This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications

Constrained Deep Reinforcement Learning for Energy Sustainable Multi-UAV based Random Access IoT Networks with NOMA

Sami Khairy*, Prasanna Balaprakash†, Lin X. Cai*, Yu Cheng*
*Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, USA
†Mathematics and Computer Science Division, Argonne National Laboratory, Illinois, USA
Email: skhairy@hawk.iit.edu, pbalapra@anl.gov, {lincai,cheng}@iit.edu

Abstract—**In this paper, we apply the Non-Orthogonal Multiple Access (NOMA) technique to improve the massive channel access of a wireless IoT network where solar-powered Unmanned Aerial Vehicles (UAVs) relay data from IoT devices to remote servers. Specifically, IoT devices contend for accessing the shared wireless channel using an adaptive p-persistent slotted Aloha protocol; and the solar-powered UAVs adopt Successive Interference Cancellation (SIC) to decode multiple received data from IoT devices to improve access efficiency. To enable an energy-sustainable capacity-optimal network, we study the joint problem of dynamic multi-UAV altitude control and multi-cell wireless channel access management of IoT devices as a stochastic control problem with multiple energy constraints. We first formulate this problem as a Constrained Markov Decision Process (CMDP), and propose an online model-free Constrained Deep Reinforcement Learning (CDRL) algorithm based on Lagrangian primal-dual policy optimization to solve the CMDP. Extensive simulations demonstrate that our proposed algorithm learns a cooperative policy in which the altitude of UAVs and channel access probability of IoT devices are dynamically controlled to attain the maximal long-term network capacity while ensuring energy sustainability of UAVs, outperforming baseline schemes. The proposed CDRL agent can be trained on a small network, yet the learned policy can efficiently manage networks with a massive number of IoT devices and varying initial states, which can amortize the cost of training the CDRL agent.**

Index Terms—**Constrained Deep Reinforcement Learning, UAV altitude control, Solar-Powered UAVs, Energy Sustainable IoT Networks, p-persistent slotted Aloha, Non-Orthogonal Multiple Access**

I. Introduction

While internet connectivity plays an increasing role in people's everyday life in densely populated areas, some rural areas and nature fields such as farms, deserts, oceans, and polar regions, typically lack expansive internet coverage. This is because network providers tend to deploy telecommunication infrastructure in areas where providing wireless service is economically profitable. Nevertheless, farmers, environmental agencies, research organizations, defense agencies, and utility companies among many others, have increasing demands for internet connectivity in such under-served areas, to support massive Internet of Things (IoT) based applications ranging from tracking animal health, agricultural growth, and marine life, to surveillance sensors for defense applications and nuclear waste site management, just to name a few. Provisioning wireless internet access for a massive number of IoT devices in under-served areas at cost effective rates is undoubtedly of great interest for governments, businesses, and end customers.

With the explosive growth of IoT based applications, new distributed channel access and coverage solutions should be conceived. Distributed random access based wireless technologies such as Wi-Fi, Zigbee, and Aloha-based Long Range Wide Access Networks (LoRaWAN), will inevitably play an important role in provisioning massive IoT access in 5G systems and beyond, due to their scalability and ease of implementation [1]–[4]. In fact, the recent decision by the Federal Communications Commission (FCC) to open the 6GHz band for unlicensed use [5], further promotes random access based unlicensed wireless networks and renders them as an integral part of the beyond 5G wireless network ecosystem. Non Orthogonal Multiple Access (NOMA), which can improve the access efficiency by exploiting Successive Interference Cancellation (SIC) to enable non-orthogonal data transmissions, is yet another promising solution to enable massive machine type communication (mMTC) in 5G networks and beyond. Recent works propose to apply power-domain NOMA in slotted-Aloha systems to support mMTC of IoT devices [6]–[8].

On the other hand, Unmanned Aerial Vehicle (UAV) based wireless relays have been recently proposed as a new on-demand coverage solution to facilitate fast and flexible deployment of communication infrastructure [9]–[17]. UAVs, especially those in the form of quadcopter drones, can be used as aerial base stations within a 5G network to provision wireless internet connectivity to IoT devices in remote areas, or enhance wireless system capacity, reliability, and energy efficiency, in urban areas, by relaying data to network servers through satellite back-haul links [18], [19]. Furthermore, UAVs equipped with wireless transceivers can dynamically adjust their location in real-time to counter environmental changes and improve system performance. UAVs, however, are battery operated which constrains the operation time of the network device. One promising direction to extend the operation time of the network is to devise solar-powered UAVs which can harvest solar energy from the sun [20], [21]. In this case, the system controller is also tasked with managing the energy evolution process of the UAV battery to ensure energy efficiency or sustainability.

In this work, we consider solar-powered multi-UAV based wireless IoT networks, where UAVs act as wireless Base

Stations (BS) for a massive number of IoT devices. IoT devices contend for access to the shared wireless channel using an adaptive p-persistent slotted Aloha MAC protocol to send data to the UAVs, which relay the received data to the internet backbone through wireless satellite back-haul links. UAVs on the other hand, are equipped with solar cells to replenish the on-board battery, and exploit power-domain SIC to decode multiple users' transmissions, thus improving the transmission efficiency. To enable an energy-sustainable and capacity-optimal massive IoT network, we study the joint problem of dynamic multi-UAV altitude control and NOMA-based multi-cell wireless channel access management of IoT devices. The objective of the stochastic control problem is to maximize the total network capacity of a massive number of IoT devices which is characterized by random uplink channel access, varying wireless channel conditions, and dynamic network topology, while satisfying multiple constraints to ensure energy sustainability of solar-powered UAVs. To the best of our knowledge, our work is the first work to study energy sustainability of a multi-UAV based wireless communication system in support of a massive number of IoT devices with NOMA and random channel access.

The main contributions of our work can be summarized as follows. First, we formulate the joint problem of multi-UAV altitude control and adaptive random channel access of massive IoT devices to attain the maximum capacity under energy sustainability constraints of UAVs over a prespecified operating horizon as a Constrained Markov Decision Process (CMDP). Second, to learn an optimal control policy for the wireless communication system, we design an online model-free Constrained Deep Reinforcement Learning (CDRL) algorithm based on Lagrangian primal-dual policy optimization to solve the CMDP. It is shown that the proposed CDRL agent learns a cooperative policy among UAVs which ensures their energy sustainability over the operating horizon, while maximizing the total network capacity under the probabilistic mutual interference of IoT devices. Third, to evaluate the effects of policy optimization in our proposed CDRL framework, we compare the learning performance with two other DRL agents which adopt different policy optimization algorithms, namely, Trust Region Policy Optimization (TRPO) [22], and Vanilla Policy Gradient (VPG) [23]. It is shown that our proposed CDRL agent which combines Lagrangian primal-dual optimization techniques with Proximal Policy Optimization (PPO) [24] outperforms other agents in terms of both the achieved rewards and constraint satisfaction. In addition, we compare the performance of the learned policy to three baseline policies learned by a 1) Deep RL (DRL) agent which is energy-unaware, a 2) DRL agent which adopts reward shaping to penalize energy dissipation, and 3) a random management policy. Compared with baseline policies, our extensive simulations demonstrate that our proposed CDRL agent learns a feasible adaptive policy which achieves a temporal average network capacity that is 81% higher than that of a feasible DRL agent with reward shaping, and only 6.7% lower than the upper bound achieved by the energy-unaware DRL agent. Last but not least, we demonstrate that the learned policy, which has been efficiently trained on a small network size, can effectively manage networks with a massive number of IoT devices and varying initial network states.

The remainder of this paper is organized as follows. A literature survey of related research work and a background of unconstrained and constrained MDPs is given in Section II. The system model is described in Section III. The problem formulation and the proposed CDRL algorithm is presented in Section IV, followed by the simulation setup and performance evaluation results in Section V. Finally our concluding remarks and future work are given in Section VI.

II. Background and Related Works

A. UAV based Wireless Networks

The deployment and resource allocation of UAV-based wireless networks has been studied in many works. In [9], a polynomial-time algorithm is proposed for successive UAV placement such that the number of UAVs required to provide wireless coverage for a group of ground terminals is minimized and each ground terminal is within the communication range of at least one UAV. The downlink coverage probability for UAVs as a function of the altitude and antenna gain is analyzed in [10]. Based on the circle packing theory, the 3D locations of the UAVs are determined to maximize the total coverage area while ensuring the covered areas of multiple UAVs do not overlap. The work of [11] studies the problem of multiple UAV deployment for on-demand coverage while maintaining connectivity among UAVs. In [12], a distributed coverage-maximizing algorithm for multi UAV deployment subject to the constraint that UAVs maintain communication is proposed for surveillance and monitoring applications.

3D trajectory design and resource allocation in UAV based wireless networks have also been studied in [13]–[17]. In [13], a mixed integer non-convex optimization problem is formulated to maximize the minimum downlink throughput of ground users by jointly optimizing multi-user communication scheduling, association, UAVs' 3D trajectory, and power control. An iterative algorithm based on block coordinate descent and successive convex optimization techniques is proposed to solve the formulated problem. [14] extends on [13] by considering heterogeneous UAVs so that each UAV can be individually controlled. Machine learning based approaches have also been recently considered for UAV 3D trajectory design. In [15], the flight trajectory of the UAV and scheduling of packets are jointly optimized to minimize the sum of Age-of-Information (sum-AoI) at the UAV. The problem is modeled as a finite-horizon Markov Decision Process (MDP) with finite state and action spaces, and a DRL algorithm is proposed to obtain the optimal policy. [16] devises a machine learning based approach to predict users' mobility information, which is considered in the trajectory design of multiple UAVs. A sense-and-send protocol is designed in [17] to coordinate multiple UAVs, and a multi-UAV Q-learning based algorithm is proposed for decentralized UAV trajectory design. Scheduling based NOMA systems with a UAV-based BS to serve terrestrial users are considered in [25], [26]. In a recent work, the performance of NOMA transmissions in a single-hop random access wireless network is investigated, and an iterative algorithm is proposed to find the optimal transmission probabilities of users to achieve the maximum throughput [27]. Furthermore, the control, motion planning, and aerodynamic power consumption of UAVs have been studied in [28]–[30].

It is worth to mention that all aforementioned works consider battery powered UAVs with limited energy storage capacity, which constrains the operating horizon. Solar-powered UAVs have great potential to extend the operation time by harvesting solar energy from the sun [20], [21]. [31] studies the optimal trajectory of solar-powered UAVs for maximizing the solar energy harvested. In their design, a higher altitude is preferable to maximize harvested energy. On the other hand, [32] studies the trade-off between solar energy harvesting and communication system performance of a single UAV based wireless network. It is shown that in order to maximize the system throughput, the solar-powered UAV climbs to a high altitude to harvest enough solar energy, and then descends to a lower altitude to improve the communication performance.

