Distributed Power Control for Delay Optimization in Energy Harvesting Cooperative Relay Networks

Vesal Hakami and Mehdi Dehghan, *Member, IEEE*

***Abstract*—We consider cooperative communications with energy harvesting (EH) relays and develop a distributed power control mechanism for the relaying terminals. Unlike prior work, which mainly deals with single-relay systems with saturated traffic flow, we address the case of bursty data arrival at the source cooperatively forwarded by multiple half-duplex EH relays. We aim at optimizing the long-run average delay of the source packets under the energy neutrality constraint on the power consumption of each relay. While EH relay systems have been predominantly optimized using either offline or online methodologies, we take on a more realistic learning-theoretic approach. Hence, our scheme can be deployed for real-time operation without assuming acausal information on channel realizations, data/energy arrivals as required by offline optimization, or relying on precise statistics of the system processes, as is the case with online optimization. We formulate the problem as a partially observable identical payoff stochastic game (PO-IPSG) with factored controllers in which the power control policy of each relay is adaptive to its channel and energy states as well as to the state of the source buffer. We equip each relay with a reinforcement learning procedure and prove that the parallel execution of this procedure is convergent to (at least) a locally optimal solution of the formulated PO-IPSG. The proposed algorithm operates without explicit message exchanges between the relays, while inducing only little source-relay signaling overhead. By simulation, we contrast the delay performance of the proposed method against existing heuristics for throughput maximization. It is shown that compared with these heuristics, the systematic approach adopted in this paper has a smaller suboptimality gap once evaluated against a centralized optimal policy armed with perfect statistics.**

***Index Terms*—Bursty traffic, cooperative relaying, energy harvesting (EH), power control, reinforcement learning, stochastic game, wireless communication.**

I. Introduction

COOPERATIVE relaying is a promising paradigm that results in broader coverage and in combating the wireless channel impairments. Relay-assisted transmission mitigates the need to use a high power at the transmitter, leading to prolonged battery life and lower level of interference [1]. Relays in wireless networks can be classified as decode-and-forward (DaF) relays, which decode and possibly re-encode the information before forwarding it, and amplify-and-forward (AaF) relays, which forward an amplified version of the signal without hard decoding. AaF relays compared with other types, which require signal detection, are less complicated, have lower implementation cost, and are thus widely utilizable [4]. While cooperative relaying results in higher network capacity, in forwarding to the destination a representation of the signal it has received from the source, a relay consumes its own energy. Since replacing batteries for such devices is either impracticable or costly in several scenarios, recent advances in energy harvesting (EH) devices [5] have paved the way for self-sustainable relays [6] that power themselves from theoretically unlimited energy sources that are present in their surrounding environment (e.g., in the form of solar, vibration, thermoelectricity, etc.). However, the harvested energy rates are typically quite low, with sporadic arrivals in random limited amounts, and it is thus desirable to accumulate the harvested energy by storing it in a buffer such as a rechargeable battery for subsequent usage. In practice, the energy buffer is restricted in size, and thus EH relays may face power outage whenever the energy consumption rate is higher than the harvesting rate. Hence, there is a need for novel power-use policies that exploit available information on the energy, channel, and data arrival processes to efficiently utilize the harvested power for meeting application-specific demands.

A. Literature Review

Exploiting both EH and cooperative communications has received a considerable interest recently [7]–[20]. The use of EH relays in cooperative communication was first introduced in [8], wherein a comprehensive performance analysis was conducted for relay selection and transmission power setting in an AaF network in terms of symbol error probability by using a probabilistic energy model. However, the results in [8] are mostly of analytical interest rather than proposing a practical optimization scheme. More recently, several studies have come up with transmission control strategies (e.g., power allocation, relay selection, etc.) to optimize different network utility functions in EH relay systems [7], [9]–[11], [13]–[15], [17]–[20], [35]. These schemes can be categorized based on two main distinguishing features.

1) *Optimization method (offline/online/learning-theoretic):* In offline optimization, it is assumed that all future realizations of data/energy arrivals as well as the channel variations are known acausally before the system starts.

Manuscript received February 7, 2016; revised June 25, 2016; accepted August 23, 2016. Date of publication September 16, 2016; date of current version June 16, 2017. The review of this paper was coordinated by C. Assi.

V. Hakami is with the Department of Computer Engineering, Iran University of Science and Technology (IUST), Tehran 16846-13114, Iran (e-mail: vhakami@iust.ac.ir).

M. Dehghan is with the Department of Computer Engineering and Information Technology, Amirkabir University of Technology (AUT), Tehran 15916-34311, Iran (e-mail: dehghan@aut.ac.ir).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2016.2610444

In general, offline optimization problems are modeled as a mathematical program and the solution obtained can be considered as an upper bound on the performance of the actually stochastic system. In contrast, online optimization is much more realistic in that only statistical knowledge but causal information on the realizations of the system states is assumed. A systematic way to approach online optimization is to formulate the problem as a stochastic dynamic program (DP) [21], and optimize the expected value of the long-run system performance. Nonetheless, in many practical scenarios, either the characteristics of the channel variations and energy/data arrival processes change over time, or it is not possible to have reliable statistical information about these processes before node deployments. For example, in a sensor field with solar EH nodes distributed over a forest, each node's solar EH profile will depend on its location and is subject to change based on the time of the day or the day of the week. To adapt the transmission scheme in real time, one should resort to learning-theoretic schemes, as they are capable of converging to optimal transmission policy over time in the absence of prior knowledge on the statistics of the processes governing the communication system.

2) *Traffic type assumption (saturated/bursty):* Under saturated traffic assumption, there are infinite data backlogs at the source, and the optimization objective is to improve the physical layer performance (e.g., throughput, outage probability or symbol error rate), by only accounting for channel and energy state processes. When traffic is bursty, however, there is a need for a buffer where packets can be queued. The "emptying" rate of the buffer then becomes the "service" rate. A physical-layer model that only captures the variation of the channel and energy completely disregards this issue, and it can result in arbitrary long average waiting time of the packets at the source buffer. When the end-to-end delay is of interest, we need to track the source queue size that develops under bursty traffic generation, and the allocation of power at relays should control the service rate to achieve delay optimization at the source data link layer.

The majority of the studies on EH relay systems lie within the offline optimization framework, and assume nonbursty source traffic type [7], [9], [10], [13]–[15], [18]–[20]. In [10], the problem of optimal power control for throughput maximization in an SRD network (one source-destination pair and one relay) is formulated as a nonlinear program in an offline setting. Both source and relay are harvesting entities, and the relay operates in half-duplex mode using AaF protocol. A similar setup is considered in [7], but only for the case that both source and relay nodes have their own data to transmit to the destination, and the optimization objective is to maximize the total throughput. Also, in [9], the transmit power is jointly optimized with relay selection to handle the case of multiple relays. In [13], source and relay power allocation is optimized for a Source Relay Destination (SRD) system with a full-duplex relay using DaF protocol. Half-duplex DaF relaying is considered in [14], wherein it is assumed that only the source node can harvest energy. The case in which both source and relay are EH nodes is handled in [15] and [18], whereas [20] considers two parallel EH relays (the so-called diamond relay channel [22]). It is also worth noting that technically, the multirelay case can be deemed equivalent to the orthogonal frequency division multiplexing (OFDM) relay with individual power constraint in each subcarrier. Accordingly, the studies in [38] and [39] have proposed optimization schemes for data and energy cooperation in relay-enhanced OFDM systems.

Some studies [9], [10], [19] propose online throughput maximization for the case of saturated source traffic. In [19], for instance, a stochastic DP formulation is given for optimal online power allocation in the case of DaF relaying. In [10], the online power allocation problem is formulated as a Markov decision process (MDP) [23] and a computationally simple scheme is provided for the special case in which power control at the nodes is limited to on-off switching. Again, within the context of saturated source traffic type, there has also been a recent study that utilizes a solar-data-driven stochastic EH model in an MDP-based design and obtains the optimal DaF relay power control policy to minimize the long-term average symbol error rate [35]. Under a bursty on-off Markovian traffic assumption, the study in [11] addresses online relay scheduling for EH wireless sensor networks. The problem is formulated as a partially observable MDP [24] in which the source node has to choose between direct or cooperative transmission modes depending on its own available energy, the states of its EH and event generation processes, and by using only partial knowledge of the relay's state.

Finally, in [17], a multisource, single relay cooperative network is considered whereby the traffic at the source nodes is assumed to be bursty and the forwarding protocol used by the relay is DaF. The transmit power of all nodes is assumed to be contributed by both the conventional ac utility power and the renewable energy. A distributed learning algorithm is proposed to minimize the sum of the average delay of the data flows by dynamic power, rate, and link selection control.

B. Motivation, Contributions, and Outline

Most prior art in optimizing the performance of EH relay systems belong to the realm of offline optimization, and primarily deal with the didactic single relay scenario [7], [10], [11], [13]–[15], [18], [19]. Also, the existing online schemes require explicit knowledge of the statistics of the system processes [9]–[11], [19] and do not address the case of bursty traffic in general, wherein the optimization of the queuing delay is necessary. Unlike [17], in this paper, we consider an EH cooperative relay system consisting of multiple AaF relays that are powered solely by an EH storage with limited capacity. The source node, on the other hand, has a continuous power supply and maintains a data buffer for the bursty traffic flow toward the destination.

We aim at proposing a learning-theoretic scheme to control the relays' power consumption for optimizing the long-run average delay experienced by the source packets. Ideally, the learning mechanism should be able to dynamically control the transmit power at the relays in adaptation to the source buffer state information (SBSI) as well as the global channel state information (CSI) and energy state information (ESI) of the relays. This calls for a principled design based on a centralized stochastic DP formulation. However, such scheme is already doomed by the curse of dimensionality due to the huge space of global CSI, global ESI, as well as the exponential growth of the number of joint action combinations with the number of relays involved. Moreover, to gain access to the global state of the system, a centralized controller would induce heavy signaling overhead. Hence, it is way more practical to empower the relays with decentralized autonomy to make their own decisions based on immediate local feedbacks and partial observability of the system state [i.e., local CSI (LCSI) and local ESI (LESI)]. These decisions are not trivial since each relay faces the uncertainty of the system state (channel, buffer, energy) and of the other relays' actions and observations. To tackle these complications, we come up with a decentralized low overhead solution by making the following contributions.

1) We rigorously formulate the delay-optimal multirelay power control problem as a partially observable identical payoff stochastic game (PO-IPSG) [25] that considers the aforementioned properties of the EH relay system. PO-IPSG is a stochastic process that is collectively controlled by a group of independent agents who lack a central view of the global system state. Nevertheless, these agents have a shared objective; i.e., they are all interested in optimizing the utility of the team as a whole. The process is decentralized because none of the agents can control the whole process, and neither of the agents has a full view of the global state. This readily corresponds to our setting in that we also assume all relays in the network collectively aim at minimizing the average number of packets waiting in the source buffer. Also, by making each relay's power control policy adaptive to a partial view of the system consisting of SBSI, its LCSI, and LESI, the formulated PO-IPSG can systematically tradeoff long-term energy-efficiency and delay performance.
2) Given our PO-IPSG formulation, we propose a *distributed learning-theoretic power control* (DLTPC) algorithm that can be used by the relays to learn their power control play strategies in the absence of statistical knowledge regarding the dynamics of channel, traffic, and energy processes. We construct DLTPC by building on and extending the classical results for gradient-based optimization of MDPs [27], [28] and PO-IPSGs [25]. We show that our algorithm harmonizes the relays' policies so that their collective behavior is provably convergent to (at least) a locally optimal solution of PO-IPSG. As it turns out, DLTPC is a particularly lightweight algorithm, and its updates on the control policy induce only little source-relay signaling overhead with no explicit message exchange between the relays.
3) By simulation, we show the suboptimality gap between DLTPC and an MDP-based optimal policy that is armed with perfect statistics. It is evidenced that DLTPC has a smaller performance margin with the centralized controller compared to existing suboptimal throughput-maximizers for EH AaF multirelay systems (e.g., [9]).

