Обучение с подкреплением на распределениях

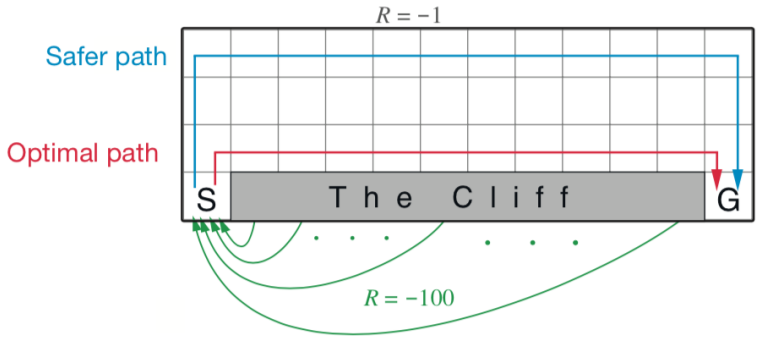
Осминин Константин

10 сентября 2019

Tinkoff.ru

Мотивация

- Более точная оценка вознаграждений
- Допускает мультимодальность вознаграждений
- Допускает управление риском
- Более стабильное обучение



Учим распределения, а не среднее

$$Q(x,a) = \mathbb{E}_\pi \sum_i \gamma^i R(x_i, a_i)$$

$$\mathbb{E}_\pi \sum_i \gamma^i R(x_i, a_i) \sim Z(x,a)$$

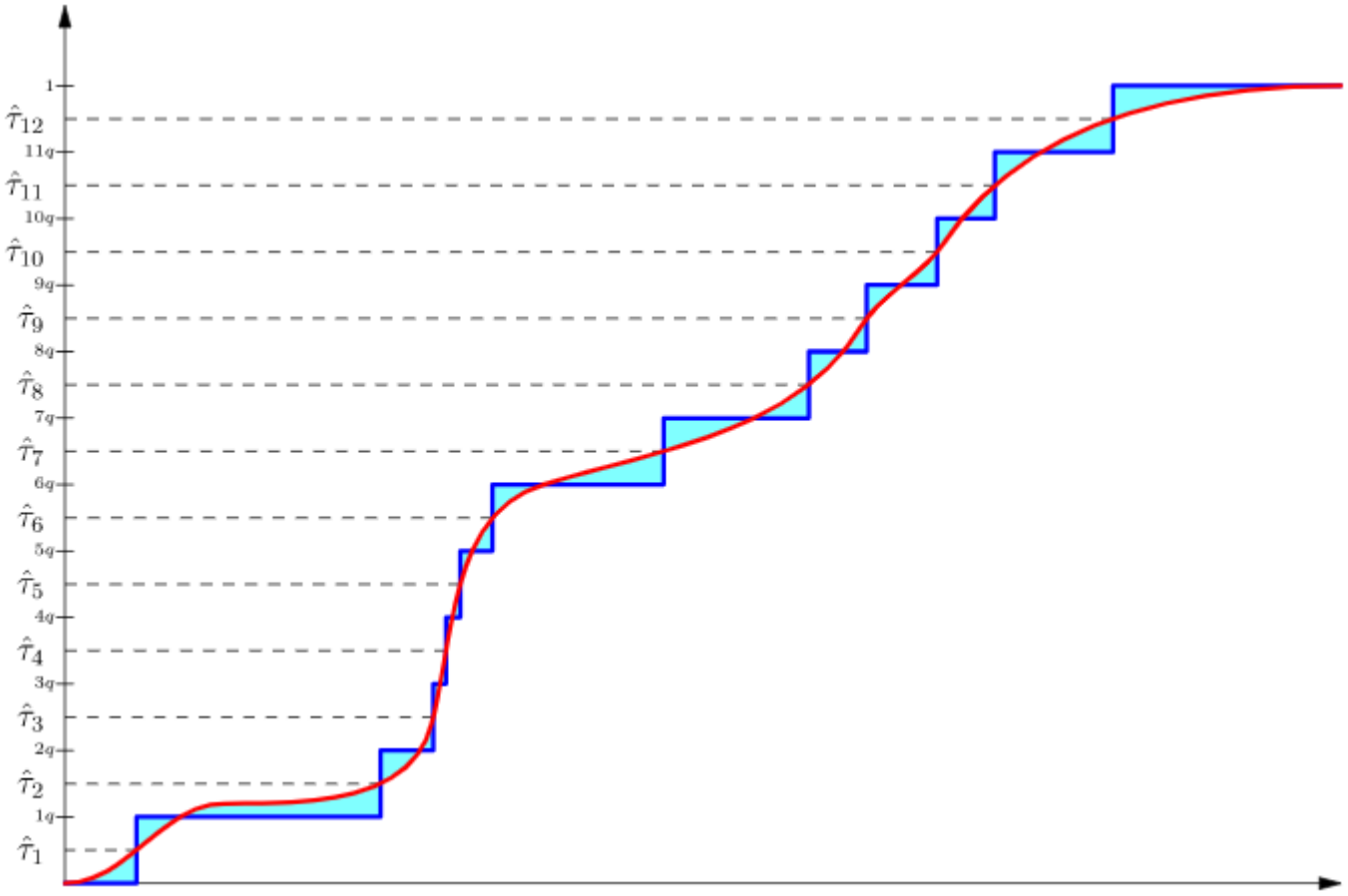
$$Q(x,a) = \mathbb{E}\, R(x,a) + \gamma\, \mathbb{E}\, Q(X', A').$$

