A Note on Softmax Deep Double Deterministic Policy Gradients

Chunyao Chiu
Department of Electronics and Electrical Engineering
National Yang Ming Chiao Tung University
david20571015.ee07@nycu.edu.tw

1 Introduction

Deep Deterministic Policy Gradients (DDPG), Lillicrap et al. [2015], use the actor-critic method to learn a deterministic policy for an environment with continuous action space. However, it is suffered from the overestimation problem. To deal with the problem, Fujimoto et al. [2018] induced the Twin Delayed Deep Deterministic Policy Gradient (TD3) which used two critic estimators. Though it avoids overestimation in DDPG, it leads to underestimation problem conversely.

This paper, Pan et al. [2020], purposes to use the Boltzmann softmax operator for value function estimation. Also, they raise Softmax Deep Deterministic Policy Gradients (SD2) and Softmax Deep Double Deterministic Policy Gradients (SD3) by apply the softmax operator on the estimated value from single and double estimators to reduce the absolute overestimation and underestimation bias. Furthermore, this paper provides a proof that the difference between the value function induced by the softmax operator and the optimal one is bounded in continuous action space, which has been proved in discrete action space, Silver et al. [2014].

2 Problem Formulation

2.1 Notations

- A Markov decision process (MDP) for the reinforcement learning problem is defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, r, p, \gamma)$.
  - $\mathcal{S}$, $\mathcal{A}$: The set of states and action. Assume the action space is continuous and bounded.
  - r : Reward. Assume the reward is continuous and bounded.
  - p : Transition probability.
  - γ : Discount factor.
- $\pi(\cdot;\phi)$ denotes the policy π parameterized by ϕ.
- $Q(\cdot,\cdot;\theta)$ denotes the Q-function parameterized by θ.
- ϕ^-, θ^- denote the parameters of the target networks for the actor and critic respectively.
- $\mathcal{T}(\cdot)$ denotes the value estimation function used to estimate the target Q-value $r + \gamma\mathcal{T}(s')$ from state s'.
- ρ denotes the sample distribution from the replay buffer.
- α denotes the learning rate.
- The Boltzmann softmax operator in continuous action space is defined as

$$\text{softmax}_\beta(Q(s,\cdot)) = \int_{a\in A} \frac{\exp(\beta Q(s,a))}{\int_{a'\in A}\exp(\beta Q(s,a'))da'} Q(s,a)da$$

  where β is the parameter of the softmax operator.

- Value iteration with the softmax operator is defined as

$$Q_{t+1}(s,a) = r_t(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V_t(s')]$$
$$V_{t+1}(s) = \mathrm{softmax}_\beta(Q_{t+1}(s,\cdot))$$

2.2 Preliminaries

- The DDPG method updates its critic according to

$$\theta' = \theta + \alpha \mathbb{E}_{s,a\sim\rho}(r + \gamma \mathcal{T}_{DDPG}(s') - Q(s,a;\theta))\nabla_\theta Q(s,a;\theta)$$

  where $\mathcal{T}_{DDPG}(s') = Q(s', \pi(s';\phi^-);\theta^-)$

3 Theoretical Analysis

3.1 Softmax Deep Deterministic Policy Gradients (SD2)

SD2 uses the softmax operator upon DDPG to estimate the value function, then update the critic by

$$\theta' = \theta + \alpha \mathbb{E}_{s,a\sim\rho}(r + \gamma \mathcal{T}_{SD2}(s') - Q(s,a;\theta))\nabla_\theta Q(s,a;\theta)$$

where $\mathcal{T}_{SD2}(s') = \mathrm{softmax}_\beta(Q(s',\cdot;\theta^-))$.

In practice, the softmax operator involves the integral which is difficult to handle in continuous action space. Thus we can obtain an unbiased estimation of $\mathcal{T}_{SD2}(s')$ by importance sampling

$$\mathbb{E}_{a'\sim p}[\frac{\exp(\beta Q(s',a';\theta^-))Q(s',a';\theta^-)}{p(a')}]/\mathbb{E}_{a'\sim p}[\frac{\exp(\beta Q(s',a';\theta^-))}{p(a')}]$$

where $p(a')$ is the probability density function of a Gaussian distribution.

To improve the exploration, they add a noise $\epsilon \sim \mathcal{N}(0,\sigma)$ and clip it to $[-c,c]$ to obtain a sampled action in $[\pi(s';\phi^-) - c, \pi(s';\phi^-) + c]$. The clipping operation is to limit the variance as $1/p(a')$ can be very large.

