Master MVA / Mathématiques, Vision, Apprentissage
*École Normale Supérieure de Paris-Saclay*

Stage 2021–2022
Alexandre Marthe

# Titre du rapport

**Résumé** : *Un résumé de 10 lignes environ du contenu du rapport, permettant de situer le sujet et les résultats principaux du stage. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.*

**Mots clefs** : *4 à 5 mots clefs*

Stage encadré par :
**Aurelien Garivier**
aurelien.garivier@ens-lyon.fr / tél. (+33) 4 72 72 81 08
UMPA, ENS de Lyon
*46, allée d'Italie*
*69364 Lyon Cedex 07, FRANCE*
http://www.umpa.ens-lyon.fr

May 17, 2022

# Remerciements

# Contents

# 1 Introduction

# 2 Related work

[introduction to reinforcement learning]

## 2.1 Markov Decision Process

We will first start by introducing the general framework of Markov Decision Process (MPD) and the basic results on dynamic programming.

**Definition 1** (Markov Decision Process and Dynamic Programming)**.** *An MPD is a tuple $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{X}$ is a finite state space, $\mathcal{A}$ a finite action space, $P$ a transition kernel, $R$ a stochastic reward, and $\gamma$ the discount*

a policy $\pi$ is a mapping from $\mathcal{X}$ to probability distribution on $\mathcal{A}$.
We are interested in the total reward:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$$

**Policy Evaluation:** The first problem when considering a MPD, is being able to evaluate a policy, i.e. compute the total reward obtained when following the policy. For this we introduce the Value function $V^\pi$ (resp. $Q^\pi$) which consist in the expected total reward with $\pi$ and starting at state $x$ (resp. starting at state $x$ and with action $a$) :

**Definition 2.** *The Value function and the Q-value function (also called action-state value function) are defined by:*

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x\right]$$

$$Q^\pi(x, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a\right]$$

*with $x_t \sim p(\cdot|x_{t-1}, a_{t-1})$ and $a_t \sim \pi(\cdot|x_t)$*

By manipulating the expressions, we can see that those two function verify the following equation called Bellman equation:

$$V^\pi(x) = \sum_{a \in \mathcal{A}} \pi(a|x) \left(\mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) V^\pi(x')\right) \tag{1}$$

$$Q^\pi(x, a) = \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x', a' \in \mathcal{X} \times \mathcal{A}} p(x'|x, a) \pi(a'|x') Q^\pi(x', a') \tag{2}$$

This suggests to introduce of the Bellman Operator:

**Definition 3** (Bellman Operator)**.** *Let $V : \mathcal{X} \mapsto \mathbb{R}$ or $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$, $\pi$ a policy. The Bellman operator $\mathcal{T}^\pi$ is defined by:*

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\pi V(x) = \sum_{a \in \mathcal{A}} \pi(a|x) \left(\mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) V(x')\right)$$

$$\forall x, a \in \mathcal{X} \times \mathcal{A}, \qquad \mathcal{T}^\pi Q(x, a) = \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x', a' \in \mathcal{X} \times \mathcal{A}} p(x'|x, a) \pi(a'|x') Q(x', a')$$

This operator happens to be a contraction (describe more)... policy evaluation algorithm

**Control:** The second problem is to find a policy that maximizes the expected return. For that we introduce the optimal value functions:

$$V^\star(x) = \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x\right]$$

$$Q^\star(x, a) = \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a\right]$$

Those functions satisfy the optimal version of the Bellman Equation:

$$V^\star(x) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) V^\star(x') \tag{3}$$

$$Q^\star(x, a) = \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) \max_{a' \in \mathcal{A}} Q^\star(x', a') \tag{4}$$

We can then introduce an optimal version of the Bellman Operator:

**Definition 4** (Optimal Bellman Operator)**.** *Let $V : \mathcal{X} \mapsto \mathbb{R}$ or $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$, $\pi$ a policy. The Bellman operator $\mathcal{T}^\star$ is defined by:*

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\star V(x) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) V^\star(x')$$

$$\forall x, a \in \mathcal{X} \times \mathcal{A}, \qquad \mathcal{T}^\star Q(x, a) = \mathbb{E}\left[r(x, a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) \max_{a' \in \mathcal{A}} Q^\star(x', a')$$

