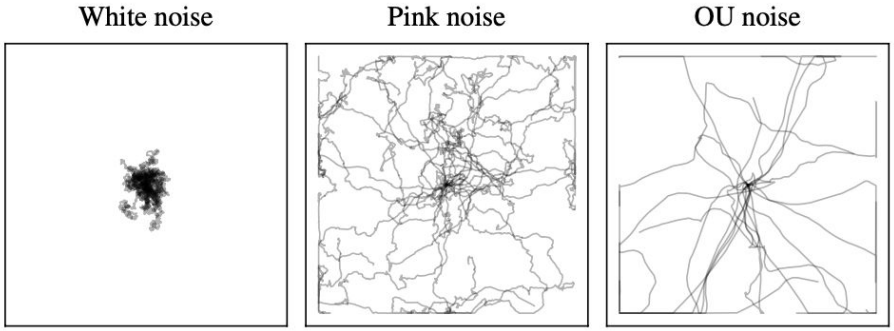
Pink Noise Is All You Need: Colored Noise Exploration in Deep Reinforcement Learning

Onno Eberhard, Jakob Hollenstein, Cristina Pinneri, Georg Martius

Горячева Екатерина Михайловна

Noises



WN:

- it *explores locally* -> if the environment was much smaller, white noise would be enough to cover the space
- identical to independently sampling from a Gaussian distribution at every time step

OU (Ornstein-Uhlenbeck):

- only *explores globally* and gets stuck at the edges
- can yield sufficient exploration to deal with hard cases, but it also introduces a different problem: strongly off-policy trajectories. Too much exploration is not beneficial for learning a good policy

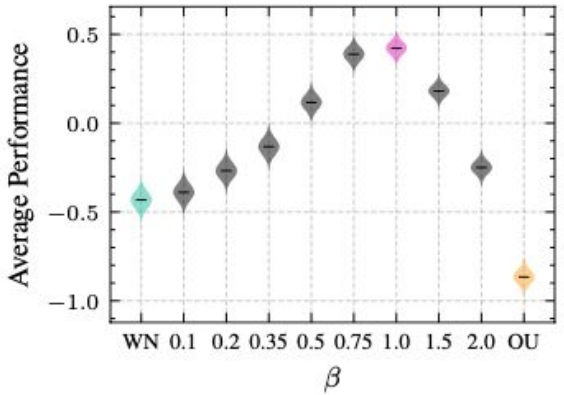
Pink:

- balance of local and global exploration, and covers the state space more uniformly
- parameter β to control the correlation strength

If an environment's dynamics are very complex, i.e. they contain many such individual parts, then the ideal action noise should score highly on each of these "sub-tasks"

Experiments

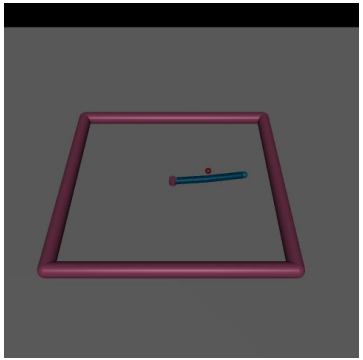
- with 20 different seeds
- normalize them to zero mean and unit variance
- β ∈ {0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2}



Environment	Best noise	p	Pink?
Pendulum	2.0	**0.01**	✗
Cartpole (b.)	1.0 (Pink)	—	✓
Cartpole (s.)	1.0 (Pink)	—	✓
Ball-In-Cup	0.75	0.88	✓
MountainCar	2.0	0.59	✓
Hopper	1.0 (Pink)	—	✓
Walker	0.5	0.36	✓
Reacher	White noise	**0.02**	✗
Cheetah	0.75	0.62	✓
Door	0.75	0.65	✓

Problems:

