bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Presynaptic Stochasticity Improves Energy Efficiency and Alleviates the Stability-Plasticity Dilemma

Simon Schug[1,*], Frederik Benzing[2,*], Angelika Steger[2]

[1]Institute of Neuroinformatics, University of Zurich & ETH Zurich, Zurich, Switzerland
[2]Department of Computer Science, ETH Zurich, Zurich, Switzerland
[*]These authors contributed equally to this work

Abstract

When an action potential arrives at a synapse there is a large probability that no neurotransmitter is released. Surprisingly, simple computational models suggest that these synaptic failures enable information processing at lower metabolic costs. However, these models only consider information transmission at single synapses ignoring the remainder of the neural network as well as its overall computational goal. Here, we investigate how synaptic failures affect the energy efficiency of models of entire neural networks that solve a goal-driven task. We find that presynaptic stochasticity and plasticity improve energy efficiency and show that the network allocates most energy to a sparse subset of important synapses. We demonstrate that stabilising these synapses alleviates the stability-plasticity dilemma, thus connecting a presynaptic notion of importance to a computational role in lifelong learning. Overall, our findings present a set of hypotheses for how presynaptic plasticity and stochasticity contribute to sparsity, energy efficiency and improved trade-offs in the stability-plasticity dilemma.

1 Introduction

It has long been known that synaptic signal transmission is stochastic (del Castillo and Katz, 1954). When an action potential arrives at the presynapse, there is a high probability that no neurotransmitter is released – a phenomenon observed across species and brain regions (Branco and Staras, 2009). From a computational perspective, synaptic stochasticity seems to place unnecessary burdens on information processing. Large amounts of noise hinder reliable and efficient computation (Shannon, 1948; Faisal et al., 2005) and synaptic failures appear to contradict the fundamental evolutionary principle of energy-efficient processing (Niven and Laughlin, 2008). The brain, and specifically action potential propagation consume a disproportionately large fraction of energy (Attwell and Laughlin, 2001; Harris et al., 2012) – so why propagate action potentials all the way to the synapse only to ignore the incoming signal there?

To answer this neurocomputational enigma various theories have been put forward, see Llera-Montero et al. (2019) for a review. One important line of work proposes that individual synapses do not merely maximise information transmission, but rather take into account metabolic costs, maximising the information transmitted *per unit of energy* (Levy and Baxter, 1996). This approach has proven fruitful to explain synaptic failures (Levy and Baxter, 2002; Harris et al., 2012), low average firing rates (Levy and Baxter, 1996) as well as excitation-inhibition balance (Sengupta et al., 2013) and is supported by fascinating experimental evidence suggesting that both presynaptic glutamate release (Savtchenko et al., 2013) and postsynaptic channel properties (Harris et al., 2015, 2019) are tuned to maximise information transmission per energy.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

However, so far information-theoretic approaches have been limited to signal transmission at single synapses, ignoring the context and goals in which the larger network operates. As soon as context and goals guide network computation certain pieces of information become more relevant than others. For instance, when reading a news article the textual information is more important than the colourful ad blinking next to it – even when the latter contains more information in a purely information-theoretic sense.

Here, we study presynaptic stochasticity on the network level rather than on the level of single synapses. We investigate its effect on (1) energy efficiency and (2) the stability-plasticity dilemma in model neural networks that learn to selectively extract information from complex inputs.

We find that presynaptic stochasticity in combination with presynaptic plasticity allows networks to extract information at lower metabolic cost by sparsely allocating energy to synapses that are important for processing the given stimulus. As a result, presynaptic release probabilities encode synaptic importance. We show that this notion of importance is related to the Fisher Information, a theoretical measure for the network's sensitivity to synaptic changes.

Building on this finding and previous work (Kirkpatrick et al., 2017) we explore a potential role of presynaptic stochasticity in the stability-plasticity dilemma. In line with experimental evidence (Yang et al., 2009; Hayashi-Takagi et al., 2015), we demonstrate that selectively stabilising important synapses improves lifelong learning. Furthermore, these experiments link presynaptically induced sparsity to improved memory.

2 Model

Our goal is to understand how information processing and energy consumption are affected by stochasticity in synaptic signal transmission. While there are various sources of stochasticity in synapses, here, we focus on modelling *synaptic failures* where action potentials at the presynapse fail to trigger any postsynaptic depolarisation. The probability of such failures is substantial (Branco and Staras, 2009; Hardingham et al., 2010; Sakamoto et al., 2018) and, arguably, due to its all-or-nothing-characteristic has the largest effect on both energy consumption and information transmission.

As a growing body of literature suggests, artificial neural networks (ANNs) match several aspects of biological neuronal networks in various goal-driven situations (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Kell et al., 2018; Banino et al., 2018; Cueva and Wei, 2018; Mattar and Daw, 2018). Crucially, they are the only known model to solve complex vision and reinforcement learning tasks comparably well as humans. We therefore choose to extend this class of models by explicitly incorporating synaptic failures and study their properties in a number of complex visual tasks.

2.1 Model Details

The basic building blocks of ANNs are neurons that combine their inputs $a_1, \ldots, a_n$ through a weighted sum $w_1 a_1 + \ldots w_n a_n$ and apply a nonlinear activation function $\sigma(\cdot)$. The weights w_i naturally correspond to *synaptic strengths* between presynaptic neuron i and the postsynaptic neuron. Although synaptic transmission is classically described as a binomial process (del Castillo and Katz, 1954) most previous modelling studies assume the synaptic strengths to be deterministic. This neglects a key characteristic of synaptic transmission: the possibility of synaptic failures where no communication between pre- and postsynapse occurs at all.

In the present study, we explicitly model presynaptic stochasticity by introducing a random variable $r_i \sim \text{Bernoulli}(p_i)$, whose outcome corresponds to whether or not neurotransmitter is

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

released. Formally, each synapse w_i is activated stochastically according to

$$w_i = \underbrace{r_i}_{\substack{\text{stochastic}\\ \text{release}}} \cdot \underbrace{m_i}_{\substack{\text{synaptic}\\ \text{strength}}}, \qquad \text{where } r_i \sim \text{Bernoulli}(\underbrace{p_i}_{\substack{\text{release}\\ \text{probability}}}) \tag{1}$$

so that it has expected synaptic strength $\bar{w}_i = p_i m_i$. The postsynaptic neuron calculates a stochastic weighted sum of its inputs with a nonlinear activation

$$\underbrace{a^{\text{post}}}_{\substack{\text{postsynaptic}\\ \text{activation}}} = \sigma\left(\sum_{i=1}^{n} w_i \underbrace{a_i^{\text{pre}}}_{\substack{i\text{-th presynaptic}\\ \text{input}}}\right). \tag{2}$$

During learning, synapses are updated and both synaptic strength and release probability are changed. We resort to standard learning rules to change the synaptic strength. For the multilayer perceptron, this update is based on stochastic gradient descent given the loss function $L(\bar{w}_i, p_i)$, i.e.

$$\bar{w}_i^{(t+1)} = \bar{w}_i^{(t)} - \eta g_i, \qquad \text{where } g_i = \frac{\partial L(\bar{w}_i^{(t)}, p_i)}{\partial \bar{w}_i^{(t)}} \tag{3}$$

where the superscript corresponds to time steps. For the standard perceptron model, g_i is given by its standard learning rule (Rosenblatt, 1958). Based on the intuition that synapses which receive larger updates are more important for solving a given task, we update p_i using the update direction g_i according to the following simple scheme

$$p_i^{(t+1)} = \begin{cases} p_i^{(t)} + p_{\text{up}}, & \text{if } |g_i| > g_{\text{lim}}, \\ p_i^{(t)} - p_{\text{down}}, & \text{if } |g_i| \leq g_{\text{lim}}. \end{cases} \tag{4}$$

Here, $p_{\text{up}}, p_{\text{down}}, g_{\text{lim}}$ are three metaplasticity parameters shared between all synapses.[1] To prevent overfitting and to test robustness, we tune them using one learning scenario and keep them fixed for all other scenarios, see Materials and Methods. To avoid inactivated synapses with release probability $p_i = 0$ we clamp p_i to stay above 0.25, which we also use as the initial value of p_i before learning.

On top of the above intuitive motivation, we give a theoretical justification for this learning rule in Materials and Methods, showing that synapses with larger Fisher Information obtain high release probabilities, also see Figure 2d.

2.2 Measuring Energy Consumption

For our experiments, we would like to quantify the energy consumption of the neural network. Harris et al. (2012) find that the main constituent of neural energy demand is synaptic signal transmission and that the cost of synaptic signal transmission is dominated by the energy needed to reverse postsynaptic ion fluxes. In our model, the component most closely matching the size of the postsynaptic current is the expected synaptic strength, which we therefore take as measure for the model's energy consumption. In the Supplementary, we also measure the metabolic cost incurred by the activity of neurons by calculating their average rate of activity.

[1] We point out that in a noisy learning setting the gradient g does not decay to 0, so that the learning rule in (4) will maintain network function by keeping certain release probabilities high. See also Material and Methods for a theoretical analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Box 1: Mutual Information

The Mutual Information $I(Y;Z)$ of two jointly distributed random variables Y, Z is a common measure of their dependence (Shannon, 1948). Intuitively, mutual information captures how much information about Y can be obtained from Z, or vice versa. Formally, it is defined as

$$I(Y;Z) \equiv H(Y) - H(Y|Z) = H(Z) - H(Z|Y)$$

where $H(Y)$ is the entropy of Y and $H(Y|Z)$ is the conditional entropy of Y given Z. In our case, we want to measure how much task-relevant information Y is contained in the neural network output Z. For example, the neural network might receive as input a picture of a digit with the goal of predicting the identity of the digit. Both the ground-truth digit identity Y and the network's prediction Z are random variables depending on the random image X. The measure $I(Y;Z)$ quantifies how much of the behaviourally relevant information Y is contained in the network's prediction Z ignoring irrelevant information also present in the complex, high-entropy image X.

