Optimization vs. Reinforcement Learning for Wirelessly Powered Sensor Networks

Ayça Özçelikkale*, Mehmet Koseoglu† and Mani Srivastava†
*Signals and Systems, Uppsala University, Uppsala, Sweden
†Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA

***Abstract*—We consider a sensing application where the sensor nodes are wirelessly powered by an energy beacon. We focus on the problem of jointly optimizing the energy allocation of the energy beacon to different sensors and the data transmission powers of the sensors in order to minimize the field reconstruction error at the sink. In contrast to the standard ideal linear energy harvesting (EH) model, we consider practical non-linear EH models. We investigate this problem under two different frameworks: i) an optimization approach where the energy beacon knows the utility function of the nodes, channel state information and the energy harvesting characteristics of the devices; hence optimal power allocation strategies can be designed using an optimization problem and ii) a learning approach where the energy beacon decides on its strategies adaptively with battery level information and feedback on the utility function. Our results illustrate that deep reinforcement learning approach can obtain the same error levels with the optimization approach and provides a promising alternative to the optimization framework.**

I. Introduction

Wireless power transfer (WPT) is a promising technology for enabling energy-autonomous future networked systems. At the moment, a significant part of the literature on WPT systems focus on linear energy harvesting (EH) models where the average power that can be harvested at the EH device is modeled as a linear function of the average power input to the device. On the other hand, practical EH hardware circuitry design is limited by the non-linear characteristics of circuit components, which yields to energy harvesting efficiencies that highly depend on the input power levels and input wave-forms.

Investigation of these issues have only recently started to appear in the communications community: Refs. [1–3] show the superior performance of multi-sine waveforms for power transfer compared to the traditional communication wave-forms. Non-linear models for power conversion efficiency in EH circuitry are investigated and performance improvements due to usage of practical models in communication system design are illustrated [4–6]. In this article, we contribute to this line of work by investigating the effect of non-linear power conversion on the performance of a remote sensing system powered with WPT.

We consider the setting in Fig. 1 where the sensor nodes are wirelessly powered by an energy beacon. Sensor nodes measure an unknown field of interest. We focus on the problem of jointly optimizing the energy allocation of the energy beacon to different sensors and the data transmission powers of the sensors in order to minimize the field reconstruction error at the sink. In contrast to the line of work that focuses on remote estimation problems under total power constraints or under wireless power transmission with linear EH models [7], we investigate this problem under non-linear EH models.

A. Özçelikkale acknowledges the support from Swedish Research Council under grant 2015-04011. M. Koseoglu acknowledges the support from Fulbright Program with grant number FY-2017-TR-PD-02.

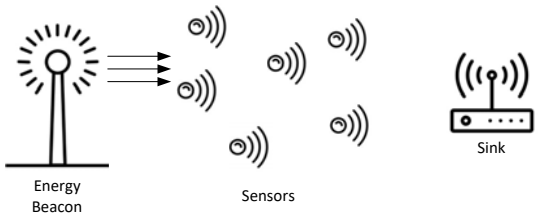

Fig. 1. Sensors powered by an energy beacon transmitting to a sink.

We consider the above resource allocation problem under two different frameworks: i) an optimization approach where the energy beacon knows the form of the utility function (i.e. average field reconstruction error), channel state information and the EH characteristics of the devices; hence can directly design resource allocation strategies using an optimization problem and ii) a reinforcement learning (RL) approach where the energy beacon decides on its strategies adaptively based on the battery level information of the nodes and feedback on the utility function. Recently, deep reinforcement learning techniques have shown state-of-the-art performance in continuous control tasks [8] and machine learning in wireless networking applications has been recently investigated [9]. Our results illustrate that although optimization and RL approaches have access to different types of knowledge on the system parameters, they are able to obtain the same error levels in the sensing problem considered here.

Notation: We denote a column vector by $\mathbf{a} = [a_1; \ldots; a_n] \in \mathbb{C}^{n\times 1}$ where semi-colon ; is used to separate the rows. The complex conjugate transpose of a matrix A is denoted by $A^\dagger$.

