Multi-agent actor-critic method for joint duty-cycle and transmission power control

Sota Sawaguchi, Jean-Frédéric Christmann, Anca Molnos, Carolynn Bernier, Suzanne Lesecq

Multi-Agent Actor-Critic Method for Joint Duty-Cycle and Transmission Power Control

Sota Sawaguchi, Jean-Frédéric Christmann, Anca Molnos, Carolynn Bernier, Suzanne Lesecq
Univ. Grenoble Alpes, CEA, LETI
MINATEC Campus, F-38054 Grenoble, France
e-mail: firstname.name@cea.fr

***Abstract*—In energy-harvesting Internet of Things (EH-IoT) wireless networks, maintaining energy neutral operation (ENO) is crucial for their perpetual operation and maintenance-free property. Guaranteeing this ENO condition and optimal power-performance trade-off under transient harvested energy and wireless channel quality is particularly challenging. This paper proposes a multi-agent actor-critic reinforcement learning for modulating both the transmitter duty-cycle and output power based on the state-of-buffer (SoB) and the state-of-charge (SoC) information as a state. Thanks to these buffers, differently from the state-of-the-art, our solution does not require any model of the wireless transceiver nor any direct measurement of both harvested energy and wireless channel quality for adapting to these uncertainties. Simulation results of a solar powered EH-IoT node using real-life outdoor solar irradiance data show that the proposed method achieves better performance without system failures throughout a year compared to the state-of-the-art that suffers some system downtime. Our approach also predicts almost no system fails during five years of operation.**

I. Introduction

Energy harvesting Internet of Things (EH-IoT) wireless systems are a recent research trend, thanks to their low maintenance cost and self-sustainability. To this end, the energy neutral operation (ENO) must be satisfied [1]. In practice, however, scavenged energy can be highly transient and unpredictable due to weather conditions and geological placement, while obstacles and movements affect the quality of wireless channels [2]. Thus, the system needs to adapt to these changes at run-time in order to meet its energy budget.

To guarantee ENO conditions, some researchers focus on adapting the duty-cycle of the wireless transmission (TX) under the constraint of quality of service (QoS) while others focus on adapting the TX output power to the volatile channel conditions to minimise retransmissions, i.e., latency and energy consumption [3] [4]. These adaptations are based on the estimation of energy budget [3] and of wireless link quality (RSSI: Received Signal Strength Indicator) [5], which often entails prediction errors. Building a control system based on such error-prone estimations may not be reliable.

To avoid these pitfalls, model-free approaches have been introduced [6]–[8]. Instead of predicting the energy budget, Aoudia et al. [6] propose maximising the packet rate under ENO conditions based on the estimate of a state value function (here, SoC) by temporal-difference error (TD-error) in a reinforcement learning (RL), or actor-critic method. This implies that the observation of SoC eliminates the observation of energy income and expenditure. Inspired by this idea, we propose that the use of RL and the observation of a data queue, referred to as state-of-buffer (SoB), can eliminate the direct observation of wireless link quality, since it affects the transmission rate. As such, this paper focuses on the possibility that the uncertainties of scavenged energy and wireless link quality can be addressed by using RL based on SoB and SoC information.

In addition, we believe that our multi-agent RL approach possesses high scalability because the SoB and SoC can be the common system parameters for any action decision within a single node. Hence, our major contributions are threefold:

1) we propose a model-free multi-agent actor-critic algorithm for joint optimisation of TX duty-cycle and output power under the ENO condition;
2) we show that, using RL, the observation of only SoB and SoC eliminates the necessity of measuring data and energy variables such as harvested and consumed energy, and wireless link conditions. This eliminates not only the energy expenditure and calibration required to perform these measurements but the questionable validity of the measurement itself;
3) simulation results show that our method yields less system failures compared with an estimation-based state-of-the-art (SotA) approach.

II. System Model

This section describes a model of an EH-IoT node that transmits data to a sink node over a wireless link. Fig. 1 illustrates the comparison of our approach with [3]'s.