This operator is also a contraction (describe more)... Value iteration algorithm

## 2.2 Metrics

[introduire]

### Wasserstein Metric

**Definition 5** ([1])**.** *Let $p \geq 1$ and $\mathscr{P}_p(\mathbb{R})$ the space of distributions with finite $p^{th}$ moment. Let $\nu_1, \nu_2 \in \mathscr{P}_p(\mathbb{R})$ and $\Lambda(\nu_1, \nu_2)$ the set of distribution on $\mathbb{R}^2$ with marginals $\nu_1$ et $\nu_2$. The p-Wasserstein distance $d_p$ is then defined as :*

$$d_p(\nu_1, \nu_2) = \left( \inf_{\lambda \in \Lambda(\nu_1, \nu_2)} \int_{\mathbb{R}^2} |x - y|^p \, \mathrm{d}\lambda(x, y) \right)^{\frac{1}{p}}$$

*Let $\eta_1, \eta_2 \in \mathscr{P}_p(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}$. We also define the supremum-p-Wasserstein distance $\overline{d}_p$ by:*

$$\overline{d}_p(\eta_1, \eta_2) = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} d_p(\eta_1^{(x,a)}, \eta_2^{(x,a)})$$

We also have another expression to compute that distance.

**Lemma 1** ([2])**.** *Let $\nu_1, \nu_2 \in \mathscr{P}_p(\mathbb{R})$ with respective cumulative distribution $F$ and $G$. Let $\mathcal{U}$ be a uniform random variable on $[0, 1]$. then*

$$d_p(\nu_1, \nu_2) = \left\| F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U}) \right\|_p$$

*which, in the case $p < \infty$ simplifies to:*

$$d_p(\nu_1, \nu_2) = \left( \int_0^1 \left| F^{-1}(u) - G^{-1}(u) \right|^p du \right)^{\frac{1}{p}}$$

the $w_1$ is the most used, and has a dual form [to invertigate for quantiles]

### Cramer Distance:

**Definition 6** ([1])**.** *Let $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$. We define the family of metrics $\ell_p$ by :*

$$\ell_p(\nu_1, \nu_2) = \left( \int_{\mathbb{R}} (F_{\nu_1}(x) - F_{\nu_2}(x))^p \mathrm{d}x \right)^{\frac{1}{p}}$$

*$\ell_2$ is called the Cramer distance.*
*We also define the suppremum version of the $\ell_p$ norm:*

$$\overline{\ell}_p(\eta_1, \eta_2) = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \ell_p(\eta_1^{(x,a)}, \eta_2^{(x,a)})$$

**Remark 1.** $l_1 = d_1$

*Proof.* argument de symétrie de graphe □

## 2.3 Distributional Reinforcement Learning

In 2017, Bellemare et al. introduces the Distributional Reinforcemennt Learning framework. The idea is to compute the full distribution of the return instead of just the expected return. In his paper, they introduce the distributional Bellman operators and prove theoretical results on their properties.

The random return is the sum of the discounted random reward:

$$Z(x,a) = \sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x, A_0 = a \tag{5}$$

The idea is that the distribution of the reward would follow similar Bellman equations:

$$Z(x,a) \overset{\mathrm{D}}{=} R(x,a) + \gamma Z(X', A') \tag{6}$$

with $X', A'$ the random next state-action.

**Policy Evaluation**

Let's consider a policy $\pi$. The distribution of the random return under $\pi$ will be written as follows:

$$\eta_\pi^{(x,a)} = \mathrm{Law}_\pi \left( \sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x, A_0 = a \right)$$

and we will write $\eta_\pi$ as the collection of distribution $(\eta_\pi^{(x,a)})_{(x,a)\in\mathcal{X}\times\mathcal{A}}$.

What makes the distributional framework worth studying, is the generalization of the Bellman equation and its properties: The random return associated to policy $\pi$ verifies the *distributional Bellman equation*:

$$\eta_\pi = \mathcal{T}^\pi \eta_\pi$$

where $\mathcal{T}^\pi$ is the Bellman operator defined by:

$$\mathcal{T}^\pi \eta^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a')} \pi(a'|x') p(r,x'|x,a) dr$$

with $(f_{r,\gamma})_\#\eta$ is the pushforward measure define by $f_\#\eta(A) = \eta(f^{-1}(A))$ for all Borel sets $A \subseteq R$ and $f_{r,\gamma}(x) = r + \gamma x$ for all $x \in R$.

