Pink Noise LQR: How does Colored Noise affect the Optimal Policy in RL?

Jakob Hollenstein[1] **Marko Zarić**[1] **Samuele Tosatto**[1,2] **Justus Piater**[1,2]

{jakob.hollenstein,marko.zaric,samuele.tosatto,justus.piater}@uibk.ac.at
[1]Department of Computer Science, University of Innsbruck, Austria
[2]Digital Science Center, University of Innsbruck, Austria

Abstract

Colored noise, a class of temporally correlated noise processes, has shown promising results for improving exploration in deep reinforcement learning for both off-policy and on-policy algorithms. However, It is unclear how temporally correlated colored noise affects policy learning apart from changing exploration properties. In this paper, we investigate the influence of colored noise on the optimal policy in a simplified linear quadratic regulator (LQR) setting. We show that the expected trajectory remains independent of the noise color for a given linear policy. We derive a closed-form solution for the expected cost and find that the noise affects both the expected cost and the optimal policy. The cost splits into two parts: a state-cost part equaling the cost for the unperturbed system and a noise-cost term independent of the initial state. Far from the goal state, the state cost dominates, and the effect due to the noise is negligible: the policy approaches the optimal policy of the unperturbed system. Near the goal state, the noise cost dominates, changing the optimal policy.

1 Introduction

Deep reinforcement learning is an approximate dynamic programming technique to derive a policy (a controller) for a given environment, i.e., reward (=−cost) and dynamics. The policy is estimated based on trajectory samples gathered from the environment. To do so, the data collection, and thus the action selection, needs to be varied. This is typically done by randomly perturbing the action selection process, i.e., by action noise. Action noise can be applied additively to the deterministically selected action of a policy or by sampling from a stochastic policy. In continuous control settings, such as robotics, the system dynamics include integrative components: the action signal (e.g., force, torque, velocity), is integrated (velocity, position). This explains why temporally correlated action noise has been found to improve learning performance in reinforcement learning (Rückstieß et al., 2008; Raffin et al., 2021; Eberhard et al., 2023; Hollenstein et al., 2022; Chiappa et al., 2023; Hollenstein et al., 2024). In particular, the temporally correlated *colored noise* processes have shown promising results for continuous control (Eberhard et al., 2023; Pinneri et al., 2020; Hollenstein et al., 2024).

While empirically, these noise processes have shown improvements in learning performance, it is less clear how this noise affects the optimal policy. In this paper, we investigate this question in a simplified setting.

ICML 2024 Workshop on Foundations of Reinforcement Learning and Control.

While reinforcement learning is able to deal with stochastic dynamics, i.e.,

$$s_{t+1} \sim p(\cdot|a_t, s_t)$$

in practice, often environments with deterministic dynamics are used (Tassa et al., 2018; Todorov et al., 2012; Brockman et al., 2016). In this paper, we make the simplifying assumption of linear deterministic dynamics:

$$s_{t+1} = Gs_t + Ha_t$$

and a linear policy:

$$a_t = -Ks_t$$

Furthermore, we assume that the actions of the policy are perturbed by colored noise:

$$a_t = -Ks_t + \varepsilon_t \tag{1}$$

and $\varepsilon_t \sim \mathcal{C}_\beta$. Additionally, we assume the cost ($= -$reward) to be quadratic and the goal state to be $s = \mathbf{0}$. That is, we study the question of the impact of colored noise in the linear quadratic regulator (LQR) setting. As is typical for reinforcement learning, we assume the initial state $s_0 \sim S$ to be sampled from a given initial state distribution. Furthermore, we limit the study to the episodic setting, limiting the length of the trajectories to T.

We are interested in the following questions:

1. Does colored noise change the expected trajectory $\mathbb{E}[s_t]$, given an expected starting state $\bar{s}_0$.
2. Does the expected cost change when the noise color (β) is changed?
3. Does colored noise change the optimal policy?

Our contributions are:

1. We show that the *expected* trajectory remains unchanged regardless of the noise (Q1).
2. We derive a closed-form solution for the expected cost.
3. We show that the cost *is* increased by the noise (Q2).
4. We show that this *can* affect the optimal policy K.

1.1 Related Work

Exploration is critical for reinforcement learning. In continuous control deep reinforcement learning, the two most prominent noise types used for exploration are uncorrelated white noise (Haarnoja et al., 2019; Fujimoto et al., 2018; Abdolmaleki et al., 2018; Schulman et al., 2017) or temporally correlated Ornstein-Uhlenbeck (Uhlenbeck & Ornstein, 1930) noise are used (Lillicrap et al., 2016). The importance of temporally correlated noise has also been shown by methods that combine random aspects with deterministic state-to-action mappings (Raffin et al., 2021; Rückstieß et al., 2008; Chiappa et al., 2023). A further type of random exploration, more similar to white noise and Ornstein-Uhlenbeck noise is action noise exploration based on colored noise processes (Pinneri et al., 2020; Eberhard et al., 2023; Hollenstein et al., 2024). In the setting of LQR, the problem of noisy observations and dynamics is approached in the setting of linear–quadratic–Gaussian (LQG) control, also in the setting of temporally correlated noise (Kwong, 1987). Escobedo-Trujillo & Garrido-Meléndez (2021) also study the setting LQR with dynamics noise of white noise and Ornstein-Uhlenbeck noise. Our setting is different in two ways: first, we focus on additive action noise $a_t = -Ks_t + \varepsilon_t$ which is typically not modeled directly in LQG (but can be modeled indirectly as observation noise, modulated by the control gain). Secondly, the colored noise process we study cannot be expressed as a recursion because it is generated in the frequency domain. We focus on this generating noise process because it is most relevant for deep reinforcement learning (Eberhard et al., 2023; Hollenstein et al., 2024; Pinneri et al., 2020). Recursive filtering approximations of colored noise are possible but have not been compared against the method of generating the noise in the frequency domain.