A. Energy Harvesting Model and State-of-Charge

While many different energy sources can be harvested from the environment, e.g. solar, wind, vibration, thermal, etc., in this work, we focus on solar energy-harvesting. The harvested power $P_h(t)$ is calculated based on the solar irradiance as $I(t)$, the size of the photovoltaic (PV) cell as A, conversion efficiency η, and the tracking factor (TF) of maximum power point tracking (MPPT) as described in [3].

To achieve the ENO condition, a supercapacitor is considered an optimal solution for energy storage [9]. It is characterised by its capacity C, nominal voltage V_{nom}, and threshold voltage V_{thrd}, which gives the maximum and minimum (i.e., failing-threshold) energy levels, E_{max} and E_{fail}. With $E(t)$

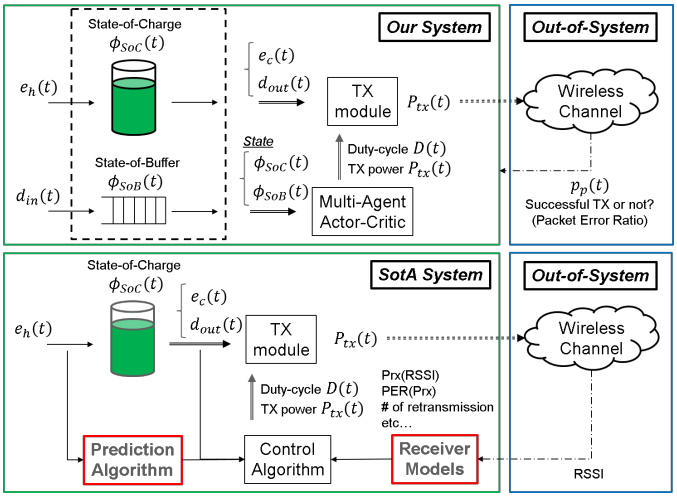

Fig. 1: Conceptual view of proposed and SotA approaches

being the residual energy of the supercapacitor at time t, the SoC $\phi_{SoC}(t)$ can be represented as follows:

$$\phi_{SoC}(t) \quad = \quad \frac{E(t) - E_{fail}}{E_{max} - E_{fail}} \tag{1}$$

Supercapacitors are exposed to a severe self-discharging. The self-discharge rate during time Δt, denoted as τ, can be up to 20% per day [9]. In our simulation, Δt is a minute basis, and the rate is simplified as $\tau = 0.8^{\frac{1}{1440}}$, which represents 20% per day. The leakage power P_{leak} is given by $P_{leak} = \frac{1}{2\Delta t} C(1-\tau^2) V_{sc}^2$, where $V_{sc}(t)$ is the voltage of the supercapacitor. With the system power consumption of P_c, which is defined in Section II-C, the total energy consumption e_c is given by $e_c = (P_c + P_{leak})\Delta t$.

B. Application Data Model and State-of-Buffer

In this paper, we assume that an embedded sensor generates data that are stored into the buffer for transmission. For example, a temperature or motion sensor is employed and characterized as a periodic or random workload, respectively. Here, the incoming data $d_{in}(t)$ during $[t, t+\Delta t]$ is generated based on the Poisson distribution with the average of λ.

A data buffer with maximum capacity B_{max} is used to temporarily store application data that awaits transmission as well as data that has suffered a transmission failure. With current buffer level $B(t)$, the SoB ϕ_{SoB} is defined as:

