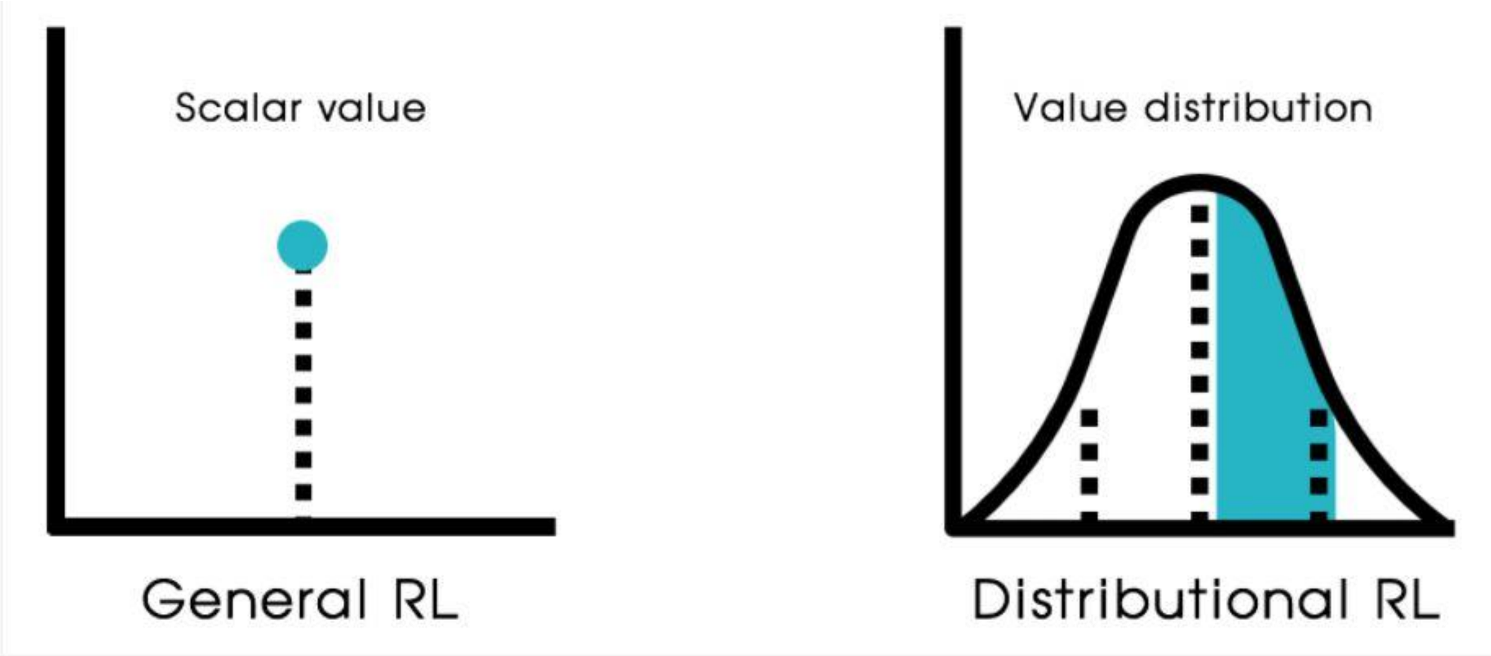
Distributional Reinforcement Learning with Quantile Regression