2 Background

Colored noise processes are a class of temporally correlated noise processes that are parameterized by the noise color β. This class includes temporally uncorrelated noise (white noise, $\beta = 0$) and temporally correlated red noise ($\beta = 2$), which is exhibited by, e.g., Brownian motion. Sequences of different noise colors are illustrated in the time domain in Figure 1. The noise color β describes how the expected power spectral density (PSD) behaves, i.e., the power components scale with $\frac{1}{f^\beta}$. This is illustrated in Figure 2.

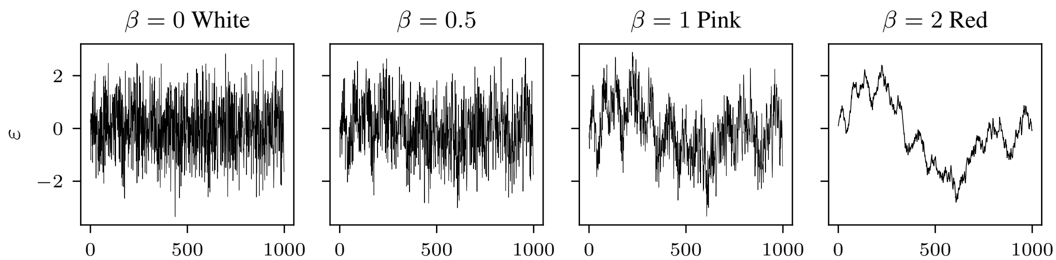

Figure 1: Colored noise processes generate temporally correlated noise with varying degrees of temporal correlation depending on the noise color β. From left to right: the temporal correlation increases with increasing β, from $\beta = 0$ (temporally uncorrelated white noise) to $\beta = 2$ (highly temporally correlated red noise).

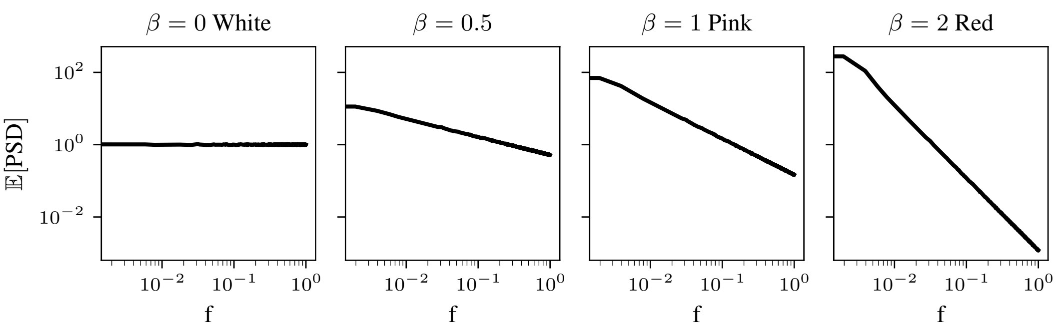

Figure 2: Colored noise is defined by the slope of the expected power spectral density: the power components scale with $\frac{1}{f^\beta}$

Generating colored noise Following the work by Eberhard et al. (2023); Hollenstein et al. (2024) on colored noise in deep reinforcement learning, we generate colored noise in the frequency domain and apply the Inverse Real (Fast) Fourier (IRFFT) to retrieve a sequence of perturbation ε_t, i.e., noise generation follows the algorithm proposed by Timmer & König (1995). This means that for a specific rollout the frequency components Φ are sampled once $\Phi \sim \mathcal{C}_\beta$, and remain fixed for the entire episode. ε_t can then be computed by the inverse Fourier transform, which amounts to a weighted summation of the components of Φ. This summation can be expressed as the inner product between $\Phi^\top$ and the time dependent Fourier coefficients f_t:

$$\varepsilon_t = \Phi^\top \cdot f_t \tag{2}$$

The perturbation ε_t can be interpreted as a discrete-time signal at time index $t \in \{0, \ldots, M - 1\}$, derived using the inverse real Fourier transform from $N = \lfloor M/2 \rfloor + 1$ frequency components. For simplicity in the colored noise generation and inverse Fourier transform, we assume both M and N to be even valued. The derivations in both cases are analogous. For details on the noise generation process see Appendix A.1.

$$\varepsilon_t = \frac{1}{N-1} \sum_{k=0}^{N-1} \left[c_k \varphi_{2k} \cos(-k\frac{t}{M} \cdot 2\pi) + c_k \varphi_{2k+1} \sin(-k\frac{t}{M} \cdot 2\pi) \right] \tag{3}$$

where $\varphi_{2k}, \varphi_{2k+1}$ denote the real, respectively imaginary part of the frequency domain Fourier coefficients and c_k denotes a scaling factor.

$$c_k = \begin{cases} 0 & \text{if } k \in \{0, N-1\} \\ 1 & \end{cases}$$

The sum can be rewritten as the dot product

$$\varepsilon_t = \Phi^\top f_t$$

where Φ and f_t are defined as:

$$\Phi = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_{2N} \end{bmatrix} \qquad f_t = \begin{bmatrix} c_0 \cos(-0 \cdot 2\pi \frac{i}{M}) \\ c_0 \sin(-0 \cdot 2\pi \frac{i}{M} \\ \vdots \\ c_{N-1} \cos(-(N-1) \cdot 2\pi \frac{i}{M} \\ c_{N-1} \sin(-(N-1) \cdot 2\pi \frac{i}{M} \end{bmatrix} \tag{4}$$

The components φ_i of Φ are independently sampled, depending on the noise color β, and the sequence length M:

procedure $\mathcal{C}(M, \beta)$
  $N \leftarrow {}^{M}\!/_{2} + 1$
  $f \leftarrow \{\frac{1}{M}, \frac{1}{M}, \ldots, \frac{i}{M}, \ldots, \frac{N-1}{M}\}$
  $\sigma^2 \leftarrow \{\ldots, f_i^{-\beta}, \ldots\}$ ▷ Calculate scales
  $\frac{1}{c^2} \leftarrow \left(\frac{2}{M}\right)^2 \cdot \sum w_i^2 | w_i \in \{\sigma_1, \ldots, \sigma_{N-2}, \frac{\sigma_{N-1}}{2}\}$
  $\Phi \sim \begin{bmatrix} \mathcal{N}(0, c \cdot \sigma_0 \cdot \sqrt{2}) \\ \mathcal{N}(0, c \cdot \sigma_0 \cdot 0) \\ \mathcal{N}(0, c \cdot \sigma_1) \\ \mathcal{N}(0, c \cdot \sigma_1) \\ \vdots \\ \mathcal{N}(0, c \cdot \sigma_{N-1}) \\ \mathcal{N}(0, c \cdot \sigma_{N-1}) \end{bmatrix}$
  return Φ
end procedure

3 Q1: Expected Trajectory remains unchanged

The addition of action noise changes the distribution of sampled trajectories. The distribution also changes when the noise color is varied. This is illustrated in Figure 3. Empirically this figure also shows, that despite the widely different distribution of trajectories, the expected trajectory remains unchanged. In this section we derive the expected trajectory.

