A Multi-Agent System Toward the Green Edge Computing with Microgrid

Md. Shirajum Munir[1], Sarder Fakhrul Abedin[1], Do Hyeon Kim[1], Nguyen H. Tran[2], Zhu Han[1,3], and Choong Seon Hong[1]

[1]*Department of Computer Science and Engineering, Kyung Hee University, Yongin-si 17104, Republic of Korea*

[2]*School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia*

[3]*Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004-4005 USA*

E-mail:{munir, saab0015, doma}@khu.ac.kr, nguyen.tran@sydney.edu.au, zhan2@uh.edu, cshong@khu.ac.kr

***Abstract*—The nature of multi-access edge computing (MEC) is to deal with heterogeneous computational tasks near to the end users, which induces the volatile energy consumption for the MEC network. As an energy supplier, a microgrid is able to enable seamless energy flow from renewable and non-renewable sources. In particular, the risk of energy demand and supply is increased due to nondeterministic nature of both energy consumption and generation. In this paper, we impose a risk-sensitive energy profiling problem for a microgrid-enabled MEC network, where we first formulate an optimization problem by considering Conditional Value-at-Risk (CVaR). Hence, the formulated problem can determine the risk of expected energy shortfall by coordinating with the uncertainties of both demand and supply, and we show this problem is NP-hard. Second, we design a multi-agent system that can determine a risk-sensitive energy profiling by coping with an optimal scheduling policy among the agents. Third, we devise the solution by applying a multi-agent deep reinforcement learning (MADRL) based on asynchronous advantage actor-critic (A3C) algorithm with shared neural networks. This approach mitigates the curse of dimensionality for state space and also, can admit the best energy profile policy among the agents. Finally, the experimental results establish the significant performance gain of the proposed model than that a single agent solution and achieves a high accuracy energy profiling with respect to risk constraint.**

***Index Terms*—green edge computing, microgrid, renewable energy, multi-agent deep reinforcement learning, conditional value-at-risk, energy profiling.**

I. INTRODUCTION

In recent years, the fifth-generation (5G) network is emerging with the expansion of multi-access edge computing (MEC) infrastructures [1], where the deployment of MEC hosts are flexible. MEC can be hosted anywhere between the edge and central data network based on technical and business requirements [2]. The goal of this expansion is to enable the low-latency, high-bandwidth communication with computational decision making feed-back of numerous IoT services and applications (e.g., smart city, autonomous vehicles, health care, industrial IoT, smart energy, and so on) [1], [2]. Energy consumption of the MEC network is not only increasing, but also random over time [3]. Meanwhile, a microgrid can be a prominent candidate to support this additional amount of energy from its renewable (e.g., solar, biofuels, wind, and so on.) and non-renewable (e.g., coal power, diesel generator, etc.) sources [4], [5]. Thus, the renewable energy generation of the microgrid/grid is also nondeterministic in nature. Hence, the uncertainties of both energy consumption and generation impose the risk of energy shortage of the MEC network.

Recently, a renewable energy aware tasks scheduling and resource allocation approach of the wireless networks has been proposed in [4]. A joint method has proposed for both base station (BS) operation and power distribution, which provides a strong relationship between BS operation power consumption and smart grid power generation [5]. Further, authors of [3] have considered a data-driven approach to manage energy demand-supply of a microgrid powered MEC networks toward the energy saving. However, literature does not focus on an approach that can assess the risk of energy shortfall for renewable energy enabled wireless networks. In this paper, we introduce a *risk-sensitive energy profiling* problem for microgrid-powered MEC network, where we incorporate a Conditional Value-at-Risk (CVaR) [6]. The CVaR captures a tail of expected return for the risk measurement of both energy consumption and generation toward the sustainable MEC network. In order to achieve this, we face several challenges:

- First, how to coordinate the risk assessment between MEC network energy consumption and renewable generation of the microgrid, where both of them are random over time. Thus, historical observations can be a possible way that can meet the CVaR confidence.
- Second, how to cope with historical observations (high dimensional energy consumption and generation data), where these data establish strong temporal dependencies (autocorrelation) among them. Markov Decision Process (MDP) can be discretized this autocorrelation.
- Third, even the Markovian properties can deal with past observations toward the future decision. However, it is hard to manage when the state space is very large.

To address the aforementioned challenges, we summarize our main contributions as follows:

- First, we formulate the *risk-sensitive energy profiling*

The research is partially supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01287, Evolvable Deep Learning Model Generation Platform for Edge Computing), US MURI AFOSR MURI 18RT0073, NSF CNS-1717454, CNS- 1731424, CNS-1702850, CNS-1646607. Dr. CS Hong is the corresponding author.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

problem of the microgrid powered MEC network, where the objective is to minimize the loss of expected energy shortfall that satisfies CVaR confidence range. We convert this problem as a discounted reward maximization problem to discretize Markovian properties of energy demand-supply.
- Second, we devise a multi-agent system, where we solve the discounted reward maximization problem under the CVaR risk assessment, and we propose a multi-agent based risk-sensitive energy profiling algorithm.
- Finally, we perform a rigorous experimental analysis of the proposed multi-agent system, where we show MADRL model outperforms single agent solution in terms of energy profiling decision. The proposed model gains up to 94.5% forecasting accuracy with CVaR 5.65% for 95% CVaR confidence.

The rest of the paper is organized as follows: we present the system model and problem formulation of the risk-sensitive energy profiling in Section II. In Section III, we represent the risk-sensitive energy profiling solution with a multi-agent system. We discuss the experimental analysis in Section IV. Finally, we conclude our discussion in Section V.

