability Metric (IPM),

$$\ell_p(\mu, \nu) \;=\; \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} f d\mu - \int_{\mathbb{R}} f d\nu \right|, \tag{3.6}$$

where $\mathbb{F}_q := \{f,\ f \text{ absolutely continuous}, \left\| \frac{df}{dx} \right\|_q \leq 1\}$, q is the conjugate exponent of p, i.e. $\frac{1}{p} + \frac{1}{q} = 1$ and $\frac{df}{dx}$ is the Radon-Nikodym derivative of f. A proof of (3.6) is given in Dedecker and Merlevède 2007 (see also Rachev 1991). We thus have,

$$\begin{aligned}
\ell_p(X + Z, Y + Z) \;&=\; \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} f d\mu \star \eta - \int_{\mathbb{R}} f d\nu \star \eta \right| \\
&=\; \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x + y) d\mu(x) - \int_{\mathbb{R}} f(x + y) d\nu(x) \right) d\eta(y) \right| \\
&\leq\; \int_{\mathbb{R}} \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} f(x + y) d\mu(x) - \int_{\mathbb{R}} f(x + y) d\nu(x) \right| d\eta(y),
\end{aligned}$$

where the second line is obtained using the fact that $\int_{\mathbb{R}} f d\mu \star \eta = \int_{\mathbb{R}^2} f(x+y) d\mu(x) d\eta(y)$ and then applying the Fubini's theorem. Since for all $y \in R$, $\mathbb{F}_q := \big\{ f(\cdot + y),\ f \text{ absolutely continuous}, \left\| \frac{df}{dx} \right\|_q \leq 1 \big\}$, $\mathbb{F}_q$ is invariant by translation and we have,

$$\begin{aligned} \ell_p(X+Z, Y+Z) &\leq \int_{\mathbb{R}} \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} f d\mu - \int_{\mathbb{R}} f d\nu \right| d\eta \\ &= \sup_{f \in \mathbb{F}_q} \left| \int_{\mathbb{R}} f d\mu - \int_{\mathbb{R}} f d\nu \right| = \ell_p(X, Y), \end{aligned}$$

hence the result. □

In order to work with value distributions, we will use the maximal form $\overline{\ell_p}$ of the ℓ_p metric, i.e. for all $\mathfrak{Z}, \mathfrak{Z}' \in \mathscr{P}_1(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$,

$$\overline{\ell_p}(\mathfrak{Z}, \mathfrak{Z}') := \sup_{s,a \in \mathcal{S} \times \mathcal{A}} \ell_p(\mathfrak{Z}(s,a), \mathfrak{Z}'(s,a)). \tag{3.7}$$

This is a metric over value distributions since we have the general fact,

Lemma 3.3. *If* $\mathbf{d}$ *is a metric over a set* E *and* I *any index set, then the maximal form* $\overline{\mathbf{d}}$ *of this metric,*

$$\overline{\mathbf{d}}(\mathbf{x}, \mathbf{y}) := \sup_{i \in I} \mathbf{d}(x_i, y_i), \quad \forall \mathbf{x}, \mathbf{y} \in E^I,$$

is a metric over E^I.

Proof. We only prove the triangular inequality, the other points are trivial. For all $\mathbf{z} \in E^I$ we have,

$$\begin{aligned} \overline{\mathbf{d}}(\mathbf{x}, \mathbf{y}) = \sup_{i \in I} \overline{\mathbf{d}}(x_i, y_i) &\leq \sup_{i \in I} \left[\overline{\mathbf{d}}(x_i, z_i) + \overline{\mathbf{d}}(z_i, y_i) \right] \\ &\leq \sup_{i \in I} \overline{\mathbf{d}}(x_i, z_i) + \sup_{i \in I} \overline{\mathbf{d}}(z_i, y_i) = \overline{\mathbf{d}}(\mathbf{x}, \mathbf{z}) + \overline{\mathbf{d}}(\mathbf{z}, \mathbf{y}). \end{aligned}$$

□

3.2 Policy evaluation

To solve the prediction problem, it remains to prove that the distributional Bellman operator is a contraction in the $\overline{\ell_p}$ metric, $1 \leq p \leq \infty$. For this purpose, it is useful to rewrite (3.3) in term of cumulative distribution functions. Let $\pi \in \overline{\Pi}$ be a stationary policy. Then, for all $\mathfrak{Z} \in \mathscr{P}(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$, and for all $s, a \in \mathcal{S} \times \mathcal{A}$, we have,

$$\begin{aligned} F_{\mathfrak{T}^\pi \mathfrak{Z}(s,a)}(x) &= \mathbb{P}(\mathfrak{T}^\pi Z(s,a) \leq x) \\ &= \sum_{s_1 \in \mathcal{S}} \left[\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z(s_1, A_1)\big) \right] (]-\infty, x[) \cdot P(s_1; s, a) \\ &= \sum_{s_1 \in \mathcal{S}} F_{R(s_1;s,a) + \gamma Z(s_1, A_1)}(x) \cdot P(s_1; s, a). \end{aligned} \tag{3.8}$$