II. System Model and Problem Statement

Sensing and Signal Model: There are n_s sensors in the system. At time slot t, sensor i obtains M realizations of the random variable x_t^i and sends it to the sink using a noisy communication channel. The aim of the sensing system at time slot t is to estimate the M realizations of the unknown complex proper zero-mean spatially correlated signal $\mathbf{x_t}$ defined as $\mathbf{x_t} = [x_t^1; \ldots, x_t^i; \ldots; x_t^{n_s}] \in \mathbb{C}^{n_s\times 1}$, with $K_{\mathbf{x}_t} = \mathbb{E}[\mathbf{x_t}\mathbf{x_t}^\dagger]$,

978-1-5386-3512-4/18/$31.00 ©2018 IEEE

$P_{x_t} \triangleq \mathrm{tr}[K_{\mathbf{x_t}}] < \infty$. The reduced eigenvalue decomposition of $K_{\mathbf{x_t}}$ is denoted by $K_{\mathbf{x_t}} = U_t \Lambda_{\mathbf{x}_t} U_t^\dagger$ where $\Lambda_{\mathbf{x}_t} \in \mathbb{R}^{s\times s}$ is the diagonal matrix of s non-zero eigenvalues and $U \in \mathbb{C}^{n_s \times s}$ is the matrix of eigenvectors. In the sequel, a realization of the random variable x_t^i is denoted with $x_{t,j}^i$ for the sake of clarity whenever needed.

Communications to the Sink: Sensors send their observations to the sink using a single cell orthogonal division multiple access (OFDMA) set-up where the spectrum is divided into n_s equal sub-channels where each sensor is assigned to one sub-channel [10]. During time slot t, M measurements of sensor i is sent to the sink in an uncoded manner as follows

$$y_{t,j}^i = g_t^i \sqrt{\frac{p_t^i}{\sigma^2_{x_t^i}}} x_{t,j}^i + w_{t,j}^i, \quad j = 1, \dots, M \tag{1}$$

where $\sqrt{p_t^i} \in \mathbb{R}$, denotes the power amplification factor adopted by sensor i at time slot t, $g_t^i \in \mathbb{R}$ is the effective channel gain, $y_{t,j}^i \in \mathbb{C}$ denotes the received observation and $w_{t,j}^i \in \mathbb{C}$ denotes the zero-mean proper white channel noise with variance σ_w^2. The channel gain and channel noise variance is assumed to be constant during transmission at time slot t.

The sink collects the measurements from sensors $i = 1, \dots, n_s$ and makes a linear minimum mean-square error (LMMSE) estimate of the unknown values $\boldsymbol{x}_{t,j} = [x_{t,j}^1; \dots; x_{t,j}^{n_s}]$, $\forall j$. The resulting average mean-square error for large M is given by $\frac{1}{M}\sum_{j=1}^M ||\boldsymbol{x}_{t,j} - \hat{\boldsymbol{x}}_{t,j}||^2 \rightarrow \mathbb{E}[||\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t||^2]$ where $\hat{\boldsymbol{x}}_{t,j}$ is the LMMSE estimate of the $\boldsymbol{x}_{t,j}$. Hence, the estimation error $\varepsilon_t(\boldsymbol{p}_t) = \mathbb{E}[||\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t||^2]$ can be written as

$$\varepsilon_t(\boldsymbol{p}_t) = \mathrm{tr}\left[(\Lambda_{\mathbf{x}_t}^{-1} + \frac{1}{\sigma_w^2} U_t^\dagger G_t \boldsymbol{P}_t U_t)^{-1}\right], \tag{2}$$

where $\boldsymbol{P}_t = \mathrm{diag}(\boldsymbol{p}_t) \in \mathbb{R}^{n_s \times n_s}$, $G_t = \mathrm{diag}(\mathbf{g}_t) \in \mathbb{R}^{n_s \times n_s}$, $\boldsymbol{p}_t = [p_t^1; \dots; p_t^{n_s}] \in \mathbb{R}^{n_s \times 1}$, $\mathbf{g}_t = [|g_t^1|^2/\sigma^2_{x_t^1}; \dots; |g_t^{n_s}|^2/\sigma^2_{x_t^{n_s}}] \in \mathbb{R}^{n_s \times 1}$ and it is assumed that channel state information g_t^i, σ_w^2 and $K_{\mathbf{x}_t}$ are known at the sink.

