International Journal of Neural Systems, Vol. 27, No. 0 (2017) 1750012 (13 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0129065717500125

Seizure Control in a Computational Model Using a Reinforcement Learning Stimulation Paradigm

Vivek Nagaraj
Graduate Program in Neuroscience
University of Minnesota – Twin Cities
312 Church St SE, Minneapolis
MN 55455, USA
nagar030@umn.edu

Andrew Lamperski
Department of Electrical and Computer Engineering
University of Minnesota – Twin Cities
200 Union Street SE
Minneapolis, MN 55455, USA
alampers@umn.edu

Theoden I Netoff*
Department of Biomedical Engineering
University of Minnesota – Twin Cities
312 Church St SE , Minneapolis
MN 55455, USA
tnetoff@umn.edu

Accepted 26 October 2016
Published Online 29 December 2016

Neuromodulation technologies such as vagus nerve stimulation and deep brain stimulation, have shown some efficacy in controlling seizures in medically intractable patients. However, inherent patient-to-patient variability of seizure disorders leads to a wide range of therapeutic efficacy. A patient specific approach to determining stimulation parameters may lead to increased therapeutic efficacy while minimizing stimulation energy and side effects. This paper presents a reinforcement learning algorithm that optimizes stimulation frequency for controlling seizures with minimum stimulation energy. We apply our method to a computational model called the epileptor. The epileptor model simulates inter-ictal and ictal local field potential data. In order to apply reinforcement learning to the Epileptor, we introduce a specialized reward function and state-space discretization. With the reward function and discretization fixed, we test the effectiveness of the temporal difference reinforcement learning algorithm (TD(0)). For periodic pulsatile stimulation, we derive a relation that describes, for any stimulation frequency, the minimal pulse amplitude required to suppress seizures. The TD(0) algorithm is able to identify parameters that control seizures quickly. Additionally, our results show that the TD(0) algorithm refines the stimulation frequency to minimize stimulation energy thereby converging to optimal parameters reliably. An advantage of the TD(0) algorithm is that it is adaptive so that the parameters necessary to control the seizures can change over time. We show that the algorithm can converge on the optimal solution in simulation with slow and fast inter-seizure intervals.

Keywords: Epilepsy; neuromodulation; deep brain stimulation; reinforcement learning; closed-loop.

*Corresponding author.

1. Introduction

Epilepsy afflicts approximately 1% of the world's population, and close to 1/3 of these patients do not respond to anti-epileptic drugs or experience debilitating side-effects.[1] Of these patients, approximately two-thirds qualify for surgical resection of the seizure focus.[2] For the remainder of patients, neuromodulation devices may provide some benefit.

Patients unable to get surgical resection often resort to neuromodulation technologies such as vagus nerve stimulation (VNS) or deep brain stimulation (DBS). Existing VNS and DBS devices showed therapeutic efficacy at approximately 40% seizure reduction rates during the clinical trial phase.[3–5] Longitudinal data from the NeuroPace RNS® study indicated an increase in responder rates over time.

The NeuroPace RNS® system applies electrical stimulation to a specific brain region when certain salient features are detected in the neural signals.[6] The methodology is computationally simple and has shown to have some efficacy, however, the large number of device parameters must be manually tuned by the physician.[6,7] Limitations of current DBS methods include a large parameter space and highly variable patient response over time. While current state-of-the-art technologies are effective in controlling seizures for some patients, there is still a need for improvement.[8]

To address these limitations, many groups have made advances in closed-loop neuromodulation. Closed-loop feedback for seizure control outcomes has been tested using three different methodologies. First, the neural signal is used in determining when to apply a stimulus. Stimulus administration in response to seizure onset has successfully suppressed seizures using electrical stimulation.[9–12] and optogenetic inhibition.[13] Secondly, feedback can be used for stimulus optimization over trials in order to maximize the therapeutic effect. Machine-learning algorithms have been used to optimize stimulus parameters by measuring the efficacy of different stimuli and their effects on seizure duration and frequency.[14] Lastly, closed-loop feedback can lead to the modulation of a stimulus parameter in real-time based on physiological measures. Seizures have been suppressed by modulating DC fields proportionally to the activity measured.[15] and by applying precisely timed transcranial electrical current stimulation at certain phases of spike-and-wave ictal behavior.[16]

We propose that a reinforcement learning algorithm that is constantly vigilant and can detect subtle changes in therapeutic efficacy may result in better patient outcomes. Furthermore, a reinforcement learning algorithm may be able to detect changes in the patient's needs and adapt accordingly. In this paper, we will present a novel approach to controlling seizures, based on the temporal difference reinforcement learning algorithm in a computational model of epilepsy called epileptor. The algorithm, through an iterative process, determines a policy so as to maximize reward while minimizing a cost function[17] (Fig. 1). In this instance, the policy is a mapping between different physiological states in an epileptic system and specific stimulus parameter combinations. The reward signal is inversely proportional to an epileptic biomarker, therefore greater reward means less epileptiform activity. The cost function tries to minimize the total amount of stimulation energy for any state-action pair. We define energy as the product of stimulation amplitude squared and the total duration of stimulation in a simulation.

We use a reduced temporal difference reinforcement learning model, TD(0), that is able to learn the optimal stimulus parameters by exploring the possible stimulation parameters and observing the results. We will show that the learning algorithm converges to a low energy solution.

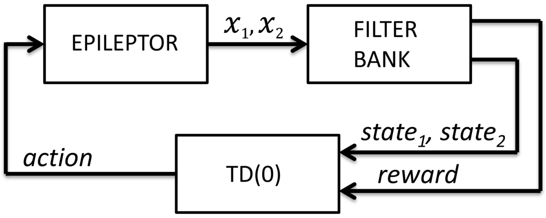


Fig. 1. *A schematic representation of the seizure control paradigm with the computational model of seizures.* LFPs generated by the Epileptor model $(-x_2 + x_1)$ are filtered (high and low pass filter) to estimate the current state of the model. A reward signal is also generated from the LFP such that it decreases with seizures and increases with time following a seizure. The reinforcement learning algorithm (TD(0)) integrates state and reward information to determine the optimal action (stimulation frequency) that will maximize the reward provided the current state. The action selected is the stimulation frequency to apply to the epileptor model.

2. Materials and Methods

2.1. *Epileptor model*

To the best of our knowledge, the epileptor is the only computational model that captures the transition from nonseizing (inter-ictal) to seizing (ictal) local field potential (LFP) activity. We tested various stimulation paradigms on the Epileptor mean-field model.[18,19] The epileptor model reproduces many of the invariant seizure characteristics documented across species including fast oscillations, spike and wave events and logarithmic increase in inter-spike interval as the seizure approaches termination.[19]

This model provides a platform in which to test closed-loop adaptive algorithms. An example of the transition between inter-ictal to ictal can be seen in Fig. 2 (top) and is confirmed by a clear increase in power between 1 and 25 Hz, Fig. 2 (bottom). For our simulations, we use a bandpass filtered signal between 2–15 Hz to calculate the reward which captures the fast oscillations in the LFP.

