Three-Dimensional Trajectory Design for Multi-User MISO UAV Communications: A Deep Reinforcement Learning Approach

Yang Wang[1], Zhen Gao[1]

[1]Beijing Institute of Technology

July 27, 2021

Outline

Outline

Outline

1. **Introduction**
   - **Introduction to UAV Wireless Communication**
   - Introduction to Deep Reinforcement Learning
2. System Model
3. Proposed Solution
   - Problem Formulation
   - MDP Formulation
   - Deep Reinforcement Learning-DDPG
4. Simulation Results
5. References

Features

Several unique features of unmanned aerial vehicle (UAV) wireless communication.

- **Low cost, flexible deployment, fast response**
  They are especially suitable for unexpected or limited-duration missions.
- **Dynamic 3D placement and movement**
  The maneuverability of UAVs offers new opportunities for performance enhancement.
- **Short-distance LoS link**
  Line-of-sight (LoS) air-to-ground communication links can be established in most scenarios.

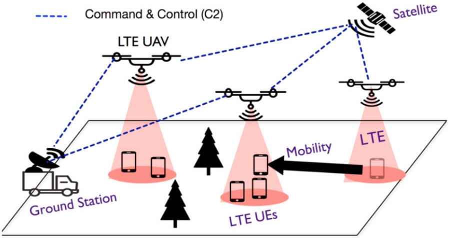


Fig. 1. An implementation of UAV communication

Functions

Typical functions.

- **UAV-aided ubiquitous coverage**
  Provide seamless wireless coverage within the serving area (discussed in this paper).

- **UAV-aided mobile relaying**
  Provide wireless connectivity between two or more distant users or user groups.

- **UAV-based IoT data collection**
  Collect delay-tolerant data from distributed Internet-of-Things (IoT) nodes.

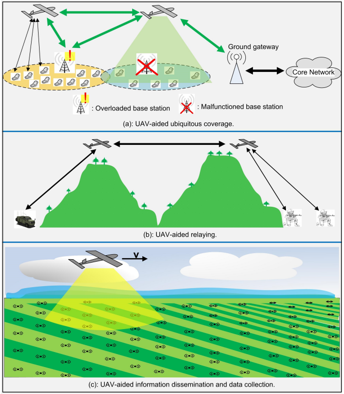


Fig. 2. Three typical functions[3]

In this paper, we consider a three-dimensional (3D) urban environment, where the UAV's 3D trajectory is designed to minimize data transmission completion time.

Outline

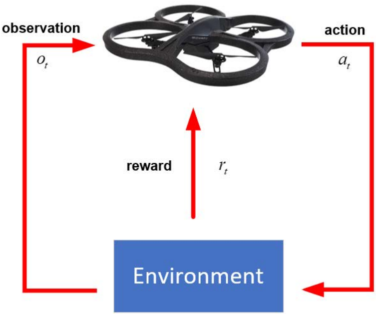


Fig. 3. The flow chart of reinforcement learning

- **Solution of sophisticated optimizations**
  DRL can obtain the solution of sophisticated optimizations.
- **Model-free learning**
  DRL allows agents to learn and build knowledge about the communication and networking environment.
- **Autonomous decision-making**
  With the DRL approaches, agents can make observation and obtain the best policy locally with minimum or without information exchange among each other.

Outline

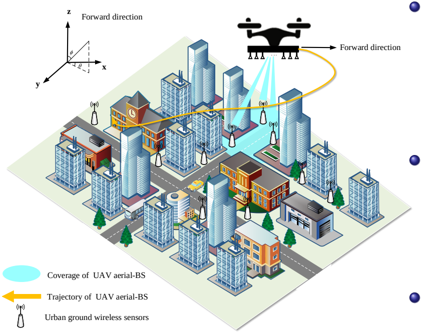


Fig. 4. Multi-antenna UAV-assisted MISO communication system.