Wireless Power Transfer: The energy beacon serves n_s sensors using an orthogonal energy transmission scheme, such as the heterogeneous scenario where devices harvest energy in different frequency bands whereas high EH efficiency in whole spectrum is challenging to achieve with practical hardware [11]. We note that this type of orthogonal energy transmission formulation also covers energy delivery by time division within time slot t with dedicated sharp energy beams to each sensor [12]. The effective channel power gain for power transfer to sensor i during time slot t is denoted by $h_t^i > 0$. The energy beacon allocates an average power of q_t^i to sensor i at time slot t. Hence, the power input to the sensor node i is given by $\bar{q}_t^i = q_t^i h_t^i$. Let the power that can be extracted by the node be denoted by d_t^i. The conversion process between $\bar{q}_t^i$ and d_t^i can be expressed as $d_t^i = \phi(\bar{q}_t^i)$. where $\phi(.)$ is a possibly non-linear function. Hence, the energy harvested by node i during time slot t can be written as

$$E_t^i = \tau_E \phi(\bar{q}_t^i) = \tau_E \phi(q_t^i h_t^i) \tag{3}$$

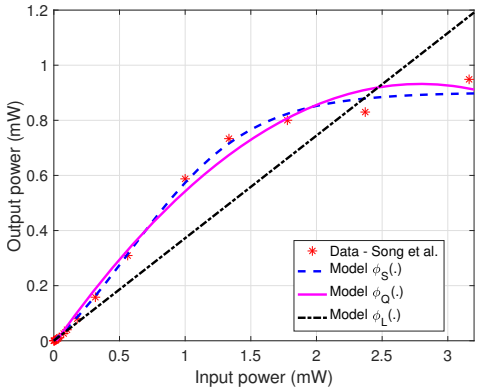

Fig. 2. Comparison of the measurement data [11] and the linear model ϕ_L, the quadratic model ϕ_Q [6], the logistic function model ϕ_S [4].

where τ_E is the length of energy harvesting time slot. We consider the following models for $\phi(.)$:

- The standard linear model with a constant power conversion efficiency
$$\phi_L(\bar{q}_t^i) = \zeta \bar{q}_t^i, \tag{4}$$
where $1 \geq \zeta \geq 0$ is the conversion efficiency. This is the typical model used in the literature [10].
- The quadratic model [6]
$$\phi_Q(\bar{q}_t^i) = \alpha_1 (\bar{q}_t^i)^2 + \alpha_2 \bar{q}_t^i + \alpha_3, \tag{5}$$
where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ are the parameters of the model.
- The logistic/sigmoid function model [4]
$$\phi_S(\bar{q}_t^i) = \frac{\bar{P} - \beta_3 S}{1 - S}, \tag{6}$$
$$\bar{P} = \frac{\beta_3}{1 + \exp(-\beta_1(\bar{q}_t^i - \beta_2))}, \tag{7}$$
where $S \triangleq \frac{1}{1+\exp(\beta_1\beta_2)}$, and $\beta_1, \beta_2, \beta_3$ are the parameters of the model.

These models are illustrated in Fig. 2, where the parameters of all the models are found by least-squares curve fitting of the measurement data from the hardware design of [11].

Energy Constraints at the Sensors: For large M, average power consumption associated with (1) is given by $\frac{1}{M}\sum_{j=1}^M \frac{p_t^i}{\sigma^2_{x_t^i}} (x_{t,j}^i)^2 \rightarrow \frac{p_t^i}{\sigma^2_{x_t^i}} \sigma^2_{x_t^i} = p_t^i$. Hence, total energy spent by the sensor at time slot t is given by $J_t = \tau_I p_t^i$, where $\tau_I = M\tau_I^v$ is the duration of the information transmission and τ_I^v is the average duration of transmission of each sensor value. Energy used by the sensor at any time slot could not exceed the available energy. Hence, we have the following energy neutrality conditions $\sum_{l=1}^t \tau_I p_l^i \leq \sum_{l=1}^t \tau_E \phi(\bar{q}_t^i)$, $t = 1, \dots, T$ where the initial energy in the battery is zero and the battery capacity is large enough so that all the energy that is delivered to the device can be stored.

