Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules

Anonymous Author(s)
Affiliation
Address
email

Abstract

To unveil how the brain learns, ongoing work seeks biologically-plausible approximations of gradient descent algorithms for training recurrent neural networks (RNNs). Yet, beyond task accuracy, it is unclear if such learning rules converge to solutions that exhibit different levels of generalization than their non-biologically-plausible counterparts. Leveraging results from deep learning theory based on loss landscape curvature, we ask: how do biologically-plausible gradient approximations affect generalization? We first demonstrate that state-of-the-art biologically-plausible learning rules for training RNNs exhibit worse and more variable generalization performance compared to their machine learning counterparts that follow the true gradient more closely. Next, we verify that such generalization performance is correlated significantly with loss landscape curvature, and we show that biologically-plausible learning rules tend to approach high-curvature regions in synaptic weight space. Using tools from dynamical systems, we derive theoretical arguments and present a theorem explaining this phenomenon. This predicts our numerical results, and explains why biologically-plausible rules lead to worse and more variable generalization properties. Finally, we suggest potential remedies that could be used by the brain to mitigate this effect. To our knowledge, our analysis is the first to identify the reason for this generalization gap between artificial and biologically-plausible learning rules, which can help guide future investigations into how the brain learns solutions that generalize.

1 Introduction

A longstanding question in neuroscience is how animals excel at learning complex behavior involving temporal dependencies across multiple timescales and thereafter generalize this learned behavior to unseen data. This requires solving the temporal credit assignment problem: how to assign the contribution of past neural states to future outcomes. To address this, neuroscientists are increasingly turning to the mathematical framework provided by recurrent neural networks (RNNs) to model learning mechanisms in the brain [1, 2]. Temporal credit assignment in RNNs is typically achieved by backpropagation through time (BPTT), or other gradient descent-based optimization algorithms, none of which are biologically-plausible (or bio-plausible for short). Therefore, the use of RNNs as a framework to understand the computational principles of learning in the brain has motivated an influx of bio-plausible learning rules that approximate gradient descent [1–3].

Performance of such rules is typically quantified by accuracy. Although the accuracy achieved by these rules is often comparable to true gradient descent, little is known about the breadth of the emergent solutions, namely how well they generalize. Broadly speaking, generalization refers to a trained model's ability to adapt to previously unseen data, and is typically measured by the so-called *generalization gap*: the difference between training and testing error. This is especially

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

important when learning complex tasks with nonlinear RNNs where the loss landscape is non-convex, and therefore, many solutions with comparable training accuracy can exist. These solutions, characterized as (local) minima in the loss landscape, can nonetheless exhibit drastically different levels of generalization (Figure 1). It is not clear if gradient-based methods like BPTT and the existing bio-plausible alternatives have a different tendency to converge to loss minima that provide better or worse generalization.

While better predictors of the generalization performance remains an open research question in deep learning [4], recent extensive studies identify flatness of the loss landscape at the solution point as a promising predictor for generalization [5–9]. Leveraging these empirical and theoretical findings, **we ask**: how do proposed biologically-motivated gradient approximations affect flatness of the converged solution on loss landscape, and thereby, generalization?

Our overarching goal is to investigate generalization trends for existing bio-plausible temporal credit assignment rules, and in particular, examine how truncation-based bio-plausible gradient approximations can affect such trends. Specifically, **our contributions** are summarized as follows:

- In numerical experiments, we demonstrate across several well-known neuroscience and machine learning benchmark tasks that state-of-the-art (SoTA) bio-plausible learning rules for training RNNs exhibit *worse* and *more variable* generalization gaps, compared to true (stochastic) gradient descent (Figure 2A-C).
- Using the same experiments, we show that bio-plausible learning rules tend to approach high-curvature regions in synaptic weight space as measured by the loss' Hessian eigenspectrum (Figure 2D-F). Further, we verify that this correlates with worse generalization (Figure 2D-F and 3), which is consistent with the literature.
- We present a theorem to explain this phenomenon by examining the weight update equation as a discrete dynamical system, which sheds light on a potential connection between gradient alignment and the preference over converging to narrower minima (Theorem 1, Figure 4).

Given the core components in designing artificial neural networks: data, objective functions, learning rules and architectures [10], we investigate different learning rules while holding the data, objective function and architecture constant. SoTA RNN learning rules investigated include a three-factor rule with symmetric feedback (symmetric e-prop [11]), a three-factor rule using random feedback weights (RFLO [12]) and a multi-factor rule using local modulatory signaling (MDGL) [13]. To our knowledge, our analysis is the first to highlight and study this gap in solution quality between artificial and bio-plausible learning rules for RNNs, thereby motivating further investigations into how the brain learns solutions that generalize.

