<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *29th IEEE International Workshop on Machine Learning for Signal Processing, October 13-16, Pittsburg, USA.*

Citation for the original published paper:

Özcelikkale, A., Koseoglu, M., Srivastava, M., Ahlén, A. (2019)
Deep reinforcement learning based energy beamforming for powering sensor networks
In: *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP 2019)*
<https://doi.org/10.1109/MLSP.2019.8918775>

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-398752>

DEEP REINFORCEMENT LEARNING BASED ENERGY BEAMFORMING FOR POWERING SENSOR NETWORKS

Ayça Özçelikkale¹, Mehmet Koseoglu², Mani Srivastava³, Anders Ahlén¹

¹Signals and Systems, Uppsala University, Sweden

²Dept. Computer Science, Hacettepe University, Turkey

³Dept. of Electrical and Computer Engineering, University of California, USA

ABSTRACT

We focus on a wireless sensor network powered with an energy beacon, where sensors send their measurements to the sink using the harvested energy. The aim of the system is to estimate an unknown signal over the area of interest as accurately as possible. We investigate optimal energy beamforming at the energy beacon and optimal transmit power allocation at the sensors under non-linear energy harvesting models. We use a deep reinforcement learning (RL) based approach where multi-layer neural networks are utilized. We illustrate how RL can approach the optimum performance without explicitly forming a system model, but suffers from slow convergence. We also quantify the importance of the number of antennas at the energy beamformer and the number of sensors.

Index Terms— Wireless communications, energy harvesting, resource allocation, reinforcement learning, neural networks.

1. INTRODUCTION

Wireless sensor networks (WSNs) is an integral part of a wide range of applications, including environmental and industrial monitoring scenarios. Autonomous optimization of WSNs is of particular interest as a part of self-organizing data acquisition systems. Motivated by the recent success of machine learning based approaches, we investigate a reinforcement learning (RL) framework for optimized resource allocation for such a sensor network. In the RL framework, only past data and feedback on the performance is used to optimize the system performance instead of the traditional approach of explicitly forming an accurate model of the system.

System Model and Methodology: We focus on a sensor network powered by wireless power transfer (WPT) from an energy beacon, as illustrated in Fig. 1. Sensors send their measurements to the sink using the energy they have harvested. The aim of the sensing system is to estimate the unknown field (i.e., unknown signal) over the area of interest as accurately as possible using these sensor measurements. We consider joint optimization of the energy beacon's beamforming strategies and sensors' data transmission power strategies over multiple time slots.

We adopt a deep RL approach, i.e. we combine reinforcement learning with deep learning [1]. In contrast to the approaches where the knowledge about the state and actions, such as the value function, is stored through look-up tables, these values are represented using multi-layer neural networks in deep RL [1]. The deep RL approach

Fig. 1: Sensors are powered by a multi-antenna energy beacon and send their measurements to the sink.

allows us to optimize energy and data transmission strategies without discretizing the state/action space. For instance, we do not discretize the action space for the possible sensor transmission powers, hence we are not restricted to a finite number of possible power values. In particular, we use a *policy-based* deep RL approach [2, 3], where multi-layer neural network representing the policy is optimized using feedback on battery level and the signal reconstruction error.

Related Work: Optimization of energy harvesting (EH) systems has recently gained increasing attention and has been studied under a wide range of scenarios, typically from a communications perspective but also from a sensing perspective [4]. Here, we emphasize the sensing aspects and focus on the field reconstruction error as the main performance metric.

Incorporating practical EH models is significant for accurate evaluation of wireless power transfer systems, as illustrated in terms of suitable waveform design [6–8] and non-linearity in harvesting efficiency [9–11]. Hence, we consider multiple practical non-linear EH efficiency models. Flexibility of our RL approach allows us to propose solutions even under non-linear EH models, which otherwise typically yield to challenging non-convex optimization problems.

Optimization of WPT for powering sensors in order to estimate an unknown signal of interest [12–15] or to optimize the sensing rate of the sensors [16] has been the focus of a number of recent works, under linear EH models [12, 13, 16] and non-linear EH models [14, 15]. In contrast to [12–14], here we focus on a reinforcement learning based approach. In contrast to [16], we consider the estimation error as the performance criterion and use a RL approach. In contrast to [15], we consider a setting with a multi-antenna energy beacon with energy beamforming; and also varying number of sensors where sensors cannot be placed at all points of interest in space; hence the number of sensors is possibly smaller than the number of unknowns (i.e., locations at which we would like to know the signal values).

Contributions: We provide a deep RL based solution for energy beamforming for powering sensor networks with wireless power transfer. Our proposed approach can handle both linear and non-

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.

linear EH efficiency models. We compare the performance of the RL approach with the performance of an off-line optimization approach which requires the knowledge of the system model. We illustrate that the RL approach can attain a performance close to the direct optimization approach after convergence. Our numerical results quantify the performance degradation until convergence due to lack of system information for the RL approach. We also investigate the effect of the number of antennas and the number of sensors; and quantify the performance gain due to an increasing number of antennas and sensors.