*Proof.* [faire la preuve] □

While this operator seems more cumbersome that the non-distributional one, it just comes down to rewriting equation 6 for distributions. The proof use the exact same idea as in the non-distributional case, but in this new formalism.

In the tabular case, it is possible to solve this fixed point equation by matrix inversion. However, it doesn't seem possible to do so when dealing with distribution. To solve it, we will use the following result, that is same used to solve the non-tabular non-distributional case.

Similarly as in the non-distributional case, this operator is a $\gamma$-contraction under a well chosen metric: the maximal $p$-Wasserstein metric $\overline{d}_p$ (for $p \geq 1$).

*Preuve:* [faire la preuve] □

This result is very important in the sense where it gives an theoretical algorithm to compute the value distribution of a policy. [+ distributional augments classical RL]

$$\forall \eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}, \quad (\mathcal{T}^\pi)^n \eta \underset{n\to\infty}{\longrightarrow} \eta_\pi$$

The Wasserstein metric is important here because the same operator is not always a contraction under the total variation distance, the Kolmogorov distance or the Kullback-Liebler divergence. (ref in article de Bellemare)

Even though this algorithm seems promising, there are several issues that arise in practice, that make it difficult to implement: It is impossible to represent exactly all the space of distributions,

which requires a parametrisation of the distribution and a projection step, then we most of the time don't have access to the exact transition of the MDP, which requires a stochastic estimation of the Bellman Operator. Those issues will be tackled in the next subsections.

### Control

Here, the goal still is to find an optimal policy. However, we will consider the full distributions of return to reach it.

We define by optimal distribution a distribution associated to an optimal policy: $\eta^\star \in \{\eta_{\pi^\star} \mid \pi^\star \in \arg\max_\pi \mathbb{E}_{R\sim\eta_\pi}[R]\}$. One of the first difference that we notice is the fact that there can be several different optimal distribution. Those optimal distribution all have the same mean, but a distribution having the optimal mean may not be an optimal distribution, because some distributions may not come from any (stationary) policy. [mettre les exemples]

As expected, the optimal distributions verify the optimal distributional Bellman equation: $\eta^\star = \mathcal{T}\eta^\star$ with

$$\mathcal{T}\eta^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a^\star(x'))} p(r,x'|x,a)dr$$

where $a^\star(x') = \arg\max_{a'\in\mathcal{A}} \mathbb{E}_{R\sim\eta^{(x',a')}}[R]$

*Proof.* [insert proof here] □

The first interesting Control result is the fact that this operator is a contraction in average:

**Lemma 2.** *Let $\eta_1, \eta_2 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we write $\mathbb{E}[\eta] := \mathbb{E}_{Z\sim\eta}[Z]$. Then:*

$$\|\mathbb{E}[\mathcal{T}\eta_1] - \mathbb{E}[\mathcal{T}\eta_2]\|_\infty \leq \gamma \|\mathbb{E}[\eta_1] - \mathbb{E}[\eta_2]\|_\infty$$

*Which means that $\mathbb{E}[\mathcal{T}^n\eta] \underset{n\to\infty}{\longrightarrow} Q^\star$ exponentially quickly.*

*Proof.* to redact □

As before this leads to a theoretical algorithm to find the optimal value function using the whole distribution. We have another result regarding the convergence of the distribution itself:

**Theorem 1.** *Let $\mathcal{X}$ and $\mathcal{A}$ be finite. Let $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. Assume that there is a single policy $\pi^\star$. Then:*

$$\mathcal{T}^n \underset{n\to\infty}{\longrightarrow} \eta_{\pi^\star} \text{ uniformly in } \overline{d}_p,\ p \geq 1$$

*Proof.* see Bellemare □

This theorem is very important to understand how distributions behave. At first it seems really convenient, with the convergence of the distribution, but there are especially two points which are worth emphasizing. The first one is that there is no exponential convergence anymore and, in fact, the speed of the convergence is unknown. The second is the condition of unicity of optimal policy. While this condition seems reasonable, it is still possible to do without, at the cost of stationarity: if there are several optimal policy, the distribution converges uniformly to one associated to a possibly nonstationary optimal policy.

