International Journal of Neural Systems, Vol. 19, No. 4 (2009) 227–240


World Scientific
www.worldscientific.com

# TREATING EPILEPSY VIA ADAPTIVE NEUROSTIMULATION: A REINFORCEMENT LEARNING APPROACH

JOELLE PINEAU*, ARTHUR GUEZ† and ROBERT VINCENT‡
*School of Computer Science, McGill University*
*Montreal, QC, Canada*
**jpineau@cs.mcgill.ca*
†*aguez@cs.mcgill.ca*
‡*bert@cs.mcgill.ca*

GABRIELLA PANUCCIO§ and MASSIMO AVOLI¶
*Montreal Neurological Institute, McGill University*
*Montreal, QC, Canada*
§*gabriella.panuccio@mail.mcgill.ca*
¶*massimo.avoli@mcgill.ca*

This paper presents a new methodology for automatically learning an optimal neurostimulation strategy for the treatment of epilepsy. The technical challenge is to automatically modulate neurostimulation parameters, as a function of the observed EEG signal, so as to minimize the frequency and duration of seizures. The methodology leverages recent techniques from the machine learning literature, in particular the reinforcement learning paradigm, to formalize this optimization problem. We present an algorithm which is able to automatically learn an adaptive neurostimulation strategy directly from labeled training data acquired from animal brain tissues. Our results suggest that this methodology can be used to automatically find a stimulation strategy which effectively reduces the incidence of seizures, while also minimizing the amount of stimulation applied. This work highlights the crucial role that modern machine learning techniques can play in the optimization of treatment strategies for patients with chronic disorders such as epilepsy.

*Keywords*: Epilepsy; neurostimulation; reinforcement learning.

## 1. Introduction

Epilepsy is one of the most common disorders of the nervous system, afflicting approximately 0.6% of the world's population. Currently, anti-convulsant drug therapies are the most popular approach to alleviate seizures, but about one third of epileptic patients have seizures that cannot be controlled by medication, highlighting the need for novel therapeutic strategies.[35] Electrical stimulation procedures have recently emerged as a promising alternative. Implantable electrical stimulation devices are now an important treatment option for patients who do not respond to anti-epileptic medication. Both direct deep brain stimulation[45,47,30,7,44,43] and vagus nerve stimulation[18,42] have demonstrated the potential to shorten or even prevent seizures. The effect has also been shown *in vitro*.[4,10] In all cases, the technology is similar: a small pacemaker-like device is implanted in the patient and sends electrical stimulation to the brain. Given this technology, there are many ways in which stimulation can be applied. For example one can vary the amplitude, duration, or frequency of the electrical stimulation. But because little is known about the optimal stimulation strategy, the most common approach is to hand-tune settings of these parameters through trial-and-error.

The main contribution of this paper is to propose a methodology to automatically learn a closed-loop stimulation strategy from experimental data. There are significant advantages to this approach.

Closed-loop strategies are in general more powerful than open-loop ones because sensory feedback (in this case, field potential recordings) is integrated into the stimulation strategy. The stimulation pattern can therefore respond in real-time to the patient's brain activity. In addition, because the strategy is optimized automatically, it can adapt to each individual, and over time. The long-term goal of this work is to build a device which, through an adaptive control system, can respond to a patient's changing condition over time without direct operator intervention.

The mathematical framework we investigate to optimize stimulation strategies is known, in the area of computer science, as *reinforcement learning*.[37] This framework is specifically designed to address the problem of optimizing action sequences in dynamic and stochastic systems. Applying reinforcement learning in the context of deep brain stimulation gives us a mathematical framework to explicitly maximize the effectiveness of stimulation, while simultaneously minimizing the overall amount of stimulation applied thus reducing cell damage and preserving cognitive and neurological functions. Reinforcement learning is particularly well suited to the problem at hand because, unlike traditional control theory, it does not require a detailed mathematical description of the relevant neural circuitry in order to optimize a stimulation strategy. Instead, it learns a control strategy through direct experience, which is advantageous given that the brain is extremely challenging to model.

The idea of applying reinforcement learning to optimize deep-brain stimulation strategies has not been sufficiently explored previously. It stands in contrast to most recent efforts by researchers to design neurostimulation devices which trigger stimulation in response to an automated seizure detection algorithm.[24] An important feature of the reinforcement learning paradigm is that it does not necessarily rely on having accurate prediction or detection of seizures. This is a significant advantage given that developing accurate methods for seizure prediction is proving to be extremely challenging and few conclusive results exist.[28]

While the long-term goal is to develop an adaptive system for therapeutic purposes, in this paper we focus on applying reinforcement learning to optimize deep-brain stimulation strategies using data collected from an *in vitro* model of epilepsy.[4,10] Animal models of epilepsy have been used extensively to analyze the biological mechanisms underlying epilepsy, as well as to study the effect of various non-adaptive stimulation strategies. An excellent review of the latter is provided by Durand and Bikson.[11]

The paper is organized as follows. Section 2 describes the particular animal model used as well as our data collection and analysis protocol. Section 3 contains a technical presentation of the reinforcement learning algorithm. Section 4 describes how the reinforcement learning algorithm can be applied to the problem of adaptive neurostimulation. Finally Section 5 analyzes the application of the reinforcement learning framework to select optimal strategies using pre-recorded data from an *in vitro* model of epileptiform behavior. Our results demonstrate that an adaptive strategy can be learned from such data. Analysis of the learned adaptive strategy on pre-recorded data show a reduction in the duration of seizures (compared to control slices), as well as a reduction in the total amount of stimulation applied compared to periodic pacing strategies. We conclude the paper with a discussion of longer term research questions that arise as we move from the animal model to treating human patients.

## 2. Model and Methods

Epilepsy is a dynamical disease, typically characterized by the sudden occurrence of hypersynchronous discharges that involve multiple neuronal networks. Seizure activity can be induced in various ways, for example, by elevating extracellular potassium ($K^+$), which has been done in both *in vivo* and *in vitro* preparations.[39]

We also know that ictal discharges can be reduced and eventually abolished by activating hippocampal outputs, a procedure that is achieved by delivering repetitive electrical stimuli. For example we have found that in pilocarpine-treated epileptic rat slices, low frequency (0.1–1.0 Hz) repetitive stimuli delivered in subiculum can reduce, but not halt, 4-aminopyridine-induced ictal discharges.[9] Overall this evidence suggests that electrical stimulation may interrupt the synchronous activity of neuronal populations.