$$\phi_{SoB}(t) \quad = \quad \frac{B(t)}{B_{max}} \tag{2}$$

C. Power Consumption Model

The power consumption and performance of the wireless transmission (TX) are tuned by the adaptation of TX duty-cycle $D(t)$ and output power $P_{tx}(t)$. Assuming the cycle period T_{cycle}, the active time T_{act}, and the sleep time T_{slp}, $D(t)$ is defined in this paper as the ratio of T_{act} to T_{cycle}; therefore, we have $T_{act} = D(t) \times T_{cycle}$. Note that, in this paper, we ignore the transceiver's wake-up time overhead; therefore, T_{act} represents the sum of time-on-air of frame transmissions (TX) and acknowledgements (RX). If P_c^{rx} and T_{ack} respectively denote the power consumption of the receiver and the time required to receive an acknowledgement packet (assumed to be equal to the acknowledgment frame's time-on-air), the power consumption in active mode P_c^{act} during T_{act} is obtained by $P_c^{act}(t) = P_c^{tx} \cdot (1 - \frac{T_{ack}}{T_{act}}) + P_c^{rx} \cdot \frac{T_{ack}}{T_{act}}$, where P_c^{tx} is the power consumption during packet transmission. If the power consumption in sleep mode is denoted as P_c^{slp}, the average TX power consumption is then defined as $P_c(t) = \frac{P_c^{act}}{\eta^{act}} \cdot D + \frac{P_c^{slp}}{\eta^{slp}} \cdot (1 - D)$, where η^{act} and η^{slp} are the efficiency of the DC-DC regulator in active and sleep mode [3], respectively. We assume P_c^{slp} is a constant value. While the overall power consumption of an IoT node typically comprises of sensing, processing and communication power, only the last one will be considered in our simulation. Also, this work neglects the power consumption overhead of the proposed actor-critic based controller.

D. Wireless Channel Model

In our simulations, we assume the following node deployment scenarios: constant node-to-sink distance with variable shadowing and no fading, or fixed node/sink positions (hence, constant distance and shadowing effect) but with mobility-induced fading. Many such models are available including analytical models based on real-world measurements [10].

The relationship between the transmitted power P_{tx} and received power P_{rx} is determined using the distance d between a transmitter and receiver. With total channel loss (in dB) $PL(d)$, we have:

$$P_{rx}(dBm) \quad = \quad P_{tx}(dBm) - PL(d)(dB) \tag{3}$$

To ease the comparison of our results, we employ an identical channel model to the one in [3]. Here, a combined path-loss and shadowing model [11] in outdoor environment is assumed in which the total channel power loss can be given by $PL(d) = K_{PL} + 10 \cdot \eta \log_{10}\left(\frac{d}{d_0}\right) + \psi_{shadow}$, where ψ_{shadow} is a Gaussian-distributed random variable with mean zero and variance $\sigma^2_{\psi_{shadow}}$ that represents the shadowing coefficient, whereas K_{PL}, η, and d_0 characterize the distance dependent path-loss: K_{PL} is a unit-less constant determined by antenna characteristics and the average channel attenuation, η is the path-loss exponent, and d_0 is a reference distance. With speed of light c and wireless carrier frequency f, we have $K_{PL} = -20 \log_{10}\left(\frac{c}{4\pi d_0 f}\right)$.

The RSSI value for the control algorithm proposed by [3] (Fig. 1) is found using (3) to deduce the packet error rate (PER) based on theoretical models or calibration data. The authors assume that the RSSI value is measured by the sink node and is piggybacked to the sender. In our approach, these models are used not in the system but only in the simulations to calculate the PER which enables a random draw that controls whether a given transmission is successful or not. This information is used to find $d_{out}(t)$.

Algorithm 1: Multi-agent actor-critic algorithm for joint TX power and duty-cycle adaptation

Require: $E(t)$, $B(t)$, $x \in \{D, P_{tx}\}$

* Observe the current state *

1: Eq.(1) and (2)

2: $\phi_+(t) = \phi_{SoB}(t) \cdot \phi_{SoC}(t)$

3: $\phi_-(t) = (1.0 - \phi_{SoB}(t)) \cdot \phi_{SoC}(t)$

4: $R_x(t) = (1.0 - a_x(t)) \cdot (1.0 - \phi_{SoB}(t)) \cdot \phi_{SoC}(t)$

5: $V_x(t-1) = \theta_x(t-1) \cdot (1.0 - \phi_{SoB}(t-1)) \cdot \phi_{SoC}(t-1)$

* TD-error for Actor-Critic *

6: $\delta_x(t) = R_x(t) + \gamma_x \theta_x(t-1)\phi_-(t) - \theta_x(t-1)\phi_-(t-1)$

