Multi-Agent Deep Reinforcement Learning based Power Control for Large Energy Harvesting Networks

Mohit K. Sharma, Alessio Zappone, Mérouane Debbah and Mohamad Assaad

***Abstract*—The goal in this work is to design online power control policies for large energy harvesting networks where it is infeasible to use a centralized policy, due to large energy overhead involved in the exchange of state information. Furthermore, typical applications of these networks concern the scenario where the statistical information about both the energy harvesting (EH) process and the wireless channel is not available. In order to address these challenges, we propose a mean-field multi-agent deep reinforcement learning framework to design the online power control policies which operate in a fully distributed fashion, i.e., it does not require the knowledge about the state of the other nodes. Using the underlying structure of the problem, we analytically establish the convergence of the proposed scheme. In particular, we show that the policies obtained using the proposed approach converge to the Nash equilibrium. Our simulation results show the efficacy of the multiple-access schemes obtained through the proposed approaches. In particular, the mean-field multi-agent reinforcement learning scheme achieves a performance close to the state-of-the-art centralized policies which operates using the information about the state of entire network.**

I. Introduction

Internet-of-things (IoT) [1] networks connect a large number of low power sensors whose lifespan is typically limited by the amount of the energy that can be stored in their batteries. The advent of the energy harvesting (EH) technology [2] promises to prolong the lifespan of IoT networks by enabling the nodes to operate using the energy harvested from environmental sources, e.g., solar, wind, etc. On the other hand, this poses new constraints in the way energy is to be managed. An EH node (EHN) operates under the energy neutrality constraint (ENC), which requires that the total energy consumed by the node up to any point in time can not exceed the total amount of energy harvested by the node until that point. This constraint is particularly challenging due to the random nature of the EH process. In addition, at a given instant, an EHN can only store an amount of energy equal to the size of its battery capacity, which further adds to the complexity of the energy management problem in EH networks. As a result, a major and challenging issue in EH-based IoT systems is to devise power control policies to maximize the communication performance while ensuring the operation under ENC.

The authors are with the CentraleSupelec, Université Paris-Saclay, 91192 Gif-sur-Y vette, France. (e-mails: {mohitkumar.sharma,alessio.zappone}@l2s.centralesupelec.fr, {merouane.debbah,mohamad.assaad}@centralesupelec.fr). Mérouane Debbah is also with the Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France (e-mail:merouane.debbah@huawei.com). This research has been partly supported by the ERC-PoC 727682 CacheMire project.

Furthermore, when only the causal information about the energy arrivals and channel states [3] is available the power control policies are termed as online policies. The online policy design problem is essentially a stochastic control problem which, upon discretizing the state space (battery state and channel gains), can be formulated as a Markov decision process (MDP) [4] and can be solved numerically to obtain the optimal long-term policy. However, the MDP based approach requires perfect knowledge of the statistics of the EH process and propagation channels, which are difficult to know in practice. In order to address this drawback, the framework of reinforcement learning (RL) [5]–[9] or that of Lyapunov optimization [10], [11] have been proposed to approximate the optimal solution. All of these previous works take a centralized approach which makes it infeasible to use them for large networks, where the presence of a large number of nodes causes inevitable feedback overheads, as well as, more importantly, a huge complexity issue. Indeed, the numerical techniques available to solve MDP problems suffer from the so-called "curse-of-dimensionality", making them computational intractable for the large EH networks.

Therefore, it is essential to develop new techniques for the design of *distributed* online policies for large EH networks in absence of any a-priori knowledge about the EH process and the channel, and that do not require each node to have any information about the state of other nodes. Distributed approaches to online power control for EH networks has been recently considered in only a handful of works [12]–[15]. In [12], the authors use a distributed Q-learning algorithm where each node independently learns its individual Q-function. However, no convergence guarantee is provided for the proposed method. Note that, in general, when multiple nodes individually use a reinforcement learning algorithm to learn the optimal power control the converges is not guaranteed. This is because in such a scenario each individual node experiences an inherently non-stationary environment [16]. The authors in [13] developed a distributed solution to minimize the communication delay in EH-based large networks, assuming the information about the statistics of the EH process and of the propagation channel are known. Interestingly, the interactions among the many distributed devices are modeled leveraging mean-field game theory, a framework specifically conceived to analyze the evolution of systems composed of a very large number of distributed decision-makers [17], [18]. A multi-agent reinforcement learning (MARL) approach is considered

in [14], where an online policy for sum-rate maximization is developed. However, there it is assumed that the system global state is available at each node, which makes the approach from [14] not applicable to large networks, due to the extensive signaling required to make the system global state available to all network nodes. In [15], a two-hop EH network is considered, and a MARL-based algorithm with guaranteed convergence is proposed to minimize the communication delay.