### 2.1. *Electrophysiological recordings*

The dataset in our experiments consists of field potential recordings of seizure-like activity in rat

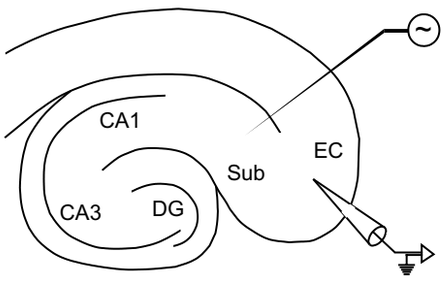


Fig. 1. Schematic of the hippocampus-EC slice. Relevant substructures are labeled.

brain slices maintained *in vitro*. A series of four recordings (each from a different slice, coming from a different animal) were made using a two-dimensional hippocampus-entorhinal cortex (EC) slice, as depicted in Fig. 1. Slices were obtained from male adult Sprague-Dawley rats (250–350 g) following standard procedures as previously described[10] and were maintained in an interface tissue chamber, where they were continuously superfused at 1 ml/min with carbogenated ($O_2$ 95%, $CO_2$ 5%) artificial cerebrospinal fluid containing the convulsant drug 4-aminopyridine.[3] Recording microelectrodes were placed in the deep layers of the EC. Recordings were sampled at a rate of 5 KHz, however for the purposes of our analysis, all recordings were filtered, to roll off frequencies above 100 Hz. In total, our analysis uses 7 hours of recorded data (roughly equally divided between the four slices).

Electrical stimulation was applied to the subiculum using low-frequency single-pulse patterns with varied timing. Each slice was subject to a stimulation protocol consisting of seven phases of stimulation patterns. Each sequence began with a control period of recording with no stimulation. Then, stimulation was applied for several minutes at a fixed low-frequency (1.0 Hz). Stimulation was then turned off and the slice was allowed to return to baseline for a period of several minutes. This process was repeated with stimulation at different rates (0.5 Hz and 2.0 Hz), always interleaving, between each stimulation phase, a prolonged recovery period during which no stimulation was performed. Stimulation intensity (100–250 $\mu$A biphasic pulse-wave width 100 $\mu$s) remained fixed throughout the experiments. Slices which did not exhibit good suppression at 1.0 Hz were excluded from the dataset because presumably presenting with weak connection between the two regions of interest, i.e. the EC and the subiculum.

Figure 2 shows a sample trace recorded from the EC while stimulating the subiculum at 0.5 Hz. An ictal event starts around $t = 20$ sec. The stimulation artifacts are also visible in this recording. In general, the actions may or may not be visible in the EEG signal, depending on the sample rate and relative electrode placement.

### 2.2. *Signal processing*

Each trace was divided into a set of overlapping frames of 65536 samples (approximately 13 seconds)

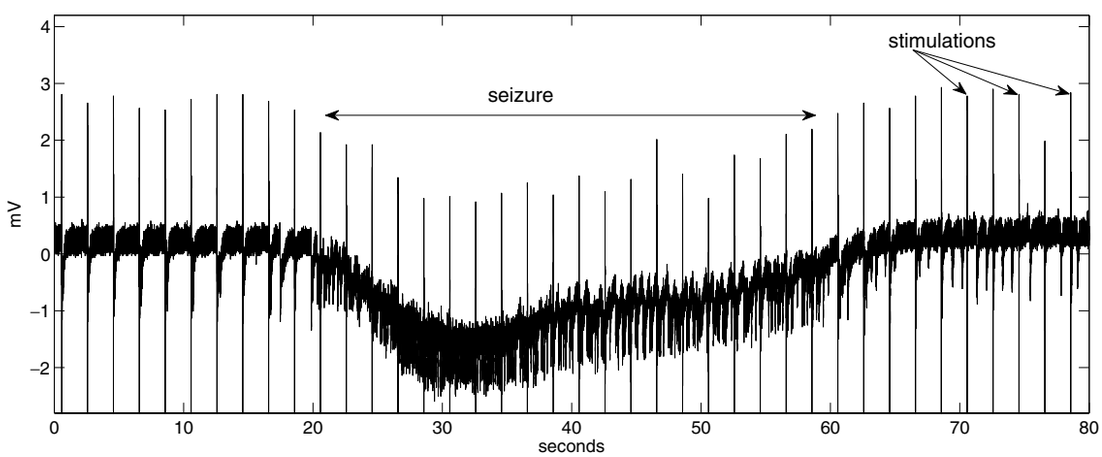


Fig. 2. Trace example recorded in the entorhinal cortex. Stimulation is applied to the subiculum at 0.5 Hz. An ictal event appears in the first half, lasting approximately 45 seconds. Periodic stimulation artifacts are observed at 2-second intervals. Inter-ictal spikes are also observed.

in length, with each frame beginning 8192 samples after the previous frame. Each frame is smoothed with a Hann window and normalized, and the mean, range, and energy of the signal is calculated. A discrete fast Fourier transform is used to extract spectral magnitude features from the frame. Within each frame, the smoothing, normalization, and Fourier transform is repeated for the final half frame (32768 samples), quarter frame (16384 samples), eighth frame (8192 samples), and sixteenth frame (4096 samples). Low frequency components are extracted from the full-frame spectrum, and high frequency components from the subframe spectra. These features are combined with the mean, range, and energy of each subframe to yield a 114-dimensional continuous feature vector. Many other features could be extracted, for example those proposed in the literature on seizure prediction and on EEG analysis.[28,1,31] In fact the question of feature selection is a challenging statistical problem, which will be the subject of future investigations. The other information which could be included is the time elapsed since beginning of the pulse train.[a] We do not include this information in the current implementation, because we assume that all recordings we use feature periodic stimulation that has been applied for a sufficiently long time to ignore edge effects.

### 2.3. *Data labeling*

The adaptive control algorithm described in the following section requires a number of traces with hand-annotated state information, for automatically learning the optimal stimulation strategy. Therefore all recorded traces were labeled by hand, indicating on each frame whether it features ictal or normal activity, as well as which stimulation protocol was used at the time. In the future, this step could be performed by an automatic seizure detection algorithm.

## 3. Adaptive Control Algorithm

Questions of prediction and control in dynamical systems have a long history in engineering and computer science. A variety of computational methods from these fields have been proposed to automatically detect or predict epileptic seizures from EEG recordings.[20,25,34,26,15,36] But much less effort has been spent on applying equally principled mathematical tools to the question of optimizing stimulation strategies. *In vitro* experiments have investigated the application of electrical stimulation based on periodic pacing,[3,21,23] nonlinear control[38] and feedback control.[46,16] However most of these methods are not automatic (in the sense that the strategy is learned directly from data), and some are neither adaptive (in the sense that the strategy evolves over time), nor optimal (in the sense that it minimizes a cost function).

More recently, models based on chaotic oscillator networks have been proposed, as a means of controlling epileptic seizures.[40,41,8] These models aim to minimize spatial synchronization via a feedback mechanism. These methods have the advantage that they require no training period. The results so far are limited to theoretical models of epilepsy, and their efficacy with animal models is not known. Nonetheless the results with the theoretical models confirm that open-loop periodic stimulation (as currently used in clinical trials) can be inefficient at achieving desynchronization, compared to a closed-loop feedback control strategy which can require much less stimulation power. Furthermore these methods generally require that the parameters of the control strategy (e.g. control gain, feedback threshold) be set by hand, or through trial-and-error. One of the advantages of the method we propose in this paper is that it uses automated learning techniques to gradually optimize the setting of the control parameters as it acquires data.

