Nonlinear Distributional Gradient Temporal Difference Learning

Chao Qu [1]   Shie Mannor [2]   Huan Xu [3]

[1]Ant Financial Services Group

[2]Faculty of Electrical Engineering, Technion

[3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Tech

The distributional reinforcement learning has gained much attention recently [Bellemare et al., 2017]. It explicitly considers the stochastic nature of the long term return $Z(s, a)$.

The recursion of $Z(s, a)$ is described by the distributional Bellman equation,

$$Z(s, a) \overset{D}{=} R(s, a) + \gamma Z(s', a'),$$

where $\overset{D}{=}$ stands for "equal in distribution"

The distributional reinforcement learning has gained much attention recently [Bellemare et al., 2017]. It explicitly considers the stochastic nature of the long term return $Z(s, a)$.

The recursion of $Z(s, a)$ is described by the distributional Bellman equation,

$$Z(s, a) \stackrel{D}{=} R(s, a) + \gamma Z(s', a'),$$

where $\stackrel{D}{=}$ stands for "equal in distribution"

Distributional gradient temporal differenct learning

We consider a distributional counterpart of Gradient Temporal Difference Learning [Sutton et al., 2008].

Properties:

- Convergence in the off-policy setting.
- Convergence with the nonlinear function approximation.
- Include distributional nature of the long term reward.

Distributional gradient temporal differenct learning

We consider a distributional counterpart of Gradient Temporal Difference Learning [Sutton et al., 2008].

Properties:

- Convergence in the off-policy setting.
- Convergence with the nonlinear function approximation.
- Include distributional nature of the long term reward.

To measure the distance between distributions $Z(s, a)$ and $\mathcal{T}Z(s, a)$, we need to introduce Cramér distance.

Suppose there are two distributions P and Q and their cumulative distribution functions are F_P and F_Q respectively, then the square root of Cramér distance between P and Q is

$$\ell_2(P, Q) := \Big(\int_{-\infty}^{\infty} (F_P(x) - F_Q(x))^2 dx \Big)^{1/2}.$$

To measure the distance between distributions $Z(s, a)$ and $\mathcal{T}Z(s, a)$, we need to introduce Cramér distance.

Suppose there are two distributions P and Q and their cumulative distribution functions are F_P and F_Q respectively, then the square root of Cramér distance between P and Q is

$$\ell_2(P, Q) := \Big(\int_{-\infty}^{\infty} (F_P(x) - F_Q(x))^2 dx\Big)^{1/2}.$$

Denote the (cumulative) distribution function of $Z(s)$ as $F_\theta(s,z)$, $G_\theta(s,z)$ as the distribution function of $\mathcal{T}Z(s)$.

D-MSPBE:

$$\underset{\theta}{\text{minimize:}} \quad J(\theta) := \|\Phi_\theta^T D(F_\theta - G_\theta)\|^2_{(\Phi_\theta^T D\Phi_\theta)^{-1}},$$

Denote the (cumulative) distribution function of $Z(s)$ as $F_\theta(s, z)$, $G_\theta(s, z)$ as the distribution function of $\mathcal{T}Z(s)$.

D-MSPBE:

$$\underset{\theta}{\text{minimize:}} \quad J(\theta) := \|\Phi_\theta^T D(F_\theta - G_\theta)\|^2_{(\Phi_\theta^T D \Phi_\theta)^{-1}},$$

- Value distribution ($F_\theta(s,z)$) is discrete within the range $[V_{\min}, V_{\max}]$ with m atoms.
- $\phi_\theta(s,z) = \frac{\partial F_\theta(s,z)}{\partial \theta}$ and $(\Phi_\theta)_{((i,j),l)} = \frac{\partial}{\partial \theta_l} F_\theta(s_i, z_j)$.
- Project onto the space spanned by Φ w.r.t. the Cramér distance and then obtain D-MSPBE.
- SGD and weight duplication trick to optimize it.