We can now state the theorem,

Theorem 3.4. *The distributional Bellman operator is a $\gamma^{\frac{1}{p}}$-contraction in the $\overline{\ell_p}$ metric.*

Proof. Let $\mathfrak{Z}$ and $\mathfrak{Z}'$ be in $\mathscr{P}(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$. Then, for all $s, a \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\ell_p\left(\mathfrak{T}^\pi\mathfrak{Z}(s,a), \mathfrak{T}^\pi\mathfrak{Z}'(s,a)\right) &= \left\| \sum_{s_1\in S} \left(F_{R(s_1;s,a)+\gamma Z(s_1,A_1)} - F_{R(s_1;s,a)+\gamma Z'(s_1,A_1)} P(s_1;s,a)\right)\right\|_p \\
&\leq \sum_{s_1\in S} P(s_1;s,a)\cdot \ell_p\Big(\mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z(s_1,A_1)\big), \mathcal{L}\left(R_{(s_1;s,a)}\right) \star \mathcal{L}\big(\gamma Z'(s_1,A_1)\big)\Big) \\
&\leq \sum_{s_1\in S} P(s_1;s,a)\cdot \gamma^{\frac{1}{p}} \ell_p\left(\mathfrak{Z}(s_1,A_1), \mathfrak{Z}'(s_1,A_1)\right) \\
&\leq \gamma^{\frac{1}{p}} \sup_{s_1\in\mathcal{S}} \ell_p\left(\mathfrak{Z}(s_1,A_1), \mathfrak{Z}'(s_1,A_1)\right),
\end{aligned}$$

where we use (3.8) for the first line, the second line is obtained by triangular inequality and the third is obtained by using both properties (**S**) and (**I**) satisfied by ℓ_p. Then, by a similar reasoning we have,

$$\ell_p\left(\mathfrak{Z}(s_1,A_1), \mathfrak{Z}'(s_1,A_1)\right) \leq \sum_{a_1\in\mathcal{A}} \pi(a_1|s_1)\cdot \ell_p\left(\mathfrak{Z}(s_1,a_1), \mathfrak{Z}'(s_1,a_1)\right) \leq \sup_{a_1\in\mathcal{A}} \ell_p\left(\mathfrak{Z}(s_1,a_1), \mathfrak{Z}'(s_1,a_1)\right).$$

So,

$$\ell_p\left(\mathfrak{T}^\pi\mathfrak{Z}(s,a), \mathfrak{T}^\pi\mathfrak{Z}'(s,a)\right) \leq \gamma^{\frac{1}{p}} \sup_{s_1,a_1\in\mathcal{S}\times\mathcal{A}} \ell_p\left(\mathfrak{Z}(s_1,a_1), \mathfrak{Z}'(s_1,a_1)\right).$$

Finally taking the supremum over $\mathcal{S}\times\mathcal{A}$ in the last equation and we have the result. □

3.3 The distributional control problem

We now turn to the control problem. We will first clarify the notion of optimality in the distributional setting and then show that the distributional Bellman optimality operator still converges in a sense that will be specified. The uniqueness of an optimal value distribution will be, however, invalidated. Recall that in the classic approach, the optimal value function defined as

$$Q^*(s,a) = \sup_{\pi\in\Pi} Q_\pi(s,a),$$

is unique even though there exists a lot of optimal policies. That's because only the expectation of the return is taken into account. In the distributional setting, the value distribution strongly depend on the underlying policy and there can exists a lot of value distributions with the same expectation. Formally, we will say that an optimal value distribution is a value distribution corresponding to an optimal policy. The set of optimal value distributions is denoted $\mathcal{Z}^*$. Hence we have,

$$\mathfrak{Z}^* \in \mathcal{Z}^* \iff \forall s,a \in \mathcal{S}\times\mathcal{A},\ \mathbb{E}\, Z^*(s,a) = Q^*(s,a) \iff \exists \pi^* \in \Pi^* \text{ such that } \mathfrak{Z}^* = \mathfrak{Z}_{\pi^*}. \tag{3.9}$$

The same differences arises for Bellman operators. The classic Bellman operator (2.8) execute a greedy update on any value function Q in a universal way, but in the distributional setting, we need to specify how the update is performed, that is to specify which greedy policy is chosen.

