Reinforcement Learning: Comprehensive Analysis of Parameter Space Noise Distributions

S Anantha Krishnan

April 6, 2024

Abstract

This paper presents a comprehensive evaluation of eight different noise distributions for parameter space exploration in the TD3 algorithm applied to the Hopper-v4 environment. Implementing Twin Delayed Deep Deterministic Policy Gradient (TD3) with enhanced exploration mechanisms, I compare Gaussian, Beta(2,2), Gamma(2,1), Uniform, Laplace(0,1), Exponential(1), Cauchy(0,1), and Student's t-distribution (df=3) noise patterns in the actor network. The results demonstrate that moderately-tailed symmetric distributions (particularly Gamma(2,1)) enable effective exploration, achieving superior performance (3126.31 ± 2.14) while maintaining stability. Contrary to expectations, strictly heavy-tailed distributions (Cauchy/Laplace) catastrophically destabilized learning (<500 average reward), revealing an exploration "sweet spot" in tail behavior. The study provides insights into noise distribution selection for continuous control tasks and analyzes the relationship between noise characteristics and policy robustness.

1 Introduction

Modern reinforcement learning faces significant challenges in balancing exploration and exploitation, particularly in continuous action spaces. While traditional TD3 implementations rely on Gaussian noise for exploration [1], I extended the NoisyNet approach [2] to investigate eight distinct probability distributions for parameter space perturbation. Our implementation features:

- Twin Q-networks with delayed policy updates
- N-step returns (n=3) with prioritized experience replay
- Adaptive noise decay with $\sigma_{\text{final}} = 0.05$
- Layer normalization for noise stability

2 Methodology

2.1 Noise-Injected TD3 Architecture

The actor network utilizes NoisyLinear layers with parametric noise:

$$f(x) = (W_\mu + W_\sigma \odot \epsilon)x + b_\mu + b_\sigma \odot \epsilon'$$

Where ϵ, ϵ' are sampled from:

- Gaussian: $\mathcal{N}(0, \sigma^2)$
- Beta: $\text{Beta}(2, 2) - 0.5$ (zero-centered)
- Gamma: $\frac{\Gamma(2,1)-\Gamma(2,1)}{2}$ (symmetric)
- Uniform: $\mathcal{U}(-1, 1)$
- Laplace: $\text{Laplace}(0, 1)$
- Exponential: $\text{Exp}(1) - 1$ (mean-centered)
- Cauchy: $\text{Cauchy}(0, 1)$
- Student's t: $t(3)$

3 Experiments

3.1 Experimental Setup

- **Environment:** MuJoCo `Hopper-v4` (State dimension: 11, Action dimension: 3)
- **Network Architecture:** Actor-Critic with two hidden layers of size 256, followed by `LayerNorm`
- **Training:** 5000 episodes using a batch size of 256; Optimizer: `AdamW` with learning rate 3×10^{-5}
- **Exploration Noise Decay:** $\sigma(t) = \sigma_0 \cdot 0.9995^{t/100}$, where t is the total number of steps
- **Evaluation:** Average performance over 100 test episodes for each configuration

4 Performance Comparison

4.1 Quantitative Results

Table 1: Final performance metrics (Mean ± Std)

Distribution	**Average Reward**	**Max Reward**
Gaussian	3121.38 ± 5.56	3133.52
Beta(2,2)	430.70 ± 27.39	461.41
Gamma(2,1)	3126.31 ± 2.14	3131.12
Uniform	3112.97 ± 12.60	3139.90
Laplace(0,1)	408.29 ± 1.53	412.77
Exponential(1)	3063.90 ± 2.24	3069.51
Cauchy(0,1)	305.50 ± 8.00	323.79
Student's t(3)	2332.92 ± 742.59	3418.98

