The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Distributional Reinforcement Learning with Quantile Regression

Dabney
pMind

Mark Rowland
University of Cambridge*

Marc G. Bellemare
Google Brain

Rémi Mur
DeepMind

【强化学习 48】Quantile Regression

张楚珩
清华大学 交叉信息院博士在读

紧接着前面Distributional RL做的工作

原文传送门

Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

特色

专栏前一篇讲了distributional RL，使用价值函数分布而不仅是价值函数期望值来做强化学习，之前文章对于分布的近似方式是把概率密度函数用格子来近似（categorical representation），这里使用了更加有效的近似方式（quantile representation），同时理论上也更有保证。

过程

1. 为什么要使用quantile representation?

这里主要研究的问题是价值函数的分布应该如何来表示。

理论上，任何的实际的表示方式都不能完全准确表示一个任意概率分布，因此任何概率分布的表示方式都可以看做真实分布向所能表示的空间上投影（基于从真实概率分布中采集样本集）。理论上关心两件事情。第一件事情，在policy evaluation下Bellman算子多轮迭代后能够收敛，但是它联合与表示有关的投影算子后（即 $\Pi \mathcal{T}^\pi$ ）是否还能在Wasserstein距离度量下收敛呢；第二件事情，考虑采样之后，其实是在缩小所表示的概率分布和样本集所代表分布之间的距离，当所表示的概率分布和样本集上距离最小的时候，它是否也和真实分布距离最小。

categorical representation在这里提到的两件事情上都不能保证（但可以证明在Cramer距离度量下 $\Pi_C \mathcal{T}^\pi$ 是contraction[1]）；而quantile representation在第一件事情上能保证收敛，即能够证明 $\Pi \mathcal{T}^\pi$ 算子是Wasserstein距离下的contraction。

实际操作上，categorical representation需要实现确定价值函数值所在的区间 $[V_{min}, V_{max}]$ ，这引入了超参数；同时，对于某些状态价值函数分布范围相对于 $[V_{min}, V_{max}]$ 很小的情况下，近似非常不准确。

而quantile representation就不存在此问题。

2. Quantile Representation

考虑一个分布的CDF（如下图），categorical representation相当于在横轴上均匀分出若干格子，然后表示每个各自里面CDF增大的量；而quantile representation相当于在纵轴上划分出若干个各自，然后表示每个各自中对应的价值函数值是多少。

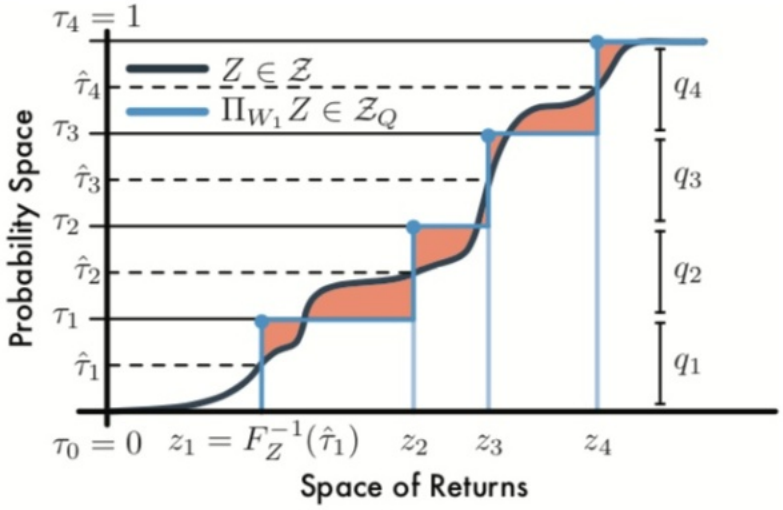


Figure 2: 1-Wasserstein minimizing projection onto $N = 4$ uniformly weighted Diracs. Shaded regions sum to form the 1-Wasserstein error.

考虑 $\tau_i = i/N, i \in [N]$ $\tau_0 = 0$ ，quantile representation相当于对于概率分布做了如下建模，其中 $\theta_i(x,a)$ 可以用神经网络来表示。

$$Z_\theta(x,a) := \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i(x,a)}, \tag{7}$$

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

把任意一个概率分布表示出来相当于就是往上面这个quantile模型上做投影。这里使用了1-Wasserstein距离。

$$\Pi_{W_1} Z := \arg\min_{Z_\theta \in \mathcal{Z}_Q} W_1(Z, Z_\theta),$$

可以证明，在1-Wasserstein距离下，其最优表示 $\theta_i = F_Z^{-1}(\frac{\tau_{i-1}+\tau_i}{2})$ 。但是实际中，真实分布函数及其CDF反函数都是得不到的，因此，我们需要quantile regression来从样本中学习到相应的表示参数。