2.3 Measuring Information Transmission

We would like to measure how well the neural network transmits information relevant to its behavioural goal. In particular, we are interested in the setting where the complexity of the stimulus is high relative to the amount of information that is relevant for the behavioural goal. To this end, we present complex visual inputs with high information content to the network and teach it to recognise the object present in the image. We then measure the mutual information between network output and object identity, see Feature Box 1.

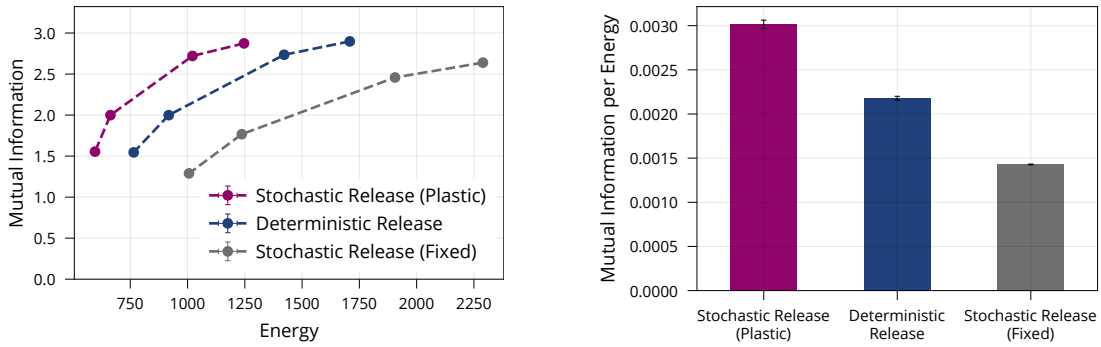
3 Results

3.1 Presynaptic Stochasticity Enables Energy-Efficient Information Processing

We now investigate the energy efficiency of a network that learns to classify digits from the MNIST handwritten digit dataset (LeCun, 1998). The inputs are high-dimensional with high entropy, but the relevant information is simply the identity of the digit. We compare the model with plastic, stochastic release to two controls. A standard ANN with deterministic synapses is included to investigate the combined effect of presynaptic stochasticity and plasticity. In addition, to isolate the effect of presynaptic plasticity, we introduce a control which has stochastic release, but with a fixed probability. In this control, the release probability is identical across synapses and chosen to match the average release probability of the model with plastic release after it has learned the task.

All models are encouraged to find low-energy solutions by penalising large synaptic weights through standard ℓ_2-regularisation. Figure 1a shows that different magnitudes of ℓ_2-regularisation induce different information-energy trade-offs for all models, and that the model with plastic, stochastic release finds considerably more energy-efficient solutions than both controls, while the model with non-plastic release requires more energy then the deterministic model. Together, this supports the view that a combination of presynaptic stochasticity and plasticity promotes energy-efficient information extraction.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) (b)

Figure 1: **Energy Efficiency of Model with Stochastic and Plastic Release.** **(a)** Different trade-offs between mutual information and energy are achievable in all network models. Generally, stochastic synapses with learned release probabilities are more energy-efficient than deterministic synapses or stochastic synapses with fixed release probability. The fixed release probabilities model was chosen to have the same average release probability as the model with learned probabilities. **(b)** Best achievable ratio of information per energy for the three models from (a). Error bars in (a) and (b) denote the standard error for three repetitions of the experiment.

In addition, we investigate how stochastic release helps the network to lower metabolic costs. Intuitively, a natural way to save energy is to assign high release probabilities to synapses that are important to extract relevant information and to keep remaining synapses at a low release probability. Figure 2a shows that after learning, there are indeed few synapses with high release probabilities, while most release probabilities are kept low. We confirm that this sparsity develops independently of the initial value of release probabilities before learning, see Supplementary Figure 6d. To test whether the synapses with high release probabilities are most relevant for solving the task we perform a lesion experiment. We successively remove synapses with low release probability and measure how well the lesioned network still solves the given task, see Figure 2b. As a control, we remove synapses in a random order independent of their release probability. We find that maintaining synapses with high release probabilities is significantly more important to network function than maintaining random ones. Moreover, we find, as expected, that synapses with high release probabilities consume considerably more energy than synapses with low release probability, see Figure 2c. This supports the hypothesis that the model identifies important synapses for the task at hand and spends more energy on these synapses while saving energy on irrelevant ones.

We have seen that the network relies on a sparse subset of synapses to solve the task efficiently. However, sparsity is usually thought of on a neuronal level, with few neurons rather than few synapses encoding a given stimulus. Therefore, we quantify sparsity of our model on a neuronal level. For each neuron we count the number of 'important' input- and output synapses, where we define 'important' to correspond to a release probability of at least $p = 0.9$. Note that the findings are robust with respect to the precise value of p, see Figure 2a. We find that the distribution of important synapses per neuron is inhomogeneous and significantly different from a randomly shuffled baseline with a uniform distribution of active synapses (Kolmogorov-Smirnoff test, $D = 0.505, p < 0.01$), see Figure 3a. Thus, some neurons have disproportionately many important inputs, while others have very few, suggesting sparsity on a neuronal level. As additional quantification of this effect, we count the number of highly important neurons, where we define a neuron to be highly important if its number of active inputs is two standard deviations below or above the mean (mean and standard deviation from shuffled baseline). We find that our model network with presynaptic stochasticity has disproportionate numbers

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

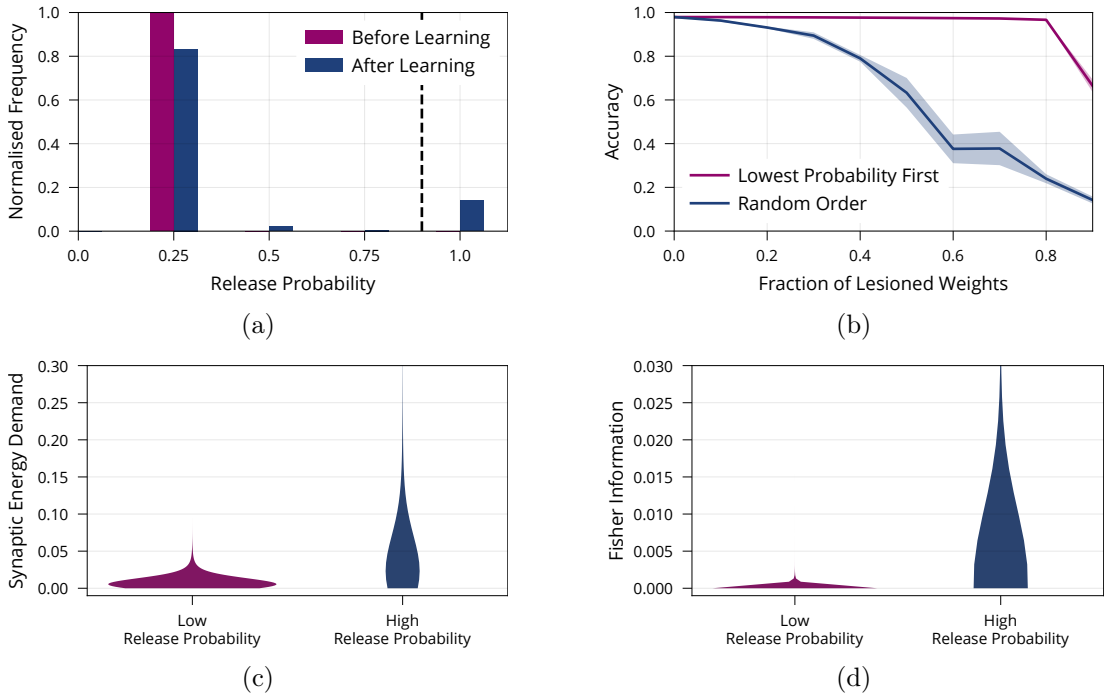


Figure 2: **Importance of Synapses with High Release Probability for Network Function.** **(a)** Histogram of release probabilities before and after learning, showing that the network relies on a sparse subset of synapses to find an energy-efficient solution. Dashed line at $p = 0.9$ indicates our boundary for defining a release probability as 'low' or 'high'. We confirmed that results are independent of initial value of release probabilities before learning (see Supplementary, Figure 6d). **(b)** Accuracy after performing the lesion experiment either removing synapses with low release probabilities first or removing weights randomly, suggesting that synapses with high release probability are most important for solving the task. **(c)** Distribution of synaptic energy demand for high and low release probability synapses. **(d)** Distribution of the Fisher information for high and low release probability synapses. It confirms the theoretical prediction that high release probability corresponds to high Fisher Information. All panels show accumulated data for three repetitions of the experiment. Shaded regions in (b) show standard error.

of highly important and unimportant neurons, see Figure 3c. Moreover, we check whether neurons with many important inputs tend to have many important outputs, indeed finding a correlation of $r = 0.93$, see Figure 3b. These analyses all support the claim that the network is sparse not only on a synaptic but also on a neuronal level.