### 3.1. *Reinforcement learning*

Reinforcement learning is one of the leading techniques in computer science and robotics for automatically learning optimal control strategies in dynamical systems. The technique was originally inspired by the trial-and-error learning studied in the psychology of animal learning (thus the term "learning"). In this setting, good actions by the animal are positively reinforced and poor actions are negatively reinforced (thus the term "reinforcement"). Reinforcement learning was formalized in computer

[a]Presumably, applying a single pulse is not the same as applying a sequence of 10, or 100, or more; the system adapts to trains and responds differently depending on whether the train is of long or short duration.

science and operations research by researchers interested in sequential decision-making for artificial intelligence and robotics, where there is a need to estimate the usefulness of taking sequences of actions in evolving, time varying system.[22,37] It is especially useful in situations in which the agent's environment is stochastic, and for poorly-modeled problem domains in which the optimal control strategy is not obvious.

Recent developments in reinforcement learning have brought about a wealth of new algorithmic techniques, which can be used to automatically learn good action strategies directly from experimental data, yet the application of reinforcement learning to medical treatment design is very recent.[19,32] In this section, we describe how reinforcement learning can be used to directly optimize stimulation patterns of a closed-loop stimulation device, without necessarily requiring accurate seizure prediction.

Informally, the learning problem can be formulated as follows: at every moment in time, given some information about what happened to the signal previously (our *state*), we need to decide *which stimulation action* we should choose (if any) so as to minimize seizures *now and in the future*.

Considering the problem more formally, we assume the underlying dynamical system can be modeled as a Markov decision process (MDP).[5,33] The MDP model is defined by a **set of states**, $\mathcal{S}$, describing the space of observable variables, and a **set of actions**, $\mathcal{A}$, describing the available input set. In our case, the states are defined by the post-processed EEG recordings (i.e. the feature vector described in Sec. 2.2). The (discrete) set of actions corresponds to the different stimulation frequencies applied during data collection (Sec. 2.1).

Upon performing an action $a \in \mathcal{A}$ in state $s$, the learning agent receives a **scalar reward**, $r = R(s,a)$. This reward serves as a reinforcement signal to the agent, indicating which actions are good (=high reward) and which actions are to be avoided (=low reward). The reward can be positive or negative, but must be finite.

After an action is performed, the environment moves to a new state $s'$ according to some **conditional probability distribution**, $P(s'|s,a)$. Time is modeled as a series of discrete steps with $0 \leq t \leq T$, corresponding to the interval at which a decision must be made regarding the choice of action. At every time step, the state is assumed to be a sufficient statistic for the past sensor observations; this is the so-called *Markov* assumption.

The primary objective is to find a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maps each state to an action such as to maximize the expected total reward over some time horizon:

$$R_T = E\left[\sum_{t=0}^{T} \gamma^t r_t\right]. \tag{1}$$

Here $\gamma \in (0,1]$ is a discount factor for future rewards (it can be thought of as the agent's probability of surviving to the next time step). For $T = \infty$, $\gamma$ must be less than one to preclude an infinite total reward. For finite $T$ we can allow $\gamma = 1$.

Given this formulation, we can write the value of a given state if the agent follows a fixed policy $\pi$ as:

$$V^{\pi}(s) = E_{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]. \tag{2}$$

We define the *optimal* value for a state $V^*(s)$ to be:

$$V^*(s) = \max_{\pi} E_{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right], \tag{3}$$

which we can expand to the recursive equation:

$$V^*(s) = \max_{a \in \mathcal{A}}\left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s')\right). \tag{4}$$

This equation is often referred to as the **value function**. Here the value of a state is the maximum of the reward possible in the current state ($s$) plus the expected value over the successor states ($s'$), presuming that the agent behaves optimally at every subsequent time step. The corresponding optimal policy $\pi^*(s)$ is defined as:

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}}\left(R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s')\right). \tag{5}$$

It is also sometimes useful to express the value of a state-action pair, which defines the expected long-term reward of applying action $a$ when in state $s$:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q^*(s',a'). \tag{6}$$

This is sometimes referred to as the **Q-function**. From this, we can directly compute the optimum policy:

$$\pi^*(s) = \operatorname*{argmax}_a Q^*(s,a). \tag{7}$$

In many real-world problems, the transition probabilities are not known in advance, thus it is not possible to solve the above equations exactly. However, if enough empirical data is collected, for example, using the protocol described in Sec. 2, it is possible to treat this as a set of training trajectories to estimate the Q-function using the Fitted Q-Iteration algorithm.

### 3.2. *Fitted Q-iteration algorithm*

To apply Fitted Q-Iteration, it is necessary to begin by pre-processing the trajectories such that the state, action and reward information are extracted in a sequence of atomic events. This produces a set $\mathcal{F}$ of 4-tuples of the form $\langle s_t, a_t, r_t, s_{t+1}\rangle$, where each tuple is an example of the one-step transition dynamics of the system. This forms the input set for the Fitted Q-Iteration algorithm. The core of this algorithm is simple. It consists of repeatedly applying the following recurrence relation:

$$\hat{Q}_k(s_t, a_t) = r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{k-1}(s_{t+1}, a'). \tag{8}$$

In cases where the set of possible states can be finitely enumerated, this sequence can converge to the optimal $Q$ function (Eq. (6)) under some conditions.[33] In cases where the state space is very large (or continuous), it is necessary to assume a functional form for $\hat{Q}_k$, and use a regression algorithm to learn the mapping $Q : S \times A \rightarrow \Re$. Throughout our experiments, the term $\hat{Q}_k$ is approximated using Extremely Randomized Tree Regression.[14,12] This method has been shown to be effective in settings with large numbers of weak variables and substantial noise, as well as being computationally efficient.

## 4. Adaptive Neurostimulation

This section describes how the reinforcement learning algorithm outlined above can be applied to automatically learn an optimal neurostimulation policy for the treatment of epilepsy.

### 4.1. *Reinforcement learning problem definition*

Our state space $\mathcal{S}$ is constructed such that each element $s_t$ is a vector of 114 continuous dimensions, summarizing past EEG activity. Our action set $\mathcal{A}$ consists of four options: no stimulation, and stimulation at one of the fixed frequencies of 0.5, 1.0, or 2.0 Hz. Each frame is assigned an action $a_t$ based on the labeling information (Sec. 2.3).

We define a reward function

$$r_t = R_{\text{seizure}}(s_t) + \alpha R_{\text{stim}}(a_t) \tag{9}$$

to penalize both stimulation and seizure occurrences. We assume $R_{\text{seizure}}(s_t)$={-1 if seizure is occurring at time $t$, 0 otherwise} and $R_{\text{stim}}(a_t)$={-1 if stimulation is applied at time $t$, 0 otherwise}. This reward function requires a quantitative trade-off between the penalty for occurrence of a seizure, and the penalty for applying stimulation. This trade-off is defined by the parameter $\alpha$. In most experiments described below, we assume that a seizure is substantially more costly than delivering a single stimulation event (unless mentioned otherwise, we assume $\alpha = 0.04$). Changing this parameter may affect the learned stimulation strategy; we investigate this further in the experiments presented below.

