Covertness-Aware Trajectory Design for UAV: A Multi-Step TD3-PER Solution

Yuanjian Li, *Student Member, IEEE* and A. Hamid Aghvami, *Fellow, IEEE*
Centre for Telecommunications Research (CTR), King's College London, London WC2R 2LS, U.K.
Email: {yuanjian.li, hamid.aghvami}@kcl.ac.uk

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9839093

***Abstract*—In the presence of Warden's detection, a maximization problem on transmission throughput from unmanned aerial vehicle (UAV) to legitimate nodes is considered and solved via UAV trajectory design, subject to covert, velocity and mobility constraints. With the building-distribution-based pathloss model and the Warden's uncertain location model, the formulated optimization problem is challenging to be tackled through standard offline optimization methods. Alternatively, a twin delayed deep deterministic policy gradient (TD3) approach enhanced by multi-step learning and prioritized experience replay (PER) techniques, termed as multi-step TD3-PER, is proposed to help the UAV adaptively select velocity from continuous action space. Numerical results demonstrate the effectiveness of the proposed multi-step TD3-PER solution and showcase the corresponding superiorities against provided baselines.**

***Index Terms*—Drone, trajectory design, covert communication, deep reinforcement learning.**

I. Introduction

Unmanned aerial vehicles (UAVs) have been widely used in current markets, such as surveillance, emergency rescue and material transport. With controllable mobility, low cost and on-demand deployment, UAVs have also been increasingly applied in the field of wireless communications, e.g., mobile relaying, data collection and wireless coverage. Compared to terrestrial communication networks that are based on fixed-location high-altitude platforms, UAV-aided wireless communication systems are able to establish short-distance transmission links within line of sight (LoS), resulting in better air-to-ground (A2G) channel qualities among transceivers [1].

Due to the broadcasting nature of radio frequency (RF) medium, the confidential information transmitted within UAV-aided wireless communications is vulnerable to be intercepted by malicious eavesdroppers, not to mention that the LoS-dominated A2G wireless links may further deteriorate this transmission security issue. However, the majority of related literature concentrates on preventing the UAVs' transmitted signals from being wiretapped by eavesdroppers, while hiding wireless transmissions initiated from UAVs, i.e., covert communication for UAV-aided networks, has been overlooked. Covert communications aim to help the transmitters achieve low probability of being detected by potential Wardens, which can undoubtedly enhance transmission security in the first place. According to the Square Root Law [2], $\mathcal{O}\left(\sqrt{n}\right)$ bits of information can be transmitted reliably and covertly in $n \to +\infty$ symbols over additive white Gaussian noise (AWGN) channels. In the case of uncertain Warden's noise power, it was proved that the transmitter is able to covertly send $\mathcal{O}\left(\sqrt{n}\right)$ bits to the intended receiver [3]. The possibility and achievable performance of low probability of detection in one-way relay system were examined in [4], in which rate-control and power-control transmission strategies are considered, respectively. In wireless relay networks, the impacts of channel uncertainty on covert communication performance was investigated in [5]. A joint design of transmit power and trajectory was proposed and solved via convex optimization technique in [6], aiming at maximizing UAV's transmission rate under covert constraint.

Although covert transmissions have been intensively studied in the field of terrestrial communications, covert transmissions in UAV-aided networks have not drawn much attention so far, especially on the topic of how to help the UAV achieve low probability of being detected via deep reinforcement learning (DRL)-aided trajectory design. Motivated by the above observations, this paper investigates transmission throughput maximization problem for UAV-mounted network via path planning, subject to covert, velocity and mobility constraints. The main contributions are concluded as follows.

- With the building-distribution-based A2G pathloss model and assuming that the Warden has no exact knowledge of its received noise power, the optimal detection threshold adopted by the Warden is derived. Considering that the UAV cannot gain perfect Warden's location information in practice, the estimated Warden's overall detection error rate on the perspective of the UAV is formulated, which then plays the role as the covert constraint.
- The considered maximization problem is difficult to be solved via standard optimization methods, which is alternatively mapped into a Markov decision process (MDP) and tackled via DRL-aided approach. Specifically, a twin delayed deep deterministic policy gradient (TD3) agent is invoked to help the UAV find proper velocity from continuous action space, alongside the UAV's flight from the initial location to the destination. Furthermore, multi-step learning and prioritized experience replay (PER) techniques are integrated to help the TD3 agent hit a neater training performance.
- To highlight the advantages offered by the proposed multi-step TD3-PER solution, performance comparisons against DRL-based and non-learning baselines, i.e., double duelling deep Q network (D3QN) and straight-line solutions, are provided in numerical results.

