Improving exploration in reinforcement learning with temporally correlated stochasticity

Dongqi Han(P)[1]

[1] Cognitive Neurorobotics Research Unit, Okinawa Institute of Science and Technology
E-mail: dongqi.han@oist.jp

Abstract— Reinforcement learning is a useful approach to solve machine learning problems by self-exploration when training samples are not provided. However, researchers usually ignore the importance of the choice of exploration noise. In this paper, I show that temporally self-correlated exploration stochasticity, generated by Ornstein-Uhlenbeck process, can significantly enhance the performance of reinforcement learning tasks by improving exploration.

Keywords— Machine learning, reinforcement learning, Ornstein-Uhlenbeck process, stochastic exploration

1 Introduction

Reinforcement learning (RL) has recently been a powerful solution to difficult tasks where well-defined sample data is not available. State-of-art RL algorithms have achieved human-level or superhuman performance in a variety of task such as the game of Go[1], video games[2, 3], and robotic controls[4].

In RL, the agents require self-exploration to extract sample data by interacting with environment. In RL studies, exploration can be done using various methods, such as ϵ-greedy[2] and actor-critic[3]. Conventionally, researchers usually use random white noise for stochastic exploration, which is not temporally self-correlated. It remains to address that how auto-correlated stochasticity affects exploration efficiency.

2 Temporally correlated stochasticity

Temporally correlated noise can be obtain by a continuous stochastic process, such as Wiener process or by the moving average of white noise[5]. One of these stochastic processes, known as Ornstein-Uhlenbeck process(OU-process)[6], is defined by the following equation:

$$dx_t = \theta(\mu - x_t)\,dt + \sigma\sqrt{2\theta}\,dW_t \tag{1}$$

where $dW_t = \xi(t)\,dt$, and $\xi(t)$ is sampled from Gaussian white noise. Parameter μ denotes the expected mean of x_t, while σ is the standard-deviation and θ indicates the inverse of its auto-correlation timescale. The distribution of x_t is Gaussian. In this study, I set $dt = 1$, which is a reasonable approximation when $\theta \ll 1$, and also can be regarded as a special case of autoregression model. I call the trajectory of x_t as "OU noise". A comparison between OU noise and Gaussian white noise is showed in **Figure** 1 (a-d).

The auto-correlation of OU noise is:

$$\frac{1}{\sigma^2}\langle(x(t_0) - \mu)(x(t_0 + t) - \mu)\rangle = \exp(-\theta|t|) \tag{2}$$

Algorithm 1: ϵ-greedy Q-learning with temporally correlated exploration

```
Initialize the table or function approximator for
  action state value $\hat{Q}$ and the hidden variable $h$;
while $episode{+}{+} < max\ episode$ do
    Initialize the environment;
    while $task\ not\ done$ do
        Sample a random number $\xi \sim N(0,1)$ ;
            $h \leftarrow h - \theta h + \sqrt{2\theta}\xi$
        if $|h| < \epsilon$ then Sample a temporally
          correlated stochastic action $a(h)$;
        else Select action $a = \text{argmax}(\hat{Q}(s,\cdot))$;
        Execute and record the state transition
          $(s, a, s', r)$;
        Compute the target Q value:
            $Q_{target}(s,a) = r + \gamma \max[\hat{Q}(s',\cdot)]$
        Update the table or approximator for $Q$;
    end
end
```

With this temporal correlation, it is obvious that the range of exploration will be larger by larger auto-correlation timescale, without changing the variance of the noise. Consider a 2-dimensional random walk problem, where the step size (velocity) at x and y direction is sampled from either Gaussian white noise or OU noise. As **Figure** 1(e) shows, the range of random walk is enlarged when the temporally correlated OU noise is used.

3 Related works

Two recent studies[4, 5] utilized temporally correlated noise to perform exploration in control tasks with continuous actions space, with actor-critic algorithms. However, they did not show the results that temporally correlated noise significantly outperforms white noise. Neither did they extend the stochastic action selection into discrete action space.

In this article, I propose the OU-process as the exploration stochasticity in discrete action space. I will show that the temporal correlation in noise significantly increase the efficiency of exploration and thus the performance of RL.

