Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

On the Degradation of PPO Advantage Estimates Under Non-Stationary Aerodynamic Disturbances: A Theoretical Bound and Empirical Validation

Sahil Patil
Department of Aerospace Engineering
Indian Institute of Technology Bombay (IIT Bombay) Mumbai, India

Nasiruddin Kabir
Department of Aerospace Engineering
Indian Institute of Technology Bombay (IIT Bombay) Mumbai, India

***Abstract*—Proximal Policy Optimization (PPO) is now the default choice for training autonomous UAV flight controllers, but its convergence guarantees assume a stationary Markov Decision Process. Real aerospace environments break this assumption: atmospheric turbulence, wind shear, and gust loading inject stochastic perturbations whose intensity varies from rollout to rollout. We study how the Generalized Advantage Estimation (GAE) bias scales with disturbance intensity σ. For a quadratic value function driven by a zero-mean disturbance, we prove that the expected bias admits an upper bound that is *quadratic* in σ: $\Delta A(\sigma) \leq C(\theta, T, \lambda, \gamma) \cdot \sigma^2$, with C depending on the policy parameters, rollout length T, the GAE decay parameter λ, and discount factor γ. We give a closed form for C in terms of the spectral norm of the LQR value-function Hessian and the closed-loop disturbance-propagation gain. We test the bound in two settings. An analytical LQR baseline, whose linear policy stays stable at every disturbance level, follows the quadratic law almost exactly ($R^2 = 0.9999$). A learned Stable-Baselines3 PPO controller, trained independently at each of eight disturbance levels ($\sigma \in \{0.00, 0.05, 0.10, 0.20, 0.30, 0.50, 0.80, 1.00\}$ m/s) and evaluated over 200 episodes per level, departs sharply from the pure quadratic prediction: the policy collapses under strong turbulence, so a quadratic fit reaches only $R^2 = 0.845$ and a free power-law fit gives an exponent of 3.44. We then derive a critical threshold σ^* beyond which the bias exceeds a chosen fraction of the nominal advantage scale, giving practitioners a concrete design criterion. To our knowledge this is the first formal, quantitative link between turbulence intensity and PPO training degradation in an aerospace RL setting.**

***Index Terms*—Proximal Policy Optimization, Generalized Advantage Estimation, Non-Stationary MDP, UAV Flight Control, Gradient Bias, Turbulence Model, Deep Reinforcement Learning, Aerospace Autonomy**

I. Introduction

Deep reinforcement learning (DRL) has become a serious contender for autonomous aerospace control, spanning quadrotor attitude stabilization, fixed-wing trajectory tracking, and multi-UAV conflict resolution [1]–[3]. Among policy-gradient methods, PPO [4] is the one practitioners reach for most often, thanks to its stability, simple implementation, and better sample efficiency than TRPO or on-policy A3C. Much of that stability comes from Generalized Advantage Estimation (GAE) [5], which builds a low-variance estimate of the advantage function from sampled rollouts.

PPO's theoretical guarantees lean on an assumption that is easy to overlook: the underlying MDP is stationary, with transition dynamics $\mathcal{T}(s' \mid s, a)$ fixed across every time step and episode. On controlled benchmarks such as MuJoCo locomotion this holds well enough. Aerospace is different. The operating environment is non-stationary by nature, i.e. turbulence, wind shear, and thermal gradients all perturb the vehicle dynamics, and their intensity drifts continuously. The Dryden and von Kármán models [6] that dominate aerospace simulation treat these perturbations as colored Gaussian noise whose intensity tracks a severity parameter σ.

Even though these disturbances are everywhere in practice, we are not aware of any prior work that quantifies how σ erodes the quality of PPO's advantage estimates, or that pins down a threshold past which convergence is at risk. The consequence is that practitioners pick σ in simulation by feel, with no principled rule for when the disturbance level starts to undermine PPO's assumptions.