Each element of the training set $\mathcal{F}$ is constructed by concatenating the experience-tuples $\langle s_t, a_t, r_t, s_{t+1}\rangle$.

We assume a discrete time step of 1.6 seconds (= 8192 samples). This is sufficient to compute our input features in real time, yet is sufficiently short to allow flexibility in the learned policy. For all of our experiments, the discount factor is $\gamma = 0.95$; this is a common choice in the reinforcement learning literature.

### 4.2. *Learning the regression function*

The algorithmic approach for the Extremely Randomized Tree regression is analogous to that proposed by Ernst *et al.*[13] (the reader is referred to that publication for details of what we outline next). A few of the parameter choices are worth discussing briefly. Throughout the experiments presented below we assume a set of $M = 70$ regression trees for each action. The estimate $\hat{Q}(s,a)$ is obtained by averaging the value returned by each of these trees at the current state $s$. We repeat this individually for each

action, and choose the action with maximal value. The number of candidate tests considered before expanding a node (defined by the parameter $K$) is set to 40. Finally, the minimum number of elements at each leaf (parameter $n_{min}$) is set to 5. We did not extensively tune these parameters; this could be done through a cross-validation procedure. In general we found that performance of the algorithm was quite robust to these parameter choices (within an order of magnitude).

Throughout an initial learning phase (lasting 30 iterations), the Fitted Q-Iteration algorithm is applied over the full set of trees and we allow the set of trees to be rebuilt entirely at each iteration. After this first phase, the structure of the trees is fixed and in subsequent iterations only the value at each node is allowed to vary. This second phase continues until the Bellman error falls below a given threshold.[b] This two-phase learning is common to ensure proper convergence.

The output of this learning phase is the regression function $\hat{Q}(s,a)$, defined for any state-action pair $(s,a)$. This function estimates the expected long-term cumulative reward that can be obtained by applying any action $a$ in any state $s$. The optimal action choice for each state can be extracted using Eq. (7). During deployment of the neurostimulation system, it is sufficient to store the $\hat{Q}(s,a)$ in memory, and repeatedly apply Eq. (7) as the state of the system evolves, so that the best possible neurostimulation action is selected at every time step. Thus we get an online control strategy which adaptively changes in response to changes in the dynamics of the system.

It is worth noting that other types of regression function could be used to fit the Q-function. We experimented also with linear regression, as well as neural networks, but found the random tree approach to yield better empirical performance.[17]

### 4.3. *Analysis method*

Finally, we turn to the question of validating the learned adaptive neurostimulation strategy. The preferred method for evaluating the performance of the strategy learned by reinforcement learning is to deploy it directly *in vitro*, and measure seizure incidence and duration, compared to other (non-adaptive) strategies of stimulation. However, this approach requires substantial time and resources, thus we begin our analysis by looking at performance metrics over the pre-recorded data.

Instead we consider quantitative measures which can be estimated using a *hold-out* test set, which is separate from our training data. This is a common technique in machine learning, whereby part of the recorded data is used to learn the regression function, and the remaining data is used to quantify the error in the estimate. Our original data set includes recordings from four animal slices. Therefore during testing we perform four-fold cross-validation, whereby the Q-function is estimated using data from three different slices, and we then measure performance on the fourth slice. We then repeat with all slice permutations. This means that data in the test set comes from a different animal than the training data. It is well-documented that epileptic seizures vary greatly between animals (and individuals), therefore this is an important test for the generalizability of our approach. In future work, an individual Q-function could be learned for each patient (or slice), using the algorithm outlined above, thereby providing a neurostimulation strategy that is specific to each individual.

There is another subtle difficulty in using a test set to validate a target policy (e.g. the learned optimal policy, $\pi^*$). That is the fact that the test set was collected *using a behavior policy*, $\pi$, which is different from the target policy. We cannot simply compute a score over the test set. Instead, we create a surrogate data set for the target policy by using rejection sampling to select only those segments of the test set which are consistent with the target policy. Recall that the test set is divided into single-step episodes: $\langle s_i, a_i, r_i, s_{i+1}\rangle$. We define an indicator function:

$$I_\pi(s_i,a_i) = \begin{cases} 1 & \text{if } \pi(s_i) = a_i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

to flag experience-tuples where the action in the test set ($a_i$) matches the target policy ($\pi(s_i)$). We exclude all experience-tuples that do not match the target policy. Using this indicator function, we consider two different scores to quantify the performance of the adaptive neurostimulation strategy.

[b]The Bellman error is defined to be $|\hat{Q}_k - \hat{Q}_{k-1}|$.

The first score is an estimated proportion of seizure steps when following a particular strategy $\pi$. Again, we compare the action selected by the policy and the action in the test trace for each experience-tuple from the test trace, and count the number of states which were labeled as "seizure":

$$\hat{S}_\pi = \frac{\sum_{i=0}^{N} I_\pi(s_i, a_i) I_{\text{seizure}}(s_i)}{\sum_{i=0}^{N} I_\pi(s_i, a_i)}, \tag{11}$$

where $I_{\text{seizure}}(s_i)$ indicates whether state $s_i$ was hand-labeled as a seizure (1 if yes, 0 if no). Recall that data instances are defined on a 1.6-second window interval.

The second score calculates the estimated value function (i.e. discounted sum of rewards). Formally,

$$\hat{V}^\pi = \frac{\sum_{i=0}^{N} I_\pi(s_i, a_i)[r(s_i) + \gamma \hat{Q}(s_{i+1}, \pi(s_{i+1}))]}{\sum_{i=0}^{N} I_\pi(s_i, a_i)}, \tag{12}$$

where $\hat{Q}$ is the estimated Q-function calculated by the regression algorithm (Eq. (8)). For fixed stimulation strategies, which were in fact deployed during data collection, we use the empirical return (Eq. (1)) instead. This second score is considered because it reflects the expected long-term accumulated reward. Since our reward function is a linear combination of the amount of both stimulation and seizure, this is an aggregate measure of the optimization over these two components.

## 5. Results

Many *in vitro* studies have investigated effectiveness of low-frequency periodic pacing for suppressing ictal events. For the particular animal model we are considering, the most effective fixed stimulation frequency was identified to be 1.0–2.0 Hz.[4,10] In this section, we evaluate the ability of our reinforcement learning framework to automatically acquire an adaptive strategy from the *in vitro* recordings. We analyze the behavior of the adaptive strategy in comparison with non-adaptive periodic stimulation strategies at low-frequencies as well as a control (no stimulation) strategy.

