Research Article

Optimal Management of Rechargeable Biosensors in Temperature-Sensitive Environments

Yahya Osais, 1 F. Richard Yu, 2 and Marc St-Hilaire 2

1 Department of Computer Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
2 Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6

Correspondence should be addressed to Yahya Osais; yosais@kfupm.edu.sa

Received 26 September 2012; Accepted 13 March 2013

Academic Editor: Nadjib Achir

Copyright © 2013 Yahya Osais et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Biological sensors (biosensors, for short) are tiny wireless devices attached or implanted into the body of a human or animal to monitor and detect abnormalities and then relay data to physician or provide therapy on the spot. They are distinguished from conventional sensors by their biologically derived sensing elements and by being temperature constrained. Biosensors generate heat when they transmit their measurements and when they are recharged by electromagnetic energy. The generated heat translates to a temperature increase in the tissues surrounding the biosensors. If the temperature increase exceeds a certain threshold, the tissues might be damaged. In this paper, we discuss the problem of finding an optimal policy for operating a rechargeable biosensor inside a temperature-sensitive environment characterized by a strict maximum temperature increase constraint. This problem can be formulated as a Markov Decision Process (MDP) and solved to obtain the optimal policy which maximizes the average number of samples that can be generated by the biosensor while observing the constraint on the maximum safe temperature level. In order to handle large-size MDP models, it is shown how operating policies can be obtained using Q-learning and heuristics. Numerical and simulation results demonstrating the performance of the different policies are presented.

1. Introduction

Biosensors are tiny wireless devices attached or implanted into the body of a human or animal to monitor and detect abnormalities and then relay data to physician or provide therapy on the spot. Unlike conventional wireless sensors, biosensors are energy as well as temperature constrained. Also, their sensing elements are biological materials such as enzymes and antibodies which are integrated into transducers for producing electrical signals in response to biological reactions and changes. Biosensors are powered by either rechargeable built-in batteries or by continuously sending electric energy to them in the form of electromagnetic waves.

The use of batteries necessitates periodic recharging which can be performed using energy resulting from vibration, motion, light, and heat. However, a more mature approach is to wirelessly collect energy from a Radio Frequency (RF) source and then convert it into usable power. This approach is widely used in the industry to transfer data and power to biosensors. It is also more practical since many biosensors can be recharged simultaneously. In essence, a charging station generates a magnetic field that can convey energy through the skin. From the penetrating magnetic field, an electric voltage is produced by induction in the receiver circuit. The induced voltage is then rectified, filtered, and stabilized to run the biosensors or recharge their batteries.

In this paper, we study a stochastic control problem which arises when a rechargeable biosensor operates in a temperature-sensitive environment like the human body. In this problem, the state of the biosensor is characterized by its current temperature and energy levels, and uncertainty exists due to the random behavior of the wireless channel between the biosensor and base station. The objective is to operate the biosensor in such a way that the average number of samples generated by the biosensor is maximized while the maximum safe temperature level is not exceeded. This control problem can be formulated as an MDP and solved to obtain an optimal operating policy.

Since the size of the MDP model increases with the number of biosensors and their states, Q-learning which is
a form of reinforcement learning is used to obtain the optimal policy. The optimal policy is learned by interacting with a simulation model of the system. Another way to handle large MDP models is through the use of heuristic policies. This paper proposes a simple heuristic policy whose performance is sufficiently close to that of the optimal policy. A greedy policy is also proposed and used as a baseline for comparing the performance of the different policies.

The remainder of the paper is organized as follows. In the next section, the necessary background information is given. Then, the limited available literature is reviewed. After that, the model of the system under study is described followed by the presentation of its MDP formulation. Next, it is shown how large-size MDP models can be handled using Q-learning. Besides, greedy and heuristic policies are described. Then, numerical and simulation results are presented and an example is given. Finally, conclusions are summarized and directions for further research are suggested.

2. Calculating the Temperature Increase

Radiation due to wireless communication and recharging are the major sources of heat in biosensor networks. The level of radiation absorbed by the human body when exposed to RF radiation is measured by the Specific Absorption Rate (SAR) which is expressed in units of W/Kg. SAR records the rate at which radiation energy is absorbed per unit mass of tissue [1]. The mathematical relationship between SAR and radiation is given by

$$\text{SAR} = \frac{\sigma |E|^2}{\rho}, \quad (1)$$

where $E$ is the induced electric field due to radiation, and $\rho$ and $\sigma$ are the density and electrical conductivity of the tissue, respectively. As an example to appreciate the importance of this measure, it was reported in [2] that an exposure to a SAR of 8 W/Kg in any gram of tissue in the head for 15 minutes may result in tissue damage.

