Abstract

- Setting: Off-Policy reinforcement learning for continuous control
- Exploration is commonly performed by adding random perturbations to the actions or, equivalently, by sampling actions from a stochastic policy.
- This white noise exploration is often not sufficient to find high reward regions
- Structurally temporally correlated alternatives like Ornstein-Uhlenbeck (OU) noise, which try to tackle this issue, can inhibit learning when not necessary
- We examine the effectiveness of colored noise of intermediate temporal correlation
- Our results show that pink noise significantly outperforms white noise and OU noise across many tasks, and should be preferred as the default choice for action noise

Action Noise for Exploration

In off-policy RL, action noise \( \epsilon_t \sim \mathcal{N}(0, I) \) may be added to a deterministic policy:

\[
\epsilon_t = \mu(s) + \sigma(s) \epsilon_t
\]

or for sampling from a stochastic policy \( \pi(a | s) = \mathcal{N}(\mu(s), \sigma^2(s)) \):

\[
\epsilon_t = \mu(s) + \sigma(s) \epsilon_t
\]

In both cases, the noise signal \( \epsilon_t \) has no temporal correlation and is called white noise. Some tasks require stronger exploration and are better served by temporally correlated noises like Ornstein-Uhlenbeck (OU) noise:

\[
\epsilon_t = (1 - \Delta t) \epsilon_t + \sigma \mathcal{N}(0, \Delta t)
\]

For many tasks OU noise is too strongly correlated → idea: intermediate correlation

Colored Noise

A stochastic process is called colored noise with color parameter \( \beta \), if signals \( \epsilon(t) \) drawn from it have the property that \( |\mathcal{F}(\epsilon(t))| \propto f^{-\beta} \), where \( \mathcal{F}(\epsilon(t)) \) denotes the Fourier transform of \( \epsilon(t) \) and \( |\mathcal{F}(\epsilon(t))| \) is called the power spectral density.

The Power of Pink

What makes pink noise a better default than white noise or OU noise? We examine this question using two simple environments which mirror common dynamics:

1. A bounded integrator:

\[
\epsilon_t = \text{clip}(\epsilon_t + a_t, -c, +c) \quad \rightarrow \text{Parameterized by area } (4c^2)
\]

2. A harmonic oscillator:

\[
\dot{x} = F - k x \quad \rightarrow \text{Parameterized by resonant frequency } f = \frac{1}{2\pi \sqrt{m}}
\]

We now vary the parameters \( (c, f) \) over the complete sensible range (for episode lengths of \( T = 1000 \) and noise with \( \text{var}[\epsilon_t] = 1 \)) and measure the quality of exploration.

Takeaway: The intermediate temporal correlation makes pink noise more general.

- It is less sensitive to the environment parameterization than white noise / OU noise.
- If the parameterization (e.g. \( c \) or \( f \)) is unknown, pink noise is the best choice.
- This explains the good average performance on the benchmark experiments.

Obtainable Noise

White noise Pink noise OU noise

We recommend pink noise as the default choice for action noise in reinforcement learning for continuous control. pip install pink-noise-rl

References


Eleventh International Conference on Learning Representations (ICLR 2023) · Kigali, Rwanda

al.is.tuebingen.mpg.de