For a given Φ, the action noise is fixed for the whole duration of a trajectory. The policy can be included in the dynamical system to make it autonomous, i.e., the system evolution only depends on the initial state. This means that the trajectory, or more precisely, the

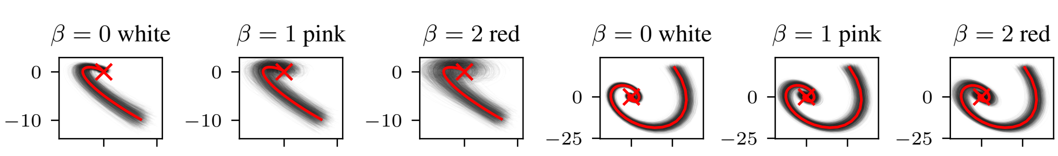

Figure 3: Effect of different noise colors on sampled trajectories. (first three plots) Double Integrator environment, (last three plots) Randomly generated test environment. While the distribution of trajectories (black) changes depending on the noise color, the expected trajectory (red) remains unchanged, reaching the same goal state (marked X) reliably.

state s_t, can be expressed as a sum capturing the recursion starting at the given state s_0 (Appendix B):

$$s_1 = Gs_0 + Ha_0 \tag{5}$$

$$s_1 = Gs_0 - HKs_0 + H\Phi^\top f_0 \tag{6}$$

$$s_2 = Gs_1 + Ha_1 \tag{7}$$

$$\vdots \tag{8}$$

$$s_t = (G - HK)^t s_0 + \sum_{i=1}^{t} (G - HK)^{t-i} H\Phi^\top f_{i-1} \tag{9}$$

Assuming that the initial state is randomly chosen, $s_0 \sim S_0$ and the expected value exists $\mathbb{E}[s_0] = \overline{s_0}$, the expected value for $\mathbb{E}[s_t]$ can be computed:

$$\mathbb{E}[s_t] = \mathbb{E}\left[(G - HK)^t s_0 + \sum_{i=1}^{t} (G - HK)^{t-i} H\Phi^\top f_{i-1}\right]$$

Because $\mathbb{E}[\Phi] = \mathbf{0}$, by definition of the chosen colored noise generation process, using the linearity property of the expectation, this simplifies to

$$\mathbb{E}[s_t] = \mathbb{E}\left[(G - HK)^t s_0\right] = (G - HK)^t \mathbb{E}[s_0]$$

That is: the *expected trajectory* under colored noise remains unchanged.

4 Q2: Effect of Colored Noise on Cost

In the previous section, we demonstrated that the noise color alters the distribution of sampled trajectories, but the expected trajectory stays the same. This raises the question of how the noise color affects the cost, particularly the expected cost.

We assume quadratic costs:

$$J = \sum_{t=0}^{\top} s_t^\top Q s_t + a_t^\top R a_t = J_Q + J_R \tag{10}$$

Because of the presence of action noise, e.g., Equation (1), we are interested in the expected cost $\overline{J} = \mathbb{E}[J]$, which we derive (see Appendix C for details) as follows:

$$\mathbb{E}[J] = \mathbb{E}[J_Q] + \mathbb{E}[J_R] = \mathbb{E}\big[\sum_{t=0}^{\top} s_t^\top Q s_t\big] + \mathbb{E}\big[\sum_{t=0}^{\top} a_t^\top R a_t\big] = \tag{11}$$

$$\sum_{t=0}^{\top} \Bigg(\overline{s}_0^\top S_t \overline{s}_0 + \operatorname{tr}\left(S_t \mathrm{Cov}[s_0]\right) \tag{12}$$

$$+ f_t^\top \mathbb{E}\left[\Phi R \Phi^\top\right] f_t + \tag{13}$$

$$+ \sum_{i=1}^{t}\sum_{j=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi(W_{i,j,t} + B_{i,j,t})\Phi^\top\right] f_{i-1} + \sum_{i=1}^{t} f_{i-1}^\top \mathbb{E}\left[\Phi Y_{t,i} \Phi^\top\right] f_t \Bigg) \tag{14}$$

$$\begin{aligned} &\text{where } C^t = (G - HK)^t && B_{i,j,t} := H^\top {C^{t-j}}^\top K^\top R K C^{t-i} H \\ &W_{i,j,t} = H^\top {C^{t-j}}^\top Q C^{t-i} H && Y_{t,i} := H^\top (C^{t-i})^\top K^\top R \\ &S_t := {C^t}^\top Q C^t + (C^t)^\top K^\top R K C^t && \end{aligned}$$

Note that the cost splits into a state dependent term equal to the cost of the *noise-free system* (12) and three noise terms, one *independent of the policy* K (13), and two *dependent on the policy* K (14). Interestingly, all three of the noise terms are independent of the expected initial state $\overline{s}_0$. These additional terms show analytically that the noise influences the cost. This is illustrated empirically in Figure 4.