II. System Model

Covertness-aware transmissions within UAV-aided network are considered, where a UAV u tries to deliver messages to legitimate nodes $l \in \mathbb{L}$ with low probability of being detected by a Warden w. Note that both the legitimate nodes and the Warden are located on the earth, while all the involved transceivers are equipped with single antenna. The UAV is supposed to reach a predefined destination from its initial location, with static flying altitude A. For clarity, a rectangle subregion $\mathbb{A} : [x_{\text{lo}}, x_{\text{up}}] \times [y_{\text{lo}}, y_{\text{up}}]$ is specified as the UAV's horizontal exploration environment, where the subscripts "lo" and "up" indicate the lower and upper boundaries, respectively. Furthermore, the UAV's horizontal location at time instant $t \in [0, T]$ remains in the range of $\vec{q}_{\text{lo}} \preceq \vec{q}_u(t) \preceq \vec{q}_{\text{up}}$, where $\vec{q}_{\text{lo}} = (x_{\text{lo}}, y_{\text{lo}})$, $\vec{q}_{\text{up}} = (x_{\text{up}}, y_{\text{up}})$, T represents the overall flight duration and $\preceq$ denotes the element-wise inequality. Then, the trajectory of UAV's flight can be fully traced by $\vec{q}_u(t) = (x_u(t), y_u(t))$, starting from the initial location $\vec{q}_u(0) \in \mathbb{R}^{1*2}$ and ending at $\vec{q}_u(T) \in \mathbb{R}^{1*2}$. Unfortunately, continuous time t makes it unaffordable to design the UAV's flight trajectory because infinite location possibilities and velocity constraints are inevitably involved. To circumvent this issue, we uniformly divide the flight duration T into N time slots. The time slot length is delicately defined as a relatively small value $\Delta_t = T/N$, and thus the velocity and the distances from the UAV to the ground nodes can be treated as unchanged within each time slot. Moreover, the horizontal coordinates of legitimate nodes and Warden are indicated as $\vec{q}_l = (x_l, y_l)$ and $\vec{q}_w = (x_w, y_w)$, respectively.

A. Pathloss Model

In accordance with 3GPP urban-micro (UMi) pathloss model [7], the A2G pathloss in dB from $i \in \{l, w\}$ to the UAV within n-th time slot is given by

$$\aleph_i\left[\vec{q}_u(n)\right] = \begin{cases} \max\{\aleph', 30.9 + [22.25 - 0.5\lg(A)] \\ \qquad \lg(d_{iu}) + 20\lg(f_c)\}, & \text{LoS} \\ \max\{\aleph_i^{\text{LoS}}\left[\vec{q}_u(n)\right], 32.4 + \\ [43.2 - 7.6\lg(A)]\lg(d_{iu}) + 20\lg(f_c)\}, & \text{NLoS} \end{cases}, \tag{1}$$

where $\aleph' = 20\lg(d_{iu}) + 20\lg(f_c) + 32.45$ represents the free space pathloss, f_c in GHz indicates the carrier frequency and $d_{iu} = \sqrt{||\vec{q}_u(n) - \vec{q}_i||^2 + A^2}$ outputs the Euclidean distance between the UAV and i. From (1), it is straightforward to conclude that the availability of $\vec{q}_i$ is of essence for the UAV to estimate the corresponding A2G pathlosses. However, it is difficult (or, even impossible) for the UAV to gain perfect location estimations of malicious equipments on the ground. Therefore, this paper adopts a practical assumption on location availability, i.e., the UAV can only obtain the Warden's location information with uncertainty, while the exact locations of the legitimate nodes are known by the UAV *a prior*. Specifically, uncertain location estimation model is invoked to characterize the noised location information (e.g., Gaussian estimation noises) of the Warden, expressed as

$$\vec{q}_w = \hat{\vec{q}}_w + \vec{\varepsilon}, \tag{2}$$

where $\hat{\vec{q}}_w = (\hat{x}_w, \hat{y}_w)$ and $\vec{\varepsilon} = (\ddot{x}_w, \ddot{y}_w) \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$ represent the estimated Warden's location and the corresponding estimation error, respectively.

To practically trace the type of experienced A2G pathlosses, building distribution within $\mathbb{A}$ should be taken into consideration. Then, the type of large-scale pathloss of A2G channels for UAV at arbitrary location $\vec{q}_u(n)$, i.e., LoS or NLoS in (1), can be accurately determined via checking the potential blockages between the UAV and ground receiver i.

B. Transmission Rate from the UAV to the Legitimate Node

For the ϖ-th channel use over the n-th time slot, the received signals at the legitimate node can be given by

$$y_l(\varpi, n) = \sqrt{P_u 10^{\frac{-\aleph_l[\vec{q}_u(n)]}{10}}} x_u(\varpi) + \varkappa_l(\varpi), \tag{3}$$

where $x_u(\varpi) \sim \mathcal{CN}(0, 1)$ is the transmitted signal from the UAV to the legitimate node, P_u means the UAV's transmit power and $\varkappa_l(\varpi) \sim \mathcal{CN}(0, \sigma_l^2)$ denotes the AWGN. Note that $\varpi = \{1, 2, \ldots, c\}$ indicates the symbol index within a time slot and c measures the slot length. Then, the transmission rate in bps/Hz from the UAV to the legitimate node over the n-th time slot can be derived as

$$R_l(n) = \log_2(1 + \Gamma_l(n)), \tag{4}$$

where $\Gamma_l(n) = P_u 10^{\frac{-\aleph_l[\vec{q}_u(n)]}{10}} / \sigma_l^2$ represents the corresponding signal-to-noise-ratio (SNR).