Algorithm 1 SD2

```
Initialize the critic network $Q$ and the actor network $\pi$ with random parameters $\theta$, $\phi$
Initialize target networks $\theta^- \leftarrow \theta$, $\phi^- \leftarrow \phi$
Initialize replay buffer $\mathcal{B}$
for $t = 1$ to $T$ do
    Select action $a$ with exploration noise $\epsilon \sim \mathcal{N}(0,\sigma)$ based on $\pi$
    Execute action $a$, observe reward $r$, new state $s'$ ans done $d$
    Store transition tuple $(s,a,r,s',d)$ in $\mathcal{B}$
    Sample a mini-batch of N transitions $(s,a,r,s',d)$ from $\mathcal{B}$
    Sample $K$ noises $\epsilon \sim \mathcal{N}(0,\overline{\sigma})$
    $\hat{a}' \leftarrow \pi(s';\phi^-) + \mathrm{clip}(\epsilon,-c,c)$
    $\mathrm{softmax}_\beta(Q(s',\cdot;\theta^-)) \leftarrow \mathbb{E}_{\hat{a}'\sim p}[\frac{\exp(\beta Q(s',\hat{a}';\theta^-))Q(s',\hat{a}';\theta^-)}{p(\hat{a}')}]/\mathbb{E}_{\hat{a}'\sim p}[\frac{\exp(\beta Q(s',\hat{a}';\theta^-))}{p(\hat{a}')}]$
    $y_i \leftarrow r + \gamma(1-d)\mathrm{softmax}_\beta(Q(s',\cdot;\theta^-))$
    Update the parameter $\theta$ of the critic according to Bellman loss: $\frac{1}{N}\sum_s(Q(s,a;\theta) - y)^2$
    Update the parameter $\phi$ of the actor by policy gradient: $\frac{1}{N}\sum_s[\nabla_\phi(\pi(s;\phi))\nabla_a Q(s,a;\theta)|_{a=\pi(s;\phi)}]$
    Update target networks: $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$, $\phi^- \leftarrow \tau\phi + (1-\tau)\phi^-$
end for
```

3.2 Softmax Deep Double Deterministic Policy Gradients (SD3)

TD3 maintains double critics with clipping the Q-value by the original Q-value to prevent overestimation. Specifically, it estimates the value function by taking the minimum from the original Q-value $y_1, y_2 = r + \gamma \min_{i=1,2}(Q_i(s', \pi(s';\phi^-);\theta_i^-)$. However, it leads to underestimation bias, and affect the performance.

The paper propose to estimate the target value for critic Q_i by $y_i = r + \gamma \mathcal{T}_{SD3}(s')$, where $\mathcal{T}_{SD3}(s') = \mathrm{softmax}_\beta(\hat{Q}_i(s',\cdot))$ and $\hat{Q}_i(s',a') = \min(Q_i(s',a';\theta_i^-), Q_{-i}(s',a';\theta_{-i}^-))$. The value of $\mathcal{T}_{SD3}(s')$ can be obtained by the same way as SD2 in Section 2.3.

Algorithm 2 SD3

```
Initialize the critic network $Q_1$, $Q_2$ and the actor network $\pi_1$, $\pi_2$ with random parameters $\theta_1$, $\theta_2$, $\phi_1$, $\phi_2$
Initialize target networks $\theta_1^- \leftarrow \theta_1, \theta_2^- \leftarrow \theta_2, \phi_1^- \leftarrow \phi_1, \phi_2^- \leftarrow \phi_2$
Initialize replay buffer $\mathcal{B}$
for $t = 1$ to $T$ do
    Select action $a$ with exploration noise $\epsilon \sim \mathcal{N}(0, \sigma)$ based on $\pi_1$ and $\pi_2$
    Execute action $a$, observe reward $r$, new state $s'$ ans done $d$
    Store transition tuple $(s, a, r, s', d)$ in $\mathcal{B}$
    for $i = 1, 2$ do
        Sample a mini-batch of N transitions $(s, a, r, s', d)$ from $\mathcal{B}$
        Sample $K$ noises $\epsilon \sim \mathcal{N}(0, \overline{\sigma})$
        $\hat{a}' \leftarrow \pi_i(s'; \theta^-) + \text{clip}(\epsilon, -c, c)$
        $\hat{Q}(s', \hat{a}') \leftarrow \min_{j=1,2}(Q_j(s', \hat{a}'; \theta_j^-))$
        $\text{softmax}_\beta(\hat{Q}(s', \cdot)) \leftarrow \mathbb{E}_{\hat{a}' \sim p}[\frac{\exp(\beta\hat{Q}(s',\hat{a}'))\hat{Q}(s',\hat{a}')}{p(\hat{a}')}]/\mathbb{E}_{\hat{a}' \sim p}[\frac{\exp(\beta\hat{Q}(s',\hat{a}'))}{p(\hat{a}')}]$
        $y_i \leftarrow r + \gamma(1 - d)\text{softmax}_\beta(\hat{Q}(s', \cdot))$
        Update the critic $\theta_i$ according to Bellman loss: $\frac{1}{N}\sum_s(Q_i(s, a; \theta_i) - y_i)^2$
        Update the actor $\phi_i$ by policy gradient: $\frac{1}{N}\sum_s[\nabla_{\phi_i}(\pi(s; \phi_i))\nabla_a Q_i(s, a; \theta_i)|_{a=\pi(s;\phi_i)}]$
        Update target networks: $\theta_i^- \leftarrow \tau\theta_i + (1 - \tau)\theta_i^-$, $\phi_i^- \leftarrow \tau\phi_i + (1 - \tau)\phi_i^-$
    end for
end for
```