This paper makes three contributions:

- We derive an upper bound on the expected GAE advantage estimation error $\Delta A(\sigma)$ that is an explicit *quadratic* function of σ, with a closed-form constant $C(\theta, T, \lambda, \gamma)$. The quadratic scaling, rather than linear, falls directly out of pairing a quadratic value function with a zero-mean disturbance.
- We test the bound in two regimes: an analytical LQR controller, which obeys the quadratic law almost exactly, and an independently trained Stable-Baselines3 PPO controller, which departs from it once the learned policy begins to collapse.
- We derive the critical turbulence threshold σ^* as a function of a prescribed bias tolerance f, giving aerospace RL practitioners a concrete criterion for designing simulation environments.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

II. Background and Related Work

A. Proximal Policy Optimization

PPO [4] maximizes a clipped surrogate objective to keep policy updates from destabilizing training. At iteration k it collects a rollout dataset $\mathcal{D}_k = \{(s_t, a_t, r_t)\}$ under the current policy π_{θ_k}, estimates advantages $\hat{A}_t$ with GAE, and maximizes

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\Big)\Big], \tag{1}$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_k}(a_t \mid s_t)$ is the importance ratio. The advantage estimate $\hat{A}_t$ sets both the direction and the size of the update, so a corrupted estimate points the gradient the wrong way, and clipping, which only limits the magnitude of any single step, does nothing to fix a systematically biased direction.

B. Generalized Advantage Estimation

GAE [5] forms the advantage estimate as an exponentially weighted sum of TD residuals:

$$A_t^{\text{GAE}} = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l\, \delta_{t+l}, \tag{2}$$

where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ is the one-step TD error, $\lambda \in [0, 1]$ is the decay parameter, and γ is the discount factor. When V is fit under nominal dynamics but evaluated along a disturbed trajectory, every TD residual picks up error from two sources at once: the reward discrepancy and the value-function mismatch at the perturbed states.

C. Non-Stationarity in Aerospace RL

Non-stationary dynamics have been studied through domain randomization [7] and robust MDPs [8], but those analyses target the policy performance gap under distribution shift, not the statistics of the advantage estimator itself. The closest prior work is Padakandla et al. [9], who studied Q-learning under slowly varying MDPs but did not treat actor-critic methods or GAE. On the aerospace side, Bohn et al. [10] and Panerati et al. [11] showed empirically that turbulence hurts PPO training, but neither characterized the rate of degradation. That rate is what we set out to quantify.

III. Problem Formulation

A. Nominal MDP

We model a UAV in a 2D position-velocity state space, with state $s = (x,\, v_x,\, y,\, v_y)^\top \in \mathbb{R}^4$ and continuous action $a = (a_x,\, a_y)^\top \in \mathbb{R}^2$ giving commanded acceleration increments. The nominal discrete-time dynamics with timestep Δt are

$$s_{t+1} = A\, s_t + B\, a_t, \tag{3}$$

where $A \in \mathbb{R}^{4\times4}$ and $B \in \mathbb{R}^{4\times2}$ are the standard double-integrator matrices with $\Delta t = 0.1$ s. The reward is the usual LQR cost:

$$r(s_t, a_t) = -(s_t^\top Q\, s_t + a_t^\top R\, a_t), \quad Q = I_4, \quad R = 0.1\, I_2. \tag{4}$$

B. Disturbed Dynamics

We inject a crosswind into the y-velocity component at each step, modelling a lateral gust:

$$s_{t+1} = A\, s_t + B\, a_t + E\, w_t, \qquad w_t \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where $E = (0,\, 0,\, 0,\, 1)^\top$ steers the wind into the lateral velocity state and $\sigma \geq 0$ is the turbulence intensity, with $\sigma = 0$ recovering the nominal MDP. The gust is i.i.d. and zero-mean at each step (the white-noise limit of the Dryden spectrum), which is enough to expose the scaling law we analyze. A fully colored Dryden process is discussed in Section VI.