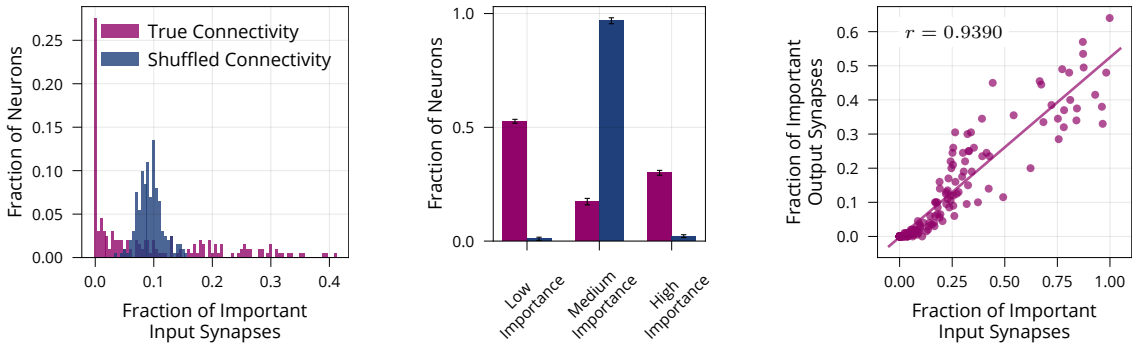
Finally, we investigate how release probabilities evolve from a theoretical viewpoint under the proposed learning rule. Note that the evolution of release probabilities is a random process, since it depends on the random input to the network. Under mild assumptions, we show (Materials and Methods) that release probabilities are more likely to increase for synapses with large Fisher Information[2]. Thus, synapses with large release probabilities will tend to have high Fisher Information. We validate this theoretical prediction empirically, see Figure 2d.

3.2 Presynaptically Driven Consolidation Alleviates the Stability-Plasticity Dilemma

While the biological mechanisms addressing the stability-plasticity dilemma are diverse and not fully understood, it has been demonstrated experimentally that preserving memories requires

[2]In this context, the Fisher Information is a measure of sensitivity of the network to changes in synapses, measuring how important preserving a given synapse is for network function.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) (b) (c)

Figure 3: **Neuron-Level Sparsity of Network after Learning.** **(a)** Histogram of the fraction of important input synapses per neuron for second layer neurons after learning for true and randomly shuffled connectivity (see Supplementary, Figure 7a for other layers). **(b)** Same data as (a), showing number of low/medium/high importance neurons, where high/low importance neurons have at least two standard deviations more/less important inputs than the mean of random connectivity. **(c)** Scatter plot of first layer neurons showing the number of important input and output synapses after learning on MNIST, Pearson correlation is $r = 0.9390$ (see Supplementary, Figure 7b for other layers). Data in (a) and (c) are from one representative run, error bars in (b) show standard error over three repetitions.

maintaining the synapses which encode these memories (Yang et al., 2009; Hayashi-Takagi et al., 2015; Cichon and Gan, 2015). In this context, theoretical work suggests that the Fisher Information is a useful way to quantify which synapses should be maintained (Kirkpatrick et al., 2017). Inspired by these insights, we hypothesise that the synaptic importance encoded in release probabilities can be used to improve the network's memory retention by selectively stabilising important synapses.

We formalise this hypothesis in our model by lowering the learning rate (plasticity) of synapses according to their importance (release probability). Concretely, the learning rate $\eta = \eta(p_i)$ used in (3) is scaled as follows

$$\eta(p_i) = \eta_0 \cdot (1 - p_i). \tag{5}$$

such that the learning rate is smallest for important synapses with high release probability. η_0 denotes a base learning rate that is shared by all synapses. We complement this consolidation mechanism by freezing the presynaptic release probabilities p_i once they have surpassed a predefined threshold p_{freeze}. This ensures that a synapse whose presynaptic release probability was high for a previous task retains its release probability even when unused during consecutive tasks. In other words, the effects of presynaptic long-term depression (PRE-LTD) are assumed to act on a slower timescale than learning single tasks. Note that the freezing mechanism ensures that all synaptic strengths $\bar{w}_i$ retain a small degree of plasticity, since the learning rate modulation factor $(1 - p_i)$ remains greater than 0.

To test our hypothesis that presynaptically driven consolidation allows the network to make improved stability-plasticity trade-offs, we sequentially present a number of tasks and investigate the networks behaviour. We mainly focus our analysis on a variation of the MNIST handwritten digit dataset, in which the network has to successively learn the parity of pairs of digits, see Figure 4a. Additional experiments are reported in the Supplementary Material, see Table 1.

First, we investigate whether presynaptic consolidation improves the model's ability to remember old tasks. To this end, we track the accuracy on the first task over the course of learning, see Figure 4b. As a control, we include a model without consolidation and with deterministic synapses. While both models learn the first task, the model without consolidation forgets more

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

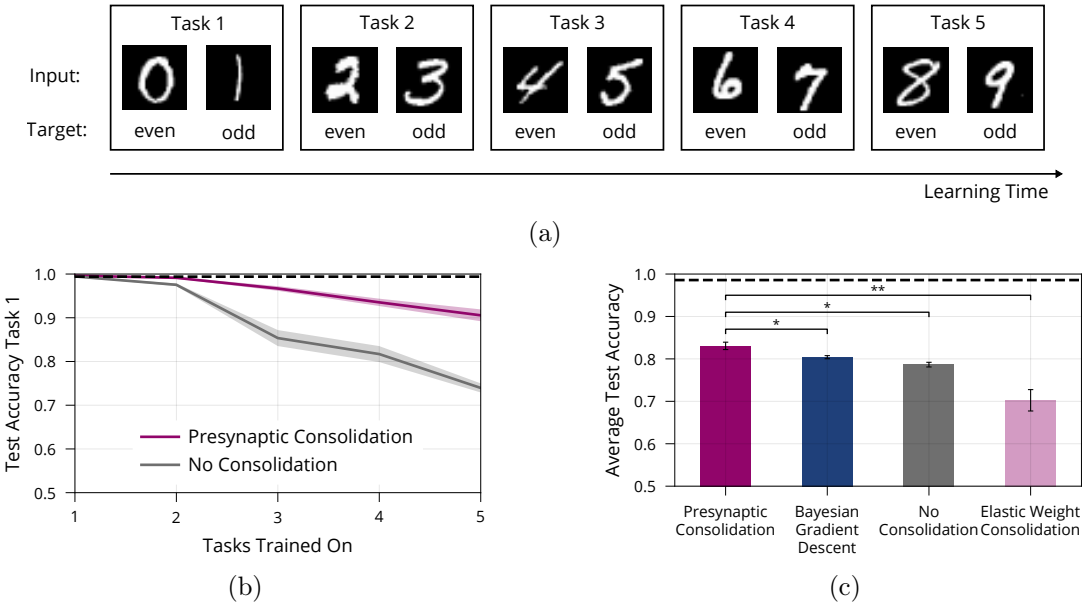


Figure 4: **Lifelong Learning in a Model with Presynaptically Driven Consolidation.** **(a)** Schematic of the lifelong learning task Split MNIST. In the first task the model network is presented 0s and 1s, in the second task it is presented 2s and 3s, etc. For each task the model has to classify the inputs as even or odd. At the end of learning, it should be able to correctly classify the parity of all digits, even if a digit has been learned in an early task. **(b)** Accuracy of the first task when learning new tasks. Consolidation leads to improved memory preservation. **(c)** Average accuracies of all learned tasks. The presynaptic consolidation model is compared to a model without consolidation and two state-of-the-art machine learning algorithms. Differences to these models are significant in independent t-tests with either $p < 0.05$ (marked with *) or with $p < 0.01$ (marked with **). Dashed line indicates an upper bound for the network's performance, obtained by training on all tasks simultaneously. Panels (b) and (c) show accumulated data for three repetitions of the experiment. Shaded regions in (b) and error bars in (c) show standard error.

quickly, suggesting that the presynaptic consolidation mechanism does indeed improve memory.

Next, we ask how increased stability interacts with the network's ability to remain plastic and learn new tasks. To assess the overall trade-off between stability and plasticity we report the average accuracy over all five tasks, see Figure 4c.

We find that the presynaptic consolidation model performs better than a standard model with deterministic synapses and without consolidation. In addition, we compare performance to two state-of-the art machine learning algorithms: The well-known algorithm Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) explicitly relies on the Fisher Information and performs a separate consolidation phase after each task. Bayesian Gradient Descent (BGD) (Zeno et al., 2018) is a Bayesian approach that models synapses as distributions, but does not capture the discrete nature of synaptic transmission. The presynaptic consolidation mechanism performs better than both these state-of-the-art machine learning algorithms, see Figure 4c. Additional experiments in the Supplementary suggest overall similar performance of Presynaptic Consolidation to BGD and similar or better performance than EWC.

To determine which components of our model contribute to its lifelong learning capabilities, we perform an ablation study, see Figure 5a. We aim to separate the effect of (1) consolidation mechanisms and (2) presynaptic plasticity.