The parameters in the model were empirically fit to generate the different behaviors seen in physiological recordings. The full equations are included in Sec. A.1.

The derivation of the equations for the epileptor model is described in two papers from the Jirsa lab.[18,19] The model is represented as a system of five ordinary differential equations with three different time scales. The fastest time scales (x_1, y_1) generate fast oscillations, the medium time scales (x_2, y_2) generate spike and wave events, and the long time scale slow permittivity variable (z) determines the seizure duration and frequency. The LFP is determined by adding $-x_1$ and x_2. They also developed a reduced two-dimensional model[18] to characterize seizure onset and offset behaviors using only the fast x_1 and slow z parameters, as described in Sec. A.2.

The reduced model does not produce the high frequency burst activity seen in the full model but it does reproduce the duration of the seizure and intervals between the seizures. The limit cycle in Fig. 3 (top) captures the reduced system cycling through inter-ictal (left) and ictal (right) phases. We used the reduced model to visualize how stimulation affects the trajectory of two state variables responsible for ictogenesis. The goal of the state-space analysis is to

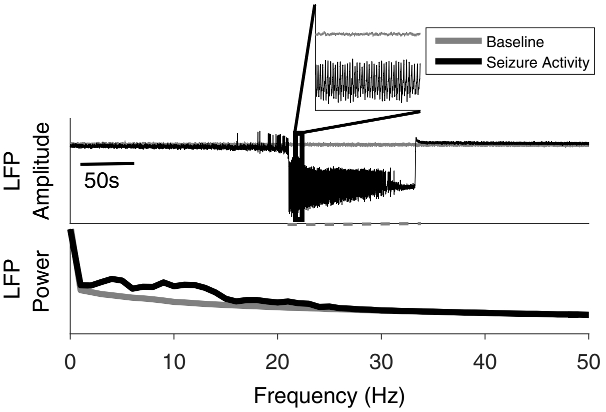


Fig. 2. Example of the LFP generated by the epileptor model. Top, sample trace of the epileptor model during a nonseizing state (gray) and seizing state (black). Seizures are clearly indicated by prominent DC shift in the LFP. Duration of the seizure, from onset to offset, is indicated by the grey dotted line. Bottom, power spectral density estimate of nonseizing versus seizure activity. The y-axis shows power at different frequency bands on a log scale.

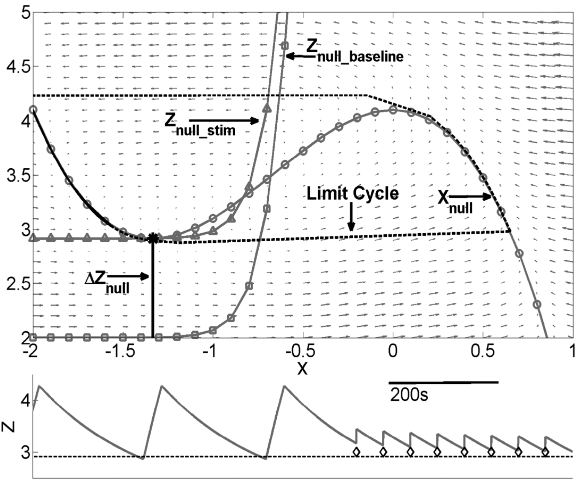


Fig. 3. Prediction of minimum energy stimulus parameters to suppress seizures using a state-space analysis of the epileptor model. Top, vector field of the reduced epileptor model with x-nullcline (gray-circles) and baseline z-nullcline (gray-square) and z-nullcline with stimulation (gray-triangle). The black solid line between the x- and z-nullclines (with stimulation) indicates the distance the baseline z-nullcline needs to move in order to generate a stable fixed point (large black asterisk) which leads to seizure suppression. The limit cycle is represented by the trapezoidal shaped dotted black line. The bifurcation, from inter-ictal to ictal, occurs when the z value crosses ≈ 2.91. Bottom, example trace of the z variable before and during periodic stimulation (diamonds). The y-axis values correspond to the changing z values as the system moves along the limit cycle in the top figure. The z variable does not descend below the seizure threshold (black dashed line) during stimulation.

determine how much stimulation results in a bifurcation that terminates seizures. Both models were integrated with a Euler–Maruyama integrator with a time step of 1 ms. Longer time steps lead to numerical errors, while smaller time steps did not improve resolution of salient features of the Epileptor.

To simulate electrical stimulation, we applied pulses of 1 ms width to the z-variable. Periodic stimulation of the z-variable using sufficient energy leads to seizures control (bottom, Fig. 3).

Changing the z time constant τ_0 modulates the duration of the seizure and the inter-seizure intervals (ISIs). For a given τ_0, the model will produce seizures that occur at almost regular intervals.

2.2. *Determining minimum stimulation frequency to suppress seizures*

In order to determine the minimum stimulus frequency while holding stimulation pulse amplitude to achieve seizure suppression in the epileptor model, we conducted a state-space analysis. The dynamics of the seizure duration and interval can be understood by analyzing the x- and z-nullclines of the reduced model. With no stimulation input, the x- and the z-nullclines have a single unstable equilibrium point at the crossing, allowing for a stable limit cycle that determines the seizure duration and ISIs. Positive stimulation current pushes the z-nullcline up. With sufficient current, the x- and the z-nullclines go through a bifurcation resulting in a stable fixed point and seizures stop. The minimum pulsatile amplitude necessary to stop the seizures can be determined by the distance of the z-nullcline to the knee of the x-nullcline (ΔZ_{null}), as shown in top of Fig. 3.

The z-variable relaxes slowly back to its resting point after each pulse according to the z time constant τ_0, as shown in the bottom of Fig. 3. Therefore, to assure seizure control throughout the simulation, the minimum z value between pulses must be greater than $\Delta Z_{\mathrm{null}} + x_0$. This will keep the z variable above the bifurcation point Fig. 3 (bottom).

Provided the stimulus amplitude (S_a), stimulus interval (S_{IPI}), and the time constant of $z(\tau_0)$ it is possible to calculate the approximate minimum distance z needs to travel to control seizures:

$$z_{\min} \approx x_0 + \frac{S_a}{1 - e^{(-\frac{S_{\mathrm{IPI}}}{\tau_0})}}. \tag{1}$$

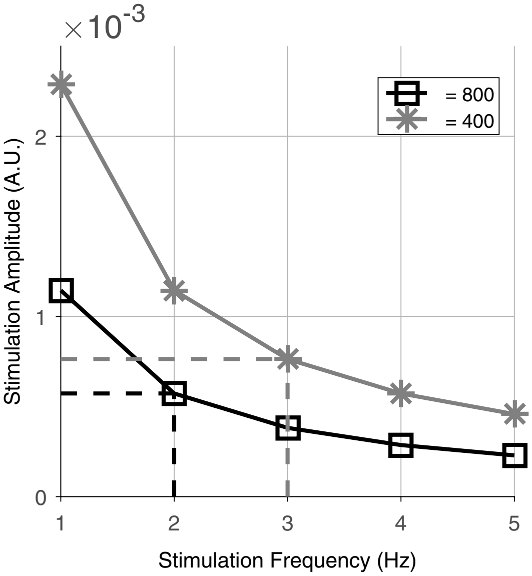


Fig. 4. Map indicating minimum stimulus frequency and amplitude to control seizures given the time constant for the z variable (legend). Stimulus parameter combinations below the solid lines will fail to suppress seizures.