SAR is a point quantity. That is, its value varies from one location to another. In this paper, we consider only the SAR in the near field (i.e., the space around the antenna of the biosensor). The extent of the near field is given by $d_\text{NF} = \lambda/2\pi$, where $\lambda$ is the wavelength of the carrier signal used in wireless communication. SAR in the near field is given by the following equation [3]:

$$\text{SAR}_{\text{NF}} = \frac{\sigma \mu \omega}{\rho \sqrt{\sigma^2 + \epsilon^2 \omega^2}} \cdot \left(\frac{1}{4\pi} \int dl \sin \theta e^{-\alpha R} \left( \frac{1}{R^2} + \frac{|y|}{R} \right)^2 \right), \quad (2)$$

where $\mu$ and $\epsilon$ are the permeability and permittivity of the tissue, respectively; $dl$ is the length of the wire representing the antenna, $I$ is the current provided to the antenna, $\alpha$ is the attenuation constant, $R$ is the distance from the biosensor to the observation point, $\theta$ is the angle between the observation point and the $x$-$y$ plane, $y$ is the propagation constant, and $\omega$ is the angular frequency. We assume that the radiation patterns are omnidirectional on the 2D plane and thus $\sin \theta = 1$.

The Pennes bioheat equation [4] is the standard for calculating the temperature increase in the body due to heating. The general form of the equation is

$$\rho C_p \frac{dT}{dt} = K V^2 T - b (T - T_b) + \rho \text{SAR} + P_c + Q_m, \quad (3)$$

where $\rho$ is the mass density, $C_p$ is the specific heat of the tissue, $dT/dt$ is the rate of temperature increase, $K$ is the thermal conductivity of the tissue, $b$ is the blood perfusion constant which indicates how fast the heat can be taken away by the blood flow inside the tissue, and $T_b$ is the temperature of the blood and the tissue. Terms on the right side indicate the heat accumulated inside the tissue. The terms $KV^2T$ and $b(T - T_b)$ are the heat transfer due to the thermal conduction and the blood perfusion, respectively. The terms $\rho \text{SAR}$, $P_c$, and $Q_m$ are the heat generated due to radiation, the power dissipation of circuitry, and the metabolic heating, respectively.

The Finite-Difference Time-Domain (FDTD) method [5] is a technique that transforms the previous bioheat equation to a discrete form with discrete time and space steps. The area under consideration is divided into cells of size $\delta$, and the temperature is evaluated in a grid of points defined at the centers of the cells. Temperatures are computed at equally spaced time instants with a time step equal to $\delta$. Therefore, from [6], the new bioheat equation is

$$T_{m+1}(i, j) = \left[1 - \frac{\delta b}{\rho C_p} - \frac{4\delta K}{\rho C_p \delta^2} \right] T_m(i, j)$$

$$+ \frac{\delta}{C_p} \text{SAR}_{\text{NF}} + \frac{\delta b}{\rho C_p} T_b + \frac{\delta}{C_p} P_c + \frac{\delta K}{\rho C_p \delta^2} \left[ T_m(i + 1, j) + T_m(i, j + 1) \right]$$

$$+ \frac{\delta}{\rho C_p} P_c$$

$$+ \frac{\delta}{\rho C_p \delta^2} \left[ T_m(i - 1, j) + T_m(i, j - 1) \right], \quad (4)$$

where $T_{m+1}(i, j)$ is the temperature of cell $(i, j)$ at time $m+1$, $\delta$ is the time step, and $\delta$ is the space step.

Using (2) and (4), the temperature increase at the location of the biosensor $(i, j)$ can be found. It is assumed that the temperature of the surrounding cells points is the normal body temperature (i.e., 37°C).

3. Related Work

The research on the possible biological effects caused by biosensors and how to mitigate those effects is very recent. Most of the existing research deals with other technical issues such as energy efficiency and quality of service. In this section, the limited available literature is briefly reviewed.

Tang et al. [6] were the first to propose rotating the cluster leadership in a cluster-based biosensor network to minimize the heating effects on human tissues. They proposed a genetic algorithm for computing a minimal temperature increase rotation sequence. Since using (4) in computing
the temperature increase due to a sequence is computationally expensive, they proposed a scheme for estimating the possible temperature increase due to a sequence.

In another work, Tang et al. [7] addressed the issue of routing in implanted biosensor networks. They proposed a thermal-aware routing protocol that routes the data away from high-temperature areas referred to as hot spots. The location of a biosensor becomes a hot spot if the temperature of the biosensor exceeds a predefined threshold. The proposed protocol achieves a better balance of temperature increase and shows the capability of load balance.