First, we remove the two consolidation mechanisms, learning rate modulation and freezing release probabilities, from the model with stochastic synapses. This yields a noticeable decrease

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

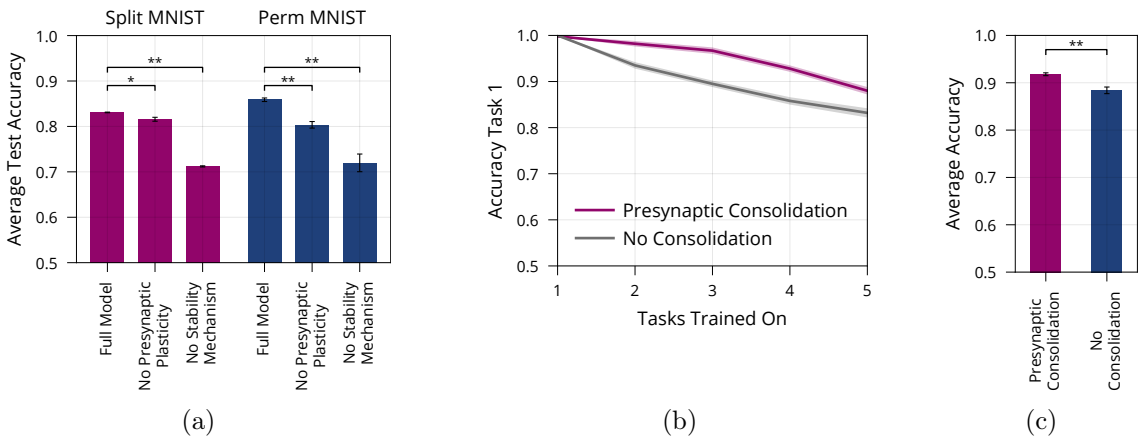


Figure 5: **Model Ablation and Lifelong Learning in a Standard Perceptron. (a)** Ablation of the Presynaptic Consolidation model on two different lifelong learning tasks, see full text for detailed description. Both presynaptic plasticity and synaptic stabilisation significantly improve memory. **(b)+(c)** Lifelong Learning in a Standard Perceptron akin to Figure 4b, 4c, showing the accuracy of the first task when learning consecutive tasks in (b) as well as the average over all five tasks after learning all tasks in (c). Error bars and shaded regions show standard error of three respectively ten repetitions, in (a) respectively (b+c). All pair-wise comparisons are significant, independent t-tests with $p < 0.01$ (denoted by **) or with $p < 0.05$ (denoted by *).

in performance during lifelong learning, thus supporting the view that stabilising important synapses contributes to addressing the stability-plasticity dilemma.

Second, we aim to disentangle the effect of presynaptic plasticity from the consolidation mechanisms. We therefore introduce a control in which presynaptic plasticity but not consolidation is blocked. Concretely, the control has 'ghost release probabilities' $\tilde{p}_i$ evolving according to equation (4) and modulating plasticity according to equation (5); but the synaptic release probability is fixed at 0.5. We see that this control performs worse than the original model with a drop in accuracy of 1.4% on Split MNIST ($t = 3.44$, $p < 0.05$) and a drop of accuracy of 5.6% on Permuted MNIST ($t = 6.72, p < 0.01$). This suggests that presynaptic plasticity, on top of consolidation, helps to stabilise the network. We believe that this can be attributed to the sparsity induced by the presynaptic plasticity which decreases overlap between different tasks.

The above experiments rely on a gradient-based learning rule for multilayer perceptrons. To test whether presynaptic consolidation can also alleviate stability-plasticity trade-offs in other settings, we study its effects on learning in a standard perceptron (Rosenblatt, 1958). We train the perceptron sequentially on five pattern memorisation tasks, see Materials and Methods for full details. We find that the presynaptically consolidated perceptron maintains a more stable memory of the first task, see Figure 5b. In addition, this leads to an overall improved stability-plasticity trade-off, see Figure 5c and shows that the effects of presynaptic consolidation in our model extend beyond gradient-based learning.

4 Discussion

4.1 Main Contribution

Information transmission in synapses is stochastic. While previous work has suggested that stochasticity allows to maximise the amount of information transmitted per unit of energy spent, this analysis has been restricted to single synapses. We argue that the relevant quantity

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to be considered is task-dependent information transmitted by entire networks. Introducing a simple model of the all-or-nothing nature of synaptic transmission, we show that presynaptic stochasticity enables networks to allocate energy more efficiently. We find theoretically as well as empirically that learned release probabilities encode the importance of weights for network function according to the Fisher Information. Based on this finding, we suggest a novel computational role for presynaptic stochasticity in lifelong learning. Our experiments provide evidence that coupling information encoded in the release probabilities with modulated plasticity can help alleviate the stability-plasticity dilemma.

4.2 Modelling Assumptions

Our model captures the occurrence of synaptic failures by introducing a Bernoulli random variable governing whether or not neurotransmitter is released. Compared to classical models assuming deterministic transmission, this is one step closer to experimentally observed binomial transmission patterns, which are caused by multiple, rather than one, release sites between a given neuron and dendritic branch. Importantly, our simplified model accounts for the event that there is no postsynaptic depolarisation at all. Even in the presence of multiple release sites, this event has non negligible probability: Data from cultured hippocampal neurons (Branco et al., 2008, Figure 2D) and the neocortex (Hardingham et al., 2010, Figure 7C) shows that the probability $(1-p)^N$ that none of N release sites with release probability p is active, is around 0.3-0.4 even for N as large as 10. More recent evidence suggests an even wider range of values depending on the extracellular calcium concentration (Sakamoto et al., 2018).

Several parts of our model require communication between pre- and postsynapse: Updating the expected strength (3), changing release probabilities based on postsynaptic changes (4) and consolidating synapses based on the release probability (5). While this property makes learning more biologically complex, it is well known that pre-post communication does underlie several forms of synaptic adaptation (Sanderson et al., 2020).

In our multilayer perceptron model, updates are computed using backpropagated gradients. Whether credit assignment in the brain relies on backpropagation – or more generally gradients – remains an active area of research, but several alternatives aiming to increase biological plausibility exist and are compatible with our model (Sacramento et al., 2018; Lillicrap et al., 2016; Lee et al., 2015). To check that the proposed mechanism can also operate without gradient information, we include an experiment with a standard perceptron and its learning rule (Rosenblatt, 1958), see Figure 5b and 5c.

4.3 Relation to Other Probabilistic Modelling Approaches

The goal of incorporating and interpreting noise in models of neural computation is shared by many computational studies. Inspired by a Bayesian perspective, neural variability is often interpreted as representing uncertainty (Ma et al., 2006; Fiser et al., 2010; Kappel et al., 2015; Haefner et al., 2016), or as a means to prevent overfitting (Wan et al., 2013). The Bayesian paradigm has been applied directly to variability of individual synapses in neuroscience (Aitchison et al., 2014; Aitchison and Latham, 2015; Aitchison et al., 2021) and machine learning (Zeno et al., 2018). It prescribes decreasing the plasticity of synapses with low posterior variance. Aitchison and Latham (2015) find that this normative prediction matches experimental data from Sjöström et al. (2001, 2007). In this context it is interesting to note that our model also decreases the plasticity of synapses with low posterior variance, see Material and Methods for a more detailed description.

In contrast to common Bayesian interpretations (Zeno et al., 2018; Aitchison and Latham, 2015; Kappel et al., 2015) which model release statistics as Gaussians and optimize complex objectives (see also Llera-Montero et al., 2019) our simple proposal represents the inherently discrete nature of synaptic transmission more faithfully.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Modelling synaptic release as a probability distribution raises the question how release statistics ought to adapt during learning (Costa et al., 2017; Llera-Montero et al., 2019). Our framework offers a new perspective on this question, suggesting that jointly learning pre- and postsynaptic parameters may address the inevitable trade-off between metabolic costs and reliable information extraction.

4.4 Synaptic Importance May Govern Energy-Information Trade-offs

Energy constraints are widely believed to be a main driver of evolution (Niven and Laughlin, 2008). From brain size (Isler and van Schaik, 2009; Navarrete et al., 2011), to wiring cost (Chen et al., 2006), down to ion channel properties (Alle et al., 2009; Sengupta et al., 2010), presynaptic transmitter release (Savtchenko et al., 2013) and postsynaptic conductance (Harris et al., 2015, 2019), various components of the nervous system have been shown to be optimal in terms of their total metabolic cost or their metabolic cost per bit of information transmitted.

Crucially, there is evidence that the central nervous system operates in varying regimes, making different trade-offs between synaptic energy demand and information transmission: Perge et al. (2009); Carter and Bean (2009); Hu and Jonas (2014) all find properties of the axon (thickness, sodium channel properties), which are suboptimal in terms of energy per bit of information. They suggest that these inefficiencies occur to ensure fast transmission of highly relevant information.

We propose that a similar energy/information trade-off could govern network dynamics preferentially allocating more energy to the most relevant synapses for a given task. Our model relies on a simple, theoretically justified learning rule to achieve this goal and leads to overall energy savings. Neither the trade-off nor the overall savings can be accounted for by previous frameworks for energy-efficient information transmission at synapses (Levy and Baxter, 2002; Harris et al., 2012).

This view of release probabilities and related metabolic cost provides a way to make the informal notion of "synaptic importance" concrete by measuring how much energy is spent on a synapse. Interestingly, our model suggests that this notion is helpful beyond purely energetic considerations and can in fact help to maintain memories during lifelong learning.