This $z_{\min}$ value is the minimum distance required to suppress seizures and is $z_{\min} = 2.91$. The parameter x_0 determines at what value the floor of the z-nullcline is at (in Fig. 3 it is at 2.0). If x_0 is set at ≈ 2.91 then there will be no seizure events. This is because a stable fixed point forms at the intersection of the z and x_1 nullclines. For all simulations the x_0 variable was set at 2.0. This, however, is the minimum pulsatile perturbation to suppress seizures, not all epileptiform activity. Increasing the stimulation above this can produce significant improvements by suppressing high frequency activity seen in the full model when the system is close to the bifurcation point.

Alternatively, we can calculate the stimulus amplitude for any arbitrary stimulus interval.

$$S_a = \Delta Z_{\mathrm{null}}(1 - e^{(-\frac{S_{\mathrm{IPI}}}{\tau_0})}), \tag{2}$$

where S_a is the minimum stimulus amplitude to suppress seizures given the stimulus intervals (S_{IPI}) and z variable time constant (τ_0). The simulations presented in the paper use integer value stimulus intervals. A map of analytically calculated minimum pulsatile stimulation amplitude necessary for seizure suppression given different integer stimulation frequencies is shown in Fig. 4. Given a particular stimulation amplitude-frequency pair, we ran simulations to compare the minimum frequency (while holding stimulation amplitude) to the analytically calculated stimulation frequency for the given amplitude.

2.3. *Temporal difference reinforcement learning algorithm implementation*

Reinforcement learning has a long and intricate history and has been reviewed in detail elsewhere.[17] These algorithms have been implemented in many different neural-engineering problems, including seizure control.[20,21] However, in these cases, the investigators develop a stimulation policy with a reinforcement learning algorithm using offline training data. Optimal parameters were determined through dynamic programming principles using the full data set, a process known as batch learning. Our approach substitutes batch learning for an iterative learning process where optimal stimulation parameters are determined as data is collected in real time. The decision to use an online learning scheme, as opposed to offline training, was motivated by the fact that the effects of stimulation for each experimental preparation (and patient) seem to be different depending on subtle details as the positioning of the stimulation electrode. Thus, optimizing parameters based on data collected from one patient or experiment and applied to another may not result in the best control.

The full temporal difference learning algorithm uses memory in the form of eligibility traces.[17] to identify sequences of actions necessary to achieve a high reward state. However, in simulations we found little benefit from using eligibility traces, presumably because the occurrence of seizures is stochastic, and therefore was not used. This way, only the current state-action pair is updated. In this study, we show how a learning algorithm can determine the optimal stimulation parameters for a seizure model using domain knowledge. Furthermore, the algorithm can respond to nonstationarities of an epileptic system and adapt accordingly.

2.3.1. *Reduced temporal difference learning algorithm*: *TD*(0)

The TD(0) algorithm generates a map relating the state of the system to the expected reward given a selected action: $Q(s, a)$, where Q is the expected reward, proportional to the time since the last seizure, a is the selected action from a set of stimulation frequencies, and s is the state of the epileptor (i.e. seizing or not seizing). When the simulation is initialized the algorithm is naïve, and has no information about the expected reward for each state-action pair. To initialize $Q(s, a)$, we use the average reward over a few seizures without stimulation, μ_0, plus additive zero mean white noise with small variance $\sigma^2 \ll 1$.

Once the algorithm is turned on, it chooses the next actiona' expected to produce the highest reward, $Q(s', a')$, given the current state s'. After the action is executed ($a' \to a$), the actual reward (R) is measured. The error between the measured reward and expected reward is calculated as follows:

$$\delta = R - Q(s, a) + \gamma * Q(s', a'), \qquad (3)$$

where γ is the delay discounting factor that accounts for how much the algorithm values future expected rewards. We set to γ to zero ($\gamma = 0$) to favor more greedy behavior to speed up convergence. Initially, because the matrix Q is arbitrarily set at a high value, the error, between the measured reward and the expected reward will be great. Using this error the map is updated to increase the accuracy of the predicted reward value with the following update equation:

$$Q(s, a) = Q(s, a) + \alpha * \delta, \qquad (4)$$

where $0 < \alpha \ll 1$ determines how quickly $Q(s, a)$ is updated given the error to the measured reward.

2.3.2. *Action selection*

If at every time step the action providing the greatest expected reward is selected, this algorithm is called "greedy". However, it may result in convergence to action selection policy at a local minimum rather than allowing for exploration to find a global minimum. Therefore, using a selection policy that explores some actions that may at first appear subminimum can help find the optimal solution more reliably than a purely greedy algorithm. An alternative to the purely greedy approach is the Softmax policy that selects actions with probability proportional to the expected reward. So, the action with the highest expected reward will have a higher probability of being selected than other actions. The probability of each action, a', selected at each state, s', can be calculated as:

$$P(a' \mid s') = \frac{e^{\frac{Q(s',a')}{\tau_s}}}{\sum_{i=1}^{n} e^{\frac{Q(s',a_{i'})}{\tau_s}}}. \tag{5}$$

The variable τ_s is the temperature value for the Softmax policy, which determines how sensitive the probability is to differences in expected reward, and n is the total number of actions.

2.3.3. *Learning rate*

The degree to which the TD(0) algorithm weights new information compared to its current estimate of the reward for a selected state-action pair is determined by the parameter α. When the model is completely naïve, using a large α results in updating the map rapidly. This method can lead to a quick convergence to an action that is effective, but the solution may not be the global minimum. We selected α to be long enough to average information over the duration of one inter seizure interval. The algorithm updates the action in windows ($w = 15\,\text{s}$). Given the duration of an ISI, we set α so that the time constant is about one seizure interval:

$$\alpha \approx 1 - e^{(-w/\text{ISI})} \tag{6}$$

A full description of the algorithm is given in Sec. 6.3.

2.4. *Ictal and inter-ictal state estimation from LFP data*

The epileptor model has no reward signal and has a continuous state-space, which is not amenable to a traditional reinforcement learning framework. Domain knowledge (i.e. spectral characteristics, Fig. 2) was used to establish the reward function and discretize the state-space. In the epileptor model, the ictal versus inter-ictal dynamics can be separated in state-space by plotting the x_1 variable against x_2 as shown in Fig. 5(B). The inter-ictal activity is in the left side of the state-space and the right side is during the ictal activity. The goal is to generate some state-space representation from the LFP that separates the inter-ictal from the ictal activity.