The work of [32] considers downlink wireless resource allocation in a centralized scheduling-based and interference-free wireless network with a single UAV. Deploying one solar-powered UAV may lead to a communication outage when the UAV ascends to high altitudes to replenish its on-board battery, and therefore, one UAV cannot satisfactorily serve wireless users and ensure its energy sustainability. On the other hand, scheduling-based networks usually suffer from the curse of dimensionality and do not scale well to massive IoT networks, as signaling overheads scale up with the network size. This has led to a growing interest in wireless networks with NOMA and distributed random access based Medium Access Control (MAC) protocols, such as Aloha-type MAC adopted in LoRaWAN networks [4]. It is very challenging to analyze the performance of distributed multi-cell random access networks with NOMA-enabled aerial base stations due to the combinatorial space of possible transmissions and interference that affect decoding events, and the time varying network topology. Machine learning provides a data driven approach for end-to-end system design, and can be therefore used to holistically study these challenging wireless systems and provide proper guidance for integration within 5G systems and beyond. In this work, we study solar-powered multi-UAV based massive IoT networks with random-access and NOMA, and propose solutions to ensure long-term energy sustainability of UAVs without causing wireless service interruption or degradation. Specifically, we propose a new framework for data-driven stochastic control under constraints, by combining DRL policy optimization methods with Lagrangian primal-dual optimization, and apply it to the problem of multi-UAV altitude control and wireless random channel access management. Our proposed solution demonstrates that by deploying multiple UAVs, it is possible to learn a cooperative policy in which multiple UAVs take turns to charge their battery and provision uninterrupted wireless service to IoT devices.

B. Constrained Deep Reinforcement Learning

One of the primary challenges faced in reinforcement learning is the design of a proper reward function which can efficiently guide the learning process. Many real world problems are multi-objective problems in which conflicting objectives should be optimized. A common approach to handling multi-objective problems in RL is to combine the objectives using a set of coefficients [33]. With this approach, there exist a set of optimal solutions for each set of coefficients, known as the Pareto optimal solutions [34]. In practice, finding the set of coefficients which leads to the desired solutions is not a trivial task. For many problems, it is more natural to specify a single objective and a set of constraints. The CMDP framework is the standard formulation for RL problems involving constraints [35]. Optimal policies for CMDPs can be obtained by solving an equivalent linear programming formulation [35], or via multi time-scale dynamic-programming based algorithms [36]–[40]. Such methods may not be applicable to large scale problems or problems with continuous state-action space. Leveraging recent advances in deep learning and policy search methods [22], some works devise multi-time scale algorithms for solving RL problems in presence of constraints [41]–[45]. Broadly speaking, these methods are either based on Lagrangian relaxation [41]–[44] or constrained policy optimization [45]. In Lagrangian relaxtion based method, primal and dual variables are updated at different time-scales using gradient ascent/descent. In these methods, constraint satisfaction is guaranteed at convergence. On the other hand, in [45] an algorithm is proposed where constraint satisfaction is enforced in every step throughout training. Our proposed algorithm is based on the PPO algorithm [24], and adopts the Lagrangian relaxation based approach to handle multiple constraints. However, we propose a new method to adapt the Lagrangian penalty multipliers during training, and show that our new approach improves the learning stability of the algorithm. In addition, our work is the first to demonstrate successful policy learning in environments with multiple constraints, and policy transferability among wireless networks of different scales.

C. Background

In this subsection, unconstrained and constrained MDPs are introduced. MDPs and CMDPs are are the classical formalization of sequential decision making and define the interaction between a learning agent and its environment in RL and constrained RL, respectively.

1) Markov Decision Process: An infinite horizon MDP with discounted-returns is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, \mathcal{R}, \zeta)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite sets of states and actions, respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the model's state-action-state transition probabilities, and $\mathcal{P}_0 : \mathcal{S} \to [0, 1]$ is the initial distribution over the states, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is the immediate reward function which guides the agent through the learning process, and ζ is a discount factor to bound the cumulative rewards and trade-off how far or short sighted the agent is in its decision making. Denote the transition probability from state $s_n = i$ at time step n to state $s_{n+1} = j$ if action $a_n = a$ is chosen by $P_{ij}(a) := P(s_{n+1} = j | s_n = i, a_n = a)$. The transition probability from state i to state j is therefore, $p_{ij} = P(s_{n+1} = j | s_n = i) = \sum_a P_{ij}(a)\pi(a|i)$, where $\pi(a|i)$

is a stochastic policy which maps states to actions. The state-value function of state i under policy π is the long-term expected discounted returns starting in state i and following policy π thereafter,

$$V_\pi(i) = \sum_{n=1}^{\infty} \sum_{j,a} \zeta^{n-1} P^\pi(s_n = j, a_n = a | s_0 = i)\mathcal{R}(j,a), \forall i \in \mathcal{S} \tag{1}$$

Denote the initial distribution over the states by the vector $\boldsymbol{\beta}$, where $\beta(i) = P(s_0 = i), \forall i \in \mathcal{S}$. The solution of an MDP is a Markov stationary policy π^* that maximizes the inner product $\langle \mathbf{V}_\pi, \boldsymbol{\beta} \rangle$,

$$\max_{\pi} \qquad \sum_{n=1}^{\infty} \sum_{j,a} \zeta^{n-1} P^\pi(s_n = j, a_n = a)\mathcal{R}(j,a) \tag{2}$$

There are several approaches to solve (2), including dynamic programming based methods such as value iteration and policy iteration [46], in addition to linear programming based methods [47]. When the model's dynamics, i.e., transition probabilities, are unknown, the Reinforcement Learning (RL) framework can be adopted to find the optimal policies. It is worth to mention that when the agent learns the optimal state-value function and/or the policy as parameterized Deep Neural Networks (DNNs), the agent is commonly referred to as a Deep RL (DRL) agent. There exists a significant body of works with state-of-the-art algorithms to solve the RL problem, which vary by design from value-based methods [48], to policy-based methods [22]–[24], and hybrid actor-critic type algorithms [49]–[53].

2) Constrained Markov Decision Process: In constrained MDPs (CMDPs), additional immediate cost functions $\mathcal{C}_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are augmented, such that a CMDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, \mathcal{R}, \mathcal{C}, \zeta)$ [35]. The state-value function is defined as in unconstrained MDPs (1). In addition, the infinite-horizon discounted-cost of a state i under policy π is defined as,

$$C_\pi^k(i) = \sum_{n=1}^{\infty} \sum_{j,a} \zeta^{n-1} P^\pi(s_n = j, a_n = a | s_o = i)\mathcal{C}_k(j,a), \forall i \in \mathcal{S}, \forall k. \tag{3}$$

The solution of a CMDP is a markov stationary policy π^* which maximizes $\langle \mathbf{V}_\pi, \boldsymbol{\beta} \rangle$ subject to the constraints $\langle \mathbf{C}_\pi^k, \boldsymbol{\beta} \rangle \leq E_k, \forall k$,

$$\max_{\pi} \qquad \sum_{n=1}^{\infty} \sum_{j,a} \zeta^{n-1} P^\pi(s_n = j, a_n = a)\mathcal{R}(j,a) \tag{4}$$

$$\sum_{n=1}^{\infty} \sum_{j,a} \zeta^{n-1} P^\pi(s_n = j, a_n = a)\mathcal{C}_k(j,a) \leq E_k, \ \forall k \tag{4a}$$

Solving for feasible and optimal policies in CMDPs is more challenging compared to unconstrained MDPs, and requires extra mathematical efforts. CMDPs can be solved by defining an appropriate occupation measure and constructing a linear program over this measure, or alternatively by using a Lagrangian relaxation technique in which the CMDP is converted into an equivalent unconstrained problem,

$$\begin{aligned} \max_{\pi} \min_{\boldsymbol{\eta} \geq 0} \ & \mathcal{L}(\pi, \boldsymbol{\eta}) \\ &= \max_{\pi} \min_{\boldsymbol{\eta} \geq 0} \ \langle \mathbf{V}_\pi, \boldsymbol{\beta} \rangle - \sum_k \eta_k \big(\langle \mathbf{C}_\pi^k, \boldsymbol{\beta} \rangle - E_k \big) \end{aligned} \tag{5}$$

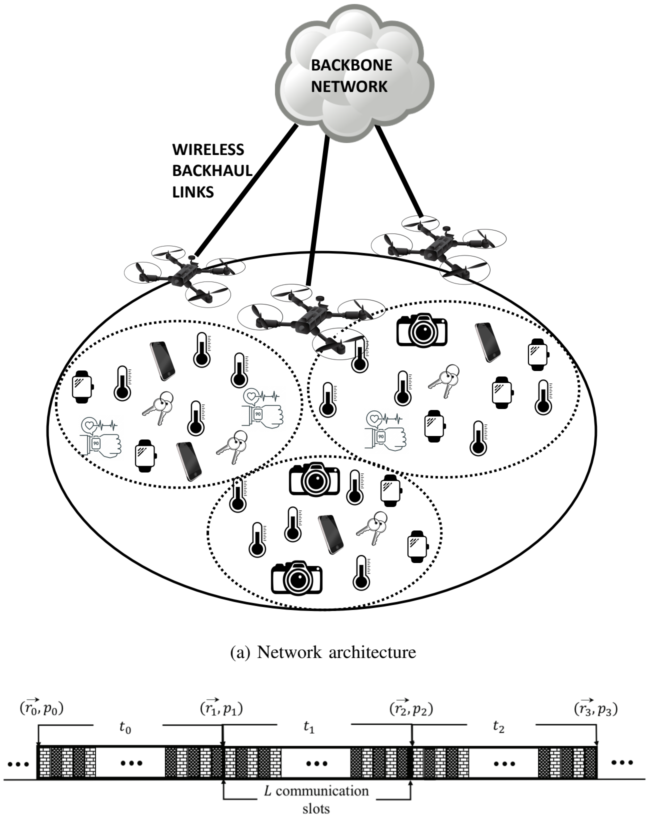
and invoking the minimax theorem,

$$\max_{\pi} \min_{\boldsymbol{\eta} \geq 0} \ \mathcal{L}(\pi, \boldsymbol{\eta}) = \min_{\boldsymbol{\eta} \geq 0} \max_{\pi} \ \mathcal{L}(\pi, \boldsymbol{\eta}) \tag{6}$$

The right hand side of (6) can be solved on two-time scales: on a faster time scale gradient-ascent is performed on state-values to find the optimal policy for a given set of Lagrangian variables, and on a slower time scale, gradient-descent is performed on the dual variables [35]. Past works explore this primal-dual optimization approach for CMDPs with known model dynamics and tabular-based RL methods with unknown model dynamics [36]–[40]. In the realm of deep RL where policies and value functions are parameterized neural networks, recent works which apply primal-dual optimization for generic benchmark problems are emerging [41]–[45]. None of these works, however, apply primal-dual optimization techniques in the wireless networking domain. Furthermore, practical wireless networking systems admit multiple constraints, which can be conflicting. This incurs extra difficulty for policy search and optimization, and may cause learning instability. Our proposed approach tackles these issues and demonstrates successful policy learning in wireless environments with multiple constraints.

III. System Model

Consider a multi-UAV based IoT network consisting of M UAVs and N IoT devices, where the UAVs collect data from a massive deployment of IoT devices, as shown in Figure 1(a). Let $\mathcal{M} = \{1, \cdots, M\}$ be the set of UAVs, and $\mathcal{N} = \{1, \cdots, N\}$ be the set of IoT devices. UAVs are connected via wireless back-haul links to a central controller, which controls the altitude of each UAV and manages the access parameters of wireless IoT devices. IoT devices are independently and uniformly distributed (i.u.d.) across a deployment area $\mathbb{A}$. Let the locations of IoT devices be $\{\hat{x}^i, \hat{y}^i\}_{i=1}^N$. Each IoT device is served by the closest UAV. Denote the subset of IoT devices which are associated with UAV m by $\mathcal{N}_m \subset \mathcal{N}$, $|\mathcal{N}_m| \leq N$, $\bigcup_{m=1}^M \mathcal{N}_m = \mathcal{N}$, $\mathcal{N}_i \cap \mathcal{N}_j = \phi, \forall i \neq j \in \mathcal{M}$. Time is slotted into fixed-length discrete time units indexed by n. For instance, the n-th time slot is $[t_n, t_{n+1})$, where $t_{n+1} - t_n = \Delta t$, $\forall n$. Each time slot n is further divided into L communication sub-slots of length $\frac{\Delta t}{L}$ each, as shown in 1(b). Denote the l-th communication sub-slot in slot n by t_n^l, $l = \{0, \cdots, L-1\}$. During these communication sub-slots, IoT devices contend for channel access based on an adaptive p-persistent slotted Aloha MAC protocol. In this protocol, an IoT device waits until the beginning of a communication sub-slot before attempting to access the channel with probability p, which is adapted every time slot by the central controller



(a) Network architecture

(b) System time model

Fig. 1. Multi-cell UAV based wireless IoT network

based on network dynamics[1]. IoT devices transmit uplink data to their associated UAV with a fixed transmission power of P_{TX} watts, and are traffic-saturated, i.e., there is always a data packet ready for transmission.