3. Quantile Regression

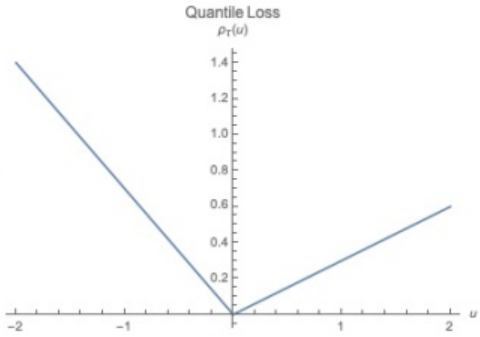
对于给定的样本 z 和参数 τ ，相应的分位数 θ 可以通过对如下损失函数做梯度下降得到。

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

怎样理解这个损失函数呢，即，当样本 z 小于当前分位数 θ 的时候，线性惩罚系数为 $1-\tau$ ；当样本 z 大于当前分位数 θ 的时候，线性惩罚系数为 τ 。这样当分位数 θ 估计准确的时候，这个损失函数刚好不会提供增大或者减小方向的梯度。

```
In[11]:= delta = Function[x, If[x, 1, 0]];
         ρ = Function[{u, τ}, u (τ - delta[u < 0])];
         Plot[ρ[u, 0.3], {u, -2, 2}, AxesLabel → {HoldForm[u], HoldForm[ρτ[u]]},
          PlotLabel → HoldForm[Quantile Loss]]
```

Out[13]=



Quantile loss在零点处导数不连续，计算上不太稳定，这里又提出了一种平滑的方案，即quantile

Huber loss。

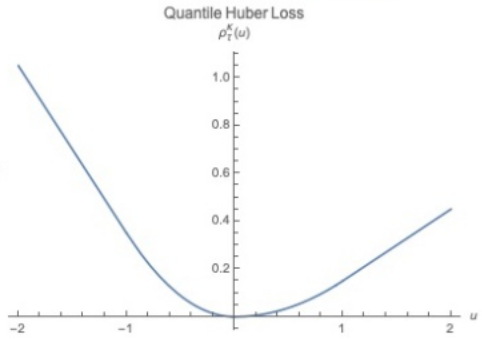
$$\mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \le \kappa \\ \kappa(|u| - \frac{1}{2}\kappa), & \text{otherwise} \end{cases}.$$

$$\rho_\tau^\kappa(u) = |\tau - \delta_{\{u<0\}}| \frac{\mathcal{L}_\kappa(u)}{\kappa}.$$

可以看到它会更平滑。

```
In[8]:= L = Function[{u, κ}, If[Abs[u] ≤ κ, 1/2 u^2, κ (Abs[u] - 1/2 κ)]];
        ρ = Function[{u, τ, κ}, Abs[τ - If[u < 0, 1, 0]] L[u, κ]/κ];
        Plot[ρ[u, 0.3, 1], {u, -2, 2}, AxesLabel → {HoldForm[u], HoldForm[ρ_τ^κ[u]]},
         PlotLabel → HoldForm[Quantile Huber Loss]]
```

Out[10]=



4. 算法

文章还提了一个做policy evaluation的quantile TD的算法，这里不说。仿照DQN，文章提出了quantile regression Q-learning。

Algorithm 1 Quantile Regression Q-Learning

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 (Equation 10)

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

其中 q_i 有点没懂，但这一行应该是求期望，即估计到的各个分位数的价值函数的均值。最后一行里面的 $\mathbb{E}_j[\cdots\mathcal{T}\theta_j\cdots]$ 就相当于是前面损失函数里面对于真是分布的采样 $\mathbb{E}_{\hat{Z}\sim Z}[\cdots\hat{Z}\cdots]$ 。

5. 理论结果

5.1. categorical representation在样本上的最优不等于相对于真实分布的最优

Theorem 1 (Theorem 1, Bellemare et al. 2017). *Let* $\hat{Y}_m := \frac{1}{m}\sum_{i=1}^{m}\delta_{Y_i}$ *be the empirical distribution derived from samples* $Y_1, \ldots, Y_m$ *drawn from a Bernoulli distribution* B*. Let* B_μ *be a Bernoulli distribution parametrized by* μ*, the probability of the variable taking the value* 1*. Then the minimum of the expected sample loss is in general different from the minimum of the true Wasserstein loss; that is,*

$$\arg\min_{\mu} \mathop{\mathbb{E}}_{Y_{1:m}} \left[W_p(\hat{Y}_m, B_\mu)\right] \neq \arg\min_{\mu} W_p(B_\mu, B_\mu)$$

5.2. 考虑policy evaluation，categorical representation的投影联合Bellman算子，在Cramer距离度量下是contraction [1]