We first report on results characterizing the performance of the learning algorithm used to acquire the adaptive strategy. All error bars correspond to 1 standard error. In the case of the control and periodic strategies, this is due to variance between the four slices in the dataset. In the case of the adaptive strategy, the standard error includes both slice-to-slice variance and variance in the randomized tree regression algorithm.

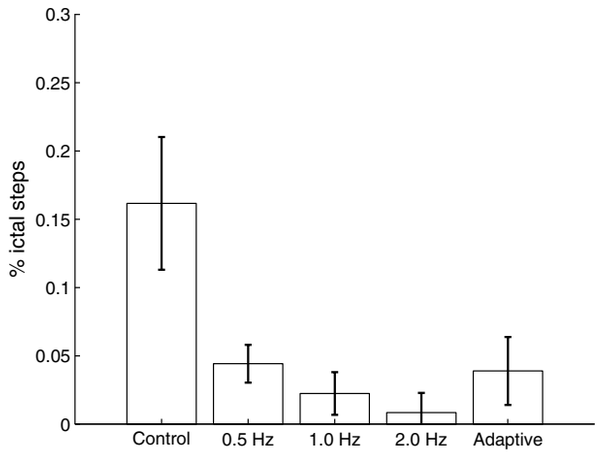


Fig. 3. Proportion of seizure steps (compared to non-seizure) under the following strategies: Control (no stimulation), Periodic pacing at 0.5 Hz, 1.0 Hz, 2.0 Hz, and Adaptive stimulation. The proportion of seizure/non-seizure for the Adaptive stimulation is estimated from Eq. (11). Proportions of seizure/non-seizure for the other strategies is calculated through hand-annotations of the EEG trace by an expert.

Figure 3 compares the proportion of states in which epileptiform behavior is observed under each of the policies. This corresponds to the score calculated in Eq. (11). We first note that under control conditions, slices in the dataset exhibit a larger rate of ictal events than under any of the stimulation strategies. Next we observe that periodic pacing at either 1 Hz or 2 Hz achieves near-complete suppression, and that performance is slightly less effective when stimulating at 0.5 Hz. Finally, we note that the adaptive strategy is able to achieve similar performance as the 0.5 Hz strategy in terms of seizure suppression.

Figure 4 shows the estimated long-term return for each of the strategies considered. This corresponds to the score calculated in Eq. (11), which is an empirical approximation of Eq. (1). The results here show a better return for the adaptive policy, compared to the periodic stimulation and control cases. Given that all strategies (except Control) achieve similar suppression efficacy, it seems reasonable to conclude that this return gain is primarily achieved through a reduction of the stimulation in the adaptive strategy (compared to the periodic strategies).

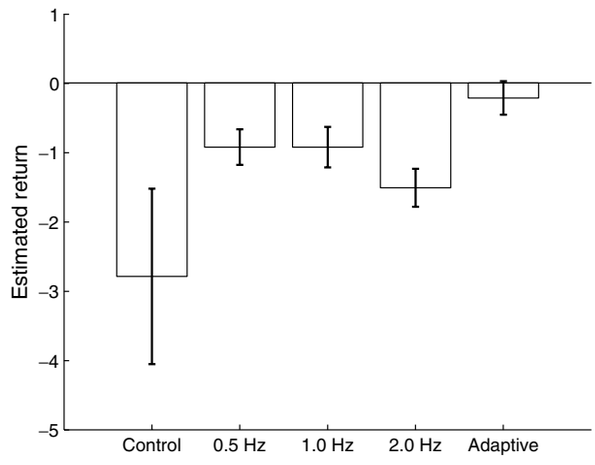


Fig. 4. Estimated long-term return under the following strategies: Control (no stimulation), Periodic pacing at 0.5 Hz, 1.0 Hz, 2.0 Hz, and Adaptive stimulation.

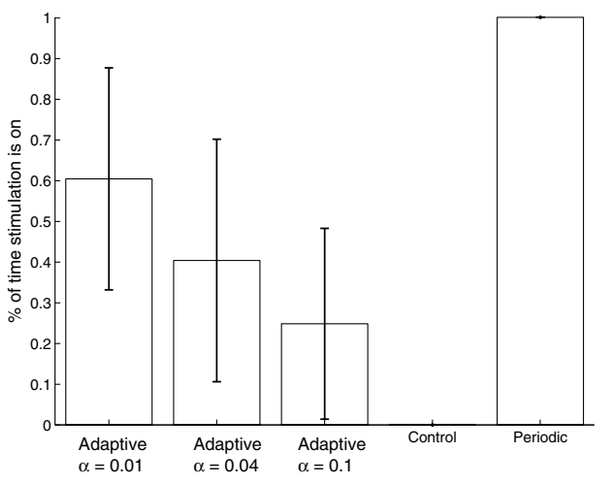


Fig. 5. Proportion of time under stimulation. All periodic strategies assume stimulation is on continuously. The proportion for the adaptive strategies is evaluated for different reward parameters.

Figure 5 supports this by showing the proportion of time during which stimulation is turned on under each of the conditions. We also show how this proportion changes as we re-train the adaptive strategy for different values of the parameter penalizing each stimulation action ($\alpha$ in Eq. (9)). As expected, when the penalty for stimulating is increased, the amount of stimulation is automatically reduced. There is substantial variation here between the different slices; in some slices some amount of stimulation would be necessary throughout most of the life of the slice to achieve reasonable suppression; in other slices it is

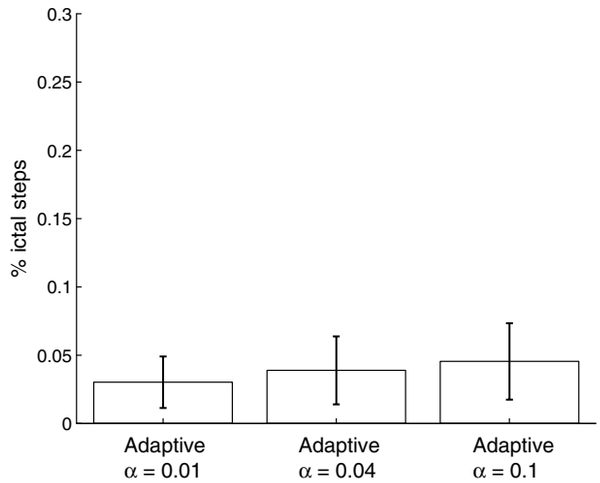


Fig. 6. Proportion of seizure steps as a function of the stimulation penalty. The result for $\alpha = 0.04$ is the same as shown in Fig. 3.

possible to turn off any stimulation for prolonged periods of time.

Lastly, it is worth considering how changes in the reward function impact the suppression efficacy. As shown in Fig. 6, the effect seems to be quite minimal.