The non stationarity of the optimal policy isn't an issue when the goal is solely to maximize the mean reward, as the greedy policy associated to its distribution will still be optimal. However, it can be more problematic if we try to find policy that takes acount of the whole distribution, such a safer or riskier policy. [ajuster avec papier de Achab et Neu]

The two previous properties are weaker that what we found in the Policy Evaluation case. To emphasize more on the differences, here are some more results that underline the pathologic cases that arise in Distributional Control:

**Proposition 1.** *The optimality operators are not always contractions.*

[look for the argument for any metric]

**Proposition 2.** *The optimality operators do not always have fixed points.*

[insert contre exemple here]

The lack of contraction is the precise result that prevents us to get the same properties as in the non-distributional case or as in the distributional, especially the existence and unicity of a fixed point, and the exponential convergence.

[detailler un peu plus] [faire l'analyse plus poussée du contre-exemple ? pour améliorer le résultat en rajoutant certaines contraites ?]

## 2.4 Distribution Parametrization

One of the main issue when dealing with distribution in practice, is the question of representation. It is not possible for a computer to represent the full extent of the distribution space. It is then necessary to restrain ourself on a parametrized space.

In their papers, [3, Morimura et al.] decide to parametrize the return distribution as a Gaussian, a Laplace or a skewed Laplace distribution. Later, [2, Bellemare et al.] and then [4, Dabney et al.] developed the theory for a richer class of parametric distributions, discrete ones, that are much more convenient. There two main approaches for that: the categorical approach, and the quantile regression approach.

### Categorical

This is the approach introduced by [2, Bellemare et al.] which led to the C51 algorithm that reached state of the art result for ALE. However, the theoretical properties of such approach were mainly developed later, by [1, Rowland et al.].

The idea is to use the hypothesis of bounded reward to use evenly spread diracs on that reward support, and use the diracs weight as the parameters.

[illustration]

More formally, let's consider $V_{\text{MIN}}, V_{\text{MAX}}$ the bounds for the reward, $N$ the number of diracs (the resolution) to use, $\Delta z = \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}$ the steps between diracs. The support of the distributions will be $\{z_i = V_{\text{MIN}} + i\Delta z \mid 0 \leq i < N\}$. The parametric family then is $\{\sum_{i=0}^{N-1} q_i \delta_{z_i} \mid \sum_{i=0}^{N-1} q_i = 1,\ 0 \leq q_i \leq 1\}$.

We define the projection operator $\Pi_C : \mathscr{P}(\mathbb{R}) \to \mathcal{P}_C$ by :

$$\Pi_C(\delta_y) = \begin{cases} \delta_{z_0} & y \leq z_0 \\ \frac{z_{i+1}-y}{z_{i+1}-z_i}\delta_{z_i} + \frac{y-z_i}{z_{i+1}-z_i}\delta_{i+1} & z_i < y < z_{i+1} \\ \delta_{z_{N-1}} & y \geq z_{N-1} \end{cases} \tag{7}$$

[explain what is the idea behind it + illustration]

Bellemare et al. introduced this projection step as an heuristic, without any theoretical motives or results related to the Wasserstein metric. It is Rowland et al. that later, found deep connection between this projection and another metric: the Cramer distance.

In fact, for the Wasserstein metric, we have the following result.

**Proposition 3.** $\Pi_C \mathcal{T}^\pi$ *is not a contraction for* $\overline{d}_p$ *with* $p > 1$.

*Proof.* to copy □

For the case $p = 1$ it is however true, but only because it is the same as the $\ell_p$ distance, for which we have much more properties:

**Proposition 4.** *For a specific subset of $\mathscr{P}(R)$ and appropriate Hilbert space structure with $\ell_2$, $\Pi_C$ is the orthogonal projection of that subset onto $\mathcal{P}_C$*

[projection particularly relevant + relevance of the other metric]