Proposition 2. *The operator $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$. Further, there is a unique distribution function $\eta_{\mathcal{C}} \in \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$ such that given any initial distribution function $\eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we have*

$$(\Pi_{\mathcal{C}}\mathcal{T}^{\pi})^m \eta_0 \to \eta_{\mathcal{C}} \text{ in } \bar{\ell}_2 \text{ as } m \to \infty.$$

其中 Π_C 表示categorical representation的投影算子，Cramer距离的定义如下

Definition 3. *The Cramér distance ℓ_2 between two distributions $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, with cumulative distribution functions F_{ν_1}, F_{ν_2} respectively, is defined by:*

$$\ell_2(\nu_1, \nu_2) = \left(\int_{\mathbb{R}} (F_{\nu_1}(x) - F_{\nu_2}(x))^2 \mathrm{d}x\right)^{1/2}.$$

Further, the supremum-Cramér metric $\bar{\ell}_2$ is defined between two distribution functions $\eta, \mu \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ by

$$\bar{\ell}_2(\eta, \mu) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \ell_2(\eta^{(x,a)}, \mu^{(x,a)}).$$

5.3. quantile representation在样本上的最优也不等于相对于真实分布的最优

Proposition 1. *Let Z_θ be a quantile distribution, and $\hat{Z}_m$ the empirical distribution composed of m samples from Z. Then for all $p \geq 1$, there exists a Z such that*

$$\arg\min \mathbb{E}[W_p(\hat{Z}_m, Z_\theta)] \neq \arg\min W_p(Z, Z_\theta).$$

证明方式是举反例，就设真实分布 $Z = \frac{1}{N}\sum_{i=1}^{N}\delta_i$ ，这样采样得到的样本里面最小值或者是1，或者大于1。假设估计分布的第一个参数已经最优了，即 $\theta_1 = 1$ ，那么这个目标对于 θ_1 求导应该为零。但实际情况是当样本的最小值等于1时，导数为零；大于1的时候，导数小于零。因此其导数的期望一定

为负，因此，θ_1 最后肯定会收敛到比 τ_1 更大的某个位置。

5.4. 最优quantile representation存在并且合理

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

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

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.

观察到优化目标是凸函数，因此直接找梯度为零的点即可解得。

5.5. 考虑policy evaluation，quantile representation的投影联合Bellman算子，仍然能形成contraction

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,*

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

其证明过程先考虑如下不影响问题实质的简化：1）认为 $r=0$ ，原因是奖励只是把概率分布做平移，在quantile表示下，这一点完全不影响；2）系数 $\gamma=1$ ，这不影响推导；3）仅针对特定的 τ 分位，不同分位的投影算子互不影响，因此只需要分析一个即可；4）原来的分布任务是单dirac分布，其他情况可以看做该分布的加和。

这样可以简化为如下引理

Lemma 3. *Consider an MDP with countable state and action spaces. Let Z, Y be value distributions such that each state-action distribution $Z(x,a)$, $Y(x,a)$ is given by a single Dirac. Consider the particular case where rewards are identically 0 and $\gamma = 1$, and let $\tau \in [0,1]$. Denote by Π_τ the projection operator that maps a probability distribution onto a Dirac delta located at its τ^{th} quantile. Then*

$$\bar{d}_\infty(\Pi_\tau \mathcal{T}^\pi Z, \Pi_\tau \mathcal{T}^\pi Y) \leq \bar{d}_\infty(Z, Y)$$

首先左手边可以只考虑一个状态 (x',a')，这样右手边只需要考虑其一步可达的状态 $\{(x_i,a_i)\}_{i\in I}$。由于 Z,Y 都是单个的dirac分布，右手边就可以直接写成 $\max_i |\theta_i - \psi_i|$，$\theta_i,\psi_i$ 分布代表不同状态随机变量 Z,Y 的分布。通过算子作用之后，$\mathcal{T}^*Z,\mathcal{T}^*Y$ 变成了多个dirac分布的加权和。使用反证法，即左手边大于 $|\theta_i - \psi_i|,\forall i$，那么左手边选出来的分位数肯定来自右手边不同的状态，但考虑到左手边都按照相同的 τ 取分位数，可以推出矛盾。

该引理可以这么理解，Bellman算子相当于是把各个不同状态的分布做了加权和，不同分布的加权和只会让分位数更加收缩。

实验

Policy evaluation 实验使用 gridworld 环境，说明 quantile regression 能够很好地拟合出多模的概率分布；Control 的实验使用 Atari 环境，效果相比之前有些提升。

参考文献

[1] Rowland, Mark, et al. "An analysis of categorical distributional reinforcement learning."*arXiv preprint arXiv:1802.08163*(2018).

编辑于 2019-03-31

强化学习 (Reinforcement Learning)