We conclude our empirical evaluation by looking at some sample traces illustrating the behavior of the adaptive stimulation strategy in real-time. In this case, a new hippocampus-EC slice was prepared as described in the Methods section. The slice was subject to a stimulation protocol consisting of four phases. First, we applied a period of recording with no stimulation (control). Then, stimulation was applied at 1.0 Hz for at least 3 times the mean observed interval of occurrence of ictal discharges. The slice was then allowed to recover for several minutes until epileptiform activity returned to baseline. Finally we applied the same adaptive stimulation protocol as evaluated throughout this section (with $\alpha$=0.04). All other parameters were fixed as described in Secs. 2–4.

Figure 7 shows a typical excerpt from each of the recording conditions (control, 1.0 Hz stimulation, and two instances of adaptive stimulation, all taken from the same slice). The four phases were time-aligned to offer a better comparison. In Fig. 7(a) we see an ictal event typical of this *in vitro* model. Under control (no stimulation) conditions, such events usually appear every 150-200 seconds. As expected, the event is preceded by a few inter-ictal spikes.

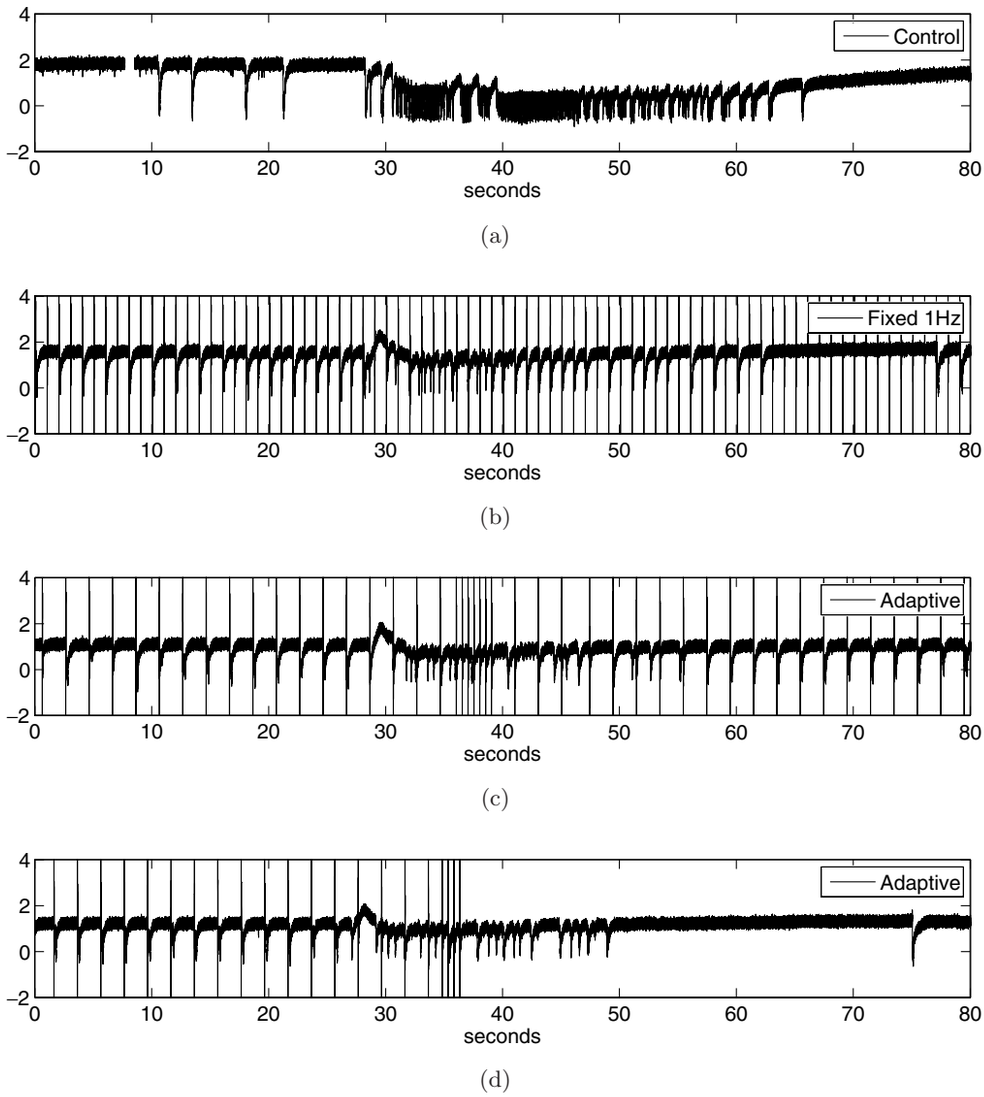


Fig. 7. Sample data traces comparing (a) epileptiform behavior under control conditions, (b) epileptiform behavior under periodic pacing conditions, (c) epileptiform behavior under adaptive stimulation (Example 1), and (d) epileptiform behavior under adaptive stimulation (Example 2). The four phases were time-aligned to offer a better comparison.

The post-ictal period is also quite characteristic of this acute *in vitro* model. In Fig. 7(b) we see typical behavior under 1.0 Hz stimulation. In this case, while there appears to be an ictal onset, it is of short duration and does not lead to a full ictal event. In Fig. 7(c) we see the effects of the adaptive strategy. First, we note that through much of the recording, the adaptive strategy maintains a slow pace of stimulation (roughly 0.5 Hz), which it interleaves with faster stimulation (roughly 2.0 Hz) following an ictal onset. The adaptive strategy is able to suppress the ictal event. It is possible this event would have been suppressed with similar effectiveness using only periodic (0.5 Hz) stimulation. Given the high degree of effectiveness of the periodic strategies on this particular model (as shown in Fig. 3), it would be surprising to see an adaptive strategy do much better in terms of suppression of ictal events. The analysis in Fig. 5 rather suggests that most of the gains to be made in this particular *in vitro* acute model of epilepsy are in terms of reducing the amount of stimulation applied. The last trace, shown in Fig. 7(d),

gives evidence in support of this. Here we see the same low-frequency (0.5 Hz) pacing being applied through an initial 35 seconds, followed by a period of faster stimulation (2.0 Hz) in reaction to an ictal onset. Once the seizure is successfully suppressed, the adaptive strategy chooses not to apply any stimulation for a prolonged period. It is not yet known what are the key characteristics of the signal that caused the difference in behavior between the two adaptive traces; this needs to be further investigated.

Considering Fig. 7 again, it seems that the primary benefit of the adaptive strategy in this particular animal model is to reduce the overall number of pulses (and not so much improve seizure suppression, which is already achieved through period pulsing.) This raises the following question: if the objective is really is to reduce the number of pulses, couldn't one use a simple feedback system to trigger stimulation upon seizure detection? Clearly, such a comparison would be very interesting. In the absence of such data, we remain skeptical that a detect-then-stimulate approach would perform as well as the reinforcement learning method, in terms of achieving an optimal balance between seizure suppression and low number of pulses. For this *in vitro* model in particular, current results suggest that delivering pulses between seizures has an important effect on suppression effectiveness, which would not necessarily be achieved with a detect-then-stimulate approach. It remains an interesting research question to verify this experimentally.