2. SYSTEM MODEL

Overview: We now present an overview of the system. Each time slot t is divided into two blocks. At the beginning of each time slot, the energy beacon transfers power to the sensors for the fixed duration of time τ_E while the sensors collect data. After the power transfer, the amount of energy accumulated in the battery of each sensor is the sum of the energy it harvested in the current time slot and the energy left on its battery from the previous time slot. (Energy consumption is dominated by the communications unit in a typical sensor network with passive sensors [17], hence the energy spent for sensing is ignored here.) In the second part of the time slot, during a fixed duration of time τ_I , the sensors transmit their data to the sink using a part of the energy accumulated in their batteries. The sink reconstructs the unknown spatial field (i.e. signal) using the data it received and the linear minimum mean-square error estimator. We would like to minimize the average the field reconstruction error at the sink over a time frame of n_t time slots under power budget constraints at the energy beacon.

Wireless Power Transfer: During the energy transmission phase of time slot t , the energy beacon with m_e antennas broadcasts the complex proper energy signal $\mathbf{z}_t \in \mathbb{C}^{m_e \times 1}$, where

$$\mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\dagger] = \mathbf{K}_{\mathbf{z}_t} = \sum_{j=1}^{n_b} \gamma_{t,j} \mathbf{e}_j \mathbf{e}_j^\dagger \quad (1)$$

and $\{\mathbf{e}_1, \dots, \mathbf{e}_{n_b}\}$ with $\mathbf{e}_j \in \mathbb{C}^{m_e \times 1}$, $\|\mathbf{e}_j\|^2 = 1$ denotes the dictionary of n_b beam vectors that the energy beacon can adopt and † denotes complex conjugate transpose. Here, $\gamma_{t,j}$ is the power allocated to the beam j during time slot t . Each sensor has a single antenna. The channel for power transfer to sensor i during time slot t is denoted by $\mathbf{h}_t^i \in \mathbb{C}^{1 \times m_e}$. Hence, sensor i receives $\mathbf{h}_t^i \mathbf{z}_t$ for energy harvesting and the input power to sensor i at time slot t is given by

$$P_{r,t}^i = \mathbb{E}[\|\mathbf{h}_t^i \mathbf{z}_t\|^2] = \text{tr}[\mathbf{h}_t^i \mathbf{K}_{\mathbf{z}_t} (\mathbf{h}_t^i)^\dagger]. \quad (2)$$

Let the power that can be extracted by the sensor node i be denoted by $P_{h,t}^i$. The conversion process between the received power $P_{r,t}^i$ and the harvested power $P_{h,t}^i$ can be expressed as

$$P_{h,t}^i = \phi(P_{r,t}^i), \quad (3)$$

where $\phi(\cdot)$ is a possibly non-linear function that denotes the power conversion efficiency. Hence, the energy harvested by node i during time slot t can be written as

$$E_t^i = \tau_E P_{h,t}^i = \tau_E \phi(P_{r,t}^i), \quad (4)$$

where τ_E is the fixed time that is allocated for energy harvesting.

We consider the standard linear model as well as recent non-linear models for $\phi(\cdot)$. The standard linear model is given by

$\phi_L(P_{r,t}^i) = \alpha_L P_{r,t}^i$, where $0 \leq \alpha_L \leq 1$ is the conversion efficiency. The quadratic model can be expressed as $\phi_Q(P_{r,t}^i) = \alpha_1 (P_{r,t}^i)^2 + \alpha_2 P_{r,t}^i + \alpha_3$, where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ are the parameters of the model [11]. The logistic/sigmoid function model is given by $\phi_S(P_{r,t}^i) = \frac{\tilde{P} - \vartheta_3 S}{1 - S}$, where $\tilde{P} = \frac{\vartheta_3}{1 + \exp(-\vartheta_1(P_{r,t}^i - \vartheta_2))}$, $S = \frac{1}{1 + \exp(\vartheta_1 \vartheta_2)}$, and $\vartheta_1, \vartheta_2, \vartheta_3$ are the parameters of the model [10].

Sensing and Communications: The aim of the sensing system is to estimate the unknown zero-mean complex proper spatial field $s(x, t) \in \mathbb{C}$ at spatial locations $x \in \mathcal{S}_s = \{x_1, \dots, x_n\}$, $|\mathcal{S}_s| = n$ over the time slots $t = 1, \dots, n_t$. Let $\mathbf{s}_t = [s(x_1, t); \dots; s(x_n, t)] \in \mathbb{C}^{n \times 1}$ denote the unknown field values at time t . Here, $\mathbf{s}_t$ has the covariance matrix $\mathbb{E}[\mathbf{s}_t \mathbf{s}_t^\dagger] = \mathbf{K}_{\mathbf{s}_t} \in \mathbb{C}^{n \times n}$, where $\mathbb{E}[\mathbf{s}_t] = 0$ and $\mathbb{E}[\mathbf{s}_t \mathbf{s}_t^\dagger] = 0$ if $t \neq \tilde{t}$, and $\text{tr}[\mathbf{K}_{\mathbf{s}_t}] < \infty$. We consider signals that are possibly spatially correlated in space, hence $\mathbf{K}_{\mathbf{s}_t}$'s are not necessarily diagonal. There are n_s sensors where sensor i is located at position ζ_i . The set of sensor locations is given by $\mathcal{S}_m = \{\zeta_1, \dots, \zeta_{n_s}\}$. We have $\mathcal{S}_m \subseteq \mathcal{S}_s$. Hence, we cannot put a sensor at all locations we are interested in estimating values of the unknown field.