* Critic: TD(λ) algorithm *

7: $v_x(t) = \gamma_x \lambda_x v_x(t-1) + \phi_-(t)$

8: $\theta_x(t) = \theta_x(t-1) + \alpha_x \delta_x(t) v_x(t)$

* Actor: Policy gradient theorem *

9: $\psi_x(t) = \psi_x(t-1) + \beta_x \delta_x(t) \frac{a_x(t-1) - \mu_x(t-1)}{\sigma_x^2} \phi_+(t-1)$

* Next TX current selection *

10: $\mu_x(t) = \psi_x(t) \cdot \phi_{SoB}(t) \cdot \phi_{SoC}(t)$

11: $\bar{\mu_x}(t) \leftarrow$ Limit $\mu_x(t)$ to $[a_x^{min}, a_x^{max}]$

12: $a_x(t+1) \sim \mathcal{N}(\bar{\mu_x}(t), \sigma_x)$

13: $a_x(t+1) \leftarrow$ Clamp $a_x(t+1)$ to $[a_x^{min}, a_x^{max}]$

14: Return the next action $a_x(t+1)$

TABLE I: Parameter set-ups for simulations

Parameter	Value
PV cell size A	$2.5 \times 10^{-4}\mathrm{m}^2$
PV conversion efficiency η	0.1
Tracking factor TF	0.963
Nominal/threshold voltage V_{nom}/V_{thrd}	2.7V/0.9V
Supply voltage V_{dd}	1.8V
Capacity C	1.0F
Mean packet arrival rate λ	1.0pkt/min
The size of SoB	500pkts
Cycle period T_{cycle}	60s
Initial duty-cycle $D(0)$	$5.0 \times 10^{-4}(i.e., 30\mathrm{ms})$
Minimum positive active time D_{min}^{pos}	5.0×10^{-4}
DCDC conversion efficiency (active mode) η^{act}	0.85 [3], [12]
DCDC conversion efficiency (sleep mode) η^{slp}	0.75 [3], [12]
TX current in sleep mode	900nA
Initial TX output power $P_{tx}(0)$	+1dB
TX packet size N	32bytes
RX acknowledgement packet size	20bytes
Channel bit rate R_b	51.2kpbs
Noise bandwidth	51.2kHz
Path-loss exponent η	4.0
Variance of shadowing ψ_{shadow}	6.0dBm
Distance d	15m
Reference distance d_0	1m
The speed of light c	$3.0 \times 10^8\mathrm{m/s}$
Signal frequency f	2.4GHz
Noise floor	−115dB

III. Actor-Critic Algorithm

Our purpose is to avoid as many system failures as possible, while providing required performance (power-performance trade-offs), under power and performance uncertainties. Such uncertainties vary between nodes that may be in different environments. These facts dictate a model-free approach that fully adapts itself to the uncertainties without any *a priori* knowledge. Hence, inspired by [6], we present a multi-agent actor-critic algorithm with linear function approximations based on SoB and SoC. The algorithm for each agent is shown in Algorithm 1, where x is the target variable to be controlled, e.g., the duty-cycle D and the TX output power P_{tx}, and a_x is the corresponding action.

The reward function R_x expresses the goal of the algorithm. As stated above, our goal is to address the power-performance trade-offs under power and performance uncertainties. While the power uncertainties are addressed only by the SoC observation in [6], we consider both the SoB and SoC to tackle both uncertainties. Hence, the goal is to minimise the power consumption (i.e., maximise the SoC) and to maximise the performance (i.e., minimise the SoB). In this work, since the system failure is more critical, minimising the action value is also considered. Thus, the reward function is formulated as in line 4 in Algorithm 1.