By filtering the data with a high pass filter and using another lowpass filter, the two signals resulted in a state-space representation very similar to that seen by plotting the x_1 variable against x_2 (Fig. 5(c)). Data was filtered using first-order Butterworth high pass (Fig. 5(A) top) and low pass filters (Fig. 5(A) bottom) each with 0.5 Hz cutoff frequency. Filter coefficients were calculated using the MATLAB *butter* command. Filter outputs were normalized to each axis in the two-dimensional state-space. This state-space is partitioned into ictal and inter-ictal states. A state index is assigned at each time step depending on where state value falls in the inter-ictal or the ictal partition. Finer resolution partitions are possible, but more partitions increased the exploration time.

2.5. *Reward*

The reward for each $Q(s, a)$ is calculated at each iteration by taking the negative log of the band-pass filtered LFP (LFP_{bp}, 2–15 Hz). The reward for each time step is calculated as follows:

$$\tau_{\text{filter}} = \left(\frac{1}{1 - e^{\left(\frac{-1}{\tau_{\text{ISI}}}\right)}}\right), \tag{7}$$

$$\text{L}\bar{\text{F}}\text{P}_t^{\text{bp}} = \frac{(\text{LFP}_t^{\text{bp}} * \text{LFP}_t^{\text{bp}})}{\tau_{\text{filter}}} + (\tau_{\text{filter}} * \text{L}\bar{\text{F}}\text{P}_{t-1}^{\text{bp}}), \tag{8}$$

$$R_t = -\log(\text{L}\bar{\text{F}}\text{P}_t^{\text{bp}} - \text{Cost}), \tag{9}$$

where τ_{filter} is the decay rate for the smoothing filter, $\text{L}\bar{\text{F}}\text{P}_t^{\text{bp}}$ is the smoothed power of the LFP at time t, and R_t is the reward at time t. The tau-filter smoothens the noisy reward signal to remove noise in the raw reward signal. Noise in the raw reward signal can lead to misleading updates of the state-action function $Q(s, a)$.

Each stimulation parameter combination has an associated cost. The cost function is as follows:

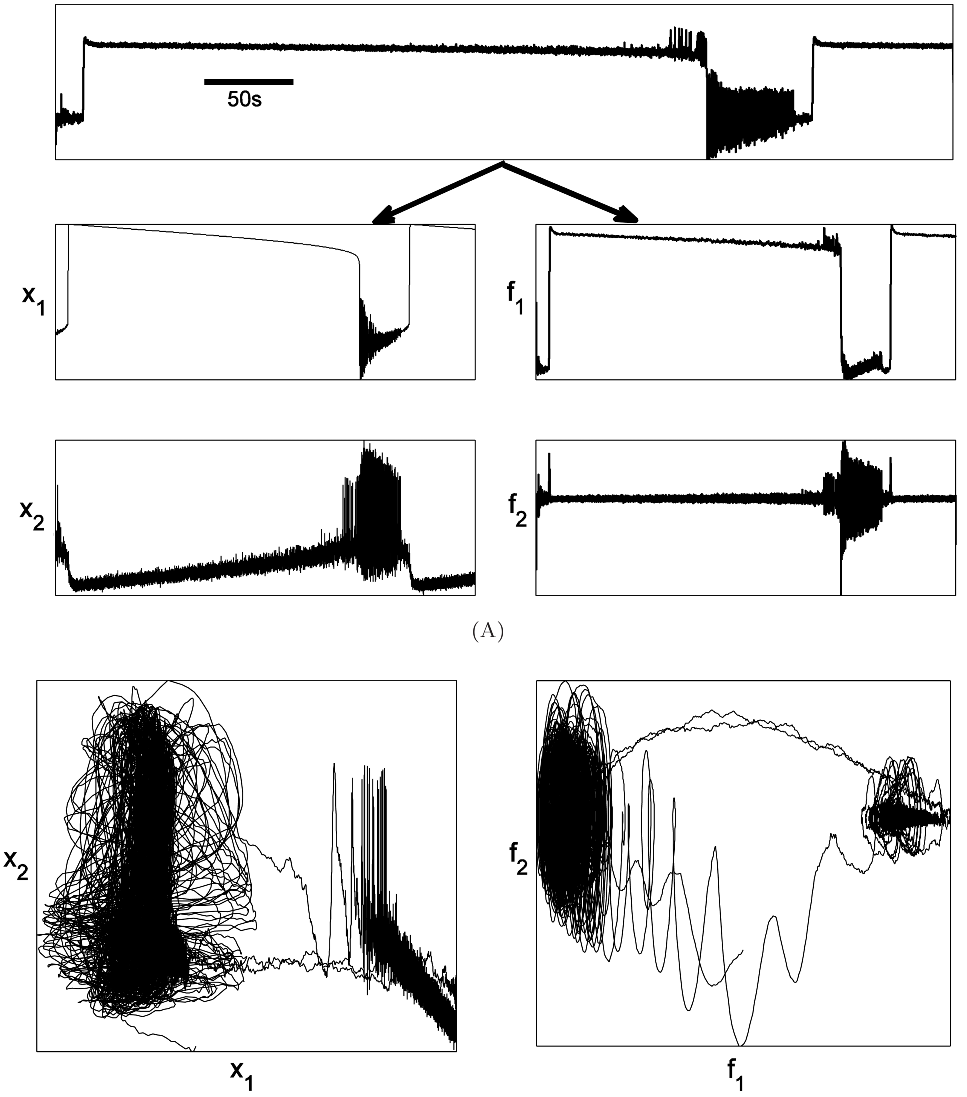
$$\text{Cost} = \left(\text{S}_a^2 * \left(\frac{1}{S_{\text{IPI}}}\right)\right) * \text{CW}, \tag{10}$$

where CW is the cost weight, S_{IPI} is the stimulation inter-pulse interval, S_a is the stimulation amplitude and μ_0 is the mean reward during the simulation when stimulation pulses were not applied.

3. Results

3.1. *Open-loop stimulation controls stationary seizures*

We first tested open-loop stimulation on seizure control in the epileptor model. Seizures were suppressed with sufficiently high stimulation frequency



(A)

(B)

(C)

Fig. 5. Decomposition of the epileptor LFP using a filter bank results in feature space closely resembling the actual state-space of the epileptor dynamic variable x_1 and x_2. (A) top, raw LFP generated by the epileptor model. (B) State-space of the epileptor model parameters x_1 and x_2. Top, the x_1 variable time series, middle, the x_2 variable time series, bottom, the state-space from x_1 and x_2 clearly distinguishes inter-ictal (right) from ictal (left) epochs. (C) Estimation of the state-space using a filter bank. Top, low-pass filtered LFP data, middle, high-pass filtered LFP data, bottom, feature space clearly shows an inter-ictal cluster (right) and ictal cluster (left). Cutoff frequency was 0.5 Hz.

and amplitude. The minimum stimulation energy required to suppress seizures depended on the time constant of the z-variable of the model; the faster the time constant, the more stimulation energy is required to suppress seizures. Examples of the epileptor model with two different time constants, $\tau_0 = 800$ and $\tau_0 = 400\,\mathrm{s}$ in response to stimulation are shown in Fig. 7. The analytically calculated stimulation (1.04 Hz) applied to the model with a time constant of 800 s is sufficient to suppress all seizure like activity (Fig. 7, top), while it is necessary to stimulate at 2.08 Hz to achieve control when the system has a time constant of 400 s (Fig. 7, middle). The high frequency activity is distinguishable from a seizure because the system did not undergo a saddle-node bifurcation at the seizure onset. Seizures in the epileptor model are characterized by a prominent DC offset, which is a key feature of saddle-node bifurcations.[19] Applying the minimum stimulus frequency to suppress seizures for $\tau_0 = 800$ in a simulation with $\tau_0 = 400$ seizure dynamics does not suppress the seizure (Fig. 7, bottom).