The above two works have motivated us to explore further the bioeffects of implanted biosensor networks. As a result, we noticed a lack of information on how to optimally operate an implanted biosensor network when bounds such as the maximum temperature increase exist. Most of the existing works assume that energy is the only limiting factor in the operation of Wireless Sensor Networks (WSNs). However, this is not the case in biosensor networks where the increase in temperature is a serious limiting factor.

We have approached the problem of how to optimally operate a biosensor network from the perspective of sensor scheduling and activation in conventional wireless sensor networks. Sensor scheduling is concerned with the problem of how to dynamically choose a sensor for communication with the base station. On the other hand, sensor activation is concerned with the problem of when a sensor should be activated. Many interesting works have been done in this regard. Next, these works are briefly reviewed.

In [8], the sensor scheduling problem is formulated as an MDP. The objective is to find an operating policy that maximizes the network lifetime. The state of a sensor is characterized by its current energy level only. Three kinds of channel state information are considered: global, channel statistics, and local. Considering only the energy level at each sensor gives rise to an acyclic (i.e., loop-free) transition graph which enables the MDP model to converge in one iteration. On the other hand, if the temperature of each sensor is included in the model, the transition graph of the underlying MDP becomes cyclic. This is because when the sensor cools down (i.e., its temperature decreases), it transitions back to a less hot state. An MDP model whose transition graph is cyclic needs more time to converge.

Dynamic sensor activation in networks of rechargeable sensors is considered in [9]. The objective is to find an activation policy that maximizes the event detection probability under the constraint of slow rate of recharge of the sensor. The state of the system is characterized by the energy level of the sensor and whether or not an event would occur in the next time slot. The recharge event is random and recharges the sensor with a constant charge. The model does not include the state of the wireless channel which is very crucial when temperature is considered.

Body sensor networks [10] with energy harvesting capabilities are another kind of WSNs in which each sensor has an energy harvesting device that collects energy from ambient sources such as vibration, light, and heat. In this way, the more costly recharging method which uses radiation is avoided. The interaction between the battery recharge process and transmission with different energy levels is studied in [11]. The proposed policies utilize the sensor’s knowledge of its current energy level and the state of the processes governing the generation of data and battery recharge to select the appropriate transmission mode for a given state of the network.

4. System Model

Figure 1 shows the system under study where a mobile subject, in this case an animal, has a biosensor implanted into its body. The biosensor has a built-in battery which is recharged by an RF power source. The role of the biosensor is to monitor and report interesting physiological events such as heart rate and blood pressure. The biosensor becomes incapable of detecting and reporting events if it does not have enough energy for transmission under any channel condition or the increase in its temperature exceeds a prespecified threshold. The latter condition causes a halt in system operation to allow the system to cool down.

Both the biosensor and RF power source are under the control of the base station which initiates the measurement process. The base station generates three control signals: Sleep and Sample targeted at the biosensor and Recharge targeted at the RF power source. The system state information is assumed to be available to the base station before it generates a control signal.

Mathematically, the system can be modeled as a discrete-state system which evolves in discrete time. Therefore, the time axis is divided into slots of equal durations $\Delta t$. At the beginning of each time slot, the state of the system is observed and a control signal is generated by the base station accordingly. Each time slot is long enough to transmit a complete packet carrying a measurement. Next, the elements of the system are described.

4.1. Biosensor. A biosensor typically contains four essential components: biorecognition, transducer, radio and battery.
The biorecognition system is made of elements such as enzymes and antibodies whose role is to produce a physiochemical change which is detected and measured by the signal transducer. The transducer carries out signal processing tasks. The radio circuitry is responsible for wireless communication. The battery provides power for all active modules in the biosensor and is recharged using RF energy. During a recharging period, the biosensor uses its radio module to collect energy and recharge the battery. Therefore, while its battery is being recharged, the biosensor cannot perform sensing and communication.

The location of a biosensor represents a critical point since it experiences the maximum temperature increase. This is because the tissues surrounding the biosensor might be heated continuously due to the local radiation generated by the biosensor itself and the radiation generated by the base station while recharging the biosensor.

In each time slot $t$, the state of the biosensor is characterized by two variables which are the current temperature $T_t$ and energy level $E_t$. There are $\tau + 1$ safe temperature levels; that is, $T_t \in \{0, 1, \ldots, \tau\}$, where the zero temperature level represents the normal body temperature and $\tau$ is an upper limit which must not be exceeded. Initially, the biosensor has a total energy of $\delta_0$ which is also the capacity of its battery. The energy required for the biosensor to successfully transmit its measurement to the base station is determined by the state of the wireless channel at the time of transmission. This transmission energy is denoted by $\delta_w$, where $w_t$ is the $i$th state of the wireless channel. The temperature increase due to a transmission energy of $\delta_w$ units is denoted by $F_w$.