1. Oscillation
2. Pendulum: OU
   - strong exploration is necessary
   - underactuated and requires a gradual build-up of momentum by slowly swinging back and forth. Strongly correlated action noise, such as red noise, makes this behavior much more likely
3. Reacher: WN
   - has neither a particularly large state space, nor does it exhibit under-actuation, such that white noise is well suited to explore the space



Alternatives

A. Rollout's noise type is selected: a color-schedule going from $\beta = 2$ to $\beta = 0$

1. start with highly correlated red noise ($\beta = 2$)

2. slowly decrease β to white noise ($\beta = 0$) over the course of training

The rationale behind this strategy is that high-reward regions can be quickly discovered at the beginning of training when β is large, while the trajectories get more on-policy over time.

The results indicate that the schedule is generally better than OU and white noise, but does not outperform pink noise

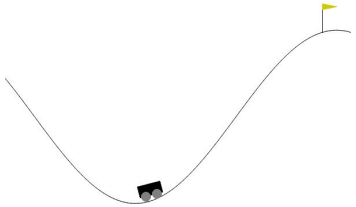
B. Bandit approach with the intention of finding the optimal color for an environment (to automatically adapt the noise to different stages of training)

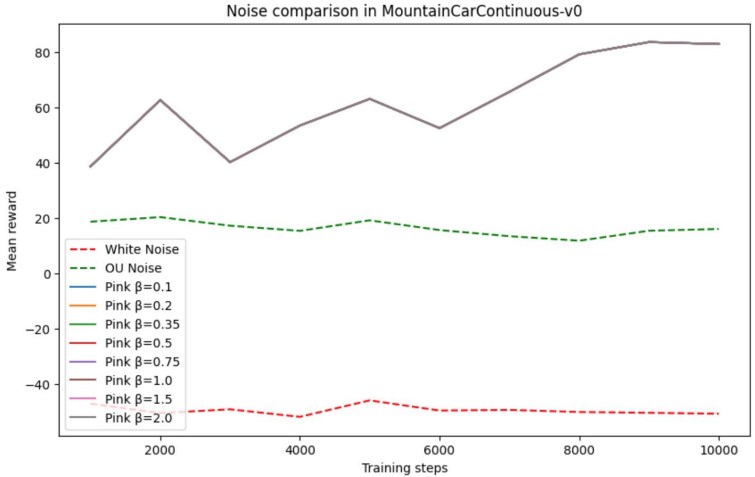
- The reasoning for this is that in environ ments where strong exploration is necessary (such as Pendulum and MountainCar), high return will only be achieved by strongly correlated actions. On the other hand, if environments do not require correlated actions, or a capable policy has been learned, the highest return should be achieved by the action noise which least disturbs the policy, i.e. noise with a low correlation.