4.5 Relation of Energy Demand to Synaptic Stability and Memories

Our framework links energy demand, synaptic stability and memory consolidation in a principled way, giving a new perspective on experimental findings observing correlations between these quantities. Firstly, there is evidence that memory formation requires energy (Morris et al., 2010; Suzuki et al., 2011; Hebert-Chatelain et al., 2016; Alberini et al., 2018). For example, Hebert-Chatelain et al. (2016) establish a detailed account of how endocannabinoids cause amnesia by altering mobility and energetic activity of axonal mitochondria. A common interpretation of this and related studies is that long-term potentiation (LTP) is required for memory formation and that LTP itself incurs a metabolic cost (Li and Van Rossum, 2020) which is consistent with our model. In addition, our model suggests that recall requires maintaining the metabolically most expensive synapses, giving a complementary view on the described experimental findings. Secondly, the presynaptic release machinery requires energy, and its functioning relies partly on mitochondria (Ivannikov et al., 2013). Some synaptic boutons contain stationary mitochondria, which keep their position for extended time periods ranging up to several weeks (Chang et al., 2006; Obashi and Okabe, 2013). Synapses with stationary mitochondria have been observed to be more stable than synapses without. On a short time scale, they have more stable vesicle release than synapses without stationary mitochondria (Sun et al., 2013). On a longer time scale (at least weeks), there is less turnover of boutons in the mouse neocortex, which contain mitochondria (Lees et al., 2020). These studies suggests a link of synaptic energy demand

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and stability. Our model suggests a computational role of this link, showing that stabilising important synapses contributes to improved memory.

4.6 Presynaptic Contribution to Sparsity and Memory

Our model suggests that presynaptic plasticity improves a network's ability to save energy as well as to maintain memories. A more detailed analysis suggests that energy savings occur as presynaptic plasticity helps to select the most relevant synapses and to sparsify the network elsewhere, supporting common believes about the energetic role of sparsity (Olshausen and Field, 2004). Furthermore, our experiment of blocking presynaptic plasticity but not consolidation suggests that sparsity is helpful for lifelong learning, likely by reducing the overlap of neural representations of different tasks. It is already known from a model of associative memory that such a mechanism improves memory performance (Willshaw et al., 1969). We extend these findings to a more complex model and task.

An increasing number of experimental studies provide evidence that the presynapse significantly contributes to memory formation (see e.g. review by Monday et al., 2018). Our simple model presents a concrete, mechanistic set of hypotheses for how the presynapse contributes to maintaining memories through sparsification and network consolidation.

5 Materials and Methods

5.1 Theoretical Analysis of Presynaptic Learning Rule

As indicated in the results section the release probability p_i is more likely to be large when the Fisher Information of the synaptic strength w_i is large as well. This provides a theoretical explanation to the intuitive correspondence between release probability and synaptic importance. Here, we formalise this link starting with a brief review of the Fisher Information.

5.1.1 Fisher Information

The Fisher Information is a measure for the networks sensitivity to changes in parameters. Under additional assumptions it is equal to the Hessian of the loss function (Pascanu and Bengio, 2013; Martens, 2014), giving an intuitive reason why synapses with high Fisher Information should not be changed much if network function is to be preserved.

Formally, for a model with parameter vector θ predicting a probability distribution $f_\theta(X, y)$ for inputs X and labels y drawn from a joint distribution $\mathcal{D}$, the Fisher Information matrix is defined as

$$\mathbb{E}_{X\sim\mathcal{D}}\mathbb{E}_{y\sim f_\theta(y|X)}\left[\left(\frac{\partial \ln f_\theta(X,y)}{\partial\theta}\right)\left(\frac{\partial \ln f_\theta(X,y)}{\partial\theta}\right)^T\right].$$

Note that this expression is independent of the actual labels y of the dataset and that instead we sample labels from the model's predictions. If the model makes correct predictions, we can replace the second expectation, which is over $y \sim f_\theta(y \mid X)$, by the empirical labels y of the dataset for an approximation called the Empirical Fisher Information. If we further only consider the diagonal entries – corresponding to a mean-field approximation – and write $g_i(X, y) = \frac{\partial \ln f_\theta(X,y)}{\partial\theta_i}$ we obtain the following expression for the i-th entry of the diagonal Empirical Fisher Information:

$$F_i = \mathbb{E}_{X,y\sim\mathcal{D}}[g_i(X, y)^2].$$

Note that this version of the Fisher Information relies on the same gradients that are used to update the parameters of the multilayer perceptron, see equations (3), (4).

Under the assumption that the learned probability distribution $f(\cdot \mid X, \theta)$ equals the real probability distribution, the Fisher Information equals the Hessian of the cross entropy loss

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(i.e. the negative log-probabilities) with respect to the model parameters (Pascanu and Bengio, 2013; Martens, 2014). The Fisher Information was previously used in machine learning to enable lifelong learning (Kirkpatrick et al., 2017; Huszár, 2018) and it has been shown that other popular lifelong learning methods implicitly rely on the Fisher Information (Benzing, 2020).

5.1.2 Link between release probabilities and Fisher Information

We now explain how our learning rule for the release probability is related to the Fisher Information. For simplicity of exposition, we focus our analysis on a particular sampled subnetwork with deterministic synaptic strengths. Recall that update rule (4) for release probabilities increases the release probability, if the gradient magnitude $|g_i|$ is above a certain threshold, $g_i > |g_{\text{lim}}|$, and decreases them otherwise. Let us denote by p_i^+ the probability that the i-th release probability is increased. Then

$$p_i^+ := \Pr[|g_i| > g_{\text{lim}}] = \Pr[g_i^2 > g_{\text{lim}}^2], \tag{6}$$

where the probability space corresponds to sampling training examples. Note that $\mathbb{E}[g_i^2] = F_i$ by definition of the Empirical Fisher Information F_i. So if we assume that $\Pr[g_i^2 > g_{\text{lim}}^2]$ depends monotonically on $\mathbb{E}[g_i^2]$, then we already see that p_i^+ depends monotonically on F_i. This in turn implies that synapses with a larger Fisher Information are more likely to have a large release probability, which is what we claimed. We now discuss the assumption made above.

Assumption: $\Pr[g_i^2 > g_{\text{lim}}^2]$ depends monotonically on $\mathbb{E}[g_i^2]$. While this assumption is not true for arbitrary distributions of g, it holds for many commonly studied parametric families and seems likely to hold (approximately) for realistic, non-adversarially chosen distributions. For example, if each g_i follows a normal distribution $g_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ with varying σ_i and $\sigma_i \gg \mu_i$, then

$$F_i = \mathbb{E}[g_i^2] \approx \sigma_i^2$$

and

$$p_i^+ = \Pr[g_i^2 > g_{\text{lim}}^2] \approx \operatorname{erfc}\left(\frac{g_{\text{lim}}}{\sigma_i\sqrt{2}}\right)$$

so that p_i^+ is indeed monotonically increasing in F_i. Similar arguments can be made for example for a Laplace distribution, with scale larger than mean.

Fisher Information and Learning Rate Modulation Note that we multiply the learning rate of each synapse by $(1 - p_i)$, see (5). Since synapses with large Fisher Information tend to have large release probability, the plasticity of synapses with large Fisher Information, and thus low posterior variance, is decreased as would also be prescribed by Bayesian modelling.

Practical Approximation The derivation above assumes that each gradient g is computed using a single input, so that $\mathbb{E}[g^2]$ equals the Fisher Information. While this may be the biologically more plausible setting, in standard ANN training the gradient is averaged across several inputs (mini-batches). Despite this modification, g^2 remains a good, and commonly used, approximation of the Fisher, see e.g. (Khan et al., 2018; Benzing, 2020).

5.2 Perceptron for Lifelong Learning

To demonstrate that our findings on presynaptic stochasticity and plasticity are applicable to other models and learning rules, we include experiments for the standard perceptron (Rosenblatt, 1958) in a lifelong learning setting.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5.2.1 Model

The perceptron is a classical model for a neuron with multiple inputs and threshold activation function. It is used to memorise the binary labels of a number of input patterns where input patterns are sampled uniformly from $\{-1, 1\}^N$ and their labels are sampled uniformly from $\{-1, 1\}$. Like in ANNs, the output neuron of a perceptron computes a weighted sum of its inputs followed by nonlinear activation $\sigma(\cdot)$:

$$\underbrace{a^{\text{post}}}_{\substack{\text{postsynaptic}\\ \text{activation}}} = \sigma\left(\sum_{i=1}^{n} w_i \underbrace{a_i^{\text{pre}}}_{\substack{i\text{-th presynaptic}\\ \text{input}}}\right). \tag{7}$$

The only difference to the ANN model is that the nonlinearity is the sign function and that there is only one layer. We model each synapse w_i as a Bernoulli variable r_i with synaptic strength m_i and release probability p_i just as before, see equation (1). The expected strengths $\bar{w}_i$ are learned according to the standard perceptron learning rule (Rosenblatt, 1958). The only modification we make is averaging weight updates across 5 inputs, rather than applying an update after each input. Without this modification, the update size g_i for each weight w_i would be constant according to the perceptron learning rule. Consequently, our update rule for p_i would not be applicable. However, after averaging across 5 patterns we can apply the same update rule for p_i as previously, see equation (4), and also use the same learning rate modification, see equation (5). We clarify that g_i now refers to the update of expected strength $\bar{w}_i$. In the case of ANN this is proportional to the gradient, while in the case of the non-differentiable perceptron it has no additional interpretation.

5.2.2 Experiments

For the lifelong learning experiments, we used 5 tasks, each consisting of 100 randomly sampled and labelled patterns of size $N = 1000$. We compared the perceptron with learned stochastic weights to a standard perceptron. For the standard perceptron, we also averaged updates across 5 patterns. Both models were sequentially trained on 5 tasks, using 25 passes through the data for each task.

We note that for more patterns, when the perceptron gets closer to its maximum capacity of $2N$, the average accuracies of the stochastic and standard perceptron become more similar, suggesting that the benefits of stochastic synapses occur when model capacity is not fully used.