The objective of this work is to develop a mechanism to learn the optimal online power control policies for fading-impaired multiple access channel (MAC) with a large number of EH transmitters, in a distributed fashion. In this context, we make the following main contributions:

- We model the problem of throughput maximization for EH MAC as a discrete-time mean-field game. Further, exploiting the structure of the problem we show that the mean-field game has a uniqueness stationary solution.
- Next, we propose to use deep reinforcement learning at each individual node to learn the stationary solution of the mean-field game. Under the proposed scheme the nodes learn the optimal power control in a completely distributed fashion without any apriori knowledge about the EH process and the propagation channel.
- Our simulation results confirm the theoretical findings and show that the throughput achieved by online power control policies learned using the proposed mean-field MARL (MF-MARL) framework is close to the throughput obtained using the state-of-the-art online policies which operate in a centralized fashion.

In the following section, we describe our system model.

II. System Model and Problem Formulation

We consider a time-slotted EH network where a large number of *identical* EHNs transmit their data over a block fading channel to an AP which is connected to the mains. The set of transmitters is denoted by $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$, where $K \gg 1$ denote the number of EHNs. In the n^{th} slot, the *fading* complex channel gain between the k^{th} transmitter and the AP is denoted[1] by $g_n^k \in \mathbf{G}_k$. In each slot, the channel between any transmitter and the AP remains constant for the entire slot duration, and changes at the end of the slot. We assume that the wireless channels between the nodes and the AP, $\mathbf{G}_k$, are identically distributed.

In a slot, the k^{th} node harvests energy according to a general stationary and ergodic harvesting process $f_{\mathcal{E}_k}(e_k)$, where the random variable $\mathcal{E}_k$ denotes the amount of energy harvested by the k^{th} transmitter and e_k denotes a realization of $\mathcal{E}_k$. We assume that the harvesting processes $\{\mathcal{E}_k\}_{k\in\mathcal{K}}$ are identically distributed across the individual nodes, but not necessarily independent. At each node, the harvested energy is stored in a perfectly efficient, finite capacity battery of size $B_{\max}$. Further, only *causal* and *local* information is available, i.e. each node knows only its *own* energy arrivals and the channel states to the AP in the current and all the previous time slots. In particular, no node has information about the battery and the channel state of the other nodes in the network. Also, no information is available about the distribution of the EH process and of the wireless channels at any node.

[1]For any symbol in the paper, the superscript and subscript represent the node index and the slot index, respectively, and if only the subscript is present then it denote either the node index or the slot index, depending on the context.

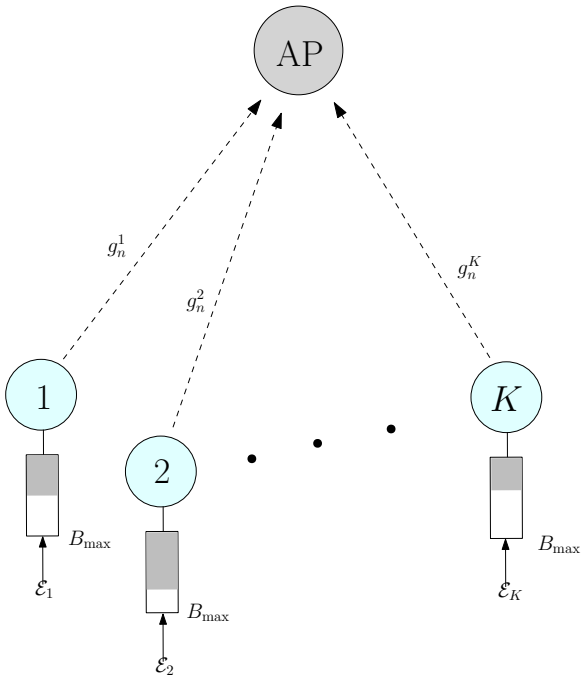

Fig. 1. System model for the EH fading multiple access network. The EH process and battery size at the k^{th} node are denoted by $\mathcal{E}_k$ and $B_{\max}$, respectively. The nodes transmit their data to an AP over a fading channel. The complex channel gain from the transmitter k to the AP, in the n^{th} slot, is denoted by g_n^k .

Let $p_n^k \leq P_{\max}$ denote the transmit energy used by the k^{th} transmitter in the n^{th} slot, where $P_{\max}$ denotes the maximum transmit energy which is determined by the RF front end. Further, $\mathcal{P}_n \triangleq \{p_n^k\}_{k=1}^K$ denote the set of transmit energies used, in the n^{th} slot, by all the transmitters. The battery at the k^{th} node evolves as

$$B_{n+1}^k = \min\{[B_n^k + e_n^k - p_n^k]^+, B_{\max}\}, \tag{1}$$

where $1 \leq k \leq K$, and $[x]^+ \triangleq \max\{0, x\}$. In the above, B_n^k and e_n^k denote the battery level and the energy harvested by the k^{th} node at the start of the n^{th} slot. An upper bound on the successful transmission rate of the EH MAC over N slots is given by [19]

$$\mathcal{T}(\mathcal{P}) = \sum_{n=1}^{N} \log\left(1 + \sum_{k\in\mathcal{K}} p_n^k g_n^k\right), \tag{2}$$

where $\mathcal{P} \triangleq \{\mathcal{P}_n | 1 \leq n \leq N\}$. Note that, the above upper bound can be achieved by transmitting independent and identically distributed (i.i.d.) Gaussian signals. In (2) for simplicity and without loss of generality, we set the power

spectral density of the AWGN at the receiver as unity. [2]

