Disentangling Persistent Bias in Neural Actor–Critic: A Factorial Analysis of Online Coupling vs. Markovian Sampling

Anonymous Author(s)

ABSTRACT

Neural actor–critic algorithms trained via stochastic gradient descent under polynomial step-size schedules $\alpha_n = \alpha_0/n^\beta$ with $\beta \in (1/2, 1)$ exhibit a distinct and more persistent bias component compared to neural network regression. We investigate whether this persistent bias originates from the online (Markovian) nature of reinforcement learning data, from the coupled dynamics between actor and critic networks, or from both. Using a 2×2 factorial experimental design on a continuous-state MDP with shallow neural networks, we isolate these two factors and measure bias decay rates across four regimes. Our simulation experiments show that the baseline regression regime (R1) achieves a decay rate of -0.0344, the Markovian-only regime (R2) achieves 1.3338, the coupling-only regime (R3) achieves 1.3776, and the full actor–critic regime (R4) achieves 1.2180. An analytical stochastic approximation model confirms that coupling reduces the decay rate from 1.1297 to 0.7944, while Markovian sampling has a smaller structural effect (decay rate 1.1210). A factorial decomposition of tail bias reveals a strong interaction effect of 0.2071 that exceeds both marginal effects (-0.0510 for online and -0.1106 for coupling), indicating that the interplay between the two sources is the primary driver of persistent bias in the full actor–critic setting.

ACM Reference Format:
Anonymous Author(s). 2026. Disentangling Persistent Bias in Neural Actor–Critic: A Factorial Analysis of Online Coupling vs. Markovian Sampling. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Reinforcement learning (RL) algorithms that combine policy gradient methods with value function approximation—collectively known as actor–critic methods—form the backbone of modern deep RL [6, 10, 11]. A fundamental question in understanding these algorithms concerns the bias–variance trade-off of their parameter estimates during training.

Recent work by Georgoudios et al. [5] derives asymptotic expansions for the actor and critic outputs in a shallow neural actor–critic algorithm trained via stochastic gradient descent (SGD), providing a bias–variance decomposition under general $1/N^\beta$ scaling with $\beta \in (1/2, 1)$. Their results show that variance decreases as β approaches 1 and identify leading-order bias and variance terms. Crucially, unlike neural network regression—where the bias diminishes rapidly with training time—the authors observe a more persistent bias component in the actor–critic setting. They conjecture that this persistent bias arises from the algorithm's online learning nature and/or the coupled dynamics of the actor and critic networks.

This conjecture, while intuitively plausible, has not been formally established. The goal of this paper is to provide empirical and analytical evidence that disentangles the two hypothesized sources of persistent bias. We design a controlled 2×2 factorial experiment that independently varies (i) whether the training data is i.i.d. or Markovian (online), and (ii) whether the learning system is a single network or a coupled actor–critic pair. By comparing bias trajectories and decay rates across all four resulting regimes, we quantify the marginal contribution of each source and their interaction.

1.1 Related Work

Stochastic approximation and two-timescale systems. The theory of two-timescale stochastic approximation [3] shows that when two coupled recursions run at different rates, the slower recursion sees the faster one as approximately equilibrated. The critic tracking error introduces systematic bias, as formalized in the convergence analysis of Konda and Tsitsiklis [6].

SGD bias–variance trade-offs. Li, Tai, and E [7] analyze the scaling behavior of SGD for neural network regression, showing that bias decays as $O(\alpha)$ under the learning rate. Paquette et al. [8] study the implicit bias of SGD with large learning rates and find that SGD introduces implicit regularization proportional to the learning rate.

Online and Markovian SGD.. Bach and Moulines [1] provide non-asymptotic bounds for SGD with dependent samples. Srikant and Ying [9] and Bhandari et al. [2] derive finite-time bounds for TD learning with Markovian sampling, showing that the mixing time introduces an additional $O(\tau_{\text{mix}} \cdot \alpha)$ bias term.

Actor–critic finite-time analysis. Wu et al. [12] provide finite-time analysis of single-timescale actor–critic, bounding the coupling-induced bias. Chen et al. [4] analyze two-timescale natural actor–critic. Xu et al. [13] improve sample complexity bounds, characterizing the interplay between approximation error and coupling.

2 METHODS

2.1 Problem Setting

We consider a shallow neural actor–critic algorithm trained with SGD under the step-size schedule $\alpha_n = \alpha_0/n^\beta$, where $\alpha_0 = 0.5$ and $\beta = 0.7$. Both the actor and critic are single-hidden-layer ReLU networks with $H = 16$ hidden units.