The value function V_x is the value of the state, which is considered more valuable when the SoC is larger and the SoB is smaller. Since linear function approximation requires less computation and memory footprint as discussed in [6], we use the same method to establish the relationship between V_x and the state by using the parameter θ_x. As such, the value function can be defined as in line 5. Thanks to this approximation, the TD(λ) algorithm can be applied to update θ_x and to find the optimal value function at run-time with learning rate α_x (line 7 and 8). Note that λ_x is the exponential weighting for the recency of the prediction, and the TD-error of the value function is obtained by the equation in line 6, where γ_x is the discount factor.

The system can afford to provide more performance, i.e., higher values of action $a_x(t)$ when the SoC level is higher. This logic should be pushed even more when the SoB level is higher in order to avoid data overflow. On the contrary, when the SoB and/or SoC level is lower, less performance may be preferable to prevent the system failure from happening. Thus, again like in [6], we suppose the linear function approximation between the mean action $\mu_x(t)$ and the multiplication of SoB and SoC, which gives the equation in line 10 with the policy parameter $\psi_x(t)$. Because of this approximation combined with the Gaussian policy of mean $\mu_x(t)$ and standard deviation σ_x, the policy, i.e., the action selection probability in the current state can also be optimised by updating the policy parameter with learning rate β_x (line 9). This equation is obtained by the policy gradient theorem. The learned policy generates the next action (line 12), which is then limited to $[a_x^{min}, a_x^{max}]$ (line 13).

IV. Simulation Results

Simulations were conducted using Python to verify the effectiveness of our proposed algorithm. To evaluate our approach, we chose [3] as a recent SotA that presents a joint TX adaptation under the ENO condition based on energy-harvesting prediction and RSSI information.

TABLE II: Hyper-parameter settings

Agent x	α_x	β_x	γ_x	σ_x	λ_x
P_{tx}	0.1	1.0×10^{-5}	0.9	1.0×10^{-3}	0.9
D	0.1	1.0×10^{-6}	0.9	1.0×10^{-4}	0.9

We use a set of real-life outdoor solar irradiance (i.e. global horizontal) data profile ending on May 31st, 2019 [13]. Since the TX adaptation and harvested energy prediction of the SotA are carried out in every 10 and 30min, respectively, our control cycle time is set to 30min. The TX profile in this paper is obtained from the data sheet of CC2500, Texas Instrument [14], and the output power can take $+1\mathrm{dB}$ and $[0, -30\mathrm{dB}]$ with step size of $2\mathrm{dBm}$. The duty-cycle $D(t)$ ranges $[0.0, 1.0]$. Every simulation uses the combinations of hyper-parameter values for each RL agent shown in Table II. The rest of the system parameters are listed in Table I.

We conducted and averaged 100 simulations to compare our approach with the SotA using real-life one-year harvesting data. The evaluation metrics are: *a*) the number of times the system fails; *b*) the ratio of system downtime to the whole simulation time $\%$; *c*) the throughput $(\mathrm{pkt/min})$. Note that the number of dropped packets was zero in every simulation. The results for the metrics are shown in Table III. Since the SotA method makes the most of the energy budget and neglects the SoB, it tend to be zero most of the time (i.e., the throughput is maximised with almost no latency). Meanwhile the proposed method strikes the balance in the SoB-SoC trade-offs in every 30 minute, which resulted in the mean latency of 7.36 min with the standard deviation of 9.63 min. This large variance can be addressed by more fine-grained control policy. Nonetheless, since the SotA frequently faces the system failures, it cannot provide the throughput constantly as opposed to the proposed RL approach which yields no system failures. Many system fails of the SotA approach can be explained by the prediction error caused by too much dependence on the "recent past" information, which may lead to an optimistic control policy. All kinds of prediction-based methods may suffer from prediction errors which would be induced by the approximations done by the chosen prediction algorithm. Also, this error cannot be learned and minimised over time by their algorithm. By contrast, the proposed method keeps minimising the TD-error and takes actions based on the "current" SoC, which is not impacted by past dissimilar experiences. To validate the self-sustainability of our algorithm, we also conducted and averaged 100 simulations using five years of real-life solar irradiance data. The results are shown in Table III. The number of times that the system failed at least once was merely 7 times out of 100. In such cases, the system failed 4.0 times in average with $8.33\mathrm{h}$ of mean system downtime. That accounts for around $2\mathrm{h}$ of mean system failure time for each failure.