In the absence of information about the statistics of the EH process and the channel, our goal in this work is to learn online energy management policies for each node to maximize the time-averaged sum throughput. The optimization problem can be expressed as follows

$$\max_{\{\mathcal{P}\}} \liminf_{N\to\infty} \frac{1}{N}\mathcal{T}(\mathcal{P}), \tag{3a}$$

$$\text{s.t. } 0 \leq p_n^k \leq \min\{B_n^k, P_{\max}\}, \tag{3b}$$

for all n and $1 \leq k \leq K$. Constraint (3b) captures the fact that the maximum energy a node can use in the n^{th} slot is limited by the minimum between the amount of energy available in the battery, B_n^k, and the maximum allowed transmit energy $P_{\max}$. Note that, since the information about the *random* energy arrivals and the channel is only *causally available* and for each node the battery evolves in a Markovian fashion, according to (1), the optimization problem (3) is essentially a stochastic control problem which, upon discretization of the state space, could be formulated as a Markov decision process (MDP). However, solving such an MDP in the considered setting poses at least three major challenges:

- Infeasible complexity, since in the considered setup a large number of nodes K is present in the network.
- Considerably large feedback overhead, since global information about the battery and channel states of each network node would be needed.
- Finally, solving the MDP also requires statistical information about the EH process and the wireless channel, which is often difficult to obtain and indeed is not assumed in this work.

For these reasons, the goal of this work is to develop a framework to learn online power control policies in a distributed fashion, i.e., each node learns the optimal online power control policy without requiring to know the battery and channel states, and actions of the other nodes. In the ensuing sections, we develop a provably convergent multi-agent reinforcement learning approach exploiting the tools of deep reinforcement learning and mean-field games.

III. Mean-field Game to Maximize the Sum Throughput

In this section, first we present a discrete time finite state mean-field game [20] formulation of the sum throughput maximization problem in (3). Next, we present preliminaries on the discrete-time mean-field games and list the key results which are useful in showing the convergence of the proposed approach to the stationary solution of the mean-field game.

[2]We note that, in a scenario when all the EHNs simultaneously transmits their data, the cumulative signal-to-noise ratio (SNR) term in (2), $\sum_{k\in\mathcal{K}} p_n^k g_n^k$, grows with the number of users in the network. In practice, this problem can be circumvented by ensuring that the transmit power of EHNs scales down in inverse proportion to the number of users, i.e., $O\left(\frac{1}{K}\right)$, as when the number of users increases the power per user must decrease in order to ensure that the total energy in the network stays finite.

A. Throughput Maximization Game

The throughput maximization game $\mathcal{G}_T \triangleq \{\mathcal{K}, \mathcal{S}, \mathcal{P}, \mathcal{R}\}$ consists of:

- The set of players $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$, each one corresponding to an EH transmitter, where $K >> 1$;
- The state space of all players $\mathcal{S} \triangleq \times_{k\in\mathcal{K}}\mathcal{S}^k$, with $\mathcal{S}^k$ denoting the space of all the states s^k for the k^{th} transmitter and $|\mathcal{S}^k| \triangleq d$. Also, let $s_n^k \triangleq (B_n^k, g_n^k, e_n^k)$ denote the state of the k^{th} transmitter in the n^{th} slot, where B_n^k, g_n^k, and e_n^k are discrete-valued;
- The set of policies of all the nodes $\mathcal{P} \triangleq \{\mathcal{P}^k\}_{k\in\mathcal{K}}$, where $\mathcal{P}^k$ denotes the policy of the k^{th} node;
- The set of reward functions of all the nodes $\mathcal{R} \triangleq \{\mathcal{R}_k\}_{k\in\mathcal{K}}$, where $\mathcal{R}_k$ is the reward function of node k.

Note that, since all the transmitters are identical, the state space of individual nodes, $\mathcal{S}^k$, is the same set for all $k = 1, \ldots, K$. In the n^{th} time slot, the k^{th} node uses p_n^k amount of energy, prescribed by its policy $\mathcal{P}^k$, and collects a reward according to its reward function $\mathcal{R}_k$ and evolves from one state to another.