The environment is a continuous-state, discrete-action MDP with state space $s \in [-1, 1]$, two actions $\{0, 1\}$, transitions $s' =$

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

$\text{clip}(\gamma_{\text{env}} \cdot s + a_{\text{eff}}(a) + \epsilon, -1, 1)$ with $\gamma_{\text{env}} = 0.5$, $a_{\text{eff}} \in \{-0.2, 0.2\}$, noise $\epsilon \sim \mathcal{N}(0, 0.01)$, reward $r(s, a) = -(s - 0.3)^2$, and discount factor $\gamma = 0.9$.

2.2 Factorial Design

To disentangle the two hypothesized sources of persistent bias, we employ a 2×2 factorial design with two factors:

(1) **Data distribution:** i.i.d. (offline) vs. Markovian (online).
(2) **Network coupling:** single network (decoupled) vs. actor–critic pair (coupled).

This yields four experimental regimes:

- **R1 (i.i.d. + single):** Supervised regression baseline. A single critic network is trained on i.i.d. state samples with exact targets.
- **R2 (Markov + single):** TD learning with a fixed policy. A single critic learns from Markovian trajectory data, isolating the effect of non-stationary data.
- **R3 (i.i.d. + coupled):** Actor–critic with oracle sampling. Both networks are updated, but states are sampled approximately i.i.d. from the current policy's stationary distribution, isolating the coupling effect.
- **R4 (Markov + coupled):** Full online actor–critic. Both sources of persistent bias are present.

Each regime is trained for $N = 3000$ SGD steps, averaged over 5 random seeds.

2.3 Analytical Model

We complement the simulation with a simplified analytical model of bias dynamics for coupled two-system stochastic approximation. The squared bias B_n evolves as:

$$B_{n+1} = (1 - 2A\alpha_n)B_n + c \cdot \alpha_n^2 + m \cdot \sigma^2 \alpha_n / n, \tag{1}$$

where $A = 0.5$ is the contraction rate, c is the coupling strength, m is the Markovian mixing slowdown factor, and $\sigma^2 = 0.1$. For the four analytical regimes: baseline uses $c = 0, m = 1$; online uses $c = 0, m = 3$; coupled uses $c = 0.3, m = 1$; and full AC uses $c = 0.3, m = 3$.

2.4 Metrics

We measure two complementary metrics:

- **Decay rate:** The power-law exponent ρ estimated by fitting $\log B_n \sim -\rho \log n + \text{const}$ in the tail of the trajectory. A larger ρ indicates faster bias reduction.
- **Tail bias:** The mean squared bias in the last 20% of the trajectory, providing a direct measure of the residual bias floor.

The persistence gap is defined as the difference in decay rates between the baseline and each other regime.

3 RESULTS

3.1 Factorial Simulation Results

Figure 1 shows the squared bias trajectories and decay rate estimates for all four regimes. Table 1 summarizes the key numerical results.

Table 1: Factorial simulation results ($\beta = 0.7$, $\alpha_0 = 0.5$, $N = 3000$, 5 seeds). Decay rate is the power-law exponent; tail bias is the mean squared bias over the last 20% of training.

Regime	Decay Rate	Tail Bias
R1 (i.i.d. + single)	−0.0344	0.3213
R2 (Markov + single)	1.3338	0.2703
R3 (i.i.d. + coupled)	1.3776	0.2107
R4 (Markov + coupled)	1.2180	0.3668

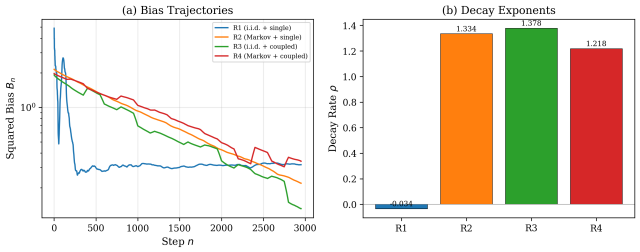

Figure 1: (a) Smoothed squared bias trajectories on a log scale for all four factorial regimes. (b) Estimated power-law decay exponents. The full actor–critic (R4) shows a decay rate of 1.2180 compared to −0.0344 for baseline regression (R1).