Denote the location of UAV m during time slot n by $r_m(t_n) = (x^m, y^m, z_n^m)$. In our system model, z_n^m is dynamically adjusted by the central controller, while the the planar location of the UAVs, $(x^m, y^m), \forall m \in \mathcal{M}$ are determined a priori by using Lloyd's K-means clustering algorithm to partition the distribution of IoT devices on the ground into $K = M$ Voronoi cells, such that the sum of squared distances between any IoT device and its nearest UAV is minimized [54] [2]. Let $u_n^{i,m} = \{0,1\}$ indicate whether IoT device i is associated with UAV m during time slot n. If UAV m located at $r_m(t_n)$ during the n-th time slot is the closest to IoT device i, $u_n^{i,m} = 1$, and $u_n^{i,l} = 0, \forall l \neq m$. The power of the signal transmitted by a wireless IoT device i to a UAV m is subject to independent Rayleigh channel fading, $h_{i,m}(t_n)$[3], and a distance-dependent free-space path-loss $d_{i,m}^{-\alpha}(t_n)$, where α is the path-loss exponent, and $d_{i,m}^{-\alpha}(t_n)$ is the propagation distance between IoT device i and UAV m, $d_{i,m}(t_n) = \sqrt{(x^m - \hat{x}^i)^2 + (y^m - \hat{y}^i)^2 + (z_n^m)^2}$. The received power at UAV m from IoT device i in a communication sub-slot t_n^l is,

$$P_{RX}^{i,m}(t_n^l) = \begin{cases} \hat{\mathcal{I}}_i(t_n^l) c_0 h_{i,m}(t_n) P_{TX} d_0^{-\alpha}, \ d_{i,m}(t_n) \le d_0, \\ \hat{\mathcal{I}}_i(t_n^l) c_0 h_{i,m}(t_n) P_{TX} d_{i,m}^{-\alpha}(t_n), d_{i,m}(t_n) \ge d_0, \end{cases} \tag{7}$$

where $c_0 = \frac{\lambda^\alpha}{(4\pi)^\alpha}$ is a constant which depends on the wavelength of the transmitted signal, λ, and d_0 is a reference distance. $\hat{\mathcal{I}}_i(t_n^l)$ is a Bernoulli random variable with parameter $p(t_n) \in (0, 1]$ which indicates whether IoT device i transmits during communication sub-slot t_n^l, i.e., $\hat{\mathcal{I}}_i(t_n^l) = 1$ with probability $p(t_n)$. UAV m first decodes the signal with the highest signal power under the interference from all other IoT devices involved in the NOMA transmissions. Without loss of generality, IoT devices $\mathcal{I}_m(t_n^l) = \{i | \hat{\mathcal{I}}_i(t_n^l) = 1\}$ are sorted in the descending order of their received signal strength at UAV m, such that $i = 1$ is the IoT device with the highest received signal to interference plus noise (SNIR) at UAV m, and $i = 2$ is the IoT device with the second highest received SNIR at UAV m [4]. The highest received SNIR at UAV m in a communication sub-slot t_n^l is therefore,

$$\text{SNIR}_{1,m}(t_n^l) = \frac{P_{RX}^{1,m}(t_n^l)}{n_0 + \sum_{k \in \mathcal{I}_m(t_n^l) \setminus 1} P_{RX}^{k,m}(t_n^l)}, \tag{8}$$

where n_0 is the noise floor power. Similarly, the second highest received SNIR at UAV m in a communication sub-slot t_n^l is,

$$\text{SNIR}_{2,m}(t_n^l) = \frac{P_{RX}^{2,m}(t_n^l)}{n_0 + \sum_{k \in \mathcal{I}_m(t_n^l) \setminus \{1,2\}} P_{RX}^{k,m}(t_n^l)}, \tag{9}$$

UAV m can decode the signal with $\text{SNIR}_{1,m}(t_n^l)$ if

1) user 1 is associated with UAV m during communication sub-slot n, $u_n^{1,m} = 1$, and,
2) $\text{SNIR}_{1,m}(t_n^l)$ is larger than the SNIR threshold, i.e., $U(\text{SNIR}_{1,m}(t_n^l)) = \text{SNIR}_{1,m}(t_n^l)$,

where $U(.)$ is a thresholding function to maintain a minimum quality of service,

$$U(\text{SNIR}_{i,m}(t_n^l)) = \begin{cases} 0, \ \text{SNIR}_{i,m}(t_n^l) < \text{SNIR}_{\text{Th}}, \\ \text{SNIR}_{i,m}(t_n^l), \text{SNIR}_{i,m}(t_n^l) \ge \text{SNIR}_{\text{Th}}. \end{cases} \tag{10}$$

In addition, UAV m can decode the signal with $\text{SNIR}_{2,m}(t_n^l)$ if

1) The signal with $\text{SNIR}_{1,m}(t_n^l)$ is successfully decoded,
2) user 2 is associated with UAV m during communication sub-slot n, $u_n^{2,m} = 1$, and,
3) $\text{SNIR}_{2,m}(t_n^l)$ is larger than the SNIR threshold, i.e., $U(\text{SNIR}_{2,m}(t_n^l)) = \text{SNIR}_{2,m}(t_n^l)$

[1] Implementing time-synchronization among IoT devices is important to enable NOMA decoders. Without time-synchronization, transmissions may partially overlap which makes the design of NOMA decoders hard.

[2] Lloyd's K-means clustering algorithm is an iterative algorithm to determine a set of K centroids given a large set of IoT device locations $\{\hat{x}^i, \hat{y}^i\}_{i=1}^N$, so as to minimize the within-cluster variance (sum of squared distances to cluster centroid), $\min_{\{(x^m, y^m)\}} \sum_{m=1}^M \sum_{i \in \mathcal{N}_m} ||(\hat{x}^i, \hat{y}^i) - (x^m, y^m)||^2$. Given that the range of vertical flight is the dominant factor in our system model, determining the planar locations of UAVs a priori reduces the number of control variables without adversely impacting network performance, as will be shown in Sec. V.

[3] The statistical channel state information, $h_{i,m}(t_n)$, is assumed to be quasi-static and is fixed during a time slot n.

[4] We consider the two highest received signals to trade-off NOMA gain and SIC decoding complexity for uplink transmissions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications

The sum rates of the received data at UAV m in communication sub-slot t_n^l is,

$$G_m(t_n^l) = \mathcal{W}\log_2\Big(1 + U(\mathrm{SNIR}_{1,m}(t_n^l))u_n^{1,m}\Big) + \mathcal{W}\log_2\Big(1 + U(\mathrm{SNIR}_{2,m}(t_n^l))u_n^{1,m}u_n^{2,m}e_n^{1,m}\Big) \tag{11}$$

where $\mathcal{W}$ is the transmission bandwidth, and $e_n^{1,m} = 1$ if $U(\mathrm{SNIR}_{1,m}(t_n^l)) = \mathrm{SNIR}_{1,m}(t_n^l)$ and 0 otherwise. The total network capacity in any given system slot t_n,

$$\mathbb{G}(t_n) = \sum_{l=0}^{L-1}\sum_{m\in\mathcal{M}} G_m(t_n^l) \tag{12}$$

On the other hand, UAVs are equipped with solar panels, which harvest solar energy to replenish the on-board battery. The attenuation of solar light passing through a cloud can be modeled based on [32],

$$\phi(d^{cloud}) = e^{-\beta_c d^{cloud}} \tag{13}$$

where $\beta_c \geq 0$ denotes the absorption coefficient of the cloud, and d^{cloud} is the distance that the solar light travels through the cloud. Following [32] and the references therein, the solar energy harvested by UAV m during time slot n can be modeled as,

$$E_{\mathrm{H}}^m(t_n) = \begin{cases} \psi\tilde{S}\tilde{G}\Delta t, & \frac{z_n^m + z_{n+1}^m}{2} \geq z_{high} \\ \psi\tilde{S}\tilde{G}\phi(z_{high} - \frac{z_n^m + z_{n+1}^m}{2})\Delta t, & z_{low} \leq \frac{z_n^m + z_{n+1}^m}{2} < z_{high} \\ \psi\tilde{S}\tilde{G}\phi(z_{high} - z_{low})\Delta t, & \frac{z_n^m + z_{n+1}^M}{2} < z_{low} \end{cases} \tag{14}$$

where ψ is a constant representing the energy harvesting efficiency, $\tilde{S}$ is the area of solar panels, and $\tilde{G}$ denotes the average solar radiation intensity on earth. z_{high} and z_{low} are the altitudes of upper and lower boundaries of the cloud. During time-slot n, UAV m can cruise upwards or downwards from $r_m(t_n)$ to $r_m(t_{n+1})$. The energy consumed by UAV m during time slot n [32] is,

$$E_{\mathrm{C}}^m(t_n) = \left(\frac{W^2/(\sqrt{2}\rho A)}{4^{0.25}V_z} + Wv_z + P_{\text{static}} + P_{\text{antenna}}\right)\Delta t, \qquad v_z = \frac{z_{n+1}^m - z_n^m}{\Delta t} \tag{15}$$

where, $V_z = \sqrt{\frac{W}{2\rho A}}$, W is the weight of the UAV, ρ is air density, and A is the total area of UAV rotor disks. P_{static} is the power consumed for maintaining the operation of UAV, and P_{antenna} is the power consumed by the receiving antenna. It is worth to mention that cruising upwards consumes more power than cruising downward and hovering.

Denote the battery energy storage of UAV m at the beginning of slot n by $B_m(t_n)$. The battery energy in the next slot is given by,

$$B_m(t_{n+1}) = \min\{[B_m(t_n) + E_{\mathrm{H}}(t_n) - E_{\mathrm{C}}^m(t_n) + \mathbb{B}(t_n)]^+, B_{\max}\}, \tag{16}$$

where $\mathbb{B}(t_n), \forall n$, are independent zero-mean gaussian random variables with variance σ_B^2 which characterizes the randomness in the battery evolution process, and $[\,]^+$ denotes the positive part.

IV. Problem Formulation and Proposed CDRL Algorithm

In order to maximize the total network capacity under stochastic mutual interference of IoT devices while ensuring energy sustainability of UAVs over the operating horizon H, the central controller decides on the altitude of each UAV m, $\forall m \in \mathcal{M}$, at the beginning of each slot n, z_n^m, as well as the channel access probability $p(t_n)$ of IoT devices considering the potential access gain provisioned by NOMA. The channel access probability will be broadcast to IoT devices through beacons at the beginning of each time slot, so that IoT devices adapt their random channel access parameter and maintain slotted-time synchronization.