At the beginning of each time slot, the base station may decide to recharge the biosensor; let it transmit its measurement or put it into sleep. The time required for a full recharge is random since it depends on the current temperature and energy levels. During this time, interesting events may occur but they will not be reported by the biosensor since it is being recharged. Also, the biosensor may be put into sleep for a random amount of time during which no measurements can be produced.

At the beginning of the next time slot (i.e., $t + 1$), the energy level at the biosensor is given by the following equation:

$$E_{t+1} = \begin{cases} E_t & \text{if } a_t = \text{Sleep} \\ E_t - \delta_w & \text{if } a_t = \text{Sample} \\ E_t + \delta_r & \text{if } a_t = \text{Recharge}, \end{cases}$$

where $a_t$ is the action taken by the base station at time $t$ and $\delta_r$ is the amount of energy gained by the biosensor. Similarly, the temperature of the biosensor at $t + 1$ is given by the following equation:

$$T_{t+1} = \begin{cases} \max\{T_t - F_s, 0\} & \text{if } a_t = \text{Sleep} \\ T_t + F_w & \text{if } a_t = \text{Sample} \\ T_t + F_r & \text{if } a_t = \text{Recharge}, \end{cases}$$

where $F_s$ is the amount by which the temperature of the biosensor decreases when it is put to sleep. In the same way, $F_r$ and $F_w$ are the amounts by which the temperature of the biosensor increases when it is recharged and when it is allowed to transmit its measurement, respectively. $F_r$ and $F_w$ can be calculated using (4). $F_s$ is constant since the SAR due to the base station is assumed to be constant. On the other hand, $F_w$ is not constant since the SAR due to the biosensor changes with the change in transmission energy. Therefore, before $F_w$ can be calculated, the SAR due to the radiation from the antenna of the biosensor is calculated using (2). Since (2) is a function of $f$, it is assumed that the current $f$ corresponding to each transmission power level is known.

4.2. Wireless Channel. The communication between the biosensor and base station occurs over a Rayleigh fading channel with additive Gaussian noise. Hence, the instantaneous received Signal-to-Noise Ratio (SNR) denoted by $\gamma$ is exponentially distributed with the following probability density function [12]:

$$P(\gamma) = \frac{1}{\gamma_0} \exp\left(-\frac{\gamma}{\gamma_0}\right),$$

where $\gamma_0$ is the average received SNR.

Such a wireless channel can be modeled as a Finite-State Markov Chain (FSMC) [13, 14]. The model can be built as follows. For a wireless channel with $K$ states, the state boundaries (i.e., SNR thresholds) are denoted by $\Gamma_1, \Gamma_2, \ldots, \Gamma_K, \Gamma_{K+1}$, where $\Gamma_1 = 0$ and $\Gamma_{K+1} = \infty$. The channel is said to be in state $s_i$ if the SNR is between $\Gamma_i$ and $\Gamma_{i+1}$, where $i = 1, 2, \ldots, K$. It is assumed that the SNR remains the same during packet transmission, and only transitions to the current or adjacent states are allowed.

The steady-state probability of the $i$th state of the FSMC is given by

$$P(s_i) = \exp\left(-\frac{\Gamma_i}{\gamma_0}\right) - \exp\left(-\frac{\Gamma_{i+1}}{\gamma_0}\right)$$

and thus the state transition probabilities are

$$P(s_{i+1} | s_i) = \frac{N(\Gamma_i)}{P(s_i)} \Delta t,$$

where $N(\Gamma_i)$ is the average number of times per unit interval that the SNR crosses level $\Gamma_i$ and $\Delta t$ is the packet duration.

$N(\Gamma_i)$ can be computed using the following equation [15]:

$$N(\Gamma_i) = \sqrt{2\pi \Gamma_i f_d e^{-\frac{\Gamma_i}{\gamma_0}}}$$

where $f_d$ is the maximum Doppler frequency defined as $f_d = v/\lambda$ with $v$ being the speed of the subject and $\lambda$ being the wavelength.

The above channel model has been verified to be precise when the fading process is slow [13], such as in biosensor applications.
5. MDP Formulation

An MDP is a model of a dynamic system whose behavior varies with time. The elements of an MDP model are the following [16]:

(1) system states,
(2) possible actions at each system state,
(3) a reward or cost associated with each possible state-action pair, and
(4) next state transition probabilities for each possible state-action pair.

