Using batch RL to optimize neurostimulation strategies

Arthur Guez, Joelle Pineau
McGill University

The idea of applying reinforcement learning to optimize deep-brain stimulation strategies for the treatment of epilepsy has not been explored previously. Our long-term goal is to develop an adaptive electrical stimulation device that controls seizures in human patients, here we present our work on developing an adaptive stimulation controller that seeks to suppress seizures in a standard *in vitro* model of epilepsy which uses rat brain slices.

It is not possible to train our agent entirely on-line because rat brain slices do not live long enough for the agent to gather enough data. Since at least part of the learning must be done off-line, a batch reinforcement learning approach is used to train an initial policy. The off-line data is collected using a specific stimulation protocol on each slice, alternating periods without stimulation with periods in which the slice is stimulated at some fixed low frequency. The goal of the agent is to learn a policy of stimulation that minimizes the duration of seizures while minimizing the number of stimulations. The policy sequentially decides which frequency of stimulation to use for a small amount of time until the next decision point. Our action space is discrete since we only consider a finite selection of those fixed frequencies of stimulation.

Our data is high-dimensional, continuous, and substantially noisy; for that reason we require a robust technique to be able to learn anything from it. We employ the well-known Fitted Q Iteration (FQI) algorithm [1] to learn an initial policy from the processed off-line data. To learn the mapping $Q : S \times A \rightarrow \Re$ at each iteration of the FQI algorithm, we apply the Extremely Randomized Trees algorithm [2] which builds an ensemble of decision trees.

Due to the expense of doing *in vitro* experiments, it is problematic to directly evaluate our policy on rat brain slices after the learning phase. For an initial validation of our results, we consider quantitative test measures that are obtained using the data from a *hold-out* slice. Those results are already available and can be consulted in [3].

An important detail in our methodology is the choice of building a single ensemble of trees that encompass all actions at each iteration of the FQI algorithm as opposed to building a set of tree forest where each one corresponds to an action. In other words, we are allowing the regression algorithm to approximate the Q-value function across actions. Because our training data is sparse and because our actions have a local effect on the system, the advantage of approximating across actions is significant. It keeps the learning agent from being too optimistic in regions of the state space where data is sparse which ultimately leads to a better policy. We tested that hypothesis empirically on a simplified version of our epilepsy problem. We also have preliminary evidence that approximating across actions in our real setting leads to a better policy.

We are currently evaluating the efficacy of our policy of stimulation *in vitro*. In Figure 1, sample traces illustrating the behavior of the learned policy in real-time are compared to other recording conditions (control, 1 Hz stimulation). In Figure 1(a),

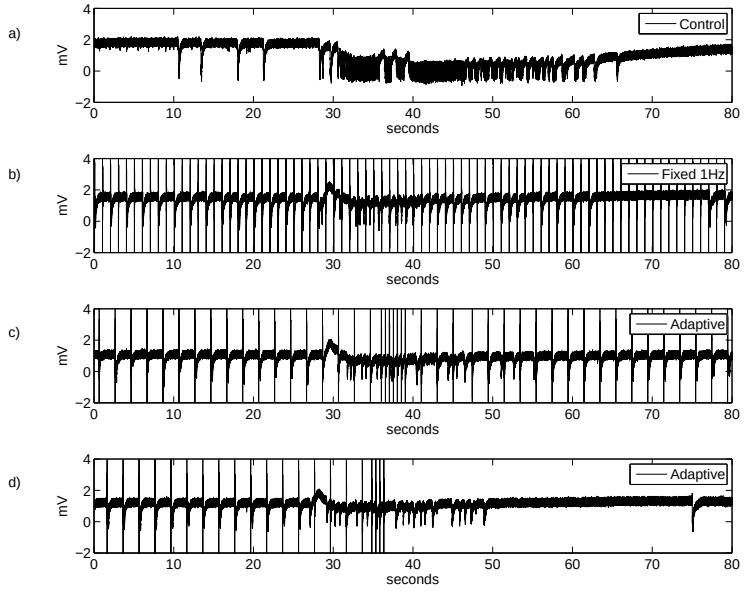

Figure 1: Sample data traces of *in vitro* experiments

we see a seizure typical of the *in vitro* model we are using. In Figure 1(b), we see the 1 Hz stimulation strategy suppressing a seizure which is also typical behavior in the model we are using. In Figure 1(c-d), we see the effects of the adaptive strategy. The learned policy is able to suppress those particular seizures with less stimulations than the fixed 1 Hz strategy. 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. We are also considering different ways to explore on-line to be able to quickly adapt to a particular rat brain slice.

References

[1] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[2] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[3] A. Guez, R.D. Vincent, M. Avoli, and J. Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *Proceedings of the Twentieth Innovative Applications of Artificial Intelligence Conference*, pages 1671–1678, 2008.