Under the mean field hypothesis [20], the reward obtained by a given node depends on the other nodes only through the distribution of all the nodes across the states. Let $\boldsymbol{\pi}_n \triangleq (\pi_n^1, \ldots, \pi_n^d)$ denote the distribution of all the nodes across the states in the n^{th} slot, where π_n^i denotes the fraction of nodes in the i^{th} state. Thus, in the n^{th} slot the reward obtained by the k^{th} node can be expressed as

$$\begin{aligned}\mathcal{R}_k(\boldsymbol{\pi}_n, p_n^k) &= \log\left(1 + p_n^k + \sum_{i=1}^{d}(K-1)\pi_n^i p_i g_i\right)\\ &= \log\left(1 + \sum_{i=1}^{d} K\pi_n^i p_i g_i\right),\end{aligned} \tag{4}$$

where g_i is the wireless channel gain between the nodes in the i^{th} state and the AP, and $p_i \in \mathcal{A}_p \triangleq \{0, p_{\min}, \ldots, P_{\max}\}$ denote the energy level used for transmission by the nodes in the i^{th} state. Here, $p_{\min}$ denotes the minimum energy required for transmission. Note that, (4) is written using the fact that under the mean-field hypothesis all the nodes are identical and hence use the same policy which, in turn, also implies that the reward function, $\mathcal{R}_k(\cdot,\cdot)$, is identical for all the nodes. Hence, to simplify the notations, in the ensuing discussion we omit the node index k. Also, (4) implicitly assumes that all nodes in state i use the energy p_i which essentially follows from the fact that for an MDP with finite state and action sets the optimal policy is a Markov deterministic policy [21, Thm. 8.4.7], i.e., in a slot the optimal transmit energy depends only on the current state.

In the n^{th} slot, when a node in state $s_n \in \mathcal{S}$ transmits using energy p_{s_n} the system evolves as

$$\pi_{n+1}^j = \sum_i \pi_n^i P_{ij}^n\left(p_i\right), \tag{5}$$

where $P_{ij}^n(\cdot)$ denotes the probability in the slot n that a node in state i transits to state j and depends on p_i, the energy used

by the nodes in the i^{th} state for transmission[3]. In addition, $P_{ij}^n(\cdot)$ is determined by the statistics of the EH process and the wireless channel. The nodes in the i^{th} state obtains a reward $\mathcal{R}\left(\boldsymbol{\pi}_n, p_i\right)$, equal to the total sum rate obtained in the slot.

For a given node, starting from the state in the n^{th} slot the expected sum-throughput obtained by following a policy $\mathcal{P}$ can be expressed as

$$V_n(\boldsymbol{\pi}_n, \mathcal{P}) = \mathcal{R}\left(\boldsymbol{\pi}_n, \mathcal{P}\right) + V_{n+1}\left(\boldsymbol{\pi}_{n+1}, \mathcal{P}\right), \tag{6}$$

where $V_{n+1}(\boldsymbol{\pi}_{n+1}, \mathcal{P})$ denotes the expected throughput obtained by following a policy $\mathcal{P}$ starting from slot $n+1$, when in the $(n+1)^{\text{th}}$ slot the distribution of the nodes across the states is $\boldsymbol{\pi}_{n+1}$. In the rest of the paper $V(\cdot,\cdot)$ is also termed as the value function. In the above, similar to an MDP [21], (6) is written using the fact that the expected sum-throughput obtained by following policy $\mathcal{P}$, starting from the time slot n, is equal to the sum of the expected sum-throughput obtained in the slot n and the slot $n+1$ onward. Note that, under the mean-field hypothesis, the expected sum-throughput in (6) is identical for all the nodes and due to special structure of the reward function, the value function of each node $V(\cdot,\cdot)$ only depends on the distribution of the nodes across the states, $\boldsymbol{\pi}_n$, not on the state of the individual nodes. In the following, we present preliminaries on the discrete-time finite state mean field games.

B. Preliminaries: discrete-time finite state Mean-field games

In the following, we define the notion of Nash equilibrium, stationary solution, and briefly summarize the key results used to prove the convergence of the proposed MARL algorithm. For a detailed exposition on the discrete-time finite state mean-field games we refer the readers to [20].

Definition 1 (Nash maximizer)**.** *For a fixed probability vector $\boldsymbol{\pi}_n$, a policy $\mathcal{P}^*$ is said to be a Nash maximizer if and only if*

$$V_n(\boldsymbol{\pi}_n, \mathcal{P}) \leq V_n(\boldsymbol{\pi}_n, \mathcal{P}^*), \text{ for all policies } \mathcal{P}.$$

Next, for a discrete-time finite state mean-field game, we define the notion of the solution and the stationary solution.

