Application of Safety-Constrained Reinforcement Learning in Robust UAV Hovering

Yong-Jie Ning, Bo-Cing Wu, Guan-Yu Chen, Po-Ju Chen, Yan-Hao Huang*

Abstract

The evolution of autonomous control in unmanned aerial vehicles (UAVs) is transitioning from algorithms that rely purely on trial-and-error learning toward safety-constrained control architectures. This study proposes and empirically validates a novel control system that fuses physical boundary constraints via Control Barrier Functions (CBFs) with the Soft Actor-Critic (SAC) algorithm. Relying exclusively on physical constraints to precisely delineate a safety manifold, this framework enforces strict hard filtering and enables safe policy exploration. Comparative experiments demonstrate that this mechanism not only eradicates collision risks during the initial stages of training but also significantly enhances the stability of the policy training and convergence process by filtering out substantial amounts of hazardous and invalid exploratory actions. These results provide a robust theoretical and empirical foundation for developing autonomous UAV systems characterized by high stability and hard safety guarantees.

Keywords: Unmanned Aerial Vehicle, Reinforcement Learning, Soft Actor-Critic (SAC), Control Barrier Function, Robust Hovering.

Yong-Jie Ning, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11122130@gm.nttu.edu.tw

Bo-Cing Wu, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11122121@gm.nttu.edu.tw

Guan-Yu Chen, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11325103@gm.nttu.edu.tw

Po-Ju Chen, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11325102@gm.nttu.edu.tw

Yan-Hao Huang (Corresponding Author), Assistant Professor, Department of Green Energy and Information Technology, National Taitung University, E-mail: yhhuang@nttu.edu.tw

安全約束強化學習在穩健無人機懸停的應用

甯勇捷、吳柏慶、陳冠佑、陳柏儒、黃彥皓*

摘要

無人機自主控制技術正從單純依賴試錯學習的演算法，逐步邁向受安全約束的控制架構。本研究提出並實證了一種新型控制系統，將基於控制屏障函數 (CBFs) 的物理邊界約束，與柔性行動者-評論家 (Soft Actor-Critic, SAC) 演算法進行深度融合。本框架僅憑藉物理約束來精準界定安全流形，藉此執行嚴格的硬性過濾並實現安全的策略探索。對比實驗結果表明，此機制不僅徹底消弭了訓練初期的碰撞風險，更透過濾除大量具危險性且無效的探索動作，顯著提升了策略訓練與收斂過程的穩定性。本研究之成果，為建構兼具高穩定性與硬性安全保證的無人機自主系統，奠定了堅實的理論與實證基礎。

關鍵字：無人機、強化學習、柔性行動者-評論家 (SAC)、控制屏障函數、穩健懸停。

甯勇捷，國立臺東大學綠能與資訊科技學系學生，E-mail: 11122130@gm.nttu.edu.tw
吳柏慶，國立臺東大學綠能與資訊科技學系學生，E-mail: 11122121@gm.nttu.edu.tw
陳冠佑，國立臺東大學綠能與資訊科技學系學生，E-mail: 11325103@gm.nttu.edu.tw
陳柏儒，國立臺東大學綠能與資訊科技學系學生，E-mail: 11325102@gm.nttu.edu.tw
黃彥皓(通訊作者)，國立臺東大學綠能與資訊科技學系助理教授，E-mail: yhhuang@nttu.edu.tw

1. Introduction

As unmanned aerial vehicle (UAV) applications grow increasingly pervasive (Rezwan & Choi, 2022), achieving safety guarantees in UAV control has emerged as a central challenge (Ma et al., 2024). To address this issue, this research explores a strictly self-constrained control architecture by defining physical boundary laws as Control Barrier Functions (CBF) (Ames et al., 2019). These functions are deployed as real-time action filters integrated within the Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018).

This unified strategy, hereafter denoted as the CBF-SAC architecture, leverages SAC's intrinsic maximum entropy mechanism to encourage robust exploration. To bridge the gap between stochastic exploration and operational safety, the CBF component serves as an execution-level safeguard. The fundamental value of this framework lies in its capacity to constrain the agent's behavior entirely within a mathematically defined safety manifold. The primary contributions of this work are summarized as follows:

- Novel Hybrid Control Architecture: We propose the CBF-SAC framework, which integrates Control Barrier Functions as a real-time execution-level filter within the SAC algorithm.
- Hard Safety Guarantees: By defining physical boundary laws through CBFs, we provide a mathematically defined safety manifold that strictly enforces UAV operational limits (30cm to 200cm) regardless of the agent's exploratory actions.
- Enhanced Training Stability: Empirical results demonstrate that the proposed architecture eliminates 99.9% of safety violations (reducing instances from 1,567 to 2) and mitigates gradient volatility, ensuring a smooth and monotonic convergence process compared to unconstrained baselines.