The map in Fig. 4 shows the necessary stimulation parameters to achieve seizure control for different time constants of the model as determined by analysis of the reduced model. Parameters used in Fig. 7 are indicated in the map shown in Fig. 4. While the parameters necessary to suppress seizures are calculated from the reduced model, it can be seen that high frequency activity may persist even though the seizures have been stopped. Increasing the stimulation frequency above this minimum threshold of seizure suppression may provide further improvement by reducing this "epileptiform" behavior. Therefore, we consider the optimal parameter combination for epileptogenic feature control to be slightly higher than the minimum necessary to suppress seizures.

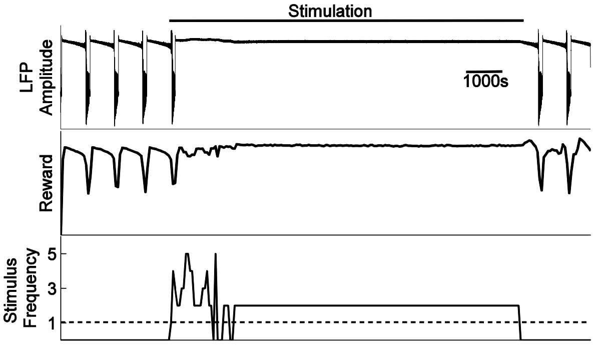


Fig. 6. TD(0) algorithm converges to optimal solution (2 Hz) when seizure dynamics are stationary. Top, raw LFP trace. Middle, reward trace associated with the LFP. Bottom, Stimulation frequency selected at each action selection interval. The TD(0) algorithm explores the state-action space at first because the initial expected values for each state-action is high. Dotted black line represents the minimum stimulus frequency required for control (1.04 Hz). Black bar indicated duration of stimulus.

3.2. *Closed-loop adaptive stimulation results in seizure control while minimizing energy*

We next tested whether the TD(0) algorithm can converge to an effective stimulus parameter solution. The goal of the TD(0) algorithm is to maximize reward through selection of actions, in this case, stimulation frequencies at each state (ictal versus inter-ictal). The reward trace increases as activity decreases, therefore suppressing seizures results in an increased reward. A small cost is attributed to each stimulation pulse which helps refine the stimulation frequency once a range of stimulation actions that suppress seizures are found. At the onset of the simulation the model is naïve and set to random initial conditions. We set the learning rate to reflect the time course of a seizure (Eq. (6)). This results in rapid learning with fairly stable behavior once good control is achieved.

An example of the TD(0) algorithm identifying optimal seizure control parameters is shown in Fig. 6. Simulations were run for 15,000 s on a Windows 7 with Intel (i7) 3.5 GHz processor. The TD(0) algorithm computation time was 160 ms. High frequency activity and seizures correspond to downward deflections of the reward signal. After probing the state-action space following stimulation onset (black bar) the policy converges to the optimal state-action solution that suppresses all epileptiform activity. Since the action choices are discrete frequencies, the optimal frequency was 2 Hz, the minimum integer frequency nearest the analytically calculated minimum to suppress seizures. While the stimulation energy is higher, the total epileptiform activity present during the simulation is lower than seen in Fig. 7 (top). Once the optimal stimulus frequency is identified, the Softmax policy for action selection using a very low temperature maintained the stimulation parameters throughout the remainder of the simulation. Turning off stimulation leads to reemergence of seizures, shown at the end of the simulation.

Increasing the temperature value ($\tau_s = 1$) in the Softmax action selection algorithm alters the action selection policy (Fig. 8). Instead of converging on a single stimulation frequency to suppress the seizures, it converges on a distribution from which the TD(0) selects stimulation frequencies probabilistically based on expected value. In this case, the average stimulation frequency was 2.05 Hz.

The effects of temperature in the Softmax algorithm on the number of escaped seizures, convergence time and stimulation frequency are shown in Fig. 9. At very low temperature values ($\tau_s = 0.01$), the TD(0) algorithm behaves in a greedy fashion and converges on a solution quickly (Fig. 6).

As the temperature value increases, the algorithm spends more time searching the parameter space, which ensures a global minimum at the cost of time

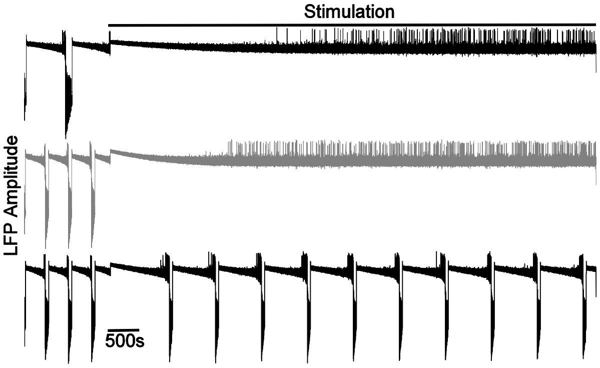


Fig. 7. Open-loop stimulation can control seizures if sufficient stimulation energy is applied. Top, applying stimulus pulses into the z variable at the minimum calculated stimulus frequency (1.04 Hz) with slow seizure dynamics ($\tau_0 = 800$) controls seizures. Middle, similarly, seizures can be controlled if the seizure dynamics are faster ($\tau_0 = 400$) when stimulation is administered at the minimum calculated frequency (2.08 Hz). Bottom, insufficient stimulus energy cannot control seizures. In this example 1.04 Hz was applied to seizures with $\tau_0 = 400$. Black bar indicates duration of stimulus. Note that the time scale has been condensed to see how dynamics evolve on a longer time scale, and the seizures appear as abrupt downward deflections (DC offset) in the LFP.

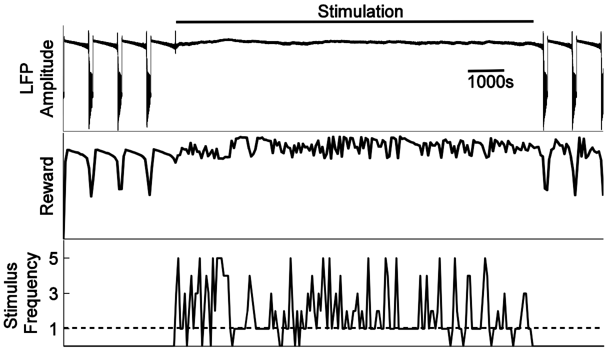


Fig. 8. TD(0) algorithm action selection with high Softmax temperature value. Softmax policy action selection is more explorative when temperature values are high ($\tau_s = 1$). Top, raw LFP trace. Middle, reward trace associated with the LFP. Bottom, Stimulation frequency selected at each action selection interval. Stimulus frequencies selected vary between 0 and 5 Hz, but the average stimulus frequency is approximately 2.05 Hz. Dotted black line represents the minimum stimulus frequency required for control. Black bar indicates duration of stimulus.