Problem Statement: Our goal is to jointly design the optimal power amplification factors p_t^i at the sensors and energy allocations q_t^i for the energy beacon in order to minimize the mean-square error over the whole time period of $1 \leq t \leq T$

$$\min_{\boldsymbol{p}_t,\boldsymbol{q}_t} \quad \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\left(\boldsymbol{p}_t\right) \tag{8a}$$

$$\text{s.t.} \quad \sum_{l=1}^{t}\tau_I p_l^i \le \sum_{l=1}^{t}\tau_E\phi(h_l^i q_l^i), \quad \forall t,\forall i \tag{8b}$$

$$\sum_{i=1}^{n_s}\tau_E q_t^i \le P_B, \quad \forall t, \tag{8c}$$

$$p_t^i \ge 0, \quad q_t^i \ge 0, \qquad \forall t,\forall i \tag{8d}$$

where $\boldsymbol{q}_t = [q_t^1;\ldots,;q_t^{n_s}] \in \mathbb{R}^{n_s\times 1}$ is the vector of power allocations by the energy beacon at time t and P_B is the power budget of the energy beacon. For notational simplicity, we normalize as $\tau_I = \tau_E = 1$ in the rest of the article.

We consider this problem under two different frameworks: In the first approach, we consider this optimization problem, i.e. (8), directly. Here, the covariance matrix of $\boldsymbol{x}_t$ and all the relevant channel state information (CSI) is assumed to be known. This off-line optimization set-up serves as a benchmark. In the second approach, neither this information nor the form of the objective function is known by the decision maker. A reinforcement learning approach that uses battery level information at the nodes and feedback on the utility (i.e. distortion) is used to solve this problem. This corresponds to a practical scenario where the sensor nodes and the sink report their battery levels and the utility function to the decision maker (for instance, at the energy beacon), respectively. The underlying assumption for the usage of RL approach is that the channels and the statistical properties of the unknown field change in a way that RL agent can learn from the previous experiences and can adaptively form a resource allocation strategy. In Section V, we illustrate this point for the case of periodically changing signal covariance matrix.

III. Optimization Approach

The objective function of (8) is a convex function of $\boldsymbol{p}_t$, since $\mathrm{tr}[X^{-1}]$ is convex over $X \succeq 0$. Whether (8) constitutes a convex optimization problem is determined by (8b). If $\phi(h_l^i q_l^i)$ is a concave function of q_l^i, then the problem becomes convex; since (8b) becomes an upper bound on a convex function. This is the case for $\phi_L(.)$ which is linear and hence concave; and for $\phi_Q(.)$ which has $\alpha_1 < 0$ and hence concave [6]. Hence, given a strictly feasible point exists, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for optimality. In the below, we illustrate the usage of KKT conditions for $\phi_L(.)$ for the special case where the channel gains and the covariance matrix of the field is constant over time but not necessarily over different nodes:

Proposition 3.1: Let $|g_t^i| = |g^i|$, $|h_t^i| = h^i$, $K_{\boldsymbol{x}_t} = K_{\boldsymbol{x}} = \mathrm{diag}(\sigma_{x^i}^2)$. Then, we have the following: i) Let $\phi = \phi_L$ or $\phi = \phi_Q$ with $\alpha_1 < 0$. Optimal p_t^i and q_t^i values do not depend on time. ii) Let $\phi = \phi_L$. Optimal values are given by

$$p^i = \left(\sqrt{\frac{1}{\kappa}\frac{h^i\zeta\sigma_{x^i}^2\sigma_w^2}{|g^i|^2}} - \frac{\sigma_w^2}{|g^i|^2}\right)^+ \tag{9}$$

and $q_i = p^i/(h^i\zeta)$ where $\kappa > 0$ is a Lagrange multiplier so that $\sum_i q_i = P_B$ and $c^+ \triangleq \max(0,c)$.
Proof is given in Appendix VII. Note that, in the optimal strategy no energy savings between the time instants occur.

We note that, for $\phi_S(.)$, the problem is in general not convex, and sufficiency of KKT conditions should be further investigated. On the other hand, optimal solutions can be determined using numerical optimization methods with convergence guarantees for $\phi_L(.)$ and $\phi_Q(.)$ due to convexity [13]. In Section V, we first solve (8) for $\phi_L(.)$ and $\phi_Q(.)$ using such tools [13]. We then use the resulting solutions as benchmarks to evaluate the success of the RL approach. Then, we investigate the problem with $\phi_S(.)$ using the RL approach.