II. System Model and Problem Formulation

A. System Model

Let us consider a microgrid-powered wireless network that includes both renewable and non-renewable energy sources, as seen in Fig. 1. We consider a set $\mathcal{B} = \{1, 2, \ldots, B\}$ of MEC-enabled small-cell base stations (SBSs) is controlled by a macro base station (MBS), and each SBS i enables a set $\mathcal{C}_i = \{1, 2, \ldots, C_i\}$ of active MEC servers with the heterogeneous computational capacity. Hence, we consider a time slot t with 15 minutes duration, which is in a finite time horizon $\mathcal{T} = 1, 2, \ldots, T$ [7]. SBS i can serve a set $\mathcal{K} = \{1, 2, \ldots, K\}$ of heterogeneous user tasks (e.g., video surveillance, emergency health care, smart transportation and so on.) during the time slot t. Tasks are associated with SBS i by a given user task association indicator $\Omega_{ik}(t)$, where $\Omega_{ik}(t) = 1$ if task k is assigned to SBS i at time t, 0 otherwise. Further, a microgrid controller is physically connected with MBS to control the energy demand-response (DR) of the MEC network.

1) Communication Model: Let us consider a task arrival rate $\lambda_i(t)$ of SBS i follows Poisson process at time slot t and average traffic load is determined by $\lambda_i(t)\chi_i(t)$ with an average traffic size $\chi_i(t)$. The uplink transmission data rate for SBS i is defined as follows [3]:

$$\varphi_i(t) = \sum_{k \in \mathcal{K}} \Omega_{ik}(t) w_{ik} \log_2 \left(1 + \frac{p_{ik} g_{ik}(t)}{\sigma_{ik}^2 + \sum_{j \in \mathcal{B}, j \neq i} I_j(t)}\right), \forall i \tag{1}$$

where w_{ik} is a fixed channel bandwidth, p_{ik} represents the transmission power, $g_{ik}(t)$ determines the channel gain, channel noise is denoted by σ_{ik}^2, and $I_j(t)$ is the channel interference. We consider $\mu_i(t) = \frac{\varphi_i(t)}{\chi_i(t)}$ is a service rate of user task execution at SBS i. We assume a set of heterogeneous tasks $\mathcal{K}$ are uniformly distributed under the SBS i at time slot t. Thus, the overall server utilization rate is defined as follows [5]:

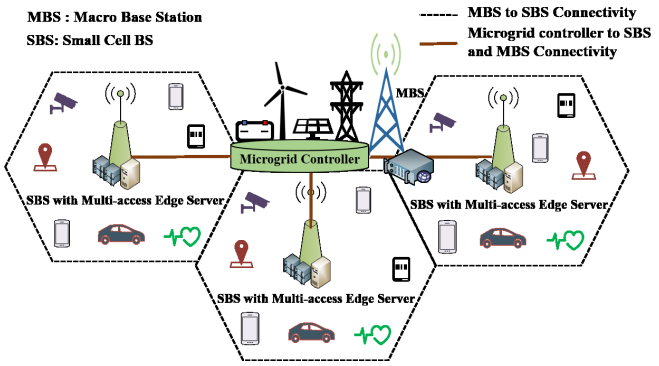

Fig. 1. Edge computing system model with microgrid.

$$\rho_i(t) = \sum_{k \in \mathcal{K}} \Omega_{ik}(t) \frac{\lambda_i(t)}{\mu_i(t)}, \tag{2}$$

where $\sum_{k \in \mathcal{K}} \Omega_{ik}(t)\lambda_i(t)$ determines the total served tasks at time slot t under the SBS i.

2) Energy Consumption Model:

SBS Network Operation Energy Consumption: Network operation energy consumption of SBS i depends on network communication and data transfer through the network, where network communication consumption is considered as a static energy consumption $\mathcal{E}_{st}^{net}(t)$ [8]. Thus, network operation energy consumption of SBS i is defined as follows:

$$\mathcal{E}_i^{net}(t) = \sum_{k \in \mathcal{K}} \left(\eta^{net}(t) \frac{p_{ik}\chi_i(t)}{\varphi_i(t)} + \mathcal{E}_{st}^{net}(t)\right), \tag{3}$$

where η^{net} determines an energy coefficient of data transfer through the network and value of η^{net} depends on types of network devices [8].

MEC Server Computational Energy Consumption: Computational energy consumption of the MEC server relies on the number of CPU cores, numbers of CPU components, CPU activity ratio, and also with the processor architecture [9]. We consider L number of CPU cores with M number of components, where activity ratio of the CPU is determined by $\delta_c(t)$. Other general operation of MEC server consumes static energy $\mathcal{E}_{st}^{mec}(t)$ and MEC servers energy consumption at SBS i is defined as follows [9]:

$$\mathcal{E}_i^{mec}(t) = \sum_{c \in \mathcal{C}_i} \left(\left(\sum_{l \in L} \sum_{m \in M} \eta^{cpu} \delta_c(t) \right) + \mathcal{E}_{st}^{mec}(t) \right), \tag{4}$$

where $\eta^{cpu}\delta_c(t)$ represents dynamic energy consumption and η^{cpu} determines an energy coefficient [9] for CPU usage at time slot t. Moreover, the relationship between CPU activity ratio $\delta_c(t)$ and server utilization $\rho_i(t)$ is, $\delta_c(t) \approx \frac{1}{\rho_i(t)}$.

