ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ETH
Eidgenössische Technische Hochschule Zü
Swiss Federal Institute of Technology Zuric

Prof. Name

Master Thesis

LaTeX -Latex Template

Spring Term 2020

Supervised by:
Supervisor 1

Author:
Nuria Armengol Urpi

Contents

Preface

Bla bla . . .

Abstract

Bla bla …

Symbols

Symbols

ϕ, θ, ψ	roll, pitch and yaw angle

Indices

x	x axis
y	y axis

Acronyms and Abbreviations

ETH	Eidgenössische Technische Hochschule

Chapter 1

Introduction

Introduction

1.1 Distributional RL

Recent research has been done demonstrating the importance of learning the value distribution, i.e, the distribution of the random return received by a RL agent. This differs from the common RL approach which is focused on learning the expected value of this return.
One of the major goals of RL is to teach an agent so that it learns how to act so that it maximizes its expected utility, Q [1] Bellman's equation describes this value Q in terms of the expected reward and expected outcome of the random transition $(x, a) \to (X', A')$, showing the particular recursive relationship between the value of a state and the values of its successor states:

$$Q(x, a) = \mathbb{E}[R(x, a)] + \gamma \mathbb{E}[Q(X', A')] \tag{1.1}$$

Distributional RL aims to go beyond the notion of *value* and training to study instead the random return Z.

1.1.1 Example showing interest in learning the distribution

Imagine the example in which we are playing a board game and we roll 2 dices. If we get a 3, we fall in prison and need to pay 2000CHF (ie reward of -2000CHF), whereas otherwise we we collect a salary of 200CHF (ie reward of +200CHF). If we consider the common reinforcement learning approach and we compute the expected immediate ($\gamma = 1$) reward:

$$\mathbb{E}[R(x)] = \frac{1}{36}(-2000\,\mathrm{CHF}) + \frac{35}{36}(200\,\mathrm{CHF}) = 138.88\,\mathrm{CHF} \tag{1.2}$$

Hence, the expected immediate return is +138.88CHF. However, in any case we will get a return of +138.88CHF. Instead:
$R(x) = \begin{cases} -2000\,\mathrm{CHF}, & \text{w.p } \frac{1}{36} \\ 200\,\mathrm{CHF}, & \text{w.p } \frac{35}{36} \end{cases}$

We define the random return $Z^{\pi}(x,a)$ as the random variable that represents the sum of discounted rewards obtained by starting from position x taking action a and thereupon following policy π.
This variable captures intrinsic randomness from:

1. Immediate rewards
2. Stochastic dynamics
3. Possibly an stochastic policy

Having defined $Z^{\pi}(x,a)$, we can clearly see that:

$$Q^{\pi}(x,a) = \mathbb{E}[Z^{\pi}(x,a)] \tag{1.3}$$

Z is also described by a recursive equation, but of a distributional nature:

$$Z^{\pi}(x,a) \overset{D}{=} R(x,a) + \gamma Z(x',a') \tag{1.4}$$

where $x' \backsim p(\cdot|x,a)$ and $a' \backsim \pi(\cdot|x')$
where $\overset{D}{=}$ denotes that the RV on both sides of the equation share the same probability distribution. The *distributional Bellman equation* defined in (1.4), states that the distribution of Z is characterized by the interaction of 3 RV's: the random variable reward R, the next state-action (X',A') and its random return Z(X',A'). From here on, we will view Z^{π} as a mapping from state-action pairs to distributions over returns, and we call this distribution the *value distribution.*

1.1.2 Distributional Bellman Operator

In the policy evaluation setting [1], one aims to find the value function $V\pi$ associated with a given fixed policy π. In the distributional case, we aim to find $Z\pi$. [2] defined the Distributional Bellman operator T^{π}. We view the reward function as a random vector R $\in \mathbb{Z}$ and define the transition operator $P^{\pi} : \mathbb{Z} \to \mathbb{Z}$

$$P^{\pi}Z(x,a) \overset{D}{=} Z(X',A') \tag{1.5}$$
$$X' \backsim P(\cdot|x,a) \text{ and } A' \backsim \pi(\cdot|X') \tag{1.6}$$

where we use capital letters to emphasize the random nature of the next state-action pair (X',A') Then, the Distributional Bellman operator T^{π} is defined as:

$$T^{\pi}Z(x,a) \overset{D}{=} R(x,a) + \gamma P^{\pi}Z(x,a) \tag{1.7}$$

[2] showed that (1.7) is a contraction mapping whose unique fixed point is the random return Z^{π}.