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

по распределению

$$Q(x,a) = \mathbb{E}\ \ Z(x,a)$$

Как учить распределение

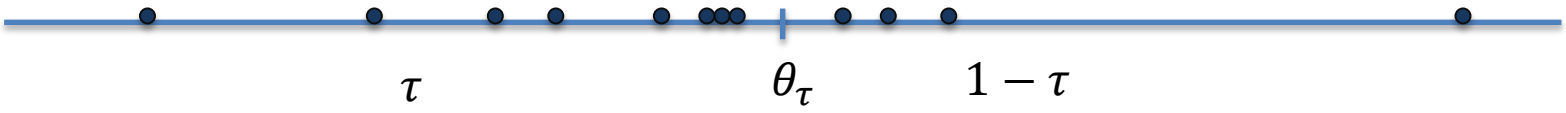


Учим N квантилей $\{\theta_i = F^{-1}(i/N)\}$

Квантильная регрессия

θ_τ есть τ-квантиль выборки X =>

$$\theta_\tau = argmax\, L_\theta = argmax\, \mathbb{E}_X |X - \theta| \cdot |\tau - \delta_{X<\theta}|$$

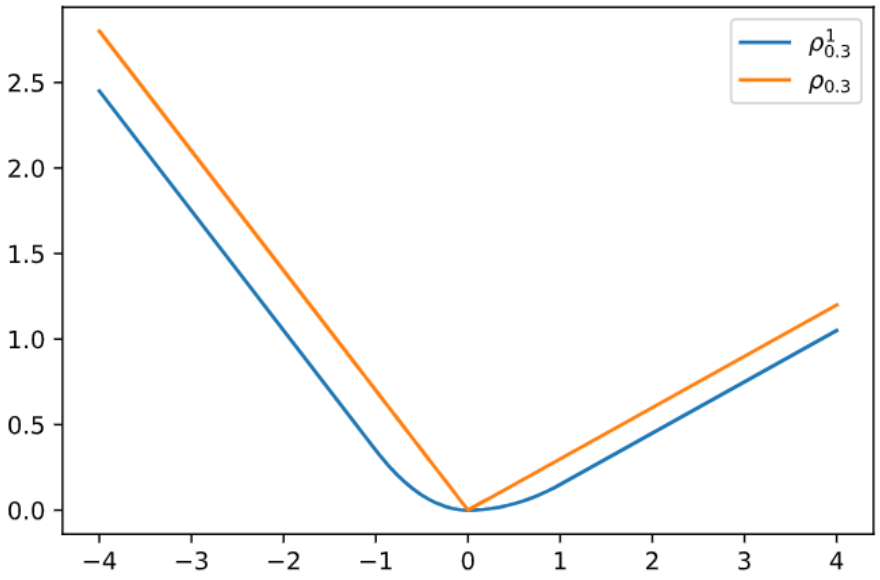


$$\nabla_\theta\, L_\theta = \begin{cases} 1-\tau, X < \theta \\ \tau, X \geq \theta \end{cases}$$

Как учить квантили

- Используем квантильную регресию, только для гладкости применяем Huber Loss

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



- Для квантили τ

$$L_\theta = \mathbb{E}_X \mathcal{L}_k(X - \theta) \cdot |\tau - \delta_{X<\theta}|$$

- Для всех квантилей

$$L_\theta = \sum_{i=1}^{N} \mathbb{E}_X \mathcal{L}_k(X - \theta) \cdot |\tau_i - \delta_{X<\theta}|$$

Алгоритм QR-DQN

- $s = s_0$