Definition 3.1. A *distributional Bellman optimality operator* is any operator $\mathfrak{T}^*$ which implements a greedy selection rule, i.e. for all $\mathfrak{Z} \in \mathcal{Z}$, there exists a greedy policy $\pi \in \mathcal{G}_Z$ such that,

$$\mathfrak{T}^*\mathfrak{Z} = \mathfrak{T}^\pi\mathfrak{Z}.$$

Note that since the policy $\pi = (\pi_1, \pi_2, \ldots)$ is not necessarily stationary, we extend the definition of the Bellman operator to non-stationary policies by setting,

$$\mathfrak{T}^\pi := \mathfrak{T}^{\pi_1}.$$

As in the prediction problem, we are interested in the behaviour of any value distribution under successive applications of $\mathfrak{T}^*$. We will denote $\mathfrak{Z}_{k+1} = \mathfrak{T}^*\mathfrak{Z}_k$ for any $\mathfrak{Z}_0 \in \mathcal{Z}$. From (3.4) we have $\mathbb{E}\,\mathfrak{T}^*Z = \mathbb{E}\,\mathfrak{T}^\pi Z = \mathcal{T}^\pi Q$ and because π is greedy w.r.t $\mathfrak{Z}$ we have $\mathbb{E}\,\mathfrak{T}^*Z = \mathcal{T}^*Q$. Then, the expected distributional Bellman operator behave like the classic Bellman optimality operator and for all $\mathfrak{Z}_1, \mathfrak{Z}_2 \in \mathcal{Z}$, we have,

$$\left\|\mathbb{E}\,\mathfrak{T}^*Z_1 - \mathbb{E}\,\mathfrak{T}^*Z_2\right\|_\infty \;\leq\; \gamma \left\|\mathbb{E}\,\mathfrak{Z}_1 - \mathbb{E}\,\mathfrak{Z}_2\right\|_\infty. \tag{3.10}$$

However, the iterates $\mathfrak{Z}_k$ cannot converges in any sense to a unique optimal value distribution. It converges in fact to the set of non-stationary optimal value distributions $\mathcal{Z}^*$.

Theorem 3.5. *For $1 \leq p \leq \infty$, the sequence of value distributions $(\mathfrak{Z}_k)_{k\in\mathbb{N}}$ converges uniformly in ℓ_p to the set of optimal non-stationary value distributions $\mathcal{Z}^*$,*

$$\overline{\ell_p}(\mathfrak{Z}_k, \mathcal{Z}^*) = \inf_{\mathfrak{Z}^*\in\mathcal{Z}^*} \overline{\ell_p}(\mathfrak{Z}_k, \mathfrak{Z}^*) \underset{k\to+\infty}{\longrightarrow} 0. \tag{3.11}$$

Note that the convergence is uniform because we imposed a finite states space. The countable states space case is treated in M. G. Bellemare, Dabney, and Munos 2017. Before proving theorem 3.5, let us state this useful lemma,

Lemma 3.6. *There exists a time after which all greedy policies with respect to $\mathfrak{Z}_k$ are optimal. In other words, there exists $k \in \mathbb{N}$ such that for all $k' \geq k$, $\mathcal{G}_{Z_{k'}} \subset \Pi^*$.*

Proof. Recall that all rewards are bounded by a constant C. So, all value distributions are bounded by $\frac{1}{1-\gamma}C$. We can then define the upper bound $B := 2\sup_{\mathfrak{Z}\in\mathcal{Z}} \|\mathfrak{Z}\|_\infty$ where $\|\mathfrak{Z}\|_\infty = \sup_{s,a\in\mathcal{S}\times\mathcal{A}} \|\mathfrak{Z}(s,a)\|_\infty$. We will show that, for any state $s \in \mathcal{S}$, the optimal actions for s induce a return which is significantly greater than returns after non-optimal actions. In this case, if we choose a time k such that the value function Q_k is close enough to Q^*, then we have the same result for Q_k : optimal actions for Q_k are a great deal better than the other and they coincide with optimal actions for Q^*. Greedy policies with respect to Q_k then must be optimal policies. The set of optimal actions from a state s is denoted $A^*(s) := \arg\max_{a\in\mathcal{A}} Q^*(s,a)$. The minimal difference between the optimal return and returns after non-optimal actions is,

$$\Delta(s) := Q^*(s,a^*) - \max_{a\notin A^*(s)} Q^*(s,a),$$

where $a^* \in \mathcal{A}^*(s)$. Taking $\varepsilon_k := \gamma^k B$, then there exists a time k_s such that,

$$\varepsilon_{k_s} = \gamma^{k_s} B < \frac{\Delta(s)}{2}.$$

So, for all $k' > k_s$ and for all optimal action $a^* \in \mathcal{A}^*(s)$,

$$\Delta(s) > 2\varepsilon_{k'}.$$

The greedy actions for state s at time $k' > k_s$ are then optimal. In fact, by (3.10), for all non-optimal actions $a \notin \mathcal{A}^*(s)$ we have,

$$|Q_{k'}(s,a) - Q^*(s,a)| \;\leq\; \gamma^{k'} \, |Q_0(s,a) - Q^*(s,a)| \;\leq\; \varepsilon_{k'},$$

