Distributional reinforcement learning

Grishin Alexander

PhD student at HSE

August 28, 2018

p(B|A)yesgroup.ru

Important concepts

Markov Decision Process

$$\langle \mathcal{X}, \mathcal{A}, R, P, \gamma \rangle$$

- $\mathcal{X}$ - state space
- $\mathcal{A}$ - action space
- $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ - reward function
- $P : \mathcal{X} \times \mathcal{A} \to \Omega(\mathcal{X})$ - transition probability map of env.
- γ - discount factor

Policy

$$\pi : \mathcal{X} \to \Omega(\mathcal{A})$$

Return (discounted reward)

$$Z^\pi \triangleq \sum_{t=0}^{\infty} \gamma^t R_t$$

Value functions

Goal

Maximize the expected return!

Expected return conditional on state (**state-value function**):

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

Expected return conditional on state and action (**action-value function**):

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

where $x_t \sim P(\cdot|x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot|x_t), x_0 = x, a_0 = a$.

Bellman equations

Bellman expectation equation

$$Q^\pi(x, a) = \mathbb{E}_R[R] + \gamma \mathbb{E}_{R,P,\pi}[Q(x', a')]$$

Operator form:

$$\mathcal{T}^\pi Q(x, a) = \mathbb{E}_R\left[R(x, a)\right] + \gamma \mathbb{E}_{R,P,\pi}\left[Q(x', a')\right]$$

Bellman optimality equation

$$Q(x, a) = \mathbb{E}_{P,R}[R + \gamma \max_{a'} Q(x', a')]$$

Operator form:

$$\mathcal{T} Q(x, a) = \mathbb{E}_{P,R}[R + \gamma \max_{a'} Q(x', a')]$$

Q-learning algorithm

Given a finite MDP $(\mathcal{X}, \mathcal{A}, P, R, \gamma)$, the Q-learning algorithm, given by the update rule:

$$Q(x, a) \leftarrow \alpha \hat{Q}(x, a) + (1 - \alpha) Q(x, a),$$

where $\hat{Q}(x, a) = r + \gamma \max_{a'} Q(x', a')$

Algorithm

- Get sample (x, a, x', r)
- Compute $\hat{Q}(x, a)$
- Update Q(x, a)
- Repeat

Under some conditions theoretical guarantees on convergence to the optimal solution

Distributional Bellman operator

$$\mathcal{T}^\pi Z(x,a) \overset{D}{:=} R(x,a) + \gamma Z(x',a'), \qquad x' \sim P(\cdot|x,a), a' \sim \pi(\cdot|x')$$

- Get the distribution $Z(x',a')$
- For each state-action pair (x,a)
- Estimate the probability to get to (x',a') from (x,a)
- Mix Zs with these probabilities
- Squash with γ
- Shift on R

Motivation

- The value function gives the expected future discounted reward
- This ignores variance and multi-modality
- Means equality doesn't mean that we have right view on distributions
- Distributional operator can possibly establish better optimization problem

Convergence

One of the key property of non-distributional Bellman operator - it is contraction in Lp $(p \geq 1)$ metric

- Convergence
- Unique fixed point
- Optimality of fixed point

Can we derive similar results for distributional operator?

Wasserstein metric

Wasserstein metric

$$W_p(U, Y) = \left(\int_0^1 |F_Y^{-1}(\omega) - F_U^{-1}(\omega)|^p d\omega \right)^{1/p},$$

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

$$F_Y^{-1}(\omega) := \inf\{y \in \mathbb{R} : \omega \leq F_Y(y)\},$$

Maximal form

$$\bar{d}_p(Z_1, Z_2) := \sup_{x,a} W_p(Z_1(x,a), Z_2(x,a)). \tag{1}$$

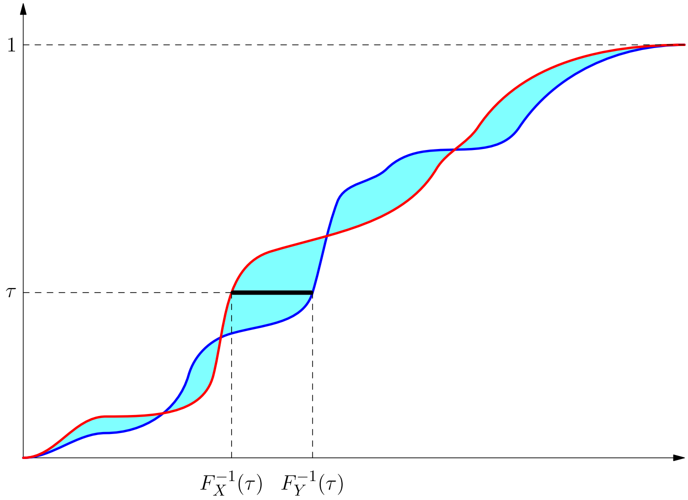

Figure: 1-Wasserstein distance as the measure of difference between the CDFs[1]

[1] mtomassoli.github.io

Theoretical results

Theorem

$\mathcal{T}^\pi$ is a γ-contraction: for any two $Z_1, Z_2 \in \mathcal{Z}$,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma \bar{d}_p(Z_1, Z_2).$$

In theory

This gives us all nice properties: convergence, unique fixed point and optimality **in theory**

In practice

The parametrization can break all results

How to define such good parametrization?