IV. Reinforcement Learning Approach

In this part, we redefine the problem as an RL problem. In an RL setting, the system dynamics is assumed to be Markovian so that the next state of the system depends solely on the current state and the action of the RL agent, i.e. it is independent of the previously visited states. In particular, we assume that both the channel gains and the unknown field are statistically independent random processes over time and the signal covariance matrix changes periodically. The *agent* aims to maximize a reward signal based on its observations by interacting with the environment without a priori knowledge of transition dynamics of the environment and its rewards. We use the following notation: At step t, the *observation* of the agent, o_t, is the limited view of the agent regarding the underlying state of the system. The *action*, a_t, is the decision made by the RL agent based on its current observation of the system. The *policy*, π, guides the decision of the agent by mapping an observation to an action. The agent's aim is to reach to an optimum policy which maximizes the sum of discounted rewards over time as given by $G = \sum_t \gamma^t r_t$ where γ is defined as the discount factor. Our aim is to minimize distortion, hence we define the reward at time t as the negative of the distortion, i.e. $r_t = -\varepsilon_t$, and the RL agent tries to minimize the sum of distortions over multiple time slots.

We assume that the RL agent can get feedback on the battery levels of the nodes and on the distortion after each step. However, it has no information on the statistics of the field to be estimated; i.e. $K_{\boldsymbol{x}_t}$. We consider the observation space of the system as the combination of the energy stored in the batteries of the nodes, b_t^i, along with the reward returned in the previous step, r_{t-1}. We include the last reward information to capture the current state of the environment. Hence, there are n_s+1 observations for the RL agent for an n_s-node system.

RL agent controls both the energy beacon and the sensors. Its actions are 1) energy allocations at the beacon, i.e. q_t^i, and 2) transmission power factors at the sensors, i.e. p_i^t. For q_i^t, to make sure that the transmitted energy to all nodes equals to the power budget of the energy beacon, P_B, as given by (8c), we define auxiliary variables s_t^i for each node where the energy transferred to a node is found using an exponential softmax operation $q_t^i = P_B\frac{\exp(s_t^i)}{\sum_j \exp(s_t^j)}$ which results in $\sum_i q_t^i = P_B$.

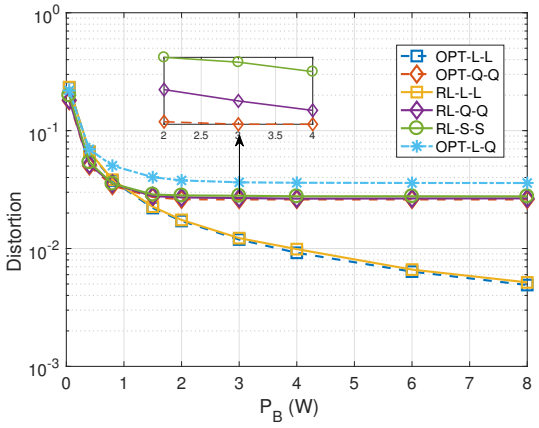

Fig. 3. Distortion (mean-square error) versus power budget (P_B)

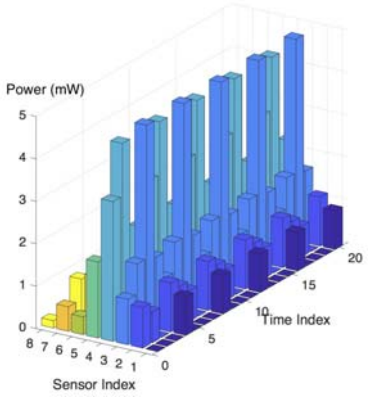

Fig. 4. Information transmission power allocation p_t^i, $n_s = 8$, $n_t = 20$

The second action, i.e. p_t^i, is limited by the energy stored in the battery of node i at time t, $b_t^i = \bar{b}_{t-1}^i + \phi(h_t^i q_t^i)$ with $\bar{b}_{t-1}^i \triangleq \sum_{l=1}^{t-1} \phi(h_l^i q_l^i) - \sum_{l=1}^{t-1} p_l^i$ as implicitly defined by (8b). The agent only knows $\bar{b}_{t-1}^i$. We define another auxiliary variable, $0 \leq \rho_t^i \leq 1$, which indicates the ratio of p_t^i to b_t^i. Then, the transmission power p_t^i of a node is given by $p_t^i = b_t^i \times \rho_t^i$. There are $2n_s$ actions to be determined for an n_s-node system.