At time slot t , sensor i has realizations of $s(\zeta_i, t)$. Sensors send their observations to the sink using an orthogonal multiple-access scheme, e.g. OFDMA [18]. The signal that is received by the sink from sensor i at time slot t is [19]

$$y_t^i = g_t^i \sqrt{\frac{p_t^i}{\sigma_{\tilde{s}_t}^2}} \tilde{s}_t^i + w_t^i, \quad (5)$$

where $\tilde{s}_t^i = s(\zeta_i, t)$ denotes the field value at the location of the sensor, $p_t^i \in \mathbb{R}$ is the (normalized) power amplification factor adopted by sensor i at time slot t , $g_t^i \in \mathbb{R}$ is the channel gain, $y_t^i \in \mathbb{C}$ is the received observation, $w_t^i \in \mathbb{C}$ denotes the zero-mean proper white channel noise with variance σ_w^2 . Here, $s(x_j, t)$ and w_t^i are uncorrelated $\forall t, \forall \tilde{t}, \forall j, \forall i$. The channel gain g_t^i is fixed during the time slot t . Note that $\sigma_{\tilde{s}_t}^2$ is the variance of the variable $\tilde{s}_t^i$ and provides normalization so that the power amplification factor p_t^i directly corresponds to the power consumption. In particular, the power consumption for communications based on (5) is given by $p_t^i = \mathbb{E}[(\sqrt{p_t^i/\sigma_{\tilde{s}_t}^2} \tilde{s}_t^i)^2]$ where $\mathbb{E}[(\sqrt{p_t^i/\sigma_{\tilde{s}_t}^2} \tilde{s}_t^i)^2] = (\sqrt{p_t^i/\sigma_{\tilde{s}_t}^2})^2 \mathbb{E}[(\tilde{s}_t^i)^2] = p_t^i$ [19].

Energy Constraints at the Sensors: At each time slot t , the sensors first harvest energy for a duration of τ_E , then communicate for a duration of τ_I , where τ_I is the fixed time used for information transmission at each time slot. The total energy spent by the sensor at time slot t is given by $\tau_I p_t^i$ [12–14]. The energy used by the sensor at any time slot could not exceed the available energy, i.e.

$$\sum_{k=1}^t \tau_I p_k^i \leq \sum_{k=1}^t \tau_E \phi(P_{r,k}^i), \forall t \quad (6)$$

where the initial energy in the battery is zero and the battery capacity is large enough so that all the energy delivered to the device can be stored [5].

Signal Reconstruction at the Sink: Using $\mathbf{y}_t = [y_t^1; \dots; y_t^{n_s}] \in \mathbb{C}^{n_s \times 1}$, the sink computes the linear minimum mean-square error (LMMSE) estimates of the unknown values $\mathbf{s}_t$. Let $\hat{\mathbf{s}}_t(\mathbf{p}_t)$ denote

¹We denote a column vector $\mathbf{a}$ by $\mathbf{a} = [a_1; \dots; a_n] \in \mathbb{C}^{n \times 1}$ where semi-colon ; is used to separate the rows.

the LMMSE estimate of $\mathbf{s}_t$ as a function of the power amplification factors $\mathbf{p}_t = [p_t^1; \dots; p_t^{n_s}] \in \mathbb{R}^{n_s \times 1}$ and let $\varepsilon_t(\mathbf{p}_t) = \mathbb{E}[|\mathbf{s}_t - \hat{\mathbf{s}}_t(\mathbf{p}_t)|^2]$ denote the associated mean-square error. Hence, $\varepsilon_t(\mathbf{p}_t)$ is the mean-square error associated with estimating $\mathbf{s}_t$ (signal values we are interested in) using $\mathbf{y}_t$ (sensor measurements received by the sink), written as a function of $\mathbf{p}_t$ (information transmission powers used by the sensors). Then, $\varepsilon_t(\mathbf{p}_t)$ can be written as [20, Ch2]

$$\varepsilon_t(\mathbf{p}_t) = \text{tr}[\mathbf{K}_{s_t} - \mathbf{K}_{s_t \mathbf{y}_t} \mathbf{K}_{\mathbf{y}_t}^{-1} \mathbf{K}_{s_t \mathbf{y}_t}^\dagger], \quad (7)$$