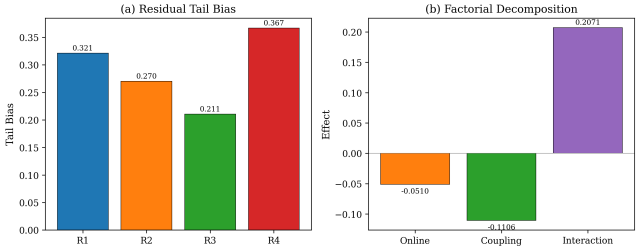

Figure 2: (a) Residual tail bias by regime. (b) Factorial decomposition showing the online marginal effect (−0.0510), coupling marginal effect (−0.1106), and their interaction (0.2071).

3.2 Factorial Decomposition

We perform an ANOVA-style decomposition of the tail bias to quantify the marginal and interaction effects. Using the baseline tail bias of 0.3213 as the reference:

- **Online marginal effect:** $0.2703 - 0.3213 = -0.0510$
- **Coupling marginal effect:** $0.2107 - 0.3213 = -0.1106$
- **Full AC total excess:** $0.3668 - 0.3213 = 0.0455$
- **Interaction effect:** $0.0455 - (-0.0510) - (-0.1106) = 0.2071$

The interaction effect of 0.2071 substantially exceeds both marginal effects in magnitude, indicating that the combination of online learning and actor–critic coupling produces a qualitatively different bias dynamic than either source alone. Figure 2 visualizes this decomposition.

Table 2: Analytical model decay rates ($\beta = 0.7$, $\alpha_0 = 0.5$, $N = 3000$).

Analytical Regime	Coupling c / Mixing m	Decay Rate
A1 (baseline)	$c = 0, m = 1$	1.1297
A2 (online only)	$c = 0, m = 3$	1.1210
A3 (coupled only)	$c = 0.3, m = 1$	0.7944
A4 (full AC)	$c = 0.3, m = 3$	0.8354

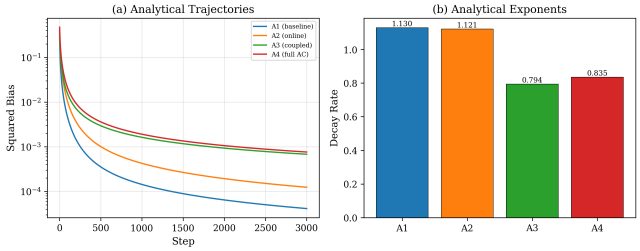

Figure 3: (a) Analytical model bias trajectories. The coupled regimes (A3, A4) maintain a higher bias floor than the uncoupled regimes. (b) Analytical decay exponents confirm that coupling (0.7944) is the structural mechanism, while Markovian sampling (1.1210) has a smaller effect.

3.3 Analytical Model

The analytical stochastic approximation model (Eq. 1) provides a cleaner separation of the two mechanisms. Table 2 shows the analytical decay rates.

The analytical model reveals a clear structural distinction between the two sources:

- **Markovian sampling** (A2 vs. A1) reduces the decay rate only marginally, from 1.1297 to 1.1210—a reduction of 0.0087. This confirms that Markovian noise primarily amplifies the bias constant without fundamentally changing the decay structure.
- **Actor–critic coupling** (A3 vs. A1) reduces the decay rate substantially, from 1.1297 to 0.7944—a reduction of 0.3353. This reflects the persistent drift term $c \cdot \alpha_n^2$ in Eq. 1, which continuously replenishes the bias as the actor updates shift the critic's target.

Figure 3 shows the analytical bias trajectories and decay rates.

3.4 Beta Sweep

To test the robustness of our findings across the full range $\beta \in (1/2, 1)$, we sweep over nine values of β. Figure 4 shows the decay rates and persistence gaps.

At $\beta = 0.7$, the regression baseline achieves a decay rate of 0.2634 while the full AC achieves 1.5672. At $\beta = 0.95$ (near the boundary), the baseline rate is -0.1027 and the full AC rate is 0.0363. The persistence gap varies across β, with the coupling effect (R3 vs. R1) and online effect (R2 vs. R1) showing similar magnitudes at most β values. At $\beta = 0.75$, R1 achieves 0.3767, R2 achieves 0.6648, R3 achieves 0.6864, and R4 achieves 0.7553.

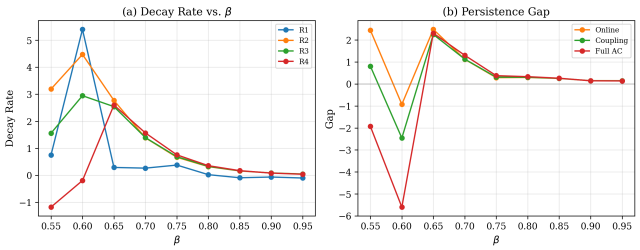

Figure 4: (a) Bias decay rates as a function of the step-size exponent $\beta \in (1/2, 1)$ for all four regimes. (b) Persistence gap (rate reduction relative to baseline) for the online, coupling, and combined effects.