To represent the policy π of the RL agent, we use artificial neural networks due to recent success of deep neural networks at representing complex policies [8]. In our scenario, both state and action spaces are continuous. Hence, RL algorithms for discrete action spaces such as deep Q-learning are not applicable. Naive discretization of the action and state spaces would result in an explosion in the number of states which would make the problem intractable. Hence, here we adopted a policy gradient approach referred as Trust Region Policy algorithm (TRPO) which is suitable for continuous control problems and has shown state-of-the-art performance in deep RL benchmarks [14], [15]. Further implementation details are presented in [16].

V. Numerical Results

Let $h_t^i = \frac{A_E A_N}{\lambda^2 {d_i^E}^\gamma_E} |Z_{t,E}^i|^2$, where λ is the wavelength, d_i^E is the distance between the energy beacon and node i, A_N and A_E are the total apertures of the sensor node and energy beacon antenna arrays, respectively [10]. The propagation is assumed to be close to line of sight with path loss coefficient $\gamma_E = 2$. Let $f = v_c/\lambda = 2.45$GHz, $v_c = 3 \times 10^8 m/s$, $A_E = 0.2m^2$, $A_S = 0.005m^2$, $Z_{t,E}^i \sim \mathcal{CN}(1, 0.2)$. We have $|g_t^i|^2 = \frac{A_I A_N}{\lambda^2 {d_i^I}^\gamma_I} |Z_{t,I}^i|^2$, where $\gamma_I = 3$, $A_I = A_E$, $Z_{t,I}^i \sim \mathcal{CN}(1, 0.2)$ and $\sigma_w^2 = 0.1$ μW. Here, $Z_{t,E}^i$ and $Z_{t,I}^i$ are statistically independent of each other and over t and i. The d_i^E and d_i^I values are set according to the following scenario in 2-D plane: Energy Beacon at $(-1, 0)$, sink at $(4, 0)$, node j at $(0, j\text{-}4)$, $j \in \mathbb{Z}$, $j \in [1, 8]$ where the unit is meters. We assume that the hardware design of [11] is used for the energy harvesting circuitry. The second order statistics of $\mathbf{x}_t$ is periodic in time with the period $\kappa = 4$: $K_{\mathbf{x_t}} = U_t \Lambda_{\mathbf{x}_t} U_t^\dagger$ where $\Lambda_{\mathbf{x}_t} = \frac{n_s}{\mathrm{tr}[\Lambda_t]} \Lambda_t$, $\Lambda_t = \mathrm{diag}(\boldsymbol{\eta}_t)$, $\boldsymbol{\eta}_t = [\eta_{1,t}; \ldots; \eta_{n_s,t}] \in \mathbb{R}^{n_s \times 1}$ $\eta_{k,t} = {\nu_t}^k, 0 \leq k \leq n_s - 1$, $\nu_t = 0.2^{\mathrm{mod}(t,\kappa)}$, where $\mathrm{mod}(t, \kappa)$ denotes modulo operation in base κ. The unitary matrices $U_t = U_{\mathrm{mod}(t,\kappa)}$ are drawn from the uniform (Haar) unitary matrix distribution. We report the normalized error with $\bar{\varepsilon} \in [0, 1]$, where $\bar{\varepsilon} = \varepsilon / P_x$, $\varepsilon = \sum_{t=1}^T \varepsilon_t(\boldsymbol{p}_t)$, $P_x = \sum_{t=1}^T \mathrm{tr}[K_{\mathbf{x_t}}]$ and $T = 20$.

We label the different scenarios as "S-AM-RM" where S $\in \{OPT, RL\}$, AM $\in \{L, Q, S\}$, RM $\in \{L, Q, S\}$ refer to the solution method (optimization versus reinforcement learning), assumed EH model for optimization and the actual EH model (ϕ_L, ϕ_Q, ϕ_S), respectively. For instance, OPT-L-Q refers to the case where optimization problem in (8) is solved using the model $\phi_L(.)$ using CVX [13] and the performance of the resulting p_t^i and q_t^i values are evaluated based on $\phi_Q(.)$. Hence, this is the scenario where the resource allocation is based on $\phi_L(.)$ whereas the actual EH hardware follows $\phi_Q(.)$. In this scenario, the nodes may not have enough energy to implement p_t^i values found for some time instants due to the erroneous model assumption. For these cases, the node sends with all the energy available. If there is remaining energy, it is used at $t = T$. It is assumed that energy harvested saturates for input values higher than 2.8mW for the actual hardware of $\phi_Q(.)$, see Fig. 2. For RL scenarios, there are no cases with discrepancy between assumed and actual models, since RL makes no assumptions on the EH models and decides on the power allocations based on the feedback on the battery levels and the distortion values.