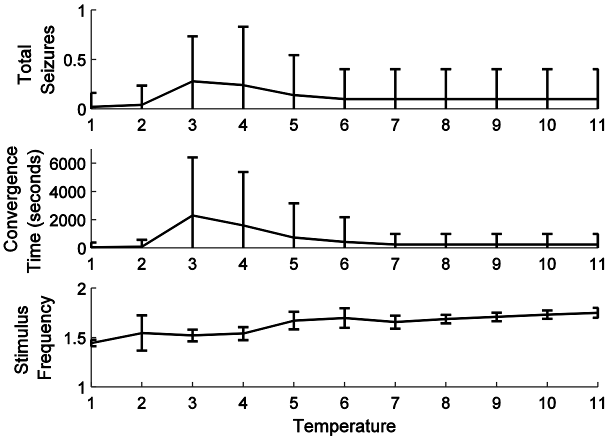


Fig. 9. Performance of reinforcement learning algorithm at different temperature values of Softmax action selection for large ISI($\tau_0 = 800$). Top, mean total number of seizures that escape "therapy" when the reinforcement learning algorithm is choosing stimulation actions. Middle, mean time to converge (seconds) on a policy that controls seizures. Bottom, average stimulus frequency during the simulation.

of convergence The mean convergence time increases to over 2000 s when the temperature value (τ_s). is at 0.2. However, at very high temperatures we found a surprising behavior. The algorithm does not settle on any single stimulus frequency, but instead uses a distribution of stimulus parameters and from which is selects probabilistically to deliver stimuli and optimizes the distribution (Fig. 8) resulting in a good convergence rate to the globally optimal solution. Ultimately, there is a tradeoff in the choice of temperature value for the Softmax action selection policy. Lower energy solutions are typically less robust. For instance, the lowest energy solutions occur when the temperature value is between 0.1 and 0.2 (Fig. 10), but this results in the highest number of seizures after onset of the controller.

3.3. *TD(0) converges to optimal solution when seizure dynamics are faster*

We tested whether the TD(0) algorithm could robustly find an optimal stimulation frequency when the parameter τ_0 of the epileptor model was changed. In these simulations, the seizure frequency increased thereby requiring a net increase in stimulation energy, as shown in Fig. 10. During the learning process, two seizures escape requiring further refinement in action selection. Towards the end of stimulation the algorithm converges on 3 Hz as the optimal stimulation frequency for controlling seizures with a fast time constant ($\tau = 400$).

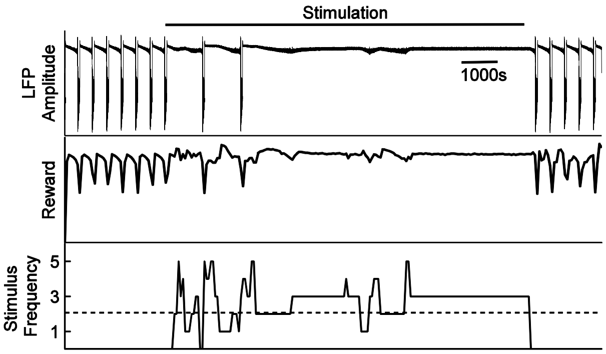


Fig. 10. TD(0) algorithm converges to optimal solution (3 Hz) when seizure dynamics are faster ($\tau = 400$). Top, raw LFP trace. Middle reward trace associated with the LFP. Bottom, stimulation frequency selected at each action selection interval. The TD(0) algorithm needs to explore the state-action space for a longer period of time when the ISI is smaller. Dotted black line represents the minimum stimulus frequency required for control (2.08 Hz). Black bar indicates duration of stimulus.

4. Discussion

This paper presents a reinforcement learning algorithm to determine the optimal stimulation parameters for seizure control. Specifically, we use a reduced TD(0) algorithm to determine low energy stimulation parameters and tested the approach in a computational model of epilepsy called epileptor.

4.1. *State-space approximation*

The TD(0) algorithm determines the optimal stimulation parameters given an estimate of the current state of the system. The optimal action for the ictal-state may be different than that found during the inter-ictal state. This allows the algorithm to optimize stimulation parameters to prevent seizures as well as find the optimal parameters to terminate seizures, which may be different. To separate seizing versus nonseizing we used a low and high pass filtered the raw LFP data. This approach was motivated by the clear partitioning of the inter-ictal an ictal epochs seen in the state-space analysis of the x_1 and x_2 variables, and the reconstructed state-space of the x_1' and x_2' variables. Seizure control is achieved when a therapy brings the system to a state where seizures are unlikely to occur. Seizure termination, on the other hand, can be defined as the stimulus induced suppression of seizures following seizure initiation. While seizure termination was not the central focus of this paper, the TD(0) algorithm nonetheless finds the optimal solutions based on the state of the epileptor (ictal versus inter-ictal), and ultimately optimizes for seizure suppression and seizure termination.

With finer state-space partitioning, or the addition of additional feature dimensions, it may be possible to identify pre-ictal states in which stimulation could prevent seizures. These states would be different from inter-ictal states where stimulation has very limited benefit on seizure suppression. Further division of the state-space, however, results in additional state-action pairs, which increases the computational cost. Without any underlying model of the system, this can result in long searches before finding an efficient policy. Therefore, in this paper, we focused only on the simple state-space partition of ictal and nonictal but finer partitioning with rapid identification of a good control policy remains a problem that needs to be further investigated.

A limitation to this work is how well the design choices (i.e. reward function and state-space discretion) would carry over to other computational and experimental models. Our design approach was specific for the Epileptor model, and implementation of the TD(0) algorithm in other contexts may require a different reward function and state-space discretization.

4.2. *Stimulation variables*

In this paper, TD(0) was used to optimize stimulus pulses applied to the z-variable, the slow permittivity variable in the Epileptor model. The z-variable may be a physiological correlate for K^+ dynamics,[19] or could potentially even model adenosine dynamics.[22] Experimental evidence shows that electrical stimulation directly modulates adenosine concentrations[22] and fast stimulation causes an increase in extracellular potassium concentration.[9]

The TD(0) algorithm can also be used to optimize stimulation applied to the x_1 and x_2 state variables, from which the LFP is directly estimated. In modeling experiments, not shown in this paper, the reinforcement learning algorithm was able to control seizures by applying stimuli to these variables. Since the time-constants of these variables are much faster than the z-variable, higher frequency stimulation was needed to achieve similar seizure control.