QR-DQN

IAN LEE

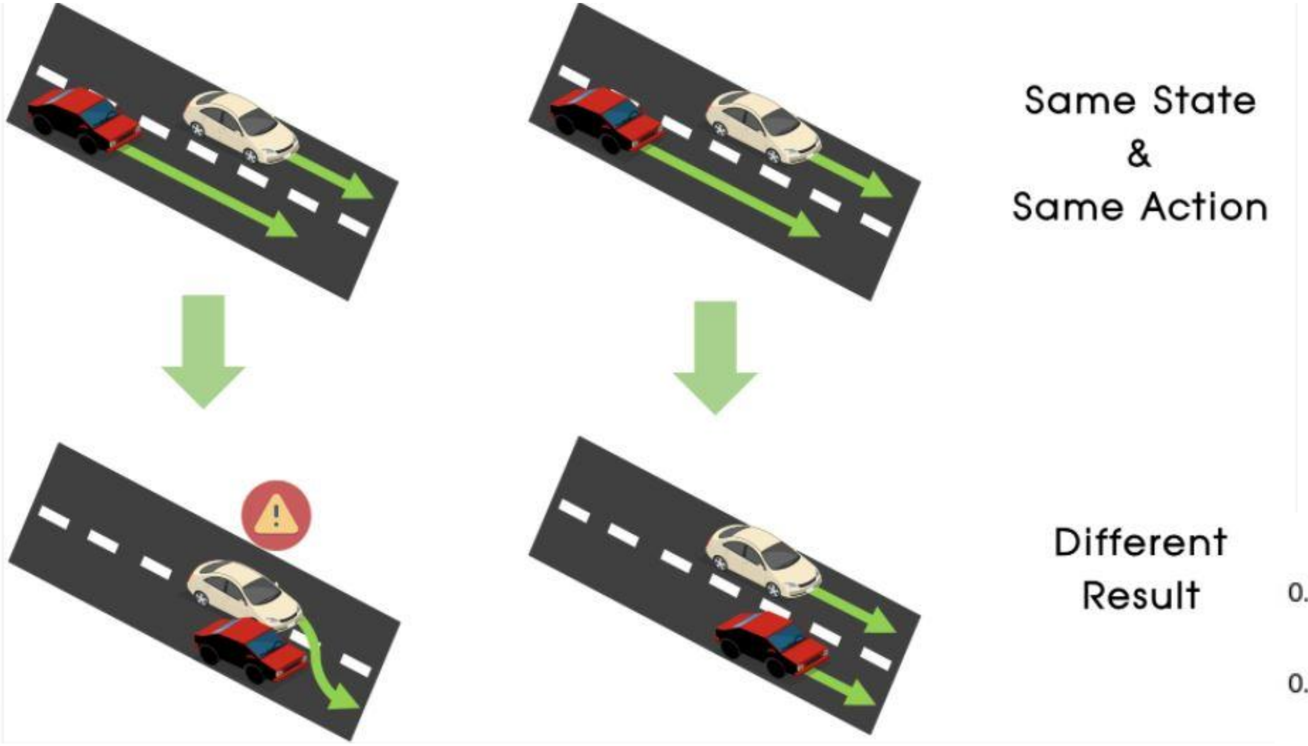
Dist.RL

기댓값만 구하지 말고, 분포를 직접 근사한 다음에 기댓값을 구하자!

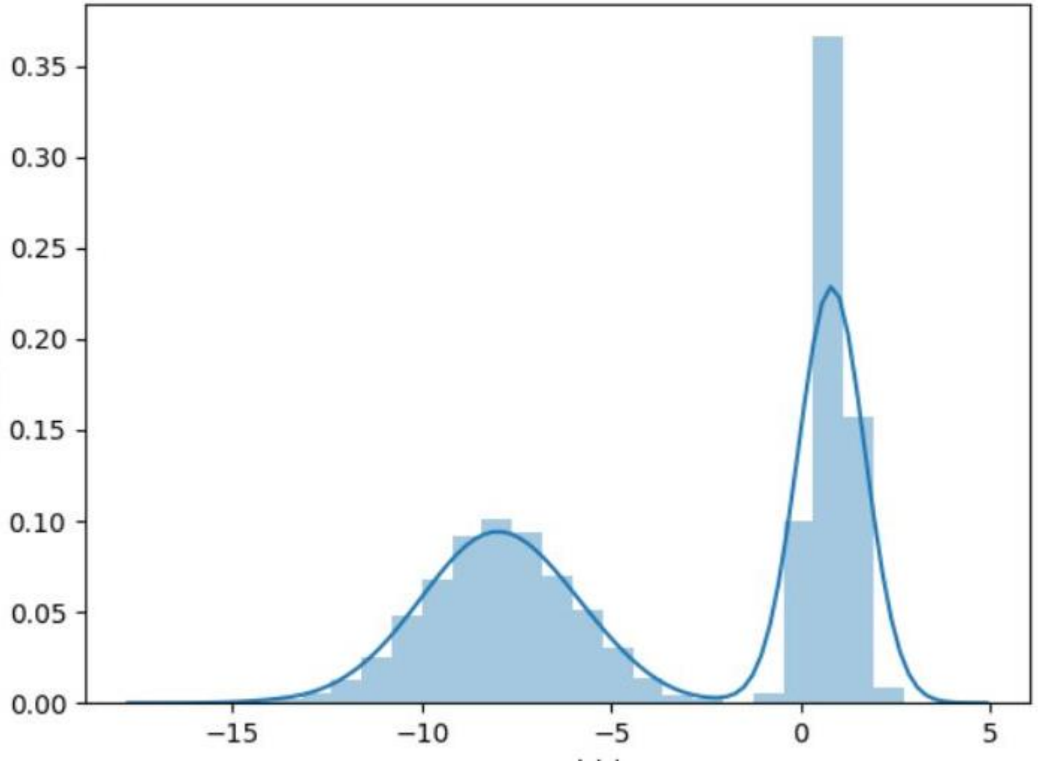


=Q(s, a)

(s, a)



환경의 Intrinsic Randomness를
고려하기 위해



왜 이렇게 하지..?!

