Implicit Quantile Networks for Distributional Reinforcement Learning

When "risk preference" is considered in RL

Will Dabney, Georg Ostrovski, David Silver & Remi Munos

by Shijie Huang

Brain, Mind and Markets Lab

November 2018

Table of Contents

1. A Quick Recap
2. Risk Preference in Economics
3. Implicit Quantile Networks
4. Implementation

Table of Contents

1. A Quick Recap
2. Risk Preference in Economics
3. Implicit Quantile Networks
4. Implementation

Distributional Reinforcement Learning

Core idea: we consider a distribution $Z^{\pi}(x,a)$ over returns instead of merely the expectation $Q^{\pi}(x,a)$.

Distributional Bellman Equation

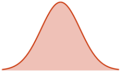
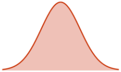
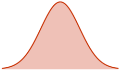
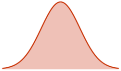
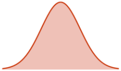
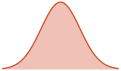
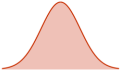
$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X^{'},A^{'})$$

where $X^{'} \sim P(.|x,a)$ and $A^{'} \sim \pi(.|x^{'})$

The rational behind is that we can obtain better estimates of $Q(x,a)$ (i.e. expected values)

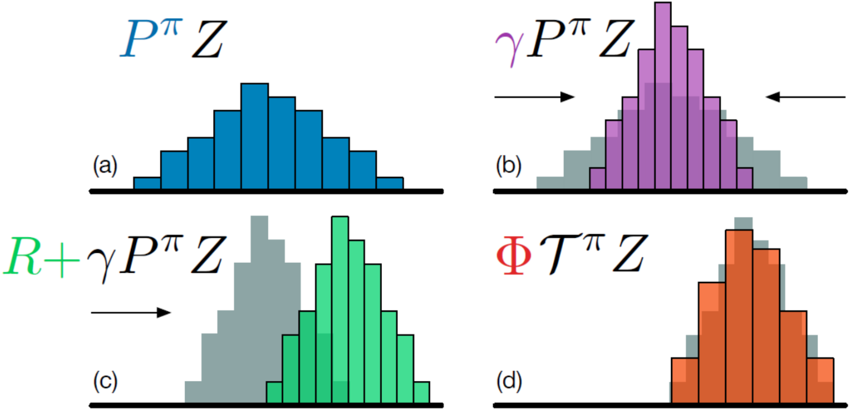
Distributional Reinforcement Learning

For discrete state-action case

Z(x, a)	S1	S2	S3
A1			
A2			
A3			

Distributional Reinforcement Learning

Fro each distribution $Z^{\pi}(x, a)$, note that Z in the later paper refers to quantile



Distributional Reinforcement Learning

A side note on Wasserstein Distance Metric:

- KL divergence, Wasserstein Metric, and Cramer Distance receive much attention recently. (Unsupervised Learning)
- Mostly apply to algorithms that intend to use probabilistic model to explain data, i.e. learn a probability distribution.
- e.g. Wasserstein Generative Adversarial Networks (W-GAN) (Arjovsky, Chintala, & Bottou, 2017)

Motivation of Implicit Quantile RL

"Their claim"

The problem: the conventional **policies** used was based entirely on the mean of return distribution.

Rationale for **new policy class**: can we expand the class of policies using information provided by the distribution over returns (i.e. to the class of risk-sensitive policies)?

Remarks

Before I started reading the rest of the paper, my thoughts were:

- So they want to expand to new policy classes, specifically, developing new policies that incorporate risk-preferences.
- Recall: policy is how we should choose action under a particular state. e.g. greedy policy.
- Towards the end, I find that their claim has puzzled me.

Motivation of Implicit Quantile RL

In short:

Previous paper: how we can have a better estimate of $Q(x, a)$

This paper: how we can have a better policy, specifically, a policy that incorporates risk-preferences (maybe?)

