A btsp model

guy 1[1] and guy 2[2]

[1]city 1
[2]city 2

Abstract

.. a beautiful abstract :)

1 Introduction

The predominant description of BTSP consists of a two-stage process, where the first is the induction of a sub-threshold instructive signal *IS* in the CA1 pyramidal neurons, and the second is overlapping with a supre-threshold eligibility trace *ET*. Consistent observations agree on the entorhinal cortex (EC) to be the source of the instructive signal *IS*. More specifically for the case of BTSP, the entorhinal afferences are observed to originate in layer III of the lateral region, and target the pyramidal layer in the stratum lacunosum modelculare of the distal CA1 through the temporo-ammonic pathway (Ito, 2012; Soltesz, 2018). The action of these afferences is in the form of dendritic plateau potentials, which consist of calcium spikes evoked by sub-threshold EPSPs in the CA1 apical dendrites (Golding, 1999). The information carried by the *IS* is thought to be related to non-spatial elements of the environment, such as the occurrence of rewards or other behaviourally relevant events. For what concerns the eligibility trace *ET*, it has been identified with the projections from region CA3 of the hippocampus, gated through the Schaffer collaterals (Soltesz, 2018). The *ET* is thought to carry spatial information, such as the perceived location of the animal in the environment. This signal is generated by the CA3 pyramidal neuron, which als receive upstream input from the entorhinal cortex, but predominantly from the medial region (MEC) through the perforant path [**cite**]. Its action in the context of BTSP is a supra-threshold dendritric depolarization. When this occurs, the generated synaptic trace gets integrated with the plateau potentials from the *IS* to produce a long-term potentiation (LTP) of the synapse (Bittner, 2017; Milstein, 2021). Importantly, this process is independent of the

post-synaptic activity, and is meant to capture the temporal contiguity of the pre-synaptic *IS* and *ET* signals. The duration is typically measured in seconds, supporting the idea of consolidating information related to behaviour.

2 Methods

The architecture of the model is meant to resemble the hippocampal formation. It is composed of an input area, being layer III of the entorhinal cortex (EC) with activation $\mathbf{x}_{\text{ECin}}$; two hidden layers $\mathbf{x}_{\text{CA3}}$, $\mathbf{x}_{\text{CA1}}$, representing CA3 and CA1; and an output $\mathbf{x}_{\text{ECout}}$ being layer IV of EC. In particular, it is identified an inner loop ECin $\rightarrow$ CA1 $\rightarrow$ ECout that forms an autoencoder structure, and an outer loop ECin $\rightarrow$ CA3 $\rightarrow$ CA1 $\rightarrow$ ECout, as showed in figure 1-**a**.

The neurons of each layer are artificial nodes, and the forward propagation follows the usual linear combination of the input, connections weights and a bias vector. Concerning the activation function, we defined a modified sigmoid with a sparsity feature, relying on a parameter β; see **??** for details.

The training protocol is defined in two stages. Firstly, the inner loop is trained to match the input pattern $\mathbf{x}_{\text{ECin}}$ with the output $\mathbf{x}_{\text{ECout}}$ using back-propagation and the Adam optimizer, as a proper autoencoder. The goal of this passage is to imprint the connections $\mathbf{W}_{\text{ECin}\rightarrow textCA1}$ with the information for compressing the entorhinal pattern, and the output connection $\mathbf{W}_{\text{CA1}\rightarrow textECout}$ with decoding abilities.

Then, in the second stage the outer loop is involved, recruiting the CA3 projections to CA1. The trained autoencoder connections are frozen, and the focus is placed on the $\mathbf{W}_{\text{CA3}\rightarrow\text{CA1}}$ connections. Here, training occurs in the form of synaptic plasticity, and relies on the ECin input, identified as the instructive signal (IS), and the CA3 activity, designated to be the eligibility trace (ET). The goal of this stage is actually mimick the process of memory consolidation and retrieval, taking care of preserving the deconding of the the neural traces despite continuosly memorizing new patterns. Moere specifically, the learning process is aimed at strengthen the CA3-CA1 synapses that would allow the outer loop, without the ECin $-$ CA1 projections, to reproduce the input activation. The role of the IS is thus to provide of mask over the CA1 neurons, highlighting those whose activity would more reliably preserve the input when propagate to ECout.