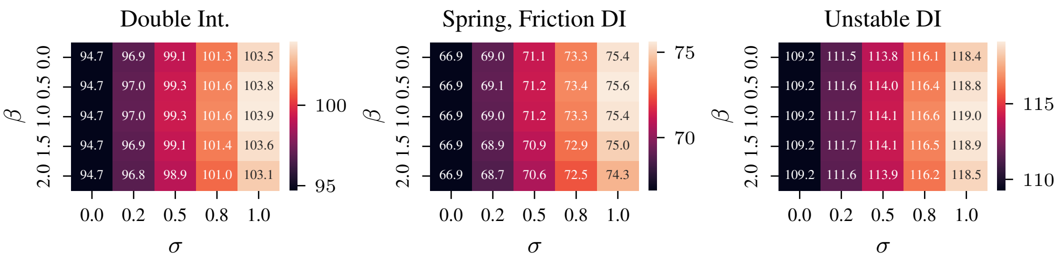

Figure 4: Effect of different noise colors β and noise scales $\sigma \cdot \Phi$ on the expected cost. Generally, the cost increases with a larger noise scale σ. Whether the cost is higher for a specific noise color β depends on the dynamics. The experiments are performed with a horizon $T = 30$.

5 Q3: Optimal Policy K under Colored Noise

In the previous section we showed that the noise color affects the expected cost and that the expected cost splits into a state dependent cost term and a noise dependent cost term.

The split of the expected cost $\overline{J}$ Equation (11), into the state dependent cost

$$J_{s_0} = \sum_{t=0}^{\top} \left(\overline{s}_0^\top S_t \overline{s}_0 + \operatorname{tr}\left(S_t \mathrm{Cov}[s_0]\right) \right)$$

and noise dependent cost (keeping only terms dependent on K)

$$J_\varepsilon = \sum_{t=0}^{\top} \left(\sum_{i=1}^{t} \sum_{j=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi(W_{i,j,t} + B_{i,j,t})\Phi^\top\right] f_{i-1} + \sum_{i=1}^{t} f_{i-1}^\top \mathbb{E}\left[\Phi Y_{t,i} \Phi^\top\right] f_t \right)$$

shows a dependency of the optimal policy, i.e., a change in optimal K in the presence of noise.

The effect of σ and β on K is illustrated empirically in Figure 5. Here, the Double Integrator is studied with initial state $s_0^\top = [0.5 \quad 0.5]$, horizon $T = 32$, the optimal policy K, ($K \in \mathbb{R}^2$) is found numerically from the closed form solution of the expected cost. With larger scale σ, or change in β the components of the gain matrix policy K change.

However, if the noise is scaled down, $\sigma \cdot \Phi$ for $\sigma \ll 1$, J_{s_0} dominates the combined cost and the optimal policy K approaches the optimal policy of the unperturbed system. On the

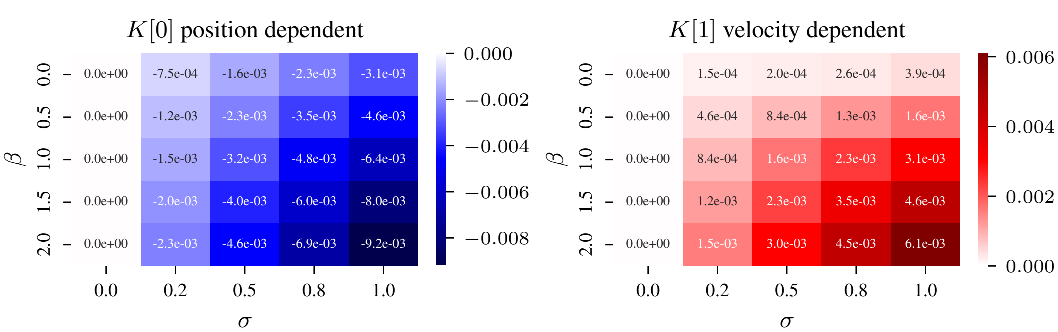

Figure 5: The optimal policy K, is affected by the noise color β and the noise scale σ. We numerically derive the optimal K for the Double Integrator for $\overline{s}_0^\top = \begin{bmatrix} 0 & 0 \end{bmatrix}$ and horizon $T = 30$. The plots show the difference to the optimal K for the action-noise free setting for both dimensions of K (state, velocity) separately. The color gradients show that the growing discrepancy to the noise-free policy is driven by the increase in noise scale and the change in noise color (i.e., temporal correlation of the noise).

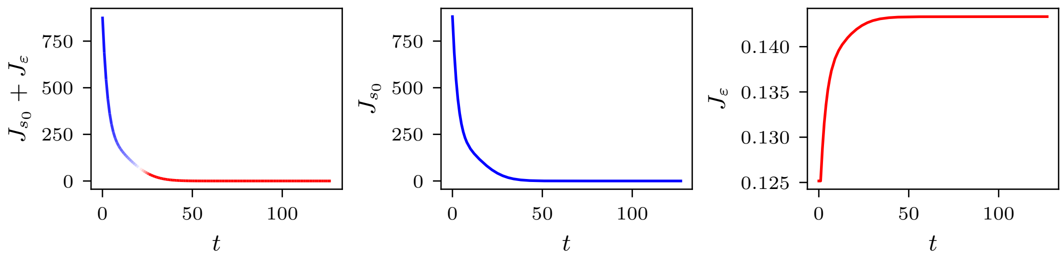

Figure 6: Visualization of the contributions to the cost for each timestep t: The closed-form solution for the expected cost consists of a state dependent term $J_{s_0}(t)$ (■) and a noise dependent term $J_\varepsilon(t)$ (■) At the beginning of the trajectory, the cost is dominated by $J_{s_0}(t)$, close to the goal state, the cost is dominated by the cost incurred by the noise $J_\varepsilon(t)$. The noise cost $J_\varepsilon(t)$ appears to reach an equilibrium. Overall J_{s_0}, which is independent of the action noise, dominates.

flip side, when J_{s_0} has little influence, J_ε dominates the combined cost, causing a shift in the optimal policy K to counteract the noise effect. The influence of J_{s_0} is small when the system is close to the goal state ($s_{\text{goal}} = \mathbf{0}$). This indicates that the cost is likely to be dominated by the state cost J_{s_0} at the beginning of the trajectory, shifting to J_ε towards the end, i.e., close to the goal.