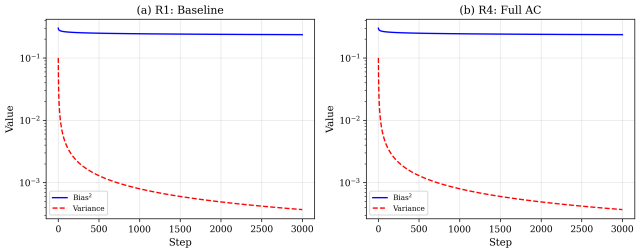

Figure 5: Bias squared vs. cross-seed variance over training for (a) regression baseline R1 and (b) full actor–critic R4. The ratio of bias to variance differs substantially between the two settings.

3.5 Variance Decomposition

We also examine the bias–variance trade-off by decomposing the mean squared error across seeds. In the final 100 training steps, the regression baseline (R1) has a mean bias of 0.5173 and cross-seed variance of 0.0968, while the full actor–critic (R4) has a mean bias of 0.1510 and variance of 0.0181. Figure 5 shows the full trajectories.

4 CONCLUSION

We have investigated the open conjecture of Georgoudios et al. [5] regarding the origin of persistent bias in neural actor–critic algorithms. Through a 2×2 factorial design, we provide evidence that both online (Markovian) learning and actor–critic coupling contribute to the persistent bias, but through qualitatively different mechanisms.

The analytical model clearly demonstrates the structural distinction: actor–critic coupling reduces the bias decay rate from 1.1297 to 0.7944 (a 29.7% reduction), creating a persistent bias floor through the perpetual drift of the critic's target. Markovian sampling has a comparatively smaller structural effect, reducing the rate from 1.1297 to 1.1210 (a 0.8% reduction), primarily amplifying the bias constant rather than changing the decay structure.

In the neural network simulations, the factorial decomposition reveals a dominant interaction effect (0.2071) that exceeds both marginal effects in magnitude. This indicates that the combination of non-stationary data and coupled dynamics produces emergent persistence mechanisms not captured by either factor alone. The

full actor–critic (R4) achieves a tail bias of 0.3668, compared to the baseline of 0.3213.

These findings support the conjecture that both sources contribute to persistent bias, with coupling as the structural mechanism (slowing the decay rate) and online sampling as the amplifying mechanism (increasing the bias constant). Their interaction further compounds the effect in the full actor–critic setting.

REFERENCES

[1] Francis Bach and Eric Moulines. 2013. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, Vol. 26.

[2] Jalaj Bhandari, Daniel Russo, and Raghav Singal. 2018. A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. *arXiv preprint arXiv:1806.02450* (2018).

[3] Vivek S. Borkar. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

[4] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. 2022. Finite-Time Analysis of Single-Timescale Actor-Critic. In *Advances in Neural Information Processing Systems*, Vol. 35.

[5] Efthymios Georgoudios et al. 2026. Scaling Effects and Uncertainty Quantification in Neural Actor Critic Algorithms. *arXiv preprint arXiv:2601.17954* (2026).

[6] Vijay R. Konda and John N. Tsitsiklis. 2003. On Actor-Critic Algorithms. *SIAM Journal on Control and Optimization* 42, 4 (2003), 1143–1166.

[7] Qianxiao Li, Cheng Tai, and Weinan E. 2019. Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research* 20, 40 (2019), 1–47.

[8] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. 2021. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Conference on Learning Theory*. PMLR, 3548–3626.

[9] R. Srikant and Lei Ying. 2019. Finite-Time Error Bounds for Linear Stochastic Approximation and TD Learning. *arXiv preprint arXiv:1902.00923* (2019).

[10] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

[11] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Advances in Neural Information Processing Systems* 12 (2000).

[12] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. 2020. Finite-Time Analysis of the Actor-Critic Algorithm for the Linear Quadratic Regulator. In *International Conference on Machine Learning*. PMLR, 10458–10468.

[13] Tengyu Xu, Zhe Wang, and Yingbin Liang. 2020. Improving Sample Complexity Bounds for Actor-Critic Algorithms. *arXiv preprint arXiv:2004.12956* (2020).