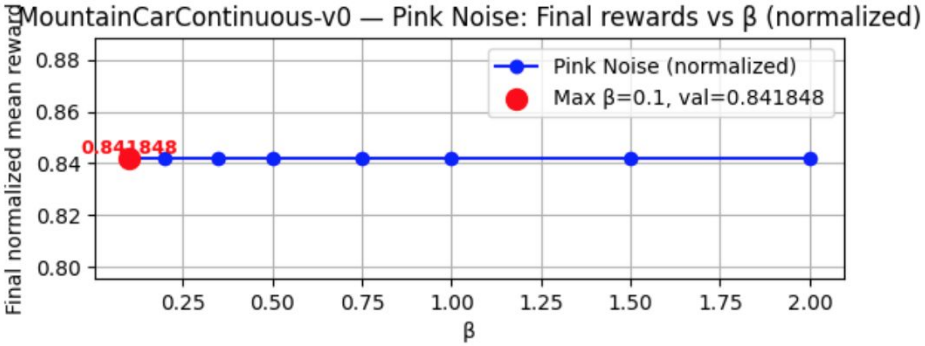
My results

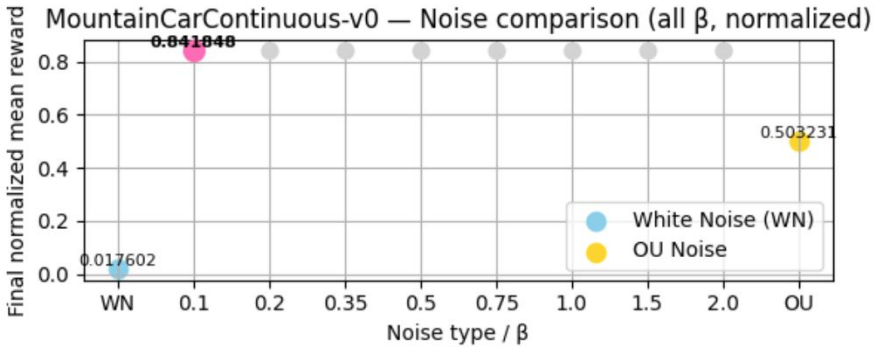
MountainCarContinuous-v0

Pink Noise beta=0.1





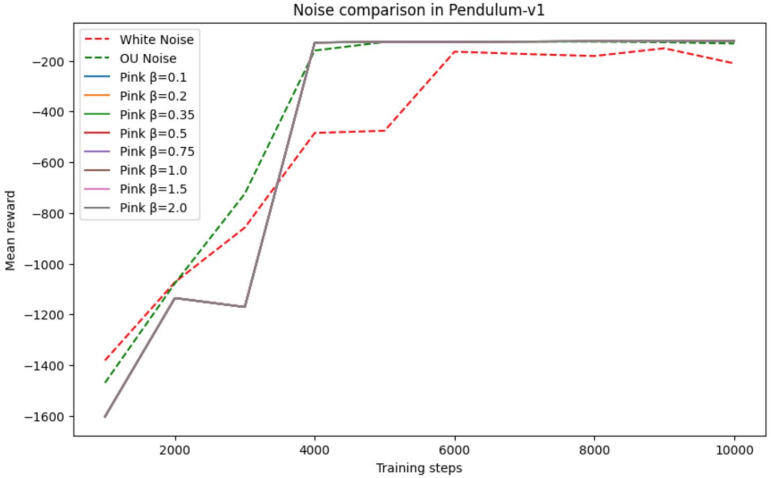


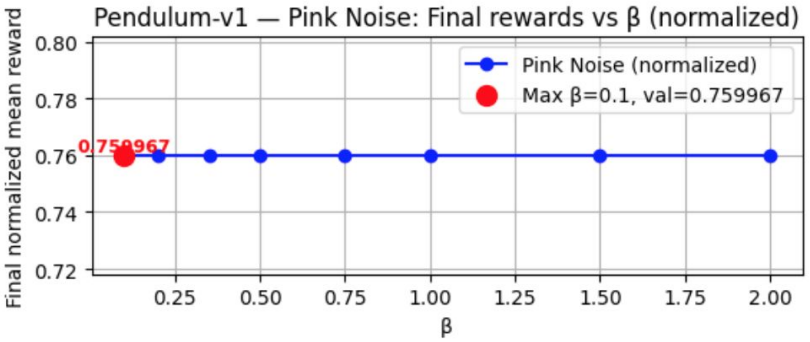


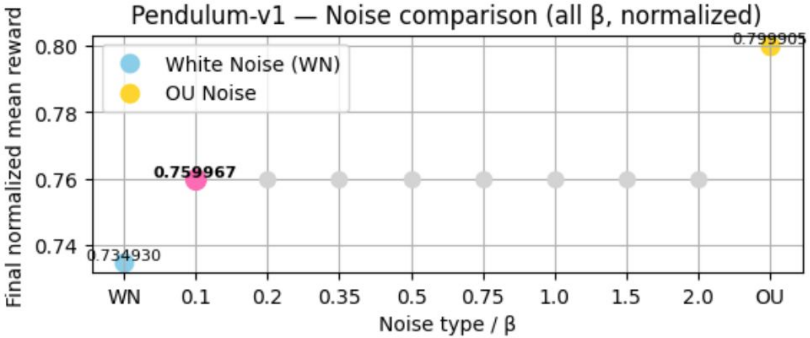
My results

Pendulum-v1