This has several interesting implications:

- J_ε is independent of s_0 and will thus not converge to zero. For an infinite horizon $\lim_{T\to\infty} J$ might diverge and average or discounted cost formulations need to be investigated.
- Hollenstein et al. (2022) suggests reducing the influence of the noise over the course of the training process, i.e., scheduling σ in $\sigma \cdot \Phi$ to shrink over the training process. This would reduce the influence of J_ε and recover the optimal policy of the unperturbed system.
- Close to the goal state $s = \mathbf{0}$, the unperturbed policy would not take any action $a = -Ks = \mathbf{0}$. In this case, J_ε would dominate over J_{s_0}. However, in practical applications, the policy will either have to take actions, suggesting $s \neq \mathbf{0}$, or, if the system is required to stay close to the goal state, the action noise scale needs to be small to prevent the system from deviating too far from $\mathbf{0}$. Both of these factors would lead to $J_{s_0} \gg J_\varepsilon$, suggesting that the optimal policy K approaches the solution of the unperturbed system.

This shift from the state dependent cost J_{s_0} to the noise dependent cost J_ε is illustrated in Figure 6 for the Double Integrator, $s_0 = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$, for K the infinite horizon LQR solution is used, and $T = 120$. In this example, the total cost of the trajectory is determined mostly by the state dependent cost accounting for 99.7% of the total cost. Here, the influence of the noise on the policy would be marginal.

6 Conclusion

In this paper, we investigated the effect of colored action noise on the optimal policy in a simplified LQR setting. We found that the expected trajectory for a given policy remains unchanged in the presence of colored noise but that the expected cost changes. Associated with this change in cost is a change in the optimal policy. The change in cost is due to an additional cost term compared to the cost of the unperturbed system, which is independent of the starting state and instead depends on the noise color, system dynamics, and policy matrix. This effect is relevant close to the goal state, but has little impact further away from the goal. This suggests that while colored noise can change the optimal policy, this change is likely to be small in practice.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. A. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. https://openreview.net/forum?id=S1ANxQW0b.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *arXiv:1606.01540 [cs]*, June 2016. http://arxiv.org/abs/1606.01540.

Chiappa, A. S., Vargas, A. M., Huang, A. Z., and Mathis, A. Latent Exploration for Reinforcement Learning. [object Object], 2023. doi: 10.48550/ARXIV.2305.20065. https://arxiv.org/abs/2305.20065.

Eberhard, O., Hollenstein, J., Pinneri, C., and Martius, G. Pink Noise Is All You Need: Colored Noise Exploration in Deep Reinforcement Learning. In *International Conference on Learning Representations*, February 2023. https://openreview.net/forum?id=hQ9V5QN27eS.

Escobedo-Trujillo, B. and Garrido-Meléndez, J. Stochastic LQR optimal control with white and colored noise: Dynamic programming technique. *Revista Mexicana de Ingeniería Química*, 20(2):1111–1127, 2021. http://rmiq.org/iqfvp/Numbers/V20/No2/Sim2353.pdf.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*, October 2018. http://arxiv.org/abs/1802.09477.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S. Soft Actor-Critic Algorithms and Applications. *arXiv:1812.05905 [cs, stat]*, January 2019. http://arxiv.org/abs/1812.05905.

Hollenstein, J., Auddy, S., Saveriano, M., Renaudo, E., and Piater, J. Action Noise in Off-Policy Deep Reinforcement Learning: Impact on Exploration and Performance. *Transactions on Machine Learning Research*, November 2022. ISSN 2835-8856. https://openreview.net/forum?id=NljBlZ6hmG&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DTMLR%2FAuthors%23your-submissions).

Hollenstein, J., Martius, G., and Piater, J. Colored noise in PPO: Improved exploration and performance through correlated action sampling. In *Conference of the Association for the Advancement of Artificial Intelligence*, February 2024.

Kwong, R. On the LQG problem with correlated noise and its relation to minimum variance control. In *26th IEEE Conference on Decision and Control*, pp. 763–767, Los Angeles, California, USA, December 1987. IEEE. doi: 10.1109/CDC.1987.272492. http://ieeexplore.ieee.org/document/4049369/.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. http://arxiv.org/abs/1509.02971.

Pinneri, C., Sawant, S., Blaes, S., Achterhold, J., Stueckler, J., Rolínek, M., and Martius, G. Sample-efficient cross-entropy method for real-time planning. In Kober, J., Ramos, F., and Tomlin, C. J. (eds.), *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1049–1065. PMLR, 2020. https://proceedings.mlr.press/v155/pinneri21a.html.

Raffin, A., Kober, J., and Stulp, F. Smooth exploration for robotic reinforcement learning. In Faust, A., Hsu, D., and Neumann, G. (eds.), *Conference on Robot Learning*, 2021. https://proceedings.mlr.press/v164/raffin22a.html.

Rückstieß, T., Felder, M., and Schmidhuber, J. State-Dependent Exploration for Policy Gradient Methods. In Daelemans, W., Goethals, B., and Morik, K. (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 5212, pp. 234–249. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-87480-5 978-3-540-87481-2. doi: 10.1007/978-3-540-87481-2_16. `http://link.springer.com/10.1007/978-3-540-87481-2_16`.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347, 2017.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. DeepMind Control Suite. *arXiv:1801.00690 [cs]*, January 2018. `http://arxiv.org/abs/1801.00690`.

Timmer, J. and König, M. On generating power law noise. *Astron. Astrophys*, 300:707–710, 1995.

Todorov, E., Erez, T., and Tassa, Y. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, October 2012. doi: 10.1109/IROS.2012.6386109.

Uhlenbeck, G. E. and Ornstein, L. S. On the theory of the Brownian motion. *Physical review*, 36(5):823, 1930.