The problem of maximizing the total network capacity while ensuring energy sustainability of each UAV is a constrained stochastic optimization problem over the operating horizon due to random channel access, the stochastic channel model, dynamic network topology, and the stochastic energy evolution in the batteries of UAVs. Moreover, this problem is mathematically intractable because of the combinatorial space of possible transmissions and channel interference events induced by the random access protocol, and hence offline solutions cannot be devised. To solve this problem and find an energy-sustainable capacity-optimal control policy, we formulate it as a CMDP and design an online Constrained Deep Reinforcement Learning (CDRL) algorithm by combining state-of-the-art DRL policy optimization algorithms with Lagrangian primal-dual techniques.

A. CMDP Formulation

To enable continuous control of UAVs altitudes and channel access probability, we consider parametrized DNN based policies with parameters $\boldsymbol{\theta}$, and parametrized state-value function with parameters $\boldsymbol{\Theta}$ henceforth. We formulate the joint problem of UAVs altitude control and random channel access of IoT devices as a discrete-time CMDP with continuous state and action spaces as follows,

1) $\forall s_n \in \mathcal{S}$,

$$\begin{aligned} s_n = \bigcap_{\mathcal{M}}\Big\{ & z_n^m, \cdots, z_{n-h_k}^m, B_m(t_n), \cdots, B_m(t_{n-h_k}), \\ & P\big(\mathrm{SNIR}_{1,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big), \\ & P\big(\mathrm{SNIR}_{2,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big), \\ & \mathbb{E}\big[\mathrm{SNIR}_{1,m}(t_n^l)|\mathrm{SNIR}_{1,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big], \\ & \mathbb{E}\big[\mathrm{SNIR}_{2,m}(t_n^l)|\mathrm{SNIR}_{2,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big], \\ & \mathrm{Var}\big[\mathrm{SNIR}_{1,m}(t_n^l)|\mathrm{SNIR}_{1,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big], \\ & \mathrm{Var}\big[\mathrm{SNIR}_{2,m}(t_n^l)|\mathrm{SNIR}_{2,m}(t_n^l) \geq \mathrm{SNIR}_{\mathrm{Th}}\big]\Big\}, \end{aligned}$$

i.e., the state space encompasses $\forall m$, the current altitude of UAV m along with h_k historical altitudes, current battery energy of UAV m along with h_k historical battery energies, probability the highest and second highest received SNIRs from associated users at UAV m is greater than or equal to $\mathrm{SNIR}_{\mathrm{Th}}$, the mean of the highest and second highest received SNIRs from associated users at UAV m given that they are greater than or equal to the SNIR threshold, and the variance of the highest and second highest received SNIRs from associated users at UAV m given that they are greater than or equal to the SNIR threshold. Here, the mean and variance are calculated over the L communication sub-slots.

2) $\forall a_n \in \mathcal{A}$, $a_n = \bigcap_{\mathcal{M}} \{\Delta z_n^m\} \cap \{p(t_{n+1})\}$, where $\Delta z_n^m = z_{n+1}^m - z_n^m$, i.e., the action space encompasses the altitude displacement of each UAV between any two consecutive time slots, and the random channel access probability in the next system slot.
3) $\mathcal{R}(s_n, a_n) = \frac{\mathbb{G}(t_n)}{H}$, i.e., the immediate reward at the end of each time slot n is the total network capacity during slot n, normalized by the operating horizon H.
4) $\mathcal{C}_m(s_n, a_n) = \frac{B_m(t_n) - B_m(t_{n+1})}{B_{max}}$, $\forall m$, i.e., the immediate cost at the end of each slot n is the change in the battery energy between any two consecutive time slots, which is caused by the displacement of each UAV m, normalized by the maximum battery energy.
5) $E_m = -B_{min}, \forall m$, i.e., the upper bound on the long-term expected cost is the negative of the minimum desired battery energy increase at the end of the planning horizon over the initial battery energy.

Based on this formulation, the objective is to find a Markov policy $\pi_{\boldsymbol{\theta}_\pi}$ which maximizes the long-term expected discounted total network capacity, while ensuring energy sustainability of each UAV m over an operating horizon,

$$\begin{aligned} &\max_{\boldsymbol{\theta}_\pi} \quad \mathbb{E}^{\beta}_{\pi_{\boldsymbol{\theta}}}\Big[\sum_{n=0}^{\infty} \zeta^n \mathbb{G}(t_n)\Big] \\ &\qquad\quad \mathbb{E}^{\beta}_{\pi_{\boldsymbol{\theta}}}\Big[\sum_{n=0}^{H} B_m(t_n) - B_m(t_{n+1})\Big] \leq -B_{min}, \forall m \end{aligned} \tag{17}$$

Problem (17) exhibits trade-offs between total system capacity and energy sustainability of UAVs. For instance, a UAV hovering at a higher altitude above the cloud cover can harvest more solar energy to replenish its on-board battery storage, as given by (14). However, at higher altitudes, the received signal power at a UAV from IoT devices will be smaller due to the log-distance path loss model, and consequently, the system capacity will be smaller. The converse is true, that is, when a UAV hovers at lower altitudes, network capacity is improved, yet solar energy harvesting is heavily attenuated. In addition, based on the network topology at any time slot n, spatial gain and NOMA overload vary. An optimal stochastic control policy for altitude control of UAVs and channel access management of IoT devices should be therefore learned online. In the following subsection, we propose an online CDRL algorithm to solve (17).

B. Proposed CDRL Algorithm

To solve (17) in absence of the state-action-state transition probabilities of the Markov model, we adopt the RL framework in which an autonomous agent learns an optimal policy by repeated interactions with the wireless environment [46].

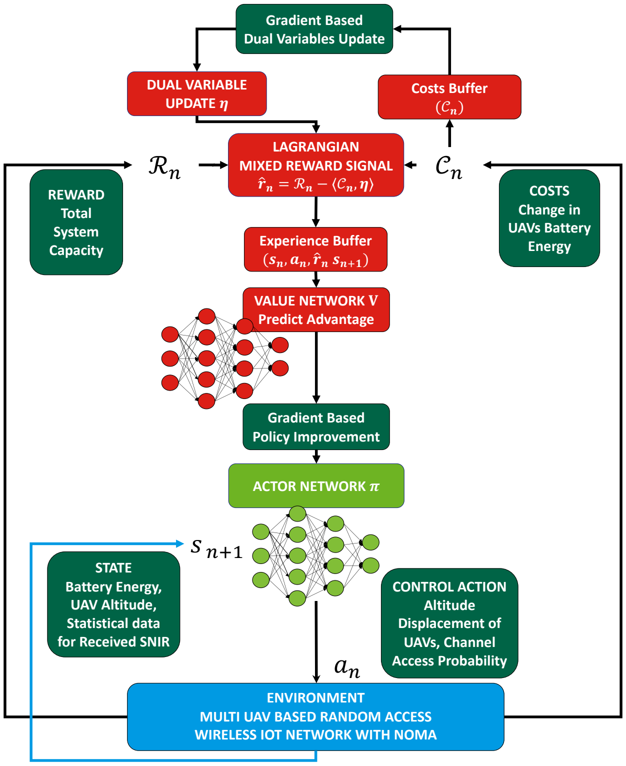


Fig. 2. Proposed constrained deep reinforcement learning architecture. Constrained policy optimization is performed on three time scales. On the fastest time scale, the state-value function is updated by minimizing (22), then the policy is updated by maximizing (18) on the intermediate time-scale, and finally, the Lagrangian multipliers are updated on the slowest time-scale by minimizing (21).

The wireless environment provides the agent with rewards and costs signals, which the agents exploit to further improve its policy. Our proposed algorithm is based on the state-of-the-art Proximal Policy Optimization (PPO) algorithm [24], and leverages the technique of primal-dual optimization [42]. The architecture of our proposed algorithm is shown in Figure 2.

In the proposed CDRL algorithm, parameterized DNN of the policy $\pi_{\boldsymbol{\theta}}(a|s)$ is learned by maximizing the PPO-clip objective function, which is a specially designed clipped surrogate advantage objective that ensures constructive policy updates,

$$O^{\text{clip}}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_n\Big[\min(\tfrac{\pi_{\boldsymbol{\theta}}(a_n|s_n)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_n|s_n)}\hat{A}_n, \text{clip}(\tfrac{\pi_{\boldsymbol{\theta}}(a_n|s_n)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_n|s_n)}, 1+\epsilon, 1-\epsilon)\hat{A}_n)\Big], \tag{18}$$

where $\boldsymbol{\theta}$ are the policy neural network parameters, ϵ is a clip fraction, and $\hat{A}_n$ is the generalized advantage estimator (GAE) [55][5],

$$\hat{A}_n = \sum_{l=0}^{\infty} (\zeta\xi)^l \big(\hat{\mathcal{R}}_{n+l} + \zeta V_{\boldsymbol{\Theta}}(s_{n+l+1}) - V_{\boldsymbol{\Theta}}(s_{n+l})\big). \tag{19}$$

Clipping in (18) acts as a regularizer which controls how much the new policy can go away from the old one while

[5] The advantage function is defined as the difference between the state-action value function and the value function, $A(s_n, a_n) = Q(s_n, a_n) - V(s_n)$. GAE makes a compromise between bias and variance in estimating the advantage.

still improving the training objective. In order to further ensure reasonable policy updates, we adopt a simple early stopping method in which gradient optimization on (18) is terminated when the mean KL-divergence between the new and old policy reaches a predefined threshold KL_{Th}. In (19), $\hat{\mathcal{R}}_n = \hat{\mathcal{R}}_n(s_n, a_n, \boldsymbol{\eta})$ is a Lagrangian penalized reward signal [41],

$$\hat{\mathcal{R}}(s_n, a_n, \boldsymbol{\eta}) = \mathcal{R}(s_n, a_n) - \sum_m \eta_m \mathcal{C}_m(s_n, a_n). \tag{20}$$

In our proposed algorithm, the Lagrangian penalty multipliers are updated adaptively according to policy feasibility by performing gradient descent on the original constraints. Given that $\boldsymbol{\eta}$ are initially set to 0, i.e., the agent is initially indifferent to the cost constraints, we propose to learn the Lagrangian penalty multipliers by minimizing the following loss function with respect to $\boldsymbol{\eta}$ in order to ensure learning stability and convergence to a local saddle point,

$$O^P(\boldsymbol{\eta}) = \textstyle\sum_m \eta_m \text{clip}\Big(- B_{min} - \mathbb{E}^{\beta}_{\pi_{\boldsymbol{\theta}}}\Big[\sum_{n=0}^{H} B_m(t_n) - B_m(t_{n+1})\Big], -\infty, 0\Big) \tag{21}$$

If a constraint m is violated, then $\frac{\partial O^P}{\partial \eta_m} < 0$, and so η_m will be increased to enforce the constraint. Due to clipping, η_m will not be updated if the constraint m is satisfied. By ensuring that $\boldsymbol{\eta}$ are monotonously increased starting from $\boldsymbol{\eta}_0 = 0$ during training, the agent avoids oscillations between the feasible and infeasible policy spaces when optimizing the policy parameters (18), which improves learning stability.

Finally, the state-value function is learned by minimizing the mean squared error loss against the policy's discounted rewards-to-go,

$$O^{\text{V}}(\boldsymbol{\Theta}) = \hat{\mathbb{E}}_n\Big[\Big(V_{\boldsymbol{\Theta}}(s_n) - \sum_{l=0}^{\infty} \zeta^l \hat{\mathcal{R}}_{n+l}(s_{n+l}, a_{n+l})\Big)^2\Big]. \tag{22}$$

The optimization in our proposed algorithm is performed over three time-scales, on the fastest time scale, the state-value function is updated by minimizing (22), then the policy is updated by maximizing (18) on the intermediate time-scale, and finally, the Lagrangian multipliers are updated on the slowest time-scale by minimizing (21). Optimization time-scales are controlled by choosing the maximum learning rate of the stochastic gradient optimizer used, e.g., adaptive moment estimation (ADAM) [56], as well as the number of gradient steps performed at the end of each training epoch. The full algorithmic procedure for training the CDRL agent is outlined in Algorithm 1.