where $\mathbf{K}_{\mathbf{y}_t} = \mathbb{E}[\mathbf{y}_t \mathbf{y}_t^\dagger]$, $\mathbf{K}_{s_t \mathbf{y}_t} = \mathbb{E}[\mathbf{s}_t \mathbf{y}_t^\dagger]$, $\mathbb{E}[\mathbf{y}_t] = 0$. The received signal $\mathbf{y}_t$ can be written as

$$\mathbf{y}_t = \mathbf{P}_t^{1/2} \mathbf{G}_t^{1/2} \tilde{\mathbf{s}}_t + \mathbf{w}_t, \quad (8)$$

where $\mathbf{w}_t = [w_t^1; \dots; w_t^{n_s}] \in \mathbb{C}^{n_s \times 1}$ is the noise vector and $\tilde{\mathbf{s}}_t = [\tilde{s}_t^1; \dots; \tilde{s}_t^{n_s}] = [s(\zeta_1, t); \dots; s(\zeta_{n_s}, t)] \in \mathbb{C}^{n_s \times 1}$ denotes the sensor measurements, $\mathbf{P}_t = \text{diag}(\mathbf{p}_t) \in \mathbb{R}^{n_s \times n_s}$, $\mathbf{G}_t = \text{diag}(\mathbf{g}_t) \in \mathbb{R}^{n_s \times n_s}$, $\mathbf{g}_t = [|g_t^1|^2/\sigma_{s_t^1}^2; \dots; |g_t^{n_s}|^2/\sigma_{s_t^{n_s}}^2] \in \mathbb{R}^{n_s \times 1}$. Here, $\tilde{\mathbf{s}}_t$ can be expressed as $\tilde{\mathbf{s}}_t = \mathbf{F} \mathbf{s}_t$ where $\mathbf{F} \in \mathbb{R}^{n_s \times n}$ is all zeros except the i th row, j th column element of $\mathbf{F}$, which is 1 if $\zeta_i = x_j$, i.e., if we make a measurement at location x_j with the sensor i . Using the Sherman-Morrison-Woodbury identity [21], $\varepsilon_t(\mathbf{p}_t)$ can be expressed as [22, Eqn. 5]

$$\varepsilon_t(\mathbf{p}_t) = \text{tr} \left[(\Lambda_{s_t}^{-1} + \frac{1}{\sigma_w^2} \mathbf{U}_t^\dagger \mathbf{F}^\dagger \mathbf{G}_t \mathbf{P}_t \mathbf{F} \mathbf{U}_t)^{-1} \right], \quad (9)$$

where $\mathbf{K}_{s_t} = \mathbf{U}_t \Lambda_{s_t} \mathbf{U}_t^\dagger$ denotes the reduced eigenvalue decomposition of $\mathbf{K}_{s_t}$ and $\Lambda_{s_t} \in \mathbb{R}^{l_t \times l_t}$ is the diagonal matrix of l_t non-zero eigenvalues and $\mathbf{U}_t \in \mathbb{C}^{n_s \times l_t}$ is the matrix of eigenvectors.

3. PROBLEM STATEMENT

The amount of energy each sensor harvests depends on the energy beamforming strategy, which is determined by the energy beam power allocations $\gamma_{t,j}$. The field reconstruction quality depends on the information power transmission values p_t^i . We would like to optimally change the energy received by the sensors using $\gamma_t = [\gamma_{t,1}, \dots, \gamma_{t,n_b}]$ and how this energy is spent using $\mathbf{p}_t$ so that the field reconstruction error is minimized.

Hence, we jointly design the energy beam power allocations $\gamma_{t,j}$ for the energy beacon and the optimal power amplification factors p_t^i at the sensors in order to minimize the mean-square error over the whole time period of $1 \leq t \leq n_t$ as follows

$$\min_{\mathbf{p}_t, \gamma_t} \sum_{t=1}^{n_t} \varepsilon_t(\mathbf{p}_t) \quad (10a)$$

$$\text{s.t.} \quad \sum_{k=1}^t \tau_I p_k^i \leq \sum_{k=1}^t \tau_E \phi(\text{tr}[\mathbf{h}_k^i \mathbf{K}_{z_k} (\mathbf{h}_k^i)^\dagger]), \quad \forall t, \forall i \quad (10b)$$

$$\text{tr}[\mathbf{K}_{z_t}] \leq P_B, \quad \forall t, \quad (10c)$$

where $p_t^i \geq 0$, $\gamma_{t,j} \geq 0$, $\mathbf{K}_{z_t} = \sum_{j=1}^{n_b} \gamma_{t,j} \mathbf{e}_j \mathbf{e}_j^\dagger$ and P_B is the total power budget of the energy beacon at each time slot. Hence, P_B gives how much power the energy beacon is allowed to use at each time slot.

4. METHODOLOGY

We consider (10) under two different frameworks: i) off-line optimization and ii) reinforcement learning. In the off-line optimization

framework, system information (including $\mathbf{K}_{s_t}$ and channel state information) is known. We use this off-line optimization approach as a benchmark. For the reinforcement learning approach, neither this information nor the form of the objective function is known.