A Colored Noise

A.1 Generating Colored Noise

procedure GenerateColoredNoise(N, β)
- $L \leftarrow \lfloor N/2 \rfloor$
- $f \leftarrow \{\frac{1}{N}, \frac{1}{N}, \ldots, \frac{i}{N}, \ldots, \frac{L}{N}\}$ ▷ Frequencies of components $0 \ldots L$
- $s \leftarrow \{\ldots, f_i^{-\beta/2}, \ldots\}$ ▷ Calculate scales
- $w_L \leftarrow \begin{cases} s_L, & \text{if } L \text{ is odd} \\ s_L/2, & \text{otherwise} \end{cases}$
- $w \leftarrow \{s_1, \ldots s_{L-1}, w_L\}$
- $\sigma \leftarrow \frac{2}{N} \cdot \sqrt{\sum w_i^2}$
- $\alpha = \{\ldots, \alpha_i, \ldots\} : \alpha_i \sim \mathcal{N}(0, s_i)$ ▷ Real part
- $\beta = \{\ldots, \beta_i, \ldots\} : \beta_i \sim \mathcal{N}(0, s_i)$ ▷ Imaginary part
- $\alpha_0 \sim \mathcal{N}(0, s_0 \cdot \sqrt{2})$
- $\beta_0 \leftarrow 0$
- $\alpha_L \sim \begin{cases} \mathcal{N}(0, s_0 \cdot \sqrt{2}), & \text{if } odd \\ \mathcal{N}(0, s_0), & \text{otherwise} \end{cases}$
- $\beta_L \sim \mathcal{N}(0, s_0) \cdot \begin{cases} 0, & \text{if } odd \\ 1, & \text{otherwise} \end{cases}$
- $\gamma \leftarrow \{\ldots, \gamma_i, \ldots\} : \gamma_i = \alpha_i + \boldsymbol{i}\beta_i$
- $\tau_\varepsilon = \mathcal{F}^{-1}[\gamma] \cdot 1/\sigma$
- **return** τ_ε ▷ Return noise sequence of length N

end procedure

B Derivation of s_t

$$a_t = -Ks_t + \Phi^\top f_t$$
$$s_1 = Gs_0 + Ha_0$$
$$s_1 = Gs_0 - HKs_0 + H\Phi^\top f_0$$
$$s_1 = (G - HK)s_0 + H\Phi^\top f_0$$
$$s_2 = Gs_1 + Ha_1$$
$$s_2 = Gs_1 - HKs_1 + H\Phi^\top f_1$$
$$s_2 = (G - HK)s_1 + H\Phi^\top f_1$$
$$s_2 = (G - HK)(G - HK)s_0 + H\Phi^\top f_0) + H\Phi^\top f_1$$
$$s_2 = (G - HK)^2 s_0 + (G - HK)(H\Phi^\top f_0) + H\Phi^\top f_1$$
$$s_3 = Gs_2 + Ha_2$$
$$s_3 = (G - HK)s_2 + H\Phi^\top f_2$$
$$s_3 = (G - HK)((G - HK)^2 s_0 + (G - HK)(H\Phi^\top f_0) + H\Phi^\top f_1) + H\Phi^\top f_2$$
$$s_3 = (G - HK)^3 s_0 + (G - HK)^2(H\Phi^\top f_0) + (G - HK)H\Phi^\top f_1 + H\Phi^\top f_2$$
$$\vdots$$
$$s_t = (G - HK)^t s_0 + \sum_{i=1}^{t} (G - HK)^{t-i}\Phi^\top f_{i-1}$$

C Derivation of closed-form expected cost $\mathbb{E}[J]$

From Equations (1) and (9) and let $C = (G - HK)$. For a given trajectory length T, the noise sample ε_t is generated by the Fourier series at time t. This amounts to a weighted sum of the frequency component random variables: $\varepsilon_t = \Phi^\top f_t$.

$$\mathbb{E}\left[J\right] = \mathbb{E}\left[J_Q\right] + \mathbb{E}\left[J_R\right] = \mathbb{E}\big[\sum_{t=0}^{T} s_t^\top Q s_t\big] + \mathbb{E}\big[\sum_{t=0}^{T} a_t^\top R a_t\big] =$$

$$\sum_{t=0}^{T} \mathbb{E}\left[s_t^\top Q s_t\right] + \sum_{t=0}^{T} \mathbb{E}\left[a_t^\top R a_t\right] = \sum_{t=0}^{T} \mathbb{E}\left[s_t^\top Q s_t\right] + \mathbb{E}\left[a_t^\top R a_t\right]$$

We derive $\mathbb{E}\left[s_t^\top Q s_t\right]$ and $\mathbb{E}\left[a_t^\top R a_t\right]$ in separate subsections and combine the results afterwards into the final closed-form expected cost solution.

C.1 Derivation of $\mathbb{E}\left[s_t^\top Q s_t\right]$

$$\begin{aligned}
\mathbb{E}\left[s_t^\top Q s_t\right] =& \mathbb{E}\left[\left(C^t s_0 + \sum_{i=1}^{t} C^{t-i} H \Phi^\top f_{i-1}\right)^\top Q \left(C^t s_0 + \sum_{i=1}^{t} C^{t-i} H \Phi^\top f_{i-1}\right)\right] = \\
=& \mathbb{E}\left[s_0^\top {C^t}^\top Q C^t s_0\right] + \mathbb{E}\left[2 s_0^\top {C^t}^\top Q \sum_{i=1}^{t} C^{t-i} H \Phi^\top f_{i-1}\right] + \\
&+ \mathbb{E}\left[\sum_{j=1}^{t} f_{j-1}^\top \Phi H^\top {C^{t-j}}^\top Q \sum_{i=1}^{t} C^{t-i} H \Phi^\top f_{i-1}\right] = \\
=& \mathbb{E}\left[s_0^\top\right] {C^t}^\top Q C^t \mathbb{E}\left[s_0\right] + \mathrm{tr}({C^t}^\top Q C^t \mathrm{Cov}[s_0]) + 2\mathbb{E}\left[s_0^\top\right] {C^t}^\top Q \sum_{i=1}^{t} C^{t-i} \mathbb{E}\left[\Phi^\top\right] f_{i-1} + \\
&+ \sum_{j=1}^{t} \sum_{i=1}^{t} f_{j-1}^\top H^\top \mathbb{E}\left[\Phi {C^{t-j}}^\top Q C^{t-i} H \Phi^\top\right] f_{i-1} = \\
=& \overline{s}_0^\top {C^t}^\top Q C^t \overline{s}_0 + \mathrm{tr}({C^t}^\top Q C^t \mathrm{Cov}[s_0]) + \sum_{j=1}^{t} \sum_{i=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi H^\top {C^{t-j}}^\top Q C^{t-i} H \Phi^\top\right] f_{i-1} = \\
=& \overline{s}_0^\top {C^t}^\top Q C^t \overline{s}_0 + \mathrm{tr}({C^t}^\top Q C^t \mathrm{Cov}[s_0]) + \sum_{j=1}^{t} \sum_{i=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi W_{i,j,t} \Phi^\top\right] f_{i-1}
\end{aligned}$$