As metaplasticity parameters we used $g_{\text{lim}} = 0.1, p_{\text{up}} = p_{\text{down}} = 0.2$ and $p_{\text{min}} = 0.25, p_{\text{freeze}} = 0.9$. These were coarsely tuned on an analogous experiment with only two tasks instead of five.

5.3 Experimental Setup

5.3.1 Code Availability

Code for all experiments will be made publicly available upon publication.

5.3.2 Metaplasticity Parameters

Our method has a number of metaplasticity parameters, namely p_{up}, p_{down}, g_{lim} and the learning rate η. For the lifelong learning experiments, there is an additional parameter p_{freeze} and the learning rate $\eta_i(p_i)$ becomes dependent on the release probability with the additional parameter η_0. We confirmed that our results are robust to the exact choice of these parameters.

For the energy experiments we fix $p_{\text{up}} = p_{\text{down}} = 0.07$, $g_{\text{lim}} = 0.001$ and choose $\eta = 0.05$. For the lifelong learning experiments we choose $\eta_0 \in \{0.01, 0.001\}$ and optimise the remaining metaplasticity parameters through a random search on one task, namely Permuted MNIST. We

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

then evaluate them on various other tasks, namely Permuted Fashion MNIST, Split MNIST and Split Fashion MNIST (see below for detailed task descriptions) to ensure that the parameters are not overfitted to a specific task. The results presented in Table 1 suggest that the parameters found in one scenario are robust and do carry over to several other scenarios. For the ablation experiment in Figure 5a, metaplasticity parameters were re-optimised for each ablation in a random search to ensure a fair comparison.

5.3.3 Lifelong Learning Tasks

For the lifelong learning experiments we tested our method as well as baselines in several scenarios on top of the Split MNIST protocol described in the main text.

Permuted MNIST In the Permuted MNIST benchmark, each task consists of a random but fixed permutation of the input pixels of all MNIST images (Goodfellow et al., 2013). We generate 10 tasks using this procedure and present them sequentially without any indication of task boundaries during training. A main reason to consider the Permuted MNIST protocol is that it generates tasks of equal difficulty.

Permuted & Split Fashion MNIST Both the Split and Permuted protocol can be applied to other datasets. We use them on the Fashion MNIST dataset (Xiao et al., 2017) consisting of $60,000$ greyscale images of 10 different fashion items with 28×28 pixels.

Continuous Permuted MNIST We carry out an additional experiment on the continuous Permuted MNIST dataset (Zeno et al., 2018). This is a modified version of the Permuted MNIST dataset which introduces a smooth transition period between individual tasks where data from both distributions is mixed. It removes the abrupt change between tasks and allows us to investigate if our method depends on such an implicit task switch signal. We observe a mean accuracy over all tasks of 0.8539 ± 0.006 comparable to the non-continuous case suggesting that our method does not require abrupt changes from one task to another.

5.3.4 Neural Network Training

Our neural network architecture consists of two fully connected hidden layers of 200 neurons without biases with rectified linear unit activation functions $\sigma(x)$. The final layer uses a softmax and cross-entropy loss. Network weights were initialised according to the PyTorch default for fully connected layers, which is similar to Kaiming uniform initialisation (Glorot and Bengio, 2010; He et al., 2015) but divides weights by an additional factor of $\sqrt{6}$. We use standard stochastic gradient descent to update the average weight $\bar{w}_i$ only altered by the learning rate modulation described for the lifelong learning experiments. We use a batch size of 100 and train each task for 10 epochs in the lifelong learning setting. In the energy-information experiments we train the model for 50 epochs.

6 Acknowledgements

S.S. was supported by grant PZ00P3186027 from the Swiss National Science Foundation. A.S. and F.B. were supported by grant CRSII5_173721 by the Swiss National Science Foundation. We would like to thank João Sacramento and Mark van Rossum for stimulating discussions and helpful comments.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7 Author Contributions

A.S., F.B., and S.S. designed the project and developed the main methodology. S.S. and F.B. performed the experiments, carried out the formal analysis, and wrote the original draft. All authors reviewed the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

8 Supplementary Material

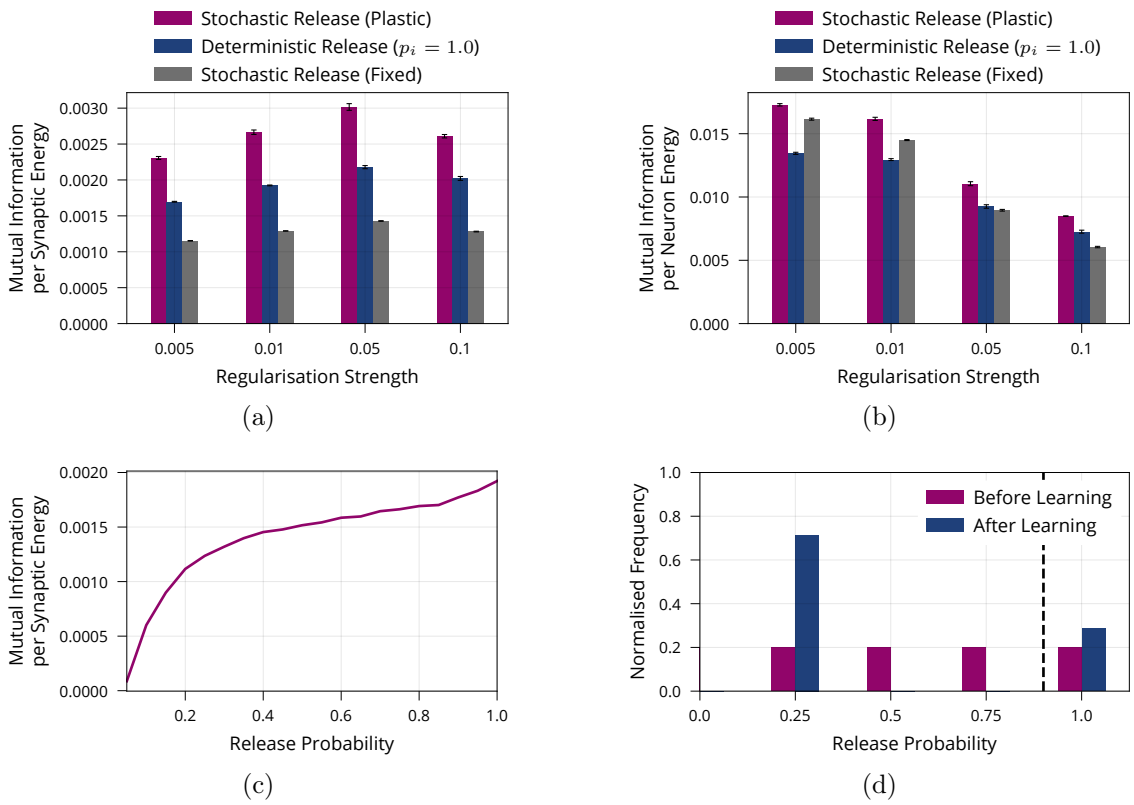
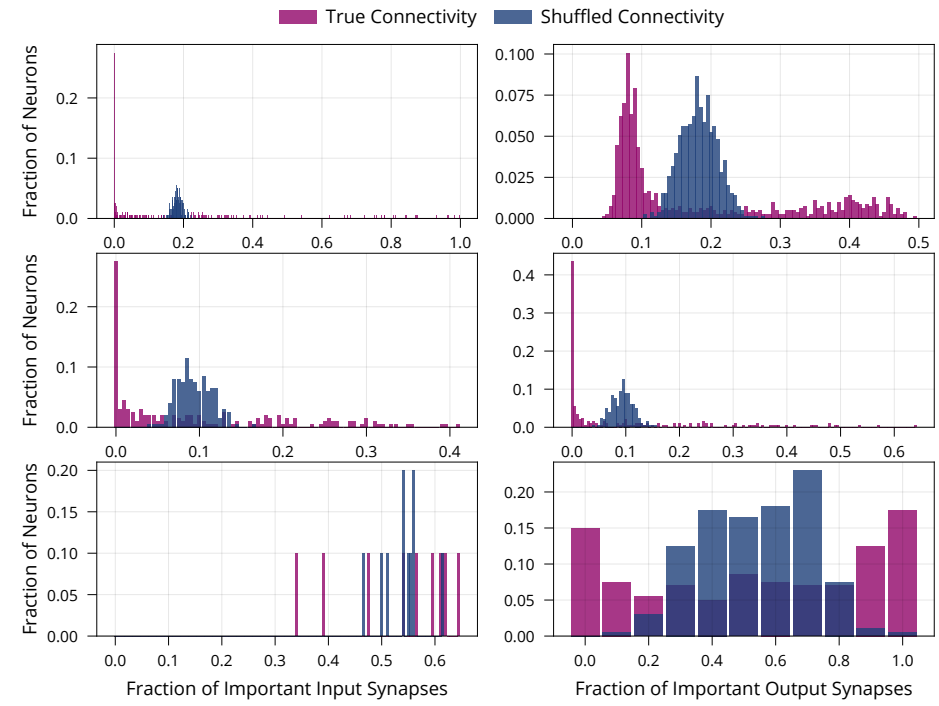
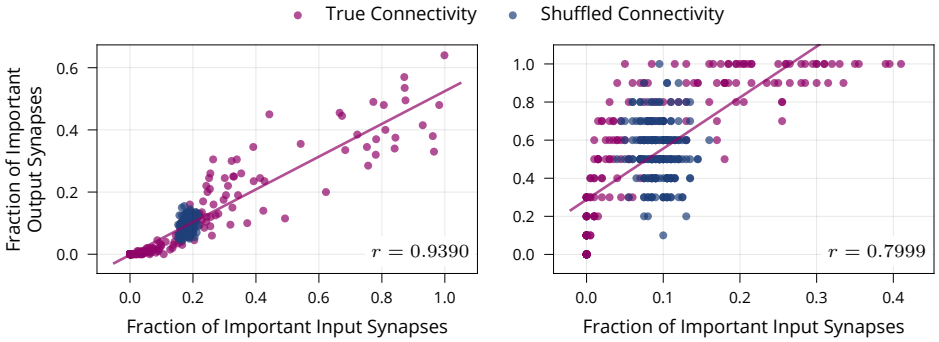