Off-line Optimization Approach: The objective function of (10) is a convex function of $\mathbf{p}_t$, since $\text{tr}[X^{-1}]$ is convex over $X \succeq 0$. In (10c), a linear function of γ_t is bounded. Hence, convexity of the optimization problem in (10) is determined by the energy constraint (10b). In particular, $\text{tr}[\mathbf{h}_k^i \mathbf{K}_{z_k} (\mathbf{h}_k^i)^\dagger] = \mathbf{h}_k^i (\sum_{j=1}^{n_b} \gamma_{k,j} \mathbf{e}_j \mathbf{e}_j^\dagger) (\mathbf{h}_k^i)^\dagger$, hence we can rewrite (10b) as

$$\sum_{k=1}^t \tau_I p_k^i - \sum_{k=1}^t \tau_E \phi(\sum_{j=1}^{n_b} \gamma_{k,j} \mathbf{h}_k^i \mathbf{e}_j \mathbf{e}_j^\dagger \mathbf{h}_k^i) \leq 0. \quad (11)$$

If $\phi(\cdot)$ is a concave function of $\gamma_{k,j}$, the constraint on (10b) becomes an upper bound on a convex function, and the problem becomes convex. Here $\phi_L(\cdot)$ is a linear, hence also concave function of $\gamma_{k,j}$ and $\phi_Q(\cdot)$ is a concave function of $\gamma_{k,j}$ for $\alpha_1 \leq 0$, which is the case for EH efficiency [11]. Hence, (10) is a convex optimization problem under $\phi_L(\cdot)$ and $\phi_Q(\cdot)$. Optimal solutions can be determined using numerical optimization methods with convergence guarantees, for instance interior point algorithms. The sigmoid function $\phi_S(\cdot)$ is neither convex nor concave, hence the problem, in general, is not convex under this EH efficiency model.

In Section 5, we use off-line optimization as a benchmark for the RL approach whenever applicable. We first solve (10) under the models $\phi_L(\cdot)$ and $\phi_Q(\cdot)$ using convex optimization tools. In these cases, we study how close the RL approach can perform to these optimal solutions and how fast it can adapt to the environment. Then, instead of $\phi_L(\cdot)$ and $\phi_Q(\cdot)$, we investigate the problem under the model $\phi_S(\cdot)$ using the RL approach. Hence, the flexibility of RL approach allows us to provide solutions under the model $\phi_S(\cdot)$, where traditional convex optimization tools are not directly applicable.

Deep Reinforcement Learning Approach: In the RL setting, there is a learning agent, i.e., decision maker, which aims to find optimum actions in an environment to maximize a long term reward function. The agent observes the state of the system at each time slot and decides on its actions based on the state and its policy. We use the following notation: At time slot t , the RL agent observes the state c_t , takes an action a_t according to its policy π and receives a reward r_t . The environment moves to the next state c_{t+1} which depends on the action a_t . The policy is defined as stochastic and gives the probability of taking an action a at a state c , i.e., $\pi(c|a)$. The return $R_t = \sum_{k=0}^{\infty} \delta^k r_{t+k}$ is the sum of discounted future rewards where δ is the discount factor between 0 and 1. The state-value function defines the expected return starting from state c under policy π , i.e., $V^\pi(c) = \mathbb{E}[R_t | c_t = c]$. The aim of the agent is to find the optimum policy, π^* which maximizes the expected sum of returns from each state, i.e., $\pi^* = \arg \max_{\pi} V^\pi(c)$ for each c .

We adopt a policy-based RL method [2, 3]. In such methods, the policy π is characterized by a function with parameters θ , i.e., $\pi(a|c) = \pi(a; c, \theta)$. In our deep RL framework, this policy is represented by a multi-layer neural network with parameters θ . The aim of the policy-based RL methods is to search for the optimum parameters θ which corresponds to the optimal value function $V^\pi(c) = V^\pi(c, \theta)$. This is typically accomplished by a gradient descent over $\mathbb{E}[R_t]$. Here, the most fundamental question is to control the change in the step size, since too small step sizes cause slow convergence and too large sizes cause fluctuating policies. Naive policy-based methods updates θ with $\nabla_{\theta} \log \pi(a_t | c_t; \theta) R_t$ resulting in slow convergence. In contrast, we utilize a proximal policy opti-

mization (PPO) method which improves convergence by penalizing the amount of change in the policy at each update [2, 3].

As in the off-line optimization framework, actions of the RL agent are the energy beam power allocations ($\gamma_{t,j}$) for the energy beacon and the optimal power amplification factors at the sensors (p_t^i). The agent's aim to minimize the mean-square error over multiple time-slots where the reward is defined as $r_t = -\varepsilon_t$. Note that ε_t is bounded, i.e. $0 \leq \varepsilon_t \leq \text{tr}[K_{s_t}]$; hence the reward is bounded. The RL agent gets feedback on the battery levels of the sensors and on the mean-square error ε_t after each step, for instance as a centralized decision maker located at the energy beacon. However, it has no information on the statistics of the field to be estimated; i.e., K_{s_t} , or channel state information. Hence, the observation space of the system is defined as the aggregation of the energy stored in the batteries of the nodes (before charging and transmission at that time step) along with the reward returned in the previous step, r_{t-1} . We include the last reward information r_{t-1} in state description to capture the current state of the environment. Hence, there are $n_s + 1$ observation variables for the RL agent for an n_s -node system.