C.2 Derivation of $\mathbb{E}\left[a_t^\top R a_t\right]$

Calculating the expected action cost for time step a_t results in three separate action cost terms:

$$\begin{aligned}
\mathbb{E}\left[a_t^\top R a_t\right] =& \mathbb{E}\left[(-K s_t + \Phi^\top f_t)^\top R (-K s_t + \Phi^\top f_t)\right] = \\
=& \mathbb{E}\left[f_t^\top \Phi R \Phi^\top f_t\right] + \mathbb{E}\left[s_t^\top K^\top R K s_t\right] - 2\mathbb{E}\left[s_t^\top K^\top R \Phi^\top f_t\right]
\end{aligned}$$

First action cost term:

$$\mathbb{E}\left[f_t^\top \Phi R \Phi^\top f_t\right] = f_t^\top \mathbb{E}\left[\Phi R \Phi^\top\right] f_t$$

Second action cost term:

$$\begin{aligned}
\mathbb{E}\left[s_t^\top K^\top R K s_t\right] = \mathbb{E}\left[s_0^\top (C^t)^\top K^\top R K C^t s_0\right] + \operatorname{tr}\left((C^t)^\top K^\top R K C^t \operatorname{Cov}[s_0]\right) + \\
+\mathbb{E}\left[2 s_0^\top {C^t}^\top K^\top R \sum_{i=1}^{t} K C^{t-i} H \Phi^\top f_{i-1}\right] + \\
+\mathbb{E}\left[\sum_{i=1}^{t}\sum_{j=1}^{t} \left(C^{j-i} H \Phi^\top f_{j-1}\right)^\top K^\top R K C^{t-i} H \Phi^\top f_{i-1}\right] = \\
= \mathbb{E}\left[s_0^\top\right] (C^t)^\top K^\top R K C^t \mathbb{E}\left[s_0\right] + \operatorname{tr}\left((C^t)^\top K^\top R K C^t \operatorname{Cov}[s_0]\right) + \\
+\mathbb{E}\left[\sum_{i=1}^{t}\sum_{j=1}^{t} f_{j-1}^\top \Phi H^\top {C^{j-i}}^\top K^\top R K C^{t-i} H \Phi^\top f_{i-1}\right] = \\
= \overline{s}_0^\top (C^t)^\top K^\top R K C^t \overline{s}_0 + \operatorname{tr}\left((C^t)^\top K^\top R K C^t \operatorname{Cov}[s_0]\right) + \\
+\sum_{i=1}^{t}\sum_{j=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi H^\top {C^{j-i}}^\top K^\top R K C^{t-i} H \Phi^\top\right] f_{i-1} = \\
[\text{substitute} : B_{i,j,t} := H^\top {C^{j-i}}^\top K^\top R K C^{t-i} H] \\
= \overline{s}_0^\top (C^t)^\top K^\top R K C^t \overline{s}_0 + \operatorname{tr}\left((C^t)^\top K^\top R K C^t \operatorname{Cov}[s_0]\right) + \sum_{i=1}^{t}\sum_{j=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi B_{i,j,t} \Phi^\top\right] f_{i-1}
\end{aligned}$$

Third action cost term:

$$\begin{aligned}
\mathbb{E}\left[s_t^\top K^\top R \Phi^\top f_t\right] = \mathbb{E}\left[(C^t s_0 + \sum_{i=1}^{t} C^{t-i} H \Phi^\top)^\top K^\top R \Phi^\top f_t\right] = \\
= \mathbb{E}\left[s_0^\top (C^t)^\top K^\top R \Phi^\top f_t\right] + \mathbb{E}\left[\sum_{i=1}^{t} f_{i-1}^\top \Phi H^\top (C^{t-i})^\top K^\top R \Phi^\top f_t\right] = \\
\left[\mathbb{E}\left[s_0^\top (C^t)^\top K^\top R \Phi^\top f_t\right] = 0 \quad \& \quad \text{substitute } Y_{t,i} := H^\top (C^{t-i})^\top K^\top R\right] \\
= \sum_{i=1}^{t} f_{i-1}^\top \mathbb{E}\left[\Phi Y_{t,i} \Phi^\top\right] f_t
\end{aligned}$$

C.3 Closed-form expected cost $\mathbb{E}[J]$

When combining the derived parts, we group them based on

(15) **Initial state dependency**: We use the linear property of the quadratic form and trace operator to merge the state cost and action cost parts with initial state dependency.

(16) **Noise dependency**: The action cost term that only depends on the noise.