- Пока не сошлись:
  - $a = argmax_{a'} \, {}^{1}/_{N} \sum_{j=1}^{N} \theta_j(s, a')$
  - Из состояния s делаем шаг a и наблюдаем s', r.
  - Обучение:
    - $a' = argmax_{a'} \, {}^{1}/_{N} \sum_{j=1}^{N} \theta_j(s, a')$
    - $\theta_j^* = r + \gamma \theta_j(s', a')$
    - $L \quad = {}^{1}/_{N^2} \sum_{i,j=1}^{N} \mathcal{L}_k(\theta_j^* - \theta(x, a)) \cdot \left| {}^{i}/_{N} - \delta_{X<\theta(x,a)} \right|$
    - Шаг градиентного спуска по параметрам $\theta(x, a)$
  - $s \leftarrow s'$

Результаты

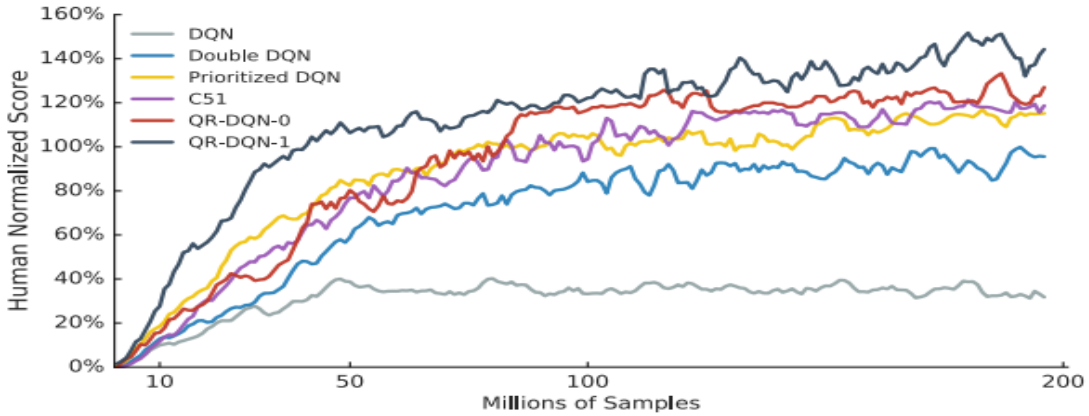


Figure 4: Online evaluation results, in human-normalized scores, over 57 Atari 2600 games for 200m train samples.

	Mean	Median	>human	>DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	701%	178%	40	50
QR-DQN-0	881%	199%	38	52
QR-DQN-1	**915%**	**211%**	**41**	**54**

Table 1: Mean and median of *best* scores across 57 Atari 2600 games, measured as percentages of human baseline (Nair et al. 2015).