Distributional GTD2

Input: step size α_t, step size β_t, policy π.
for $t = 0, 1, \ldots$ **do**

$$w_{t+1} = w_t + \beta_t \sum_{j=1}^{m} \big(- \phi_{\theta_t}^T(s_t, z_j) w_t + \boldsymbol{\delta_{\theta_t}} \big) \phi_{\theta_t}(s_t, z_j)$$

$$\theta_{t+1} = \Gamma[\theta_t + \alpha_t \{ \sum_{j=1}^{m} \big(\phi_{\theta_t}(s_t, z_j) - \boldsymbol{\phi_{\theta_t}(s_{t+1}, \frac{z_j - r_t}{\gamma}}) \big)$$

$$\phi_{\theta_t}^T(s_t, z_j) w_t - h_t \}]$$

$\Gamma : \mathbb{R}^d \to \mathbb{R}^d$ is a projection onto an compact set C with a smooth boundary.
$h_t = \sum_{j=1}^{m} (\boldsymbol{\delta_{\theta_t}} - w_t^T \phi_{\theta_t}(s_t, z_j)) \nabla^2 F_{\theta_t}(s_t, z_j) w_t$,
where $\boldsymbol{\delta_{\theta_t}} = F_{\theta_t}(s_{t+1}, \frac{z_j - r_t}{\gamma}) - F_{\theta_t}(s_t, z_j)$.
end for

Some remarks:

- Use the temporal distribution difference $\boldsymbol{\delta_{\theta_t}}$ instead of the temporal difference in GTD2.
- Summation over z_j, which corresponds to the integral in the Cramér distance.
- h_t results from the nonlinear function approximation, which is zero in the linear case. it can be evaluated using forward and backward propagation.

Theoretical Result

Theorem

Let $(s_t, r_t, s'_t)_{t\geq 0}$ be a sequence of transitions. The positive step-sizes in the algrithm satisfy $\sum_{t=0}^{\infty} a_t = \infty$, $\sum_{t=0}^{\infty} \beta_t = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2, \sum_{t=1}^{\infty} \beta_t^2 < \infty$ and $\frac{\alpha_t}{\beta_t} \to 0$, as $t \to \infty$. Assume that for any $\theta \in C$ and $s \in \mathcal{S}$ s.t. $d(s) > 0$, F_θ is three times continuously differentiable. Further assume that for each $\theta \in C$, $\left(\mathbb{E} \sum_{j=1}^{m} \phi_\theta(s, z_j)\phi_\theta^T(s, z_j)\right)$ is nonsingular. Then the Algorithm converges with probability one, as $t \to \infty$.

Distributional Greedy GQ

Input: step size α_t, step size β_t, $0 \leq \eta \leq 1$

for $t = 0, 1, \ldots$ **do**

$Q(s_{t+1}, a) = \sum_{j=1}^{m} z_j p_j(s_t, a)$, where $p_j(s_t, a)$ is the density function w.r.t. $F_\theta((s_t, a))$. $a^* = \arg\max_a Q(s_{t+1}, a)$.

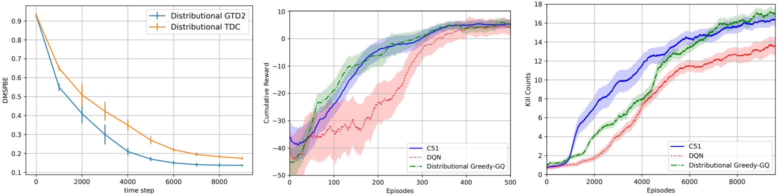
$$w_{t+1} = w_t + \beta_t \sum_{j=1}^{m} \big(-\phi_{\theta_t}^T((s_t, a_t), z_j) w_t + \delta_{\theta_t} \big) \times \phi_{\theta_t}((s_t, a_t), z_j).$$

$$\theta_{t+1} = \theta_t + \alpha_t \{ \sum_{j=1}^{m} \big(\delta_{\theta_t} \phi_{\theta_t}((s_t, a_t), z_j) - \eta \phi_{\theta_t}((s_{t+1}, a^*), \frac{z_j - r_t}{\gamma}) \big) \big(\phi_{\theta_t}^T((s_t, a_t), z_j) w_t \big) \big) \}.$$

where $\delta_{\theta_t} = F_{\theta_t}((s_{t+1}, a^*), \frac{z_j - r_t}{\gamma}) - F_{\theta_t}((s_t, a_t), z_j)$.

end for

Experimental Result



Thank you!

Visit our poster today at pacific Ballroom #33.