Direct inclusion of constraints on functions of optimization variables (such as constraints imposed on $\gamma_{t,i}$ by the total power constraint in (10c)) is not straightforward in a typical RL framework. Hence, we adopt the following approach to guarantee these constraints are satisfied by the actions found by the RL framework: Instead of $\gamma_{t,i}$, we define corresponding auxiliary variables $\tilde{\gamma}_{t,i}$ which are optimized by the RL approach. Then, to guarantee that the total energy allocated to all beams equals to the power budget of the energy beacon, the power allocated to a beam is found using an exponential softmax operation $\gamma_{t,i} = P_B \frac{\exp(\tilde{\gamma}_{t,i})}{\sum_j \exp(\tilde{\gamma}_{t,j})}$ which results in $\sum_{j=1}^{n_b} \gamma_{t,j} e_j e_j^\dagger = P_B$, as desired. From the RL agent's point of view, this operation is a part of the unknown environment, hence the RL agent learns to adapt to this additional processing similar to the other environment conditions.

Another constraint we need to impose is the power constraint at the sensors. Here, p_t^i 's are limited by the energy stored in the battery of node i at time t , $b_t^i = \tilde{b}_{t-1}^i + \tau_E \phi(P_{r,t}^i)$ with

$$\tilde{b}_{t-1}^i \triangleq \sum_{l=1}^{t-1} \tau_E \phi(P_{r,l}^i) - \tau_I \sum_{l=1}^{t-1} p_l^i \quad (12)$$

as implicitly defined by (10b). The agent only knows $\tilde{b}_{t-1}^i$. We define another auxiliary variable, $0 \leq \rho_t^i \leq 1$, which indicates the ratio of p_t^i to b_t^i . The RL agent optimizes the variable ρ_t^i , where the constraint $0 \leq \rho_t^i \leq 1$ is a direct constraint on the optimization variable and can be imposed in a straightforward manner. Then, the transmission power p_t^i of a node is given by $p_t^i = b_t^i \times \rho_t^i$.

5. NUMERICAL RESULTS

To model the unknown signal $s(x, t)$, we consider the Gaussian-Schell model (GSM), a random optical field model of central importance in optics [23]. This model allows us to vary the spatial correlation between signal values in a systematic manner. Spatial correlation of a zero-mean GSM source is characterized by the covariance function

$$K_{s_t}(\tilde{x}_1, \tilde{x}_2) = \mathbb{E}[s(\tilde{x}_1, t)s(\tilde{x}_2, t)^\dagger] \quad (13)$$

$$= A \exp\left(-\frac{\tilde{x}_1^2 + \tilde{x}_2^2}{4\sigma_{I,t}^2}\right) \exp\left(-\frac{(\tilde{x}_1 - \tilde{x}_2)^2}{2\sigma_{\nu,t}^2}\right), \quad (14)$$

Fig. 2: The mean-square error vs. power budget

Fig. 3: The mean-square error vs. number of sensors (n_s), $m_e = 4$

which gives the covariance between the signal value at location $\tilde{x}_1$ and the signal value at location $\tilde{x}_2$, at time slot t . Here, $A > 0$, $\sigma_{I,t} > 0$, $\sigma_{\nu,t} > 0$ are the parameters of the field, where $\sigma_{I,t}$ and $\sigma_{\nu,t}$ determine the width of the intensity profile (consider $K_{s_t}(\tilde{x}_1, \tilde{x}_2)$ with $\tilde{x}_1 = \tilde{x}_2$) and the width of the degree of spatial correlation (consider how the correlation coefficient, i.e., normalized $K_{s_t}(\tilde{x}_1, \tilde{x}_2)$, varies with fixed $\tilde{x}_1$ and varying $\tilde{x}_2$), respectively [23]. For a GSM source, $\beta_t = \sigma_{\nu,t}/\sigma_{I,t}$ is interpreted as a measure of degree of global coherence (i.e., correlatedness) of the field [23]. As β_t increases/decreases, the field becomes more correlated/uncorrelated [23]. We set $\sigma_{I,t} = 1$, $A = 1$ and vary β_t in time.

We are interested in estimating $s(x, t)$ at $n = 33$ positions in space uniformly distributed over $[-5\sigma_{I,t}, 5\sigma_{I,t}]$ on a line at $y = 0$ in the 2-D plane. The n_s sensor nodes are uniformly distributed over this grid with $n = 33$ elements, where we vary n_s across experiments. We consider $n_s \in \{2, 4, 8, 16, 33\}$, for which sensors can be exactly placed on (a subset of) these grid points. Here β_t varies periodically in time with the period 4 over the values $\beta = [1, 1/2, 1/4, 1/16]$, hence the spatial correlation of the field varies periodically in time. This type of varying correlation model allows us to study to which extent RL can track changes in the system model. We report the normalized error with $\bar{\varepsilon} \in [0, 1]$, where $\bar{\varepsilon} = \varepsilon/P_s$, $\varepsilon = \sum_{t=1}^{n_t} \varepsilon_t(\mathbf{p}_t)$, $P_s = \sum_{t=1}^{n_t} \text{tr}[K_{s_t}]$ and $n_t = 16$.