The solution of an MDP model (referred to as a policy) gives a rule for choosing an action at each possible system state. If the policy chooses an action at time \( t \) depending only on the state of the system at time \( t \), it is referred to as a stationary policy. An optimal stationary policy exists over the class of all policies if every stationary policy gives rise to an irreducible Markov chain. This means that one can limit the attention to the class of stationary policies.

In order to obtain a policy from an MDP model, it is necessary to form and solve the so-called optimality equation (or Bellman’s equation). The following is the standard form of this equation with the maximization operator [17]:

\[
V_n(s) = \max_{a \in A(s)} \left[ f(s,a) + \sum_{s' \in S} \mathbb{P}(s,s',a)V_{n-1}(s') \right],
\]

where \( n \) is the iteration index, \( S \) is the set of system states (\( s \in S \)), \( A(s) \) is the set of actions possible when the system is at state \( s \), \( f(s,a) \) is the reward/cost per step, \( \mathbb{P} \) is the system state transition probability matrix, and \( V(s) \) is the optimal value of the objective function when the system is started at state \( s \) and the optimal policy is followed. Equation (II) can be solved using the classical policy iteration, value iteration, and relative value iteration algorithms [17]. Next, the details of the MDP model are given.

5.1. State Set. The state of the system at time \( t \) is described by the following 3-dimensional vector:

\[
s_t = (T_t, E_t, W_t),
\]

where \( T_t \), \( E_t \), and \( W_t \) are the current temperature of the biosensor, its energy level, and transmission power required for successful transmission at time \( t \), respectively. The total number of possible states is \(|S| = |T| \times |E| \times |W|\), where \( |T| \), \( |E| \), and \( |W| \) are the numbers of possible temperatures, residual energies, and transmission energy levels, respectively.

5.2. Action Set. In each time slot, the base station chooses an action based on the current state of the system. In each state \( s \), there are three possible actions:

\[
A(s) = \{\text{Sample}, \text{Recharge}, \text{Sleep}\},
\]

where the \text{Sample} action lets the biosensor generate a measurement and transmit it to the base station, \text{Recharge} action recharges the biosensor, and \text{Sleep} action puts the biosensor into sleep.

The \text{Sleep} action can be performed at every system state. The other two actions, however, can only be performed at system states, where the next temperature of the biosensor is within the safe temperature range. In addition, the \text{Sample} action can only be performed at system states, where the remaining energy is sufficient to make a successful transmission.

5.3. Reward Function. Since the objective is to maximize the expected number of samples that can be generated by the biosensor, the reward function is defined as

\[
R(s, \text{Sample}) = 1.
\]

This means that one unit of reward is earned every time the \text{Sample} action is performed. The long-run expected sum of rewards represents the average number of samples that can be generated by the biosensor with an initial energy of \( \mathcal{E}_0 \) units and maximum temperature increase of \( \tau \) units.

5.4. Transition Probability Function. After the action taken by the base station is performed, the system transits to a new state according to the transition probabilities of the present state of the wireless channel. Thus, the behavior of the system is described by \(|A|\) transition probability matrices, and each such matrix is of size \(|S| \times |S|\). Each matrix is denoted by \( P_{s,s'_{t+1}}(a) \) which is the probability that choosing an action \( a \) when in state \( s_t \) will lead to state \( s_{t+1} \). More formally, \( P_{s,s'_{t+1}}(a) \) can be written as follows:

\[
P_{s,s'_{t+1}}(a) = P(W_{t+1} = W' | W_t).
\]

5.5. Value Function. The problem of finding an optimal policy for maximizing the average number of samples is formulated as an infinite-horizon MDP using the average reward criterion [16]. So, let \( V_n(s_0) \) be the expected number of samples given that the policy \( \pi \) is used with an initial state \( s_0 \). Then, the maximum expected number of samples \( V^*(s_0) \) starting from state \( s_0 \) is given by

\[
v_n(s) = \max_{a \in A(s)} \left[ R(s,a) + \sum_{s_{t+1} \in S} \mathbb{P}(s_t, s_{t+1}, a)V_{n-1}(s_{t+1}) \right].
\]

The optimal policy \( \pi^* \) is the one that achieves the maximum expected number of samples at all system states.

The famous value iteration algorithm [17] is used to numerically solve the following recursive equation for \( n > 0 \):

\[
v_n(s) \to \infty, V_n \to V^*.
\]
6 International Journal of Distributed Sensor Networks

6. Handling Large-Size MDP Models

The size of the proposed MDP model depends on the number of biosensor states which is a function of the number of possible temperature and energy levels. As the number of biosensor states increases, the process of computing the transition probability matrices for the system becomes very time consuming. Also, the value iteration algorithm used for solving the MDP model becomes impractical. This section presents two methods (namely, Q-learning and heuristics) for handling MDP models with a large number of states.