C. The UAV's Estimated Detection Performance for Covert Transmissions

To detect potential transmissions initiated from the UAV, the Warden has to perform a binary hypothesis testing, where $\mathcal{H}_0$ represents the null hypothesis implicating that the UAV does not transmit and $\mathcal{H}_1$ denotes the alternative hypothesis indicating that the UAV transmits. Then, the false alarm probability (type I error) and the miss detection probability (type II error) are defined as $\mathbb{P}_{fa} = \Pr(\mathcal{D}_1|\mathcal{H}_0)$ and $\mathbb{P}_{md} = \Pr(\mathcal{D}_0|\mathcal{H}_1)$, respectively. Hereby, $\mathcal{D}_0$ indicates that the detection performed by the Warden is in favor of $\mathcal{H}_0$, while $\mathcal{D}_1$ means that it is in favor of $\mathcal{H}_1$. The Warden is assumed to posses no *a prior* knowledge of both hypotheses, and thus the overall detection error rate is defined as $\mathbb{P}_{de} = \mathbb{P}_{fa} + \mathbb{P}_{md}$. In covert communications, covert transmission is treated as achieved if any communication scheme stands with $\mathbb{P}_{de} \geq 1 - \varsigma$, where $\varsigma \ll 1$ is a positive value representing the covert requirement.

For the ϖ-th channel use over the n-th time slot, the received signals at the Warden can be derived as

$$y_w(\varpi, n) = \begin{cases} \sqrt{P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}} x_u(\varpi) + \varkappa_w(\varpi), & \mathcal{H}_1 \\ \varkappa_w(\varpi), & \mathcal{H}_0 \end{cases}, \tag{5}$$

where $\varkappa_w(\varpi) \sim \mathcal{CN}(0, \sigma_w^2)$ is the AWGN. Then, we have

$$\sum_{\varpi=1}^{c} |y_w(\varpi, n)|^2 = \begin{cases} \left(P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + \sigma_w^2\right) \mathcal{X}_{2c}^2, & \mathcal{H}_1 \\ \sigma_w^2 \mathcal{X}_{2c}^2, & \mathcal{H}_0 \end{cases}, \tag{6}$$

where $\mathcal{X}_{2c}^2$ represents chi-squared random variable with $2c$ degrees of freedom. According to Neyman-Pearson criterion [8], likelihood ratio test (LRT) is the optimal method for the Warden to minimize its detection error probability. Furthermore, invoking stochastic ordering theory [9] and Neyman-Fisher factorization corollary [10], the LRT applied by the Warden is equivalent to a test on its average received power within each time slot [11]. Therefore, the binary hypothesis testing performed by the Warden can be given by

$$H_w(n) = \frac{1}{c} \sum_{\varpi=1}^{c} |y_w(\varpi, n)|^2 \underset{\mathcal{D}_0}{\overset{\mathcal{D}_1}{\gtrless}} \tau(n), \tag{7}$$

where $\tau(n)$ represents the detection threshold. This paper adopts infinite slot-length assumption, which means that each time slot accommodates $c \to \infty$ channel uses (symbols). According to Lebesgue's Dominated Convergence Theorem, $\lim_{c\to\infty} \mathcal{X}_{2c}^2/c = 1$ holds. Then, we have

$$H_w(n) = \begin{cases} P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + \sigma_w^2, & \mathcal{H}_1 \\ \sigma_w^2, & \mathcal{H}_0 \end{cases}. \tag{8}$$

In practice, it is impractical for transceivers to gain perfect information regarding their received AWGN. Therefore, uncertain AWGN model is adopted to characterize the dynamics of Warden's AWGN power. Specifically, the Warden only knows the distribution of its received AWGN's variance, given by

$$f_{\sigma_w^2}(x) = \begin{cases} \frac{1}{2\ln\left(10^{\frac{\iota}{10}}\right)x}, & x \in [10^{-\frac{\iota}{10}}\hat{\iota}, 10^{\frac{\iota}{10}}\hat{\iota}] \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where ι in dB measures the degree of noise uncertainty and $\hat{\iota}$ indicates nominal noise power. Based on (7), (8) and (9), false alarm and miss detection probabilities can be derived as

$$\mathbb{P}_{fa}(n) = \Pr\left(\sigma_w^2 \geq \tau(n)\right) = \begin{cases} 0, & \tau(n) > 10^{\frac{\iota}{10}}\hat{\iota} \\ 1, & \tau(n) < 10^{-\frac{\iota}{10}}\hat{\iota} \\ \frac{1}{2\ln\left(10^{\frac{\iota}{10}}\right)} \ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{\tau(n)}\right), & \text{otherwise} \end{cases}, \tag{10}$$

$$\mathbb{P}_{md}(n) = \Pr\left(P_u(n) 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + \sigma_w^2 \leq \tau(n)\right) = \begin{cases} 1, & \tau(n) > 10^{\frac{\iota}{10}}\hat{\iota} + P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \\ 0, & \tau(n) < 10^{-\frac{\iota}{10}}\hat{\iota} + P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \\ \frac{\ln\left(\left(\tau(n) - P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}\right) 10^{\frac{\iota}{10}}\hat{\iota}^{-1}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, & \text{otherwise} \end{cases}, \tag{11}$$

respectively. In contrast to the uncertain Warden's location model on the perspective of the UAV as expressed in (2), this paper considers the worst case of covert communications, i.e., the Warden is able to obtain the perfect location information of the UAV. As such, $\vec{q}_u(n)$ and $\vec{q}_w$ in (11) are treated as accurately-known parameters for the Warden. Invoking (10) and (11), it is straightforward to calculate the Warden's overall detection error rate as $\mathbb{P}_{de}(n) = \mathbb{P}_{fa}(n) + \mathbb{P}_{md}(n)$. Then, the Warden tries to derive the optimal detection threshold $\tau^*(n)$ to perform covert detections, aiming at minimizing $\mathbb{P}_{de}(n)$.