Definition 2 (Solution of a mean-field game)**.** *Suppose that for each $\boldsymbol{\pi}_n$ there exists a Nash maximizer $\mathcal{P}^*$. Then a sequence of tuples $\{(\boldsymbol{\pi}_n, V_n)$ for $n \in \mathbb{N}\}$ is a solution of the mean field game if for each $n \in \mathbb{N}$ it satisfies* (5) *and* (6) *for some Nash maximizer of V_n.*

Definition 3 (Stationary solution)**.** *Let $\mathcal{G}_{\boldsymbol{\pi}}$ and $\mathcal{K}_V$ be defined as $\mathcal{G}_{\boldsymbol{\pi}_n}(V_{n+1}) = V_n(\boldsymbol{\pi}_n, \mathcal{P})$, and $\mathcal{K}_{V_n}(\boldsymbol{\pi}_n) = \boldsymbol{\pi}_{n+1}$. A pair of tuple $(\tilde{\boldsymbol{\pi}}, \tilde{V})$ is said to be a stationary solution if and only if*

$$\mathcal{G}_{\tilde{\boldsymbol{\pi}}}(\tilde{V}) = \tilde{V} \text{ and } \mathcal{K}_{\tilde{V}}(\tilde{\boldsymbol{\pi}}) = \tilde{\boldsymbol{\pi}}.$$

Note that, the operators $\mathcal{K}_{V_n}(\cdot)$ and $\mathcal{G}_{\boldsymbol{\pi}_n}(\cdot)$ are compact representations of (5) and (6), respectively. The stationary solution of a mean-field game, $(\tilde{\boldsymbol{\pi}}, \tilde{V})$, is a fixed-point of operators $\mathcal{G}_{\boldsymbol{\pi}}$ and $\mathcal{K}_V$ which are essentially discrete time counterparts of Hamilton-Jacobi-Bellman and Fokker-Planck equations. Next, we list the results which identifies the conditions under which a stationary solution exists. We omit the proofs for brevity. These results are later used for proving the convergence of our mean-field MARL (MF-MARL) algorithm.

Theorem 1 (Uniqueness of Nash maximizer (Theorem 2 [20]))**.** *Let $f_i(p_i) = \frac{\partial V(\boldsymbol{\pi}, \mathcal{P})}{\partial p_i}$ where $p_i \in [0, P_{\max}]$ for all $1 \leq i \leq d$. If the value function V_n is convex and continuous with respect to p_i, and f_i is strictly diagonally convex, i.e., it satisfies*

$$\sum_{i=1}^{d} (p_i^1 - p_i^2)(f_i(\mathcal{P}^1) - f_i(\mathcal{P}^2)) > 0, \tag{7}$$

then there exists a unique policy which is a Nash maximizer for the value function V. Here, p_i^1 and p_i^2 denote the actions prescribed in the i^{th} state by two arbitrary policies $\mathcal{P}^1$ and $\mathcal{P}^2$, respectively.

The following result shows that if the reward function is monotonic with respect to both the variables then the mean-field game admits a unique stationary solution.

Theorem 2 (Uniqueness of stationary solution (Proposition 4.3.1, [22]))**.** *Let value function be a continuous function with respect to both its arguments and also assume that there exists a unique Nash maximizer $\mathcal{P}_n$ for all $n \in \{0, 1, 2, \cdots\}$. Further, let the reward function be monotone with respect to the distribution $\boldsymbol{\pi}$, i.e,*

$$\sum_{i=1}^{d} (\pi_i^2 - \pi_i^1)(\mathcal{R}_i(\mathcal{P}^1, \pi^2) - \mathcal{R}_i(\mathcal{P}, \pi^1)) \geq 0, \tag{8}$$

then there exists a unique stationary solution for the mean-field game. In the above $\mathcal{R}_i(\cdot,\cdot)$ denotes the reward obtained by the nodes in the i^{th} state.

In the following subsection, we establish that the mean-field game $\mathcal{G}_T$ admits a unique stationary solution.

C. Unique Stationary Solution for $\mathcal{G}_T$

Theorem 3. *The throughput maximization mean-field game $\mathcal{G}_T$ has a unique stationary solution.*

Proof: Proof is relegated to appendix A ∎

In the next section, we present an algorithm to learn the stationary solution of the mean-field game $\mathcal{G}_T$ and also the corresponding Nash maximizer power control policy. The proposed approach uses reinforcement learning for this purpose and is termed as MF-MARL approach.

IV. MF-MARL Approach to Power Control

In this section, we present our mean-field MARL approach to learn the online energy management policies to maximize the throughput of a fading EH-MAC with large number of users. We show that the policies obtained using the proposed approach not only are amenable to be learned in a distributed fashion but also converge to the *stationary* Nash equilibrium.

[3]Note that, in a general mean-field game the transition probabilities P_{ij}^n may also depend on the actions of the other players.