To parameterize the space of continuous control policies, a parameterized stochastic Gaussian policy is adopted [46],

$$\pi_{\boldsymbol{\theta}}(a_n|s_n) = \frac{1}{\sigma(s_n, \boldsymbol{\theta_\sigma})\sqrt{2\pi}} \exp\Big(- \frac{(a_n - \mu(s_n, \boldsymbol{\theta_\mu}))^2}{2\sigma(s_n, \boldsymbol{\theta_\sigma})^2}\Big). \tag{23}$$

where $\boldsymbol{\theta_\mu}$ are the DNN parameters for the mean of the policy, and $\boldsymbol{\theta_\sigma}$ are the DNN parameters for the variance of the policy. The choice of a gaussian policy is commonly adopted in DRL because the gradient of the policy as well as the log probabilities necessary for KL-divergence computation can be derived in a closed form [46]. At the beginning of training, the variance of the policy network encourages exploration. As training progresses, the variance of the policy is reduced due to maximizing (18) and in doing so, the policy shifts slowly towards the deterministic policy given by $\mu(s_n, \boldsymbol{\theta_\mu})$.

Algorithm 1: Constrained PPO-Clip

```
Input: Initial policy network parameters $\boldsymbol{\theta}$, initial value
       network parameters $\boldsymbol{\Theta}$, initial Lagrange
       multipliers $\boldsymbol{\eta} = \mathbf{0}$
for epoch $= 0, 1, \cdots$ do
    for $n = 0, 1, \cdots, H$ do
        Observe initial state $s_n$
        Sample action $a_n \sim \pi_{\boldsymbol{\theta}}(a_n|s_n)$
        Take action $a_n$
        Receive reward $\mathcal{R}(s_n, a_n)$, $M$ costs $\mathcal{C}_m(s_n, a_n)$,
          and new state $s_{n+1}$
        Compute penalized reward $\hat{\mathcal{R}}(s_n, a_n, \boldsymbol{\eta})$ using
          (20)
        Store transition $(s_n, a_n, \hat{\mathcal{R}}(s_n, a_n, \boldsymbol{\eta}), s_{n+1})$ in
          policy training buffer
        Store $\mathcal{C}_{k,\forall k}(s_n, a_n)$ in Lagrange multiplier
          training buffer
    end
    Compute rewards to go
    Compute advantage estimate $\hat{A}_n$ using GAE (19) and
      current value network
    for $k = 0, 1, \cdots$ do
        Update policy parameters $\boldsymbol{\theta}_k$ via stochastic
          gradient ascent with ADAM on the PPO-Clip
          objective (18)
        Compute KL-divergence between new policy and
          old policy
        Break if KL-divergence hits $\text{KL}_{\text{Th}}$
    end
    for $k = 0, 1, \cdots$ do
        Fit the value network via stochastic gradient
          descent with ADAM on (22)
    end
    Update Lagrangian multipliers via stochastic gradient
      descent with ADAM on (21)
end
```

CDRL Implementation and Training: A fully connected multi-layer perceptron network with three hidden layers for both the policy and value networks are used. Each hidden layer has 128 neurons. $Tanh$ activation units are used in all neurons. The range of output neurons responsible for the altitude displacement of each UAV is linearly scaled to $[\Delta z_{min}, \Delta z_{max}]$ so as to limit the maximum cruising velocity of the UAVs. The range of the output neuron controlling the random channel access probability is linearly scaled to $[0, \frac{2}{N}]$. The weights of the policy network are heuristically initialized to generate a feasible policy. The variance of the Gaussian policy is state-independent, $\sigma(s_n, \boldsymbol{\theta_\sigma}) = \boldsymbol{\theta_\sigma}$, with initial value $\boldsymbol{\theta^0_\sigma} = e^{-0.5}$.

Training of the proposed CDRL agent has been performed over 1000 epochs, where each epoch corresponds to 32 episodes, and each episode corresponds to trajectories of length $H = 360$ time steps, amounting for about 11.5 million training samples. At the end of each episode, the trajectory is cut-off and the wireless system is reinitialized. Episodes in each epoch are rolled-out in parallel by 32 Message Passing Interface (MPI) ranks, to sample 32 trajectories and reduce gradient estimation variance. After each MPI rank completes its episodic roll-out, Lagrangian primal-dual policy optimization is performed as outlined in Algorithm 1 based on the average gradients of the MPI ranks, such that the DNN parameters $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$, and the Lagrange multipliers $\boldsymbol{\eta}$, remain synchronized among the 32 MPI ranks during training. At the end of training, the trained policy network corresponding to the mean of the learned Gaussian policy, $\mu(s_n, \boldsymbol{\theta}_\mu)$, is used to test its performance through the simulated environment.

V. Performance Evaluation

We have developed a simulator in Python for the solar-powered multi-UAV based Wireless IoT network with NOMA described in section III, and implemented the proposed CDRL algorithm based on OpenAI's implementation of PPO [57]. We trained the proposed CDRL agent in a multi-cell wireless IoT network of $M = 2$ solar-powered UAVs and $N = 200$ IoT devices. At testing, N is varied to demonstrate the efficacy of the trained CDRL in managing massive number of IoT devices. IoT devices were deployed independently and uniformly within a grid of $[0, 0] \times [1500, 500]$m. The two UAVs were initially deployed at $(250, 250, 750)$m and $(750, 250, 1250)$m with 50% initial battery energy, i.e., 111 Wh. Notice that the x and y coordinates of the two UAVs, i.e., $(250, 250)$m and $(750, 250)$m, are minimizers of the sum of squared planar distances between IoT devices and the closest UAV, as determined by Lloyd's K-means clustering algorithm for the uniform random deployment of IoT devices on the ground. The impacts of different planar UAV deployments on network performance will also be investigated. UAVs were allowed to cruise vertically between 500m and 1500m. The main simulation parameters for the experiments are outlined in Table I. UAV related parameters are based on the work of [32].

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
P_{TX}	30 dBm	B_{min}	22 Wh
α	2	ψ	0.4
Δt	$10s$	$\tilde{S}$	$1m^2$
f_0	900 MHz	$\tilde{G}$	$1367W/m^2$
d_0	$1m$	W	$39.2kg * m/s^2$
$h_{i,m}$	$exp(1)$	ρ	$1.225kg/m^3$
n_0	$-80dBm$	A	$0.18m^2$
$\mathcal{W}$	$1Hz$	$P_{static} + P_{antenna}$	5 watts
SNIR$_{\text{Th}}$	$10dB$	ϵ	0.2
L	1000	h_k	4
β_c	0.01	H	360 (1hr)
z_{high}, z_{low}	$1.3, 0.7$km	$\Delta z_{min}, \Delta z_{max}$	$-40m, 40m$
z_{min}, z_{max}	$0.5, 1.5$km	ξ	0.97
B_{max}	222 Wh	$\boldsymbol{\eta}_{\text{lr}}$	3×10^{-3}
ζ	0.999	$\boldsymbol{\Theta}_{\text{lr}}$	10^{-3}
KL$_{\text{Th}}$	0.01	$\boldsymbol{\theta}_{\text{lr}}$	3×10^{-4}
σ_B^2	500		

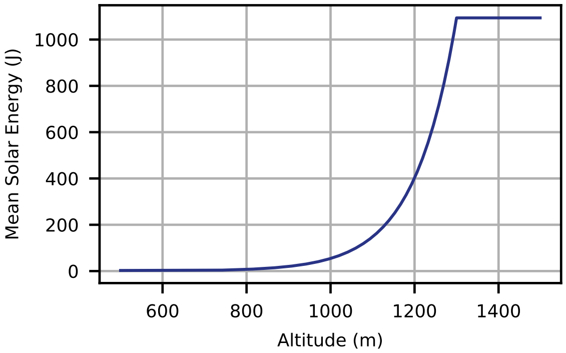


Fig. 3. Solar energy output of solar panels versus UAV altitude. Solar energy decays exponentially through the cloud cover between 1300m and 700m.

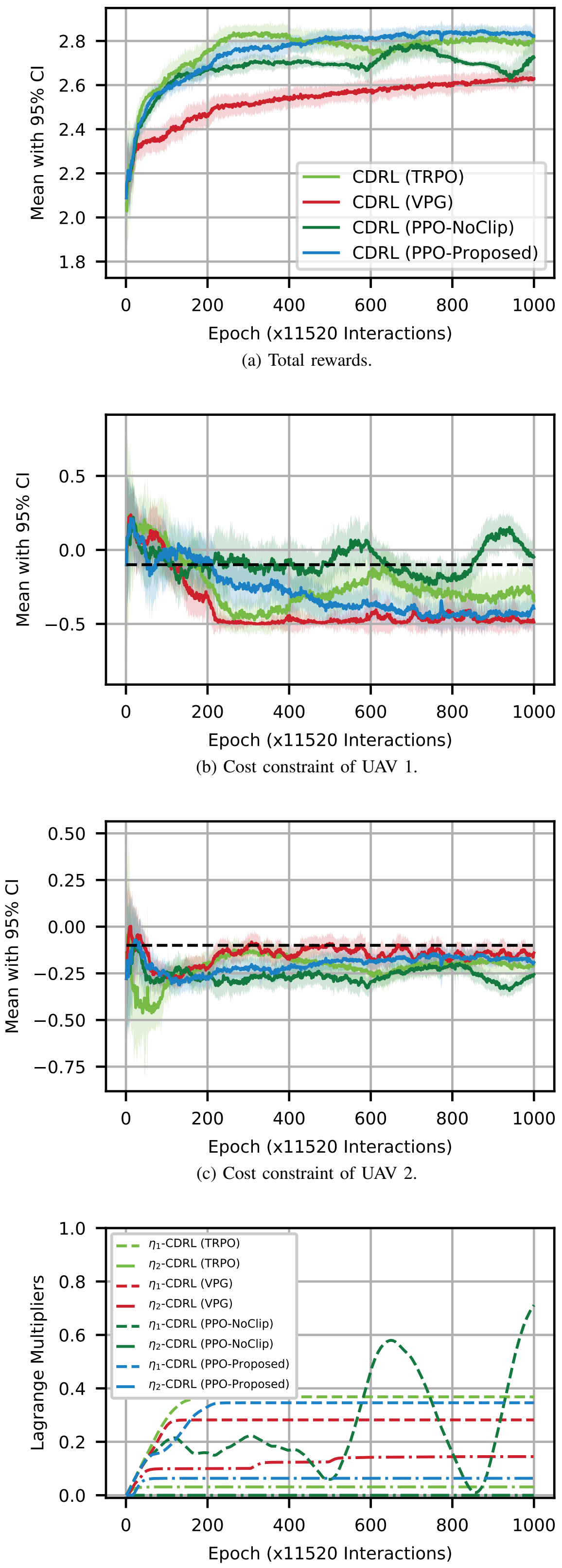
The output energy of solar panels mounted on the UAV as a function of altitude according to (14) is shown in Figure 3. The highest solar energy is achieved when the UAV is above the cloud cover at $z_{high} = 1300$m. Notice how solar energy decays exponentially through the clouds which extend down to $z_{low} = 700$m. When the UAV is below 700m, the output energy of its solar panels is zero. UAVs cruising above 1300m harvest the most solar energy, while UAVs cruising below 700m harvest the least energy. Hovering at low altitudes reduces distance-dependant path-loss and improves the wireless channel capacity, however, it is not energy sustainable.