The rest of the paper is organized as follows. In Section II, we present the system model along with the general characteristics of the channel, traffic, and EH processes we assume in this paper. In Section III, we give our PO-IPSG-based formulation of the multirelay delay optimization problem. In Section IV, the DLTPC algorithm is proposed for convergence to a locally optimal solution of the formulated PO-IPSG. Section V is dedicated to the comparative evaluation of the DLTPC algorithm. The paper ends with a concluding epilogue.

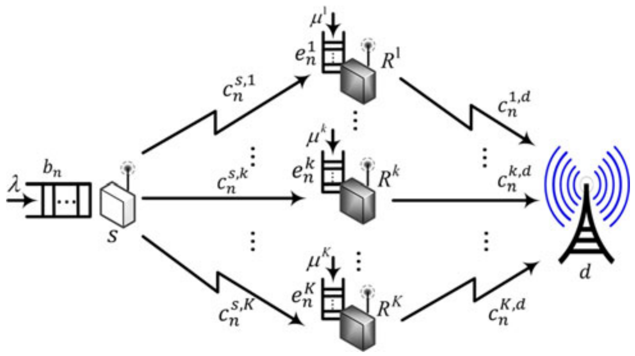

Fig. 1. Two-hop energy-harvesting cooperative relaying network.

II. System Model

In this section, we describe the two-hop relay communication system, as well as the channel, traffic, and EH models. As a notational convention, the time index appears as a subscript, while a relay's index is always a superscript. Bold symbols are used for nonscalars (i.e., vectors or sets) at the social level, collecting quantities across all relays. A symbol associated with an individual relay (be it a scalar, a vector, or a set) is never in bold.

A. Energy-Harvesting Relay Communication System

The system under consideration is a two-hop relay network with one source node s, K energy-harvesting relay terminals (each denoted by $R^k, k \in \mathcal{K} \triangleq \{1, \ldots, K\}$) and one destination node d, as illustrated in Fig. 1. It is assumed that the source node's signal cannot reach the destination directly due to its limited transmission radius, and instead relies on the relays' assistance to transmit to d. We assume that all relays operate in half-duplex mode. A two-phase AaF protocol is used for s -to- d packet delivery; more specifically, each time slot n is split into two subslots, each with duration $\tau/2$. In the first subslot, the source broadcasts its own data with full transmission power a^s to relay nodes. In the second subslot, according to the power control policy (defined in Section III-A and calculated by Algorithm 1), each relay decides whether to remain silent or to amplify the signal it has received from the source and forwards it to d. It is further assumed that the second hop transmissions by the relays are over orthogonal channels (e.g., using frequency division multiple access).

B. Channel and Physical Layer Model

We consider a frequency nonselective block fading model, where $c^{s,k} \in \mathcal{C}^{s,k}$ denotes the channel fading gain from node s to relay R^k. We use $\mathcal{C}^{s,k}$ to refer to the *local source-to-relay channel state information* space; similarly, $c^{k,d} \in \mathcal{C}^{k,d}$ is used to denote the channel gain on the R^k - d link, and $\mathcal{C}^{k,d}$ represents the *local relay-to-destination CSI* space. We define the *LCSI* space for the kt relay as $\mathcal{C}^k = \mathcal{C}^{s,k} \times \mathcal{C}^{k,d}$, where

$c_n^k = \langle c_n^{s,k}, c_n^{k,d} \rangle \in \mathcal{C}^k$ is referred to as relay R^k's LCSI at the n th time slot. Also, we use $\quad = \times_{k=1}^{K} \mathcal{C}^k$ to denote the space of the global CSI, collecting the channel gains across all the relays $R^k, k \in \mathcal{K}$.

Assumption 1: The global CSI $\boldsymbol{c}_n = \langle c_n^k \rangle_{k \in \mathcal{K}} \in \quad$ is quasi-static in each time slot. Furthermore, the process $\{\boldsymbol{c}_n\}_{n \in \mathbb{N}}$ is i.i.d. between slots with distribution $\mathbb{P}\{\boldsymbol{c}\}$. It is assumed that $\mathbb{P}\{\boldsymbol{c}\}$ is unknown and that each relay R^k is aware of only its local CSI c_n^k at time n, which can be estimated using channel reciprocity, assuming a time-division duplexing system. ■

Let x represent the broadcast information symbol with unit energy from node s. The signal received by R^k is given by

$$y_n^{s,k} = \sqrt{a^s c_n^{s,k}}\, x + \eta$$

where η is the additive white Gaussian noise. Without loss of generality, we assume that the noise power is the same over all links, denoted by σ^2. In phase 2, relay R^k amplifies $y_n^{s,k}$, and forwards it to node d with the chosen power $a_n^k \in \mathcal{A}^k$. The received signal $y_n^{k,d}$ at d is as follows:

$$y_n^{k,d} = \sqrt{a_n^k c_n^{k,d}}\, x_n^{k,d} + \eta$$

where $x^{k,d}$ is the signal sent from R^k to d, normalized to have unit energy; i.e., $x_n^{k,d} = \frac{y_n^{s,k}}{|y_n^{s,k}|}$.

Given the power profile $\boldsymbol{a}_n = \langle a_n^k \rangle_{k \in \mathcal{K}}$, the end-to-end AaF cooperative service rate is as [34]

$$r_n^{s,\mathcal{K},d} = \gamma_L W \log_2 \left(1 + \frac{\sum_{k \in \mathcal{K}} \Gamma_n^{s,\mathcal{K},d}}{\Upsilon} \right) \tag{1}$$

where W is the bandwidth for transmission, γ_L denotes a bandwidth factor that is set to 1 for energy-constrained settings, Υ is a constant denoting the capacity gap, and

$$\Gamma_n^{s,\mathcal{K},d} = \frac{a_n^k a^s c_n^{s,k} c_n^{k,d}}{\sigma^2 \left(a^s c_n^{s,k} + a_n^k c_n^{k,d} + \sigma^2 \right)} \tag{2}$$

is the relayed signal-to-noise ratio (SNR) for source node s, which is helped by relay node R^k.

C. Traffic Model and Source Buffer Dynamics

There is one buffer at the source for the storage of packets. Let l be the size of each packet and A_n be the random new packet arrival at the nth slot.

Assumption 2: The arrival process $\{A_n\}_{n \in \mathbb{N}}$ is i.i.d. with distribution $\mathbb{P}\{A\}$ and mean $\lambda = \mathbb{E}[A]$. Also, packet arrivals occur at the end of each time slot. It is further assumed that the specific form of $\mathbb{P}\{A\}$ is unknown a priori. ■

We use $b_n \in \mathcal{B}$ to denote the ***SBSI***, which is the number of packets in the source buffer at the beginning of the n th time slot. N_B denotes the maximum buffer size. When the buffer is full ($b_n = N_B$), new arrivals will be dropped. Finally, the buffer dynamics follow Lindley's equation:

$$b_{n+1} = \min \left(\left(b_n - \frac{\tau r_n^{s,\mathcal{K},d}}{2l} \right)^+ + A_n, N_B \right) \tag{3}$$

where $(.)^+$ stands for $\max(.,0)$.

D. EH and Relay Energy Storage Dynamics

The EH process at each relay is modeled as a packet arrival process (e.g., see [37]) such that each energy packet is an integer multiple of a fundamental energy unit. The relay R^k is capable of harvesting a random number H_n^k of energy packets from the environment at each time slot. The relay stores its harvested energy in its battery or a super-capacitor [26] with a finite capacity denoted by N_E^k (energy packets), and all the energy harvested when the battery is full is lost. Also, the leakage within the battery or super-capacitor and the inefficiency in storing harvested energy are assumed to be negligible. Let $e_n^k \in \mathcal{E}^k$ be the amount of renewable energy in relay R^k's energy storage at the beginning of the n th time slot. We refer to e_n^k as *local energy state information* (LESI). Also, we use $\quad = \times_{k=1}^{K} \mathcal{E}^k$ to denote the space of the global ESI, collecting all possible LESI combinations across all the relays. Similarly, $\boldsymbol{e}_n = \langle e_n^k \rangle_{k \in \mathcal{K}} \in \mathcal{E}$ is referred to as the system's global ESI at the n th time slot.

Assumption 3: The arrival process $\{H_n^k\}_{n \in \mathbb{N}}\ \forall k \in \mathcal{K}$ is i.i.d. with respect to n, and has distribution $\mathbb{P}\{H^k\}$ and mean $\mu^k = \mathbb{E}[H^k]$. We assume that the new energy arrivals are observed after the control actions are performed at each slot. It is assumed that $\mathbb{P}\{H^k\}$ and $\mathbb{E}[H^k]$ are unknown and each relay R^k is only aware of its LESI e_n^k at each time slot. ■

Let a_n^k denote the chosen power level by relay R^k at time n. The LESI dynamics for each relay R^k is as follows:

$$e_{n+1}^k = \min \left(e_n^k - a_n^k \frac{\tau}{2} + H_n^k, N_E^k \right) \tag{4}$$

where a_n^k must satisfy the following energy availability constraint:

$$a_n^k \frac{\tau}{2} \leq e_n^k \quad \forall k \in \mathcal{K}. \tag{5}$$

Finally, it is implicitly assumed that $a_n^k = 0$ means that relay R^k remains inactive in time n.

III. Problem Formulation

In this section, we formulate a decentralized power control policy for the relays to cooperatively optimize the average delay incurred by the source packets. In our system model, the dynamics of the source buffer depends, in part, on the packet arrival intensity λ, but it also depends on the cooperative service rate $r^{s,\mathcal{K},d}$ it receives from the relays, which is affected by their channel states as well as their EH profile. Accordingly, we define the power control policy at each relay to be adaptive to SBSI, as well as its LCSI and LESI. In particular, adaptation to LCSI is needed to opportunistically exploit the channel dynamics and gain more value for the power invested. SBSI-adaptability is needed to make the policy delay-aware under the conditions of unsaturated traffic and finite-length buffer at the source. Finally, given that the relays rely on EH for their operation, their control policies are subject to instantaneous energy availability constraints. An LESI-adaptive policy avoids inadvertent consumption of the harvested energy, and increases the odds that on urgent occasions a larger number of relays are available for rendering their service (i.e., higher diversity order), and they have more feasible power options at their disposal.