**Proposition 5.** *$\Pi_C \mathcal{T}^\pi$ is a $\sqrt[p]{\gamma}$-contraction in $\overline{\ell}_p$.*

*Proof.* to do □

The Banach fixed point theorem thus provides with a proof a convergence of the iterated projected Bellman operators:

$$\exists! \eta_C \in \mathcal{P}_C^{\mathcal{X}\times\mathcal{A}},\ \forall \eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}, \quad (\Pi_C \mathcal{T}^\pi)^m \eta_0 \underset{m\to\infty}{\longrightarrow} \eta_C \quad \text{exponentially quickly in } \overline{\ell}_p \tag{8}$$

It is important to notice that this does not have to converge to $\eta_\pi$, for the simple reason that this operator is convergent in the parametrized space $\mathcal{P}_C$, which may not contain $\eta_\pi$. The question that arises next, is how well does $\eta_C$ approximates $\eta_\pi$.

**Lemma 3.** *Let $\eta_C$ defined as in (8). Assume that $\eta_\pi$ is supported on $[z_0, z_{N-1}]$. Then:*

$$\overline{\ell}_2(\eta_C, \eta_\pi) \leq \frac{1}{1-\gamma}\Delta z$$

[result that we want + increase in resolution get us closer][case when no garantuee on the reward]

### Quantile regression

This approach was first introduced by [4, Dabney et al.] and led to the QR-DQN algorithm that outperformed C51.

The idea is to do the opposite of the categorical approach: instead of having fixed reward support with variable weight, it considers fixed weight for variable support.

[illustration]

More formally, let's consider $N$ the resolution. The parametric family is $\{\frac{1}{N}\sum_{i=0}^{N-1} \delta_{z_i} \mid (z_i) \in \mathbb{R}^n\}$

As the Wasserstein metric seems to be a metric of choice for this framework, it seems natural to try to project a distribution on the parametrized space by minimizing the Wasserstein distance between both. In this subsection we will use the 1-Wasserstein distance. The projection operator $\Pi_{d_1} : \mathscr{P}(\mathbb{R}) \to \mathcal{P}_Q$ is thus defined by:

$$\Pi_{d_1}\nu = \underset{\nu_Q \in \mathcal{P}_Q}{\arg\min}\, d_1(\nu, \nu_Q) \tag{9}$$

This is actually possible to compute, and the minimizers are exactly :

$$\Pi_{d_1}\nu = \frac{1}{N}\sum_{i=0}^{N-1} \delta_{z_i}, \quad F_\nu(z_i) = \frac{2i+1}{2N}$$

where $F$ is the cumulative distribution funcition of $\nu$.

*Proof.* to copy □

[illustration/exemple]

**Proposition 6.** *$\Pi_{d_1}\mathcal{T}^\pi$ is $\gamma$-contraction in $\overline{d}_\infty$ :*

$$\overline{d}_\infty(\Pi_{d_1}\mathcal{T}^\pi\eta_1, \Pi_{d_1}\mathcal{T}^\pi\eta_2) \leq \gamma\overline{d}_\infty(\eta_1, \eta_2)$$

*Proof.* □

[comment (about the implications of the results and about the norm)]

[faire la comparaison entre des deux : quantile regression utilise moins de paramètres et de conditions, mais les résultats sont pas exactement les mêmes.]

### Diatomic AVaR

[Achab et Neu, motives about keeping the mean]

## 2.5 Quantile Optimization

Issues with greedy policy for implementation (Defourny 2008)

### Quantile

[Morimura et al.]

### Superquantile

[results by Achab et Neu]

## 2.6 unformal Distributional Approach (with random variable)[TO REMOVE]

**Policy Evaluation:** We here want to look at the full distribution of the return when following a certain policy $\pi$. We define the random return as follow (recall that $R$ is a stochastic reward):

**Definition 7** (Value distribution function)**.** *The random value function associated with policy $\pi$ is defined as follow:*

$$\mathcal{V}^\pi(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \qquad x_0 = x$$

$$\mathcal{Q}^\pi(x,a) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \qquad x_0 = x, a_0 = a$$