The learning curves of the trained CDRL agent, (PPO-Proposed), are shown in Figure 4. To evaluate the effects of policy optimization in our proposed CDRL framework, we have compared the learning performance with two other DRL agents which adopt different policy optimization algorithms, namely, (TRPO) [22] and (VPG) [23]. In these algorithms, the Lagrangian penalty multipliers are adapted during training as in Algorithm 1 by minimizing (21) with respect to $\boldsymbol{\eta}$. In addition, we demonstrate the effects of clipping in (21) on learning stability by training a CDRL agent (PPO-NoClip), which adopts PPO for policy optimization, yet it adapts the Lagrange penalty multipliers by minimizing an unclipped version of (21) with respect to $\boldsymbol{\eta}$. It can be seen from Figure 4(a) that the CDRL agent becomes more experienced as training progresses and collects higher expected total rewards. In addition, the proposed CDRL agent becomes more experienced in satisfying energy constraints as can be seen from Figures 4(b)-(c), and learns a policy whose expected costs fall below the normalized energy constraint upper bound $-B_{min}/B_{max} = -0.1$. The proposed CDRL agent outperforms both CDRL (TRPO) and CDRL (VPG) in terms of the achieved total rewards during training. Compared with CDRL (TRPO) and CDRL (VPG), the proposed CDRL agent also exhibits a relatively more stable constraint satisfaction during training. This finding is consistent with previous works which demonstrates the superiority of PPO as a policy optimization algorithm on a variety of benchmark tasks [24]. On the other hand, the convergence of the Lagrangian multipliers to non-negative values during the training of the proposed CDRL algorithm is shown in Figure 4(d). It can be observed that the two cost constraints are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications

penalized differently, which is primarily due to the different initial conditions and different terminal states. The Lagrangian multiplier of the energy constraint corresponding to UAV 1 is larger than that of UAV 2, therefore it is expected that UAV 1 will end up its flight with more harvested energy in its battery. Notice that in the case of (PPO-NoClip), where the Lagrange penalty multipliers are adapted during training by minimizing an unclipped version of (21), the Lagrange multipliers fluctuate which leads to oscillations between the feasible and infeasible policy spaces as can be seen from Figure 4(b). On the other hand, agents in which the Lagrangian penalty multipliers are adapted by minimizing the clipped objective (21), have monotonously increasing Lagrange multipliers which converge early during training. Clipping in (21) results in a better learning stability and constraint satisfaction compared with (PPO-NoClip), as can be seen from Figures 4(a)-(c).

The learned policy by our proposed CDRL algorithm is shown in Figure 5. It can be seen from Figures 5(a) and 5(b) that the CDRL agent learns a policy in which the two UAVs take turns in cruising upwards to recharge their on-board batteries, and in serving IoT devices deployed on the ground. UAV 2 first climbs up to recharge its battery, while UAV 1 descends down to improve communication performance for IoT devices on the ground. Since IoT devices are associated with the closest UAV, in this case, all IoT devices are associated with UAV 1. When UAV 2's battery is fully charged, it descends down gradually to switch roles with UAV 1: UAV 2 becomes the BS with which all IoT devices are associated, while UAV 1 climbs up to recharge its battery. Such a policy ensures that the battery energy of the two UAVs is not drained throughout the operating horizon as can be seen from Figure 5(c). Note that the terminal energy in UAV's 1 battery is higher than that of UAV 2, which is expected based on the observation that the Lagrange multiplier corresponding to UAV's 1 energy constraint is larger than that of UAV 2. In Figure 5(d), the random channel access probability based on the learned CDRL policy is shown. It can be observed that when either of the two UAVs is fully serving all the 200 IoT devices, the wireless networking system is overloaded with $p > 1/N$, thanks to spatial gain and NOMA. Note that $p = 1/N$ is the optimal transmission probability in single-cell p-persistent slotted Aloha systems, which is oblivious to NOMA and the heterogeneous channel conditions of IoT devices. The channel access probability is dynamically adapted when the two UAVs cruise upward and downward to exchange roles in the wireless system. At times when both UAVs have associated users, the channel access probability can be seen to spike higher to maintain NOMA overload, as can be observed from Figures 6(a)-(b). It can be seen from these two figures that NOMA's gain is higher in steady states when all IoT devices are associated with the same UAV, compared to transient states when the two UAVs exchange roles and are both serving IoT devices. This is because it is less likely that the second highest received SINR to a UAV is from within the same cell at times when both UAVs provision wireless service. By deploying multiple UAVs, it is therefore possible to learn a cooperative policy in which UAVs take turns to charge their battery and provision uninterrupted wireless service.



(a) Total rewards.

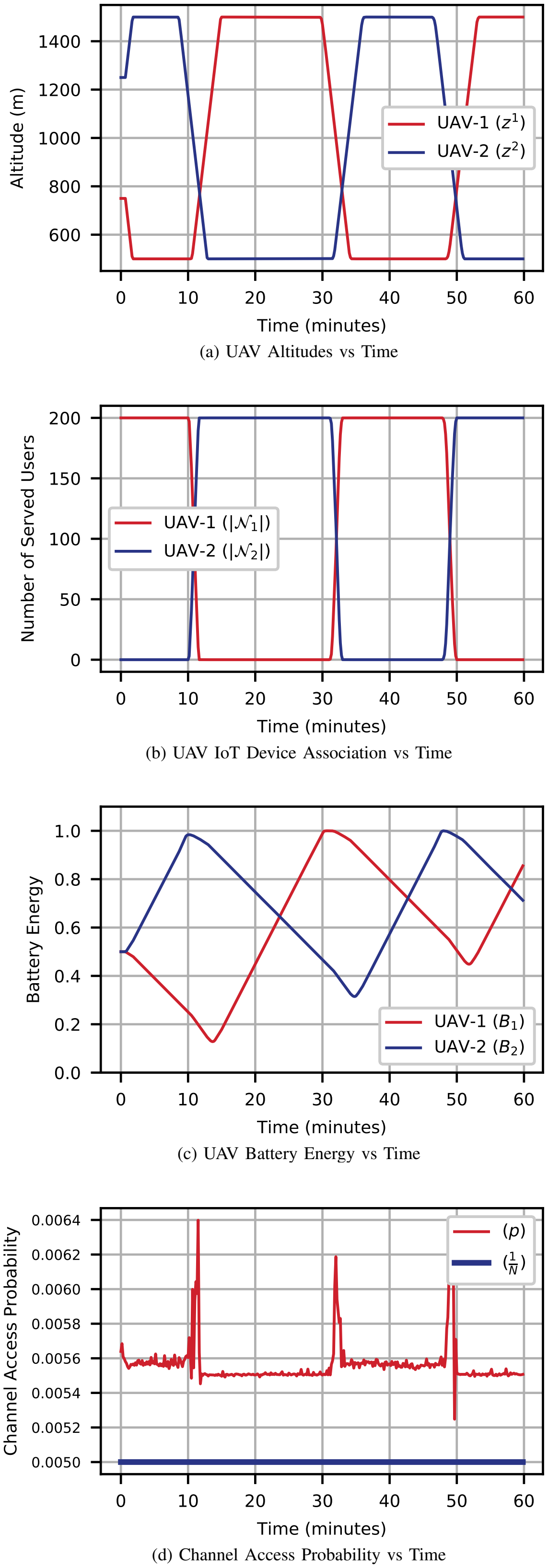
(b) Cost constraint of UAV 1.

(c) Cost constraint of UAV 2.

(d) Lagrange Multipliers.

Fig. 4. Training results of the proposed CDRL agent. As training progresses, the proposed CDRL agent becomes more experienced; collects higher expected total rewards, and better satisfies the energy constraints. Proposed CDRL agent outperforms other agents and possesses an improved learning stability thanks to clipping in (18) and (21).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications



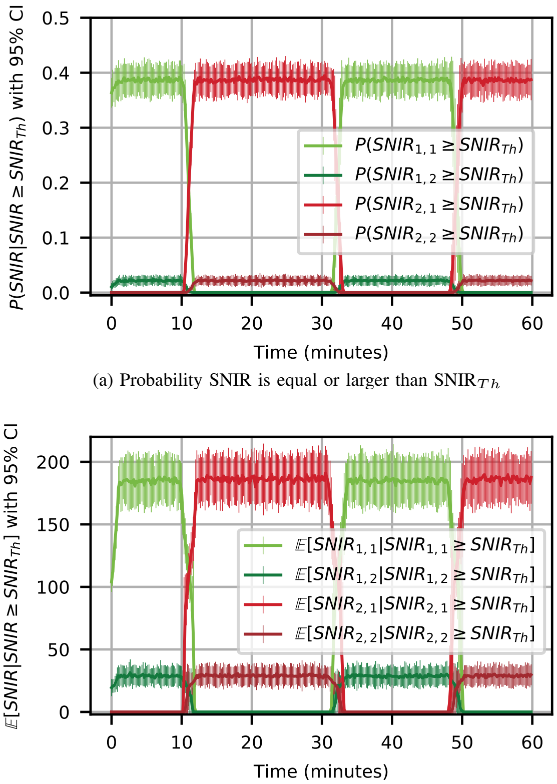
(a) UAV Altitudes vs Time

(b) UAV IoT Device Association vs Time

(c) UAV Battery Energy vs Time

(d) Channel Access Probability vs Time

Fig. 5. Learned policy by the CDRL agent during the operating horizon of the two UAVs. A cooperative policy among UAVs is learned in which the two UAVs take turns to recharge the battery and provision uninterrupted wireless service to IoT devices.



(a) Probability SNIR is equal or larger than SNIR$_{Th}$

(b) Conditional Expected SNIR

Fig. 6. NOMA performance based on the learned CDRL policy during the operating horizon of the two UAVs. NOMA's gain is higher in steady states when all IoT devices are associated with the same UAV, compared to transient states when the two UAVs exchange roles and are both serving IoT devices.