Notations..

$$(\text{MDP})\ (\mathcal{X}, \mathcal{A}, R, P, \gamma)$$ 상태, 행동, 보상...

$$Z^{\pi} = \sum_{t=0}^{\infty} \gamma^t R_t$$ **Random Variable**

이것의 기댓값이 바로 우리가 쓰던 **Q**와 **V**

Dist-RL은
Z의 **확률분포**를 직접 계산
하겠다는 것!

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

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

어떻게 Z를 Learning하지?

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

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

Z에도 Q처럼 벨만 방정식이 있을까..?

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

있다!!

저 식에서 $:\overset{D}{=}$ 의 의미는 두 확률변수가 같다는 것!

그리고 심지어 Q처럼 γ-contraction도 만족한다.

γ-contraction에 관하여

- $\mathcal{T}^{\pi}$는 L2-distance에 대해 γ-contraction이다. (Bellman Operator)

그 말인즉슨.. $d(Q_1, Q_2) = \sup_{\mathrm{x,a}} |\mathrm{Q}_1(\mathrm{x,a}) - \mathrm{Q}_2(\mathrm{x,a})|^2$ 으로 정의할 때

$d(\mathcal{T}^{\pi}Q_1, \mathcal{T}^{\pi}Q_2) \le \gamma d(Q_1, Q_2)$

Z에 대한 Bellman Operator $\mathcal{T}^{\pi}$는 **Wasserstein Distance**에 대해 γ-contraction이다.

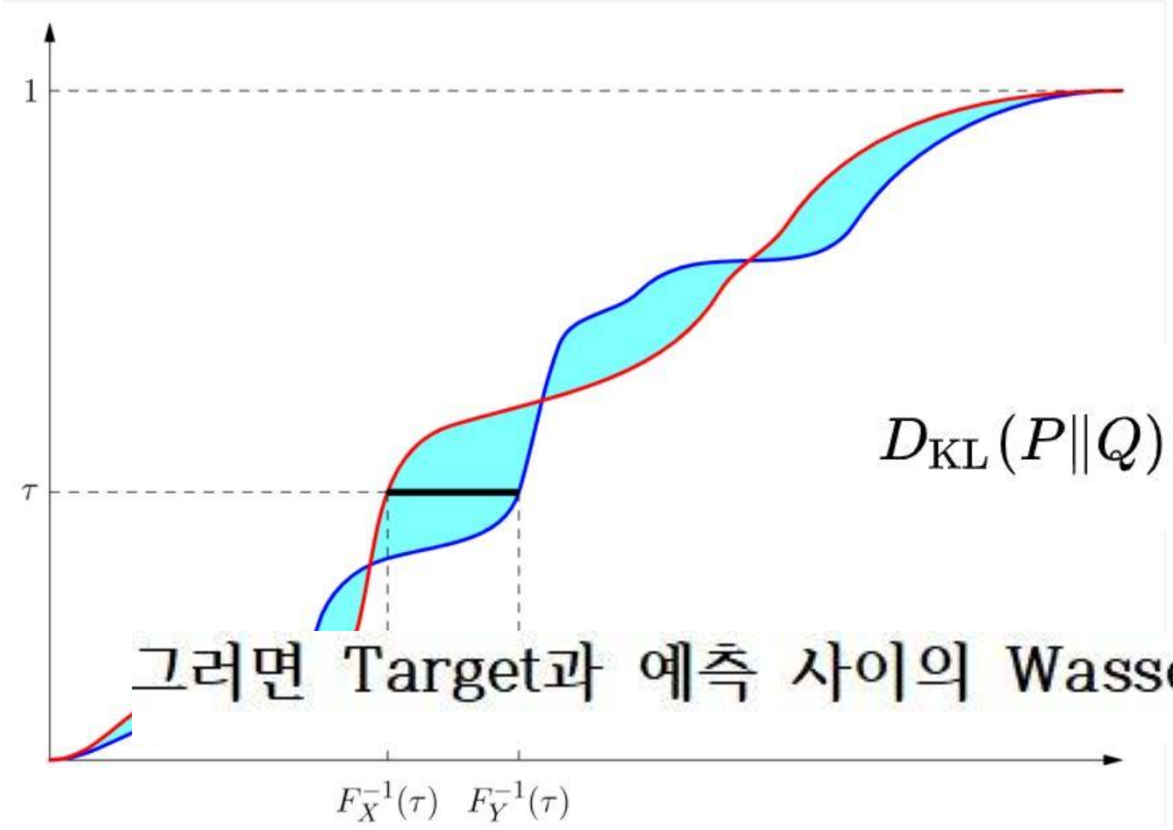
$$Z_1, Z_2 \in \mathcal{Z} \quad \bar{d}_p(Z_1, Z_2) := \sup_{x,a} W_p(Z_1(x,a), Z_2(x,a)).$$