***Proposition** 1:* The optimal detection threshold $\tau^*(n)$ adopted by the Warden within the n-th time slot is given by

$$\tau^*(n) = \begin{cases} 10^{\frac{\iota}{10}}\hat{\iota}, & P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \geq 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota} \\ P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}, & \text{otherwise} \end{cases}. \tag{12}$$

Then, the corresponding minimum overall detection error rate can be calculated as

$$\mathbb{P}_{de}^*(n) = \begin{cases} 0, & P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \geq 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota} \\ \frac{1}{2\ln\left(10^{\frac{\iota}{10}}\right)} \ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}}\right), & \text{otherwise} \end{cases}. \tag{13}$$

Proof: If $P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} < 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota}$ holds, the Warden's overall detection error rate can be expressed as

$$\mathbb{P}_{de}(n) = \begin{cases} \frac{\ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{\tau(n)}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, \tau(n) \in \left[10^{-\frac{\iota}{10}}\hat{\iota}, P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}\right) \\ \frac{\ln\left(10^{\frac{\iota}{5}} - \frac{P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} 10^{\frac{\iota}{5}}}{\tau(n)}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, \tau(n) \in \left[P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}, 10^{\frac{\iota}{10}}\hat{\iota}\right] \\ \frac{\ln\left(\left(\tau(n) - P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}\right)\frac{10^{\frac{\iota}{10}}}{\hat{\iota}}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, \tau(n) \in \left(10^{\frac{\iota}{10}}\hat{\iota}, P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{\frac{\iota}{10}}\hat{\iota}\right] \\ 1, \quad \tau(n) < 10^{-\frac{\iota}{10}}\hat{\iota} \text{ or } \tau(n) > P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{\frac{\iota}{10}}\hat{\iota} \end{cases}. \tag{14}$$

Based on (14), the Warden will never allow $\tau(n) < 10^{-\frac{\iota}{10}}\hat{\iota}$ or $\tau(n) > P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{\frac{\iota}{10}}\hat{\iota}$ because $\mathbb{P}_{de}(n) = 1$ stands, which means that the Warden gets the worst detection performance. Furthermore, it is easy to verify that $\mathbb{P}_{de}(n)$ is a monotonically-decreasing function with respect to (w.r.t.) $\tau(n)$ in the interval $\left[10^{-\frac{\iota}{10}}\hat{\iota}, P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}\right)$, while $\mathbb{P}_{de}(n)$ is a monotonically-increasing function w.r.t. $\tau(n)$ in intervals $\left[P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}, 10^{\frac{\iota}{10}}\hat{\iota}\right]$ and $\left(10^{\frac{\iota}{10}}\hat{\iota}, P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{\frac{\iota}{10}}\hat{\iota}\right]$. Then, we can conclude that the optimal detection threshold should be $\tau^*(n) = P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota}$, given the fact that (14) is continuous w.r.t. $\tau(n)$. Via invoking $\tau^*(n)$ into (14), the corresponding minimum overall detection error rate can be derived as $\frac{1}{2\ln\left(10^{\frac{\iota}{10}}\right)} \ln\left(10^{\frac{\iota}{10}}\hat{\iota}/(P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} + 10^{-\frac{\iota}{10}}\hat{\iota})\right)$.

Similarly, in the case of $P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \geq 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota}$, we have

$$\mathbb{P}_{de}(n)=\begin{cases} \frac{\ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{\tau(n)}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, & \tau(n)\in\left[10^{-\frac{\iota}{10}}\hat{\iota}, 10^{\frac{\iota}{10}}\hat{\iota}\right) \\ 0, & \tau(n)\in\left[10^{\frac{\iota}{10}}\hat{\iota}, P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}+10^{-\frac{\iota}{10}}\hat{\iota}\right] \\ 1, & \tau(n)<10^{-\frac{\iota}{10}}\hat{\iota} \text{ or } \tau(n)>P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}+10^{\frac{\iota}{10}}\hat{\iota} \\ \frac{\ln\left(\left(\tau(n)-P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}\right)\frac{10^{\frac{\iota}{10}}}{\hat{\iota}}\right)}{2\ln\left(10^{\frac{\iota}{10}}\right)}, & \text{otherwise} \end{cases}. \tag{15}$$

Based on (15), it is straightforward to get that the optimal detection threshold should be $\tau^*(n) = 10^{\frac{\iota}{10}}\hat{\iota}$ and the corresponding minimum overall detection error rate is 0. ■

***Proposition** 1* enables the Warden minimize its overall detection error rate within arbitrary time slot, via providing the optimal detection threshold. Then, the UAV needs to estimate the Warden's overall detection error rate within each time slot, based on which it can make adaptive decisions to counter detections from the Warden, e.g., trajectory design. With the considered uncertain location estimation model (2), the UAV can estimate the expected version of overall detection error rate suffered by the Warden.