Figure 6: **Additional Results on Energy Efficiency of Model with Stochastic and Plastic Release.** **(a)** Mutual information per energy analogous to Figure 1b, but showing results for different regularisation strengths rather than the best result for each model. As described in the main part, energy is measured via its synaptic contribution. **(b)** Same experiment as in (a) but energy is measured as the metabolic cost incurred by the activity of neurons by calculating their average rate of activity. **(c)** Maximum mutual information per energy for a multilayer perceptron with fixed release probability and constant regularisation strength of 0.01. This is the same model as "Stochastic Release (Fixed)" in (a), but for a range of different values for the release probability. This is in line with the single synapse analysis in Harris et al. (2012). For each model, we searched over different learning rates and report the best result. **(d)** Analogous to Figure 2a, but release probabilities were initialised independently, uniformly at random in the interval $[0.25, 1]$ rather than with a fixed value of 0.25. Error bars in (a) and (b) denote the standard error for three repetitions of the experiment. (c) shows the best performing model for each release probability after a grid search over the learning rate. (d) shows aggregated data over three repetitions of the experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a)



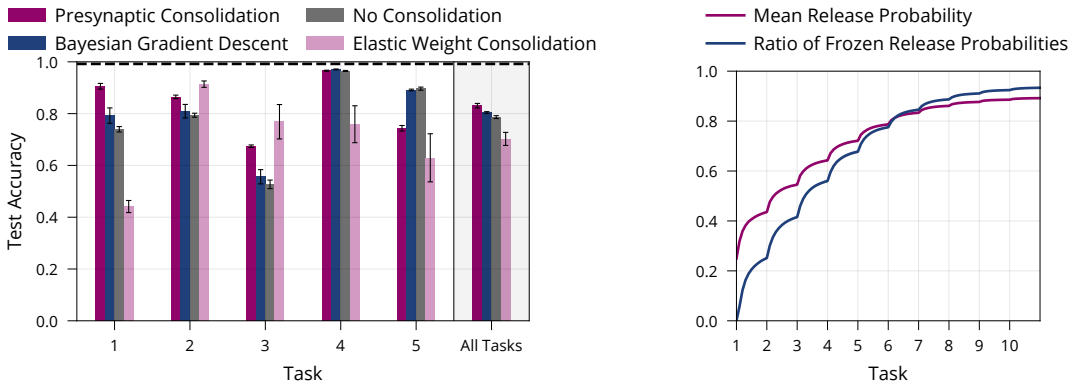
(b)

Figure 7: **Additional Results on Neuron-Level Sparsity of Network after Learning.** **(a)** Number of important synapses per neuron for all layers after learning on MNIST. The i-th row shows data from the i-th weight matrix of the network and we compare true connectivity to random connectivity. Two-sample Kolmogorov-Smirnov tests comparing the distribution of important synapses in the shuffled and unaltered condition are significant for all layers ($p < 0.01$) except for the output neurons in the last layer (lower-left panel) ($p = 0.41$). This is to be expected as all 10 output neurons in the last layer should be equally active and thus receive similar numbers of active inputs. **(b)** Scatter plot showing the number of important input and output synapses per neuron for both hidden layers after learning on MNIST. First hidden layer (left) has a Pearson correlation coefficient of $r = 0.9390$. Second hidden layer (right) has a Pearson correlation coefficient of $r = 0.7999$. Data is from one run of the experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: **Lifelong Learning Comparison on Additional Datasets.** Average test accuracies (%, higher is better, average over all sequentially presented tasks) and standard errors for three repetitions of each experiment on four different lifelong learning tasks for the Presynaptic Consolidation mechanism, BGD (Zeno et al., 2018) and EWC (Kirkpatrick et al., 2017).

	Split MNIST	Split Fashion	Perm. MNIST	Perm. Fashion
Presynaptic Consolidation	$82.90^{\pm 0.01}$	$91.98^{\pm 0.12}$	$86.14^{\pm 0.67}$	$75.92^{\pm 0.37}$
No Consolidation	$77.68^{\pm 0.31}$	$88.76^{\pm 0.45}$	$79.60^{\pm 0.43}$	$72.13^{\pm 0.75}$
Bayesian Gradient Descent	$80.44^{\pm 0.45}$	$89.54^{\pm 0.88}$	$89.73^{\pm 0.52}$	$78.45^{\pm 0.15}$
Elastic Weight Consolidation	$70.41^{\pm 4.20}$	$76.89^{\pm 1.05}$	$89.58^{\pm 0.53}$	$77.44^{\pm 0.41}$
Joint Training	$98.55^{\pm 0.10}$	$97.67^{\pm 0.09}$	$97.33^{\pm 0.08}$	$87.33^{\pm 0.07}$



(a) (b)

Figure 8: **Additional Results on Lifelong Learning in a Model with Presynaptically Driven Consolidation.** **(a)** Detailed lifelong-learning results of various methods on Split MNIST, same underlying experiment as in Figure 4c. We report the test accuracy on each task of the final model (after learning all tasks). Error bars denote the standard error for three repetitions of the experiment. **(b)** Mean release probability and percentage of frozen weights over the course of learning ten permuted MNIST tasks. Error bars in (a) and shaded regions in (b) show standard error over three repetitions of the experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

J. del Castillo and B. Katz. Quantal components of the end-plate potential. *The Journal of Physiology*, 124(3):560–573, 1954. doi:10.1113/jphysiol.1954.sp005129. URL `https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1954.sp005129`.

Tiago Branco and Kevin Staras. The probability of neurotransmitter release: variability and feedback control at single synapses. *Nature Reviews Neuroscience*, 10(5):373–383, 2009. doi:10.1038/nrn2634.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x`.

A Aldo Faisal, John A White, and Simon B Laughlin. Ion-channel noise places limits on the miniaturization of the brain's wiring. *Current Biology*, 15(12):1143–1149, 2005.

Jeremy E Niven and Simon B Laughlin. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008. doi:10.1242/jeb.017574.

David Attwell and Simon B Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10):1133–1145, 2001. doi:10.1097/00004647-200110000-00001.

Julia J. Harris, Renaud Jolivet, and David Attwell. Synaptic Energy Use and Supply. *Neuron*, 75(5):762–777, 2012. ISSN 0896-6273. doi:10.1016/j.neuron.2012.08.019. URL `http://www.sciencedirect.com/science/article/pii/S0896627312007568`.

Milton Llera-Montero, João Sacramento, and Rui Ponte Costa. Computational roles of plastic probabilistic synapses. *Current Opinion in Neurobiology*, 54:90 – 97, 2019. ISSN 0959-4388. doi:https://doi.org/10.1016/j.conb.2018.09.002. URL `http://www.sciencedirect.com/science/article/pii/S0959438818301028`. Neurobiology of Learning and Plasticity.

William B Levy and Robert A Baxter. Energy efficient neural codes. *Neural computation*, 8 (3):531–543, 1996. doi:10.1162/neco.1996.8.3.531.

William B Levy and Robert A Baxter. Energy-efficient neuronal computation via quantal synaptic failures. *Journal of Neuroscience*, 22(11):4746–4755, 2002. doi:10.1523/JNEUROSCI.22-11-04746.2002.

Biswa Sengupta, Simon B Laughlin, and Jeremy E Niven. Balanced excitatory and inhibitory synaptic currents promote efficient coding and metabolic efficiency. *PLoS Comput Biol*, 9 (10):e1003263, 2013. doi:10.1371/journal.pcbi.1003263.

Leonid P Savtchenko, Sergiy Sylantyev, and Dmitri A Rusakov. Central synapses release a resource-efficient amount of glutamate. *Nature neuroscience*, 16(1):10–12, 2013. doi:10.1038/nn.3285.

Julia J Harris, Renaud Jolivet, Elisabeth Engl, and David Attwell. Energy-efficient information transfer by visual pathway synapses. *Current Biology*, 25(24):3151–3160, 2015. doi:10.1016/j.cub.2015.10.063.

Julia Jade Harris, Elisabeth Engl, David Attwell, and Renaud Blaise Jolivet. Energy-efficient information transfer at thalamocortical synapses. *PLoS computational biology*, 15(8):e1007226, 2019. doi:10.1371/journal.pcbi.1007226.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. doi:10.1073/pnas.1611835114.

Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924, 2009. doi:10.1038/nature08577.

Akiko Hayashi-Takagi, Sho Yagishita, Mayumi Nakamura, Fukutoshi Shirai, Yi I Wu, Amanda L Loshbaugh, Brian Kuhlman, Klaus M Hahn, and Haruo Kasai. Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, 525(7569):333–338, 2015. doi:10.1038/nature15257.