Our formulation is founded on the assumption that the relays would be working toward a common goal, i.e., the optimization of the incurred delay by the source packets. Altogether, our setup comes down to the coupled interaction of a number of agents with identical interest in a Markovian environment based on partial knowledge of the system state information and without explicit awareness of the action choices of the other agents. A systematic way to formulate this problem is to cast the system as a PO-IPSG [25]. We denote the PO-IPSG as a quintuple $\mathcal{G} = \langle \mathcal{K}, \ \ , \ \ , \mathrm{T}, r\rangle$. $\ \ = \mathcal{B} \times \ \ \times \mathcal{E}$ is the global system state space, where each $\boldsymbol{s}_n \in \ \ $ denotes the global system state at the nth time slot, i.e., $\boldsymbol{s}_n = \langle b_n, \boldsymbol{c}_n, \boldsymbol{e}_n \rangle$ consists of the SBSI, global CSI, and global ESI; likewise, we use $\mathcal{S}^k = \mathcal{B} \times \mathcal{C}^k \times \mathcal{E}^k$ to represent the space of partially observed system states from the viewpoint of relay R^k, $k \in \mathcal{K}$. Similarly, $s_n^k = \langle b_n, c_n^k, e_n^k \rangle$ denotes the k th relay's observed state at the nth time slot. $\ \ (\boldsymbol{e}) = \times_{k=1}^{K} \mathcal{A}^k(e^k) \forall \boldsymbol{e} \in \ \ $ is the battery state-dependent joint action space, i.e., different combinations of feasible power levels that can be chosen by the relays [see (5)]. The mapping $\mathrm{T} : \ \ \times \ \ \times \ \ \rightarrow [0,1]$ denotes the global state transition probabilities, and is discussed in more detail in Section III-B. Finally, $r : \ \ \times \ \ \times \ \ \rightarrow \mathbb{R}$ is the instantaneous reward function that is defined to be identical across all relays. More specifically, we define r as a function of the number of vacant places in the source buffer; i.e.,

$$r(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}_{n+1}) = \nu (N_B - b_{n+1}) \tag{6}$$

where ν is a positive constant. The dynamics of the game $\mathcal{G}$ proceeds as follows: at each time slot n, each relay R^k observes its local state s_n^k and selects an action a_n^k according to its power control policy u^k (to be specified in Section III-A). A composite action profile $\boldsymbol{a}_n = \langle a_n^k \rangle_{k\in\mathcal{K}}$ from the joint action space $\ \ $ is executed, the system probabilistically transitions to the next state $\boldsymbol{s}_{n+1}$ according to the law $\mathrm{T}(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{a}_n)$, and all relays receive the identical reward $r(\boldsymbol{s}_n, \boldsymbol{a}_n, \boldsymbol{s}_{n+1})$. The system-wide objective is to maximize the *value of the game*, i.e., the long-run average of the received rewards.

A. Factored Control Policy

We assume that the system is controlled by stationary policies. The stationarity of a policy implies that it depends on the history of the game only through the current state. Moreover, we parameterize the policy space by a set of continuous parameters $\boldsymbol{\Theta} \in \mathbb{R}^{\mathcal{D}}$ of some dimension $\mathcal{D}$. In particular, as we are interested in decentralized optimization with partial state observability by the relays, we restrict ourselves to the space of *factored* joint controllers $\ \ ^{\boldsymbol{\Theta}}$, where each $\ \ ^{\boldsymbol{\Theta}} \in \ \ ^{\boldsymbol{\Theta}}$ is a probabilistic mapping of the form $\ \ ^{\boldsymbol{\Theta}} : \ \ \times \ \ \rightarrow [0,1]$ and it holds that $\ \ ^{\boldsymbol{\Theta}} = \prod_{k=1}^{K} u^{\theta^k}$. Basically, $\boldsymbol{\Theta}$ is defined to be the concatenation of individual relay policy parameters, i.e., $\boldsymbol{\Theta} = \langle \theta^1, \ldots, \theta^K \rangle$, and $u^{\theta^k} : \mathcal{S}^k \times \mathcal{A}^k \rightarrow [0,1]$ is relay R^k's individual power control policy. θ^k is taken to be a $\mathcal{D}^k \triangleq |\mathcal{S}^k \times \mathcal{A}^k|$-dimensional vector of the form $\theta^k = \langle \theta_{s,a}^k \rangle_{s\in\mathcal{S}^k, a\in\mathcal{A}^k}$; i.e., the joint policy space is of dimension $\mathcal{D} = \sum_{k=1}^{K} \mathcal{D}^k$.

Remark 1: The factorization of action choice allows for parallel computation of the control policy by the relays as stated in Theorem 2 (Section IV). It also helps overcome the curse of dimensionality associated with the huge size of the joint state-action space $\ \ \times \ \ $; however, as argued in [25], a side-effect is that only a subset of policies from the full space of joint policies (corresponding to, e.g., a central nonfactored controller) can be represented. Hence, we can at the best yield the best set of policies from within the restricted space $\ \ ^{\boldsymbol{\Theta}}$. ■

A common way to express parametric policies in the literature (e.g., see [27]) is to assume a Gibbs-like distribution for the shape of $u^{\theta^k}(.)$; more precisely, the probability of choosing power level $a \in \mathcal{A}^k(e)$ by relay R^k in state $s = \langle b, c, e \rangle \in \mathcal{S}^k$ is expressed as follows:

$$u^{\theta^k}(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{\acute{a}\in\mathcal{A}^k(e)}^{\prime} \exp\left(\theta_{s,\acute{a}}^{\prime}\right)} \tag{7}$$

Note that the denominator in (7) is ensured to be nonzero by always having $a = 0$ as the feasible choice.

B. State Transition Laws

Assume a joint parametric control policy $\ \ ^{\boldsymbol{\Theta}} \in \ \ ^{\boldsymbol{\Theta}}$ is given. The probabilistic dynamics of the system state can be characterized in terms of $\ \ ^{\boldsymbol{\Theta}}$ and the mapping T, which denotes the controlled transition probabilities; more specifically, we have

$$\mathbb{P}\left\{ \boldsymbol{s}_{n+1} | \boldsymbol{s}_n, \ \ ^{\boldsymbol{\Theta}}(\boldsymbol{a}_n|\boldsymbol{s}_n) \right\} = \mathrm{T}(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{a}_n) \ \ ^{\boldsymbol{\Theta}}(\boldsymbol{a}_n|\boldsymbol{s}_n) \tag{8}$$

where (recalling Assumption 1 on i.i.d. channels), we have

$$\begin{aligned} &\mathrm{T}(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{a}_n) \\ &= \mathbb{P}\{\boldsymbol{c}_{n+1}\} . \mathrm{T}(b_{n+1}|\boldsymbol{s}_n, \boldsymbol{a}_n) \, \mathrm{T}(\boldsymbol{e}_{n+1}|\boldsymbol{e}_n, \boldsymbol{a}_n) \end{aligned} \tag{9}$$

and the source buffer state transition is as follows:

$$\mathrm{T}(b_{n+1}|\boldsymbol{s}_n, \boldsymbol{a}_n) = \\ \begin{cases} P\left\{ A_n = b_{n+1} - \left(b_n - \frac{\tau r_n^{s,\mathcal{K},d}}{2l} \right)^+ \right\}, & b_{n+1} < N_B \\ \sum\limits_{A = N_B - \left(b_n - \frac{\tau r_n^{s,\mathcal{K},d}}{2l} \right)^+}^{\infty} \mathbb{P}\{A_n = A\}, & b_{n+1} = N_B. \end{cases} \tag{10}$$

For the probabilistic transition of the global ESI, we have

$$\mathrm{T}(\boldsymbol{e}_{n+1}|\boldsymbol{e}_n, \boldsymbol{a}_n) = \prod_{k=1}^{\mathcal{K}} \mathrm{T}^k \left(e_{n+1}^k | e_n^k, a_n^k \right).$$

where

$$\mathrm{T}^k \left(e_{n+1}^k | e_n^k, a_n^k \right) = \\ \begin{cases} P\left\{ E_n^k = e_k^{n+1} - \left(e_n^k - \frac{\tau a_n^k}{2} \right) \right\}, & e_{n+1}^k < N_E^k \\ \sum\limits_{E = N_E^k - \left(e_n^k - \frac{\tau a_n^k}{2} \right)}^{\infty} \mathbb{P}\left\{ E_n^k = E \right\}, & e_{n+1}^k = N_E^k \end{cases}. \tag{11}$$

C. System-Wide Objective

As is common in infinite-horizon stochastic DP problems [21], we may seek policies that choose actions to optimize either the expected total discounted reward or the expected average-reward per step criterion. In this study, we opt for the time-averaged metric due to the following reasons.

1) The average reward criterion puts more emphasis on the long-run performance of the system and does not discount its future behavior; without prior knowledge, each byte of a file or voice packet is of equal significance and it is hardly justified to discount later packets as inherently less important.
2) Moreover, even if a formulation based on discounted-reward maximization is employed to trade off the delay experienced by recent and later packets, the discount factor needs to be chosen heuristically, which affects the performance of the derived power control policy.
3) Finally, we set the goal in PO-IPSG $\mathcal{G}$ to be the maximization of the long-run average number of empty slots in the source buffer. As we clarify in the sequel (see Remark 3), this time-averaged metric in our problem is naturally related to the mean waiting time in the source buffer, and correlates well with an objective judgment of the system performance.

Now that we have stated our rationale for choosing a time-averaged criterion, in Remark 2, we impose a mild assumption on the set of admissible policies to ensure that the time-average criterion is well-defined:

Remark 2: Similar to other literature in MDP [12], [28], we restrict our consideration to unichain policies in this paper. The stationary policy ${}^{\Theta}$ is said to be unichain if the controlled Markov chain $\{\boldsymbol{s}_n\}_{n\in\mathbb{N}}$ under ${}^{\Theta}$ is ergodic [33]. In this case, $\{\boldsymbol{s}_n\}_{n\in\mathbb{N}}$ has a unique steady state probability distribution $\boldsymbol{\pi}$, where for all $\boldsymbol{s}\in\mathcal{S}, \pi(\boldsymbol{s}) = \lim_{n\to\infty}\mathbb{P}(\boldsymbol{s}_n = \boldsymbol{s})$ [28]. Now, we may define the optimization objective as follows:

$$\max_{\Theta}\bar{\mathcal{R}}\left({}^{\Theta}\right) \triangleq \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}^{\ \Theta}\left\{r_n\right\} = \mathbb{E}^{\boldsymbol{\pi}}\left\{\nu\left(N_B - b\right)\right\} \tag{12}$$

where the $\mathbb{E}^{\boldsymbol{\pi}}$ denotes expectation w.r.t. the underlying probability $\boldsymbol{\pi}$. ■