Total Energy Demand: Energy consumption of MEC-enabled SBS i consists of two type of energy consumption, one is a static energy $\mathcal{E}_i^{st}(t)$ and another is a dynamic energy $\mathcal{E}_i^{dyn}(t)$ [8]. Using (3) and (4), we can capture the

dynamic energy consumption of SBS and the dynamic energy consumption of SBS i at time slot t is determined as follows:

$$\mathcal{E}_i^{dyn}(t) = \mathcal{E}_i^{mec}(t) + \mathcal{E}_i^{net}(t). \tag{5}$$

Thus, the total energy demand of all SBSs B under the MBS is defined as follows:

$$\mathcal{E}^{dem}(t) = \sum_{\forall i \in \mathcal{B}} \big(\mathcal{E}_i^{dyn}(t)\eta^{dyn}(t) + \mathcal{E}_i^{st}(t)\big), \tag{6}$$

where $\eta^{dyn}(t)$ represents an energy coefficient for dynamic energy consumption and $\mathcal{E}_i^{st}(t)$ determines idle operation [8] energy consumption of SBS i at time slot t.

3) Microgrid Energy Generation: Microgrid consists of two distinct energy sources: renewable (e.g., solar, biofuels, wind, and so on) and non-renewable energy (e.g., coal, diesel etc.). $\mathcal{E}^{ren}(t)$ and $\mathcal{E}^{non}(t)$ denote the amount of renewable and non-renewable energy generation at time slot t, respectively. The total amount of energy generation $\mathcal{E}^{tot}(t)$ at time slot t is defined as follows:

$$\mathcal{E}^{tot}(t) = \mathcal{E}^{ren}(t) + \mathcal{E}^{non}(t). \tag{7}$$

Microgrid is capable of buying additional energy from a main grid to fulfill an extra demand $\mathcal{E}^{buy}(t) = \mathcal{E}^{dem}(t) - \mathcal{E}^{tot}(t)$ of the MEC network [7] and also can store a surplus amount of generated energy $\mathcal{E}^{sto}(t) = \mathcal{E}^{tot}(t) - \mathcal{E}^{dem}(t)$ at time slot t for the future usages.

B. Problem Formulation

1) Risk Assessment with Conditional Value-at-Risk: We consider a function $\Upsilon(\mathbf{x}, \mathbf{z})$, where loss associated with a $q = 2$ dimensional decision (i.e., buy or store) vector $\mathbf{x} \in \mathbb{R}^q$ to determine energy buying and store decision. A set of decision vectors $\mathbf{X}$ represents the available energy profiles such that $\mathbf{x} \in \mathbf{X}$ and a $p = 4$ dimensional random vector $\mathbf{z} \in \mathbb{R}^p$, which stands for uncertainties (i.e., energy demand $\mathcal{E}^{dem}(t)$, renewable energy generation $\mathcal{E}^{ren}(t)$, non-renewable energy generation $\mathcal{E}^{non}(t)$ and storage energy $\mathcal{E}^{sto}(t)$) that affects on the loss at time slot t. Therefore, the loss function is defined as follows:

$$\Upsilon(\mathbf{x}, \mathbf{z}) = \min_{\mathbf{x} \in \mathbf{X}} \mathbb{E}\big[\sum_{\mathbf{x} \in \mathbf{X}} \sqrt{\big(\mathcal{E}^{dem}(t) - \mathcal{E}^{tot}(t)\big)^2}\big]. \tag{8}$$

For each $\mathbf{x} \in \mathbb{R}^q$, the loss $\Upsilon(\mathbf{x}, \mathbf{z})$ is a random variable, where $P(\Upsilon(\mathbf{x}, \mathbf{z}))$ is a probability distribution for $\mathbf{z} \in \mathbb{R}^p$. For a given CVaR confidence level ξ, the probability distribution of $\Upsilon(\mathbf{x}, \mathbf{z})$ is denoted by $\psi(\mathbf{x}, \xi)$ and the distribution of $\Upsilon(\mathbf{x}, \mathbf{z})$ always satisfy the CVaR confidence ξ, where the loss $\Upsilon(\mathbf{x}, \mathbf{z})$ is inversely proportional to CVaR confidence ξ. Thus, for any specified probability level $\alpha \in (0, 1)$, Value-at-Risk (VaR) $\xi_\alpha(\mathbf{x})$ and CVaR $\Phi_\alpha(\mathbf{x})$ are associated with random variable $\mathbf{x}$. VaR $\xi_\alpha(\mathbf{x})$ of the energy profile is defined as follows [6]:

$$\xi_\alpha(\mathbf{x}) = \arg\min_{\xi \in \mathbb{R}} \psi(\mathbf{x}, \xi) \geq \alpha, \tag{9}$$

where $\xi_\alpha(\mathbf{x})$ determines the value ξ such that $\psi(\mathbf{x}, \xi) = \alpha$ and CVaR $\Phi_\alpha(\mathbf{x})$ is defined as follows [6]:

$$\Phi_\alpha(\mathbf{x}) = \min_{\xi \in \mathbb{R}} \frac{1}{(1-\alpha)} \mathbb{E}_{\Upsilon(\mathbf{x}, \mathbf{z}) \geq \xi_\alpha(\mathbf{x})}\big[\Upsilon(\mathbf{x}, \mathbf{z})\big]. \tag{10}$$

Here, the probability $P(\Upsilon(\mathbf{x}, \mathbf{z})) \geq P(\xi_\alpha(\mathbf{x}))$ is equal to $1 - \alpha$. Thus, $\Phi_\alpha(\mathbf{x})$ is a conditional expectation and the loss is associated with random variable $\mathbf{x}$ (i.e., energy store and buying decision), where the loss is $\xi_\alpha(\mathbf{x})$ or greater than that.