and the same for all optimal actions $a^* \in \mathcal{A}^*(s)$, then

$$|Q_{k'}(s,a^*) - Q^*(s,a^*)| \;\leq\; \varepsilon_{k'}.$$

So,

$$Q_{k'}(s,a^*) - Q_{k'}(s,a) \;\geq\; Q^*(s,a^*) - Q^*(s,a) - 2\varepsilon_{k'},$$

and $Q_{k'}(s,a^*) > Q_{k'}(s,a)$ for all $a \notin \mathcal{A}^*(s)$. Therefore, greedy actions for s at time k' are in $\mathcal{A}^*(s)$. We have finitely many states so we can take K the maximum of k_s over S. From then on, for all $k' > K$, greedy actions for any states at time k' are optimal. So any greedy policies $\pi^{k'} \in \mathcal{G}_{Z_{k'}}$ are in Π^*. □

Proof of theorem 3.5. Recall that for all $k \in \mathbb{N}$, there exists a greedy policy $\pi^k \in \mathcal{G}_{Z_k}$ such that $\mathfrak{T}^* \mathfrak{Z}_k = \mathfrak{T}^{\pi^k} \mathfrak{Z}_k$. Let us fix $i \in \mathbb{N}$. By lemma 3.6, there exists $K \in \mathbb{N}$ such that $\pi^K, \pi^{K+1}, \ldots, \pi^{K+i}$ is a familly of optimal policies. Moreover,

$$\mathfrak{Z}_{K+i+1} = \mathfrak{T}^{\pi^{K+i}} \circ \ldots \circ \mathfrak{T}^{\pi^K} \mathfrak{Z}_K. \tag{3.12}$$

In the same way, let us choose an optimal value distribution $\mathfrak{Z}^* \in \mathcal{Z}^*$ and define $\mathfrak{Z}^*_{i+1} := \mathfrak{T}^{\pi^{K+i}} \circ \ldots \circ \mathfrak{T}^{\pi^K} \mathfrak{Z}^*$. Since $\mathfrak{Z}^*$ corresponds to some optimal policy $\pi^* = (\pi^*_1, \pi^*_2, \ldots)$, it is easily checked that $\mathfrak{Z}^*_1 = \mathfrak{T}^{\pi^K} \mathfrak{Z}^*$ corresponds to the optimal policy $(\pi^K_1, \pi^*_1, \pi^*_2, \ldots)$ so $\mathfrak{Z}^*_1 \in \mathcal{Z}^*$. Then, by induction we deduce that $\mathfrak{Z}^*_{i+1}$ is also an optimal value distribution. We can now prove that the distance between $\mathfrak{Z}_{K+i}$ and $\mathcal{Z}^*$ tends to zero when i tends to infinity. By theorem 3.4, we have,

$$\overline{\ell_p}\left(\mathfrak{Z}_{K+i+1}, \mathfrak{Z}^*_{i+1}\right) \;=\; \overline{\ell_p}\left(\mathfrak{T}^{\pi^{K+i}} \mathfrak{Z}_{K+i}, \mathfrak{T}^{\pi^{K+i}} \mathfrak{Z}^*_i\right) \;\leq\; \gamma \overline{\ell_p}\left(\mathfrak{Z}_{K+i}, \mathfrak{Z}^*_i\right).$$

All value distributions are bounded by $C/(1-\gamma)$ so by induction on i, we have

$$\overline{\ell_p}\left(\mathfrak{Z}_{K+i+1}, \mathfrak{Z}^*_{i+1}\right) \;\leq\; \gamma^{i+1} \overline{\ell_p}\left(\mathfrak{Z}_K, \mathfrak{Z}^*\right) \;\leq\; C \frac{\gamma^{i+1}}{1-\gamma}.$$

Thus, for all $\varepsilon > 0$, there exists $i \in \mathbb{N}$ such that $C\frac{\gamma^{i+1}}{1-\gamma} < \varepsilon$. Hence, $\overline{\ell_p}(\mathfrak{Z}_k, \mathcal{Z}^*) \underset{k \to +\infty}{\longrightarrow} 0$. □

This last theorem suggest a certain instability in the distributional Bellman optimality operators. Indeed, the non-stationarity of the underlying greedy policy could leads to operators that, after reaching the set $\mathcal{Z}^*$, indefinitely switch between optimal value distributions. For those interested, a number of negative results were given by M. G. Bellemare, Dabney, and Munos 2017 in the distributional setting. Nevertheless, it should be noted that, under the hypothesis of existence of a unique optimal policy, the theorem states that the Bellman optimality operator is a contraction and that the sequence $(Z_k)_{k \in \mathbb{N}}$ converges exponentially quickly to the optimal value distribution.