2. Methodology

2.1 SAC algorithm

The autonomous hovering task is modeled as a Markov Decision Process (MDP). In this research, we employ the SAC algorithm (Haarnoja et al., 2018) to navigate the continuous state and action spaces of a UAV. SAC operates within a maximum entropy framework by modifying the standard reinforcement learning objective to incorporate the expected entropy of the policy. This yields the following augmented objective:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\pi}\left[r(s_t,a_t) + \alpha\mathcal{H}\left(\pi(\cdot\,|s_t)\right)\right]. \tag{1}$$

where $r(s_t, a_t)$ represents the reward function designed to penalize altitude error, and $\mathcal{H}(\pi)$ is the entropy term that encourages the UAV to explore diverse thrust commands. The temperature coefficient α balances the trade-off between reward exploitation and entropy exploration, directly influencing the stochasticity of the UAV's flight behavior. In our implementation, the state s_t is defined as the one-dimensional altitude error, while the action a_t corresponds to the normalized vertical thrust within the range of [-1.0, 1.0]. The actor network updates its parameters $\emptyset$ by minimizing the expected Kullback-Leibler (KL) divergence, effectively projecting the policy toward the exponential of the soft Q-values:

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[D_{KL}\left(\pi_\phi(\cdot\,|s_t) \| \frac{exp\big(Q_\theta(s_t,\cdot)\big)}{Z_\theta(s_t)}\right)\right] \tag{2}$$

Here, $Q_\theta(s_t,\cdot)$ denotes the soft Q-value that estimates the long-term return of a specific thrust action, and $Z_\theta(s_t)$ is the partition function that normalizes the target distribution. By integrating these objectives, the SAC agent learns to map altitude errors to optimal motor commands without prior knowledge of the vehicle's mass. This maximum entropy approach ensures that the UAV can reliably discover the equilibrium thrust needed for a stable hover within the AirSim environment.

2.2 Safety Filtering Mechanism of CBF

Although the SAC approach excels at driving exploratory behavior and optimizing cumulative returns, its fundamental reliance on stochastic sampling cannot inherently ensure operational safety during the policy learning process. To address this critical deficiency and enforce absolute physical boundaries, our methodology incorporates a CBF (Ames et al., 2019) acting as an execution-level safeguard. Building upon the robust safety manifold and dynamic velocity constraints validated in recent research (Huang et al., 2025), we define a safe altitude range of 30cm to 200cm. Consequently, only safe actions filtered by the CBF are transmitted to the UAV for execution.

Specifically, as the drone approaches either the upper or lower altitude limits, the permissible velocity toward that boundary decays quadratically to ensure smooth deceleration. Furthermore, if the distance to the boundary falls below a critical 10% threshold of the operational range, the system enforces a strict, constant repulsion limit to absolutely prevent boundary violations.

3. Experiments and Discussion

The experiments and training in this study are all conducted in the Microsoft AirSim virtual simulation environment (Shah et al., 2017) equipped with a high-fidelity physics engine. Regarding the core architecture definition, this research is divided into two groups. The first is the pure SAC baseline group, where the agent learns the control policy from scratch without introducing any safety constraints, outputting an uncorrected raw action based on the current observation space at each time step. The second is the CBF-SAC safety group, which incorporates the CBF as a real-time safety filter.

3.1 Learning Convergence and Safety Boundary Performance

To evaluate the learning dynamics of the control policies, we conducted training sessions where the UAV's initial altitude varied randomly between 50 and 80 cm, with a target hovering altitude fixed at 100 cm. The convergence stability was assessed by recording the actor loss over one million training steps, as illustrated in **Figure 1**.