***Proposition** 2:* On the perspective of UAV, the expected overall detection error rate under uncertain Warden's location estimation model within the n-th time slot can be derived as

$$\bar{\mathbb{P}}_{de}(n) = \begin{cases} 0, & P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \geq 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota} \\ \hat{\mathbb{P}}_{de}(n), & \text{otherwise} \end{cases}. \tag{16}$$

Proof: From ***Proposition** 1*, we know that the Warden can always achieve accurate detection performance without suffering from detection errors, i.e., $\mathbb{P}^*_{de}(n) = 0$, via letting $\tau^*(n) = 10^{\frac{\iota}{10}}\hat{\iota}$ in the case of $P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} \geq 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota}$. Therefore, in this case, the expected overall detection error rate should be $\bar{\mathbb{P}}_{de}(n) = 0$.

In the case of $P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}} < 10^{\frac{\iota}{10}}\hat{\iota} - 10^{-\frac{\iota}{10}}\hat{\iota}$, the expected overall detection error rate can be given by

$$\bar{\mathbb{P}}_{de}(n) = \frac{\mathbb{E}_{\{\vec{q}_w=\hat{\vec{q}}_w+\vec{\varepsilon}\}}\left[\ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{P_u 10^{\frac{-\aleph_w[\vec{q}_u(n)]}{10}}+10^{-\frac{\iota}{10}}\hat{\iota}}\right)\right]}{2\ln\left(10^{\frac{\iota}{10}}\right)}. \tag{17}$$

However, it is mathematically intractable to derive the closed-form expression for (17) because the pathloss factor $\aleph_w[\vec{q}_u(n)]$ is located in the denominator of logarithmic function and the uncertain location estimation model is involved in the complex A2G pathloss model (1). Alternatively, numerical evaluation on location uncertainty is applied to approximate (17), expressed as

$$\bar{\mathbb{P}}_{de}(n) \simeq \hat{\mathbb{P}}_{de}(n) = \frac{\sum_{\check{c}=1}^{\hat{c}} \ln\left(\frac{10^{\frac{\iota}{10}}\hat{\iota}}{P_u 10^{\frac{-\aleph_w[\vec{q}_u(n),\check{c}]}{10}}+10^{-\frac{\iota}{10}}\hat{\iota}}\right)}{2\hat{c}\ln\left(10^{\frac{\iota}{10}}\right)}, \tag{18}$$

where $\aleph_w\left[\vec{q}_u(n), \check{c}\right]$ indicates one realization of location error $\vec{\varepsilon}$ and $\hat{c}$ measures the amount of numerical evaluations. Note that $\hat{c} \gg 1$ should be delicately chosen to hit a good trade-off between accuracy and complexity of numerical evaluations. ■

D. Problem Formulation

This paper aims to maximize transmission throughput from the UAV to legitimate nodes alongside the UAV's trajectory from the initial UAV location to the destination, via designing UAV's marching direction within each time slot, subject to covert, velocity and mobility constraints. Then, the corresponding optimization problem can be formulated as

$$(\text{P1}): \max_{\vec{v}_u(n)} \sum_{n=1}^{N}\sum_{l\in\mathbb{L}} R_l(n), \tag{19a}$$

$$\text{s.t.} \bar{\mathbb{P}}_{de}(n) \geq 1-\varsigma, \tag{19b}$$

$$\vec{q}(n+1) = \vec{q}(n) + \Delta_t \vec{v}_u(n), \|\vec{v}_u(n)\| = V, \tag{19c}$$

$$\vec{q}_{\text{lo}} \preceq \vec{q}_u(n) \preceq \vec{q}_{\text{up}}, \vec{q}_u(0) = \vec{q}_u(I), \vec{q}_u(N) = \vec{q}_u(D). \tag{19d}$$

The constraint (19b) makes sure that a certain level of covert transmission can be achieved, while the constraints (19c) and (19d) indicate the velocity and mobility regulations. For simplicity, the factor Δ_t in (19a) is omitted. Intuitively, the UAV needs to adopt proper flying direction within arbitrary time slot during its flight, for not only avoiding "covert holes" where the covert requirement cannot be satisfied but also directing itself to visit possible locations where greater transmission rate is achievable. Hence, it is non-trivial for the UAV to carefully design its velocity for arbitrary time slot.

Because of the building-distribution-based pathloss and uncertain location estimation models, it is challenging to tackle (P1) via standard optimization techniques (e.g., convex optimization), if not impossible. Alternatively, this paper aims to design a DRL-aided approach to efficiently solve the formulated optimization goal.

III. The Proposed Multi-Step TD3-PER Solution

To design the DRL-aided solution, the first step is to map the considered problem into a MDP, stated as follows.

- $\mathcal{S}$: The state space is continuous, which contains possible UAV locations within $\mathbb{A}$, subject to $\vec{q}_{\text{lo}} \preceq \vec{q}_u \preceq \vec{q}_{\text{up}}$.
- $\mathcal{A}$: The action space is continuous, which involves possible velocity options $\vec{v}_u \in \mathbb{R}^{1*2}$, subject to $\|\vec{v}_u\| = V$.
- $\mathcal{T}$: State transition is deterministic, governed by (19c).
- r: According to the optimization objective (19a), it is direct to design the reward as $r(\vec{q}_u) = \sum_{l\in\mathbb{L}} R_l(\vec{q}_u) - 1$, where the penality -1 is used to encourage the UAV to reach the destination with fewer steps.