2) Modeling with Multi-Agent Deep Reinforcement Learning: Let us consider a multi-agent reinforcement learning setting with a set of agents $\mathcal{N} = \{1, 2, \ldots, N\}$, which contains a set of states $\mathcal{S} = \{1, 2, \ldots, S\}$, a set of actions $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_N\}$, and a set of observations $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_N\}$ [10]. The state-space consist of a four elements tuple $s_t \colon \langle \mathcal{E}^{dem}(t), \mathcal{E}^{ren}(t)\mathcal{E}^{sto}(t), \alpha \rangle \in \mathcal{S}$ (i.e., demand, renewable, stored energy, and maximum risk tolerance) at time slot t. The action space $a_t \in \mathcal{A}$ contains a binary decision variable and is defined as,

$$a_t = \begin{cases} 1, \text{ if } \mathcal{E}^{sto}(t) \geq \mathcal{E}^{buy}(t), \forall t \in \mathcal{T}, \\ 0, \qquad \text{otherwise}, \end{cases} \tag{11}$$

where $a_t = 1$ if microgrid is capable of fulfilling energy demand from its own sources, and 0 otherwise.

Actions for each agent $n \in \mathcal{N}$ with the parameter θ_n is determined by a stochastic policy π_{θ_n}, where $\pi_{\theta_n} \colon \mathcal{O}_n \times \mathcal{A}_n \in [0, 1]$. A state transition function $\Gamma \colon \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2, \cdots \times \mathcal{A}_N \in \mathcal{S}$ determines the next state $s_{t'}$ according to policy π_{θ_n}. Each agent n determines reward as a function of state and its action $r_n \colon \mathcal{S} \times \mathcal{A}_n \in \mathbb{R}$. Observation of agent n is defined as, $o_n \colon \langle s_t, a_t, r_t, s_{t'} \rangle$ that correlates with state space $\mathcal{S}$ such that $o_n \colon \mathcal{S} \in \mathcal{O}_n$. Agents follow a discrete time slot $t \in \{1, 2, \ldots, T\}$ and the objective of each agent n is to maximize the total expected reward. The individual rewards for each agent n is defined as follows [10]:

$$r_n(a_t, s_t) = \max_{r_{t'}} \mathbb{E}\big[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}(a_{t'}, s_{t'})\big], \tag{12}$$

where $\gamma \in (0, 1)$ is a discount factor. The expectation of action value function for agent n from taking action a_t in state s_t is defined as follows:

$$Q^{\pi_{\theta_n}}(s_t, a_t) = \mathbb{E}_{\pi_{\theta_n}}\big[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}(a_{t'}, s_{t'}) | s_t, a_t\big], \tag{13}$$

where $\gamma^{t'-t}$ ensures the convergence of action value function $Q^{\pi_{\theta_n}}(s_t, a_t)$ estimation. Therefore, the expectation of state value function for agent n is given as follows:

$$V^{\pi_{\theta_n}}(s_t) = \mathbb{E}_{a_t \sim \pi_{\theta_n}(a_t | s_t)}\big[Q^{\pi_{\theta_n}}(s_t, a_t)\big], \tag{14}$$

where the action value function $Q^{\pi_{\theta_n}}(s_t, a_t)$ is determined by (13). Here, we can rewrite the loss function (8) as a form of constraint with a discounted reward maximization problem

[6]. The objective is to maximize the reward (state value) and the problem formulation is as follows:

$$\max_{\pi_{\theta_n}, a_t \in \mathcal{A}} \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{C}_i}\sum_{k\in\mathcal{K}} \Omega_{ik}(t) V^{\pi_{\theta_n}}(s_t), \tag{15}$$

$$\text{s.t.} \quad P(\Upsilon(a_t, s_t)) \leq \alpha, t \in \mathcal{T}, \tag{15a}$$

$$a_t \mathcal{E}^{tot}(t) + (1 - a_t)\mathcal{E}^{dem}(t) \geq 0, t \in \mathcal{T}, \tag{15b}$$

$$a_t \mathcal{E}^{sto}(t) + (1 - a_t)\mathcal{E}^{buy}(t) \geq 0, t \in \mathcal{T}, \tag{15c}$$

$$\sum_{t\in\mathcal{T}} a_t \leq T, \tag{15d}$$

$$a_t \in \{0, 1\}, \forall t \in \mathcal{T}. \tag{15e}$$

In problem (15), π_{θ_n} and a_t are decision variables, where π_{θ_n} determines the energy profiling policy and a_t decides energy storing or buying decision at time slot t. Constraint (15a) in problem (15) ensures the CVaR for a given state s_t and action a_t, where α contains maximum tolerable loss and the value of α is very small. Constraint (15b) assures a coordination between MEC network energy consumption and microgrid generation at time $t \in \mathcal{T}$. Here, constraint (15c) provides a guarantee for fulfilling the energy demand to the network. Finally, constraints (15d) and (15e) ensure the energy store/buy decision a_t is a binary variable.

Since the decision variables π_{θ_n} (policy) and a_t (action) extend the problem (15) to a mixed-integer programming problem with the corresponding constraints in (15a) - (15e). The problem (15) can be reduced to a 0/1 multiple-knapsack as base problem [11], which is NP-Complete [12] in general. Similar to the 0/1 multiple-knapsack problem, (15) holds combinatorial nature, which can be devised a feasible energy profiling of the MEC network. However, the complexity of problem (15) leads to exponentially $O(2^{T\times B\times C\times K})$, where there is no known algorithm that can solve (15) in polynomial time, whether it is optimal. As a result, we can infer that energy profiling decision of (15) belongs to the same category of multiple-knapsack problem, which is proven to be NP-hard [11]. Thus, we design a multi-agent system to solve (15) and detailed discussion of the risk-sensitive energy profiling solution is explained in the following section.