3.4 The projected distributional Bellman operator

The previous analysis cannot lead directly to a practical algorithm. Indeed, it seems natural to start from any value distribution, then successively compute its Bellman updates to finally obtain an almost optimal value distribution. However, it is impossible to perform computations on general probability distributions (in our case, distributions over $\mathbb{R}$) due to the potentially infinite number of data to be processed. Following the framework proposed by M. G. Bellemare, Dabney, and Munos 2017, we will work on the set of discrete probability distributions $\mathcal{P}$ whose support is the set of atoms $\{z_1, \dots, z_N\} \subset \mathbb{R}$,

$$\mathcal{P} = \left\{ \sum_{i=1}^{N} p_i \delta_{z_i} \;\middle|\; p_1, \dots, p_N \geq 0,\ \sum_{i=1}^{N} p_i = 1 \right\}.$$

If $\mathcal{R} \subset \mathbb{R}$ is finite, then given a probability distribution $\mathfrak{Z}$ in $\mathcal{P}^{\mathcal{S}\times\mathcal{A}}$ with $\mathfrak{Z}(s,a) = \sum_{i=1}^{N} p_i(s,a)\delta_{z_i}$ for all $s, a \in \mathcal{S} \times \mathcal{A}$, it is then possible to compute its Bellman update $\mathfrak{T}^\pi \mathfrak{Z}$:

$$\begin{aligned} \mathfrak{T}^\pi \mathfrak{Z}(s,a) &= \sum_{s_1, a_1 \in \mathcal{S}\times\mathcal{A}} \left[\left(\sum_{r \in \mathcal{R}} \frac{P(r, s_1; s, a)}{P(s_1; s, a)} \cdot \delta_r \right) \star \left(\sum_{i=1}^{N} p_i(s_1, a_1) \delta_{\gamma z_i} \right) \right] \pi(a_1|s_1) P(s_1; s, a) \\ &= \sum_{\substack{r \in \mathcal{R} \\ 1 \leq i \leq N}} \left(\sum_{s_1, a_1 \in \mathcal{S}\times\mathcal{A}} P(r, s_1; s, a) p_i(s_1, a_1) \pi(a_1|s_1) \right) \delta_{r + \gamma z_i}, \end{aligned} \tag{3.13}$$

for all $s, a \in \mathcal{S} \times \mathcal{A}$. The distribution thus calculated obviously no longer lies in the family $\mathcal{P}$, as its support contains about $\mathbf{card}(\mathcal{R})$ times more atoms. We will therefore project this new distribution on the support $\{z_1, \dots, z_N\}$ using a *projection operator* $\mathbf{P} : \mathscr{P}(\mathbb{R}) \to \mathcal{P}$. The *categorical projection* $\mathbf{P}_\mathcal{C}$, first used in M. G. Bellemare, Dabney, and Munos 2017, has proven to be particularly well suited when the ℓ_2 norm is used (Rowland et al. 2018). Since the Bellman update is also discrete, we just have to define this operator over the categorical distributions. Given $y \in \mathbb{R}$, $\mathbf{P}_\mathcal{C}(\delta_y)$ is defined as

$$\mathbf{P}_\mathcal{C}(\delta_y) = \begin{cases} \delta_{z_1} & \text{if } y \leq z_1, \\ \frac{z_{i+1} - y}{z_{i+1} - z_i} \delta_{z_i} + \frac{y - z_i}{z_{i+1} - z_i} \delta_{z_{i+1}} & \text{if } z_i < y \leq z_{i+1}, \quad 1 \leq i \leq N-1, \\ \delta_{z_N} & \text{if } y > z_N. \end{cases}$$

This operator is extended affinely to any categorical distributions. That is, for such a distribution $\sum_{i=1}^{K} p_i \delta_{y_i}$, we have $\mathbf{P}_\mathcal{C}(\sum_{i=1}^{K} p_i \delta_{y_i}) = \sum_{i=1}^{K} p_i \mathbf{P}_\mathcal{C}(\delta_{y_i})$. Finally, we extend $\mathbf{P}_\mathcal{C}$ to categorical value distributions by applying the projection coordinate by coordinate, so that, for any categorical value distribution $\mathfrak{Z} \in \mathcal{P}^{\mathcal{S}\times\mathcal{A}}$, we have $(\mathbf{P}_\mathcal{C}(\mathfrak{Z}))(s,a) = \mathbf{P}_\mathcal{C}(\mathfrak{Z}(s,a))$.

To summarize, starting from a categorical value distribution, we calculate at each step its Bellman update and then project this new distribution on the support $\{z_1, \dots, z_N\}$ with $\mathbf{P}_\mathcal{C}$. The operator to iterate is thus $\mathbf{P}_\mathcal{C} \circ \mathfrak{T}^\pi$. One may wonder whether the convergence of the sequence $(\mathfrak{Z}_k)_k$ defined recursively by $\mathfrak{Z}_0 := \mathfrak{Z}, \mathfrak{Z}_{k+1} := \mathbf{P}_\mathcal{C} \mathfrak{T}^\pi(\mathfrak{Z}_k)$, is preserved. Before answering in the affirmative, let us introduce a result by Rowland et al. 2018.