TABLE III: Simulation results

Method	# of system fails	System fail time (hrs)	Throughput (pkt/min)
Ju et al. 2018 [3]	3.5e3	4,275	0.51
Our algorithm (one year)	0	0.0	1.00
Our algorithm (10-years)	0.28	0.57	1.00
Worst case	10	21.91	1.00
Avg. of failed cases	4.00	8.33	1.00

V. Conclusions

This paper proposed a multi-agent actor-critic algorithm for joint optimisation of transmitter output power and duty-cycle using only SoB and SoC information. Thanks to these information, the system adapts itself to all the uncertainties regarding data and energy, especially wireless link conditions and harvested energy, to satisfy the ENO condition and to provide an optimal performance. Simulation results using real-life solar irradiance data show that our algorithm enables an EH-IoT system to operate with almost no system fail and an optimal performance for several years.

References

[1] G. V. Merrett and B. M. Al-Hashimi, "Energy-driven computing: Rethinking the design of energy harvesting systems," in *Design, Automation Test in Europe (DATE), 2017*, pp. 960–965, March 2017.

[2] S. Rekik, N. Baccour, M. Jmaiel, and K. Drira, "Experiencing low power wireless links in distribution smart grid environments," in *2018 IEEE/ACS 15th Int. Conf. on Computer Systems and Applications (AICCSA)*, pp. 1–8, Oct 2018.

[3] Q. Ju and Y. Zhang, "Predictive power management for internet of battery-less things," *IEEE Trans. on Power Electronics*, vol. 33, pp. 299–312, Jan 2018.

[4] A. Castagnetti, A. Pegatoquet, T. N. Le, and M. Auguin, "A joint duty-cycle and transmission power management for energy harvesting wsn," *IEEE Trans. on Industrial Informatics*, vol. 10, pp. 928–936, May 2014.

[5] M. Li, X. Zhao, H. Liang, and F. Hu, "Deep reinforcement learning optimal transmission policy for communication systems with energy harvesting and adaptive mqam," *IEEE Trans. on Vehicular Technology*, vol. 68, pp. 5782–5793, June 2019.

[6] F. Ait Aoudia, M. Gautier, and O. Berder, "Rlman: An energy manager based on reinforcement learning for energy harvesting wireless sensor networks," *IEEE Trans. on Green Communications and Networking*, vol. 2, pp. 408–417, June 2018.

[7] Y. Li, K. K. Chai, Y. Chen, and J. Loo, "Smart duty cycle control with reinforcement learning for machine to machine communications," in *2015 IEEE Int. Conf. on Communication Workshop (ICCW)*, pp. 1458–1463, June 2015.

[8] A. Masadeh, Z. Wang, and A. E. Kamal, "Reinforcement learning exploration algorithms for energy harvesting communications systems," in *2018 IEEE Int. Conf. on Communications (ICC)*, pp. 1–6, May 2018.

[9] R. Dekimpe, P. Xu, M. Schramme, D. Flandre, and D. Bol, "A battery-less ble iot motion detector supplied by 2.45-ghz wireless power transfer," in *2018 28th Int. Symp. on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 68–75, July 2018.

[10] M. Zuniga and B. Krishnamachari, "Analyzing the transitional region in low power wireless links," in *2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004.*, pp. 517–526, Oct 2004.

[11] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Press, 2005.

[12] Q. Ju and Y. Zhang, "Charge redistribution-aware power management for supercapacitor-operated wireless sensor networks," *IEEE Sensors Journal*, vol. 16, pp. 2046–2054, April 2016.

[13] "Oak ridge national laboratory (rsr) daily plots and raw data files." https://midcdmz.nrel.gov/apps/sitehome.pl?site=ORNL.

[14] Texas Instruments, *CC2500 Low-Cost Low-Power 2.4GHz RF Transceiver*, 2019. Available at http://www.ti.com/lit/ds/swrs040c/swrs040c.pdf.