Improving exploration in reinforcement learning with temporally correlated stochasticity

Dongqi Han, Cognitive Neurorobotics Research Unit,
Okinawa Institute of Science and Technology

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.

Background & 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, video games, and robotic controls.
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 and actor-critic. 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.

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

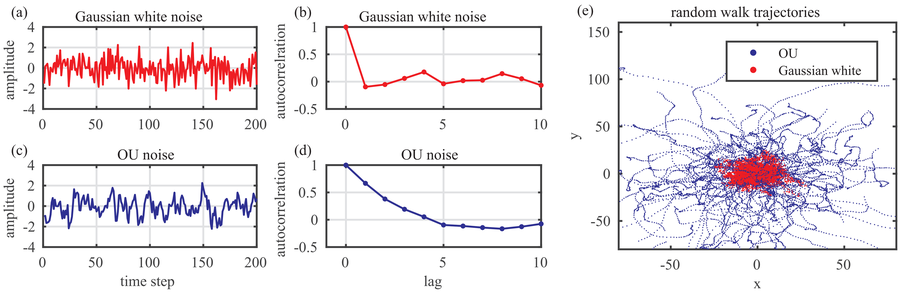
where $dW_t = \xi(t)\,dt$, and $\xi(t)$ is sampled from Gaussian white noise. Parameter μ denotes the expected mean of x_t, while σ is the standard-deviation and θ indicates the inverse of its auto-correlation timescale.

OU-process:

a temporally self-correlated random process

Q: Why we need temporally correlated noise?

A: Temporally correlated noise in motor bubbling, e.g. OU process, makes exploration range much larger while keep variance



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 $a = \mathrm{argmax}(\hat{Q}(s,\cdot))$;
    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

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

Experimental results

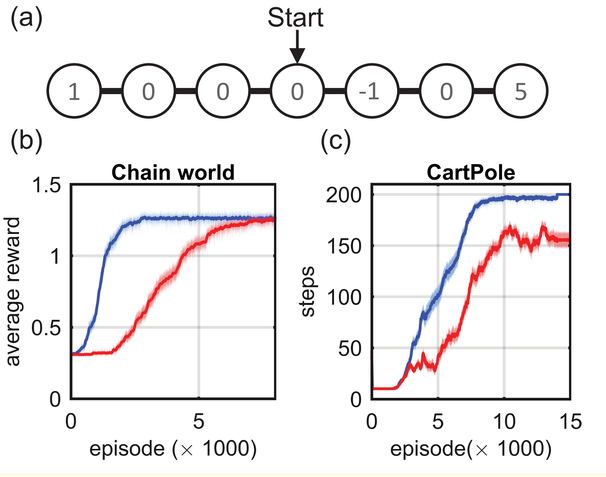


Figure: Experimental results. (a) Setting of chain world task, where the numbers indicate the reward at each state. (b,c) Performance comparison between the Q-learning using OU noise(blue, θ=0.05) and Gaussian white noise(red). The last 1000 episodes for cartpole are without training and all-greedy.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

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