Fig. 4: The MSE vs. power budget with varying number of energy beacon antennas (m_e), $n_s = 33$

We have $h_t^{i,j} = (\frac{A_E A_S}{\lambda^2 (d_i^E)^2 \eta_E})^{1/2} Z_t^{i,j}$, where $h_t^{i,j}$ is the j^{th} element of $h_t^i \in \mathbb{C}^{1 \times m_e}$, λ is the wavelength, d_i^E is the distance between the energy beacon and sensor i , A_S and A_E are the total apertures of the sensor and energy beacon antennas, respectively [18]. Let $\eta_E = 2$, $f = v_c/\lambda = 2.45\text{GHz}$, $v_c = 3 \times 10^8\text{m/s}$, $A_E = 0.8\text{m}^2$, $A_S = 0.005\text{m}^2$, $\tau_I = \tau_E = 1$. $Z_t^{i,j}$'s are generated i.i.d. with $\mathcal{N}(0.5, 0.05)$ and e_j 's are randomly chosen from the columns of uniformly distributed unitary matrices [24] with $n_b = 50$, where both are kept fixed throughout the experiments for a given m_e . We have $g_t^i = (\frac{A_I A_S}{\lambda^2 (d_i^I)^2 \eta_I})^{1/2}$, $\eta_I = 2.5$, $A_I = 0.2\text{m}^2$, $\sigma_w^2 = 0.1 \mu\text{W}$. The d_i^E and d_i^I values are set according to the following scenario in the 2-D plane: Energy beacon at (0,-1), sink at (0,4), sensors are on the line $y=0$ where the unit is meters. Design of [25] is used for EH parameters where it is assumed that $\phi_Q(\cdot)$ saturates for inputs larger than 2.8mW.

We consider a policy optimization RL algorithm [2,3]. Both the value and policy functions are approximated by 3-layer neural networks with $\tanh(\cdot)$ activations. Further increasing the number of layers were observed to provide negligible performance gain. The widths of the hidden layers are adapted to the size of the sensor network. For $n_s = 33$, widths of the hidden layers are $\{340, 41, 5\}$ and $\{340, 534, 840\}$ for the value function and the policy, respectively. The Adam optimization method [26] is used for both networks.

The mean-square error (MSE) versus power budget P_B curves are provided in Fig. 2 for $n_s = 33$, $m_e = 4$. Here “S-M” refers to the scenario with the solution approach $S \in \{OPT, RL\}$ and the EH model $M \in \{L, Q, S\}$. We also present *OPT-L-Q* where the problem is solved assuming the idealistic linear model whereas performance evaluation is done under the practical quadratic model. RL approach uses feedback from the system instead of assumptions on the EH models, hence there are no cases with model mismatch under RL. The RL and OPT curves are on top of each other for both linear and quadratic models when there is no model mismatch. Hence, the RL approach successfully learns to minimize the MSE without a priori knowledge of system parameters or dependence of the error on these parameters. Consistent with the fact that both ϕ_S and ϕ_Q provide a good fit to the measurement data [11], the performance gap between *RLQ* and *RL-S* is small. Comparison of *OPT-L-Q* and *OPT-L* reveals the gap between the performance that would have been expected under the idealistic model and the performance that will be obtained under practical scenarios.

Fig. 5: Convergence behaviour of RL: The MSE vs. iterations with $n_s = 33$, $P_B = 3W$

Fig. 3 illustrates that the MSE depends significantly on the number of sensors n_s . This is consistent with the fact that spatial field becomes periodically uncorrelated, hence more measurements should be acquired at these time instants. Again, RL and OPT curves are on top of each other. The effect of the number of antennas at the energy beacon is investigated in Fig. 4 for $n_s = 33$. To avoid clutter, we have only presented performance of the OPT approach since RL curves are again on top of the OPT curves. Under the linear EH model, a high performance gain is obtained due to the increasing number of antennas at the energy beacon. On the other hand, performance saturates quickly with the number of antennas under the quadratic model. This is consistent with the fact that under the nonlinear models the energy that can be harvested by the devices saturate. Hence, gains due to sharper beamforming with higher number of antennas is limited.

Convergence behaviour of RL is illustrated in Fig. 5. The initial performance gap can be interpreted as the cost of lack of system information in the RL approach and decreases as the RL agent has more interactions with the system. With four 3.5 GHz cores and a Quadro K620 GPU, direct optimization and RL (10^5 iterations, utilizing GPU) takes 15 and 62 minutes, respectively. Although the performance gap between the direct optimization and RL approach can be further decreased with increasing number of iterations, these results are not shown here due to slow convergence of RL approach.