In this study, we held stimulus amplitude constant while optimizing over the stimulus frequency. The map in Fig. 4 indicates that the opposite is also possible: stimulus frequency could be fixed while optimizing the stimulus amplitude. We used a discrete pulse because it was the simplest and most commonly used stimulation waveform class in neuroscience. It is possible that other waveform classes could result in a stable fixed point with less energy, but exploration into this is outside the scope of the manuscript. Ultimately, the epileptor model is sensitive to total energy, therefore optimizing over stimulus frequency and amplitude can be done, but the number of state-action pairs grows rapidly with the number of parameters to be optimized resulting in long convergence times to a solution.

In the epileptor model, the amount of activity was monotonically dependent on the amount of energy applied to the z-variable. The only cost to high frequency stimulation was from the stimulation energy subtracted from the reward function. However, in experiments, we have found that the effect of stimulation is not monotonic and there are bands of stimulation frequencies that suppress seizures while other frequencies induce seizures. The TD(0) algorithm will find the global minimum for each state-action pair, but at the cost of occasionally inducing seizures that occur while testing stimulation frequencies that are ictogenic. Black box modeling, specifically input–output models could be empirically determined to predict the effects of periodic pulsatile stimuli at different frequencies. Frequencies that predict emergence of epileptogenic biomarkers could then be cut out of the optimization process, thereby reducing the parameter space.

4.3. *Action selection*

Under the TD(0) framework, action selection is a tradeoff between exploration and exploitation. If the temperature value of the Softmax policy (Eq. (5)) is very low, the TD(0) quickly explores the state-action pairs and then converges on the optimal solution from this estimation. This is clearly evident in the simulation presented in Fig. 6. Sometimes, the quick solution may not be the global maximum, therefore, making greedy choices more likely to be less cost effective in the long run. This approach focuses on finding a single action that it considers as optimal.

High temperature values, on the other hand, lead to a policy that does not converge to a single action. Instead, actions are selected probabilistically from distribution of actions, as seen in Fig. 9. There is no one action, rather the algorithm optimizes the selection probability across all the actions resulting in an average stimulus frequency close to optimal to obtain seizure control.

The action selection paradigms explored in the paper are discrete in nature (i.e. from a set of stimulation frequencies). This limits the ability of the algorithm to find an optimal solution, especially if it is not in the defined set. One way to circumvent this is to implement other reinforcement learning methods which can be applied to a continuous action space.

4.4. *Algorithm robustness*

To the best of our knowledge there are no other computational models that capture the transition from inter-ictal to ictal activity. While we have not tested if our method is robust to changes in the class of models, we have tested robustness to parameter variations. In order to evaluate the robustness, we evaluated the algorithm's performance as parameters were varied. Specifically, we decided to evaluate the algorithm by changing the ISIs and seizure durations. These parameters are modulated by the τ_0 parameter. Reducing the value of the τ_0 parameter resulted in smaller ISIs and shorter seizure durations as shown in Fig. 10. The TD(0) was able to determine the optimal stimulation frequency even when seizure rates increased dramatically. These simulations indicate that our method is robust to changes in the seizure dynamics.

5. Conclusion

The results presented in this paper show that a TD(0) algorithm can effectively converge to the optimal stimulus parameter. Here optimization was performed only over stimulation frequency. This approach can be used to optimize over both stimulation frequency and amplitude at the cost of increased computation and an increase in convergence time. The method presented is computationally efficient and could potentially be programmed into a small bidirectional neural interface. Future studies will require testing of the algorithm in an *in vivo* epilepsy model.

Acknowledgments

The authors gratefully acknowledge funding support from the Epilepsy Foundation pre-doctoral fellowship (VN), 3M Science and Technology pre-doctoral fellowship (VN), NSF CAREER Award (0954797, TIN), NSF Collaboration grant (1264535, TIN), UMN Digital Technology Center grant (TIN, AL).

Appendix A

A.1. *Epileptor model*

$$\dot{x}_1 = \frac{1}{\tau_1}(y_1 - f_1(x_1, x_2) - z), \tag{A.1}$$

$$\dot{y}_1 = \frac{1}{\tau_1}(1 - 5x_1^2 - y_1), \tag{A.2}$$

$$\dot{z} = \frac{1}{\tau_0}(h(x_1) - z - I_1), \tag{A.3}$$

$$\begin{aligned}\dot{x}_2 = \frac{1}{\tau_1}(&-y_2 + x_2 - x_2^3 + I_2 \\ &+ 1.8u(x_1) - 0.3(z - 3.5)),\end{aligned} \tag{A.4}$$

$$\dot{y}_2 = \frac{1}{\tau_2}(-y_2 + f_2(x_1, x_2)), \tag{A.5}$$

$$f_1(x_1, x_2) = \begin{cases} x_1^3 - 3x_1^2, & \text{if } x_1 < 0, \\ (x_2 - 0.6(z-4)^2 x_1), & \text{if } x_1 \geq 0, \end{cases} \tag{A.6}$$

$$f_1(x_1, x_2) = \begin{cases} 0, & \text{if } x_2 < -0.25, \\ 6(x_2 + 0.25)x_1, & \text{if } x_2 \geq -0.25, \end{cases} \tag{A.7}$$

$$\dot{u} = -\gamma(u - 0.1x_1), \tag{A.8}$$

$$h(x_1) = x_0 + \frac{10}{(1 + e^{\frac{-x_1 - 0.5}{0.1}})}, \tag{A.9}$$

$$I_1 = 3.1, \quad I_2 = 0.45, \quad \tau_0 = 800 \text{ or } 400,$$

$$\tau_1 = 0.005, \quad \tau_2 = 0.01, \quad \gamma = 0.01.$$

A.2. *Reduced epileptor model*

We used a two-dimensional epileptor model reduction to analytically compute the minimum energy stimulus parameters to suppress seizures given the time constant of the z variable.[18] The z variable equation remained the same (Eq. (A.3)) while the $\dot{x}_1$ variable was changed.

$$\dot{x}_1 = -x_1^3 - 2x_1^2 + 1 - z + I_1, \tag{A.10}$$

$$\dot{z} = \frac{1}{\tau_0}(h(x_1) - z), \tag{A.11}$$

where I_1 is the constant current used in the full epileptor model.

A.3. *TD*(0) *algorithm*

The expected reward matrix Q is initialized to a value greater than the average reward μ_0 plus some Gaussian noise disturbance with variance ≈ 0.001.

$$\text{Initialize } Q(s, a) > \mu_0. \tag{A.12}$$

Select action a'

    Begin loop
    Apply action $a' \to a$
    Measure new state s'
    Measure actual reward R
    Calculate error:

$$\delta = R - Q(s, a). \tag{A.13}$$

    Update Expected Reward:

$$Q(s, a) = Q(s, a) + \alpha * \delta. \tag{A.14}$$

    Select new action:

$$P(a' \mid s') = \frac{e^{\frac{Q(s', a')}{\tau_s}}}{\sum_{i=1}^{n} e^{\frac{Q(s', a')}{\tau_s}}}, \tag{A.15}$$

Where,

$$\alpha = (\alpha_0 - a_\infty)\beta^{\text{count}} + \alpha_\infty, \tag{A.16}$$

$$\beta \approx e^{-w/\text{ISI}}, \tag{A.17}$$

$$\text{count} = \begin{cases} \text{count} + 1, & \text{if } s' = s \text{ and } a' = a, \\ \text{count}, & \text{otherwise}, \end{cases} \tag{A.18}$$

    End loop.