Remark 3: We have from the *extended Little's law* (c.f., [30, Lemma 1]) that the long-run average delay $\bar{\mathcal{D}}({}^{\Theta})$ of the source packets under the (unichain) policy ${}^{\Theta}$ verifies the following inequality:

$$\bar{\mathcal{D}}\left({}^{\Theta}\right) \leq \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\frac{\mathbb{E}^{\ \Theta}\left\{b_n\right\}}{\left(1-\mathbb{P}_{drop}\right)\lambda}$$

where $\mathbb{E}^{\ \Theta}$ is the expectation under stationary policy ${}^{\Theta}$ and $\mathbb{P}_{drop}$ is the packet drop rate due to source buffer overflow. Here, we argue that since in practice we target reasonable (e.g., 0.1%) drop rates, it holds that $\mathbb{P}_{drop} \ll 1$, and therefore the following is a good approximation for the average delay:

$$\bar{\mathcal{D}}\left({}^{\Theta}\right) \approx \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\frac{\mathbb{E}^{\ \Theta}\left\{b_n\right\}}{\lambda}.$$

Furthermore, this approximation is asymptotically tight as the data buffer size increases. Therefore, for sufficiently large buffer size and low load regime, maximizing $\bar{\mathcal{R}}({}^{\Theta})$ is a valid alternative to minimizing the average delay. ■

Definition 1 (Local Optimal of PO-IPSG $\mathcal{G}$): A profile of power control policies ${}^{\Theta^*} = u^{\theta_1^*}, \ldots, u^{\theta_K^*} \in \mathcal{U}^{\Theta}$ is the local optimal of the game $\mathcal{G}$ if it satisfies the following condition:

$$\nabla_{\Theta}\bar{\mathcal{R}}\left({}^{\Theta^*}\right) = \vec{0}.$$

Theorem 1: The gradient in Definition 1 can be computed as follows:

$$\begin{aligned}&\nabla_{\Theta}\bar{\mathcal{R}}\left({}^{\Theta}\right)\\ &\quad = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\frac{\nabla_{\Theta}\mathbb{P}\left\{\left.\boldsymbol{s}_{n+1}\right|\boldsymbol{s}_n,\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right\}}{\mathbb{P}\left\{\left.\boldsymbol{s}_{n+1}\right|\boldsymbol{s}_n,\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right\}}Q\left(\boldsymbol{s}_n,\boldsymbol{a}_n\right)\end{aligned} \tag{13}$$

where the function $Q(.,.)$ is the so-called *differential reward* function defined as follows:

$$\begin{aligned}&Q\left(\boldsymbol{x},\boldsymbol{y}\right)\\ &\quad = \lim_{N\to\infty}\mathbb{E}^{\ \Theta}\left\{\sum_{n=0}^{N-1}\left(r_n - \bar{\mathcal{R}}\left({}^{\Theta}\right)\right)|\boldsymbol{s}_0 = \boldsymbol{x},\boldsymbol{a}_0 = \boldsymbol{y}\right\}.\end{aligned} \tag{14}$$

Proof: The proof follows immediately from the derivation in [28, Section 3.2]. ■

Note that (13) can be written in a more convenient form by realizing that

$$\begin{aligned}&\frac{\nabla_{\Theta}\mathbb{P}\left\{\left.\boldsymbol{s}_{n+1}\right|\boldsymbol{s}_n,\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right\}}{\mathbb{P}\left\{\left.\boldsymbol{s}_{n+1}\right|\boldsymbol{s}_n,\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right\}}\\ &\quad = \nabla_{\Theta}\ln\left[\mathbb{P}\left\{\left.\boldsymbol{s}_{n+1}\right|\boldsymbol{s}_n,\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{v}_n\right)\right\}\right]\\ &\quad = \nabla_{\Theta}\ln\left[\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right].\end{aligned} \tag{15}$$

It is worth noting that a function such as $\nabla_{\Theta}\ln[\ {}^{\Theta}(\boldsymbol{a}_n|\boldsymbol{s}_n)]$, which is the gradient of a log-likelihood, is also known as a *score function* in classical statistics [31]. Finally,

$$\begin{aligned}&\nabla_{\Theta}\bar{\mathcal{R}}\left({}^{\Theta}\right)\\ &\quad = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\nabla_{\Theta}\ln\left[\ {}^{\Theta}\left(\boldsymbol{a}_n|\boldsymbol{s}_n\right)\right]Q\left(\boldsymbol{s}_n,\boldsymbol{a}_n\right)\end{aligned} \tag{16}$$

In what follows, we present a distributed learning-theoretic procedure to steer the relays' behavior toward a delay-optimal power control policy u^{Θ^*} in the sense of Definition 1.

IV. A Multiagent Reinforcement Learning Solution

In our PO-IPSG formulation, it is desired that the relays make coordinated decisions despite their independence of one another and despite their lack of omniscience (i.e., each single relay is unaware of the other relays' local states, and the policies they are pursuing). To harmonize the relays' behavior, in this section, we present a *DLTPC* algorithm to be executed in parallel by each relay involved.

In fully observable IPSGs, *value function-based learning* methods (e.g., [32]) have been proposed for discounted reward problems, which are convergent to the optimal Nash equilibrium. As for our PO-IPSG problem, however, we resort to *policy search* methods, which have been shown to be a reasonable alternative to value-based methods for partially observable environments [36]. In particular, we follow the lead of Peshkin *et al.* [25], which introduce a general method for using gradient ascent in multiagent policy spaces to guarantee convergence to local optima (i.e., gradient zero operating points) of the game. Through a sketchy analysis, it has been shown in [25] that: when the search space is restricted to factored social policies $\mathcal{U}^{\boldsymbol{\Theta}}$, joint gradient ascent performed by a central controller (with access to observation histories of the whole system) is equivalent to parallel gradient ascent performed by individual agents (with access only to their own partial view of the system history). Key to the argument in [25] is to show that:

1) the parallel algorithm samples gradients $\nabla_{\boldsymbol{\Theta}} \bar{\mathcal{R}}$ from the correct distribution, and
2) the update increments used in gradient ascent are the same in the parallel algorithm as in the joint one.

Moreover, to satisfy these two conditions, an underlying requirement is that the agents perform synchronized updates on the estimates of their own components of the global gradient vector. Although the study in [25] is conducted in the context of discounted reward PO-IPSGs, but as we show in this paper, their line of argument can be extended to average-reward settings as well. However, the discussion in [25] is more of an outline lacking most details on the machinery of gradient estimation. We thus turn to standard techniques for estimation of the gradient of the average-reward in MDP literature [27], [28]. These algorithms typically exploit the regenerative structure of the system's underlying Markov process to obtain unbiased gradient estimates based on the observations made between regeneration times (i.e., between visits to a certain recurrent state). Applied to our PO-IPSG formulation, corresponding to every global regenerative cycle, we may define a local cycle for each relay during which it collects local observations to form an estimate of its own component of the global gradient vector. We show that at the expense of a very low signaling overhead, it can be arranged for the relays to agree on the termination of global regenerative cycles, thus satisfying the underlying requirement of synchronized updates in [25]. We then rigorously apply the line of argument in [25] to show that conditions I and II will be satisfied by our derivation (see Theorem 2 in Section IV). Based on this result, in Section IV-B, we discuss the update rules to be executed iteratively by each relay, and present DLTPC's pseudocode.

A. Decentralized Computation of the Performance Gradient

Assume that the relay communication system is controlled via some factored joint parametric control policy ${}^{\boldsymbol{\Theta}} \in \mathcal{U}^{\boldsymbol{\Theta}}$ (c.f., Section III-A). The global system history is realized as an infinite-length trajectory of the form

$$\begin{aligned} \boldsymbol{h}_{\infty} = & \left[\boldsymbol{s}_0, \boldsymbol{a}_0, r_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-1}, \boldsymbol{a}_{n-1}, r_{n-1}, \boldsymbol{s}_n, \ldots\right] \\ & \in \mathcal{H}_{\infty} \triangleq (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\infty}. \end{aligned}$$

Now, fix some $e^* \in \mathcal{E}^k \; \forall k$ and let $\boldsymbol{e}^* \in \mathcal{E}$ be the global ESI, where $e_n^k = e^* \forall k$; likewise, fix some $b^* \in \mathcal{B}$. Finally, let $\mathcal{S}^* \triangleq \{\langle b^*, \boldsymbol{c}, \boldsymbol{e}^* \rangle, \; \forall \boldsymbol{c} \in \mathcal{C}\}$. With $\{\boldsymbol{s}_n\}_{n \in \mathbb{N}}$ being ergodic, elements of $\mathcal{S}^*$ recur infinitely often within any realization of the global system history. Let t_m be the time of the m-th visit to $\mathcal{S}^*$. We refer to the following portion of history:

$$\begin{aligned} \boldsymbol{h}_m^* = & \left[\boldsymbol{s}_{t_m}, \boldsymbol{a}_{t_m}, r_{t_m}, \boldsymbol{s}_{t_m+1}, \ldots, \boldsymbol{s}_{t_{m+1}-1}, \right. \\ & \left. \boldsymbol{a}_{t_{m+1}-1}, r_{t_{m+1}-1}, \boldsymbol{s}_{t_{m+1}}\right] \end{aligned}$$

as the mth *global renewal cycle* ($m \geq 1$). Under Assumption 1 for CSI and by *regenerative property* (e.g., see [29]), these pieces of system trajectory are i.i.d. We denote by $\ell(\boldsymbol{h}_m^*)$ the length of $\boldsymbol{h}_m^*$ that is equal to $\Delta t_m = t_{m+1} - t_m$. It is also convenient to introduce local versions of a renewal cycle observed through the prism of each relay R^k. In fact, corresponding to the m-th global renewal cycle $\boldsymbol{h}_m^*$, the relay R^k's *local renewal cycle* is realized as follows:

$$\begin{aligned} h_m^{*,k} = & \left[s_{t_m}^k, a_{t_m}^k, r_{t_m}, s_{t_m+1}^k, \ldots, s_{t_{m+1}-1}^k, \right. \\ & \left. a_{t_{m+1}-1}^k, r_{t_{m+1}-1}, s_{t_{m+1}}^k\right] \end{aligned}$$

where by definition of t_m, it holds that for all $k \in \mathcal{K}$: $s_{t_m}^k, s_{t_{m+1}}^k \in \mathcal{S}_k^* \triangleq \{\langle b^*, c^k, e^* \rangle \; \forall c^k \in \mathcal{C}^k\}$; i.e., $h_m^{*,k}$ is of the same length as $\boldsymbol{h}_m^*$. Now, more generally, define $\mathcal{H}^*$ to be the space of all global renewal cycles; accordingly, $\mathcal{H}^{*,k}$ is used to refer to the space of all local renewal cycles for relay R^k. For $\boldsymbol{h}^* \in \mathcal{H}^*$, it holds that