Table of Contents

1 A Quick Recap

2 Risk Preference in Economics

3 Implicit Quantile Networks

4 Implementation

von Neumann & Morgenstern Expected Utility Theory

Economists model risk as uncertainty (probability)

Expected Utility Theory (in RL context)

If a decision policy is consistent with a particular set of axioms [a] regarding its choices, then the decision policy behaves as if it is maximizing the expected value of some utility function U.

$$\pi(x) = \arg\max_{a} \mathbb{E}_{Z(x,a)}[U(z)]$$

Linear, convex, concave $U \Leftrightarrow$ risk-neutral, risk-seeking, risk-averse

[a] Axioms: Completeness, Transitivity, Continuity, Substitution, Independence

Paradoxical Behaviour

Independent Axiom

Given r.v. X, Y, Z, such that $X \geq Y$, any mixture between X and Z is preferred to the same mixture between Y and Z.

$$\alpha F_X + (1-\alpha) F_Z \geq \alpha F_Y + (1-\alpha) F_Z \quad \forall \alpha \in [0,1]$$ [a]

where F is the CDF.

[a] I think they made a mistake here, ≥ should be ≥

Often being criticized to consistently violate human behaviour: Allais paradox (Allais, 2008)

Paradoxical Behaviour

So the Expected Utility Theory do not accurately reflect human choices.

"This axiom (i.e. independent axiom) can be replaced by one in terms of convex combination of outcome values, instead of mixtures of distributions..."

Dual Theory of Choice Under Risk (Yaari, 1987)

Reasons to look at an alternative theory:

- Risk aversion and diminishing marginal utility of wealth are synonymous under the expected utility theory.
- Paradoxical behaviour that cannot be explained by the expected utility theory

Dual Theory of Choice Under Risk (Yaari, 1987)

Setup:

- decumulative distribution function of r.v. v [1]

$$G_v(t) = Pr\{v > t\}, \quad 0 \leq t \leq 1$$

$$F_v(t) = 1 - G_v(t) \quad \text{is the usual CDF}$$

- G_v is always non-increasing, right-continuous, and satisfies $G_v(1) = 0$, and

$$\forall v \in V, \quad \int_0^1 G_v(t)dt = \mathbb{E}\, v$$

[1] Can be interpreted as payments, denominated in some monetary unit, effectively, a gamble or a lottery

Dual Theory of Choice Under Risk (Yaari, 1987)

Theorem 1

A preference relation $\succeq$ satisfies Axioms $A1 - A5$ [a] iff there exists a continuous and non-decreasing real function f, defined on the unit interval, such that, for all u and v belonging to V,

$$u \succeq v \quad \Leftrightarrow \quad \int_0^1 f(G_u(t))dt \geq \int_0^1 f(G_v(t))dt$$

Function f, which is unique up to a positive affine transformation, can be selected in such a way that, for all p satisfying $0 \leq p \leq 1$, $f(p)$ solves the preference equation, $[1; p] \sim [f(p); 1]$

[a] Neutrality, Complete weak order, Continuity, Monotonicity, Dual independence

Dual Theory of Choice Under Risk (Yaari, 1987)

$[1; p]$ represents a r.v. that takes the values x and 0 with probabilities p and $1 - p$.

Function f effectively transform the $r.v.$ to an equivalent value with probability of 1.

"Certainty equivalent"

Dual Theory of Choice Under Risk (Yaari, 1987)

Axiom A5 Direct Dual Independence

Let u, v and w belong to V and assume that u, v and w are pairwise comonotonic [a]. Then, for every real number α satisfying $0 \leq \alpha \leq 1$, $u \succeq v$ implies $\alpha u + (1-\alpha)w \succeq \alpha v + (1-\alpha)w$

[a] set V of r.v. (s, Σ, P) is the underlying probability space. $\forall s$ and $s^{'}$ in S, $\frac{(u(s)-u(s^{'}))}{(v(s)-v(s^{'}))} \geq 0$

Independent Axiom in EUT and DT

Expected Utility Theory:

$$\alpha F_X + (1-\alpha) F_Z \geq \alpha F_Y + (1-\alpha) F_Z$$

- Mixture of distributions

Dual Theory of Choice:

$$\alpha u + (1-\alpha) w \geq \alpha v + (1-\alpha) w$$

- Convex combination of real functions
- This is a key theorem they employed.

Dual Theory of Choice Under Risk (Yaari, 1987)

How is "risk preference" characterized under the dual theory? ⇒ by the convexity of function f.

Theorem 2

Consider the class of preference relations on V satisfying Axiom $A1-A5$. A preference relation $\succeq$ in this class is risk averse iff the function f representing $\succeq$ (see theorem 1) is **convex** [a].

[a] In Expected Utility Theory, risk aversion ⇔ concave utility function

Will be clearer in the implementation part.