III. Risk-Sensitive Energy Profiling Solution with Multi-Agent System

We devise a solution of the problem (15) using a MADRL-based Asynchronous A3C approach (as seen in Fig. 2) [13]. Let us consider parameters θ_t, where state action function (13), state value function (14), and parameterized policy are defined as, $Q^{\pi_{\theta_n}}(s_t, a_t) \approx Q^{\pi_{\theta_n}}(s_t, a_t; \theta_t)$, $V^{\pi_{\theta_n}}(s_t) \approx V^{\pi_{\theta_n}}(s_t; \theta_t)$, and $\pi_{\theta_n}(a_t|s_t) \approx \pi_{\theta_n}(a_t|s_t; \theta_t)$, respectively. In order to find optimal policy π_{θ_n}, we learn the action value function $Q^{\pi_{\theta_n}}(s_t, a_t)$, where we use deep Q-networks (DQN) [13]. Therefore, the loss function is defined as follows:

$$\mathcal{L}(\theta_n) = \mathbb{E}_{o_n\in\mathcal{O}_n}\left[(Q^{\pi_{\theta_n}}(s_t, a_t|\theta_t) - y_t)^2\right], \tag{16}$$

where $y_t = r_n(a_t, s_t) + \gamma \max_{a_{t'}\in\mathcal{A}} \hat{Q}^{\pi_{\theta_n}}(s_{t'}, a_{t'})$ is an ideal target. Here, $o_n : \langle s_t, a_t, r_t, s_{t'}\rangle$ represents an observational tuple for

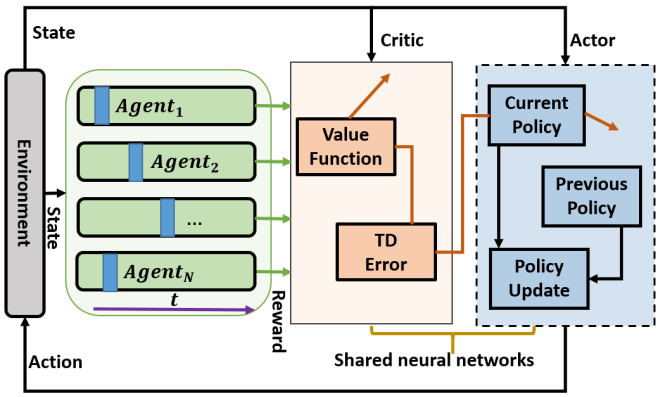

Fig. 2. Risk-sensitive energy profiling model based on multi agent asynchronous advantage actor-critic (A3C) algorithm with shared neural networks.

current state $s_{t'}$ and value $\hat{Q}^{\pi_{\theta_n}}(s_{t'}, a_{t'})$. Hence, parameters θ_n are periodically updated with the recent values that ensures the stability of DQN learning process. In MADRL settings, each agent n updates its own policy π_{θ_n} independently, where this non-stationary nature violates the convergence characteristics of the learning process. Thus, observations from the experience cannot be used for general environment settings. To overcome this challenge, we use a policy gradient method that can adjust parameters θ_n directly to determine a policy π_{θ_n} and enables maximum expected reward $\mathbb{E}[r_n]$. Therefore, we consider a set of policies $\pi = \{\pi_{\theta_1}, \pi_{\theta_2}, \ldots, \pi_{\theta_N}\}$ with parameters set $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ of N agents and expected return for policy π_{θ_n} is defined as follows:

$$\mathcal{J}(\theta_n) = \max_{\pi_{\theta_n}} \mathbb{E}[r_n]. \tag{17}$$

Here, the gradient of expected return (17) is determined as follows:

$$\nabla_{\theta_n}\mathcal{J}(\theta_n) = \mathbb{E}\left[\nabla_{\theta_n} \log \pi_{\theta_n}(a_n|o_n) Q_n^{\pi}(\mathcal{O}, a_1, \ldots, a_N)\right], \tag{18}$$

where, $Q_n^{\pi}(\mathcal{O}, a_1, \ldots, a_N)$ is a centralized action-value function, $a_1, \ldots, a_N$ determine the actions of all agents N, and $\mathcal{O}$ represents all the observations $\mathcal{O} = \{o_1, \ldots, o_N\}$ of N agents. The gradient (18) generates high bias and lower variance due to deterministic observations. Hence, to overcome this, we consider N continuous policies ϑ_{θ_n} with respect to parameters θ_n and a policy gradient function can be presented as follows:

$$\nabla_{\theta_n}\mathcal{J}(\theta_n) = \mathbb{E}_{O,a\sim\mathcal{M}_n} \left[\nabla_{\theta_n}\vartheta_n(a_n|o_n)\nabla_{a_n} Q_n^{\vartheta}(O, a_1, \ldots, a_N)|_{a_n=\vartheta_n(o_n)}\right], \tag{19}$$

where $\mathcal{M}_n$ represents experiences for all agents $(\mathcal{O}, \mathcal{O}', a_1, \ldots, a_N, r_1, \ldots, r_N)$ that includes both previous $\mathcal{O}$ and current $\mathcal{O}'$ observations. Centralized action value function for all agents can be represented as follows [10]:

$$\mathcal{L}(\theta_n) = \min_{\vartheta} \mathbb{E}_{o_n, a_n}\left[(Q_n^{\vartheta}(O, a_1, \ldots, a_N) - y)^2\right], \tag{20}$$

where $y = r_n + \gamma Q_n^{\vartheta'}(\mathcal{O}', a'_1, \ldots, a'_N)|_{a'_j=\vartheta'_j(o'_j)}$ and for parameters θ'_n, a set of target polices is determined by $\vartheta' = \left\{\vartheta_{\theta'_1}, \vartheta_{\theta'_2}, \ldots, \vartheta_{\theta'_N}\right\}$. Thus, to execute a centralized action