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{h}^* | \boldsymbol{\Theta}\right) = & \prod_{n=0}^{\ell(\boldsymbol{h}^*)-1} \tau\left(\boldsymbol{s}_{[n+1, \boldsymbol{h}^*]} | \boldsymbol{s}_{[n, \boldsymbol{h}^*]}, \boldsymbol{a}_{[n, \boldsymbol{h}^*]}\right) \\ & \times \; {}^{\boldsymbol{\Theta}}\left(\boldsymbol{a}_{[n, \boldsymbol{h}^*]} | \boldsymbol{s}_{[n, \boldsymbol{h}^*]}\right) \end{aligned} \tag{17}$$

where the notation $x_{[n, \boldsymbol{h}^*]}$ is used to refer to the component of x realized at time $0 \leq n \leq \ell(\boldsymbol{h}^*)$ within $\boldsymbol{h}^*$. Now, by *renewal-reward theorem* (e.g., see [29]), the performance gradient $\nabla_{\boldsymbol{\Theta}} \bar{\mathcal{R}}(\; {}^{\boldsymbol{\Theta}})$ defined in (16) can be calculated as follows, (18) shown at the bottom of the page,

i.e., the expected total quantity earned during one cycle, normalized by the expected cycle duration. Similarly, the differential

$$\nabla_{\boldsymbol{\Theta}} \bar{\mathcal{R}}\left(\; {}^{\boldsymbol{\Theta}}\right) = \frac{\mathbb{E}^{\; \boldsymbol{\Theta}}\left\{\sum_{n=0}^{\ell(\boldsymbol{h}^*)-1} \nabla_{\boldsymbol{\Theta}} \ln\left[\; {}^{\boldsymbol{\Theta}}\left(\boldsymbol{a}_{[n, \boldsymbol{h}^*]} | \boldsymbol{s}_{[n, \boldsymbol{h}^*]}\right)\right] Q\left(\boldsymbol{s}_{[n, \boldsymbol{h}^*]}, \boldsymbol{a}_{[n, \boldsymbol{h}^*]}\right)\right\}}{\mathbb{E}^{\; \boldsymbol{\Theta}}\left\{\ell\left(\boldsymbol{h}^*\right)\right\}} \tag{18}$$

reward for $0 \le n < \ell(\boldsymbol{h}^*)$ can be written as follows:

$$Q\left(\boldsymbol{x}, \boldsymbol{y}\right) = \\ \mathbb{E}^{\;\,\Theta}\left\{\sum_{j=n}^{\ell(\boldsymbol{h}^*)-1}\left(r_{[j,\boldsymbol{h}^*]} - \bar{\mathcal{R}}\left(\;^{\Theta}\right)\right)|\boldsymbol{s}_{[n,\boldsymbol{h}^*]} = \boldsymbol{x}, \boldsymbol{a}_{[n,\boldsymbol{h}^*]} = \boldsymbol{y}\right\}. \tag{19}$$

Replacing Q with its estimate $\hat{Q}(\boldsymbol{s}_{[n,\boldsymbol{h}^*]}, \boldsymbol{a}_{[n,\boldsymbol{h}^*]}) \triangleq \sum_{j=n}^{\ell(\boldsymbol{h}^*)-1}(r_{[j,\boldsymbol{h}^*]} - \bar{\mathcal{R}}(\;^{\Theta}))$ in (18), we have

$$\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}} \triangleq \mathbb{E}^{\;\,\Theta}\left[\ell\left(\boldsymbol{h}^*\right)\right]\nabla_{\boldsymbol{\Theta}}\bar{\mathcal{R}}\left(\;^{\Theta}\right) = \sum_{\boldsymbol{h}^*\in\mathcal{H}^*}\mathbb{P}\left(\boldsymbol{h}^*|\boldsymbol{\Theta}\right) \\ \times\left\{\sum_{n=0}^{\ell(\boldsymbol{h}^*)-1}\nabla_{\boldsymbol{\Theta}}\ln\left[\;^{\Theta}\left(\boldsymbol{a}_{[n,\boldsymbol{h}^*]}|\boldsymbol{s}_{[n,\boldsymbol{h}^*]}\right)\right]\right. \\ \left.\times\,\hat{Q}\left(\boldsymbol{s}_{[n,\boldsymbol{h}^*]}, \boldsymbol{a}_{[n,\boldsymbol{h}^*]}\right)\right\} \tag{20}$$

where given that $\mathbb{E}^{\;\,\Theta}[\ell(\boldsymbol{h}^*)]$ is a positive number, $\mathbb{E}^{\;\,\Theta}[\ell(\boldsymbol{h}^*)]\nabla_{\boldsymbol{\Theta}}\bar{\mathcal{R}}(\;^{\Theta})$ can be viewed as the expected gradient direction, and the zeroes of $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}}$ are the same as those of $\nabla_{\boldsymbol{\Theta}}\bar{\mathcal{R}}(\;^{\Theta})$.

Theorem 2 in the sequel establishes that the calculation of the direction of the performance gradient $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}}$ can be done in a decentralized manner across the relays; i.e., each relay can independently calculate its individual gradient direction $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^k}}$ based on local information contained within its local renewal cycles $h^{*,k} \in \mathcal{H}^{*,k}$, and yet the ensemble of individual gradient directions recover the whole vector $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}}$.

Theorem 2: Assume $\;^{\Theta} \in \mathcal{U}^{\Theta}$. The gradient direction $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}}$ can be expressed as the vector

$$\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\boldsymbol{\Theta}}} = \langle\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^1}}, \ldots, \overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^K}}\rangle$$

in which each component $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^k}}$, $k \in \mathcal{K}$ is calculated as follows:

$$\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^k}} \triangleq \mathbb{E}^{\;\,\Theta}\left[\ell\left(\boldsymbol{h}^*\right)\right]\nabla_{\theta^k}\bar{\mathcal{R}}\left(\;^{\Theta}\right) = \sum_{h^{*,k}\in\mathcal{H}^{*,k}}\mathbb{P}\left(h^{*,k}|\boldsymbol{\Theta}\right) \\ \times\left\{\sum_{n=0}^{\ell\left(h^{*,k}\right)-1}\nabla_{\theta^k}\ln\left[u^{\theta^k}\left(a_{[n,h^{*,k}]}|s_{[n,h^{*,k}]}\right)\right]\right. \\ \left.\times\,\hat{Q}\left(s_{[n,h^{*,k}]}, a_{[n,h^{*,k}]}\right)\right\} \tag{21}$$

and

$$\hat{Q}\left(s_{[n,h^{*,k}]}, a_{[n,h^{*,k}]}\right) \triangleq \sum_{j=n}^{\ell\left(h^{*,k}\right)-1}\left(r_{[j,h^{*,k}]} - \bar{\mathcal{R}}\left(\;^{\Theta}\right)\right). \tag{22}$$

Proof: See the Appendix. ■

In essence, Theorem 2 states that if at each renewal cycle, all relays $R^k, k \in \mathcal{K}$ update their policy parameters θ^k along the gradient direction sampled from their distribution $\mathbb{P}(h^{*,k}|\boldsymbol{\Theta})$ in parallel, the parameter vector $\boldsymbol{\Theta}$ gets updated along the gradient direction sampled from $\mathbb{P}(\boldsymbol{h}^* = \langle h^{*,1}, .., h^{*,K}\rangle|\boldsymbol{\Theta})$; i.e., the distributed algorithm is sampling from the correct distribution. Also, due to factorization, the update increments $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^k}}$ to be used in relay R^k's gradient ascent are independent of the parameters in other relays' policies. Hence, the policy learning and control can be distributed among relays without requiring that they be informed of each others' states and choices of actions.

B. Distributed Learning-Theoretic Power Control (DLTPC)

In this section, we present DLTPC (Algorithm 1), which can lead the relays' collective behavior to a locally optimal delay performance. DLTPC relies on sample estimates of the performance gradient obtained during the actual system runtime to perform gradient-ascent in policy space. Hence, our algorithm does not need the explicit knowledge of the CSI, SBSI, and ESI statistics, and is an instance of model-free learning. This is as opposed to doing exact gradient-ascent, which requires the explicit knowledge of the transition laws T to analytically compute the gradient direction. In DLTPC, each relay updates its policy parameter θ^k_m at the end of each renewal cycle, i.e., between visits to $\mathcal{S}^*$ (see (27) in Algorithm 1). To understand (27), note that according to (21) and (22), we can use

$$F^k_m \triangleq \sum_{n=t_m}^{t_{m+1}-1}\left.\frac{\partial\ln\left[u^{\theta^k}\left(a^k_n|s^k_n\right)\right]}{\partial\theta^k}\right|_{\theta^k=\theta^k_m} \\ \times\sum_{j=n}^{t_{m+1}-1}\left(r_j - \bar{\mathcal{R}}\left(\;^{\Theta}\right)\right) \tag{23}$$

as the mth cycle estimate of $\overrightarrow{\nabla^{\bar{\mathcal{R}}}_{\theta^k}}$, which is obtained by each relay R^k from the sample renewal cycle $h^{*,k}_m$. To allow for more efficient recursive implementation of the summation (23) in Algorithm 1, we rewrite F^k_m as follows:

$$F^k_m = \sum_{n=t_m}^{t_{m+1}-1}\left(r_n - \bar{\mathcal{R}}\left(\;^{\Theta}\right)\right) \\ \times\sum_{j=t_m}^{n}\left.\frac{\partial\ln\left[u^{\theta^k}\left(a^k_n|s^k_n\right)\right]}{\partial\theta^k}\right|_{\theta^k=\theta^k_m} \tag{24}$$

which makes it possible to incrementally construct F^k_m using transient quantities z^k_n and g^k_n before reaching the end of each cycle. Accordingly, (27) in the pseudocode is basically the standard rule for stochastic gradient-ascent in which the parameter $\alpha_m \in \mathbb{R}^+$ denotes a learning rate. Also, similarly to [27], $\bar{\mathcal{R}}(\;^{\Theta})$ in (24) is replaced via its estimate $\hat{\mathcal{R}}_m$, which is also updated at each renewal cycle via the recursion

$$\hat{\mathcal{R}}_{m+1} := \hat{\mathcal{R}}_m + \alpha_m\sum_{n=t_m}^{t_{m+1}-1}\left(r_n - \hat{\mathcal{R}}_m\right). \tag{25}$$

Equation (25) is a stochastic approximation of the average reward $\bar{\mathcal{R}}(\quad^{\Theta})$, and is consistent with the observation that for the mth cycle, it holds

$$\bar{\mathcal{R}}_{\Theta_m}\left(\approx\hat{\mathcal{R}}_m\right) = \frac{\mathbb{E}\quad^{\Theta}\left\{\sum_{n=t_m}^{t_{m+1}-1} r_n\right\}}{\mathbb{E}\quad^{\Theta}\left\{\Delta t_m\right\}}. \tag{26}$$

Theorem 3: Choose α_m such that the sequence $\{\alpha_m\}$ be diminishing (i.e., $\alpha_m \overset{m\uparrow\infty}{\rightarrow} 0$), un-summable (i.e., $\sum_m \alpha_m = \infty$) but square summable (i.e., $\sum_m \alpha_m^2 < \infty$). Also, consider the sequence of parameters $\{\boldsymbol{\Theta}_m\}$ generated by Algorithm 1. Then, $\{\hat{\mathcal{R}}_m\}$ converges (with probability 1), and the profile of power control policies $\{u^{\boldsymbol{\Theta}_m}\}$ converges to the local optimal of PO-IPSG $\mathcal{G}$, i.e., $\nabla_{\boldsymbol{\Theta}}\bar{\mathcal{R}}(u^{\boldsymbol{\Theta}_m}) \overset{m\uparrow\infty}{\rightarrow} 0\ (w.p.1)$.