Dual Theory of Choice Under Risk (Yaari, 1987)

If f is differentiable, then the utility:

$$U(v) = \int_0^1 f(G_v(t))dt$$

can be written as [2]

$$U(v) = \int_0^1 tf^{'}(G_v(t))dF_v(t)$$

which is the core idea for the following slide

[2] Proof can be done via integration by part, see appendix

Distortion Risk Measure

Back to Implicit Quantile RL:

Under axioms of the dual theory of choice, the decision policy behaves as though it is maximizing a distorted expectation, for some continuous monotonic function h:

$$\pi(x) = \arg\max_{a} \int_{-\infty}^{\infty} z \frac{\partial}{\partial z} (h \circ F_{Z(x,a)})(z) dz$$

where function h is known as a **distortion risk measure**.

Distortion Risk Measure

Some remarks:

- The argument here is not really clear to me, I do not see the flow from the dual theory of choice to distorted expectation maximization.
- Dual theory is NOT "paradox free", in fact, Yaari specifically mentioned that there could be "dual paradoxical" behaviour that was entirely consistent with EUT.
- which raises a question: The "distortion expectation" story should be, in theory, no better than risk-sensitive RL that adopt utility function approach, but then why?

Table of Contents

1. A Quick Recap
2. Risk Preference in Economics
3. Implicit Quantile Networks
4. Implementation

Setup

Quantile Function for r.v. Z:

$$Z_\tau := F_Z^{-1}(\tau) \quad \tau \in [0,1]$$

Let $\beta : [0,1] \to [0,1]$ be a distortion risk measure, with identity [3] corresponding to risk-neutrality. Then the distorted expectation of $Z(x,a)$ under the distortion risk measure β is given by:

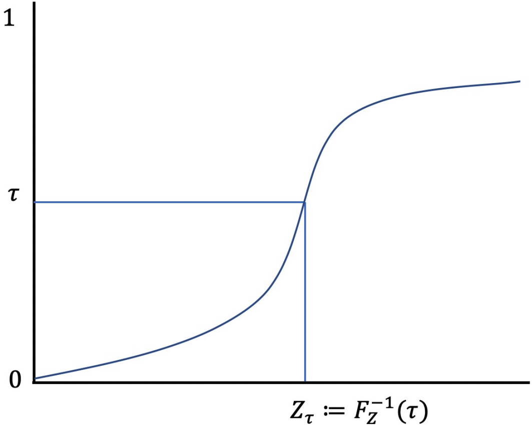
$$Q_\beta(x,a) := \mathbb{E}_{\tau \sim U([0,1])}[Z_{\beta(\tau)}(x,a)]$$

[3] Identity mapping

Quantile Function

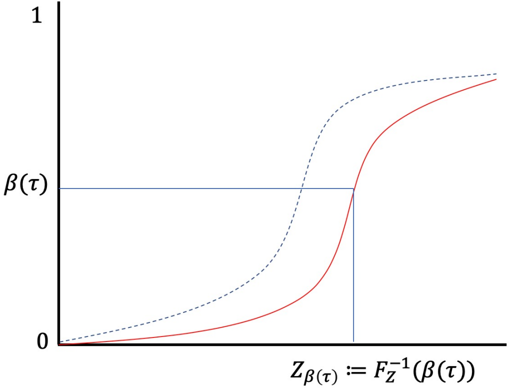
Quantile Function for r.v. Z:

$$Z_\tau := F_Z^{-1}(\tau) \quad \tau \in [0,1]$$



Distorted Expectation of $Z(x, a)$

$$Q_\beta(x,a) := \mathbb{E}_{\tau \sim U([0,1])}[Z_{\beta(\tau)}(x,a)]$$



Distorted Expectation of $Z(x, a)$

Remarks:

- Effectively, change of (probability) measure
- $\frac{\partial F_{\beta(\tau)}}{\partial F_{\tau}}$ is the Nikodym derivative.

Distorted Expectation of $Z(x, a)$

Distorted distribution $Q_\beta(x, a)$ is equal to the expected value of $F^{-1}_{Z(x,a)}$ weighted by β,

$$Q_\beta = \int_0^1 F_Z^{-1}(\tau) d\beta(r)$$

Therefore, for any distorted risk measure β, there exists a sampling distribution for τ such that the mean of Z_τ is equal to the distorted expectation of Z under β, aka, any distorted expectation can be represented as a weighted sum over the quantiles (more later).