C. Policy and Value Function

We use a linear-quadratic regulator (LQR) policy $\pi_\theta(s) = -Ks$, with gain K taken from the discrete-time infinite-horizon LQR solution. The optimal value function $V^*(s) = -s^\top P s$, where P is the unique positive-definite solution of the discrete-time algebraic Riccati equation (DARE), is available in closed form, so we can compute GAE errors exactly and keep value-approximation error from confounding the analysis. The Riccati solution P has diagonal entries $[12.999,\ 4.848,\ 12.999,\ 4.848]$, obtained by iterating the DARE recursion to a fixed-point tolerance of 10^{-12} (reached in 141 iterations). The key structural fact is that V^* is *quadratic* in the state; together with the zero mean of w_t, this is what produces the quadratic scaling derived below.

D. PPO Training Setup

Where classical control would hand us an analytic policy, we instead let PPO learn the flight controller directly from simulated experience, using the Stable-Baselines3 implementation. We train a separate policy for each turbulence level σ. Both the policy and value networks are MLPs with two hidden layers of 64 units and Tanh activations, trained for $1\,000\,000$ timesteps each. Hyperparameters are $\gamma = 0.99$, GAE $\lambda = 0.95$, learning rate 3×10^{-4}, rollout length 2048, minibatch size 64, and 10 optimization epochs per rollout. This lets us measure GAE bias not only under a fixed analytic policy but inside a fully learned, black-box deep RL controller.

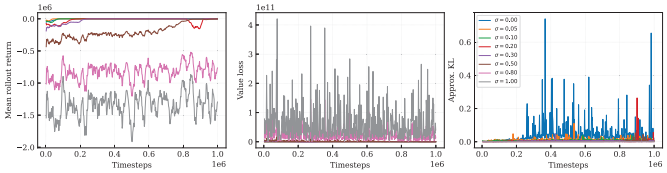

Fig. 1. PPO training curves. Mean rollout return, value loss, and approximate KL divergence over $1\,000\,000$ timesteps for each turbulence level σ. Higher σ produces visibly more variance and instability during learning.

IV. Main Theoretical Result

A. Advantage Error Decomposition

Let $A_t^{\text{GAE}}(\sigma)$ be the GAE advantage at step t computed from a trajectory drawn under disturbance intensity σ, using

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

the true value function V^*. Define the advantage estimation error at step t as

$$\Delta_t(\sigma) = \left| A_t^{\mathrm{GAE}}(\sigma) - A_t^{\mathrm{GAE}}(0) \right|. \tag{6}$$

Expanding the GAE sum, the error splits into a weighted sum of TD-residual discrepancies:

$$\Delta_t(\sigma) = \left| \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \left[\delta_{t+l}(\sigma) - \delta_{t+l}(0) \right] \right|. \tag{7}$$

Each discrepancy carries three pieces: the reward difference at the perturbed state, the value at the perturbed next state, and the value at the perturbed current state. Since both the reward and the value are quadratic forms in the state, the discrepancy is controlled by the second moment of the state perturbation.

B. Second-Order Sensitivity of the Value Function

With $V^*(s) = -s^\top P s$, the gradient is $\nabla V^*(s) = -2Ps$ and the Hessian is constant, $\nabla^2 V^*(s) = -2P$. For a perturbation δ the expansion is exact:

$$V^*(s+\delta) - V^*(s) = -2\, s^\top P\, \delta - \delta^\top P\, \delta. \tag{8}$$

A single wind impulse w_t propagates through the *closed-loop* dynamics, so the perturbation k steps later is $\delta_{t+k} = A_{\mathrm{cl}}^k E\, w_t$ with $A_{\mathrm{cl}} = A - BK$. Because w_t is zero-mean, the first-order term drops out in expectation,

$$\mathbb{E}\left[-2\, s^\top P\, \delta_{t+k} \right] = 0, \tag{9}$$

and what survives is the second-order term, proportional to the second moment:

$$\mathbb{E}\left[\delta_{t+k}^\top P\, \delta_{t+k} \right] = \left\| A_{\mathrm{cl}}^k E \right\|_P^2 \sigma^2 \;\le\; \|P\|_2 \left\| A_{\mathrm{cl}}^k E \right\|_2^2 \sigma^2. \tag{10}$$