Importantly, the set of patterns used in the two stages are samples from the same distribution, such that the inductive biases developed during from the autoencoder training can guide the selection process enacted by IS.

The specific formulation of the learning rule is inspired by BTSP, in that it solely depends on IS ($\mathbf{x}_{\text{ECin}}$) and ET ($\mathbf{x}_{\text{CA3}}$) while discarding the post-synaptic activation $\mathbf{x}_{\text{CA1}}$:

$$\mathbf{W}_{\text{CA3}\rightarrow\text{CA1}} = (1 - \mathbf{x}_{\text{ECin}} * \alpha) * \mathbf{W}_{\text{CA3}\rightarrow\text{CA1}} + \alpha * (\mathbf{x}_{\text{ECin}} \cdot \mathbf{x}_{\text{CA3}}) \qquad (1)$$

The hyperparameter β effectively functions as a learning rate, modulating the integration of the new signal into the previous state of the connections.

Similarly to other accounts of the BTSP rule, the strenghtening and weakening of synaptic weights is influenced by the relative timing of the input activations it relies on. In particular, synaptic potentiation takes place in the case of temporal coincidence, but also factoring in the magnitude of the current weight value. In fact, above a certain threshold synaptic depression takes on regardless of the relative timing. In our context, given the discrete nature of the plasticity training protocol during the second stage, the partition of the kernel landscape is also markedly discrete, as visualized in figure 1-**b**.

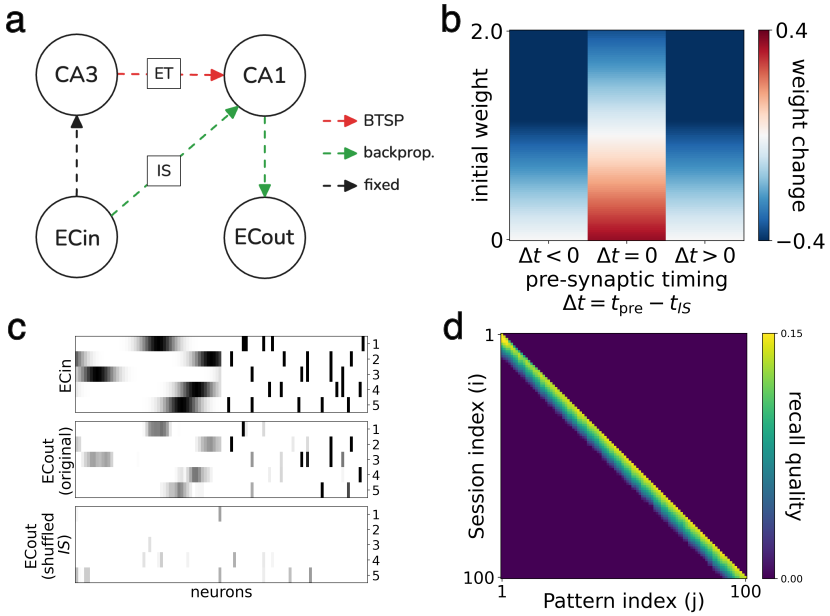

Figure 1: FIGURE 1 - **a**: *architecture, composed of two input/output entorhinal layers and two hidden hippocampal layers.* **b**: *kernel of the weight update resulting from the plasticity rule, blue represents synaptic depression while red synaptic potentiation.* **c**: *reconstruction of the* EC_{in} *input pattern from a model with IS intact (original) and shuffled.* **d**: *recall quality for all memories* j*learned before pattern* i*.*

3 Results

4 Appendix

Algorithm 1: Sparsemoid Activation Function

Input: Tensor z, Integer K, Parameter β, Boolean $flag$

Output: Transformed tensor z

if $K > 0$ **then**

Sort z in descending order along dimension 1:

$$z_{\text{sorted}} \leftarrow \text{sort}(z, \text{descending})$$

Extract the $(K-1)$-th and K-th elements along dimension 1:

$$\alpha \leftarrow z_{\text{sorted}}[:, K-1 : K+1]$$

Compute the mean of α along axis 1:

$$\alpha \leftarrow \text{mean}(\alpha, \text{axis} = 1) \text{ reshaped to } (-1, 1)$$

Apply the transformation:

$$z \leftarrow \beta \cdot (z - \alpha)$$

Return $\sigma(z) = \frac{1}{1+e^{-z}}$ (Sigmoid activation)