Wasserstein metric:

The p-Wasserstein metric W_p, for $p \in [1,\infty]$, also known as the Earth Mover's Distance when $p=1$ is an integral probability metric between distributions. The p-Wasserstein distance is characterized as the L^p metric on inverse cumulative distribution functions (CDF). Tht is, the p-Wasserstein metric between distributions U and Y is given by:

$$W_p(U,Y) = \Big(\int_0^1 |F_Y^{-1}(w) - F_U^{-1}(w)|^p dw\Big)^{\frac{1}{p}} \tag{1.8}$$

where for a random variable Y, the inverse CDF F_Y^{-1} of Y is defined by:

$$F_Y^{-1}(w) := \inf\{y \in \mathbb{R} \,|w \;\leq F_Y(w)\} \tag{1.9}$$

where $F_Y(w) = Pr(y \leq Y)$.

Add Figure 2.1 in [3]

Unlike the Kullback-Leibler divergence , the Wasserstein metric is a true probability metric and considers both the probability of and the distance between various outcome events, which makes it well-suited to domains where an underlying similarity in outcome is more important than exactly matching likelihoods.

Contraction in $\hat{d}_p$:

Let $\mathcal{Z}$ be the space of action-value distributions:

$$\mathcal{Z} = \{Z \mid \mathcal{X} \times \mathcal{A} \to \wp(\mathbb{R}) \tag{1.10}$$

$$\mathbb{E}[|Z(x,a)|^p < \infty, \forall(x,a), p \geq 1]\} \tag{1.11}$$

check first line the $\wp$

Then, for two action-value distribution $Z_1, Z_2 \in \mathcal{Z}$, the maximal form of the Wasserstein metric is defined by:

$$\hat{d}_p(Z_1, Z_2) := \sup_{x,a} W_p(Z_1(x,a), Z_2(x,a)) \tag{1.12}$$

[2] showed that $\hat{d}_p$ is a metric over value distributions and furthermore, the distributional Bellman operator T^π is a contraction in $\hat{d}_p$. Consider the process $Z_{k+1} := T^\pi Z_k$, starting with some $Z_0 \in \mathcal{Z}$.
$T^\pi Z : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in the Wasserstein metric $\hat{d}_p$, which implies that not only the first moment (expectation) converges exponentially to Q^π, but also in all moments.
Lemma 1: (Lemma 3 in [2])
T^π is a γ-contraction: for any two $Z_1, Z_2 \in \mathcal{Z}$,

$$\hat{d}_p(T^\pi Z_1, T^\pi Z_2) \leq \gamma \hat{d}_p(Z_1, Z_2) \tag{1.13}$$

Using Banach's fixed point theorem, it is proven that T^π has a unique fixed point, which by inspection must be Z^π.
Hence the $\hat{d}_p$ metric is shown to be useful metric for studying behavior of distributional RL algorithms, and to showed their convergence to a fixed point. Moreover, shows than en effective way to learn a value distribution is to attempt minimize the Wasserstein distance between a distribution Z and its distributional Bellman update $T^\pi Z$, analogously to the way that TD-learning attempts to iteratively minimize the L^2 distance between Q and TQ.
We have so fare considered a policy evaluation setting, ie trying to learn a value distribution for a fixed policy π, and we studied the behavior of its associated distributional operator T^π. In the control setting, ie, when we try to find a policy π^* that maximizes a value, or its distributional analogous, ie that induces an optimal value distribution. However, while all optimal policies attain the same value Q^*, in general there are many optimal value distributions.
The distributional analogue of the Bellman optimality operator converges, in a weak sense, to the set of optimal value distributions, but this operator is *not a contraction in any metric between distributions.*.
Let Π^* be the set of optimal policies.
Definition 1: An optimal value distribution is the value distribution of an optimal policy. The set of optimal value distributions is