We first consider the case with $|Z_{t,I}^i| = |Z_{t,E}^i| = 1$. The distortion versus power budget P_B curves are presented in Fig 3. Comparing the RL and OPT curves, we observe that the curves are on top of each other for both L-L and Q-Q scenarios. This illustrates that RL approach successfully learns how to minimize the distortion even if it does not know the form of this function or the channel gain values. It is also observed there is no significant performance difference between RL-Q-Q and RL-S-S, which is consistent with the good fit of both models with the measurement data as illustrated in Fig. 2. Comparing OPT-L-Q and OPT-Q-Q, we observe that there is a performance gap due to the wrong assumption on the EH model, which illustrates the need to design resource allocation based on realistic EH models.

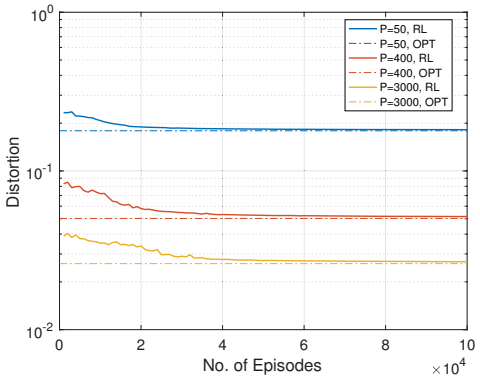

Fig. 5. Distortion as a function of the number of episodes that the RL algorithm executes.

An illustration of the optimal p_t^i values for $OPT\text{-}L\text{-}L$ are presented in Fig. 4. The nodes that are closest to the energy beacon and the sink ($j = 3, 4, 5$) are transmitting with the highest power. We observe that nodes save power to be able to transmit with higher power in the subsequent time instants. We note that the periodic nature of the power allocation scheme over time is consistent with the periodically changing correlation function of the unknown field.

Fig. 5 shows the convergence behavior of the RL algorithm which also includes the optimum values from the optimization approach as the lower bound. The distortion reduces with more episodes and approaches to the 2-3% of the optimum value at the $\approx 10^5$th episode. Considering that the time horizon is $T = 20$ time slots, RL algorithm needs to interact with the system for $\approx 2 \times 10^6$ time slots. Further discussions are provided in [16].

We now discuss the case with $Z_{t,I}^i \sim \mathcal{CN}(1, 0.2)$ and $Z_{t,E}^i \sim \mathcal{CN}(1, 0.2)$. An average over 100 channel realizations are reported. We compare the following scenarios: i) Performance of the direct solution of (8) where the CSI for all t are known, ii) Performance of the solution of (8) for $|Z_{t,I}^i| = |Z_{t,E}^i| = 1$ under stochastic channel realizations, iii) Performance of RL which does not know CSI and the form of the utility function. For $P_B = 3W$, we obtain the following normalized distortion values: i) 1.342×10^{-2}, ii) 3.178×10^{-2}, iii) 2.89×10^{-2} for $\phi_L(.)$ and i) 2.73×10^{-2}, ii) 4.54×10^{-2}, iii) 4.35×10^{-2} for $\phi_Q(.)$. We observe that the distortion values obtained by the RL based approach is reasonably close to these benchmark values.

VI. Conclusions

A comparison of the RL and optimization based approaches for resource allocation in wirelessly powered sensor networks is presented. Practical non-linear EH models are an important part of the setting. Our results illustrate that RL based approaches show promising performance with non-linear EH models and partial CSI scenarios.