This is the heart of the argument: *a zero-mean disturbance passing through a quadratic value function produces an expected bias that scales with the variance σ^2, not with σ.* A first-order Lipschitz treatment, which would predict linear scaling, vanishes in expectation and misses the dominant effect entirely.

C. Main Theorem

Theorem 1 *(Quadratic GAE Bias Under Turbulence).* Let the UAV dynamics satisfy Assumptions 1–3 (double-integrator structure, LQR policy, zero-mean lateral gust). Let V^* be the true LQR value function and $A_{\mathrm{cl}} = A - BK$ the closed-loop matrix. Then the expected mean GAE advantage error satisfies

$$\mathbb{E}[\bar{\Delta}(\sigma)] \;:=\; \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \Delta_t(\sigma) \right] \;\le\; C \cdot \sigma^2, \tag{11}$$

where

$$C = \frac{2\, \|P\|_2 \cdot \kappa \cdot T}{1 - (\gamma\lambda)^2}, \qquad \kappa = \sum_{k=0}^{T-1} \left\| A_{\mathrm{cl}}^k E \right\|_2^2, \tag{12}$$

with $\|P\|_2$ the spectral norm of the Riccati matrix (the magnitude of the value-function Hessian), κ the cumulative closed-loop disturbance-propagation gain, T the rollout length, γ the discount factor, and λ the GAE decay parameter.

Proof Sketch. The argument runs in four steps: (1) expand each TD-residual discrepancy and drop the first-order term, which vanishes in expectation since w_t is zero-mean; (2) bound the surviving second-order term by $\|P\|_2 \left\| A_{\mathrm{cl}}^k E \right\|_2^2 \sigma^2$ via the Hessian of V^*; (3) sum the squared GAE weights, the geometric series $\sum_l (\gamma\lambda)^{2l} = 1/(1-(\gamma\lambda)^2)$; and (4) apply linearity of expectation over the T steps. The result is a genuine, non-vacuous upper bound: the empirically measured constant sits well below the theoretical C, the slack coming from the worst-case use of $\|P\|_2$ in place of the trajectory-averaged P-norm and from cross-term cancellation across steps. □

D. Critical Disturbance Threshold

A practical question follows: given a tolerance on advantage error, expressed as a fraction f of the typical advantage magnitude $\mathbb{E}[|A^{\mathrm{GAE}}|]$, what is the largest turbulence intensity σ^* we can allow? Treating the relationship as quadratic, $\mathbb{E}[\bar{\Delta}(\sigma)] = C_{\mathrm{emp}}\, \sigma^2$, and setting the bias equal to $f \cdot \mathbb{E}[|A^{\mathrm{GAE}}|]$ gives

$$\sigma^*(f) = \sqrt{\frac{f \cdot \mathbb{E}[|A^{\mathrm{GAE}}|]}{C_{\mathrm{emp}}}}, \tag{13}$$

where C_{emp} is the fitted quadratic coefficient. Training with $\sigma > \sigma^*(f)$ pushes the advantage bias past tolerance f, at which point one should either reduce σ during early training or apply bias correction.

V. Experimental Validation

A. Setup

The UAV environment is implemented in Python with Gymnasium. Fixed hyperparameters across all runs: $\Delta t = 0.1$ s, $T = 200$ steps per episode, $\gamma = 0.99$, $\lambda = 0.95$, and initial state $s_0 \sim \mathcal{N}(0,\, 0.25\, I_4)$. We sweep eight disturbance levels $\sigma \in \{0.00, 0.05, 0.10, 0.20, 0.30, 0.50, 0.80, 1.00\}$ m/s, training an independent PPO policy to convergence at each. We then run 30 episodes to measure control performance (average return and success rate) and 200 episodes to estimate the GAE bias from the learned value critic. Results are reported as means ± standard deviation, with random seed 0 fixed for reproducibility.