A. Twin Delayed Deep Deterministic Policy Gradient

The optimization problem (P1) involves velocity management from continuous domain, which leads to infinite action space. Unfortunately, deep Q network (DQN)-based algorithms, e.g., D3QN, are not suitable to tackle problems containing continuous actions, because it is extremely inefficient to find the maximum Q value over continuous action space [12]. To deal with this obstacle, policy gradient approach and

actor-critic architecture are invoked. In actor-critic algorithms, the actor is a policy network which takes states as inputs and reproduces a specific action, instead of a probability distribution over possible action options. Besides, the critic is a state-action value network, in which action and state are treated as the inputs and state-action values are the corresponding outputs. The actor network in deep deterministic policy gradient (DDPG) can eliminate the need of locating the action maximizing the state-action function given the next state, which can robustly solve problems with continuous action space. However, DDPG may suffer from one common and fundamental obstacle, i.e., overestimation bias induced by unavoidable function approximation errors, which is then propagated through the Bellman equation and can result in broken policy. To relieve the aforementioned side-effect, TD3 algorithm introduces three techniques to further improve the performance of DDPG, shown as follows.

1) Target Policy Smoothing: To compute the target of critic network's loss function, unlike DDPG approach, TD3 adds additional noise to the action chosen by the target actor network for the next state, shown as

$$\mu\left(s_{t+1}|\boldsymbol{\theta}_{\mu}^{-}\right) \leftarrow \mu\left(s_{t+1}|\boldsymbol{\theta}_{\mu}^{-}\right) + \vartheta^{-}, \tag{20}$$

where $\boldsymbol{\theta}_{\mu}^{-}$ denotes the parameter vector of the target actor network and $\vartheta^{-} \sim \mathcal{N}(0, \sigma_{\mu^{-}}^{2})$ indicates the added additional noise. Note that target policy smoothing technique serves as a regularizer for TD3 algorithm, which is designed to smooth the estimated Q values over similar actions and thus can help address the overfit issue caused by some actions with sharp-peak estimations of Q value.

2) Clipped Double Q Learning: In contrast to DDPG approach where one single critic network is applied to estimate the Q function, TD3 maintains two critic networks (i.e., the twin) and utilize the critic network with smaller estimated Q value to form the target of loss function. Specifically, both critic networks of TD3 algorithm are updated via stochastic gradient decent approach to minimize their loss functions with the same target, given by

$$\mathcal{L}(\boldsymbol{\theta}_{C_j}) = \Big[r_t + \gamma \min_{j=1,2} Q(s_{t+1}, \mu\left(s_{t+1}|\boldsymbol{\theta}_{\mu}^{-}\right) + \vartheta^{-}|\boldsymbol{\theta}_{C_j}^{-}) - Q(s_t, a_t|\boldsymbol{\theta}_{C_j}) \Big]^2, \tag{21}$$

where $j \in \{1, 2\}$ is introduced to distinguish the twin critics, $\boldsymbol{\theta}_{C_j}$ and $\boldsymbol{\theta}_{C_j}^{-}$ mean the parameter vectors of critic networks and target critic networks, respectively. Note that the clipped double Q learning technique can help relieve the overestimation issue via adopting the smaller estimated Q value of the twin critic networks to realize critic network updates.

3) Delayed Policy Updates: Similar to DDPG, the actor network of TD3 algorithm is updated to maximize the expected return via gradient ascent approach, where the expected return's gradient is calculated via the chain rule, shown as

$$\nabla_{\boldsymbol{\theta}_{\mu}} J\left(\boldsymbol{\theta}\right) = \mathbb{E}_{s_t}\{\nabla_a Q\left(s_t, a|\boldsymbol{\theta}_{C_1}\right) \nabla_{\boldsymbol{\theta}_{\mu}} \mu\left(s_t|\boldsymbol{\theta}_{\mu}\right)\}, \tag{22}$$

in which $\boldsymbol{\theta}_{\mu}$ indicates the parameter vector of actor network. However, in TD3, the actor network, target actor network and the twin target critic networks are updated less frequently than the twin critic networks, which can help damp the volatility issue in policy gradient algorithms.

B. The Proposed Multi-Step TD3-PER Solution

To efficiently tackle the considered trajectory design problem with continuous action space, this paper proposes a multi-step TD3-PER approach, via leveraging multi-step learning and PER techniques to further improve the training performance of TD3-based algorithm. Different from TD3 where randomly sampled one-step transitions are used to train the actor and critic networks, the proposed multi-step TD3-PER solution takes prioritized multi-step transitions to help the agent achieve a better training performance, of which the pseudo-code is summarized in **Algorithm 1** and the high-level workflow is depicted in Fig. 1. Note that the non-uniform priority is calculated w.r.t. transition's temporal difference error and then applied to weight the twin critics' updates in an importance sampling (IS) manner.