Neil R Hardingham, Jenny CA Read, Andrew J Trevelyan, J Charmaine Nelson, J Julian B Jack, and Neil J Bannister. Quantal analysis reveals a functional correlation between presynaptic and postsynaptic efficacy in excitatory connections from rat neocortex. *Journal of Neuroscience*, 30(4):1441–1451, 2010. doi:10.1523/JNEUROSCI.3244-09.2010.

Hirokazu Sakamoto, Tetsuroh Ariyoshi, Naoya Kimpara, Kohtaroh Sugao, Isamu Taiko, Kenji Takikawa, Daisuke Asanuma, Shigeyuki Namiki, and Kenzo Hirose. Synaptic weight set by munc13-1 supramolecular assemblies. *Nature neuroscience*, 21(1):41–49, 2018. doi:10.1038/s41593-017-0041-9.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015. doi:10.1146/annurev-vision-082114-035447.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016. doi:10.1038/nn.4244.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018. doi:10.1016/j.neuron.2018.03.044.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018. doi:10.1038/s41586-018-0102-6.

Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*, 2018. URL `http://arxiv.org/abs/1803.07770`.

Marcelo G Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018. doi:10.1038/s41593-018-0232-z.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. doi:10.1037/h0042519.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Joseph Cichon and Wen-Biao Gan. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180–185, 2015. doi:10.1038/nature14251.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv preprint arXiv:1803.10123*, 2018. URL `http://arxiv.org/abs/1803.10123`.

Tiago Branco, Kevin Staras, Kevin J Darcy, and Yukiko Goda. Local dendritic activity sets release probability at hippocampal synapses. *Neuron*, 59(3):475–485, 2008. doi:10.1016/j.neuron.2008.07.006.

Thomas M Sanderson, John Georgiou, and Graham L Collingridge. Illuminating relationships between the pre-and post-synapse. *Frontiers in neural circuits*, 14, 2020.

Joao Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*, 31:8721–8732, 2018.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):1–10, 2016. doi:10.1038/ncomms13276.

Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer, 2015. doi:10.1007/978-3-319-23528-8_31.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006. doi:10.1038/nn1790.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3): 119–130, 2010. doi:10.1016/j.tics.2010.01.003.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network plasticity as bayesian inference. *PLoS Comput Biol*, 11(11):e1004485, 2015. doi:10.1371/journal.pcbi.1004485.

Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016. doi:10.1016/j.neuron.2016.03.020.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066, 2013.

Laurence Aitchison, Alex Pouget, and Peter E Latham. Probabilistic synapses. *arXiv preprint arXiv:1410.1029*, 2014. URL `http://arxiv.org/abs/1410.1029`.

Laurence Aitchison and Peter E. Latham. Synaptic sampling: A connection between PSP variability and uncertainty explains neurophysiological observations. *arXiv preprint arXiv:1505.04544*, 2015. URL `http://arxiv.org/abs/1505.04544`.

Laurence Aitchison, Jannes Jegminat, Jorge Aurelio Menendez, Jean-Pascal Pfister, Alexandre Pouget, and Peter E Latham. Synaptic plasticity as bayesian inference. *Nature Neuroscience*, pages 1–7, 2021.

Per Jesper Sjöström, Gina G Turrigiano, and Sacha B Nelson. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164, 2001. doi:10.1016/S0896-6273(01)00542-6.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Per Jesper Sjöström, Gina G Turrigiano, and Sacha B Nelson. Multiple forms of long-term plasticity at unitary neocortical layer 5 synapses. *Neuropharmacology*, 52(1):176–184, 2007. doi:10.1016/j.neuropharm.2006.07.021.

Rui Ponte Costa, Zahid Padamsey, James A D'Amour, Nigel J Emptage, Robert C Froemke, and Tim P Vogels. Synaptic transmission optimization predicts expression loci of long-term plasticity. *Neuron*, 96(1):177–189, 2017. doi:10.1016/j.neuron.2017.09.021.

Karin Isler and Carel P van Schaik. The expensive brain: a framework for explaining evolutionary changes in brain size. *Journal of Human Evolution*, 57(4):392–400, 2009. doi:10.1016/j.jhevol.2009.04.009.

Ana Navarrete, Carel P van Schaik, and Karin Isler. Energetics and the evolution of human brain size. *Nature*, 480(7375):91–93, 2011. doi:10.1038/nature10629.

Beth L Chen, David H Hall, and Dmitri B Chklovskii. Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences*, 103(12):4723–4728, 2006. doi:10.1073/pnas.0506806103.

Henrik Alle, Arnd Roth, and Jörg RP Geiger. Energy-efficient action potentials in hippocampal mossy fibers. *Science*, 325(5946):1405–1408, 2009. doi:10.1126/science.1174331.

Biswa Sengupta, Martin Stemmler, Simon B Laughlin, and Jeremy E Niven. Action potential energy efficiency varies among neuron types in vertebrates and invertebrates. *PLoS Comput Biol*, 6(7):e1000840, 2010. doi:10.1371/journal.pcbi.1000840.

János A Perge, Kristin Koch, Robert Miller, Peter Sterling, and Vijay Balasubramanian. How the optic nerve allocates space, energy capacity, and information. *Journal of Neuroscience*, 29(24):7917–7928, 2009. doi:10.1523/JNEUROSCI.5200-08.2009.

Brett C Carter and Bruce P Bean. Sodium entry during action potentials of mammalian neurons: incomplete inactivation and reduced metabolic efficiency in fast-spiking neurons. *Neuron*, 64(6):898–909, 2009. doi:10.1016/j.neuron.2009.12.011.

Hua Hu and Peter Jonas. A supercritical density of na+ channels ensures fast signaling in gabaergic interneuron axons. *Nature neuroscience*, 17(5):686–693, 2014. doi:10.1038/nn.3678.

Ken A Morris, Qing Chang, Eric G Mohler, and Paul E Gold. Age-related memory impairments due to reduced blood glucose responses to epinephrine. *Neurobiology of aging*, 31(12):2136–2145, 2010. doi:10.1016/j.neurobiolaging.2008.12.003.

Akinobu Suzuki, Sarah A Stern, Ozlem Bozdagi, George W Huntley, Ruth H Walker, Pierre J Magistretti, and Cristina M Alberini. Astrocyte-neuron lactate transport is required for long-term memory formation. *Cell*, 144(5):810–823, 2011. doi:10.1016/j.cell.2011.02.018.

Etienne Hebert-Chatelain, Tifany Desprez, Román Serrat, Luigi Bellocchio, Edgar Soria-Gomez, Arnau Busquets-Garcia, Antonio Christian Pagano Zottola, Anna Delamarre, Astrid Cannich, Peggy Vincent, et al. A cannabinoid link between mitochondria and memory. *Nature*, 539(7630):555–559, 2016. doi:10.1038/nature20127.

Cristina M Alberini, Emmanuel Cruz, Giannina Descalzi, Benjamin Bessières, and Virginia Gao. Astrocyte glycogen and lactate: New insights into learning and memory mechanisms. *Glia*, 66(6):1244–1262, 2018. doi:10.1002/glia.23250.

Ho Ling Li and Mark Cw Van Rossum. Energy efficient synaptic plasticity. *Elife*, 9:e50804, 2020. doi:10.7554/eLife.50804.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Maxim V Ivannikov, Mutsuyuki Sugimori, and Rodolfo R Llinás. Synaptic vesicle exocytosis in hippocampal synaptosomes correlates directly with total mitochondrial volume. *Journal of Molecular Neuroscience*, 49(1):223–230, 2013. doi:10.1007/s12031-012-9848-8.

Diane TW Chang, Anthony S Honick, and Ian J Reynolds. Mitochondrial trafficking to synapses in cultured primary cortical neurons. *Journal of Neuroscience*, 26(26):7035–7045, 2006. doi:10.1523/JNEUROSCI.1012-06.2006.

Kazuki Obashi and Shigeo Okabe. Regulation of mitochondrial dynamics and distribution by synapse position and neuronal activity in the axon. *European Journal of Neuroscience*, 38 (3):2350–2363, 2013. doi:10.1111/ejn.12263.

Tao Sun, Haifa Qiao, Ping-Yue Pan, Yanmin Chen, and Zu-Hang Sheng. Motile axonal mitochondria contribute to the variability of presynaptic strength. *Cell reports*, 4(3):413–419, 2013. doi:10.1016/j.celrep.2013.06.040.

Robert M Lees, James D Johnson, and Michael C Ashby. Presynaptic boutons that contain mitochondria are more stable. *Frontiers in synaptic neuroscience*, 11:37, 2020. doi:10.3389/fnsyn.2019.00037.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004. doi:10.1016/j.conb.2004.07.007.

David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969. doi:10.1038/222960a0.

Hannah R Monday, Thomas J Younts, and Pablo E Castillo. Long-term plasticity of neurotransmitter release: emerging mechanisms and contributions to brain function and disease. *Annual review of neuroscience*, 41:299–322, 2018. doi:10.1146/annurev-neuro-080317-062155.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. URL `http://arxiv.org/abs/1301.3584`.

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014. URL `http://arxiv.org/abs/1412.1193`.

Ferenc Huszár. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, page 201717042, 2018. doi:10.1073/pnas.1717042115.

Frederik Benzing. Understanding regularisation methods for continual learning. *arXiv preprint arXiv:2006.06357*, 2020. URL `http://arxiv.org/abs/2006.06357`.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. *arXiv preprint arXiv:1806.04854*, 2018. URL `http://arxiv.org/abs/1806.04854`.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. URL `http://arxiv.org/abs/1312.6211`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL `http://arxiv.org/abs/1708.07747`.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.05.442708; this version posted May 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. doi:10.1109/ICCV.2015.123.