The proposed MF-MARL algorithm to obtain the online policies exploits the fact that the discrete time mean field finite state games has the fictitious play property (FPP) [22]. The FPP for a discrete time mean field game is described in the following. Let m denote the iteration index and $\bar{\boldsymbol{\pi}}_1$ denote an arbitrary probability vector denoting the initial distribution of the nodes across the states. Let

$$\mathcal{P}_m^* \triangleq \arg\max_{\mathcal{P}} V_m\left(\bar{\boldsymbol{\pi}}_m, \mathcal{P}\right), \tag{9}$$

$$\boldsymbol{\pi}_{m+1} = \mathcal{K}_{V_m(\mathcal{P}_m^*)}(\boldsymbol{\pi}_m), \tag{10}$$

$$\text{and } \bar{\boldsymbol{\pi}}_{m+1} = \frac{m}{m+1}\bar{\boldsymbol{\pi}}_m + \frac{1}{m+1}\boldsymbol{\pi}_{m+1}. \tag{11}$$

The procedure described by (9), (10) and (11) is called the fictitious play procedure. As described in (9), at the m^{th} iteration, a node attempts to learn the Nash maximizer, given that its belief about the distribution of the nodes across the states is $\bar{\boldsymbol{\pi}}_m$. Further, at each iteration of the fictitious play procedure the belief about the distribution is updated using (10) and (11). A discrete-time mean field game is said to have a FPP if and only if the procedure described by (9), (10) and (11) converges. The following result provides the conditions under which the fictitious play procedure converges to the unique solution of the discrete-time mean field game.

Theorem 4 (Convergence of FPP to unique stationary solution (Theorem 4.3.2 [22]))**.** *Let $(\boldsymbol{\pi}_m, V_m)$ denote the sequence generated through the FPP. If a mean-field game has a unique Nash maximizer at each stage of the game and the reward function is continuous and monotone with respect to probability vector $\boldsymbol{\pi}$ then the sequence $(\boldsymbol{\pi}_m, V_m)$ converges to $(\tilde{\boldsymbol{\pi}}, \tilde{V})$ the unique stationary solution of the mean-field game.*

For the throughput maximizing mean-field game $\mathcal{G}_T$, the convergence of the FPP to the stationary solution of the game directly follows from the above result. As a consequence of this result, the stationary solution of the $\mathcal{G}_T$ can be learned through the fictitious play procedure, provided the Nash maximizer can be found at each iteration of the fictitious play and the belief about the distribution is updated accordingly. The MF-MARL proposes to use the reinforcement learning to learn the Nash maximizer at each iteration, i.e., for a given belief distribution $\bar{\boldsymbol{\pi}}$ each node uses a reinforcement learning algorithm to learn the Nash maximizer. The proposed MF-MARL approach is described in Algorithm 1.

In step 2 of algorithm, the estimate of π_{m_n} could be computed using the empirical distribution at the AP which periodically broadcasts it to the entire network. Also, in order to run the Q-learning algorithm a node need to know the reward, i.e., the sum-throughput, obtained in each slot. Since the reward function is same across the nodes, this could be accomplished by using the belief about the distribution. In particular, each node uses its own policy and the belief about the distribution to build an estimate of the reward obtained in each slot. Alternatively, in each slot the AP can directly broadcast the total number of bits successfully decoded by the AP. The latter method obviates the need to maintain a belief about the distribution of the nodes, albeit at the cost of a higher feedback overhead. The latter method gives rise to cooperative multi-agent Q-learning [23] where nodes attempts to maximize a common reward function. As observed in our simulations that the proposed MF-MARL based approach performs better than the latter cooperative Q-learning method. In the following section, we present the simulation results.

Algorithm 1 : MF-MARL approach to learn optimal online policies

Initialize: $\bar{\boldsymbol{\pi}}_{\mathbf{1}}$ to a valid probability vector, ϵ_1, $\tilde{\epsilon}$, N and $m \leftarrow 0, n \leftarrow 0$

do

1) Set $m \leftarrow m+1$; at each node run Q-learning algorithm to learn Nash maximizer $\mathcal{P}_m^k*$
2) In the n^{th} time-slot, $n \leq N$, of Q-learning episode the AP estimate $\boldsymbol{\pi}_{\boldsymbol{m_n}}$.
3) $n \leftarrow n+1$. If $\|\boldsymbol{\pi}_{m_n} - \boldsymbol{\pi}_{m_{n+1}}\|_2 \geq \epsilon_1$ or $n > N$, broadcast $\boldsymbol{\pi}_{\boldsymbol{m+1}} = \boldsymbol{\pi}_{m_{n+1}}$; else go to step 2.
4) Update $\bar{\boldsymbol{\pi}}_{m+1}$ using (11) and $n \leftarrow 0$.

while $\|\bar{\boldsymbol{\pi}}_{\boldsymbol{m+1}} - \bar{\boldsymbol{\pi}}_{\boldsymbol{m}}\|_2 \leq \tilde{\epsilon}$.

Output: The near-optimal policies and distribution are given by $\mathcal{P}^*$ and $\tilde{\pi}$, respectively.

V. Simulation Results

We consider an EH MAC with $K = 5$ EH transmitters where each EHN harvests energy according to a non-negative truncated Gaussian distribution with mean m and variance $v = 3.5$, independently of the other nodes. The capacity of the battery at each transmitter is $B_{\max} = 20$ and the maximum amount of energy allowed to be used for transmission in a slot is $P_{\max} = 15$. Note that, the unit of energy is 10^{-2} J. We benchmark the performance of the proposed MF-MARL and cooperative Q-learning approach against the state-of-the-art centralized scheme [24]. In the centralized scheme the online policies are learned by training the deep neural networks using the data obtained by solving the jointly optimal offline policies. The performance is evaluated by averaging the sum-throughput obtained over 2×10^4 slots.