Proposition 3.7 (Rowland et al. 2018)**.** *The Cramér metric ℓ_2 endows a subset of $\mathscr{P}(\mathbb{R})$ containing all bounded probability distributions with a notion of orthogonal projection, and the orthogonal projection onto the subset $\mathcal{P}$ is exactly the heuristic projection $\mathbf{P}_\mathcal{C}$. Consequently, $\mathbf{P}_\mathcal{C}$ is a non-expansion with respect to ℓ_2.*

Hence, using the fact that $\mathfrak{T}^\pi$ is a $\sqrt{\gamma}$-contraction in the $\overline{\ell_2}$ metric, we have that, for two categorical value distributions $\mathfrak{Z}$ and $\mathfrak{Z}'$ in $\mathcal{P}^{\mathcal{S}\times\mathcal{A}}$,

$$\overline{\ell_2}(\mathbf{P}_\mathcal{C}\mathfrak{T}^\pi(\mathfrak{Z}), \mathbf{P}_\mathcal{C}\mathfrak{T}^\pi(\mathfrak{Z}')) \;\leq\; \overline{\ell_2}(\mathfrak{T}^\pi(\mathfrak{Z}), \mathfrak{T}^\pi(\mathfrak{Z}')) \;\leq\; \sqrt{\gamma}\cdot\overline{\ell_2}(\mathfrak{Z}, \mathfrak{Z}').$$

Consequently, $\mathbf{P}_\mathcal{C}\mathfrak{T}^\pi$ is a $\sqrt{\gamma}$-contraction in the $\overline{\ell_2}$ metric and the sequence $(\mathfrak{Z}_k)_k$ converges exponentially quickly to the unique categorical value distribution $\mathfrak{Z}_\pi$ satisfying $\mathbf{P}_\mathcal{C}\mathfrak{T}^\pi\mathfrak{Z}_\pi = \mathfrak{Z}_\pi$.

As for the control problem, if we suppose the uniqueness of the optimal policy π^*, then $\mathfrak{T}^* = \mathfrak{T}^{\pi^*}$ so the projected Bellman update 3.13 becomes,

$$\begin{aligned} \mathfrak{T}^*\mathfrak{Z}(s,a) \;&=\; \sum_{\substack{r\in\mathcal{R}\\ 1\leq i\leq N}} \Big(\sum_{s_1\in\mathcal{S}} P(r,s_1;s,a)p_i(s_1,a^*_{s_1})\Big)\delta_{r+\gamma z_i}, \\ &\text{with}\quad a^*_{s_1} \in \arg\max_{a'\in\mathcal{A}} \sum_{i=1}^N p_i(s_1,a')z_i, \end{aligned} \tag{3.14}$$

and we derive the same conclusion.

Therefore, when $\mathcal{S}, \mathcal{A}$ and $\mathcal{R}$ are finite, we deduce the following algorithm,

Algorithm 2: Categorical distributional dynamic programming

Parameters: $\{z_1, \dots, z_N\}$ a support for categorical distributions, $\varepsilon > 0$ a small threshold determining accuracy of estimation.

Input : $\mathfrak{Z} = \left(\sum_{i=1}^N p_i(s,a)\delta_{z_i}\right)_{s,a\in\mathcal{S}\times\mathcal{A}}$ a categorical value distribution in $\mathcal{P}^{\mathcal{S}\times\mathcal{A}}$.

$\Delta \leftarrow \varepsilon + 1$
while $\Delta > \varepsilon$ **do**
  for $s, a \in \mathcal{S}\times\mathcal{A}$ **do**
    # Compute the distributional Bellman update according to 3.14
    $\mathfrak{Z}'(s,a) \leftarrow \mathfrak{T}^*\mathfrak{Z}(s,a)$
    # Project $\mathfrak{Z}'(s,a)$ onto the set of atoms $\{z_1, \dots, z_N\}$
    $\mathfrak{Z}'(s,a) \leftarrow \mathbf{P}_\mathcal{C}(\mathfrak{Z}'(s,a))$
  end
  $\Delta \leftarrow \overline{\ell_2}(\mathfrak{Z}, \mathfrak{Z}')$
  $\mathfrak{Z} \leftarrow \mathfrak{Z}'$
end
Output an almost optimal stationary policy π such that for all $s \in \mathcal{S}$, $\pi(s)$ is in $\arg\max_{a\in\mathcal{A}} \mathbb{E}\,\mathfrak{Z}(s,a)$.