To evaluate the efficacy of the learned policy, we have compared its performance with three baseline policies which are learned by 1) an unconstrained PPO agent which is energy-unaware (RL Energy-Unaware), i.e., the agent attempts to maximize the wireless network capacity with indifference to battery energy dissipation of the UAVs, 2) an unconstrained PPO agent that penalizes energy dissipation due to flight via fixed reward shaping (RLWS), where the training reward signal is $\hat{\mathcal{R}}(s_n, a_n, \boldsymbol{\eta}) = \mathcal{R}(s_n, a_n) - \eta_1 \mathcal{C}_1(s_n, a_n) - \eta_2 \mathcal{C}_2(s_n, a_n)$, and $\eta_1 = 10, \eta_2 = 10$ or $\eta_1 = 0, \eta_2 = 10$, and 3) a random management policy. The achieved mean total network capacity with 95% confidence interval versus flight time based on these policies is shown in Figure 7. The statistical results are based on 32 roll-outs of the learned policy in the simulated wireless IoT environment with NOMA. Compared with baseline policies, it can be seen that the proposed CDRL agent learns an adaptive policy which achieves a temporal average network capacity that is 81% higher compared to the feasible energy sustainable policy learned by the conservative RLWS agent with $\eta_1 = \eta_2 = 10$. Compared to the policy learned by the (RL Energy-Unaware) agent which provides an upper bound on the achievable wireless network capacity, only 6.7% of the system capacity is sacrificed in order to maintain energy sustainability of UAVs. In Table II, we provide statistics which characterize the flight-time of the learned policies to explain their behaviour. Specifically, the minimum flying altitude $\min_n z_m(t_n)$, geometric mean of the altitude during flight $\mathcal{G}_m[z_m(t_n)] = \left(\prod_{n=1}^{H} z_m(t_n)\right)^{\frac{1}{H}}$, and the maximum flying altitude $\max_n z_m(t_n)$, for $m = \{1, 2\}$, are given. These statistics help explain the trade-off between energy sustainability and the achievable wireless network capacity. Notice that the policy learned by the RLWS agent with $\eta_1 = \eta_2 = 10$ is the most conservative. In this policy, both UAVs climb upwards and hover above the cloud cover to maintain high battery energy throughout the flight. This can be inferred by looking at the temporal geometric mean of the UAV altitude and normalized battery energy during the flight. Such a policy however, results in a poor wireless network capacity with an average of $1.48\mathcal{W}$ bit per second (bps) due to large scale fading. The behavior of this policy can be justified by noting that the choice of high penalty multipliers make energy dissipation the dominant part in the reward signal. Thus, to maximize the total rewards, the agent will attempt to minimize energy dissipation, which can be achieved by hovering above the cloud cover. On the other hand, the (RL Energy-Unaware) agent learns a policy which is indifferent to battery energy. In this policy, both UAVs descend and hover close to the minimum allowable altitude in order to maximize the achieved system capacity. Notice how the temporal geometric mean of the normalized battery energy attained by this policy is 0. This is because the geometric mean is proportional to the multiplication of the normalized battery energy at each time step, and so if at any time step the normalized battery energy is 0, so will be the geometric mean. The RLWS agent with $\eta_1 = 0, \eta_2 = 10$ behaves as expected, UAV 1 is indifferent to its energy dissipation, hovering close to the minimum allowable altitude to maximize wireless network capacity, where as UAV 2 hovers above the cloud cover to maintain high battery energy throughout the flight, thus minimizing energy dissipation. These experiments demonstrate that choosing the set of Lagrange penalty multipliers to obtain the desired results can be a tedious task. In contrast, the proposed CDRL agent automatically adapts the Lagrange penalty multipliers to maximize the rewards and satisfy the constraints. The policy learned by the proposed CDRL agent indeed strikes a balance; it ensures energy sustainability while slightly sacrificing the network capacity performance.

Policy Generalizability: To demonstrate the generalizability and robustness of the learned CDRL policy, we test its performance on networks with different initial states and massive number of IoT devices. The ability to transfer pre-trained models on small scale networks to larger scale networks is desirable because it is very promising for scalable network management, can amortize the cost of training the RL agent, and addresses sampling complexity issues which arise in larger systems. Recall that the CDRL policy has been trained given that 200 IoT devices are uniformly deployed on ground, and that the two UAVs are initially present at altitudes of 250m and 750m. In Figure 8(a), we demonstrate the learned policy performance given different initial altitudes of UAVs. Specifically, we consider two extreme cases: both UAVs are initially deployed at the altitude of 500m (case A), or at 1500m (case B). For those two cases, Figure 8(a) shows that the two UAVs tend to fully de-synchronize their vertical flight trajectories, such that when one of them is charging its battery at 1500m, the other is provisioning wireless service at 500m. The temporal average system capacity for cases A and B are $2.649\mathcal{W}$ bps and $2.619\mathcal{W}$ bps, respectively. Notice that case B achieves a slightly lower temporal average network capacity because both UAVs are initially farther away from IoT devices. In the legend, the temporal geometric mean of the battery energy of each UAV is reported to demonstrate energy sustainability of UAVs throughout the operating horizon.

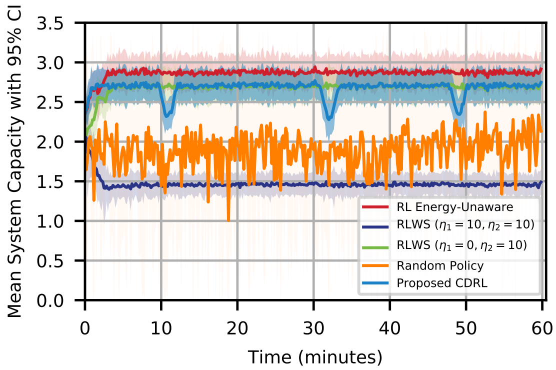


Fig. 7. Performance comparison of the proposed CDRL algorithm with baseline policies. The proposed CDRL agent learns an adaptive feasible policy which achieves a temporal average network capacity that is 81% higher than that of a feasible DRL agent with reward shaping (RLWS ($\eta_1 = \eta_2 = 10$)), and only 6.7% lower than the upper bound achieved by the RL Energy-Unaware agent.

In Figure 8(b), we test how the learned CDRL policy scales with varying number of IoT devices. As can be seen from Figure 8(b), the channel access probability is scaled appropriately given the number of IoT devices vary in $\{100, 200, 600, 1000\}$, maintaining comparable temporal average network capacity around $2.67\mathcal{W}$ bps. In addition, the temporal geometric mean of the battery energy of each UAV is also reported in the legend to demonstrate energy sustainability of UAVs. In Figure 8(c), we plot a boxplot of the achieved system capacity for massive deployments of IoT devices up to 10000 users. It can be seen that the learned policy network maintains high performance which is not compromised by increasing the number of IoT devices. Last but not least, in Figure 8(d), we test the performance of the learned policy given different horizontal deployment of the two UAVs. We consider three cases: (A) the two UAVs are deployed at $(250, 250)$m and $(750, 250)$m, as determined by the K-means clustering algorithm, (B) the two UAVs are deployed farthest from each other at $(0, 0)$m and $(1000, 500)$m, and (C) the two UAVs are randomly deployed on the xy-plane. It can be observed from Figure 8(d) that the mean network capacity is highest when the K-means algorithm is employed to determine the xy-planar deployment of the two UAVs. In addition, it is shown that randomly deploying the two UAVs in the xy-plane, case (C), achieves a mean network capacity which is slightly ($\approx 2\%$) lower than that in case (A), whereas the extreme case of deploying the two UAVs on the diagonal, case (B), achieves the lowest network capacity ($\approx 10\%$ lower compared

TABLE II
PERFORMANCE COMPARISON WITH BASELINE ALGORITHMS

	RLWS $\eta_1 = 0, \eta_2 = 0$	RLWS $\eta_1 = 0, \eta_2 = 10$	RLWS $\eta_1 = 10, \eta_2 = 10$	Random Policy	Proposed CDRL
$(\min_n\ z_1(t_n),\ \mathcal{G}_n[z_1(t_n)],\ \max_n\ z_1(t_n))$, m	$500, 504, 750$	$500, 509, 750$	$750, 1462, 1500$	$500, 803, 1500$	$500, 865, 1500$
$(\min_n\ z_2(t_n),\ \mathcal{G}_n[z_2(t_n)],\ \max_n\ z_2(t_n))$, m	$500, 520, 1250$	$1250, 1489, 1500$	$1250, 1494, 1500$	$500, 1053, 1500$	$500, 815, 1500$
$(\min_n\ B_1(t_n),\ \mathcal{G}_n[B_1(t_n)],\ \max_n\ B_1(t_n))$, %	$0.0, 0.0, 0.5$	$0.0, 0.0, 0.5$	$0.435, 0.902, 1.0$	$0.0, 0.134, 1.0$	$0.128, 0.529, 1.0$
$(\min_n\ B_2(t_n),\ \mathcal{G}_n[B_2(t_n)],\ \max_n\ B_2(t_n))$, %	$0.0, 0.0, 0.506$	$0.5, 0.942, 1.0$	$0.5, 0.942, 1.0$	$0.0, 0.548, 1.0$	$0.315, 0.686, 1.0$
$(\min_n\ \frac{\mathbb{G}(t_n)}{\mathcal{W}},\ \frac{\mathbb{E}_n[\mathbb{G}(t_n)]}{\mathcal{W}},\ \max_n\ \frac{\mathbb{G}(t_n)}{\mathcal{W}})$, bps	$2.35, 2.86, 2.94$	$2.08, 2.68, 2.75$	$1.40, 1.48, 2.11$	$1.0, 1.89, 2.37$	$2.28, 2.68, 2.78$
UAV 1 Energy Sustainable	No	No	Yes	No	Yes
UAV 2 Energy Sustainable	No	Yes	Yes	No	Yes

to that of case (A)). These results support our argument that the range of vertical flight is the dominant factor in our system model, and hence determining the planar locations of UAVs a priori reduces the number of control variables without adversely impacting network performance. In all cases, the learned policy still ensures energy sustainability of the two UAVs as indicated by the temporal geometric mean of the battery energy of each UAV, which is reported in the legend.

VI. CONCLUSION

In this paper, we study the joint problem of dynamic multi-UAV altitude control and random channel access management of a multi-cell UAV-based wireless network with NOMA, in support of a massive number of IoT devices. To enable an energy-sustainable capacity-optimal IoT network, we have formulated this constrained stochastic control problem as a constrained markov decision process, and proposed an online model-free constrained deep reinforcement learning algorithm to learn an optimal control policy for wireless network management. Our extensive simulations have demonstrated that the proposed algorithm outperforms baseline solutions, and learns a cooperative policy in which the altitude of UAVs and channel access probability of IoT devices are dynamically adapted to maximize the long-term total network capacity while ensuring energy sustainability of UAVs. In our future work, we will build a testbed in order to investigate data-driven control of wireless networks based on real hardware and real data.

ACKNOWLEDGMENT

This work is supported in part by the NSF grants ECCS1554576, ECCS-1610874, and CNS-1816908. This material is based in part upon work supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357. We gratefully acknowledge the computing resources provided on Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory.

REFERENCES

[1] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021,” Feb. 2017.

[2] S. Khairy, M. Han, L. X. Cai, Y. Cheng, and Z. Han, “A renewal theory based analytical model for multi-channel random access in IEEE 802.11 ac/ax,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 5, pp. 1000–1013, 2018.

[3] S. Khairy, M. Han, L. X. Cai, and Y. Cheng, “Sustainable wireless IoT networks with RF energy charging over Wi-Fi (CoWiFi),” *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 205–10 218, 2019.

[4] C. El Fehri, M. Kassab, S. Abdellatif, P. Berthou, and A. Belghith, “LoRa technology MAC layer operations and research issues,” *Procedia computer science*, vol. 130, pp. 1096–1101, 2018.

[5] “FCC opens 6 GHz band to Wi-Fi and other unlicensed uses,” Apr 2020. [Online]. Available: https://www.fcc.gov/document/fcc-opens-6-ghz-band-wi-fi-and-other-unlicensed-uses

[6] J. Choi, “NOMA-based random access with multichannel ALOHA,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2736–2743, 2017.

[7] J.-B. Seo, B. C. Jung, and H. Jin, “Nonorthogonal random access for 5G mobile communication systems,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7867–7871, 2018.

[8] J. Choi, “A game-theoretic approach for NOMA-ALOHA,” in *2018 European Conference on Networks and Communications (EuCNC)*. IEEE, 2018.

[9] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, “Placement optimization of UAV-mounted mobile base stations,” *IEEE Communications Letters*, vol. 21, no. 3, pp. 604–607, 2016.

[10] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage,” *IEEE Communications Letters*, vol. 20, no. 8, pp. 1647–1650, 2016.

[11] H. Zhao, H. Wang, W. Wu, and J. Wei, “Deployment algorithms for UAV airborne networks toward on-demand coverage,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2015–2031, 2018.