References

1. WH, Organization Epilepsy Fact Sheet (2015).
2. J. Engel, S. Wiebe, J. French, M. Sperling, P. Williamson, D. Spencer, R. Gumnit, C. Zahn, E. Westbrook and B. Enos, Practice parameter: Temporal lobe and localized neocortical resections for epilepsy, *Epilepsia* **44**(6) (2003) 741–751.
3. A. Handforth, C. M. DeGiorgio, S. C. Schachter, B. M. Uthman, D. K. Naritoku, E. S. Tecoma, T. R. Henry, S. D. Collins, B. V. Vaughn, R. C. Gilmartin, D. R. Labar, G. L. Morris, M. C. Salinsky, I. Osorio,

R. K. Ristanovic, D. M. Labiner, J. C. Jones, J. V. Murphy, G. C. Ney and J. W. Wheless, Vagus nerve stimulation therapy for partial-onset seizures: A randomized active-control trial, *Neurology* **51**(1) (1998) 48–55.

4. R. Fisher, V. Salanova, T. Witt, R. Worth, T. Henry, R. Gross, K. Oommen, I. Osorio, J. Nazzaro, D. Labar, M. Kaplitt, M. Sperling, E. Sandok, J. Neal, A. Handforth, J. Stern, A. DeSalles, S. Chung, A. Shetter, D. Bergen, R. Bakay, J. Henderson, J. French, G. Baltuch, W. Rosenfeld, A. Youkilis, W. Marks, P. Garcia, N. Barbaro, N. Fountain, C. Bazil, R. Goodman, G. McKhann, K. Babu Krishnamurthy, S. Papavassiliou, C. Epstein, J. Pollard, L. Tonder, J. Grebin, R. Coffey, N. Graves and S. S. Group, Electrical stimulation of the anterior nucleus of thalamus for treatment of refractory epilepsy, *Epilepsia* **51**(5) (2010) 899.
5. M. J. Morrell, Group RNSSiES. Responsive cortical stimulation for the treatment of medically intractable partial epilepsy, *Neurology* **77**(13) (2011) 1295.
6. F. T. Sun, M. J. Morrell and R. E. Wharen, Responsive cortical stimulation for the treatment of epilepsy, *Neurotherapeutics* **5**(1) (2008) 68–74.
7. C. T. Anderson, T. K. Tcheng, F. T. Sun and M. J. Morrell, Day–night patterns of epileptiform activity in 65 patients with long-term ambulatory electrocorticography, *J. Clin. Neurophysiol.* **32**(5) (2015) 406–412.
8. C. N. Heck, D. King-Stephens, A. D. Massey, D. R. Nair, B. C. Jobst, G. L. Barkley, V. Salanova, A. J. Cole, M. C. Smith, R. P. Gwinn, C. Skidmore, P. C. Van Ness, G. K. Bergey, Y. D. Park, I. Miller, E. Geller, P. A. Rutecki, R. Zimmerman, D. C. Spencer, A. Goldman, J. C. Edwards, J. W. Leiphart, R. E. Wharen, J. Fessler, N. B. Fountain, G. A. Worrell, R. E. Gross, S. Eisenschenk, R. B. Duckrow, L. J. Hirsch, C. Bazil, C. A. O'Donovan, F. T. Sun, T. A. Courtney, C. G. Seale and M. J. Morrell, Two-year seizure reduction in adults with medically intractable partial onset epilepsy treated with responsive neurostimulation: Final results of the RNS System Pivotal trial, *Epilepsia* **55**(3) (2014) 432–441.
9. 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, *J. Physiol.* **531**(1) (2001) 181–191.
10. Y. Schiller and Y. Bankirer, Cellular mechanisms underlying antiepileptic effects of low- and high-frequency electrical stimulation in acute epilepsy in neocortical brain slices *in vitro*, *J. Neurophysiol.* **97**(3) (2007) 1887–1902.
11. L. B. Good, S. Sabesan, S. T. Marsh, K. Tsakalis, D. Treiman and L. Iasemidis, Control of synchronization of brain dynamics leads to control of epileptic seizures in rodents, *Int. J. Neural Syst.* **19**(3) (2009) 173.
12. T. S. Nelson, C. L. Suhr, D. R. Freestone, A. Lai, A. J. Halliday, K. J. McLean, A. N. Burkitt and M. J. Cook, Closed-loop seizure control with very high frequency electrical stimulation at seizure onset in the GAERS model of absence epilepsy, *Int. J. Neural Syst.* **21**(2) (2011) 163–173.
13. E. Krook-Magnuson, C. Armstrong, M. Oijala and I. Soltesz, On-demand optogenetic control of spontaneous seizures in temporal lobe epilepsy, *Nat. Commun.* **4** (2013) 1376.
14. G. Panuccio, A. Guez, R. Vincent, M. Avoli and J. Pineau, Adaptive control of epileptiform excitability in an *in vitro* model of limbic seizures, *Exp. Neurol.* **241** (2013) 179.
15. B. J. Gluckman, H. Nguyen, S. L. Weinstein and S. J. Schiff, Adaptive electric field control of epileptic seizures, *J. Neurosci. Off. J. Soc. Neurosci.* **21**(2) (2001) 590.
16. A. Berenyi, M. Belluscio, D. Mao and G. Buzsaki, Closed-loop control of epilepsy by transcranial electrical stimulation, *Science* **337**(6095) (2012) 735.
17. R. Sutton and A. Barto, *Reinforcement Learning: An Introduction* (MIT press, Cambridge, 1998).
18. T. Proix, F. Bartolomei, P. Chauvel, C. Bernard and V. K. Jirsa, Permittivity coupling across brain regions determines seizure recruitment in partial epilepsy, *J. Neurosci.* **34**(45) (2014) 15009–15021.
19. V. K. Jirsa, W. C. Stacey, P. P. Quilichini, A. I. Ivanov and C. Bernard, On the nature of seizure dynamics, *Brain* **137**(8) (2014) 2210–2230.
20. K. Bush and J. Pineau, Manifold embeddings for model-based reinforcement learning under partial observability, *Adv. Neural Inf. Process. Syst.* **22** (2009) 189–197.
21. J. Pineau, A. Guez, R. Vincent, G. Panuccio and M. Avoli, Treating epilepsy via adaptive neurostimulation: A reinforcement learning approach, *Int. J. Neural Syst.* **19**(4) (2009) 227–240.
22. J. J. Van Gompel, M. R. Bower, G. A. Worrell, M. Stead, S. Y. Chang, S. J. Goerss, I. Kim, K. E. Bennet, F. B. Meyer, W. R. Marsh, C. D. Blaha and K. H. Lee, Increased cortical extracellular adenosine correlates with seizure termination, *Epilepsia* **55**(2) (2014) 233–244.