Risk-sensitive greedy policy π_β

$$\pi_\beta(x) = \arg\max_{a \in A} Q_\beta(x, a)$$

where

$$Q_\beta = \int_0^1 F_Z^{-1}(\tau) d\beta(\tau)$$

Temporal Difference (TD) error

For two samples $\tau, \tau^{'} \sim U([0,1])$, and policy π_β, the sampled TD error at step t is:

$$\delta_t^{\tau,\tau^{'}} = r_t + \gamma Z_{\tau'}(x_{t+1}, \pi_\beta(x_{t+1})) - Z_\tau(x_t, a_t)$$

Implicit Quantile Network Loss

Quantile Huber Loss:

$$\mathcal{L}(x_t, a_t, r_t, x_{t+1}) = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{k}(\delta_t^{\tau_i, \tau'_j})$$

where N and N' denote the respective number of iid samples $\tau_i, \tau_j \sim U([0,1])$ and

$$\rho_{\tau_i}^{k}(\delta_t^{\tau_i, \tau'_j}) = |\tau - \mathrm{I}\{\delta_t^{\tau_i, \tau'_j} < 0\}| \frac{H_k(\delta_t^{\tau_i, \tau'_j})}{k} \quad {}^{4}$$

with

$$H_k(\delta_t^{\tau_i, \tau'_j}) = \begin{cases} \frac{1}{2}\delta_t^{\tau_i, \tau'_j} & \text{if } \delta_t^{\tau_i, \tau'_j} \le k \\ k(|\delta_t^{\tau_i, \tau'_j}| - \frac{1}{2}k) & \text{otherwise} \end{cases}$$

[4] I is indicator function

Quantile Huber Loss

Why?

- First they want quantile regression loss in order to obtain unbiased stochastic approximation of the quantile function.
- But the quantile regression loss is not smooth at zero → miserable gradient of the loss function
- Propose to combine the quantile regression loss with Huber loss
- Quantile Huber loss is simply the asymmetric variant of Huber loss.
- see Dabney, Rowland, Bellemare, and Munos (2017)

Table of Contents

How do we approximate $Q_\beta(x, a)$

True Q_β

$$Q_\beta = \int_0^1 F_Z^{-1}(\tau) d\beta(\tau)$$

Risk-sensitive greedy policy:

$$\pi_\beta(x) = \arg\max_{a \in A} Q_\beta(x, a)$$

Sample Q_β

$$Q_\beta(x, a) = \frac{1}{K} \sum_{k=1}^{K} Z_{\beta(\tau_k)}(x, a)$$

Sample-based risk-sensitive policy:

$$\pi_\beta(x) = \arg\max_{a \in A} \frac{1}{K} \sum_{k=1}^{K} Z_{\beta(\tau_k)}(x, a)$$

i.e. average over the quantiles

One thing that puzzles me

Back to their claim: "can we expand the class of policies using information provided by the distribution over returns (i.e. to the class of risk-sensitive policies)?"

However, their policy is still greedy policy

$$\pi_\beta(x) = \arg\max_{a \in A} Q_\beta(x, a)$$

At maximum, risk-preference is incorporated in Q values, i.e. again, we have better estimate of the state-action value.

"risk-preference adjusted Q-value"

IQN Network Structure

Back to the Quantile Function for r.v. Z: $Z_\tau = F_Z^{-1}(\tau)$, now the problem is that we need to specify the inverse CDF F_Z^{-1}.

Recall in Deep Q Network: $Q(x, a) \approx f(\psi(x))_a$

Basically, many composition functions to approximate Q value. To be more specific, ψ is a set of convolutional neural networks and $f : R^d \rightarrow R^{|A|}$ is fully connected layers that maps $\psi(x)$ to the estimated action-values.

"Black Box"

IQN Network Structure

Remarks:

- It is not necessarily a black box per se.
- Value distribution can be a distribution which we cannot write down a formula (compare to normal distribution)
- Composition of functions (Neural Networks) can fit the data very well (on the other hand, that is why NN can easily overfit)
- But again the results really depend on composition functions.

IQN Network Structure

Introducing another function $\phi : [0,1] \to R^d$, which computes an embedding for the sample point τ, e.g.

$$\phi_j(r) := ReLU(\sum_{i=0}^{n-1} \cos(\pi i \tau) w_{ij} + b_j)$$ [5]

The inverse CDF can be directly approximated by:

$$Z_\tau(x,a) \approx f(\psi(x) \odot \phi(\tau))_a$$ [6]

[5] ReLU: Rectified Linear Unit, $\max(0, w^T x + b)$

[6] $\odot$ denotes element-wise (Hadamard) product

IQN Network Structure

Remarks:

- In essence, a set of composition functions that take state x as input and output inverse CDF, given a probability $\tau \in [0,1]$, for a particular state-action pair.