Algorithm 1 Risk-Sensitive Energy Profiling Based on Multi-Agent System

Input: $\forall s_t \in \mathcal{S}$, $\mathcal{T}$, *Maxepc*
Output: π_{θ_n}, $\mathcal{O}$
Initialization: $\mathcal{N}$, $\forall o_t \in \mathcal{O}$, DQN, *epc*
1: **for Until:** $epc \leq Maxepc$ **do**
2: **if** ($epc == 0$) **then**
3: **Initialization**: $o_t = \langle s_t, a_t, r_t, s_{t'} \rangle$
4: **end if**
5: **for Until:** $\forall t \in \mathcal{T}$ **do**
6: **Step 1: Preprocessing**
7: *CVaR calculation using normal linear model*
8: **if** $P(\Upsilon(s_t, a_t)) \leq \alpha$ **then**
9: **Constraints:** (15b) to (15c)
10: **Update:**
$s_t : \langle \mathcal{E}^{dem}(t), \mathcal{E}^{ren}(t), \mathcal{E}^{sto}(t), P(\Upsilon(s_t, a_t)) \rangle \in \mathcal{S}$
11: **end if**
12: **Step 2: MADRL model**
13: **for** $\forall n \in \mathcal{N}$ **do**
14: **Action:** $a_t \sim \pi_\theta(a_t|s_t)$
15: **Receive:** $o_t = \langle s_t, a_t, r_t, s_{t'} \rangle$
16: **Evaluate:** $\mathcal{L}(\phi_n)$ using eq. (22)
17: **Agent policy:** $\nabla_{\phi_n} \mathcal{J}(\phi_n)$ using eq. (23)
18: **Update:** $\phi_n = \phi_n + \nabla_{\phi_n} \mathcal{J}(\phi_n)$
19: **Global policy:** $\nabla_{\theta_n} \mathcal{J}(\theta_n)$ using eq. (19)
20: **end for**
21: **Update:** $\theta_n = \theta_n + \nabla_{\theta_n} \mathcal{J}(\theta_n)$
22: **Append:** $o_t \in \mathcal{O}$
23: **end for**
24: **end for**
25: **return** π_{θ_n}, $\mathcal{O}$

value function (20), actions for all agents are known to us, whereas the nature of the environment is stationary while policies are changing. In this scenario, we can easily learn the other agent's polices from the observations. Therefore, parameters ϕ can approximate a policy $\hat{\vartheta}_{\phi_n^j}$ for each agent n from an observed policy ϑ_j of agent j and the centralized loss is defined as follows:

$$\mathcal{L}(\phi_n^j) = -\mathbb{E}_{o_j, a_j}\big[\log \hat{\vartheta}_n^j(a_j|o_j) + \beta h(\hat{\vartheta}_n^j)\big], \quad (21)$$

where β is a coefficient for the magnitude of regularization that able to solve the bias problem and $h(\cdot)$ determines entropy for a policy distribution $\hat{\vartheta}$. Thus, we can rewrite the true approximation as, $y \approx \hat{y} = r_n + \gamma Q_n^{\vartheta'}(O', \hat{\vartheta}_n'(o_1), \ldots, \vartheta_n'(o_n), \ldots \hat{\vartheta}_n'(o_N))$, where $\hat{\vartheta_n}'$ determines the target policy networks of policy $\hat{\vartheta_n}$. Therefore, the loss function for the policy ϑ is redefined as follows:

$$\mathcal{L}(\phi_n) = \min_{\vartheta} \mathbb{E}_{o_j, a_j}\big[\frac{1}{2}(Q_n^{\vartheta}(O, a_1, \ldots, a_N) - \hat{y})^2\big], \quad (22)$$

TABLE I
SUMMARY OF EXPERIMENT SETUP

Description	Value
No. of SBSs	10
No. of servers in each SBS	5
No. of CPU cores in one MEC server	4 with 1.2 GHz [3]
No. of solar units	40 [14]
Task sizes (uniformly distributed)	[31,1546060] bytes [15]
No. of task requests at SBS i	[1,10000] [3]
CVaR confidence levels	$\{0.90, 0.95, 0.99\}$
Learning rate	10^{-3}
Discount factor γ	0.99
Number of hidden layers and neurons	2, 100
Optimizer	ADAM [16]

and policy gradient can be redefined as follows:

$$\nabla_{\phi_n} \mathcal{J}(\phi_n) \approx \frac{1}{N} \mathbb{E}_{O, a_n \sim \mathcal{M}_n}\big[\sum_{n \in N} \nabla_{\phi_n} \log \vartheta_n(a_n|o_n) \nabla_{a_n} Q_n^{\vartheta}(O, a_1, \ldots, a_N)|_{a_n = \vartheta_n(o_n)}\big]. \quad (23)$$

The proposed multi-agent risk-sensitive energy profiling algorithm (in Algorithm 1) runs by microgrid controller and energy demand information is captured from MBS for each time slot t duration. In Algorithm 1, first we preprocess the energy data (consumption and generation) to cope with the CVaR risk and constraints (in lines 7 to 11), where we update state information after satisfying the constants (15a), (15b), and (15c) (in lines 8 to 10). In the second step, we model the MADRL with shared neural networks from lines 13 to 22. In line 16, we evaluate the DQN loss function (22) toward the risk-sensitive energy scheduling policy estimation. We estimate the local agent's policy in line 17 and update parameter values in line 18. We calculate global policy in line 19 and update parameters accordingly in line 21. The current observation o_t appends in the observational set in line 22. Finally, this algorithm provides a policy and observations (in line 25), which forecast a risk-sensitive energy profiling of the microgrid-powered MEC network.