Proof: With this setup, DLTPC's update equations in (27) and (28) are exactly along the lines of the single-agent iterates in [27, (15) and (16)]; hence, the convergence of the gradient components (with respect to $\theta^k\ \forall k$) of the performance measure $\bar{\mathcal{R}}(u^{\boldsymbol{\Theta}_m})$ to zero can be established via the same arguments made in ([27], Proposition 3). Combine this with Theorem 2 to conclude. ■

C. Discussion and Directions for Future Research

In this section, we give a few remarks about the underlying assumptions in this paper, and discuss how relaxing these assumptions can serve as a basis for future research.

The first issue has to do with our assumption on altruistic participation of the relays in forwarding the source signal. In fact, a relay's willingness to cooperate is taken for granted and our game-theoretic formulation is only a means to perform decentralized coordination and control and not a means of cooperation stimulation. A potential future direction thus includes extensions to systems with self-interested relaying terminals, where acquiring service from the relays requires an incentive mechanism.

The second issue is regarding the extension of our system model to the case in which the source node also uses a state-dependent law to control its transmit power for minimizing the delay at its queue. While, ideally, the source power should be treated as yet another "degree of freedom," we argue, however, that such extension is nontrivial as an adaptive source would induce nonstationary dynamics on the power adjustment procedure performed by the relays. In fact, proposing a systematic mechanism for jointly controlling the source and relays' power is beyond the scope of this paper since we cannot naively consider the source node as another player in our PO-IPSG formulation. Therefore, in Section II, we have explicitly restricted our system model to the case in which the source is transmitting with a constant power (e.g., maximum allowed power). That being said, there exists, however, some fair justifications in support of our simplifying assumption: the source node in our system model does not rely on harvested energy but is instead connected to a fixed power supply. Also, no direct communication link is assumed between the source and the destination node. As such, it is fairly reasonable that the source can tap into its energy supply to power its transmission with little concern for replenishment of its energy budget. When the source node is a nonharvesting entity, there are several works in the context of EH relay systems, where the source power is assumed fixed [8].

Algorithm 1: Distributed Learning-Theoretic Power Control

```
Initialization: Set iteration index $n := 0$, renewal cycle
index $m := 0$, initial transient differential reward $\hat{Q}_0 := 0$,
initial estimate for the average reward $\hat{\mathcal{R}}_0 := 0$; Initialize
parameter vector $\theta_0^k$ randomly and set
$z_0^k := \vec{0},\ g_0^k := \vec{0},\ \forall k \in \mathcal{K}$;
Source $s$ broadcasts data and its buffer state $b_0$;
while (TRUE)
    for each relay $k \in \mathcal{K}$ do
        1) Choose power $a_n^k : u_k^{\theta_m^k}(.|s_n^k)$;
        2) Transmit data to destination $d$ with power $a_n^k$;
        3) Inform $s$ only if battery level $e_{n+1}^k$ has reached $e^*$;
        4) Receive data from $s$ along with the next buffer state
        $b_{n+1}$, and the cycle termination signal
        $\sigma_n \underline{\underline{\mathrm{def}}} \begin{cases} 1, & \boldsymbol{e}_{n+1} = \boldsymbol{e}^* \text{ and } b_{n+1} = b^* \\ 0, & \text{default} \end{cases};$
        5) Update transient quantities for gradient and
           differential reward:
           // Calculate immediate reward:
           $r_n := \nu(N_B - b_{n+1})$;
           // Update the transient differential reward estimate:
           $\hat{Q}_{n+1} := \hat{Q}_n + (r_n - \hat{\mathcal{R}}_m)$;
           // Update the transient gradient estimate:
           $z_{n+1}^k := z_n^k + \frac{\partial \ln[u^{\theta^k}(a_n^k|s_n^k)]}{\partial\theta^k}\Big|_{\theta^k=\theta_m^k}$;
           $g_{n+1}^k := g_n^k + (r_n - \hat{\mathcal{R}}_m)z_{n+1}^k$;
        6) if $(\sigma_n == 1)$ // The end of the $m$-th renewal cycle
           // Update policy parameter:
                   $\theta_{m+1}^k := \theta_m^k + \alpha_m g_{n+1}^k$;          (27)

           // Update the average reward estimate:
                   $\hat{\mathcal{R}}_{m+1} := \hat{\mathcal{R}}_m + \alpha_m \hat{Q}_{n+1}$;          (28)

           // Reset transient quantities:
           $g_{n+1}^k = \vec{\mathbf{0}},\ \hat{Q}_{n+1} := 0,\ z_{n+1}^k = \vec{\mathbf{0}}$;
           // Update the cycle index:
           $m := m + 1$;
        end if
    end for
    $n := n + 1$;        // Update the time index.
end while
```

Finally, we need to discuss the case of buffer-aided relaying, where the relay nodes have data queues as well. Cooperative networks with buffer-aided relays have the advantage that their achievable diversity is not bottlenecked by transmission order (unlike the stream-like communication in the conventional case where at each time slot, signal transmission starts from source and is then relayed to the destination) [41]. However, these relays may also incur larger packet delays, which can be quite diverse for different packets. Hence, from the application point of view, the lack of a data buffer at the relays in our work can be justified by arguing that it is to advocate a simple relay design while also minimizing packet delay, which is desirable in certain applications. There are also some technical complications in the way of extending the proposed approach to the case of relays with buffers. Reasonably enough, in buffer-aided relaying, it is typically the case that at each slot, only one relay is selected for either transmission or reception. This necessitates an explicit link selection mechanism that does not fit well with the collaborative all-playing nature of our PO-IPSG formulation and its identical-payoff structure. The systematic way to account for buffer-aided relaying is again a formulation based on stochastic DP; however, to come up with a realistic scalable solution, we need to take on a different approach for problem decomposition. There are some studies along this line (e.g., see [17]) that address delay optimization in the context of buffer-aided relaying by exploiting the structural properties inherent to the problem. The setup considered in [17], however, only consists of a single relay that gives the problem a nice weakly coupled structure amenable to decomposition into sub-problems.

V. Performance Evaluation

In this section, we evaluate the performance of our proposed DLTPC algorithm for decentralized power control in EH multi-relay systems. We compare DLTPC's performance with three other power control schemes:

1) *Centralized MDP with perfect statistics*: We assume that an MDP controller exists that is aware of the probability distributions of the channel fading $\mathbb{P}\{\boldsymbol{c}\}$, traffic arrival $\mathbb{P}\{A\}$, as well as the energy arrival processes $\mathbb{P}\{H^k\}$ for all relays $k \in \mathcal{K}$. Armed with this knowledge, one can use standard solution methods (e.g., relative value iteration [23]) to solve for an optimal joint power control policy $u : \quad \rightarrow \quad$, which maximizes an average reward measure defined similarly as (12). While in principle, this method can obtain superior performance compared to DLTPC, it suffers from both curses of dimensionality and modeling, and therefore has no practical relevance. However, the reward measure obtained using this procedure can serve as an upper bound against which to compare the DLTPC's performance.
2) *Harvesting rate* (*HR*) *assisted scheme* [9]: The online-HR scheme proposed in [9] is a centralized online (suboptimal) algorithm for joint relay selection and power allocation in multi-relay AaF EH cooperative communication systems. However, unlike DLTPC, online-HR assumes infinite backlog at the source (saturated traffic assumption), and aims at maximizing the throughput. To make online decisions, the approach in [9] uses the causal information of ESI and CSI but also needs the statistics of the harvesting and channel processes. The setup in [9] considers the case wherein the source node is also an EH entity; therefore, in our simulation, we remove this restriction and assume a continuous power supply for the source to make it comparable with DLTPC. At each slot, using the knowledge of mean HR and average channel SNRs, online-HR first determines the transmit power of the relays via a closed-form formula, and then a simple (centralized) optimization is solved to determine the relay with the maximum throughput.
3) *Naive scheme* [9]: This algorithm is also centralized and online; however, it does not require the statistics of the harvesting and channel processes. At each time slot, the relays use their stored energies as their transmit powers. Using these transmit powers, the equivalent SNRs for all links are calculated. Then, the relay with the maximum equivalent SNR among all is selected to forward the signal to destination.

In what follows, we first compare the computational complexity of DLTPC with Online-HR and Online-Naive, and then present our numerical results in Section V-B.

A. Comparison of Computational Complexity

At each time step, the Online-HR algorithm [9] has to compute the maximum system throughput achievable by every relay and then select the relay with the best value. Hence, its complexity is $O(K)$ in each time step (i.e., linear in the number of relays). The Online-Naive algorithm has also the complexity of $O(K)$ per time step as it needs to select the relay, which provides the maximum equivalent SNR among all the relays. Both these algorithms are centralized and need to gather global information from the whole network for their operations. On the other hand, DLTPC is a particularly lightweight algorithm, working with minimal message signaling overhead between source and relays (see steps 3 and 4 in Algorithm 1). The algorithm's update rules are written in terms of efficient recursive formulae, which lead to negligible complexity. Also, if the policy function for each relay is chosen to have the convenient form in (7), the score function at step 5 can simply be calculated as follows:

$$\left.\frac{\partial \ln\left[u^{\theta^k}\left(a_n^k|s_n^k\right)\right]}{\partial \theta^k}\right|_{\theta^k=\theta_m^k} = \begin{cases} 1-u^{\theta^k}\left(a|s\right)\Big|_{\theta^k=\theta_m^k}, & a=a_n^k, s=s_n^k \\ -u^{\theta^k}\left(a|s\right)\Big|_{\theta^k=\theta_m^k}, & a\neq a_n^k, s=s_n^k \\ 0, & s\neq s_n^k \end{cases}.$$

Therefore, at each time step, DLTPC needs just a few standard algebraic operations, along with one random number generation to calculate the next action.

B. Numerical Evaluation

We consider a setup with a total of $K = 8$ relays. The time slot duration is $\tau = 2\,\text{ms}$. We assume Poisson packet arrival with mean rate λ pkt/ms, and the packet size is 1024 B. The total bandwidth is $W = 2.5$ MHz. The source buffer is quantized to have ten states (i.e., $N_B = 9$ pkts). Moreover, we assume that all relays harvest energy according to a Poisson energy arrival with mean rate $\mu^k = 0.25$ energy pkt/ms $\forall k$, and the renewable

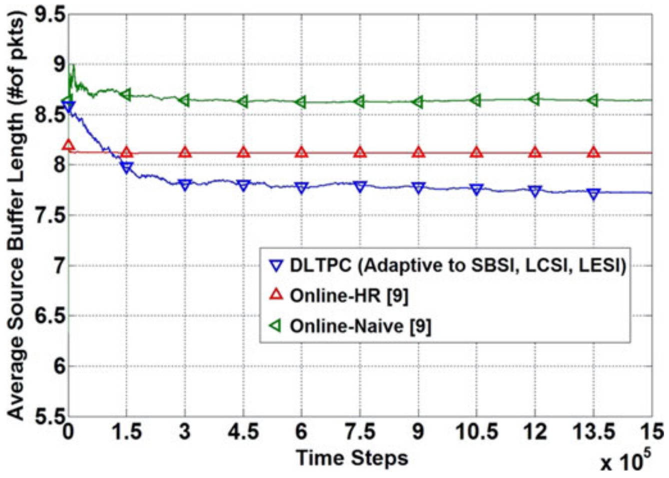

Fig. 2. Progression of average source buffer length.