## 6. Discussion

The main contribution of this paper is to propose a new methodology for automatically learning adaptive neurostimulation strategies for the treatment of epilepsy. We have demonstrated that an adaptive stimulation policy can be learned through pre-recorded data of low-frequency single-pulse fixed stimulation, using a reinforcement learning methodology. Analysis of the learned adaptive strategy using pre-recorded data indicates a substantial reduction in the total amount of stimulation applied, compared to fixed stimulation strategies. Our analysis also indicates that the expected incidence of seizure under the adaptive policy is similar to that under periodic pacing strategies. It is worth emphasizing that suppression efficacy in this *in vitro* model is very high; in cases where suppression is not as effective, it may be possible for the adaptive strategy to outperform the periodic strategies in this respect. We have reported such results when using *in vitro* stimulation in the amygdala (with microelectrode recording in the perirhinal cortex).[17]

The results presented above suggest that reinforcement learning is a promising methodology for learning adaptive stimulation strategies online. One of the key advantages of this methodology is its ability to trade-off between minimizing incidence (and/or duration) of seizures, and the quantity of stimulation delivered.

Most of the evaluation presented on this paper is based on pre-recorded epileptiform behavior. Thus it is too early to draw conclusions regarding effectiveness of deploying this method in real-time. Evidence from the few experiments we were able to conduct in real-time show good correspondence between the policy's performance on pre-recorded data, and in the online setting. The results also show that the adaptive strategy does exploit information about the signal to determine when to increase (or turn off) stimulation. A full characterization of the adaptive strategy, in terms of understanding when and why it selects actions, is worthy of further investigation; this may shed some light into developing better seizure prediction mechanisms.

The methodology we present is not limited to the particular stimulation protocol we investigated. The results presented in this paper were obtained using low-frequency single-pulse patterns delivered to the subiculum. In previous work, we performed a similar analysis using stimulation of the amygdala.[17] The algorithm outlined in Sec. 4 could be directly applied to learn an adaptive stimulation strategy for a variety of other cases, including:

- other animal models (e.g. high potassium,[39] low calcium,[2] low magnesium[27]),
- different placement of the stimulating electrode (e.g. CA1, EC-subiculum[10]),
- various patterns of stimulation (e.g. high-frequency electric fields[6]).

In those cases, the Fitted Q-learning algorithm would be the same as described above, however the action set (and possibly the state set also) would have to be changed to reflect the new model.

We are now planning a series of experiments, whereby the adaptive stimulation strategy learned

using the batch data will be evaluated online, using live *in vitro* slices which match the conditions under which the data used so far has been recorded. Performing such experiments is very time-consuming and expensive. This highlights the value of developing good computational models of dynamical diseases. Such models exist for some diseases, such as HIV/AIDS and cancer. However to date there are few good generative models of temporal-lobe epilepsy, and many of the existing state-of-the-art models, e.g.,[29] do not include spontaneous transition into, and out of, seizures, nor do they include mechanisms for applying electrical stimulation. Other recent models[40,8] seem to provide more flexibility for investigating control of epileptic seizures and will be the subject of future empirical studies.

A final important question is whether the methodology outlined in this paper will carry over to *in vivo* models of epilepsy. From a technical perspective, we do not anticipate any major technical obstacles. The reinforcement learning framework is well suited to handling larger state representations, as would be necessary in cases where there are multiple sensing electrodes, placed at different (possibly unknown) locations. The framework is also able to deal with a larger set of possible stimulation parameters (intensity, duration, higher frequencies). However we do foresee two major practical challenges. First, it may be necessary to collect larger amounts of data to accurately learn the Q-function. Second, it is imperative to ensure that the action strategy used during the data collection (i.e. before the learning) is "safe." Neither of these issues arises when working with *in silico* or even *in vitro* models of epilepsy, but they are of definite concern when dealing with *in vivo* subjects. It is worth noting that there are substantial ongoing efforts in the computer science community to address precisely those problems, namely in developing algorithms that can efficiently learn from very small data sets, and in providing formal guarantees regarding the safety (or worst-case performance) of the system during the data collection process. We hope to leverage such results as they become available.

## Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their thoughtful comments, and help in improving this manuscript. The authors also gratefully acknowledge financial support by the Natural Sciences and Engineering Council Canada, as well as the Canadian Institutes of Health Research.

## References

1. H. Adeli, Z. Zhou and N. Dadmehr, Analysis of EEG records in an epileptic patient using wavelet transform, *Journal of Neuroscience Methods* **123**(1) (2003) 69–87.
2. N. Agopyan and M. Avoli, Synaptic and non-synaptic mechanisms underlying low calcium bursts in the in vitro hippocampal slice, *Exp. Brain Res.* **73**(3) (1988) 533–40.
3. M. Avoli, M. D'Antuono, J. Louvel, R. Kohling, G. Biagini, R. Pumain, G. DArcangelo and V. Tancredi, Network and pharmacological mechanisms leading to epileptiform synchronization in the limbic system *in vitro*, *Prog. Neurobiol.* **68**(3) (2002) 167–207.
4. M. Barbarosie and M. Avoli, Ca3-driven hippocampal-entorhinal loop controls rather than sustains in vitro limbic seizures, *J. Neurosci.* **17**(23) (1997) 9308–14.
5. R. Bellman, A markovian decision process, *Journal of Mathematics and Mechanics* **6** (1957).
6. M. Bikson, J. Lian, P. J. Hahn, W. C. Stacey, C. Sciortino and D. M. Durand, Suppression of epileptiform activity by high frequency sinusoidal fields in rat hippocampal slices, *Journal of Physiol.* **531**(1) (2001) 181–191.
7. P. Boon, K. Vonck, V. De Herdt, A. Van Dycke, M. Goethals, L. Goossens, M. Van Zandijcke, T. De Smedt, I. Dewaele, R. Achten, W. Wadman, F. Dewaele, J. Caemaert and D. Van Roost, Deep brain stimulation in patients with refractory temporal lobe epilepsy, *Epilepsia* **48**(8) (2007) 1551–1560.
8. N. Chakravarthy, S. Sabesan, L. Iasemidis and K. Tsakalis, Controlling synchronization in a neuron-level population model, *International Journal of Neural Systems* **17**(2) (2007) 123–138.
9. M. D'Antuono, G. Biagini and M. Avoli, Unpublished data.
10. G. D'Arcangelo, G. Panuccio, B. Tancredi and M. Avoli, Repetitive low-frequency stimulation reduces epileptiform synchronization in limbic neuronal networks, *Neurobiology of Disease* **19**(1–2) (2005) 119–128.
11. D. Durand and M. Bikson, Suppression and control of epileptiform activity by electrical stimulation: A review, *Proceedings of the IEEE* **89**(7) (2001) 1065–1082.
12. D. Ernst, P. Geurts and L. Wehenkel, Tree-based batch mode reinforcement learning, *Journal of Machine Learning Research* **6** (2005) 503–556.