In this paper, we consider a cellular network communication for Algorithm 1. Moreover, Algorithm 1 is capable of running on other types of communication protocols such as ZigBee, Powerline communication (PLC), Digital Subscriber Lines (DSLs), and so on [17]. In Algorithm 1, the goal of all agent N is always same and the policy gradient increases linearly with respect to the number of iteration (i.e., the total number of weight ϑ_n updates in each time slot t). Thus, the overall computational complexity of Algorithm 1 leads to $O(|\mathcal{S}|^2|\mathcal{A}||\mathcal{N}|)$ [18], where a single agent complexity goes in $O(|\mathcal{S}|^2|\mathcal{A}|)$ since learning time is decreasing [19] at a rate $O(\frac{1}{n}), \forall n \in \mathcal{N}$.

IV. EXPERIMENTAL ANALYSIS AND DISCUSSION

The proposed *risk-sensitive energy profiling* model is implemented on the Python platform, where we use TensorFlow APIs. In order to evaluate the proposed model, we have used state-of-the-art UMass solar panel dataset [14] for energy generation information and CRAWDAD nyupoly/video dataset

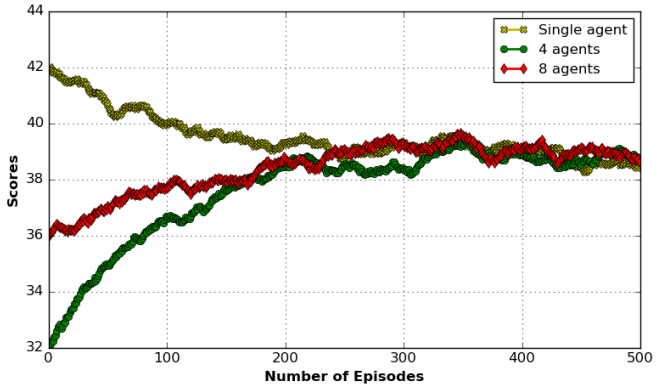

Fig. 3. Score value illustration of single agent and multi-agent deep reinforcement learning model.

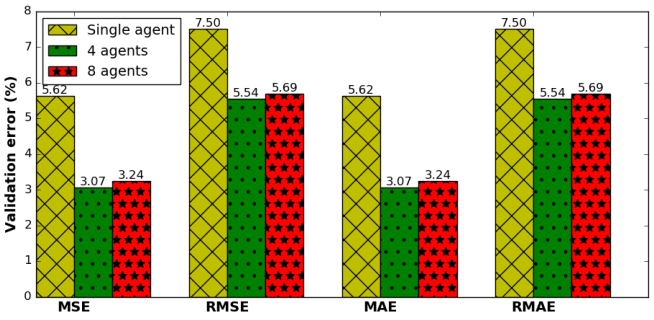

Fig. 4. Validation error analysis of the risk-sensitive energy profiling.

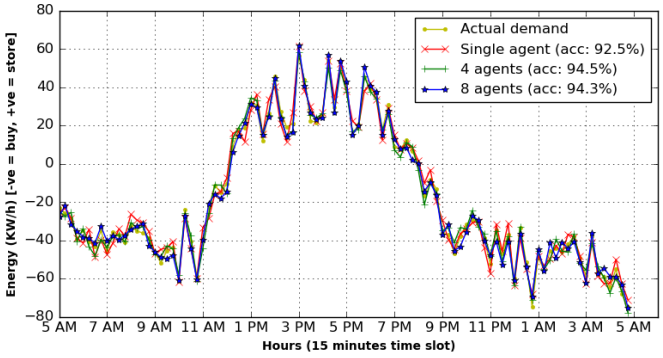

Fig. 5. Comparison of multi-agent system with single agent model in respect to buy or store energy forecasting.

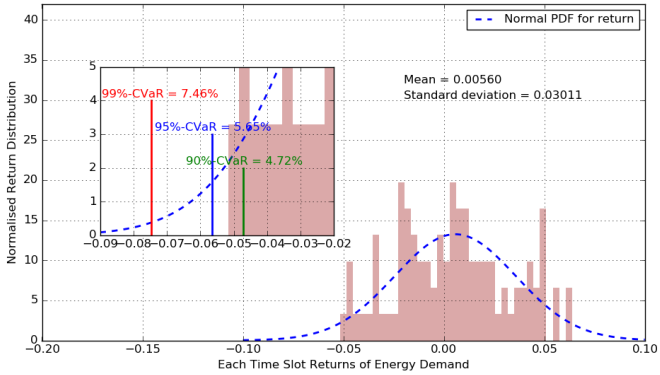

Fig. 6. Conditional Value-at-Risk (CVaR) analysis for energy demand-supply.

[15] for calculating computational task-based energy consumption of the MEC network. We divide datasets into 70% and 30% for training and testing, respectively [3]. Meanwhile, we preprocess both the datasets in order to provide the necessary information (i.e., state-space) to the implemented MADRL model. We employee single agent as a baseline method with 4-agents, and 8-agents MADRL model for a comprehensive experimental analysis. In Table I, we present the major parameters of the experiment setup. Other networks parameters are considered as similar in dataset [15] environment.