4 Discussions

In this last section we discuss the scope of our results, their extension and practical applications. The analysis carried out in section 2 allowed us to exhibit an algorithm (Algorithm 1) that approximates the optimal value function Q^* by successively applying the Bellman operator $\mathcal{T}$ to an initial value function Q and then derives an almost optimal policy from it. The method used has two drawbacks :

it needs a model of the environment and any update on $Q_k = \mathcal{T}^k Q$ requires a sweep over the entire states space. In practice, the model is not always available and the space of states can be so large that the algorithm becomes prohibitively long when performed by any existing computer (take the example of a chessboard). When no models are available, we are restricted to learn online, that is to learn by directly interacting with the environment and consider samples rewards and transitions. Recall that given a state $s \in \mathcal{S}$ and any action $a \in \mathcal{A}$, then the optimal value function satisfies the Bellman optimality equation,

$$Q^*(s,a) = \mathbb{E}_{S_1 \sim P(\cdot;s,a)} \left[R(s,a) + \gamma \max_{a' \in \mathcal{A}} Q^*(S_1, a') \right],$$

so $R(s,a) + \gamma \max_{a' \in \mathcal{A}} Q^*(S_1, a')$ is an unbiased estimate of $Q^*(s,a)$. This observation leads to a new method for estimating $Q^*(s,a)$. Considering a value function Q_k; if we are in a state s, taking an action a, moving to the state s' and receiving the reward r, then, the Bellman update takes the form,

$$Q_{k+1}(s,a) \;=\; Q(s,a) + \boldsymbol{\alpha} \underbrace{\big[r + \gamma \max_{a' \in \mathcal{A}} Q_k(s_1, a') }_{\text{target}} - Q(s,a) \big],$$

where $r + \gamma \max_{a' \in \mathcal{A}} Q_k(s_1, a')$ is the target return and $\boldsymbol{\alpha}$ can be interpreted as the learning rate. In other words, for each sample transition (s, a, s_1, r), we move the old estimation $Q(s,a)$ towards the target. This method, called *Q-learning*, is a special case of *temporal difference* (TD) learning methods and is one of the most popular in reinforcement learning. The iterates still converge to Q^* provided that all states are visited an infinite number of times. This condition is for example verified if the agent follows a policy which acts according to the current knowledge most of the time and sometimes randomly choose an exploratory action. Q-learning was first introduced in C. J. C. H. Watkins 1989 and the proof was made rigorous by C. Watkins and Dayan 1992 and by Tsitsiklis 1994.

The second problem concerns the size of the states space. When the states space is small, it is possible to store all entries of the value function Q and to use them for updates. These methods are refered to as *tabular methods*. When the space is overwhelmingly large, the best we can do is computing an approximate version of Q. This is commonly achieved by invoking a rich class of functions $\{Q_\theta, \theta \in \mathbb{R}^n\}$ parametrized by a weight vector $\theta \in \mathbb{R}^n$. Given a target value function Q_t, the weights that minimize the difference between Q_θ and Q_t is then computed using various methods, such as gradient descent. When the approximate function is linear with respect to the weight vector, then the precedent method is still guaranteed to converge (see Tsitsiklis and Van Roy 1997). Deep Neural Networks are also used as approximators and are partly responsible for the impressive performances of some recent agents (see Mnih et al. 2015) even if theoretical properties of such combinations are still poorly understood.

As for the distributional setting, we have obtained in section 3.4 an algorithm that approximates the optimal value distribution but the two previous problems persist. The first concrete solution was given by Bellemare in M. G. Bellemare, Dabney, and Munos 2017 when the categorical distribution approximation has regularly spaced outcomes. Their algorithm, `C51`, uses a deep neural network which takes as input the current state s and output the approximate categorical distribution of the return. It then compute the target value distribution from a sample, as in classic Q-learning. They then faced a problem when trying to minimize the distance between the old distribution estimation and the target. In their setting, an obvious choice for the loss distance could have been the Wasserstein metric. However, they showed that a Wasserstein loss cannot be minimized by stochastic gradients methods (see M. Bellemare et al. 2017). Instead, they minimized the Kullback-Leibler divergence between the

projected target and the distribution estimation. Their algorithm performed state-of-the-art results but they did not provide theoretical justifications and left readers with several questions, namely whether there is a guarantee of convergence in a distributional algorithm that learns from samples and how the combination of a projection step and a KL minimization affects performance.

This theory-practice gap was partially elucided in Rowland et al. 2018. They provided a theoretical framework for the analysis of categorical distributional reinforcement learning and derive the first proof of samples based distributional algorithms in the tabular case. In concrete terms, they proved the convergence of a distributional version of Q-learning when there exists a unique optimal policy π^*. Moreover, the Cramér metric ℓ_2^2 made its first appearance in DRL as they proved that the projection step in the `C51` has the property to minimize the Cramér distance between the projected distribution and the original one (see proposition 3.7). However, the role of the Kullback-Leibler divergence remained unclear.