3.3 Upper bound of the difference between max operator and softmax operator

Theorem 1 *Let* $\mathcal{C}(Q, s, \epsilon) = \{a|a \in A, Q(s, a) \geq max_a Q(s, a) - \epsilon\}$ *and* $F(Q, s, \epsilon) = \int_{a \in C(Q,s,\epsilon)} 1da$ *for any* $\epsilon 0$ *and any state s. The difference between the max operator and the softmax operator is* $0 \leq max_a Q(s, a) - softmax_\beta(Q(s, \cdot)) \leq \frac{\int_{a\in A} 1da - 1 - \ln F(Q,s,\epsilon)}{\beta} + \epsilon$.

3.4 Upper bound of the difference between softmax value function and optimal value function

Theorem 2 *For any iteration t, the difference between the optimal value function* V^* *and the value function induced by softmax value iteration at the t-th iteration* V_t *satisfies:*

$$||V_t - V^*||_\infty \leq \gamma^t ||V_0(s) - V^*(s)||_\infty + \frac{1}{1-\gamma}\frac{\beta\epsilon + \int_{a\in A} 1da - 1}{\beta} - \sum_{k=1}^{t} \gamma^{t-k} \frac{\min_s \ln F(Q_k, s, \epsilon)}{\beta}$$

For any $\epsilon > 0$, the error between the value function induced by the softmax operator and the optimal can be bounded, which converges to $\epsilon/(1 - \gamma)$, and can be arbitrarily close to 0 as β approaches to infinity.

3.5 SD2 enables a better value estimation by reducing the overestimation bias in DDPG

Theorem 3 *Denote the bias of the value estimate and the true value induced by* $\mathcal{T}$ *as* $\text{bias}(\mathcal{T}) = \mathbb{E}[\mathcal{T}(s')] - \mathbb{E}[Q(s', \pi(s'; \phi^-); \theta^{true})]$. *Assume that the actor is a local maximizer with respect to the critic, then there exists noise clipping parameter* $c > 0$ *such that* $\text{bias}(\mathcal{T}_{SD2}) \leq \text{bias}(\mathcal{T}_{DDPG})$.

3.6 SD3 improves the underestimation bias from TD3

Theorem 4 *Denote* $\mathcal{T}_{TD3}$, $\mathcal{T}_{SD3}$ *the value estimation functions of TD3 and SD3 respectively. then we have* $\text{bias}(\mathcal{T}_{TD3}) \leq \text{bias}(\mathcal{T}_{SD3})$.

4 Conclusion

This paper induced some advantage to use the softmax operator in continuous control.

- Provide a new analysis for the error bound between the value function induced by the softmax operator and the optimal in continuous action space.
- Show that the softmax operator reduces the overestimation bias of DDPG.
- Show that the softmax operator improve the underestimation bias of TD3.

And the author also propose some potential future work.

- Adaptive scheduling of the parameter β in SD2 and SD3.
- Quantify the bias reduction for overestimation and underestimation.

References

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Ling Pan, Qingpeng Cai, and Longbo Huang. Softmax deep double deterministic policy gradients. *Advances in Neural Information Processing Systems*, 33:11767–11777, 2020.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.