OU



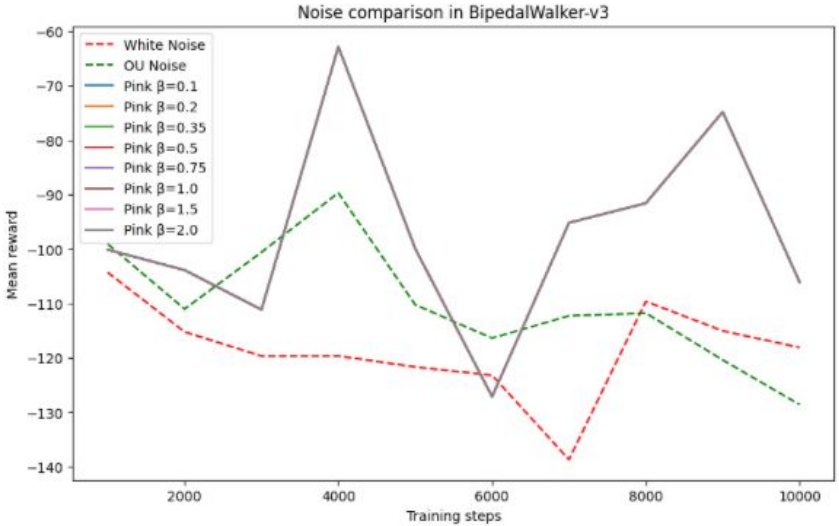


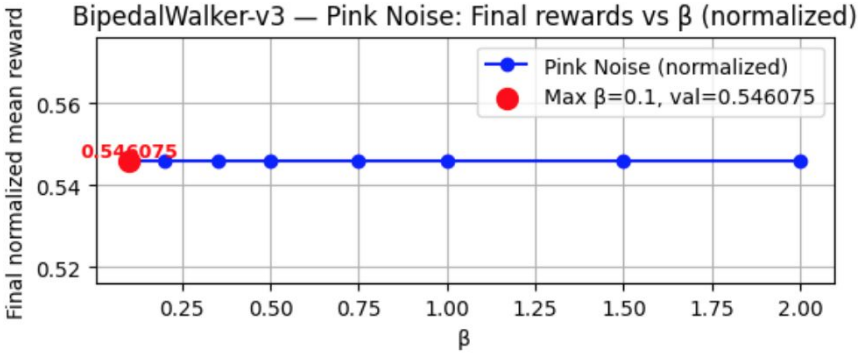


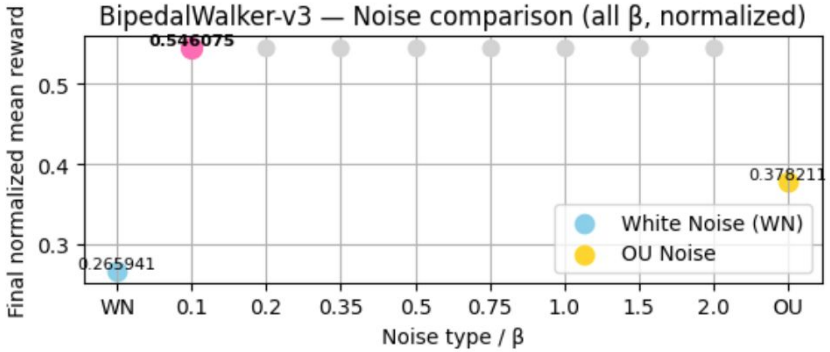
My results

BipedalWalker-v3

Pink Noise beta = 0.1



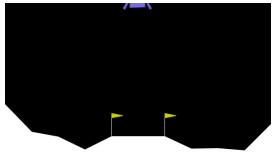


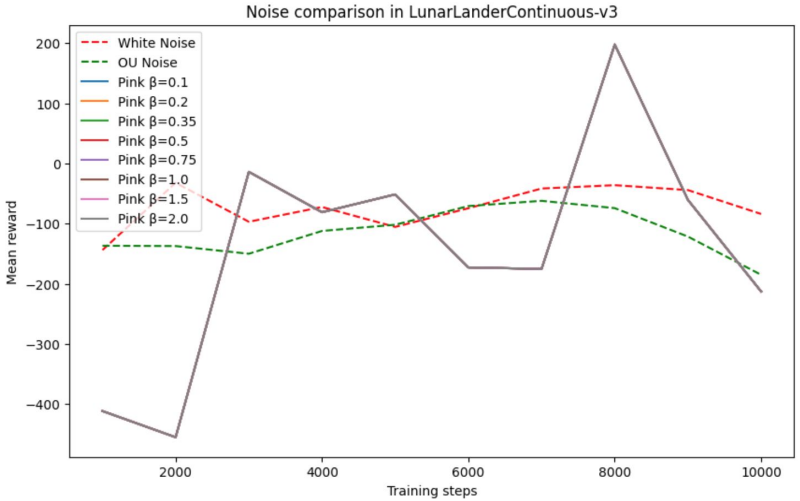