Algorithm 1: The Proposed Multi-Step TD3-PER Solution

Initialization: Initialize the twin critic networks $Q(s, a|\boldsymbol{\theta}_{C_1})$, $Q(s, a|\boldsymbol{\theta}_{C_2})$ and the actor network $\mu(s|\boldsymbol{\theta}_{\mu})$, then update their target networks via $\boldsymbol{\theta}_{C_1}^{-} \leftarrow \boldsymbol{\theta}_{C_1}$, $\boldsymbol{\theta}_{C_2}^{-} \leftarrow \boldsymbol{\theta}_{C_2}$ and $\boldsymbol{\theta}_{\mu}^{-} \leftarrow \boldsymbol{\theta}_{\mu}$. Initialize the PER buffer R with capacity C. Set the size of mini-batch as N_{mb}. Set the step length of multi-step learning as N_{ms}. Set the policy update delay as N_{pud}. Set target network update factor as τ;
for $te = [1, te_{\max}]$ **do**
  Set time step $n = 0$. Rest the UAV to its initial location as $\vec{q}_u(n) \in \mathcal{S}$. Initialize a sliding buffer $\hat{\mathrm{R}}$ with capacity N_{ms};
  repeat
    Select and execute action $a_n = \mu(\vec{q}_u(n)|\boldsymbol{\theta}_{\mu}) + \vartheta$, then observe the next state $\vec{q}_u(n+1)$ and the immediate reward $r_n = r[\vec{q}_u(n+1)]$;
    Get and record 1-step transition $\{\vec{q}_u(n), a_n, r_n, \vec{q}_u(n+1)\}$ into $\hat{\mathrm{R}}$;
    if $n \geq N_{ms}$ **then**
      Generate the N_{ms}-step reward $r_{n-N_{ms}} = \sum_{n_{ms}=0}^{N_{ms}-1} \gamma^{n_{ms}} r_{n-N_{ms}+n_{ms}}$ from $\hat{\mathrm{R}}$ and record N_{ms}-step experience $\{\vec{q}_u(n-N_{ms}), a_{n-N_{ms}}, r_{n-N_{ms}:n}, \vec{q}_u(n)\}$ into R;
    end
    Sample a mini-batch of N_{mb} N_{ms}-step transitions from R with priorities $p_{n_{mb}}$;
    $y_{n_{mb}} = r_{n_{mb}:n_{mb}+N_{ms}} + \gamma^{N_{ms}} \min_{i=1,2} Q(\vec{q}_u(n_{mb}+N_{ms}), \mu(\vec{q}_u(n_{mb}+N_{ms})|\boldsymbol{\theta}_{\mu}^{-}) + \vartheta^{-}|\boldsymbol{\theta}_{C_j}^{-})$;
    Compute the mean squared losses of the twin critics as $\mathcal{L}^{sq}(\boldsymbol{\theta}_{C_j}) = \frac{1}{N_{mb}} \sum_{n_{mb}} \frac{1}{CP_{n_{mb}}} (y_{n_{mb}} - Q(\vec{q}_u(n_{mb}), a_{n_{mb}}|\boldsymbol{\theta}_{C_j}))^2$;
    Update the twin critic networks via gradient decent aiming to minimize $\mathcal{L}^{sq}(\boldsymbol{\theta}_{C_j})$;
    Every N_{pud} times the twin critic netowrks are trained, update the actor network via gradient ascent to maximize $J(\boldsymbol{\theta}) = \frac{1}{N_{mb}} \sum_{n_{mb}} Q(s_{n_{mb}}, \mu(s_{n_{mb}}|\boldsymbol{\theta}_{\mu})|\boldsymbol{\theta}_{C_1})$, then update the target actor and target twin critics in a soft copy fashion as $\boldsymbol{\theta}^{-} \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}^{-}$;
    Let $n += 1$;
  until $\vec{q}_u(n) = \vec{q}_u(T)$ || $\vec{q}_u(n) \notin \mathcal{S}$ || $n = N_{\max}$;
end

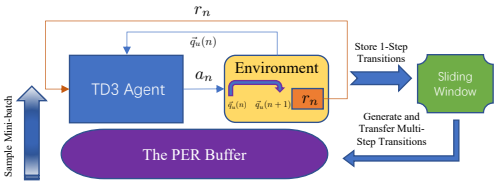


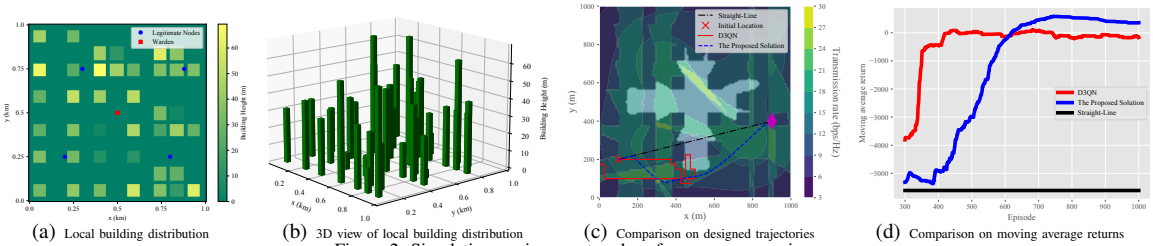
Figure 1: High-level workflow of the proposed multi-step TD3-PER solution

IV. Numerical Results

As shown in Fig. 2(a), the UAV's horizontal exploration airspace is set as $\mathbb{A} : [0, 1] \times [0, 1]$ km, in which the building distribution and the locations of $l \in \mathbb{L}$ and w are specified. Note that the considered building distribution is generated according to the statistical model recommended by the International Telecommunication Union (ITU) [13], of which the 3-dimensional (3D) view is depicted in Fig. 2(b). Both the actor and the twin critics involve 3 hidden layers with 512, 256, 128 neurons and *relu* activations, while the activation functions for the output layers of the actor