- **Simulated 3D Urban Map**
  The location and height of the buildings are generated according to a statistical model recommended by the international telecommunication union (ITU)[14].
- **The Large-Scale Fading**

$$\mathrm{PL}_k(t) = \begin{cases} L_k^{\mathrm{FS}}(t) + \eta_{\mathrm{LoS}}, \\ L_k^{\mathrm{FS}}(t) + \eta_{\mathrm{NLoS}}, \end{cases} \tag{1}$$

- **The Ground-Air (G2A) Channel Gain**

$$\boldsymbol{h}_k(t) = 10^{-\mathrm{PL}_k(t)/20} \boldsymbol{g}_k(t), \tag{2}$$

Outline

- We consider the time domain is discretized into N time steps. During each time step, the UAV's moving strategy can be expressed as

$$x_{n+1} = x_n + m_n \sin(\phi_n) \cos(\theta_n), \tag{3}$$
$$y_{n+1} = y_n + m_n \sin(\phi_n) \sin(\theta_n), \tag{4}$$
$$z_{n+1} = z_n + m_n \cos(\phi_n), \tag{5}$$

- For the downlink data transmission service associated with the k-th GT, we define a binary variable to indicate whether the k-th GT can be served in the n-th time step, i.e.,

$$\tilde{b}_{k,n} = \begin{cases} 1, & \text{if } b_{k,n} = 1, \text{and } c_{k,n} = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $b_{k,n} \in \{0,1\}$ denotes whether the k-th GT can satisfy the SNR requirement by the UAV in the n-th time step and $c_{k,n} \in \{0,1\}$ denotes whether the k-th GT has been served by the UAV.

- We define the serving flag $c_{k,n}$ as

$$c_{k,n/0} = \min\left\{\sum_{i=0}^{n} \tilde{b}_{k,i}, 1\right\}, c_{k,0} = 0 \tag{7}$$

where if $c_{k,n} = 1$, the k-th GT has been served during the mission; otherwise, the k-th GT has not been served.

Problem Formulation

- We adopt the zero-forcing (ZF) precoder as it can obtain a near-optimal solution at a low complexity. Thus, the received signal at the active GTs in the n-th time step can be written by

$$\boldsymbol{y}_n = \boldsymbol{H}_n \boldsymbol{W}_n \boldsymbol{s}_n + \boldsymbol{q}, \tag{8}$$

- With the ZF precoding, the transmission SNR for the k-th GT can be expressed as

$$\rho_{k,n}^2 = \frac{P\left\|\boldsymbol{h}_{k,n}\boldsymbol{w}_{k,n}\right\|^2}{\sigma^2}, k \in \mathcal{K}_n. \tag{9}$$

- The transmission rate between the UAV and the k-th GT can be expressed as

$$R_{k,n} = W \log_2\left(1 + \rho_{k,n}^2\right), k \in \mathcal{K}_n, \tag{10}$$

- The hovering time of UAV in the n-th time step, which equals to the maximum transmission data duration from the $\mathcal{K}_n$ GTs, can be expressed as

$$\delta_{\mathrm{ht},n} = \max_{k \in \mathcal{K}_n}\left\{\frac{D_k}{R_{k,n}}\right\}, \tag{11}$$

Problem Formulation

- The completion criterion of the data transmission mission is that all GTs has been served, which can be expressed as

$$\sum_{k=1}^{K} c_{k,N} = K. \tag{12}$$

- The problem to minimize the mission completion time via trajectory optimization can be formulated as

$$\begin{aligned}
\underset{\{\upsilon_n,\phi_n,\theta_n\},N}{\text{minimize}} \quad & \textstyle\sum_{n=0}^{N} \delta_n \\
\text{s.t.} \quad & c_{k,n} = \min\left\{\textstyle\sum_{i=0}^{n} \tilde{b}_{k,i}, 1\right\}, \forall n, k, \\
& \textstyle\sum_{k=1}^{K} c_{k,N} = K, \\
& 0 \le \upsilon_n \le \upsilon_{\max}, \forall n, \\
& 0 \le \phi_n \le \pi, \forall n, \\
& 0 < \theta_n \le 2\pi, \forall n, \\
& 0 \le x_n \le D, \forall n, \\
& 0 \le y_n \le D, \forall n, \\
& z_{\min} \le z_n \le z_{\max}, \forall n,
\end{aligned} \tag{13}$$

The above optimization problem is a mixed-integer non-convex problem, which is known to be NP-hard.

Outline

MDP Formulation

We reformulate the optimization problem as a Markov decision process (MDP) structure so that deep reinforcement learning can be applied. Specifically, an MDP $\mathcal{M}$ can be defined by four elements, $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the state transition probability, and $\mathcal{R}$ is the reward in each time step.