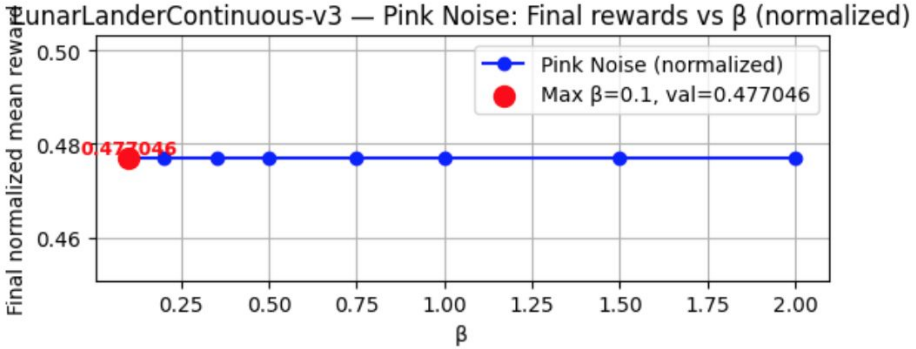
My results

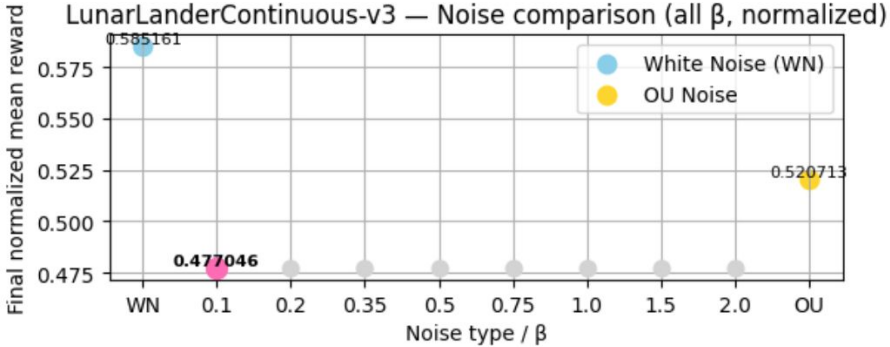
LunarLanderContinuous-v3

WN









Notebook with experiments:

https://colab.research.google.com/drive/1N1vhl6lOkBKe0KRczDuGt6u9_B4fiHK1?usp=sharing

GitHub with Pink Noise:

https://github.com/martius-lab/pink-noise-rl