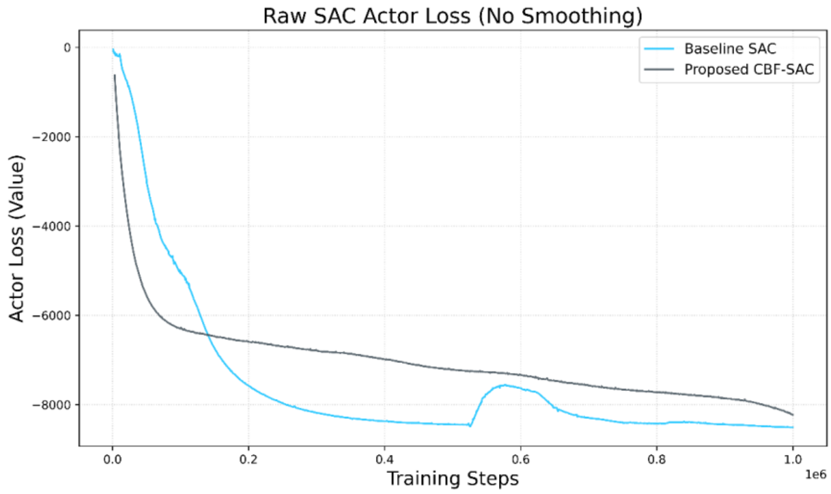

Figure 1. Comparison of raw actor loss during the training phase

In the SAC architecture, the actor network minimizes this loss to maximize the union of expected returns and policy entropy. As shown by the light blue line in Figure 1, the baseline SAC exhibits an initial rapid decline but suffers from a severe gradient rebound near the 500,000-step mark. This instability is fundamentally rooted in unconstrained stochastic exploration; without physical boundaries, the agent frequently enters high-risk states, triggering drastic fluctuations in policy entropy and destabilizing the gradient updates.

In contrast, our proposed CBF-SAC framework (dark gray line) maintains a remarkably smooth and monotonic convergence. This superior stability stems from the real-time pruning of hazardous exploratory actions. By integrating the Control Barrier Function as a persistent safety filter, the agent is prevented from executing invalid actions that would otherwise lead to catastrophic failures and gradient volatility.

Table 1
Number of safety violations.

	SAC	CBF-SAC
Safety Violation Count	1567	2

The quantitative results in **Table 1** further validate the efficacy of this mechanism. While the baseline SAC recorded 1,567 safety violations, the CBF-SAC reduced this figure to a negligible two instances. These rare occurrences are attributed to sampling delays in discrete-time control rather than a failure of the safety manifold itself. By effectively filtering out the vast majority of hazardous actions, the CBF provides an indispensable hard physical guarantee, ensuring that the reinforcement learning process yields a highly reliable and operationally safe control policy.

3.2 Hovering Precision and Dynamic Response Analysis

To evaluate the operational efficacy of the proposed safety-constrained framework, this section conducts a comparative analysis of hovering precision and transient response characteristics. Precise altitude maintenance is a fundamental requirement for UAV autonomous missions, serving as a critical benchmark for both control accuracy and system stability. By subjecting both the baseline SAC and the proposed CBF-SAC models to a standardized hovering task, we aim to investigate how the integration of Control Barrier Functions (CBFs) influences the agent's ability to minimize steady-state error while maintaining a stable vertical trajectory. The following results highlight the distinct behaviors of the two algorithms in reconciling exploratory freedom with rigorous physical constraints.

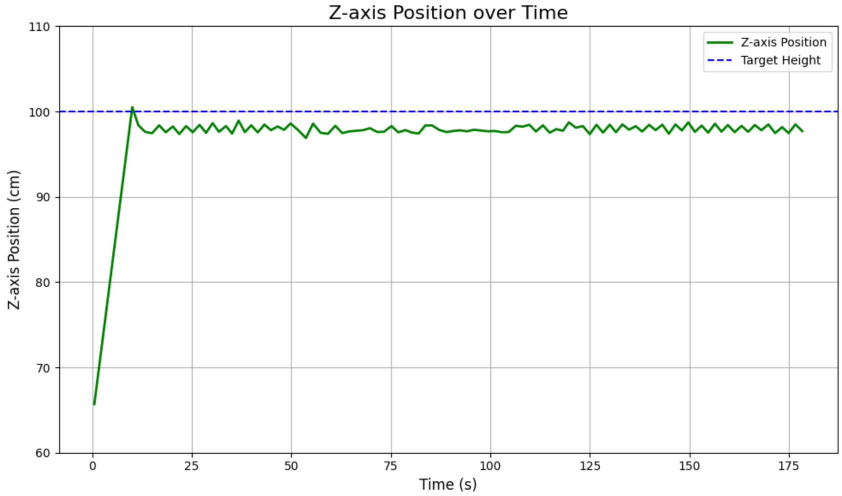

Figure 2. Altitude tracking performance of the baseline SAC algorithm.

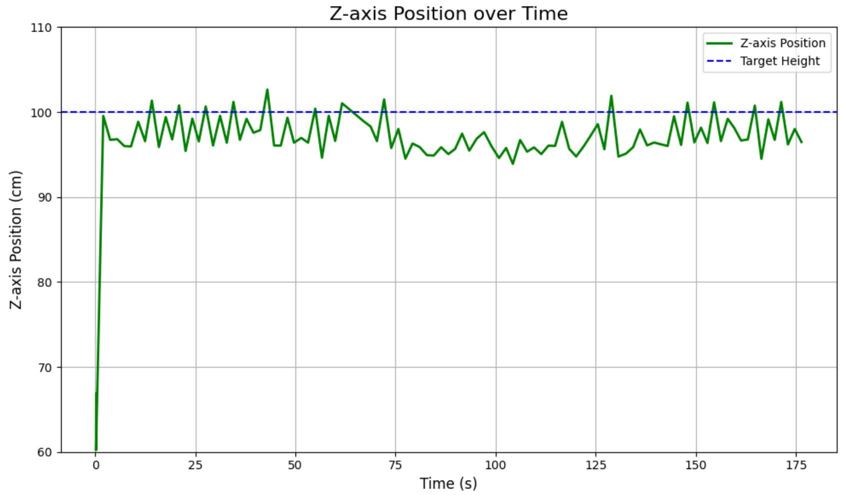

Figure 3. Altitude tracking performance of the baseline CBF-SAC algorithm.