Parameters	Values	Parameters	Values	Parameters	Values
Side length of $\mathbb{A}$ D	1 km	Duration of time slot Δ_t	0.5 s	Speed of the UAV V	50 m/s
Carrier frequency f_c	2 GHz	Flying altitude of UAV A	100 m	Location estimation error variance σ_e^2	0.025
AWGN variance σ_1^2	-90 dBm	UAV's transmit power P_u	30 dBm	Covert requirement ς	0.001
Noise uncertainty degree ι	3dB	Nominal noise power $\hat{\iota}$	10^{-6}	Amount of numerical evaluation $\hat{c}$	1000
ITU building distribution parameter $\hat{\alpha}$	0.2	ITU building distribution parameter $\hat{\beta}$	40	ITU building distribution parameter $\hat{\gamma}$	25
Amount of buildings $\hat{\beta}D^2$	40	Expected size of each building $\hat{\alpha}/\hat{\beta}$	0.005 km^2	Maximum height of buildings	70 m
Replay buffer capacity C	10^6	Mini-batch size N_{mb}	32	Multi-step learning length N_{ms}	6
Policy update delay N_{pud}	10	Target network update factor τ	10^{-5}	Actor noise power variance σ_μ^2	1
Target Actor noise power variance $\sigma_{\mu^-}^2$	1	UAV exploration step threshold $N_{\max}$	150	Reaching destination bonus	4000
Hitting boundary penality	-10000	Visiting covert hole penality	-1000	Learning rates for actor/critic	$10^{-4}/10^{-3}$

Table I: Simulation Parameter Settings



(a) Local building distribution (b) 3D view of local building distribution (c) Comparison on designed trajectories (d) Comparison on moving average returns

Figure 2: Simulation environment and performance comparison

and the twin critics are *tanh* and *linear*, respectively. Unless otherwise mentioned, the parameter settings for simulation are in line to Table I. One DRL-related baseline, i.e., D3QN and another non-learning baseline, i.e., straight-line, are considered to conduct the performance comparison. To enable the D3QN, we have to discretize the action space. In the simulation, the discrete action space for D3QN-based solution is chosen as $\mathcal{A} = \{[-1, 0], [0, 1], [1, 0], [0, -1]\}$, corresponding to move left, forward, right and backward, respectively.

Fig. 2(c) demonstrates the comparison on designed trajectories, where the purple diamond indicates the destination and the transparent white region means the covert hole. From this figure, it is clear that the proposed multi-step TD3-PER solution can direct the UAV to reach the destination, while the covert hole is always being avoided alongside the whole flight. Meanwhile, positions with higher transmission rate are kept being preferred to be visited, even the narrow tunnel (around coordinate (0.6, 0.1) km) between the covert hole and the region with relatively lower transmission rate coverage is recognized by the proposed multi-step TD3-PER solution. Although the D3QN baseline tries to navigate the UAV to avoid the covert hole, it fails to direct the UAV to hit the destination and the UAV ends up crashing into the boundary, which showcases the inflexibility caused by the compromise from limited action options. Unsurprisingly, the straight-line approach can always lead the UAV to reach the destination, but it never tries to help the UAV bypass the covert hole or visit ideal regions with stronger transmission rate quality. For more insights, Fig. 2(d) delivers comparison on moving average return generated via a moving window with length of 300 episodes. From this figure, one can observe that the proposed solution converges to the highest score after training, while D3QN baseline converges to a worse position and straight-line approach gets the worst performance.

V. Conclusion

This paper investigated covertness-aware transmission throughput maximization problem, which was solved via UAV trajectory design from the continuous velocity domain. After mapping the optimization problem into a MDP, a multi-step TD3-PER solution was proposed to help the UAV find the proper velocity within each time slot, directing the UAV from its initial location to the destination. The effectiveness of the proposed multi-step TD3-PER solution was validated by the simulation results, while the corresponding advantages were showcased via performance comparison against D3QN and straight-line baselines.

References

[1] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4205–4220, 2021.

[2] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on AWGN channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1921–1930, 2013.

[3] D. Goeckel, B. Bash, S. Guha, and D. Towsley, "Covert communications when the warden does not know the background noise power," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 236–239, 2015.

[4] J. Hu, S. Yan, X. Zhou, F. Shu, J. Li, and J. Wang, "Covert communication achieved by a greedy relay in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4766–4779, 2018.

[5] J. Wang, W. Tang, Q. Zhu, X. Li, H. Rao, and S. Li, "Covert communication with the help of relay and channel uncertainty," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 317–320, 2018.

[6] X. Zhou, S. Yan, J. Hu, J. Sun, J. Li, and F. Shu, "Joint optimization of a UAV's trajectory and transmit power for covert communications," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4276–4290, 2019.

[7] 3GPP TR 36.777, "Enhanced LTE support for aerial vehicles," Dec. 2017.

[8] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[9] M. Shaked and J. G. Shanthikumar, *Stochastic orders and their applications*. New York: Academic Press, 1994.

[10] M. H. DeGroot and M. J. Schervish, *Probability and statistics*. Pearson Education, 2012.

[11] Y. Li, R. Zhao, Y. Deng, F. Shu, Z. Nie, and A. H. Aghvami, "Harvest-and-opportunistically-relay: Analyses on transmission outage and covertness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7779–7795, 2020.

[12] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. of ICLL*, 2015.

[13] P. Series, "Propagation data and prediction methods required for the design of terrestrial broadband radio access systems operating in a frequency range from 3 to 60 GHz," *Recommendation ITU-R*, pp. 1410–1415, 2013.