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

Wasserstein Distance?

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

F는 누적분포함수!



Does not suffer from disjoint-support

$$W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$$

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} = \begin{cases} +\infty & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$$

그러면 Target과 예측 사이의 Wasserstein Distance 줄이면 되겠네!

그렇게 간단하지 않더라…

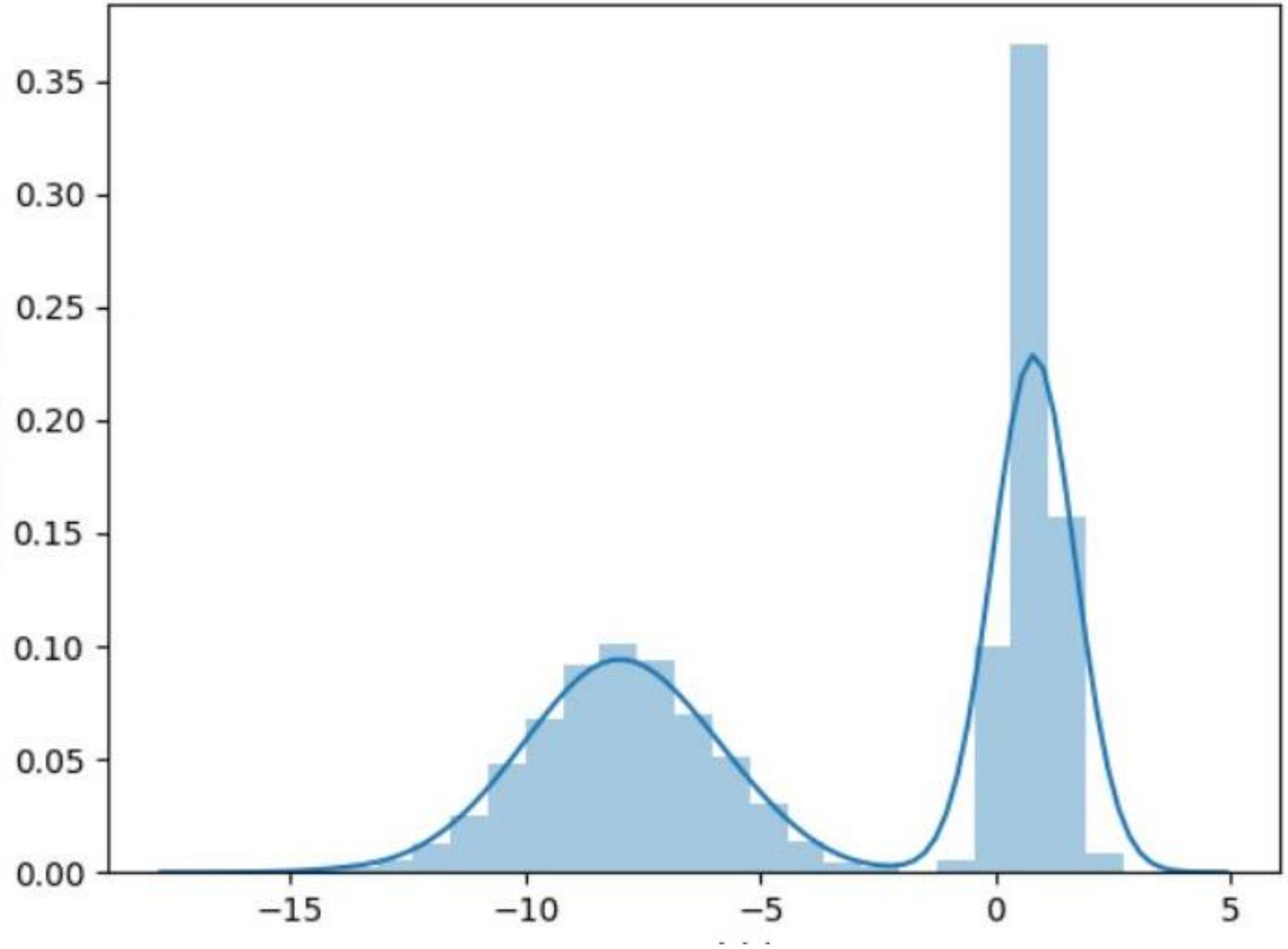
$\arg\min_{\mu}$ ͏ ͏ ͏

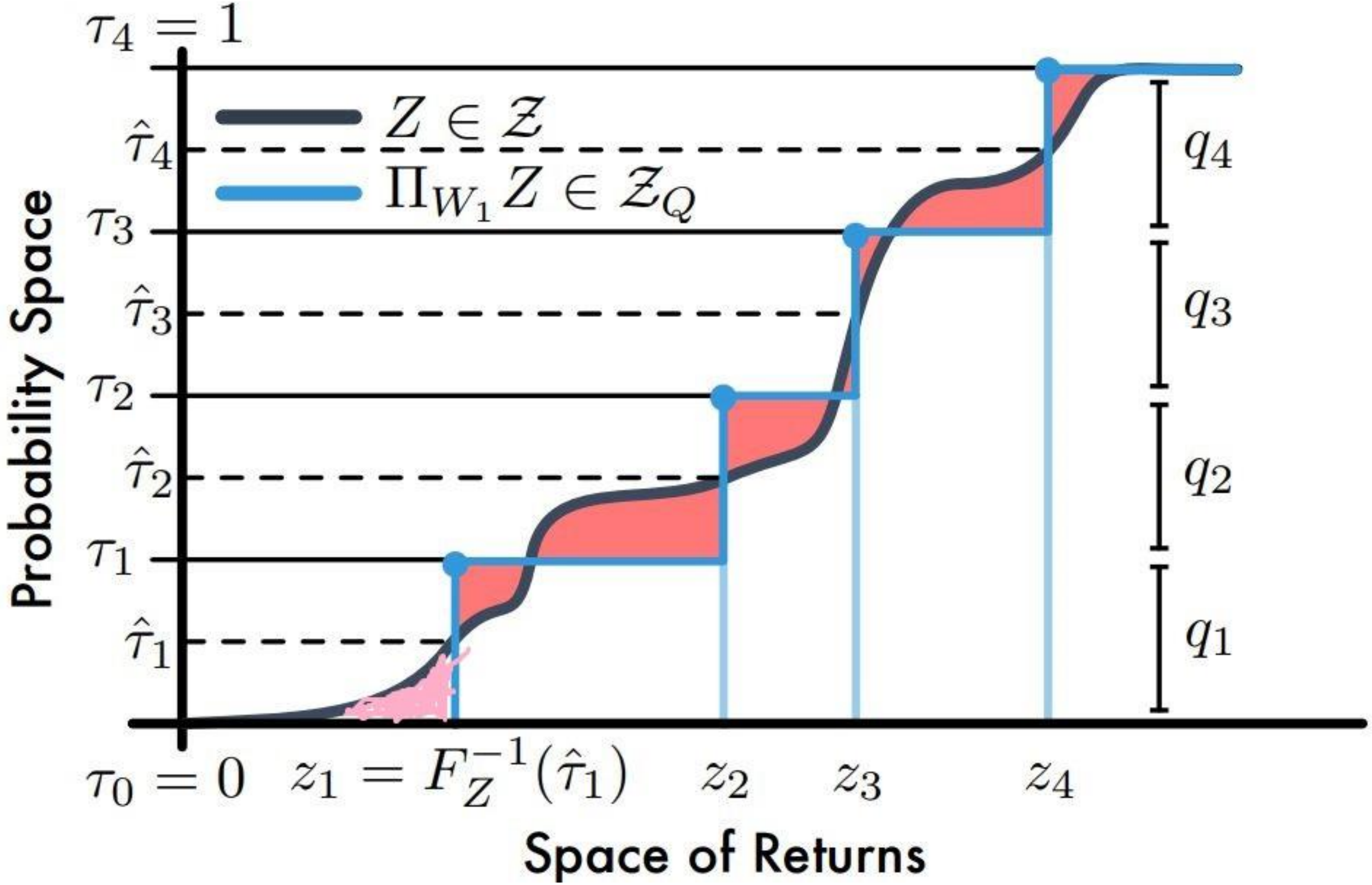
샘플링을

ᅳ vergence를 줄였음

여기ᄉ
Regre

"Fixe ,





Discrete하게 근사하다보니 저렇게 삐쭉삐쭉 올라감!