$$\mathcal{Z}^* := \{Z^{\pi^*} | \pi^* \in \Pi^*\}$$

Not all value distributions with expectation Q^* are optimal, but they must match the full distribution of the return under some optimal policy. **Definition 2:** A greedy policy π for Z $in \mathcal{Z}$ maximizes the expectation of Z. The set of greedy policies for Z is:

$$\mathcal{G}_{\mathcal{Z}} := \{\pi \mid \sum_{a} \pi(a|x)\mathbb{E}(Z(x,a) = \max_{a' \in \mathcal{A}} Q(x',a')\}$$

We will call a *distributional Bellman optimality operator* any operator $\mathcal{T}$ which implements a greedy selection rule, ie:

$$\mathcal{T}Z = \{\mathcal{T}^{\approx}Z \text{ for some } \pi \in \mathcal{G}_{\mathcal{Z}}\}$$

As in the policy evaluation setting, we are interested in the behavior of the iterates $Z_{k+1} := \mathcal{T}Z_k, Z_O \in \mathcal{Z}$. Lemma 4 in [2] shows that $\mathbb{E}Z_k$ behaves as expected: **Lemma 4:** Let $Z_1, Z_2 \in \mathcal{Z}$. Then:

$$\|\mathbb{E}\mathcal{T}Z_1 - \mathbb{E}\mathcal{T}Z_2\|_\infty \leq \gamma\|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty$$

and in particular $\mathbb{E}Z_1 \to Q^*$ exponentially quickly. However, Z_k is not assured to converge to a fixed point. Specificially, they provide a number of negative results concerning $\mathcal{T}$:
Proposition 1: The operator $\mathcal{T}$ is not a contraction.
Proposition 2: Not all optimality operators have a fixed point $Z^* = \mathcal{T}Z^*$ **Proposition 3**: That $\mathcal{T}$ has a fixed point $Z^* = \mathcal{T}Z^*$ is insufficient to guarantee the convergence of $\{Z_k\}$ to Z^*
Another result, shows that we cannot in general minimize the Wasserstein metric, viewed as a loss, using stochastic gradient descent methods. This limitation, is crucial in a practical context, when the value distribution needs to be approximated.

1.1.3 Quantile approximatio

[3] used the theory of quantile regression [**?**], to design an algorithm applicable in a stochastic approximation setting. Quantile regression is used to estimate the quantile function at precisely chosen points. Then the Bellman update is applied onto this parameterized quantile distribution. This combined operator is proven to be a contraction and the estimated quantile function is shown to converge to the true value distribution when minimized using stochastic approximation.

1.1.4 Quantile Regression

Quantile regression is a method for approximating quantile functions of distributions. The quantile regression loss, for quantile $\tau \in [0,1]$, is an asymmetric convex lox function that penalizes underestimation errors with weight τ and overestimation errors with weight $1-\tau$. For a distribution Z, and given quantile τ, the value of the quantile function $F_Z^{-1}(\tau)$ may be characterized as the minimizer of the quantile regression loss:

$$\mathcal{L}^{\tau}_{QR}(\theta) = \mathbb{E}_{\hat{Z} \backsim Z}[\rho_\tau(\hat{Z} - \theta)] \tag{1.14}$$