6.1. Q-Learning. Reinforcement Learning (RL) offers an alternative for obtaining the optimal policy at a significantly lower computational cost. Using a simulation model of the system under study, the decision maker in an MDP is viewed as a learning agent whose task is to learn the optimal action in each possible state of the system. As Figure 2 shows, the optimal policy is learned while the system is being driven (i.e., simulated) by the actions selected by the learning agent which stores the results of its actions in a knowledge base. The actions of the decision maker become better over time as new knowledge is obtained. Eventually, the RL algorithm converges to an optimal policy which can be used in the physical system.

Q-Learning is an RL algorithm which was introduced in [18]. It is used for learning from experience. It requires that each entry in the decision-maker’s knowledge base corresponds to a state-action pair. The value stored in each entry is referred to as the Q-value and is a measure of the goodness of executing an action in a particular system state. The Q-value for a state-action pair \((s, a)\) is updated as follows:

\[
Q(s, a) = Q(s, a) + \alpha \left( r + \gamma \max_{j \in A(s')} Q(s', j) - Q(s, a) \right),
\]

where \(r\) is the immediate reward obtained after executing action \(a\) in state \(s\), \(s'\) is the next state, and \(A(s')\) is the set of possible actions in state \(s'\). \(\alpha\) and \(\gamma\) denote the learning rate and discount factor \((0 < \gamma < 1)\), respectively.

The Q-learning algorithm is shown as Algorithm 1. The interaction between the learning agent and the simulator (or environment) is divided into episodes. In each episode, the system transits through a sequence of states. The length of this sequence is controlled by the parameter \(Num\) which is the number of simulated time slots. In each simulated time slot, based on the current state of the system, the learner chooses an action either based on the \(e\)-policy or randomly. If the former is selected, the action with the highest Q-value is selected. After that, if the action is \(Sample\), a reward of one unit is earned; otherwise, the reward is zero. The action is then simulated and the next system state is observed. Next, the Q-value is updated using (18). Also, the new system state becomes the current one and the cycle repeats.

Although Q-learning is theoretically guaranteed to obtain an optimal policy, it requires that each state-action pair be tried infinitely often in order to learn the optimal policy. The quality of the learned policy depends on how much time is spent in learning and if every state-action pair can be tried. On the other hand, depending on the application, a certain percental difference between the learned and optimal policies might be tolerated. This is because the system states differ in the likelihood of being visited. Thus, a default action (like Sleep) can be assigned to system states with a low likelihood of being visited.

![Diagram](https://via.placeholder.com/150)

**Figure 2:** Using an RL algorithm (like Q-learning), the decision-making agent gradually learns the optimal policy. An action is applied to the system or simulated and then the resulting reward/cost is fed back into the knowledge base of the decision maker. The new knowledge obtained over time helps the decision maker to make better actions.

**Algorithm 1:** The Q-Learning algorithm.

```plaintext
for i = 1 to NumEpisodes do
    ps ← InitialState; \{E.g., \([0, 0, 0, 1]\)\}
    for j = 1 to NumIter do
        if rand ≤ 1 − \(\epsilon\) then
            \(a ← \max_{j \in A(ps)} Q(ps, j)\); \{(rand ∈ (0,1))\}
        else
            \(a ← Random(A(ps))\); \{Random Action\}
        end if
        if \(a == Sample\) then
            \(r ← 1\); \{Reward\}
        else
            \(r ← 0\);
        end if
        ns ← SimulateAction(ps, a);
        Update \(Q(ps, a)\); \{Using (18)\}
        ps = ns
    end for
end for
```
6.2. Heuristic. Since it is difficult to describe the structure of the optimal policy, a heuristic policy is proposed in this section. The goal is to design a policy which mimics the behavior of the optimal policy as close as possible. However, before presenting such a policy, a greedy one is given to provide insight into the design of any heuristic policy.

The greedy policy is computed using Algorithm 2. The inputs to this algorithm are the set of possible system states and the set of feasible actions for each system state. The computed policy is greedy in the sense that for each system state, the feasibility of actions is checked in the following order: Sample, Recharge, and then Sleep. The first feasible action is associated with the corresponding system state.