어떻게 분포를 모델링할 것인가

$$W_1(Y,U) = \sum_{i=1}^{N} \int_{\tau_{i-1}}^{\tau_i} |F_Y^{-1}(\omega) - \theta_i| d\omega.$$

Lemma 2. *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

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

수학, 수학 수학... ㅠㅠ

요약: $\boldsymbol{\tau}$가 **아닌 $(\boldsymbol{\tau} + \boldsymbol{\tau}')/2$에 대한 support를 구해야 한다**

이렇게 모델링 하면 끝나나요..?

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

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

N: Number of quantiles

$$\rho_\tau(u) = \begin{cases} u(\tau - 1) & if\ u < 0 \\ u\ (\tau) & if\ u \geq 0 \end{cases}$$

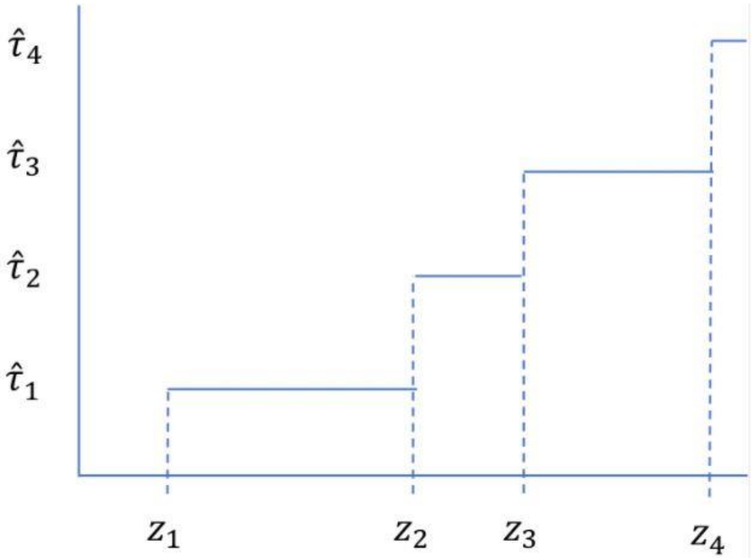
be a quantile distribution, and $\hat{Z}_m$
on composed of m samples from Z.
e exists a Z such that

$$_m, Z_\theta)] \neq \arg\min W_p(Z, Z_\theta).$$

method, more widely used in eco-

Quantile Regression Loss

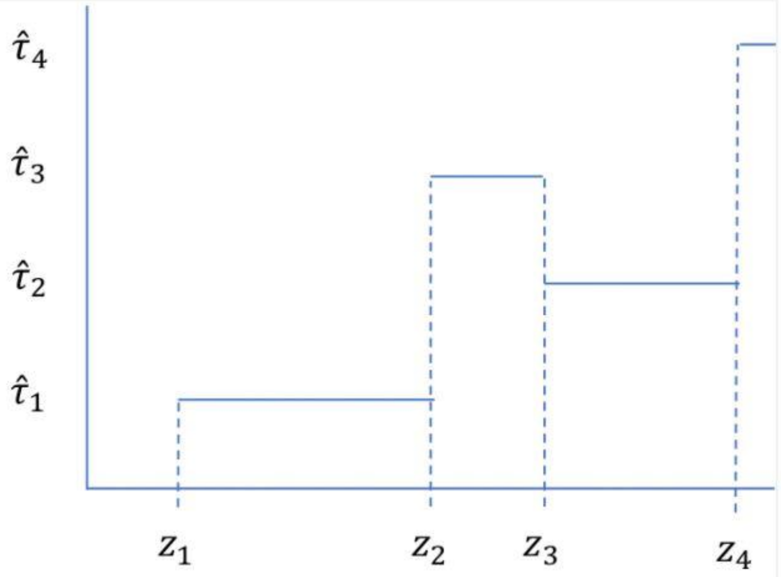
Modeling Example



$$\tau = [0.25, 0.5, 0.75, 1]$$

$$\hat{\tau} = [0.125, 0.375, 0.625, 0.875]$$

$$z = [1, 4, 5, 7]$$



$$\tau = [0.25, 0.5, 0.75, 1]$$

$$\hat{\tau} = [0.125, 0.375, 0.625, 0.875]$$

$$z = [1, 5, 4, 7]$$

오른쪽은 아예 CDF의 정의에도 안 부합하므로 왼쪽보다 더 큰 페널티가 부여되어야 함!