$$h \tag{1.15}$$

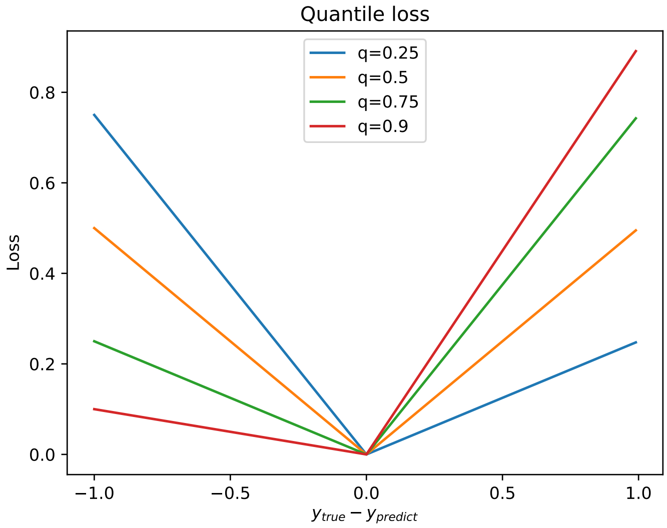

Figure 1.1: Car trajectory using DDPG algorithm and velocity penalization with probability P=0.2

1.2 IDEA AND CAR EXPERIMENT

We focus on the importance of *value distribution*, the distribution of the random return received by a RL agent, in contrast to the common approach in RL of modelling the expectation of this return. The latter neither takes int account the variability of the cost (i.e fluctuations around the mean), nor its sensitivity to modeling errors. [4]
We aim to learn this distribution and try to minimize other metrics rather than its mean, which can be crucial for some environments in which ensuring that the cost is always above a certain value with certain probability is crucial.
A metric that has recently gained a lot of popularity is the Conditional Value at Risk, eg in finance, due to its favorable computation properties and superior ability to safeguard a decision maker from the "outcomes that hurt the most" [5]

1.2.1 Conditional Value-at-Risk (CVaR)

Let Z be a bounded-mean random variable, i.e $\mathbb{E}[|Z|] < \infty$, on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, with cumulative distribution function $F(z) = \mathbb{P}(Z \leq z)$. We interpret Z as a reward. The value-at-risk (VaR) at confidence level $\alpha \in (0,1)$ is the α quantile of Z, i.e, $\mathrm{VaR}_\alpha(Z) = \inf\{z \mid F(z) \geq \alpha\}$. The conditional value-at-risk (CVaR) at confidence level $\alpha \in (0,1)$ is defined as the expected reward of outcomes worse than the α-quantile (VaR_α):

$$\mathrm{CVaR}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha F_Z^{-1}(\beta)d\beta = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(Z)d\beta \tag{1.16}$$

Rockafellar and Uryasev [6] also showed that CVaR is equivalent to the solution of the following optimization problem:

$$\mathrm{CVaR}_\alpha(Z) = \min_\nu \big\{\nu + \frac{1}{\alpha}\mathbb{E}_Z[[Z-\nu]^-]\big\} \tag{1.17}$$

where $(x)^- = \min(x,0)$. In the optimal point it holds that $\nu^* = \mathrm{VaR}_\alpha(Z)$.
A useful property of CVaR, is its alternative dual representation [7]:

$$\mathrm{CVaR}_\alpha(Z) = \min_{\xi \in U_{\mathrm{CVaR}}(\alpha,\mathbb{P})} \mathbb{E}_\xi[Z] \tag{1.18}$$

where $\mathbb{E}_\xi[Z]$ denotes the ξ-weighted expectation of Z, and the risk envelope U_{CVaR} is given by:

$$U_{\mathrm{CVaR}}(\alpha,\mathbb{P}) = \Big\{\xi | \xi(w) \in \big[0, \frac{1}{\alpha}\int_{w\in\Omega} \xi(w)\mathbb{P}(w)dw = 1\big]\Big\} \tag{1.19}$$

Thus, the CVaR of a random variable may be interpreted as the worst case expectation of Z, under a perturbed distribution $\xi\mathbb{P}$.

Bibliography

[1] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, 1998.

[2] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *34th International Conference on Machine Learning, ICML 2017*, 2017.

[3] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018.

[4] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: A CVaR optimization approach," *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 1522–1530, 2015.

[5] G. Serraino and S. Uryasev, "Conditional Value-at-Risk (CVaR)," in *Encyclopedia of Operations Research and Management Science*, 2013.

[6] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[7] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, 1999.

Appendix A

Anything

Bla bla ...