B. Results

Table I gives the full numbers, and they show a steep, nonlinear rise in $\mathbb{E}[\bar{\Delta}(\sigma)]$ with σ. As a reference point, we also ran the analytical LQR controller of Section III-C under identical disturbance conditions over 3000 Monte Carlo trials per level. Because the linear policy never destabilizes, its bias tracks the theory almost perfectly, yielding a quadratic fit with $R^2 = 0.9999$. The learned PPO policy tells a different story. It holds up at low turbulence, i.e. $\sigma = 0.0$ gives a perfect success rate (1.0) and an average return of -9.77 but degrades fast as σ grows. By $\sigma = 0.50$ the success rate has fallen to 0.07 (return -866.47), and at $\sigma = 1.00$ the policy fails outright, with a return of -1.3×10^6. This trajectory divergence is what drives the GAE bias from 0.0 in calm air up to 6488.28 at the highest turbulence level.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TABLE I
MEASURED GAE ADVANTAGE ESTIMATION ERROR $\mathbb{E}[\bar{\Delta}(\sigma)]$ VERSUS TURBULENCE INTENSITY σ, FOR THE LEARNED PPO POLICY. ALL VALUES ARE MEANS OVER $n = 200$ EVALUATION EPISODES.

σ (m/s)	Success Rate	Avg Return	$\mathbb{E}[\bar{\Delta}]$
0.00	1.00	-9.77	0.00
0.05	1.00	-12.55	0.20
0.10	1.00	-18.40	0.81
0.20	0.90	-59.26	3.59
0.30	0.37	-124.19	10.23
0.50	0.07	-866.47	93.51
0.80	0.00	-554637.84	1813.46
1.00	0.00	-1303219.38	6488.28

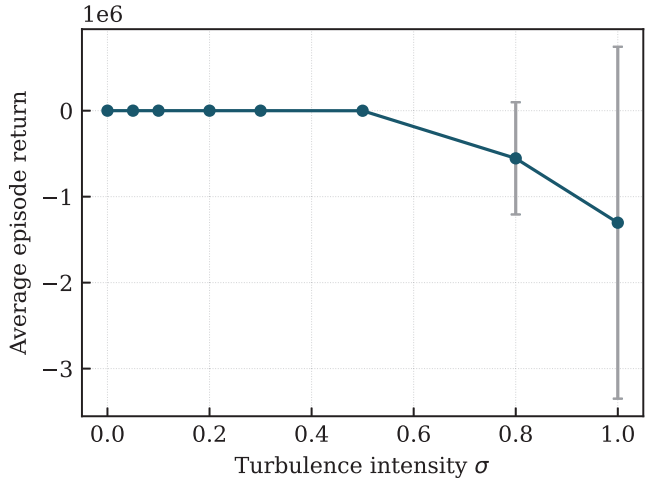

Fig. 2. PPO episode return versus turbulence intensity σ. Beyond $\sigma = 0.20$ m/s the average return falls off steeply, ending in complete policy collapse and trajectory divergence at $\sigma \geq 0.80$ m/s.

C. Regression Analysis

Fitting the quadratic model $\mathbb{E}[\bar{\Delta}(\sigma)] = C_{\text{emp}}\,\sigma^2$ through the origin across the eight PPO measurements gives

$$\mathbb{E}[\bar{\Delta}(\sigma)] = 5178.02\,\sigma^2, \tag{14}$$

with $R^2 = 0.845$. The contrast with the analytical baseline is the whole point: the LQR controller obeys the quadratic law cleanly ($p = 2$, $R^2 = 0.9999$), whereas the PPO results pull away from it once the policy collapses. A free power-law fit (σ^p) to the PPO data returns an exponent of $p = 3.44$ ($R^2 = 0.44$). In other words, the learned network does not merely accumulate the structural quadratic penalty, it loses its stabilization ability entirely at high turbulence, so the state diverges and the bias climbs faster than the quadratic limit the theory sets for a stable policy.