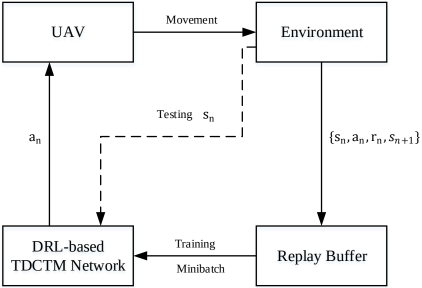


Fig. 5. DRL-based TDCTM

We define the following state, action, and reward for this problem.

- State s_n, $\forall\, n$:

$$s_n = [b_{1,n}, \cdots, b_{K,n}; c_{1,n}, \cdots, c_{K,n}; x_n, y_n, z_n; \zeta_n], \tag{14}$$

  $b_{k,n}$ and $c_{k,n}$: the coverage indicators reflecting the data transmission situation of the k-th GT in the n-th time step.
  $[x_n, y_n, z_n]$: the three-dimensional position of the UAV in a given region.
  ζ_n: the merged information between environment and UAV agent during the mission. Specifically, ζ_n can be expressed as

$$\zeta_n = \zeta_{n-1} + K_n \cdot \kappa_{\mathrm{cov}} - \kappa_{\mathrm{dis}} - P_{\mathrm{ob}}, \tag{15}$$

  where ζ_{n-1} is the remaining pheromone in the $(n-1)$-th time step, κ_{cov} is a positive constant that is used to express the captured pheromone per GT, κ_{dis} is a positive constant expressing the lost pheromone, and P_{ob} is a penalty when an action causes the boundary violation of the UAV.

- Action a_n, $\forall\, n$:
  The action is defined as $a_n = [\upsilon_n, \phi_n, \theta_n]$. Since all action variables take continuous values, the UAV's trajectory optimization is a continuous control problem.

MDP Formulation

- Reward r_n, $\forall\, n$:

$$r_n = \begin{cases} r_{\tanh}(\zeta_n) + N_{\mathrm{re}}, & \text{if } \sum_{k=1}^{K} c_{k,n} = K, \\ r_{\tanh}(\zeta_n), & \text{otherwise}, \end{cases} \tag{16}$$

where the former part can be expressed as

$$r_{\tanh}(\zeta_n) = \frac{2}{1 + \exp\left(-\zeta_n / \left(K \cdot \kappa_{\mathrm{cov}}\right)\right)} - 1, \tag{17}$$

which is a shaped reward function of the pheromone ζ_n. And $r_{\tanh}(\cdot)$ approximates $\tanh(\cdot)$ function, but the gradient is smoother than the latter. Besides, at the mission completion time step, the UAV would obtain a remaining time reward, i.e.,

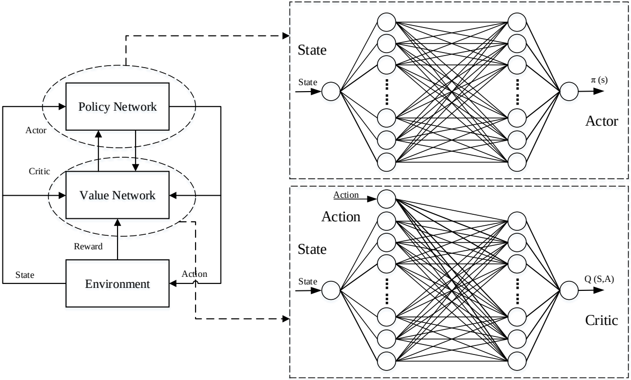
$$N_{\mathrm{re}} = N_{\max} - n, \tag{18}$$

which thus encourages the UAV to complete the data transmission mission as soon as possible.

Outline

Deep Deterministic Policy Gradient

As shown in Fig. 6, to cope with the continuous control problem with an infinite action space, the TDCTM network is conceived based on an actor critic algorithm, deep deterministic policy gradient (DDPG)[11].



Platform	
Python 3.6 + CUDA + RTX 2080TI	
Parameter	**Value**
Max Episode	8000
Max Episode Step	200
Learning Rate Actor	0.001
Learning Rate Critic	0.001
Memory Capacity	125,000
Batch Size	256
Soft-update Rate	0.0005
Reward Discount	0.99

Actor

Layer	Output Shape
Input	(None,State_num)
Dense_1	(None,200)
Dense_2	(None,200)
Output	(None,1)

Critic

Layer	Output Shape
Input_state	(None,State_num)
Dense_1	(None,200)
Input_action	(None,Action_num)
Dense_2	(None,200)
Output	(None,1)

Fig. 6. DRL-based TDCTM network architecture

DRL-based TDCTM algorithm

Fig. 7 shows the UAV can complete the data transmission service for all GTs.