As will be shown by simulations in the next section, the greedy policy is poor since it is based on a fixed order of actions. Therefore, Algorithm 2 needs to be extended to allow for a dynamic selection of actions. This objective is accomplished by introducing two control parameters: $\alpha$ and $\beta$. With these two control parameters, the Sample and Recharge actions are not selected in a specific order or whenever they are feasible. Algorithm 3 shows how the control parameters and new heuristic policy are computed.

The essence of Algorithm 3 is as follows. If the current temperature (denoted by $T$) of the biosensor is low and its current energy level (denoted by $E$) is high, then the condition $\alpha \leq \beta$ would more likely be true and thus the Sample action could be executed. However, this would not be the case when the available energy is very close to zero. In this case, the opposite condition (i.e., $\alpha \geq \beta$) would more likely be true and thus a Recharge could be performed. If neither of the two conditions is true, the biosensor is put to sleep and thus its temperature decreases.

7. Numerical and Simulation Results

In this section, an example is first presented to illustrate the viability of the proposed MDP model. Then, the performance of the optimal policy is compared to that of the approximate policies using simulation. The impact of various system parameters on the performance of the system is also evaluated. The simulation was performed using a simulator written in Matlab [19]. Each simulation was run for a duration of 100000 time slots, and each data point is the average of 10 simulation runs. The number of channel states ($W'$) is four, and the channel state boundaries are randomly generated.

7.1. Illustrative Example. In this example, a wireless channel with two states is considered. The channel state transition probabilities are calculated using (9). Table 1 shows the values of the parameters involved. Figures 3(a) and 3(b) show the expected number of samples when there is no recharge and when recharge is allowed, respectively. The expected number of samples is expressed as a function of the maximum safe temperature level ($T^*$) and initial energy ($E_0^*$). The first observation is that if recharge is allowed, more samples are expected to be generated by the biosensor. In the case when recharge is not allowed, the expected number of samples is limited only by the amount of initial energy. This is confirmed by Figure 3(a) where for the same initial energy, the same expected number of samples is obtained when $\tau$ is varied.
When recharge is allowed, the maximum safe temperature level ($\tau$) plays a critical role. This is due to the temperature increase caused by the recharge action. In Figure 3(b), for the same initial energy, the expected number of samples increases as $\tau$ is varied. Increasing $\tau$ enables the Recharge action to be performed more often. On the other hand, as one would expect, if $\tau$ is fixed and $(E_0)$ is varied, the expected number of samples slightly increases when $\tau$ is small. However, when $\tau$ is large ($\geq 6$), the maximum possible expected number of samples can be achieved when $E_0$ is at its maximum value. Therefore, for this particular example, if $E_0 = 10$, the optimal value for $\tau$ is 6.

Figures 4(a) and 4(b) show the optimal action for each possible system state. In Figure 4(a), for channel state 1, the Sample action is performed in 70% of the system states. The Sleep action is performed whenever the temperature reaches the maximum safe level ($\tau$), and the Recharge action is performed when the remaining energy is zero and the temperature is below $\tau$.

By contrast, in Figure 4(b), for channel state 2, the Sample action is performed only once at the initial system state. For this channel state, due to the higher cost of transmission, the biosensor is put in the sleep mode most of the time. However, since the cost of the Recharge action is independent of the channel state, the system recharges itself more often to enable more samples to be generated when the wireless channel switches to a state with a lesser transmission energy requirement (i.e., channel state 1).

7.2. Comparative Analysis. In order to be able to appreciate the merit of any approximate policy, a more meaningful performance criterion is needed. In this work, the average number of time slots needed to generate a sample is used as a criterion to distinguish between the different policies.
available to run a system. It is calculated as the total simulation time divided by the average number of samples generated by the system while being operated by a certain policy. This measure takes into account the effect of the Recharge and Sleep actions.

For example, consider Figure 5. In this figure, $\tau$ is fixed at five while $E_0$ is varied. The greedy policy is very costly since it requires the largest amount of time before a sample can be generated. The difference in the amount of time required by the heuristic policy and that required by the optimal policy stays around two time slots. This is a 75% reduction in time compared to the greedy policy. The $Q$ policy is the best approximate policy. On average, the difference with the optimal policy stays around 1.1 time slots.

Figure 6 shows the amount of time required to generate a sample when $E_0$ is fixed at five and $\tau$ is varied. In this figure, when $\tau = 1$, the greedy policy outperforms both the $Q$ policy and heuristic policy. A difference of three time slots is observed. This can be explained as follows. In the $Q$ and heuristic policies, the Recharge action can be performed in one state only (i.e., when $T = E = 0$). On the other hand, with the greedy policy, the Recharge action can be performed in more than one state (i.e., whenever $T = 0$). This, of course, leads to a reduction in the average amount of time needed to generate a sample. Other than that, for $\tau \geq 2$, the $Q$ and heuristic policies are always better than the greedy policy, and their performance is close to that of the optimal policy.