(17) **Noise + Policy dependency**: State and action cost have a quadratic double sum term, which we combine to one (linearity of expectation). This term and the linear action cost term are both noise and policy dependent but have no dependents on the initial state.

$$\mathbb{E}\left[J\right] = \sum_{t=0}^{T} \mathbb{E}\left[s_t^\top Q s_t\right] + \sum_{t=0}^{T} \mathbb{E}\left[a_t^\top R a_t\right] = \sum_{t=0}^{T} \Bigg($$

$$\overline{s}_0^\top \left({C^t}^\top Q C^t + (C^t)^\top K^\top R K C^t\right) \overline{s}_0 + \mathrm{tr}\left(\left({C^t}^\top Q C^t + (C^t)^\top K^\top R K C^t\right) \mathrm{Cov}[s_0]\right) \tag{15}$$

$$+ f_t^\top \mathbb{E}\left[\Phi R \Phi^\top\right] f_t + \tag{16}$$

$$+ \sum_{i=1}^{t} \sum_{j=1}^{t} f_{j-1}^\top \mathbb{E}\left[\Phi (W_{i,j,t} + B_{i,j,t}) \Phi^\top\right] f_{i-1} + \sum_{i=1}^{t} f_{i-1}^\top \mathbb{E}\left[\Phi Y_{t,i} \Phi^\top\right] f_t \Bigg) \tag{17}$$

$$\text{where } C^t = (G - HK)^t \qquad B_{i,j,t} := H^\top {C^{t-j}}^\top K^\top R K C^{t-i} H$$

$$W_{i,j,t} = H^\top {C^{t-j}}^\top Q C^{t-i} H \qquad Y_{t,i} := H^\top (C^{t-i})^\top K^\top R$$

To make this a closed-form solution we have to evaluate all expectations in the formula which all can be evaluated in the following way analogously:

$$\begin{aligned}
&\mathbb{E}\left[\Phi Z \Phi^\top\right] = \\
&\mathbb{E}\left[\begin{bmatrix} \varphi_{1,1} & \cdots & \varphi_{M,1} \\ & \vdots & \\ \varphi_{1,N} & \cdots & \varphi_{M,N} \end{bmatrix} Z \begin{bmatrix} \varphi_{1,1} & \cdots & \varphi_{M,1} \\ & \vdots & \\ \varphi_{1,N} & \cdots & \varphi_{M,N} \end{bmatrix}\right] = \\
&\mathbb{E}\left[\begin{bmatrix} \sum_{k,l} Z_{k,l} \varphi_{k,0} \varphi_{l,0} & \cdots & \sum_{k,l} Z_{k,l} \varphi_{k,0} \varphi_{l,N} \\ & \vdots & \\ \sum_{k,l} Z_{k,l} \varphi_{k,N} \varphi_{l,0} & \cdots & \sum_{k,l} Z_{k,l} \varphi_{k,N} \varphi_{l,N} \end{bmatrix}\right] = \\
&\begin{bmatrix} \sum_{k,l} Z_{k,l} \mathbb{E}\left[\varphi_{k,0} \varphi_{l,0}\right] & \cdots & \sum_{k,l} Z_{k,l} \mathbb{E}\left[\varphi_{k,0} \varphi_{l,N}\right] \\ & \vdots & \\ \sum_{k,l} Z_{k,l} \mathbb{E}\left[\varphi_{k,N} \varphi_{l,0}\right] & \cdots & \sum_{k,l} Z_{k,l} \mathbb{E}\left[\varphi_{k,N} \varphi_{l,N}\right] \end{bmatrix} = \\
&\left[\mathbb{E}\left[\varphi_{i,k} \varphi_{j,k}\right] = 0 \quad \text{for} \quad i \neq j\,\right] \\
&\begin{bmatrix} \sum_{k} Z_{k,k} \mathbb{E}\left[\varphi_{k,0} \varphi_{k,0}\right] & \cdots & \sum_{k} Z_{k,k} \mathbb{E}\left[\varphi_{k,0} \varphi_{k,N}\right] \\ & \vdots & \\ \sum_{k} Z_{k,k} \mathbb{E}\left[\varphi_{k,N} \varphi_{k,0}\right] & \cdots & \sum_{k} Z_{k,k} \mathbb{E}\left[\varphi_{k,N} \varphi_{k,N}\right] \end{bmatrix} = \\
&\left[\mathbb{E}\left[\varphi_{k,i} \varphi_{k,j}\right] = 0 \quad \text{for} \quad i \neq j\,\right] \\
&\begin{bmatrix} \sum_{k} Z_{k,k} \mathrm{Var}[\varphi_{k,0}] & & 0 \\ & \ddots & \\ 0 & & \sum_{k} Z_{k,k} \mathrm{Var}[\varphi_{k,N}] \end{bmatrix} \\
&= \mathrm{diag}_{V \to M}\left(\mathrm{diag}_{M \to V}(Z) \cdot \mathrm{Var}[\Phi]\right)
\end{aligned} \tag{18}$$

By definition of the colored noise generation process Var[Φ] is known which results in the following closed form solution:

$$
\begin{gathered}
\mathbb{E}\left[J\right] = \sum_{t=0}^{T} \mathbb{E}\left[s_t^\top Q s_t\right] + \sum_{t=0}^{T} \mathbb{E}\left[a_t^\top R a_t\right] = \\
\sum_{t=0}^{T} \Bigg(\overline{s}_0^\top S_t \overline{s}_0 + \operatorname{tr}\left(S_t \mathrm{Cov}[s_0]\right) + \\
+ f_t^\top \operatorname{diag}_{V\to M}\left(\operatorname{diag}_{M\to V}(R) \cdot \mathrm{Var}[\Phi]\right) f_t + \\
+ \sum_{i=1}^{t} \sum_{j=1}^{t} f_{j-1}^\top \operatorname{diag}_{V\to M}\left(\operatorname{diag}_{M\to V}(W_{i,j,t} + B_{i,j,t}) \cdot \mathrm{Var}[\Phi]\right) f_{i-1} + \\
+ \sum_{i=1}^{t} f_{i-1}^\top \operatorname{diag}_{V\to M}\left(\operatorname{diag}_{M\to V}(Y_{t,i}) \cdot \mathrm{Var}[\Phi]\right) f_t \Bigg)
\end{gathered}
$$

$$
\begin{aligned}
\text{where } C^t &:= (G - HK)^t & B_{i,j,t} &:= H^\top {C^{t-j}}^\top K^\top R K C^{t-i} H \\
W i, j, t &:= H^\top {C^{t-j}}^\top Q C^{t-i} H & Y_{t,i} &:= H^\top (C^{t-i})^\top K^\top R \\
S_t &:= {C^t}^\top Q C^t + (C^t)^\top K^\top R K C^t
\end{aligned}
$$