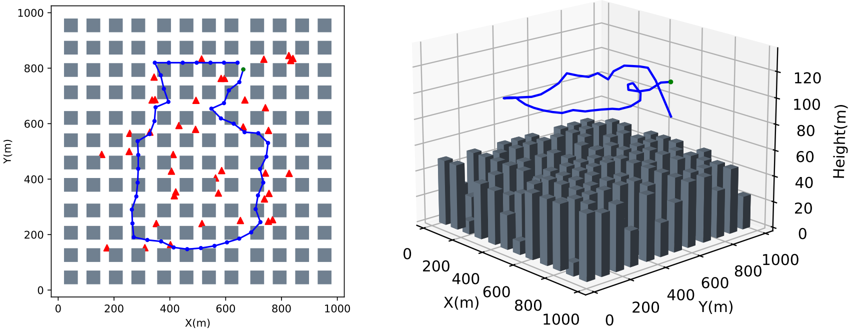


Fig. 7. UAV's 2D and 3D flight trajectories according to the proposed DRL-TDCTM algorithm, where 40 GTs are considered.

We compare the average mission completion time of different methods and the convergence performance versus different numbers of GTs in Fig. 8.

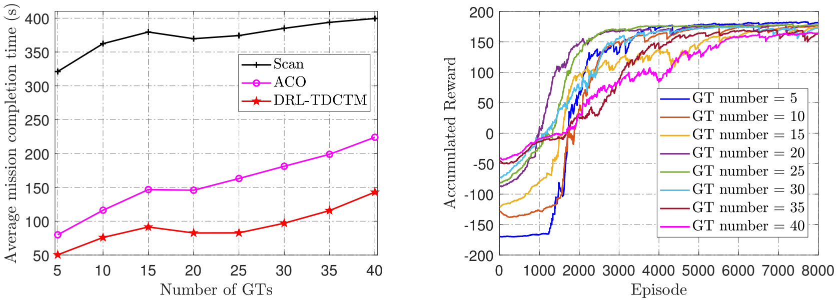


Fig. 8. The impact of the number of GTs on (a) average mission completion time and (b) convergence performance (i.e., accumulated reward versus episode).

Outline

References I

[1] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241-2263, Apr. 2019.

[2] A. E. A. A. Abdulla, Z. M. Fadlullah, H. Nishiyama, N. Kato, F. Ono, and R. Miura, "An optimal data transmission technique for improved utility in UAS-aided networks," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Toronto, Canada, May 2014, pp. 736-744.

[3] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36-42, May 2016.

[4] J. Zhang, Y. Zeng, and R. Zhang, "Multi-antenna UAV data harvesting: Joint trajectory and communication optimization," *J. Commun. Inf. Netw.*, vol. 5, no. 1, pp. 86-99, Mar. 2020.

[5] D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Multiuser MISO UAV communications in uncertain environments with no-fly zones: Robust trajectory and resource allocation design," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3153-3172, May 2020.

[6] Y. Zeng, X. Xu, and R. Zhang, "Trajectory design for completion time minimization in UAV-enabled multicasting," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2233-2246, Apr. 2018.

[7] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[8] C. H. Liu, Z. Chen, and Y. Zhan, "Energy-efficient distributed mobile crowd sensing: A deep learning approach," *IEEE J. Sel. Areas in Commun.*, vol. 37, no. 6, pp. 1262-1276, Jun. 2019.

[9] R. Ding, F. Gao, and X. S. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796-7809, Dec. 2020.

References II

[10] Y. Zeng and X. Xu, "Path design for cellular-connected UAV with reinforcement learning," *IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, 2019, pp. 1-6.

[11] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," *Comput. Sci.*, vol. 8, no. 6, 2015, Art. no. A187.

[12] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: Optimization by a colony of cooperating agents," *IEEE Trans. Sys.*, Man, and Cybernetics, Part B (Cybernetics), vol. 26, no. 1, pp. 29-41, Feb. 1996.

[13] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[14] ITU-R, Rec. P.1410-5, "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," Radiowave propagation, Feb. 2012.

[15] Q. Zhang, H. Sun, Z. Feng, H. Gao, and W. Li, "Data-aided Doppler frequency shift estimation and compensation for UAVs," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 400-415, Jan. 2020.

[16] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, 2020.

Thanks for your attention!
Q & A