Main result

The **combination** of the projection (defined further) with the Bellman operator is a **contraction** [BDM17]

Properties

- Unique fixed point
- Convergence
- Optimality?

Formally, let $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$ be some parametric model. A quantile distribution $Z_\theta \in \mathcal{Z}_Q$ maps each state-action pair (x, a) to a uniform probability distribution supported on $\{\theta_i(x, a)\}$. That is,

$$Z_\theta(x, a) := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i(x,a)},$$

where δ_z denotes a Dirac at $z \in \mathbb{R}$.

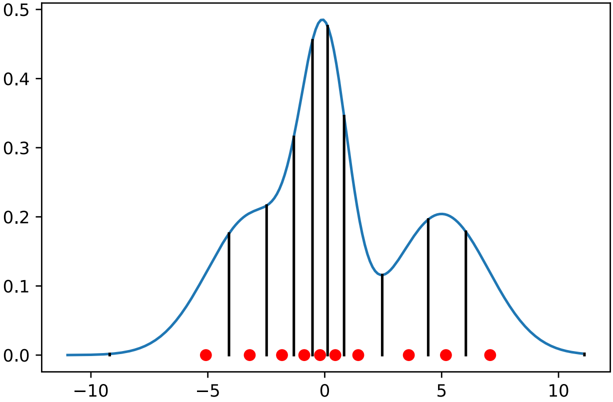

Figure: PDF. The distribution has been sliced up into slices of equal probability mass and red points have been placed in the center of mass of each slice. [3]

[3] mtomassoli.github.io

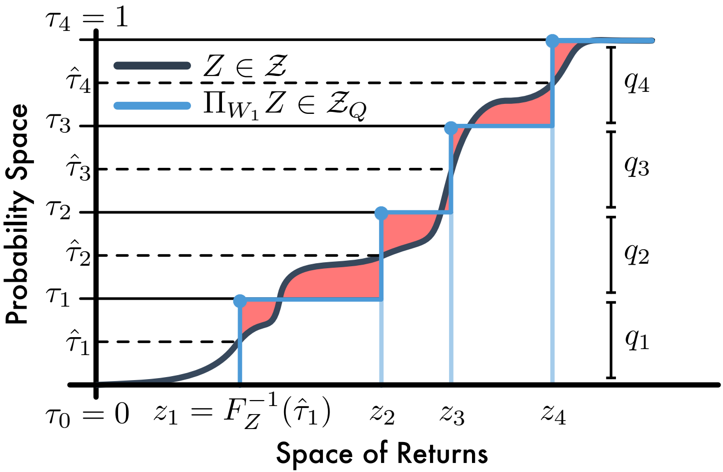

Figure: CDF. 1-Wasserstein minimizing projection onto $N = 4$ uniformly weighted Diracs. Shaded regions sum to form the 1-Wasserstein error.[DRBM17]

For any $\tau, \tau' \in [0,1]$ with $\tau < \tau'$ and cumulative distribution function F with inverse F^{-1}, the set of $\theta \in \mathbb{R}$ minimizing

$$\int_{\tau}^{\tau'} |F^{-1}(\omega) - \theta| d\omega \,,$$

is given by

$$\left\{ \theta \in \mathbb{R} \middle| F(\theta) = \left(\frac{\tau + \tau'}{2} \right) \right\}.$$

In particular, if F^{-1} is the inverse CDF, then $F^{-1}((\tau + \tau')/2)$ is always a valid minimizer, and if F^{-1} is continuous at $(\tau + \tau')/2$, then $F^{-1}((\tau + \tau')/2)$ is the unique minimizer.

Quantile regression

$$\mathcal{L}_{\mathrm{QR}}^{\tau}(\theta) := \mathbb{E}_{\hat{Z}\sim Z}[\rho_\tau(\hat{Z}-\theta)]\,, \text{ where}$$
$$\rho_\tau(u) = u(\tau - \delta_{\{u<0\}}), \forall u \in \mathbb{R}$$

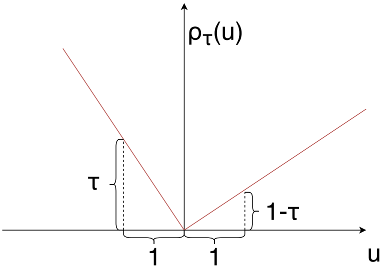

Figure: Quantile loss function

Require: N, κ

input $x, a, r, x', \gamma \in [0, 1)$

Compute distributional Bellman target

$Q(x', a') := \sum_j q_j \theta_j(x', a')$

$a^* \leftarrow \arg\max_{a'} Q(x', a')$

$\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(x', a^*), \quad \forall j$

Compute quantile regression loss

output $\sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{N} \left[\rho_{\hat{\tau}_i}^{\kappa}(\mathcal{T}\theta_j - \theta_i(x, a))\right]$

References I

[BDM17] Marc G Bellemare, Will Dabney, and Rémi Munos, *A distributional perspective on reinforcement learning*, arXiv preprint arXiv:1707.06887 (2017).

[DRBM17] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos, *Distributional reinforcement learning with quantile regression*, arXiv preprint arXiv:1710.10044 (2017).