[12] A. V. Savkin and H. Huang, “A method for optimized deployment of a network of surveillance aerial drones,” *IEEE Systems Journal*, vol. 13, no. 4, pp. 4474–4477, 2019.

[13] Q. Wu, Y. Zeng, and R. Zhang, “Joint trajectory and communication design for multi-UAV enabled wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, 2018.

[14] Y. Xu, L. Xiao, D. Yang, Q. Wu, and L. Cuthbert, “Throughput maximization in multi-UAV enabled communication systems with difference consideration,” *IEEE Access*, vol. 6, pp. 55 291–55 301, 2018.

[15] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, “Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks,” *arXiv preprint arXiv:1905.02993*, 2019.

[16] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, “Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, 2019.

[17] J. Hu, H. Zhang, and L. Song, “Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol,” *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6177–6189, 2018.

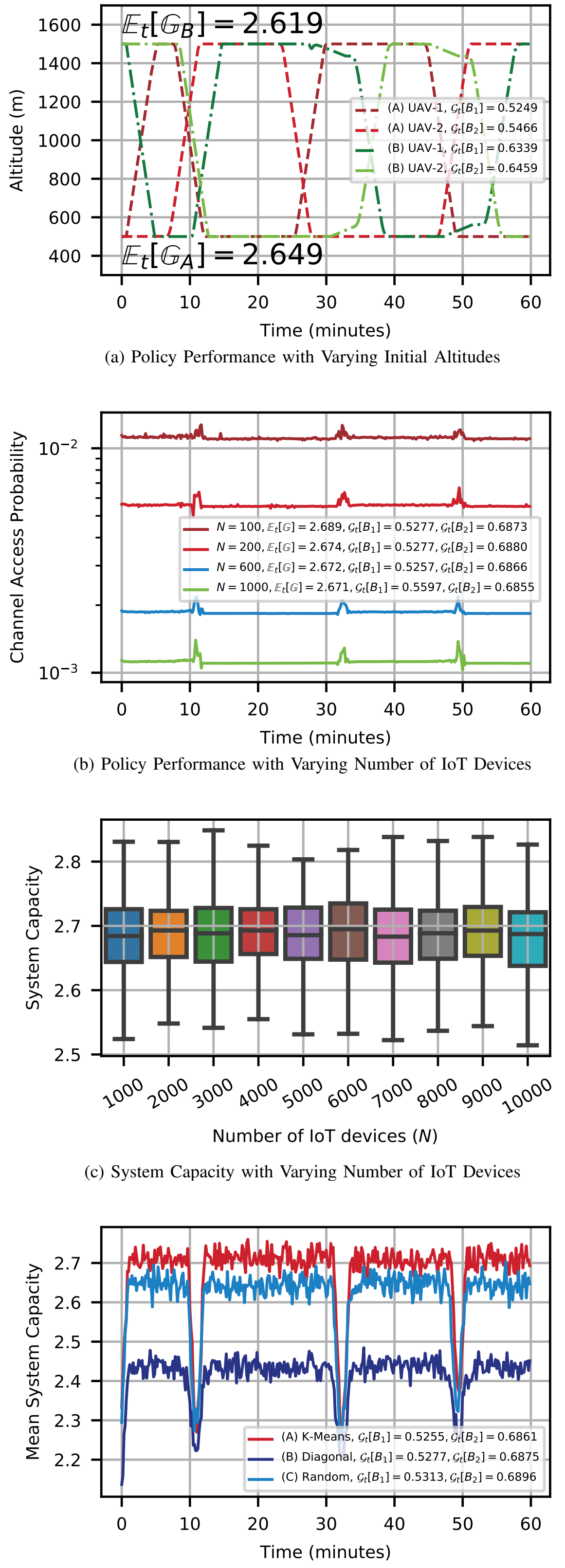
[18] Z. Yuan, J. Jin, L. Sun, K.-W. Chin, and G.-M. Muntean, “Ultra-reliable IoT communications with UAVs: A swarm use case,” *IEEE Communications Magazine*, vol. 56, no. 12, pp. 90–96, 2018.

[19] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, “A tutorial on UAVs for wireless networks: Applications, challenges, and open problems,” *IEEE communications surveys & tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

[20] S. Morton, R. D’Sa, and N. Papanikolopoulos, “Solar powered UAV: Design and experiments,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[21] P. Oettershagen, A. Melzer, T. Mantel, K. Rudin, T. Stastny, B. Wawrzacz, T. Hinzmann, K. Alexis, and R. Siegwart, “Perpetual flight

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications



(a) Policy Performance with Varying Initial Altitudes

(b) Policy Performance with Varying Number of IoT Devices

(c) System Capacity with Varying Number of IoT Devices

(d) Policy Performance with Different (x^m, y^m) Selection Schemes

Fig. 8. Learned policy generalizability with varying initial states and network scale. The proposed CDRL agent has been train on a network deployment of $N = 200$ IoT devices, yet the learned policy can efficiently manage networks with with different initial states and massive number of IoT devices.

with a small solar-powered UAV: Flight results, performance analysis and model validation," in *2016 IEEE Aerospace Conference*, 2016.

[22] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015.

[23] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000.

[24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[25] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, and M.-S. Alouini, "Joint trajectory and precoding optimization for UAV-assisted NOMA networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3723–3735, 2019.

[26] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "Uav-enabled communication using noma," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 5126–5138, 2019.

[27] Z. Chen, Y. Liu, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, "Optimizing non-orthogonal multiple access in random access networks," in *2020 IEEE 91st Vehicular Technology Conference (VTC-Spring)*, 2020.

[28] C. Goerzen, Z. Kong, and B. Mettler, "A survey of motion planning algorithms from the perspective of autonomous UAV guidance," *Journal of Intelligent and Robotic Systems*, vol. 57, no. 1-4, p. 65, 2010.

[29] M. Bangura and R. Mahony, "Thrust control for multirotor aerial vehicles," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 390–405, 2017.

[30] G. Hoffmann, H. Huang, S. Waslander, and C. Tomlin, "Quadrotor helicopter flight dynamics and control: Theory and experiment," in *AIAA guidance, navigation and control conference and exhibit*, 2007.

[31] J.-S. Lee and K.-H. Yu, "Optimal path planning of solar-powered UAV using gravitational potential energy," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 3, pp. 1442–1451, 2017.

[32] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4281–4298, 2019.

[33] S. Mannor and N. Shimkin, "A geometric approach to multi-criterion reinforcement learning," *Journal of machine learning research*, vol. 5, no. Apr, pp. 325–360, 2004.

[34] K. Van Moffaert and A. Nowé, "Multi-objective reinforcement learning using sets of pareto dominating policies," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3483–3512, 2014.

[35] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[36] S. Bhatnagar and K. Lakshmanan, "An online actor–critic algorithm with function approximation for constrained markov decision processes," *Journal of Optimization Theory and Applications*, vol. 153, no. 3, pp. 688–708, 2012.

[37] S. Bhatnagar, "An actor–critic algorithm with function approximation for discounted cost constrained markov decision processes," *Systems & Control Letters*, vol. 59, no. 12, pp. 760–766, 2010.

[38] V. S. Borkar, "An actor-critic algorithm for constrained markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[39] P. Geibel and F. Wysotzki, "Learning algorithms for discounted mdps with constraints," *International Journal of Mathematics, Game Theory, and Algebra*, vol. 21, no. 2/3, p. 241, 2012.

[40] P. Geibel and F. Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence Research*, vol. 24, pp. 81–108, 2005.

[41] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.

[42] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv preprint arXiv:1802.06480*, 2018.

[43] M. Fu *et al.*, "Risk-sensitive reinforcement learning: A constrained optimization viewpoint," *arXiv preprint arXiv:1810.09126*, 2018.

[44] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[45] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2020.3018804, IEEE Journal on Selected Areas in Communications

[47] L. Kallenberg, "Markov decision processes," *Lecture Notes. University of Leiden*, 2011.

[48] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[49] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," 2014.

[50] 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," *arXiv preprint arXiv:1509.02971*, 2015.

[51] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.

[52] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016.

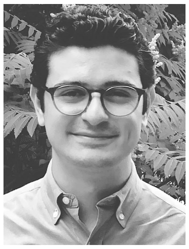
[53] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[54] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[55] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[57] "OpenAI spinning up in deep RL repository," https://github.com/openai/spinningup/tree/master/spinup/algos/ppo/, [Online; accessed January 15, 2020].



Sami Khairy (S'16) received the B.S. degree in Computer Engineering from the University of Jordan, Amman, Jordan, in 2014 and the M.S. degree in Electrical Engineering from Illinois Institute of Technology, Chicago, IL, USA, in 2016. He is currently working towards the Ph.D. degree in Electrical Engineering at Illinois Institute of Technology. His research interests span the broad areas of analysis and protocol design for next generation wireless networks, AI powered wireless networks resource management, reinforcement learning, statistical learning, and statistical signal processing. He received a Fulbright Predoctoral Scholarship from JACEE and the U.S. Department of State in 2015, and the Starr/Fieldhouse Research Fellowship from IIT in 2019. He is an IEEE student member and a member of IEEE ComSoc and IEEE HKN.

Prasanna Balaprakash is a computer scientist at the Mathematics and Computer Science Division with a joint appointment in the Leadership Computing Facility at Argonne National Laboratory. His research interests span the areas of artificial intelligence, machine learning, optimization, and high-performance computing. Currently, his research focuses on the development of scalable, data-efficient machine learning methods for scientific applications. He is a recipient of U.S. Department of Energy 2018 Early Career Award. He is the machine-learning team lead and data-understanding team co-lead in RAPIDS, the SciDAC Computer Science institute. Prior to Argonne, he worked as a Chief Technology Officer at Mentis Sprl, a machine learning startup in Brussels, Belgium. He received his PhD from CoDE-IRIDIA (AI Lab), Université Libre de Bruxelles, Brussels, Belgium, where he was a recipient of Marie Curie and F.R.S-FNRS Aspirant fellowships.

Lin X. Cai (S'09–M'11–SM'19) received the M.A.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the University of Waterloo, Waterloo, Canada, in 2005 and 2010, respectively. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, Illinois, USA. Her research interests include green communication and networking, intelligent radio resource management, and wireless Internet of Things. She received a Postdoctoral Fellowship Award from the Natural Sciences and Engineering Research Council of Canada (NSERC) in 2010, a Best Paper Award from the IEEE Globecom 2011, an NSF Career Award in 2016, and the IIT Sigma Xi Research Award in the Junior Faculty Division in 2019. She is an Associated Editor of IEEE Transaction on Wireless Communications, IEEE Network Magazine, and a co-chair for IEEE conferences.

Yu Cheng (S'01–M'04–SM'09) received the B.E. and M.E. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2003. He is currently a Full Professor with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA. His current research interests include wireless network performance analysis, network security, big data, cloud computing, and machine learning. Dr. Cheng was a recipient of the Best Paper Award at QShine 2007, the IEEE ICC 2011, the Runner-Up Best Paper Award at ACM MobiHoc 2014, the National Science Foundation CAREER Award in 2011, and the IIT Sigma Xi Research Award in the Junior Faculty Division in 2013. He has served as several Symposium Co-Chairs for IEEE ICC and IEEE GLOBECOM, and the Technical Program Committee Co-Chair for WASA 2011 and ICNC 2015. He was a founding Vice Chair of the IEEE ComSoc Technical Subcommittee on Green Communications and Computing. He was an IEEE ComSoc Distinguished Lecturer from 2016 to 2017. He is an Associate Editor for the IEEE Transactions on Vehicular Technology, IEEE Internet of Things Journal, and IEEE Wireless Communications.