At each node we use the deep Q-learning [25] method. Each deep Q network consists of 10 hidden layers and one input and output layer. The input layer contains 3 neurons, while the number of neurons in the output layer is equal to $|\mathcal{A}| = 150$, where $\mathcal{A} = \{0, 0.1, 0.2, \ldots, 15\}$. The first, third, fifth, seventh, and ninth hidden layer consists of $60, 58, 56, 54$, and 52 neurons, respectively. The number of neurons in each even indexed hidden layer remains same as in the previous odd indexed hidden layer. At each layer, except the output layer, the rectified linear unit (ReLu) is used as activation function. The output layer uses a linear activation function. The deep Q-learning algorithm uses $\gamma = 0.99$, and uses the exploration probability $\epsilon_{\max} = 1$ at the start which decays to $\epsilon_{\min} = 0.01$

TABLE I
PERFORMANCE OF THE MF-MARL BASED APPROACH FOR AN EH MAC WITH $K = 5$ USERS AND $v = 3.5$. PERFORMANCE OF THE CENTRALIZED POLICY CORRESPONDS TO 100%.

Mean (m)	Centralized Policy (RPS in nats)	MF-MARL policy (RPS in nats)	MF-MARL policy (Percentage)	Cooperative Q-learning (RPS in nats)	Cooperative Q-learning (Percentage)
4	3.1498	2.9390	93.30%	2.9354	93.19%
5	3.3107	3.1311	94.57%	3.0046	90.75%
6	3.4410	3.1072	90.29%	3.1852	92.56%
7	3.5102	3.2960	93.89%	3.2417	92.35%
8	3.6146	3.3973	93.98%	3.3064	91.47%
9	3.6166	3.5179	95.68%	3.4528	93.90%

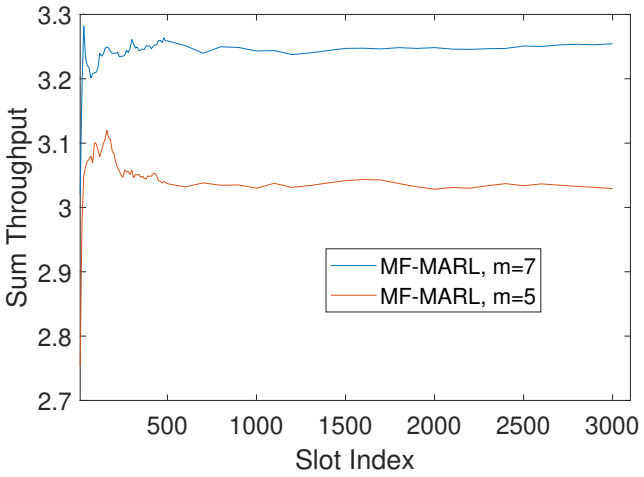

Fig. 2. Convergence of MF-MARL algorithm.

with a decay factor equal to 0.995. The replay memory of length 2000 is used. In Algorithm 1, we use $\epsilon_1 = 0.01$, $\tilde{\epsilon} = 0.001$, and $N = 1000$.

As observed from the results in Table I, the policies obtained using the proposed MF-MARL based approach achieve the sum-throughput which is close to the throughput achieved by the centralized policies. Note that, in order to implement MF-MARL (or deep Q-learning) the actions space, $\mathcal{A}$, has to be quantized which, in turn, leads to the loss in the throughput, compared to the centralized scheme where the output transmit powers are continuous. We observe that the proposed MF-MARL based approach performs marginally better than the cooperative multi-agent Q-learning based scheme. However, in contrast to cooperative multi-agent Q-learning approach, the MF-MARL based procedure requires significantly less feedback. Also, it is interesting to note that the proposed MF-MARL algorithm achieves the near-optimal throughput even for a network with small number of nodes. Further, the result in Fig. 2 show the throughput achieved by our MF-MARL algorithm as a function of slot index. It is interesting to observe that the MF-MARL algorithm converges very fast, i.e., the obtained throughput stabilizes within first 500 slot. A similar trend is observed fo cooperative Q-learning also.

VI. CONCLUSIONS

In this work, we proposed a mean-field multi-agent reinforcement learning based framework to learn the online power control policies for large EH networks. Using the structure of the underlying problem we analytically showed that the learning process converges to a unique stationary solution of the underlying mean-field game. Moreover, the proposed approach enables the nodes to learn the policies in a completely distributed fashion. Our simulation results corroborated the theoretical findings. The future work could involve characterizing the convergence speed of the proposed MF-MARL algorithm. The proposed MF-MARL framework could be useful for optimization of large wireless networks.