8. Conclusions

The increase in temperature due to the heat generated by biosensors is a limiting factor in the operation of biosensor networks. This problem can be modeled as a stochastic control problem using the framework of Markov decision processes. The solution is an optimal policy which ensures that the maximum safe temperature level is not exceeded. In order to handle large-size MDP models, it is shown how $Q$-learning can be used for obtaining the optimal policy. In addition, a heuristic policy is proposed. Its performance is comparable to that of the policies obtained by the MDP model and $Q$-learning.

This work can be extended in the following directions. First, the scenario of more than one rechargeable biosensor should be studied. In this case, the number of possible system states is exponentially huge. Thus, techniques for eliminating equivalent states would be necessary. Second, the performance of other reinforcement learning techniques should be investigated, especially for models with a huge state space. Third, algorithms for computing better heuristic policies should be developed to mitigate the problem of finding better approximate policies.

Acknowledgment

The first author would like to acknowledge the financial support of King Fahd University of Petroleum and Minerals (KFUPM) while conducting this research.

References

[1] National Council on Radiation Protection and Measurements (NCRP), "A practical guide to the determination of human exposure to radiofrequency fields," Tech. Rep. 119, NCRP, 1993.
[2] International Electrotechnical Commission (IEC), "Medical electrical equipment, part 2–33: particular requirements for the safety of magnetic resonance equipment for medical diagnosis," IEC 60601-2-33, 1995.
[3] S. K. S. Gupta, S. Lahwani, Y. Prakash, E. Elsharawy, and L. Schwiebert, "Towards a propagation model for wireless biomedical applications," in Proceedings of the International
[4] H. H. Pennes, “Analysis of tissue and arterial blood temperatures in the resting human forearm,” Journal of Applied Physiology, vol. 85, no. 1, pp. 93–122, 1948.

[5] D. M. Sullivan, Electromagnetic Simulation Using the FDTD Method, IEEE Press, New York, NY, USA, 2000.

[6] Q. Tang, N. Tummala, S. K. S. Gupta, and L. Schwiebert, “Communication scheduling to minimize thermal effects of implanted biosensor networks in homogeneous tissue,” IEEE Transactions on Biomedical Engineering, vol. 52, no. 7, pp. 1285–1294, 2005.

[7] Q. Tang, N. Tummala, S. K. S. Gupta, and L. Schwiebert, “TARA: Thermal-Aware Routing Algorithm for implanted sensor networks,” in Proceedings of the 1st IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS ’05), vol. 3560, pp. 206–217, Springer, July 2005.

[8] Y. Chen, Q. Zhao, V. Krishnamurthy, and D. Djonin, “Transmission scheduling for optimizing sensor network lifetime: a stochastic shortest path approach,” IEEE Transactions on Signal Processing, vol. 55, no. 5, pp. 2294–2309, 2007.

[9] N. Jaggi, K. Kar, and A. Krishnamurthy, “Rechargeable sensor activation under temporally correlated events,” Wireless Networks, vol. 15, no. 5, pp. 619–635, 2009.

[10] G. Z. Yang, Body Sensor Networks, Springer, Berlin, Germany, 2006.

[11] A. Seyedi and B. Sikdar, “Energy efficient transmission strategies for Body Sensor Networks with energy harvesting,” in Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS’08), pp. 704–709, IEEE, March 2008.

[12] J. G. Proakis, Digital Communications, McGraw-Hill, New York, NY, USA, 2000.

[13] H. S. Wang and N. Moayeri, “Finite-state Markov channel—a useful model for radio communication channels,” IEEE Transactions on Vehicular Technology, vol. 44, no. 1, pp. 163–171, 1995.

[14] Q. Zhang and S. A. Kassam, “Finite-state markov model for rayleigh fading channels,” IEEE Transactions on Communications, vol. 47, no. 11, pp. 1688–1692, 1999.

[15] W. C. Jakes, Microwave Mobile Communications, John Wiley & Sons, New York, NY, USA, 1974.

[16] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, New York, NY, USA, 2005.

[17] D. P. Bertsekas, Dynamic Programming and Optimal Control, Volume I, Athena Scientific, Belmont, Mass, USA, 2000.

[18] C. Watkins, “Learning from delayed rewards,” 1989, Springer, http://www.cs.rhul.ac.uk/~chrisw/thesis.html.

[19] The MathWorks Inc., http://www.mathworks.com/.
Submit your manuscripts at http://www.hindawi.com