First, we justify the convergence of the proposed MADRL model in Fig. 3, where 4-agents (circle mark with a green line), and 8-agents (diamond mark with red line) achieve fast convergence with higher scores as compared with a single agent (cross mark with yellow line) model. Even though, at the beginning of training, the single-agent model gains a higher score than the other two; whereas, for the long term energy profiling, the MADRL model performs better. The number of episodes for the single-agent, 4-agents, and 8-agents models seems nearly the same toward convergence in Fig. 3. However, the single-agent method has more fluctuation, where scores decrease around 4.1% between episodes 400 and 450 while scores increase around 2.7% between episodes 451 and 500. Since more variance between exploitation and exploration that affects the accuracy of the learning process of the single-agent model.

Second, we validate the training error, where we apply Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Root Mean Absolute Error (RMAE) as the performance metrics in Fig. 4. In summary, the multi-agent model achieves up to 2.55% and 2% less error than that the single-agent model with respect to MSE and RMSE, respectively.

Third, in Fig. 5, we illustrate the comparison of 24 hours energy buying or storing decision using test datasets for three models, where the multi-agent (4,8 agents) system performs better (accuracy is around $94.3\% - 94.5\%$) due to knowledge (policy) sharing nature of the proposed model. As a result, we choose 4-agents model for further experimental analysis. Finally, Fig. 6 describes the tail of CVaR of the proposed risk-sensitive energy profiling model, where CVaR admits 4.72%, 5.65%, and 7.46% for confidence levels 90%, 95%, and 99% (shows in zoom part in Fig. 6), respectively. This assures that the proposed MADRL model can handle the risk of energy outage under the uncertainties of both demand and supply. As a result, the proposed risk-sensitive energy profiling enables a green MEC network as well as guarantees the maximum utilization of renewable energy from the microgrid.

V. Conclusions

In this paper, we have introduced a Conditional Value-at-Risk based energy profiling for the microgrid-enabled MEC network, where we have proposed a multi-agent system. The MADRL model has reflected on both wireless network energy consumption and microgrid generation to overcome the uncertainties in case of energy demand and supply. This

model also mitigates the risk of energy shortfall for the MEC network. Finally, experimental results demonstrate a significant performance gain in order to reduce the risk of energy failure for the MEC network, where the forecasting accuracy reaches up to 94.5% for 95% CVaR confidence and 5.65% risk.

REFERENCES

[1] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *in IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2961-2991, Fourth quarter 2018.

[2] S. Kekki et al., "MEC in 5G networks," *ETSI White Papers*, June 2018 (Visited on 4 December, 2018)

[3] M. S. Munir, S. F. Abedin, N. H. Tran and C. S. Hong, "When Edge Computing Meets Microgrid: A Deep Reinforcement Learning Approach," *in IEEE Internet of Things Journal*, Early Access, February. 2019.

[4] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User Scheduling and Resource Allocation in HetNets With Hybrid Energy Supply: An Actor-Critic Reinforcement Learning Approach," *in IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680-692, January 2018.

[5] X. Huang, T. Han and N. Ansari, "Smart Grid Enabled Mobile Networks: Jointly Optimizing BS Operation and Power Distribution," *in IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1832-1845, June 2017.

[6] Y. Chow, and M. Ghavamzadeh, "Algorithms for CVaR optimization in MDPs," *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3509-3517, Montreal, Canada, December 2014.

[7] Y. Zhang, M. H. Hajiesmaili, S. Cai, M. Chen, and Q. Zhu, "Peak-Aware Online Economic Dispatching for Microgrids," *in IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 323-335, January 2018.

[8] G. Auer et al., "How much energy is needed to run a wireless network?," *in IEEE Wireless Communications*,vol. 18, no. 5, pp. 40-49, October 2011.

[9] R. Bertran, M. Gonzalez, X. Martorell, N. Navarro, and E. Ayguade, "A Systematic Methodology to Generate Decomposable and Responsive Power Models for CMPs," *in IEEE Transactions on Computers*, vol. 62, no. 7, pp. 1289-1302, July 2013.

[10] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *In Advances in Neural Information Processing Systems*, California, USA, December 2017, pp. 6379-6390.

[11] M. G. Lagoudakis, "The 0-1 Knapsack Problem An Introductory Survey", *[Online]: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.4378*, 1996 (Visited on 11 March, 2019).

[12] S. Aaronson, "Guest column: NP-complete problems and physical reality," *ACM SIGACT News*, vol. 36, no. 1, pp. 30-52, March 2005.

[13] V. Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 1928-1937, New York, NY, USA, June 19 2016.

[14] Online:"Solar panel dataset", *UMassTraceRepository: http://traces.cs.umass.edu/index.php/Smart/Smart*, (Visited on 3 July, 2018).

[15] F. Fund, C. Wang, Y. Liu, T. Korakis, M. Zink, and S. Panwar, "CRAWDAD dataset nyupoly/video (v. 20140509)", *downloaded from: https://crawdad.org/nyupoly/video/20140509*, May 2014 (Visited on 3 July, 2018).

[16] D.P. Kingma, and J. Ba, "Adam: A Method for Stochastic Optimization," *in Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pp. 1-41, San Diego, USA, May 2015.

[17] V. C. Gungor et al., 'Smart Grid Technologies: Communication Technologies and Standards," *in IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 529-539, November 2011.

[18] L. P. Kaelbling, M. L. Littman, and A. W. Moore, 'Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 237-285, January 1996.

[19] S. D. Whitehead, "A complexity analysis of cooperative mechanisms in reinforcement learning," *Proceeding AAAI'91 Proceedings of the ninth National conference on Artificial*, vol. 2, pp. 607-613, July 1991.