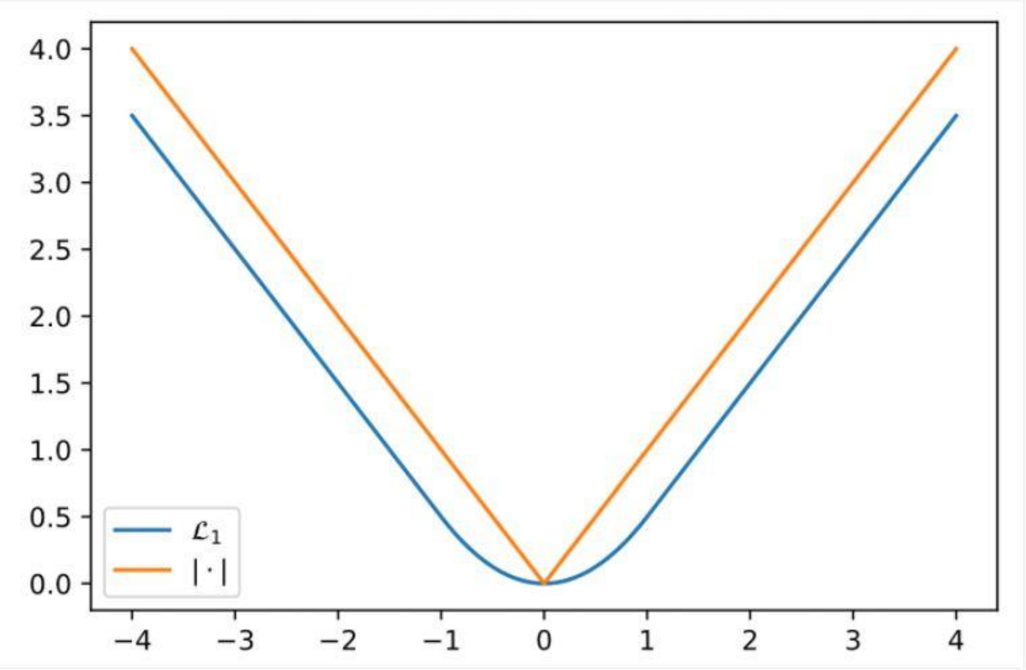
Quantile Regression Loss의 계산

$$\rho_{\hat{\tau}_i}\left(\mathcal{T}\theta_j - \theta_i(x,a)\right) \longrightarrow \rho_{\tau}(u) = \begin{cases} u(\tau - 1) & if\ u < 0 \\ u\,(\tau) & if\ u \geq 0 \end{cases}$$

Sum

$$\begin{bmatrix} 0.125 & 1.25 & 1.125 & 0.375 \\ 0.375 & 0 & 0.375 & 0.125 \\ 0.875 & 1.5 & 1.875 & 0 \\ 1 & 1.875 & 2.5 & 0.875 \end{bmatrix} \Rightarrow \begin{bmatrix} 2.875 \\ 0.875 \\ 4.25 \\ 6.25 \end{bmatrix} \xrightarrow{\text{Mean}} 3.6525$$

Quantile Huber Loss



$$\mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & if\ |u| \le \kappa \\ \kappa\left(|u| - \frac{1}{2}\kappa\right), & otherwise \end{cases}$$

$$\rho_\tau(u) = \begin{cases} u(\tau - 1) & if\ u < 0 \\ u\ (\tau) & if\ u \ge 0 \end{cases}$$

$$\rho_\tau(u) = \begin{cases} \mathcal{L}_\kappa(u)(1 - \tau) & if\ u < 0 \\ \mathcal{L}_\kappa(u)(\tau) & if\ u \ge 0 \end{cases}$$