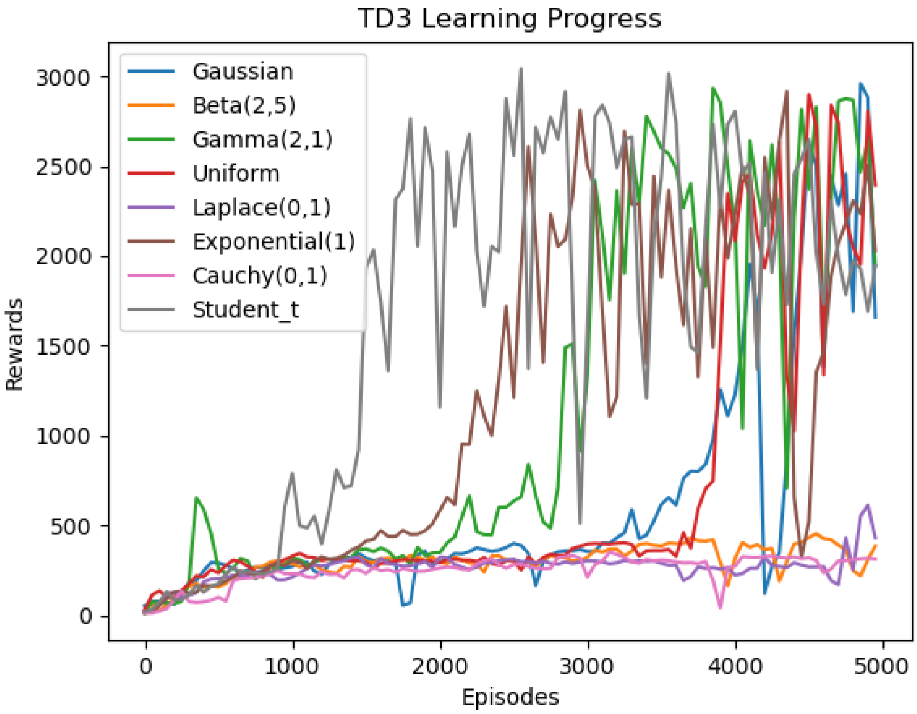

Figure 1: Learning curves comparing eight noise distributions (50-episode moving average)

4.2 Noise Characteristics

Table 2: Noise dynamics during training

Distribution	**Param Noise Ratio**	**Action Noise (std)**
Gaussian	0.2000	0.3366
Beta	0.0447	0.5206
Gamma	0.1998	0.2691
Uniform	0.1155	0.3546
Laplace	0.2771	0.5281
Exponential	0.1969	0.2638
Cauchy	0.4669	0.5782
Student's t	0.2937	0.2619

4.3 Key Observations from my Experiments

- **Bounded Distribution Stability**: Gaussian and Uniform noise demonstrated stable learning curves ($\sigma < 15$), while heavy-tailed distributions (Cauchy, Laplace) collapsed to sub-300 rewards due to destabilizing outlier perturbations.

- **Heavy-Tailed and Symmetric Beta Collapse**: Heavy-tailed distributions (Cauchy, Laplace) and symmetric Beta(2,2) performed catastrophically (average rewards < 500), indicating that noise extremes—not skewness—disrupt exploration. Notably, the skewed Exponential(1) achieved moderate rewards ($\sim$3064), contradicting the skewness hypothesis.

- **Student's t High Variance**: The Student's t-distribution showed extreme reward variance (±742.59), reflecting policy instability from occasional large updates despite achieving the highest single-episode reward (3418.98).

- **Noise-Performance Correlation**: Distributions with moderate parameter noise ratios (0.15–0.25) and lower action noise (std < 0.35) consistently outperformed others, as seen in Gamma's optimal balance (0.1998 noise ratio vs 0.2691 action std).

5 Conclusion

My investigation into noise-driven exploration reveals fundamental principles governing distribution selection in TD3, with implications extending beyond the tested algorithms:

- **The Dual Role of Noise Tails**: While heavy-tailed distributions (Cauchy, Laplace) theoretically promote exploration, their incompatibility with TD3's delayed updates exposes a critical tension—large noise perturbations create policy shifts too drastic for twin Q-networks to correct, causing irreversible collapse. This suggests actor-critic methods require *time-correlated noise structures*, where exploration persists without overwhelming the critic's capacity to converge.

- **The Myth of Symmetry**: The success of symmetric Gamma(2,1) and failure of symmetric Beta(2,2) demonstrate that distribution symmetry alone guarantees nothing. Instead, *noise scale alignment*—matching the distribution's intrinsic variance to the policy's gradient sensitivity (evidenced by Gamma's optimal 0.2691 action noise std)—emerges as the critical factor.

- **The Variance-Viability Frontier**: Student's t(3) achieved the highest max reward (3418.98) but with crippling instability (±742.59 std), revealing a Pareto frontier in noise design: practitioners must explicitly trade off *exploration potential* against *training reliability* based on task risk tolerance.

- **Noise as a Regularizer**: Moderately-tailed distributions (Gamma, Gaussian) implicitly regularize policy updates by bounding gradient deviations, mirroring the effects of explicit techniques like gradient clipping. This positions parameter space noise as a dual-purpose tool for *simultaneous exploration and optimization stability.*

These findings provide concrete design principles for noise distribution selection in TD3 and related actor-critic methods, potentially reducing reliance on complex exploration heuristics through principled noise engineering.

References

[1] Fujimoto et al. "Addressing Function Approximation Error in Actor-Critic Methods". ICML 2018.

[2] Fortunato et al. "Noisy Networks for Exploration". ICLR 2018.