6. DISCUSSIONS

The training cost, i.e. the large number of iterations the RL approach typically requires before convergence, can be seen as a hindrance for straightforward application of RL methods in low-power sensor network settings. On the other hand, offline optimization approaches have also hidden training costs since they are based on known system models; and forming accurate system models also requires multiple interactions with the system. Hence, the best trade-off between treating modelling and optimization separately (as in offline optimization) and direct optimization (as in typical RL methods) is not a priori clear. Moreover, the RL approach does not have to interact with the sensor network directly. Instead, it can interact with a comprehensive system simulation, similar to the case of self-driving cars where accurate simulators are used for initial training due to the challenges with the real-world road training. We consider investigation of the above perspectives for sensor networks as promising future directions.

7. CONCLUSIONS

We have presented a RL based approach for energy beamforming in order to minimize the signal reconstruction error in a sensor network. Our results illustrated that the deep RL method can approach the optimum performance without the explicit knowledge of the system model, but suffers from slow convergence. We have quantified the performance gap between direct optimization and the deep RL approaches after convergence as well as the importance of the number of antennas at the energy beamformer and the number of sensors. We have illustrated how the performance gains due to sharper beamforming is limited under realistic non-linear sigmoid and quadratic EH efficiency models. Our results also quantified the significant performance gains that can be obtained with increasing number of sensors. Extensions of our work to decentralized decision scenarios where the data processing and the design of resource allocation strategies are done in a distributed manner are considered as a future work.

8. REFERENCES

- [1] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.
- [2] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *ArXiv e-prints*, July 2017.
- [3] P. Coady, "pat-coady/trpo: First release.," Feb. 2018. doi: 10.5281/zenodo.1183378.
- [4] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 360–381, March 2015.
- [5] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE Journal on Selected Areas in Comm.*, vol. 29, pp. 1732–1743, Sep. 2011.
- [6] 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.
- [7] 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.
- [8] Z. B. Zawawi, Y. Huang, and B. Clerckx, "Multiuser wirelessly powered backscatter communications: Nonlinearity, waveform design, and SINR-energy tradeoff," *IEEE Trans. on Wireless Comm.*, vol. 18, pp. 241–253, Jan 2019.
- [9] P. N. Alevizos and A. Bletsas, "Sensitive and nonlinear far-field RF energy harvesting in wireless communications," *IEEE Trans. on Wireless Comm.*, vol. 17, pp. 3670–3685, June 2018.
- [10] 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.
- [11] 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 ICC Workshops*, pp. 179–184, May 2017.
- [12] Y.-W. P. Hong, T.-C. Hsu, and P. Chennakesavula, "Wireless Power Transfer for Distributed Estimation in Wireless Passive Sensor Networks," *IEEE Trans. on Signal Process.*, vol. 64, pp. 5382–5395, Oct. 2016.
- [13] 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.
- [14] N. K. D. Venkatesgoda, H. Lee, and I. Lee, "Joint transceiver designs for MSE minimization in MIMO wireless powered sensor networks," *IEEE Trans. on Wireless Comm.*, vol. 17, pp. 5120–5131, Aug 2018.
- [15] A. Özçelikkale, M. Koseoglu, and M. Srivastava, "Optimization vs. reinforcement learning for wirelessly powered sensor networks," in *IEEE 19th Int. Workshop on Signal Process. Advances in Wireless Comm. (SPAWC)*, June 2018.
- [16] R. Du, A. Özçelikkale, C. Fischione, and M. Xiao, "Towards immortal wireless sensor networks by optimal energy beamforming and data routing," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 5338–5352, Aug 2018.
- [17] V. Raghunathan, C. Schurgers, Sung Park, and M. B. Srivastava, "Energy-aware wireless microsensor networks," *IEEE Signal Processing Magazine*, vol. 19, pp. 40–50, March 2002.
- [18] K. Huang and E. Larsson, "Simultaneous information and power transfer for broadband wireless systems," *IEEE Trans. Signal Process.*, pp. 5972–5986, Dec. 2013.
- [19] I. Bahceci and A. Khandani, "Linear estimation of correlated data in wireless sensor networks with optimum power allocation and analog modulation," *IEEE Trans. Commun.*, vol. 56, pp. 1146–1156, July 2008.
- [20] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Dover Publishing, 1979.
- [21] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.
- [22] A. Özçelikkale, S. Yüksel, and H. M. Ozaktas, "Unitary precoding and basis dependency of MMSE performance for Gaussian erasure channels," *IEEE Trans. Inf. Theory*, vol. 60, pp. 7186–7203, Nov 2014.
- [23] A. T. Friberg and J. Turunen, "Imaging of Gaussian-Schell model sources," *J. Opt. Soc. Am. A*, vol. 5, pp. 713–720, Jan 1988.
- [24] F. Mezzadri, "How to generate random matrices from the classical compact groups," *Notices of the AMS*, vol. 54, pp. 592 – 604, 2007.
- [25] 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.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Inter. Conf. on Learning Representations (ICLR)*, 2015.