*with $x_t \sim p(\cdot|x_{t-1}, a_{t-1})$ and $a_t \sim \pi(\cdot|x_t)$*

By doing the same computation as in the expected reward case, we notice that the value distribution functions verifies an extended version of the Bellman equation:

$$\mathcal{V}^\pi(x) = \sum_{a\in\mathcal{A}} \pi(a|x) \left( R(x,a) + \gamma \sum_{x'\in\mathcal{X}} p(x'|x,a)\mathcal{V}^\pi(x') \right) \tag{10}$$

$$\mathcal{Q}^\pi(x,a) = R(x,a) + \gamma \sum_{x',a'\in\mathcal{X}\times\mathcal{A}} p(x'|x,a)\pi(a'|x')\mathcal{Q}^\pi(x',a') \tag{11}$$

This leads to the distributional Bellman operator:

**Definition 8** (Distributional Bellman Operator)**.**

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\pi\mathcal{V}(x) = \sum_{a\in\mathcal{A}} \pi(a|x) \left( R(x,a) + \gamma \sum_{x'\in\mathcal{X}} p(x'|x,a)\mathcal{V}(x') \right)$$

$$\forall x,a \in \mathcal{X}\times\mathcal{A}, \qquad \mathcal{T}^\pi\mathcal{Q}(x,a) = R(x,a) + \gamma \sum_{x',a'\in\mathcal{X}\times\mathcal{A}} p(x'|x,a)\pi(a'|x')\mathcal{Q}(x',a')$$

Due to the distribution persepective, it is not possible to solve for the fixed point equation with a matrix inversion anymore. But the following result still enable theoretical computation of the exact random value distribution:

**Lemma 4.** *$\mathcal{T}$ is a gamma-contraction in $d_p$*

leads to an algorithm to find the compute the distribution

**Control:**

**Definition 9** (Optimal Value distribution function)**.**

$$\mathcal{V}^\star \in \{\mathcal{V}^{\pi^\star} \in arg\max_\pi \mathbb{E}\left[\mathcal{V}^\pi\right]\}$$

$$\mathcal{Q}^\star \in \{\mathcal{Q}^{\pi^\star} \in arg\max_\pi \mathbb{E}\left[\mathcal{Q}^\pi\right]\}$$

This distribution also verify an extended version of the Optimal Bellman Equation:

$$\mathcal{V}^\star(x) = R(x, a^\star(x)) + \gamma \sum_{x'\in\mathcal{X}} p(x'|x, a^\star(x))V^\star(x')$$

$$\mathcal{Q}^\star(x,a) = R(x,a) + \gamma \sum_{x'\in\mathcal{X}} p(x'|x,a)Q^\star(x', a^\star(x))$$

and leads to the optimal bellman operator:

**Definition 10** (Optimal Distributional Bellman Operator)**.**

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\star \mathcal{V}(x) = R(x, a^\star(x)) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a^\star(x)) V(x')$$

$$\forall x, a \in \mathcal{X} \times \mathcal{A}, \qquad \mathcal{T}^\star \mathcal{Q}(x, a) = R(x, a) + \gamma \sum_{x' \in \mathcal{X}} p(x'|x, a) Q(x', a^\star(x))$$

However, this operator is not a contraction (see Bellemare 2017). But we still have some practical theoretical results: contraction in mean. Also, convergence in sequence of optimal policy (details ?)

# 3 Travail de recherche et résultats

## 3.1 Quantile

In this part, we are going to change the goal, and try to maximise a quantile of the distribution, instead of the mean. Indeed, sometimes we may have specific applications where the mean doesn't matter so much, but where it is very important to have a safe policy, i.e. to have higher quantiles, even at the cost of lower mean. In the same way, it can be interesting to find more risky behavior on some environment, with higher possible reward, but at the cost of failing more often.

Quantile Optimization is a topic well studied in finance, in the framework of portfolios. However very few tackled this issue in the case of MDP. In their paper [3, Morimura et al.] try to apply their first distributional approach on an Q-learning algorithm trying to optimise quantiles. Even though they obtained empirically promising results, no theoretical results have been obtained se far. A main reason for that is because quantiles are particularly hard to compute. Fearless, we will still try to tackle this topic.

First it is important to notice that, as the goal changed, many assumptions that were made in the original case are to be studied again in this case. The theory as to be redone from 0.