VII. Appendix: Proof of Prop. 3.1

i) Since $K_{\boldsymbol{x}} = \text{diag}(\sigma_{x^i}^2)$, (8a) can be written as $\sum_{t=1}^T \sum_{i=1}^{n_s} \varepsilon_i(p_t^i)$, where $\varepsilon_i(p_t^i) = (\sigma_w^2 \sigma_{x^i}^2)/(|g^i|^2 p_t^i + \sigma_w^2)$. Suppose that $Q^i \geq 0$ is the total power allocated to node i over the whole time frame, i.e. $\sum_t q_t^i = Q^i$. Hence, for node i, (8) reduces to: $\min_{p_t^i} \sum_{t=1}^T \varepsilon_i(p_t^i)$ such that $\sum_{l=1}^t p_l^i \leq \sum_{l=1}^t \phi(h^i q_l^i)$, $\forall t$ and $\sum_{t=1}^T q_t^i = Q^i$. We consider the following relaxation of this problem $\min_{p_t^i} \sum_{t=1}^T \varepsilon_i(p_t^i)$ such that $\sum_{l=1}^T p_l^i \leq P^i$ where $P^i = \max_{q_l^i} \sum_{l=1}^T \phi(h^i q_l^i)$ over $\sum_t q_t^i = Q^i$. The result follows from the Schur-convexity/Schur-concavity of the objective function of these optimization problems. Details can be found in [16].

ii) Using $\phi = \phi_L$ and part (i), i.e. $p_t^i = p^i$ and $p_t^i = q^i$, (8) now can be written as the minimization of $\sum_{t=1}^T \sum_{i=1}^{n_s} \varepsilon_i(p^i) = T \sum_{i=1}^{n_s} \varepsilon_i(p^i)$ such that $\sum_{i=1}^{n_s} q^i = \sum_{i=1}^{n_s} \frac{p^i}{h^i \zeta} \leq P_B$. Solving the KKT conditions reveals the solution in (9). Details can be found in [16].

References

[1] B. Clerckx and E. Bayguzina, "Waveform design for wireless power transfer," *IEEE Trans. on Signal Process.*, vol. 64, pp. 6313–6328, Dec 2016.

[2] Y. Huang and B. Clerckx, "Waveform Design for Wireless Power Transfer With Limited Feedback," *IEEE Trans. on Wireless Comm.*, vol. 17, pp. 415–429, Jan. 2018.

[3] B. Clerckx, "Wireless information and power transfer: Nonlinearity, waveform design, and rate-energy tradeoff," *IEEE Trans. on Signal Process.*, vol. 66, pp. 847–862, Feb 2018.

[4] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Comm. Letters*, vol. 19, pp. 2082–2085, Dec 2015.

[5] E. Boshkovska, D. W. K. Ng, N. Zlatanov, A. Koelpin, and R. Schober, "Robust Resource Allocation for MIMO Wireless Powered Communication Networks Based on a Non-Linear EH Model," *IEEE Trans. on Commun.*, vol. 65, pp. 1984–1999, May 2017.

[6] X. Xu, A. Özçelikkale, T. McKelvey, and M. Viberg, "Simultaneous information and power transfer under a non-linear RF energy harvesting model," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 179–184, May 2017.

[7] V. V. Mai, W.-Y. Shin, and K. Ishibashi, "Wireless Power Transfer for Distributed Estimation in Sensor Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 549–562, Apr. 2017.

[8] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *International Conference on Machine Learning*, pp. 1329–1338, 2016.

[9] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, pp. 98–105, April 2017.

[10] K. Huang and E. Larsson, "Simultaneous information and power transfer for broadband wireless systems," *IEEE Trans. Signal Process.*, pp. 5972–5986, Dec. 2013.

[11] C. Song, Y. Huang, J. Zhou, J. Zhang, S. Yuan, and P. Carter, "A High-Efficiency Broadband Rectenna for Ambient Wireless Energy Harvesting," *IEEE Trans. Antennas Propag.*, vol. 63, pp. 3486–3495, Aug 2015.

[12] R. Du, C. Fischione, and M. Xiao, "Lifetime maximization for sensor networks with wireless energy transfer," in *Proc. IEEE Inter. Conf. on Communications (ICC)*, pp. 1–6, 2016.

[13] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[14] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, pp. 1889–1897, 2015.

[15] P. Coady, "pat-coady/trpo: First release.," Feb. 2018. doi: 10.5281/zenodo.1183378.

[16] A. Özçelikkale, M. Koseoglu, and M. Srivastava, "Optimization vs. reinforcement learning for wirelessly powered sensor networks," *Technical Report: Available at https://sites.google.com/site/ aycaozcelikkale/opt-rl-wpt*, 2018.