4 Algorithm

I call the algorithm used in this paper "ϵ-greedy Q-learning with temporally correlated exploration"(**Al-**

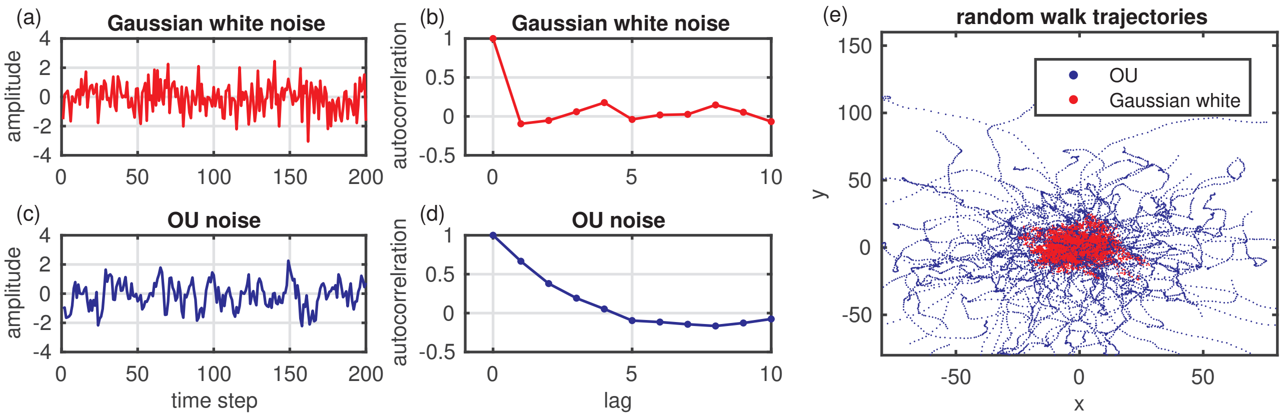

Figure 1: Comparison between Gaussian white noise($\sigma = 1$) and OU noise($\theta = 0.3, \sigma = 1$) . (a) An example trajectory of Gaussian white noise, within 200 time steps. (b) Auto-correlation of the Gaussian white noise in (a). (c-d) Same as (a-c), but an example of OU noise plotted. (e) Trajectory of a random walk on a 2-D plane, where the velocity is sampled from 2-D Gaussian white noise or OU noise. 100 trials of 100 steps are plotted.

gotithm 1). To generate temporally correlated noise, a hidden variable h is used. This algorithm is mostly the same as standard ϵ-greedy Q-learning[7] but the stochastic action for exploration is temporally self-correlated. Note that there are various ways to sample a temporally correlated stochastic action, one of them is used in this study: Assuming the action space is $\{-1, 1\}$, then $a = \text{sign}(h)$ can be sampled as the auto-correlated stochastic action.

5 Experimental results

I performed the RL experiments with ϵ-greedy Q-learning with temporally correlated exploration, compared with the same algorithm but using Gaussian white noise.

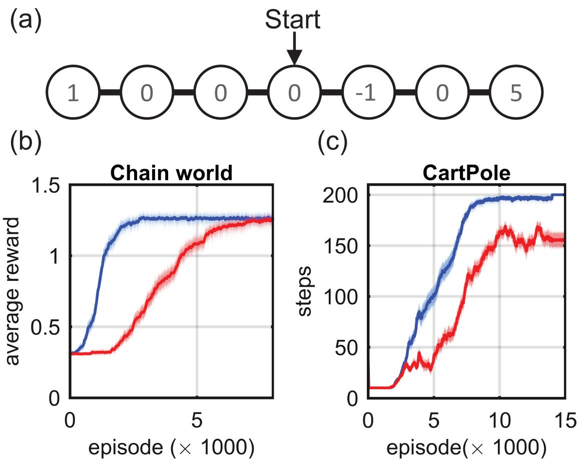

Figure 2: Experimental results. (a) Setting of chain world task, where the numbers indicate the reward at each state. (b,c) Performance comparison between the Q-learning using OU noise(blue, $\theta = 0.05$) and Gaussian white noise(red). The last 1000 episodes for cartpole are without training and all-greedy. Simulations were run for 100 trials.

In the first experiment, I tested a chain world task, as showed in **Figure** 2(a). I used a table for estimating Q value, which was updated for each step. The second experiment is the classic continuous control task "cartpole", where I use an single-layer perceptron to approximate the Q-value. State triansitions were recorded in a replay buffer, and the synaptic weights of the network were updated by randomly sampling a batch of data from the buffer at each time step.

In both two experiments, the agent using OU noise for exploration has achieved better performance than that using white noise. For the chain world, the reason is straightforward: the agent needs higher exploration range to reach the large reward at most right. But for cartpole, the case becomes more complicated. OU noise works better when learning rate is small and $\epsilon \sim 0.1$. Further detailed and systematic study will be the future direction.

The codes can be found at *https://github.com/oist-cnru/temporally_correlated_exploration_stochasticity.*

References

[1] Silver, D., Huang, A., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.

[2] Mnih, V., Kavukcuoglu, K., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529.

[3] Mnih, V., Badia, A. P., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. *In International Conference on Machine Learning* (pp. 1928-1937).

[4] Lillicrap, T. P., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint* arXiv:1509.02971.

[5] Wawrzynski, P. (2015). Control policy with autocorrelated noise in reinforcement learning for robotics. *International Journal of Machine Learning and Computing*, 5(2), 91.

[6] Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the Brownian motion. *Physical review*, 36(5), 823.

[7] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1). Cambridge: MIT press.