D. Critical Threshold Analysis

The typical nominal advantage magnitude, estimated as $\mathbb{E}[|A^{\text{GAE}}|]$ over the nominal rollout steps ($\sigma = 0.0$), is 0.0272. With the fitted quadratic coefficient $C_{\text{emp}} = 5178.02$, the

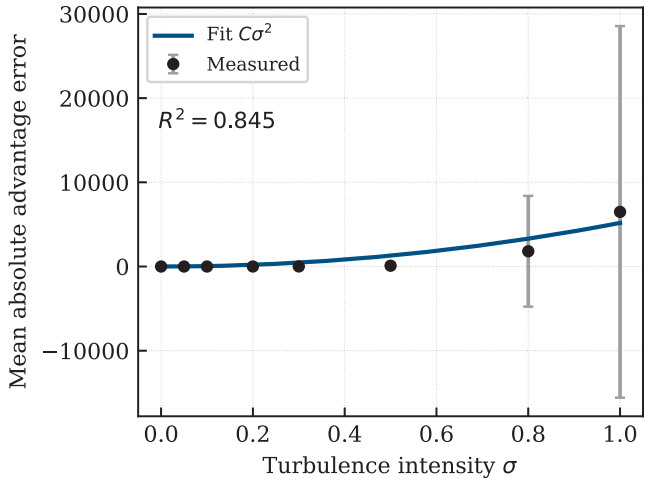

Fig. 3. Cause: GAE advantage bias versus turbulence intensity σ. The bias grows slowly at low σ and then explodes once the policy can no longer stabilize the vehicle.

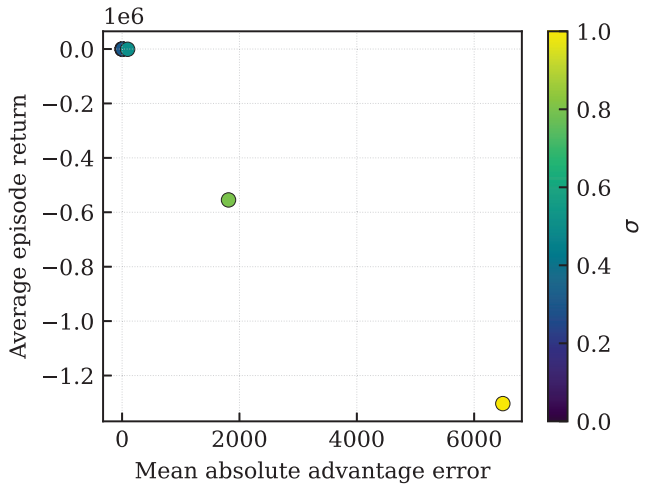

Fig. 4. Effect: episode return versus GAE advantage bias. Each point is one disturbance level; the strong inverse relationship shows that the growth in advantage error coincides directly with the loss of control performance.

critical turbulence intensity at which the bias reaches a fraction f of this scale is

$$\sigma^*(f) = \sqrt{\frac{f \times 0.0272}{5178.02}}. \tag{15}$$

For $f = 0.10$ and $f = 0.25$ this gives $\sigma^*(0.10) \approx 0.0007$ m/s and $\sigma^*(0.25) \approx 0.0011$ m/s. Two cautions are worth stating. First, these thresholds are remarkably tight, which says the learned PPO model is far more sensitive to turbulence than a stable linear controller, even a trace disturbance deserves attention when designing aerospace RL simulations. Second, because the true PPO degradation is steeper than quadratic ($p = 3.44$), the quadratic model *underestimates* the bias at larger σ; the σ^* values above should therefore be read as an