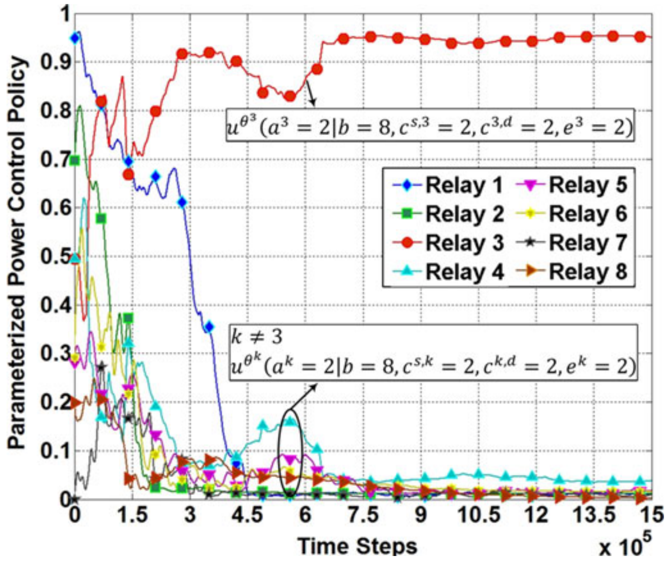

Fig. 3. Progression of power control policies.

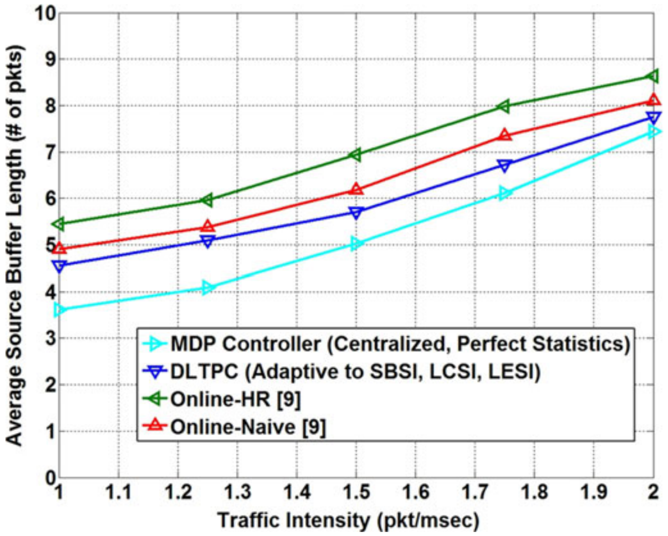

Fig. 4. Impact of input traffic intensity on delay performance.

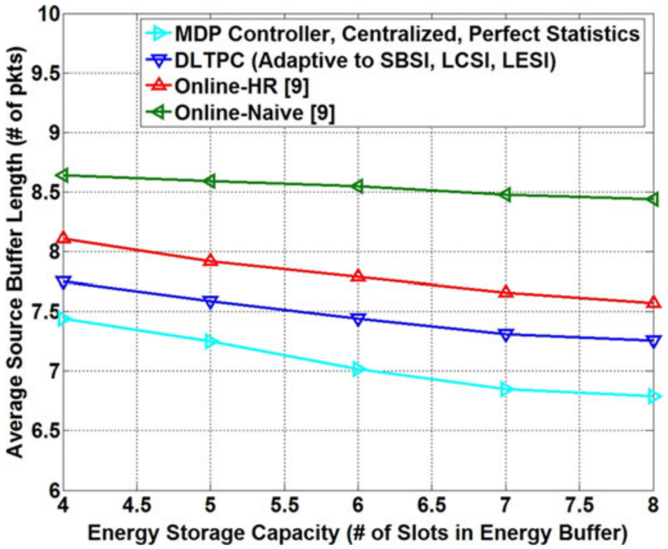

Fig. 5. Impact of energy storage capacity on delay performance.

energy is stored in a battery with maximum capacity $N_E^k = 4$ (energy pkts). The source transmission power is fixed at 5 (energy $\mathrm{pkt/ms}$). Although our algorithm does not use the knowledge of the channel model, for the purpose of experiments, we simulate Rayleigh fading for each link. In this model, the channel states $c^{s,k}$ and $c^{k,d}$ $(\forall k)$ are exponentially distributed random variables. However, as we consider a finite number of possible states, digital quantization is used to discretize the channel states. In particular, all the channel states are quantized into six probability bins with the boundaries specified as: $\{(-\infty, -5.41 \text{ dB}), [-5.41, -1.59 \text{ dB}), [-1.59, -0.08 \text{ dB}), [-0.08, 1.42 \text{ dB}), [1.42, 3.18 \text{ dB}), [3.18 \text{ dB}, \infty)\}$. Over these bins, the stochastic evolution of channel states is i.i.d. across time and independent across users. This discretization of channel states have been justified in [40]. We choose $\langle b^*, \boldsymbol{c}, \boldsymbol{e}^* \rangle = \langle N_B, \ldots, (N_E^k)_k \rangle$ as the recurrent state marking the renewal cycles for DLTPC. Also, the initial learning rate is taken to be $\alpha_0 = 2.5 \times 10^{-4}$ and is diminished every 100 renewal cycles by a factor of 0.9.

Fig. 2 plots the progression of the average source buffer length over time under DLTPC along with the two other suboptimal policies. The mean data arrival rate is fixed at 2.0 pkt/ms. As can be seen, both the online-HR and online-naive schemes converge much more quickly, but are outperformed by DLTPC in the limit. In Fig. 3, we plot the policy of all relays (for one particular state-action pair) as the joint policy is driven toward the local optimal of the PO-IPSG.

Fig. 4 illustrates the average number of occupied slots in source buffer under various traffic intensities (λ is varied from 1 to 2 pkt/ms). As a general trend, the source buffer gets more occupied as packet arrival rate increases. As expected, the MDP controller has the best performance gain among the

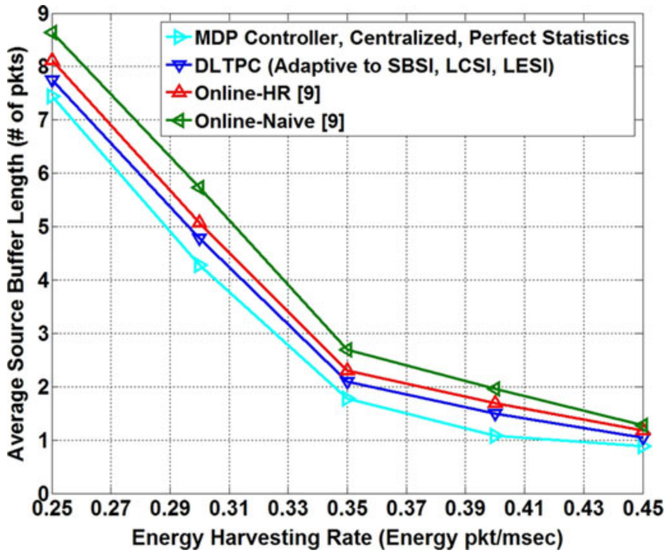

Fig. 6. Impact of energy HR on delay performance.

four schemes. However, compared to the other two suboptimal policies, our SBSI-adaptive DLTPC algorithm maintains a smaller suboptimality gap.

Next, we investigate the impact of the relays' HR μ^k and battery capacity N_E^k on delay performance. The mean Poisson data packet arrival rate is assumed to be 2.0 pkt/ms. In Fig. 5, we assume that the mean Poisson energy arrival rate for all relays is 0.25 energy pkt/ms, and plot the average number of occupied slots in source buffer for different values of battery size N_E^k (from 4 to 8 energy pkts). The delay performance generally improves as battery capacity increases. However, DLTPC and online-HR can better exploit the enlarged energy storage with respect to the naive policy.

In Fig. 6, we fix the battery size N_E^k to 4 energy pkts, and instead vary the mean Poisson energy arrival rate for all relays from 0.25 to 0.45 energy $\mathrm{pkt/ms}$. As expected, the source buffer receives a higher service rate as the relays' HR increases. In both plots, it is observed that our DLTPC algorithm maintains a better performance margin with respect to the centralized MDP controller.

VI. Conclusion

The design of new protocols for cooperative networks with EH nodes is a promising research direction that incorporates cooperative benefits (diversity, capacity, etc.) with the EH concept. In pure EH relay systems, the nodes run on the energy harvested from the environment and so are limited by their generation and storage capacities. This, together with the stochastic nature of the profile of the harvested energy, calls for the design of novel control policies, which optimally utilize the power for meeting the application demands. However, the majority of the existing schemes have considered the case of single-relay SRD systems and have focused on the optimization of the physical layer throughput by assuming nonbursty traffic arrival at the source. Also, the dominant methodologies for the optimization of these systems have been either offline optimizations assuming the availability of acausal information on the exact energy arrival instants and amounts or online optimizations that rely on precise statistical knowledge of the system. In this paper, we considered an EH relaying system consisting of a bursty source with finite data buffer size whose transmission is cooperatively assisted by multiple EH relays. To optimize the average delay experienced by the source packets, we proposed a learning-theoretic solution that operates in the absence of prior knowledge of the statistics of the channel variation, traffic arrival, and EH processes. The proposed method is highly decentralized and induces very low control overhead. Numerical evaluations demonstrated the superior delay performance of our solution compared to existing heuristics.

Appendix
Proof of Theorem 2

First, note that

$$\hat{Q}\left(\boldsymbol{s}_{[n,\boldsymbol{h}^*]},\boldsymbol{a}_{[n,\boldsymbol{h}^*]}\right)=\hat{Q}\left(s_{[n,h^{*,k}]},a_{[n,h^{*,k}]}\right). \tag{29}$$

Now, by substituting $^{\boldsymbol{\Theta}}=\prod_{i=1}^{K}u^{\theta^i}$ in (20), it holds that

$$\mathbb{E}^{\boldsymbol{\Theta}}\left[\ell\left(\boldsymbol{h}^*\right)\right]\nabla_{\theta^k}\bar{\mathcal{R}}\left(\ ^{\boldsymbol{\Theta}}\right)=\sum_{\boldsymbol{h}^*\in\mathcal{H}^*}\mathbb{P}\left(\boldsymbol{h}^*|\boldsymbol{\Theta}\right)\left\{\sum_{n=0}^{\ell\left(h^{*,k}\right)-1}\nabla_{\theta^k}\right.$$
$$\left.\times\ln\left[\prod_{i=1}^{K}u^{\theta^i}\left(a_{[n,h^{*,i}]}|s_{[n,h^{*,i}]}\right)\right]\hat{Q}\left(s_{[n,h^{*,k}]},a_{[n,h^{*,k}]}\right)\right\} \tag{30}$$

$$=\sum_{\boldsymbol{h}^*\in\mathcal{H}^*}\mathbb{P}\left(\boldsymbol{h}^*|\boldsymbol{\Theta}\right)\left\{\sum_{n=0}^{\ell\left(h^{*,k}\right)-1}\left[\sum_{i=1}^{K}\nabla_{\theta^k}\right.\right.$$
$$\left.\left.\times\ln\left[u^{\theta^i}\left(a_{[n,h^{*,i}]}|s_{[n,h^{*,i}]}\right)\right]\right]\hat{Q}\left(s_{[n,h^{*,k}]},a_{[n,h^{*,k}]}\right)\right\} \tag{31}$$

$$=\sum_{\boldsymbol{h}^*\in\mathcal{H}^*}\mathbb{P}\left(\boldsymbol{h}^*|\boldsymbol{\Theta}\right)\left\{\sum_{n=0}^{\ell\left(h^{*,k}\right)-1}\nabla_{\theta^k}\ln\left[u^{\theta^k}\left(a_{[n,h^{*,k}]}|s_{[n,h^{*,k}]}\right)\right]\right.$$
$$\left.\times\hat{Q}\left(s_{[n,h^{*,k}]},a_{[n,h^{*,k}]}\right)\right\} \tag{32}$$

where the last equality is due to $\nabla_{\theta^k}\ln[u^{\theta^i}(a_{[n,h^{*,i}]}|s_{[n,h^{*,i}]})]=0$ for all $i\neq k$. Now, the entire term within the curly brackets in (32) can be written as a function $\phi(.)$ of relay k's local renewal cycle $h^{*,k}$, i.e.,

$$\phi(h^{*,k})\triangleq\left\{\sum_{n=0}^{\ell(h^{*,k})-1}\nabla_{\theta^k}\ln[u^{\theta^k}(a_{[n,h^{*,k}]}|s_{[n,h^{*,k}]})]\right.$$
$$\left.\times\hat{Q}(s_{[n,h^{*,k}]},a_{[n,h^{*,k}]})\right\}.$$