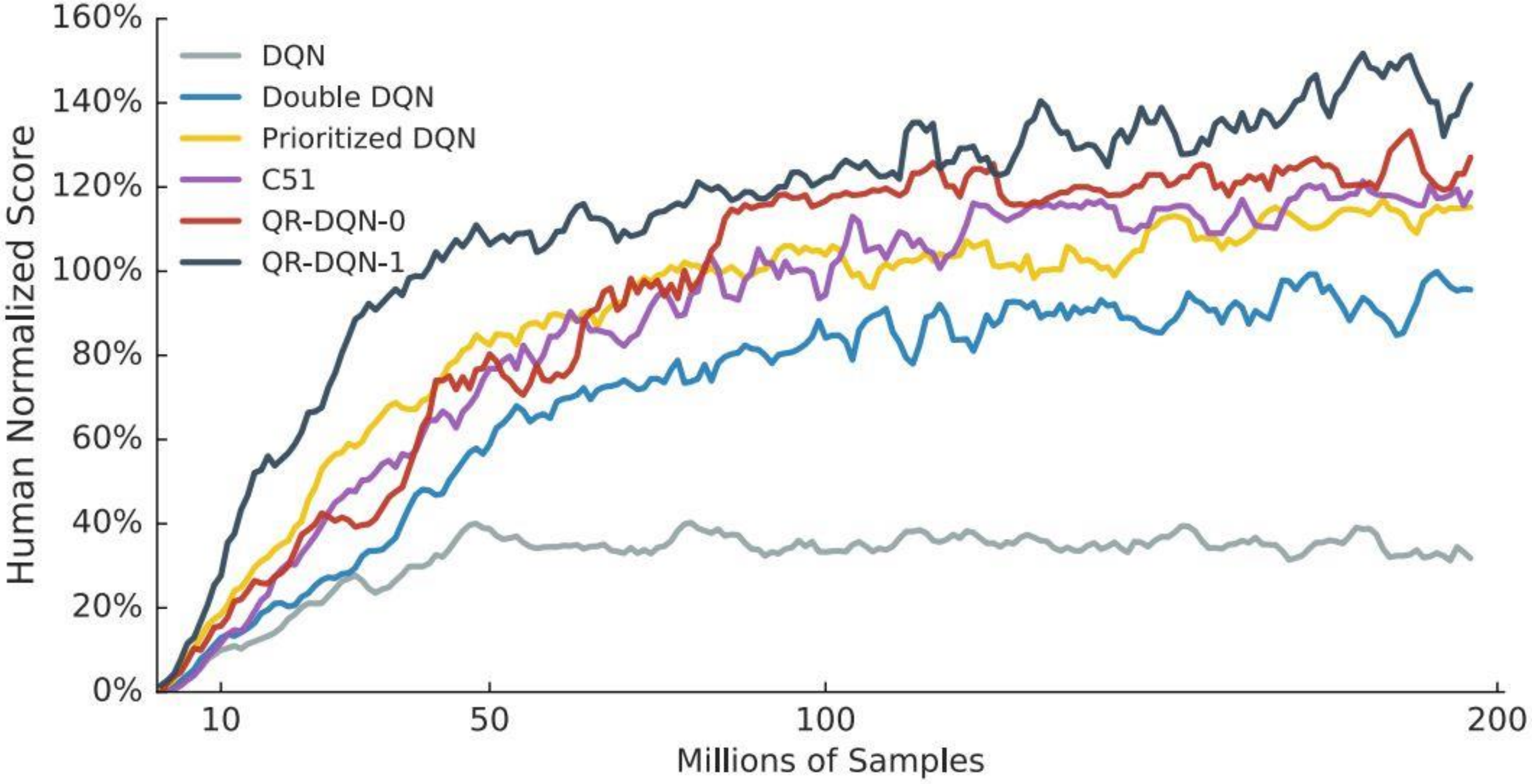
알고리즘!

Algorithm 1 Quantile Regression Q-Learning

Proposition 2. *Let Π_{W_1} be the quantile projection defined as above, and when applied to value distributions gives the projection for each state-value distribution. For any two value distributions $Z_1, Z_2 \in \mathcal{Z}$ for an MDP with countable state and action spaces,*

$$\bar{d}_\infty(\Pi_{W_1}\mathcal{T}^\pi Z_1, \Pi_{W_1}\mathcal{T}^\pi Z_2) \leq \gamma \bar{d}_\infty(Z_1, Z_2). \tag{11}$$

이론적 Contribution!


160%
140%
120%
100%
80%
60%
40%
20%
0%
Human Normalized Score
DQN
Double DQN
Prioritized DQN
C51
QR-DQN-0
QR-DQN-1
10
50
100
200
Millions of Samples

좋더라!!

감사합니다!!

QR-DQN 구현

- https://github.com/rl-max/deep-reinforcement-learning-pytorch/blob/main/qr-dqn.py
- Pytorch로 바닥부터 직접 구현한 코드는 위에 있습니다만…
- 성능이 잘 안나옵니다 ㅠ
- 관심있는 분들/고수분들께서 한번 살펴봐주시면 감사하겠습니다!!