13. D. Ernst, G.-B. Stan, J. Gonçalves and L. Wehenkel, Clinical data based optimal STI strategies for HIV: A reinforcement learning approach, in *15th Machine Learning Conference of Belgium and The Netherlands*, Ghent, Belgium (May 2006).
14. P. Geurts, D. Ernst and L. Wehenkel, Extremely randomized trees, *Machine Learning* **63**(1) (2006) 3–42.
15. S. Ghosh-Dastidar, H. Adeli and N. Dadmehr, Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection, *IEEE Transactions on Biomedical Engineering* **54**(9) (2007) 1545–1551.
16. B. J. Gluckman, H. Nguyen, S. L. Weinstein and S. J. Schiff, Adaptive electric field control of epileptic seizures, *Journal of Neuroscience* **21**(2) (2001) 590–600.
17. A. Guez, R. Vincent, M. Avoli and J. Pineau, Adaptive treatment of epilepsy via batch-mode reinforcement learning, in *Innovative Applications of Artificial Intelligence* (2008).
18. A. Handforth, C. M. DeGiorgio, S. C. Schachter *et al.*, Vagus nerve stimulation therapy for partial-onset seizures. a randomized active-control trial, *Neurology* **51** (1998) 48–55.
19. M. Hauskrecht and H. Fraser, Planning treatment of ischemic heart disease with partially observable markov decision processes, *Artificial Intelligence in Medicine* **18** (2000) 221–244.
20. L. Iasemidis, Epileptic seizure prediction and control, *IEEE Trans Biomed Eng.* **50**(5) (2003) 549–558.
21. K. Jerger and S.J. Schiff, Periodic pacing an in vitro epileptic focus, *J of Neurophysiol* **73** (1995) 876–879.
22. L. P. Kaelbling, M. L. Littman and A. W. Moore, Reinforcement learning: A survey, *Journal of Artificial Intelligence Research* **4** (1996) 237–285.
23. H. Khosravani, P. L. Carlen and J. L. Perez Velazquez, The control of seizure-like activity in the rat hippocampal slice, *Biophysical Journal* **84** (2003) 687–695.
24. E. H. Kossoff, E. A. Ritzl, J. M. Politsky, A. M. Murro, J. R. Smith, R. B. Duckrow, D. D. Spencer and G. K. Bergey, Effect of an external responsive neurostimulator on seizures and electrographic discharges during subdural electrode monitoring, *Epilepsia* **45**(12) (2004) 1560–1567.
25. M. Le Van Quyen, J. Martinerie, V. Navarro, P. Boon, M. D'Have, C. Adam, B. Renault, F. Varela and M. Baulac, Anticipation of epileptic seizures from standard EEG recordings, *The Lancet* **3567** (2001).
26. B. Litt and J. Echauz, Prediction of epileptic seizures, *The Lancet Neurology* **1**(1) (2002) l22–30.
27. I. Mody, J.D. Lambert and U. Heinemann, Low extracellular magnesium induces epileptiform activity and spreading depression in rat hippocampal slices, *J. Neurophysiol.* **57**(3) (1987) 869–88.
28. F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger and K. Lehnertz, On the predictability of epileptic seizures, *Clinical Neuropsychology* **116** (2005) 569–587.
29. T. I. Netoff, R. Clewley, Arno. S., T. Keck and J. A. White, Epilepsy in small world networks, *Journal of Neuroscience* **24**(37) (2004) 8075–8083.
30. I. Osorio, J. Overman, J. Giftakis and S. B. Wilkinson, High frequency thalamic stimulation for inoperable mesial temporal epilepsy, *Epilepsia* **48**(8) (2007) 1561–1571.
31. H. Osterhage, F. Mormann, T. Wagner and K. Lehnertz, Measuring the directionality of coupling: Phase versus state space dynamics and application to EEG time series, *International Journal of Neural Systems* **17**(3) (2007) 139–148.
32. J. Pineau, M.G. Bellemare, A. J. Rush, A. Ghizaru and S.A. Murphy, Constructing evidence-based treatment strategies using methods from computer science, *Drug and Alcohol Dependence* **88S** (2007) S52–S60.
33. M. L. Puterman, *Markov Decision Processes* (Wiley, 1994).
34. H. Qu and J. Gotman, A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: Possible use as a warning device, *IEEE Trans Biomed Engineering* **44** (1997) 115–112.
35. G. Regesta and P. Tanganelli, Clinical aspects and biological bases of drug-resistant epilepsies, *Epilepsy Res* **34** (1999) 109–122.
36. Ghosh-Dastidar S. and H. Adeli, A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection, *Neural Networks* **22** (2009).
37. Sutton R. S. and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
38. S.J. Schiff, K. Jerger, D.H. Duong *et al.*, Controlling chaos in the brain, *Nature* **370** (1994) 615–620.
39. S.F. Traynelis and R. Dingledine, Potassium-induced spontaneous electrographic seizures in the rap hippocampal slice, *J. Neurophysiol.* **59** (1988) 259–276.
40. K. Tsakalis, N. Chakravarthy and L. Iasemidis, Control of epileptic seizures: Models of chaotic oscillator networks, in *44th IEEE Conference on Decision and Control*, Vols. 12–15 (2005), pp. 2975–2981.
41. K. Tsakalis and L. Iasemidis, Control aspects of a theoretical model for epileptic seizures, *International Journal of Bifurcation and Chaos* **16**(7) (2006) 2013–2027.
42. B. M. Uthman, A. M. Reichl, J. C. Dean, S. Eisenschenk, R. Gilmore, S. A. Reid, S. N. Roper and B. J. Wilder, Effectiveness of vagus nerve stimulation in epilepsy patients: A 12 year observation, *Neurology* **63** (2004) 1124–1126.
43. A. L. Velasco, F. Velasco, M. Velasco, D. Trejo, G. Castro and J. D. Carrillo-Ruiz, Electrical stimulation of the hippocampal epileptic FOCI for seizure

control: A double-blind, long-term follow-up study, *Epilepsia* **48**(10) (2007) 1895–1903.
44. J. Vesper, B. Steinhoff, S. Rona, C. Wille, S. Bilic, G. Nikkhah and C. Ostertag, Chronic high-frequency deep brain stimulation of the STN/SNR for progressive myoclonic epilepsy, *Epilepsia* **48**(8) (2007) 1984–1989.
45. K. Vonck, P. Boon, E. Achten, J. De Reuck and J. Caemaert, Long-term amygdalohippocampal stimulation for refractory temporal lobe epilepsy, *Ann Neurol* **52**(5) (2002) 556–565.
46. R.J. Warren and E. Durand, Effects of applied currents on spontaneous epileptiform activity induced by low calcium in the rat hippocampus, *Brain Res* **806** (1998) 186–195.
47. D. Zumsteg, A. M. Lozano, H. G. Wieser and R. A. Wennberg, Cortical activation with deep brain stimulation of the anterior thalamus for epilepsy, *Clinical Neurophysiology* **117** (2006) 192–207.