Also, given that the global renewal cycle $\boldsymbol{h}^*$ can be described as the collection $\langle h^{*,1}, .., h^{*,K} \rangle$ of local renewal cycles across all relays, we have

$$\begin{aligned}
&\sum_{\boldsymbol{h}^* \in \mathcal{H}^*} \mathbb{P}\left(\boldsymbol{h}^* | \boldsymbol{\Theta}\right) \phi\left(h^{*,k}\right) \\
&\quad = \sum_{\langle h_1^*, .., h_K^* \rangle \in \mathcal{H}^*} \mathbb{P}\left(\langle h^{*,1}, \ldots, h^{*,K} \rangle | \boldsymbol{\Theta}\right) \phi\left(h^{*,k}\right) \\
&\quad = \sum_{h_k^* \in \mathcal{H}_k^*} \left[\sum_{\langle h^{*,1}, \ldots h^{*,k-1}, h^{*,k+1}, \ldots, h^{*,K} \rangle} \mathbb{P}\left(\langle h^{*,1}, .., h^{*,K} \rangle | \boldsymbol{\Theta}\right) \right] \\
&\qquad \times \phi\left(h^{*,k}\right) = \sum_{h^{*,k} \in \mathcal{H}^{*,k}} \mathbb{P}\left(h^{*,k} | \boldsymbol{\Theta}\right) \phi\left(h^{*,k}\right). \qquad (33)
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
\overrightarrow{\nabla_{\theta^k}^{\bar{\mathcal{R}}}} = &\sum_{h^{*,k} \in \mathcal{H}^{*,k}} \mathbb{P}\left(h^{*,k} | \boldsymbol{\Theta}\right) \left\{ \sum_{n=0}^{\ell\left(h^{*,k}\right)-1} \nabla_{\theta^k} \right. \\
&\left. \times \ln\left[u^{\theta^k}\left(a_{[n,h^{*,k}]} | s_{[n,h^{*,k}]}\right)\right] \hat{Q}\left(s_{[n,h^{*,k}]}, a_{[n,h^{*,k}]}\right) \right\}. \qquad (34)
\end{aligned}$$

References

[1] J. N. Laneman and G. W. Wornell, "Energy efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Chicago, IL, USA, Oct. 2000, pp. 7–12.

[2] M. Dail *et al.*, "Survey on cooperative strategies for wireless relay channels," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 9, pp. 926–942, 2014.

[3] A. Ikhlef, D. S. Michalopoulos, and R. Schober, "Buffers improve the performance of relay selection," in *Proc. IEEE Global Telecommun. Conf.* Houston, TX, USA, Dec. 2011, pp. 1–6.

[4] K. J. R. Liu, A. K. Sadek, W. Su, and A. Kwasinski, *Cooperative Communications and Networking*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[5] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Trans. Embedded Comput. Syst.*, vol. 6, no. 4, pp. 1–35, Sep. 2007.

[6] K-H. Liu and P. Lin, "Toward self-sustainable cooperative relays: state of the art and the future," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 56–62, Jun. 2015.

[7] Y. Xia, H. Chen, L. Fan, and F. Dai, "Optimal power control for source and relay in energy harvesting relay networks," in *Proc. 8th Int. ICST Conf. Commun. Netw.*, Guilin, China, Aug. 2013, pp. 942–947.

[8] B. Medepally and N. B. Mehta, "Voluntary energy harvesting relays and selection in cooperative wireless networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3543–3553, Nov. 2010.

[9] I. Ahmed, A. Ikhlef, R. Schober, and R. K. Mallik, "Joint power allocation and relay selection in energy harvesting Aaf relay systems," *IEEE Wireless Commun. Lett.*, vol. 2, no. 2, pp. 239–242, Apr. 2013.

[10] A. Minasian, S. ShahbazPanahi, and R. S. Adve, "Energy harvesting cooperative communication systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6118–6131, Nov. 2014.

[11] H. Li, N. Jaggi, and B. Sikdar, "Relay scheduling for cooperative communications in sensor networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 2918–2928, Sep. 2011.

[12] I. Krikidis, T. Charalambous, and J. S. Thompson, "Stability analysis and power optimization for energy harvesting cooperative networks," *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 20–23, Jan. 2012.

[13] D. Gunduz and B. Devillers, "Two-hop communication with energy harvesting," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Dec. 2011, pp. 201–204.

[14] Y. Luo, J. Zhang, and K. B. Letaief, "Optimal scheduling and power allocation for two-hop energy harvesting communication systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4729–4741, Sep. 2013.

[15] O. Orhan and W. Erkip, "Energy harvesting two-hop networks: optimal policies for the multi-energy arrival case," in *Proc. IEEE 35th Sarnoff Symp.*, Newark, NJ, USA, May 2012, pp. 1–6.

[16] Z. Ding, S. M. Perlaza, I. Esnaola, and H.V. Poor, "Power allocation strategies in energy harvesting wireless cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 846–860, Feb. 2014.

[17] F. Zhang and V. K. N. Lau, "Delay-optimal multi-flow buffered decode-and-forward relay communications with limited renewable energy storage," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, 2012, pp. 1351–1355.

[18] C. Huang, R. Zhang, and S. Cui, "Throughput maximization for the gaussian relay channel with energy harvesting constraints," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1469–1479, Aug. 2013.

[19] I. Ahmed, A. Ikhlef, R. Schober, and R. K. Mallik, "Power allocation for conventional and buffer-aided link adaptive relaying systems with energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1182–1195, Mar. 2014.

[20] O. Orhan and E. Erkip, "Energy harvesting two-hop communication networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2658–2670, Dec. 2015.

[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th. ed. Belmont, MA, USA: Athena Scientific, vol. II, 2012.

[22] B. Schein and R. G. Gallager, "The Gaussian parallel relay network," in *Proc. Int. Symp. Inf. Theory*, Sorrento, Italy, Jun. 2000, p. 22.

[23] M. L. Putterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley-Interscience, 2005.

[24] A. Cassandra, L. Kaelbling, and M. Littman, "Acting optimally in partially observable stochastic domains," in *Proc. Nat. Conf. Artif. Intell.*, 1994, vol. 2. pp. 1023–1028.

[25] L. Peshkin, K-E. Kim, N. Meuleau, and L.P. Kaelbling, "Learning to cooperate via policy search," in *Proc. Conf. Uncertainty Artif. Intell.,* 2000, pp. 489–496.

[26] F. Simjee and P. H. Chou, "Everlast: Long-life, supercapacitor-operated wireless sensor node," in *Proc. Int. Symp. Low Power Electron. Des.*, 2006, pp. 197–202.

[27] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of Markov reward processes," *IEEE Trans. Autom. Control*, vol. 46, no. 2, pp. 191–209, Feb. 2001.

[28] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. New York, NY, USA: Springer, 2007.

[29] P. Bremaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York, NY, USA: Springer, 1999.

[30] Y. Cui, V. K. N. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems—large deviation theory, stochastic lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1677–1701, Mar. 2012.

[31] R. Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*. New York, NY, USA: Wiley, 1993.

[32] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 9–14, 2002, vol. 15, pp. 1571–1578.

[33] J. N. Tsitsiklis, "NP-hardness of checking the unichain condition in average cost MDPs," *Operations Res. Lett.*, vol. 35, no. 3, pp. 319–323, May 2007.

[34] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[35] M.-L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "On energy harvesting gain and diversity analysis in cooperative communications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2641–2657, Dec. 2015.

[36] P. L. Bartlett and J. Baxter, "Stochastic optimization of controlled partially observable Markov decision processes," in *Proc. IEEE 39th Conf. Decision Control*, 2000, pp. 124–129.

[37] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: optimal policies," *IEEE J. Sel. Areas Commun.* vol. 29, no. 8, pp. 1732–1743, Sep. 2011.

[38] X. Huang, T. Han, and N. Ansari, "On green-energy-powered cognitive radio networks," *IEEE Commun. Surveys Tut.*, vol. 17, no. 2, pp. 827–842, 2nd Quarter 2015.

[39] X. Huang and N. Ansari, “Data and energy cooperation in relay-enhanced OFDM systems,” in *Proc. IEEE Int. Conf. Commun.*, 2016, pp. 1–6.

[40] H. Wang and N. Mandayam, “A simple packet transmission scheme for wireless data over fading channels,” *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1055–1059, Jul. 2004.

[41] G.-X. Li, C. Dong, D. Liu, G. Li, and Y. Zhang, “Outage analysis of dual-hop transmission with buffer aided amplify-and-forward relay,” in *Proc. IEEE 80th Veh. Technol. Conf.*, 2014, pp. 1–5.

Vesal Hakami received the B.S. degree in computer engineering (software) and the M.S. and Ph.D. degrees in information technology (computer networking), all from Amirkabir University of Technology, Tehran, Iran, in 2004, 2008, and 2015, respectively.

Following graduation, he has served as a Research Consultant in Iran Telecommunications Research Center (ITRC), working on stadardization issue for future wireless networks. In 2016, he joined as an Assistant Professor to the Department of Computer Engineering, Iran University of Science and Technology, Tehran. His current research mainly focuses on cognitive control of computer networks using stochastic control theory and game-theoretic learning.

Mehdi Dehghan (M’10) received the B.S. degree in computer engineering from Iran University of Science and Technology, Tehran, Iran, in 1992, and the M.S. and Ph.D. degrees from Amirkabir University of Technology (AUT), Tehran, in 1995 and 2001, respectively.

He is currently a Professor with the Department of Computer Engineering and Information Technology at AUT. Before joining AUT in 2004, he was a Research Scientist at Iran Telecommunication Research Center working in the area of network quality-of-service and management. His research interests are in wireless networks, pattern recognition, fault-tolerant computing, and distributed systems.