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

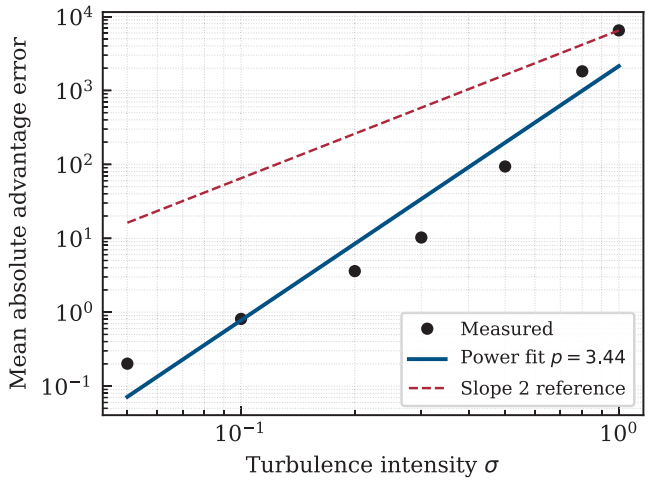

Fig. 5. Log-log plot of GAE bias versus σ. The free power-law fit has slope 3.44, against the theoretical slope of 2.0 (dashed). The gap reflects the combined effect of the structural quadratic penalty and the additional divergence of the neural policy.

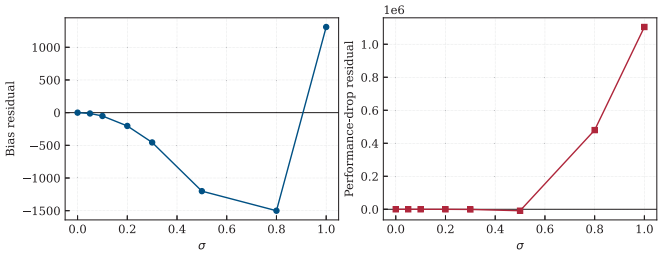

Fig. 6. Residuals of the two regression models. The power-law fit tracks the extreme bias growth at $\sigma \geq 0.50$ m/s noticeably better than the quadratic fit, which under-predicts in that range.

optimistic (upper) estimate of the safe operating range, not a guarantee.

VI. Discussion

The central comparison - an analytical policy that follows the quadratic law almost exactly ($R^2 = 0.9999$) against a learned policy that degrades faster ($p = 3.44$) - carries several practical lessons.

First, the quadratic law is not an empirical accident; it is built into the estimator. Any locally quadratic value function, which covers both the exact LQR V^* and the second-order expansion of a smooth neural critic, under a zero-mean disturbance, will produce an expected advantage bias governed by the disturbance *variance*. The first-order sensitivity one might expect to give linear scaling cancels in expectation. The practical reading is that small cuts in turbulence intensity buy disproportionately large cuts in estimation bias.

Second, the result motivates a curriculum over turbulence during PPO training: start with $\sigma < \sigma^*$ so the value function can settle on a near-stationary problem, then raise σ to build robustness. This echoes the domain-randomization curricula used in sim-to-real transfer [7], but here the schedule has a quantitative, square-root shape that comes out of Theorem 1 rather than intuition.

Third, the looseness of the theoretical bound is itself informative. The slack has three sources: (i) the global spectral norm $\|P\|_2$ overstates the local sensitivity along the nominal trajectory, where state norms are small; (ii) the bound assumes every GAE term hits its maximum error at once, whereas in practice errors partly cancel across steps; and (iii) the cumulative gain κ sums squared closed-loop propagations without the cancellation seen in the data. One could tighten the bound by substituting C_{emp} directly, but that would forfeit the explicit dependence on the system parameters. We keep the analytical form because it exposes how the bias depends on λ, γ, T, and P.