Risk-Sensitive Reinforcement Learning

Recall: identity mapping $\beta(\tau) \Leftrightarrow$ Risk-Neutrality.
Different risk preferences can be reflected via other linear/non-linear β functions, which further leads to different risk-sensitive policies $\pi_{\beta}(x)$

- Cumulative probability weighting parametrization proposed in cumulative prospect theroy (Wu & Gonzalez, 1996)

$$CPW(\eta,\tau) = \frac{\tau^{\eta}}{(\tau^{\eta} + (1-\tau)^{\eta})^{\frac{1}{\eta}}}$$

- Distortion risk measure (S. S. Wang, 2000)

$$Wang(\eta,\tau) = \Phi(\Phi^{-1}(\tau) + \eta)$$

where Φ is standard Normal CDF.

Risk-Sensitive Reinforcement Learning

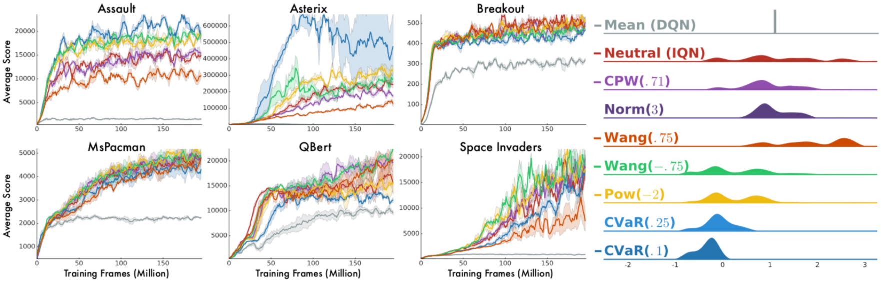
- Power function

$$Pow(\eta,\tau)=\begin{cases}\tau^{\frac{1}{1+|\eta|}}, & \text{if } \eta\geq 0\\ 1-(1-\tau)^{\frac{1}{1+|\eta|}}, & \text{otherwise}\end{cases}$$

- Conditional Value at Risk (Chow & Ghavamzadeh, 2014)

$$CVaR(\eta,\tau)=\eta\tau$$

Results


Assault
Average Score
20000
15000
10000
5000
0
50
100
150
Asterix
600000
500000
400000
300000
200000
100000
0
50
100
150
Breakout
500
400
300
200
100
0
50
100
150
MsPacman
Average Score
5000
4000
3000
2000
1000
0
50
100
150
Training Frames (Million)
QBert
20000
15000
10000
5000
0
50
100
150
Training Frames (Million)
Space Invaders
20000
15000
10000
5000
0
50
100
150
Training Frames (Million)
Mean (DQN)
Neutral (IQN)
CPW(.71)
Norm(3)
Wang(.75)
Wang(−.75)
Pow(−2)
CVaR(.25)
CVaR(.1)
-2
-1
0
1
2
3

References I

Allais, M. (2008). *Allais paradox*. Springer.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223).

Chow, Y., & Ghavamzadeh, M. (2014). Algorithms for cvar optimization in mdps. In *Advances in neural information processing systems* (pp. 3509–3517).

Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2017). Distributional reinforcement learning with quantile regression. *arXiv preprint arXiv:1710.10044*.

Von Neumann, J., & Morgenstern, O. (1964). *Theory of games and economic behavior*. Princeton university press.

Wang, S. (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, *26*(1), 71–92.

References II

Wang, S. S. (2000). A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, 15–36.

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management science*, *42*(12), 1676–1690.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, 95–115.

Distortion function and distortion expectations

Distortion Function

A distortion function is a non-decreasing function $g : [0,1] \to [0,1]$ such that $g(0) = 0$ and $g(1) = 1$.

Distortion function and distortion expectations

$$
\begin{aligned}
U(v) &= \int_0^1 f(G_v(t))dt \\
&= [tf(G_v(t))]_0^1 - \int_0^1 tdf(G_v(t)) \\
&= 0 - \int_0^1 tf^{'}(1 - F_v(t))(-F_v^{'}(t))dt \\
&= \int_0^1 tf^{'}(G_v(t))dF_v(t)
\end{aligned}
$$