Материалы

- Marc G Bellemare, Will Dabney, and R´emi Munos, A distributional perspective on reinforcement learning, arXiv preprint arXiv:1707.06887 (2017).

- Will Dabney, Mark Rowland, Marc G Bellemare, and R´emi Munos, Distributional reinforcement learning with quantile regression, arXiv preprint arXiv:1710.10044 (2017).

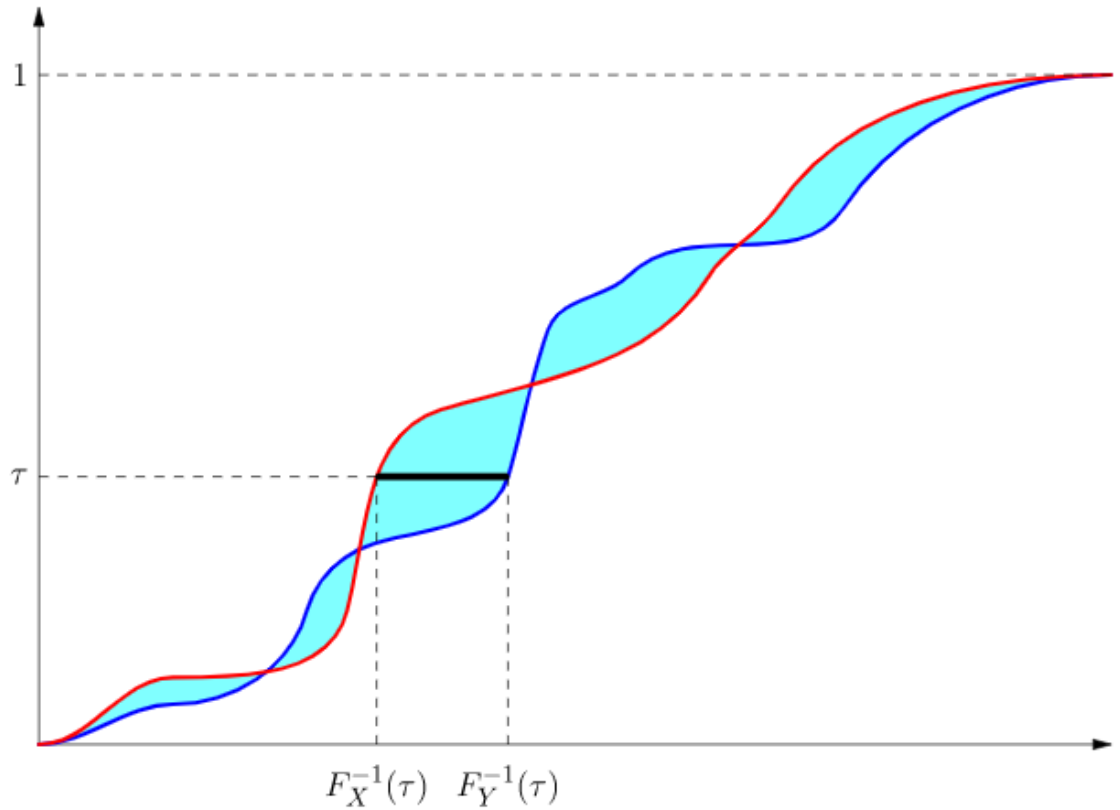
Семинар

https://github.com/bayesgroup/deepbayes-2018/blob/master/day3_qr-qnetwork/qr_dqn-local.ipynb

Метрика Вассерштайна

Let X and Y be two *scalar* random variables and F_X and F_Y their *CDFs*. Then, their p-*Wasserstein distance* is

$$\mathcal{W}_p(X,Y) = \left(\int_0^1 \left| F_X^{-1}(u) - F_Y^{-1}(u) \right|^p du \right)^{1/p}$$



Спасибо