### Framework

We are still considering MDPs of the form $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, R, \gamma)$, but with another value to optimize. We consider $x \in \mathcal{X}$ a specific state, and $\tau \in [0, 1]$ the quantile of interest. Our objective is:

$$\max_{\pi} V_\tau(x) = q_\tau \left( \sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x \right)$$

In the *average* framework, we try to optimize for every single state and action. However, in this new case, it is not possible as optimizing for a specific state may require to lower the quantiles for the next states.

insert contre exemple

This first result is particularly probematic since every method previously used would mainly profit of this property with the mean, optimizing every state separately.

### Policy Evaluation

The first question, just as before, is how to evaluate a policy. Let $\pi$ a policy. We want to compute:

$$Q_\tau(x, a) = q_\tau \left( \sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x, A_0 = a, \pi \right)$$

When working with the mean, we would profit of its linearity to find an equation verified by this quantity, and solve this equation. Here, there is no linearity. In fact, is not even possible to compute a quantile solely knowing the quantiles for the next statest and action. We require the full distribution of the reward and the full distribution of reward and next state-action return.

Luckily, with the development of the distributional approach, we have a way to compute the full distribution of the return for a specific policy. And once the distribution is known, so is the quantile.

In the case where it only requires a finite number of (distributional) Bellman operator application to get the exact distribution, we get to compute the quantile exactly. This happens for instance in MDP that ends after a finite number of steps.

if

$$\exists k \in \mathbb{N},\ \forall \eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}, (\mathcal{T}^\pi)^k \eta_0 = \eta_\pi$$

then

$$\forall \eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}, q_\tau\left((\mathcal{T}^\pi)^k \eta_0\right) = q_\tau(\eta_\pi)$$

In the general case, even though we have the convergence of the distribution,it may not be enough for the convergence of the quantile.

$$(\mathcal{T}^\pi)^n \eta \underset{n\to\infty}{\longrightarrow} \eta_\pi \quad \not\Rightarrow \quad q_\tau\left((\mathcal{T}^\pi)^n \eta\right) \underset{n\to\infty}{\longrightarrow} q_\tau(\eta_\pi)$$

We would need at least point-wise convergence of the cumulative distribution function.

[counter exemple for Wasserstein metric]

### Control

different questions arise: deterministic policy (surement oui, mais trouver une preuve? bellman equation ? easy way to policy evaluate except by computing the whole distribution ? Existance of a optimal policy for every state ? How to control (take max on what ?) ? Une policy iteration augmente forcément le q10 ? petit résultat: pour un simple mdp: tout un ensemble de policy optimal, dont une deterministic, mais si on veut en plus maximiser la mean, il faut du undeterministic

### Quantiles and distribution parametrization

The 1-Wasserstein metric (equivalent to $\ell_1$) works well with as minimizing it leads to the quantiles of the distribution

# 4 Conclusion et perspectives

# 5 Bibliographie

# 6 Annexes éventuelles.

# References

[1] M. Rowland, M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh, "An Analysis of Categorical Distributional Reinforcement Learning," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 29–37, PMLR, Mar. 2018. ISSN: 2640-3498.

[2] M. G. Bellemare, W. Dabney, and R. Munos, "A Distributional Perspectiv e on Reinforcement Learning," *arXiv:1707.06887 [cs, stat]*, July 2017. arXiv: 1707.06887.

[3] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka, "Parametric Return Density Estimation for Reinforcement Learning," *arXiv:1203.3497 [cs, stat]*, Mar. 2012. arXiv: 1203.3497.

[4] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional Reinforcement Learning with Quantile Regression," *arXiv:1710.10044 [cs, stat]*, Oct. 2017. arXiv: 1710.10044.

[5] M. Achab and G. Neu, "Robustness and risk management via distributional dynamic programming," *arXiv:2112.15430 [cs, math]*, Dec. 2021. arXiv: 2112.15430.

[6] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *29th Annual Conference on Learning Theory* (V. Feldman, A. Rakhlin, and O. Shamir, eds.), vol. 49 of *Proceedings of Machine Learning Research*, (Columbia University, New York, New York, USA), pp. 998–1027, PMLR, 23–26 Jun 2016.

[7] C. Lyle, M. G. Bellemare, and P. S. Castro, "A Comparative Analysis of Expected and Distributional Reinforcement Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4504–4511, July 2019. Number: 01.