On the other hand, Dabney et al. 2018 created an algorithm that directly minimize the Wasserstein metric between the output of the approximator (for example a deep neural network) and the target. This enabled them to do away the undesired projection step since the Wasserstein metric prevent the disjoint-support issue and to remove the fixed range constraint imposed in the `C51` algorithm. The minimization of the Wasserstein metric was performed by using quantile regression instead of gradient descent. These improvements allowed their algorithm to outperform the original `C51`.

The most recent theoretical contribution to distributional reinforcement learning is from M. G. Bellemare, Roux, et al. 2019. They created an algorithm that operate end-to-end on the Cramér metric and they offered theoretical guarantees of the behaviour of this algorithm when combined with linear functions approximation.

Acknowledgements

The authors thank Hélène Guérin for her thoughtful feedback on this paper.

References

Bellemare, Marc et al. (May 2017). "The Cramer Distance as a Solution to Biased Wasserstein Gradients". In: *ArXiv* **1705.10743**,

Bellemare, Marc G., Will Dabney, and Rémi Munos (2017). "A Distributional Perspective on Reinforcement Learning". In: *ArXiv* **1707.06887**,

Bellemare, Marc G., Yavar Naddaf, et al. (2013). "The Arcade Learning Environment: An Evaluation Platform for General Agents". In: *Journal of Artificial Intelligence Research* **47**, pp. 253–279.

Bellemare, Marc G., Nicolas Le Roux, et al. (Feb. 2019). "Distributional reinforcement learning with linear function approximation". In: *Proceedings of AISTATS 2019*.

Bellman, Richard (1957). *Dynamic Programming*. 1st ed. Princeton, NJ, USA: Princeton University Press.

Bertsekas, D.P. and S.E. Shreve (1996). *Stochastic Optimal Control: The Discrete Time Case*. Athena scientific optimization and computation series. Athena Scientific.

Bertsekas, Dimitri P. (2005). *Dynamic Programming and Optimal Control*. 3rd ed. Vol. I & II. Belmont, MA, USA: Athena Scientific.

Billingsley, P. (2013). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley.

Dabney, Will et al. (Feb. 2018). "Distributional Reinforcement Learning with Quantile Regression". In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dearden, Richard, Nir Friedman, and Stuart Russell (1998). "Bayesian Q-Learning". In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence*.

Dedecker, Jerome and F. Merlevède (2007). "The empirical distribution function for dependent variables: asymptotic and non asymptotic results in L^p". In: *ESAIM: Probability and Statistics* **11**, Publisher: EDP Sciences, pp. 102–114.

Denardo, Eric (Apr. 1967). "Contraction Mappings in the Theory Underlying Dynamic Programming". In: *SIAM Review* **9(2)**, pp. 165–177.

Jaquette, Stratton C. (1973). "Markov Decision Processes with a New Optimality Criterion: Discrete Time". In: *The Annals of Statistics* **1(3)**, pp. 496–505.

Mnih, V. et al. (2015). "Human-level control through deep reinforcement learning". In: *Nature* **518**, pp. 529–533.

Morimura, Tetsuro et al. (2010). "Nonparametric Return Distribution Approximation for Reinforcement Learning". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 799–806.

— (2012). "Parametric Return Density Estimation for Reinforcement Learning". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Rachev, Svetlozar T. (Mar. 1991). *Probability Metrics and the Stability of Stochastic Models*. Chichester ; New York: John Wiley & Sons Ltd.

Ross, Sheldon (1983). *Introduction to Stochastic Dynamic Programming*. New York: Academic Press.

Rowland, Mark et al. (2018). "An Analysis of Categorical Distributional Reinforcement Learning". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*.

Sobel, Matthew J. (Dec. 1982). "The variance of discounted Markov decision processes". In: *Journal of Applied Probability* **19(4)**, pp. 794–802.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book.

Tsitsiklis, John N. (1994). "Asynchronous Stochastic Approximation and Q-Learning". In: *Machine Learning* **16**, pp. 185–202.

Tsitsiklis, John N. and Benjamin Van Roy (May 1997). "An Analysis of Temporal-Difference Learning with Function Approximation". In: *IEEE TRANSACTIONS ON AUTOMATIC CONTROL* **42(5)**, pp. 674–690.

Watkins, Chris and Peter Dayan (1992). "Q-learning". In: *Machine Learning* **8**, pp. 279–292.

Watkins, Christopher John Cornish Hellaby (May 1989). "Learning from Delayed Rewards". PhD thesis. Cambridge, UK: King's College.

White, D. J. (Jan. 1988). "Mean, variance, and probabilistic criteria in finite Markov decision processes: A review". In: *Journal of Optimization Theory and Applications* **56**, pp. 1–29.