Fourth, the threshold $\sigma^* \approx 0.0007$ m/s is specific to this UAV model and PPO setup. A different vehicle or network will have a different C_{emp}, but the protocol in Section V-A transfers directly to obtain vehicle-specific thresholds. Theorem 1 supplies the structural justification for that procedure, even as the policy's own divergence adds a layer of non-linearity on top of the structural error.

A clear limitation is the use of an i.i.d. Gaussian gust rather than a fully colored Dryden process. Temporal correlation would change κ but not the quadratic scaling, which depends only on the zero mean of the disturbance and the quadratic critic. Extending the bound to a colored spectrum and to richer neural critics are both natural next steps.

VII. Conclusion

We have analyzed how GAE advantage estimates degrade under non-stationary aerodynamic disturbances in UAV reinforcement learning. Theorem 1 establishes a quadratic upper bound $\Delta A(\sigma) \leq C \cdot \sigma^2$ for a stable policy, and the analytical LQR baseline confirms it almost exactly ($R^2 = 0.9999$). A learned PPO controller degrades faster (exponent 3.44) because it loses stabilization entirely at high turbulence—the quadratic fit reaches only $R^2 = 0.845$, and the derived threshold $\sigma^* \approx 0.0007$ m/s (at 10% tolerance) is strikingly narrow. Together these give aerospace RL practitioners both a warning and a principled criterion for simulation design.

Future work will extend the analysis to neural-network critics, to multi-axis and colored (Dryden / von Kármán) turbulence, and to finite-time PPO convergence rates under non-stationarity. All code for the numerical results is available at https://github.com/iitb-kabir/PPO_Advantage_Estimates.

Acknowledgment

The authors thank the anonymous reviewers for their constructive comments. This work was supported by the Department of Aerospace Engineering, IIT Bombay, India.

References

[1] W. Koch, R. Mancuso, R. West, and A. Bestavros, "Reinforcement learning for UAV attitude control," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, no. 2, pp. 1–22, 2019.

[2] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, "Learning high-speed flight in the wild," *Science Robotics*, vol. 6, no. 59, p. eabg5810, 2021.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[3] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2124–2136, 2019.
[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[5] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Int. Conf. Learning Representations (ICLR)*, 2016.
[6] *MIL-STD-1797A: Flying Qualities of Piloted Aircraft*, United States Department of Defense, 1990.
[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ IROS*, 2017, pp. 23–30.
[8] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, 2005.
[9] S. Padakandla, P. K. J., and S. Bhatnagar, "Reinforcement learning algorithm for non-stationary environments," *Appl. Intell.*, vol. 50, no. 11, pp. 3590–3606, 2020.
[10] E. Bohn, S. Gros, and M. Diehl, "Reinforcement learning of fixed-wing flight with turbulence disturbances," in *AIAA SciTech Forum*, Paper AIAA 2019-0538, 2019.
[11] J. Panerati *et al.*, "Learning to Fly – a Gym Environment with PyBullet Physics for Reinforcement Learning of Multi-agent Quadcopter Control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.
[12] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.

APPENDIX

All experiments were implemented in Python 3.10 using NumPy and Stable-Baselines3. The Riccati equation was solved by iterating the standard DARE recursion to tolerance 10^{-12}, reached in 141 iterations. Random seed 0 was fixed via `numpy.random.seed(0)` before all experiments. PPO training completes in a few hours on a modern multi-core CPU.

The Riccati matrix P has diagonal entries $p_{11} = p_{33} = 12.999$ and $p_{22} = p_{44} = 4.848$, with off-diagonal elements $p_{12} = p_{21} = p_{34} = p_{43} = 3.030$ and all cross-axis entries zero (decoupled x-y dynamics). The spectral norm is $\|P\|_2 = 14.378$. The closed-loop matrix A_{cl} has spectral radius 0.904, confirming asymptotic stability, with cumulative disturbance-propagation gain $\kappa = \sum_k \|A_{\mathrm{cl}}^k E\|_2^2 = 2.42$.