Figures 2 and 3 illustrate the altitude tracking performance of both algorithms during the Z-axis autonomous hovering task, with the target altitude fixed at 100 cm. In the baseline SAC trajectory (Figure 2), the agent converges near 98 cm, resulting in a steady-state error of approximately 2 cm. This suggests that without hard boundary constraints, the SAC agent tends to converge toward a slightly conservative policy to prioritize trajectory smoothness.

In contrast, the CBF-SAC architecture (Figure 3) maintains the UAV much closer to the 100 cm target, yet the trajectory exhibits noticeable high-frequency chattering effects. This phenomenon is fundamentally a manifestation of the intrinsic conflict between SAC's stochastic exploration and the CBF's deterministic safety interventions. While the SAC agent attempts diverse and potentially aggressive actions to maximize policy entropy, the CBF's quadratic decay mechanism imposes strict, non-linear velocity restrictions. Consequently, whenever the agent proposes a radical exploratory action that threatens boundary safety, the CBF instantaneously performs action clipping. This rapid alternation between stochastic exploration and rigid safety truncation interferes with the policy's smoothness, thereby inducing the observed chattering.

Table 2

Comparison of altitude root mean square error (RMSE).

	SAC	CBF-SAC
Altitude Root Mean Square Error (cm)	4.00	6.13

The impact of this chattering is quantified in **Table 2**, which compares the Altitude Root Mean Square Error (RMSE). The data reveals that the baseline SAC achieves an RMSE of 4.00 cm, whereas the CBF-SAC yields a slightly higher RMSE of 6.13 cm. This marginal increase in tracking error is the direct physical consequence of the active interventions required to ensure absolute safety.

Ultimately, these results highlight the core design philosophy of our framework. While the active interventions of the CBF induce a minor drop in trajectory smoothness, they successfully guarantee the absolute physical integrity of the system by eliminating catastrophic boundary violations. In practical UAV applications, a hovering error within six centimeters remains a high standard of precision. Thus, the CBF-SAC architecture represents a justifiable trade-off, accepting a minimal compromise in smoothness to achieve hard safety guarantees during the neural network's exploration process.

4. Conclusion

This study presented and empirically validated a safety-constrained reinforcement learning architecture, CBF-SAC, designed for robust UAV autonomous hovering. By integrating CBF directly into the maximum entropy framework of the SAC algorithm, the proposed methodology successfully enforces strict physical boundaries. Comparative evaluations demonstrated that the CBF-SAC framework drastically reduces catastrophic boundary violations from 1,567 instances in the baseline SAC to merely two—during the critical early training phases. Furthermore, the active pruning of the hazardous action space significantly improves training stability by mitigating the severe gradient fluctuations observed in the unconstrained baseline, ensuring a smooth and monotonic convergence process. Although the instantaneous intervention of the safety filter induces minor high-frequency chattering, the resulting centimeter-level compromise in precision is a justifiable trade-off for absolute physical safety. Future research will focus on extending this architecture to address highly dynamic external disturbances, such as turbulent wind fields, to further advance the robustness of hybrid intelligence in safety-critical autonomous systems.

References

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., & Tabuada, P. (2019, June). Control barrier functions: Theory and applications. In 2019 18th European control conference (ECC) (pp. 3420-3431). Ieee.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (pp. 1861-1870). Pmlr.

Huang, Y. H., Liu, E. J., Wu, B. C., & Ning, Y. J. (2025). Safe UAV Control Against Wind Disturbances via Demonstration-Guided Reinforcement Learning. *Drones*, *10*(1), 2.

Ma, Z., You, J., Zhang, Y., Cheng, Y., & Shao, J. (2024). Reinforcement learning-based dynamic coverage control of multi-rotor UAVs with safety priority. IEEE Transactions on Automation Science and Engineering.

Rezwan, S., & Choi, W. (2022). Artificial intelligence approaches for UAV navigation: Recent advances and future challenges. IEEE access, 10, 26320-26339.

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2017, November). AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In Field and service robotics: Results of the 11th international conference (pp. 621-635). Cham: Springer International Publishing.