APPENDIX

Proof: The proof follows directly from the result in Theorem 2, provided there exists a unique Nash maximizer and the reward function is monotone in variable $\boldsymbol{\pi}$. The uniqueness of Nash maximizer can established using the result in Theorem 1. It is easy to verify that the reward and value function of the game $\mathcal{G}_T$ satisfies the strictly diagonally concavity property. In order to complete the proof we just need to show that the reward function is monotone with parameter $\boldsymbol{\pi}$, i.e.,

$$\sum_{i=1}^{d}(\pi_i^2 - \pi_i^1)(\mathcal{R}_i(\mathcal{P}, \pi^2) - \mathcal{R}_i(\mathcal{P}, \pi^1)) \geq 0. \qquad (12)$$

The proof follows by noting the fact that since the reward obtained by a node does not depend on the state of the node, i.e., $\mathcal{R}_i(\mathcal{P}, \pi^2) = \mathcal{R}(\mathcal{P}, \pi^2)$. Hence, the RHS in the above can be expressed as $(\mathcal{R}(\mathcal{P}, \pi^2) - \mathcal{R}(\mathcal{P}, \pi^1))\left(\sum_{i=1}^{d}\pi_i^1 - \sum_{i=1}^{d}\pi_i^1\right) = 0$. ■

REFERENCES

[1] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios," *IEEE Wireless Commun. Mag.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[2] M. L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1384–1412, Second Quarter 2016.

[3] M. K. Sharma and C. R. Murthy, "Distributed power control for multi-hop energy harvesting links with retransmission," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4064–4078, Jun. 2018.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2017, vol. II, http://www.athenasc.com/dpbook.html.

[5] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, Apr. 2013.

[6] M. Chu, H. Li, X. Liao, and S. Cui, "Reinforcement learning based multi-access control and battery prediction with energy harvesting in iot systems," *IEEE J. Internet Things*, vol. PP, no. 99, pp. 1–1, 2018.

[7] A. Masadeh, Z. Wang, and A. E. Kamal, "Reinforcement learning exploration algorithms for energy harvesting communications systems," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[8] 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," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.

[9] Y. Xiao, Z. Han, D. Niyato, and C. Yuen, "Bayesian reinforcement learning for energy harvesting communication systems with uncertainty," in *Proc. IEEE ICC*, Jun. 2015, pp. 5398–5403.

[10] L. Huang, "Fast-convergent learning-aided control in energy harvesting networks," in *Proc. of IEEE Conf. Dec. and Control (CDC)*, Dec. 2015, pp. 5518–5525.

[11] M. Gatzianas, L. Georgiadis, and L. Tassiulas, "Control of wireless networks with rechargeable batteries," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 581–593, Feb. 2010.

[12] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Switch-on/off policies for energy harvesting small cells through distributed Q-learning," in *Proc. WCNC*, Mar. 2017, pp. 1–6.

[13] D. Wang, W. Wang, Z. Zhang, and A. Huang, "Delay-optimal random access for large-scale energy harvesting networks," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[14] A. Ortiz, H. Al-Shatri, T. Weber, and A. Klein, "Multi-agent reinforcement learning for energy harvesting two-hop communications with full cooperation," 2017. [Online]. Available: arXiv:1702.06185v1

[15] V. Hakami and M. Dehghan, "Distributed power control for delay optimization in energy harvesting cooperative relay networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4742–4755, Jun. 2017.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. of Conf. Neural Inf. Process. Syst. (NIPS)*, 2017. [Online]. Available: arXiv:1706.02275v3

[17] A. F. Hanif, H. Tembine, M. Assaad, and D. Zeghlache, "Mean-field games for resource sharing in cloud-based networks," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 624–637, Feb. 2016.

[18] M. Larranaga, M. Assaad, and K. DeTurck, "Queue-aware energy efficient control for dense wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2018, pp. 1570–1574.

[19] Z. Wang, V. Aggarwal, and X. Wang, "Iterative dynamic water-filling for fading multiple-access channels with energy harvesting," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 382–395, Mar. 2015.

[20] D. A. Gomes, J. Mohr, and R. R. Souza, "Discrete time, finite state space mean field games," *Journal de Mathématiques Pures et Appliquées*, vol. 93, no. 3, pp. 308 – 328, 2010.

[21] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 2017.

[22] S. Hadikhanloo, "Learning in mean field games," Ph.D. dissertation, Université Paris-Dauphine, Paris, France, Jan. 2018. [Online]. Available: http://www.cmap.polytechnique.fr/~saeed.hadikhanloo/PhD_Thesis.pdf

[23] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. of 17th Int. Conf. on Autonomous Agents and MultiAgent Systems*, 2018.

[24] M. K. Sharma, A. Zappone, M. Debbah, and M. Assaad, "Deep learning based online power control for large energy harvesting networks," in *submitted to ICASSP*, 2019.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236