2 Related works

2.1 Bio-plausible gradient approximations

Investigating bio-plausible learning rules is of interest both to identify more efficient training strategies for artificial networks and to better understand learning in the brain [2, 3, 14–42]. In order for learning to reach a certain goal quantified by an objective, learning algorithms often minimize a loss function [10]. The error gradient, if available, tells us how each parameter should be adjusted in order to have the steepest local descent. For training RNNs, which are widely used as a model for neural circuits [43–62], standard algorithms that follow this gradient — real time recurrent learning (RTRL) and BPTT — are not bio-plausible and have overwhelming memory storage demands [3, 63]. However, learning rules that only approximate the true gradient can sometimes be as effective as those that follow the gradient exactly [10, 64]. Because of that, bio-plausible learning rules that approximate the gradient using known properties of real neurons have been proposed and led to successful learning outcomes across many different tasks in feedforward networks [1, 65–74], with recent extensions to recurrently connected networks [11–13]. These existing bio-plausible rules for RNNs update rules estimate the gradient using known biological learning ingredients: eligibility traces, which maintain preceding activity on the molecular levels [75–80], combined with top-down instructive signaling [75, 76, 81–87] as well as local cell-to-cell modulatory signaling within the network [13, 88, 89]. For efficient online learning in RNNs, other approximations (not necessarily bio-plausible) to RTRL [90–95] have also demonstrated to produce good performance. Given the impressive accuracy achieved

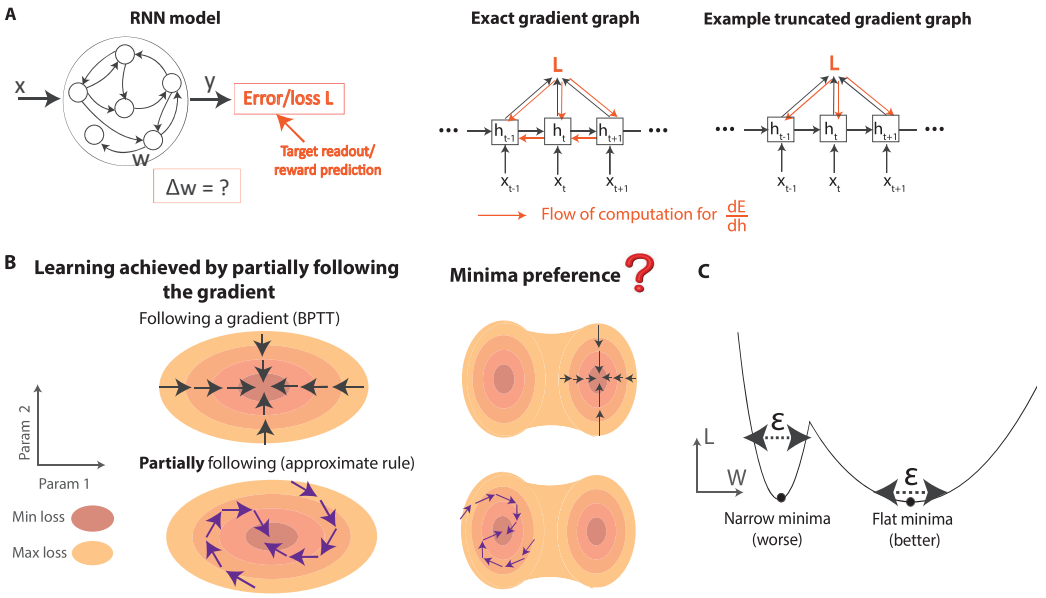

Figure 1: **Setup**. A) Illustration of an RNN trained to minimize error/loss function L (left). Existing bio-plausible proposals for RNNs estimate the gradient by neglecting dependencies that are biologically implausible to compute (right). B) Low training error/loss can be achieved by *partially* following a gradient (right), but preference for converging to minima with certain generalization properties remains underexamined for these learning rules (left). C) Minima flatness matters: 1-D loss landscape illustration with two solutions that equally minimize loss L, but exhibit drastically different generalization properties: the narrower minima is more sensitive to perturbation.

by these approximate rules, several studies began to investigate their convergence properties [96], e.g. for random backpropagation weights in feedforward networks [97, 98]. However, the trend of generalization capabilities of these rules, especially in RNNs, is underexamined.

2.2 Loss landscape curvature and generalization performance

Given the central importance of understanding how neural networks perform in situations unseen during training [4, 99–104], the deep learning community has made tremendous efforts to develop tools for understanding generalization that we leverage here. That flat minima could lead to better generalization was observed more than two decades ago [105]. Intuitively, under the same amount of perturbation ϵ in parameter space (e.g. loss landscape changes due to addition of new data) worse performance degradation will be seen around the narrower minima (see Figure 1C). We note that perturbations in parameter space can be linked to that in the input space [6]. Recently, many empirical studies have consistently supported the usefulness of this predictor [106–113]. In particular, the authors of [5] performed an extensive study and found that flatness-based measures have a higher correlation with generalization than other alternatives. Motivated by this, several studies have characterized properties of the loss functions's Hessian — whose eigenspectrum carries information about curvature [114–118]. Connections between flatness and generalization performance have shed light on the reason for greater generalization gaps in large batch training [109, 119–122], and also have inspired optimization methods to favor flatter minima [106, 107, 123–130]. Despite the criticism of scale-dependence of flatness [131], where parameter rescaling can drastically change flatness but not always generalization quality, flatness — with parameter scales taken into account [132, 133] — are connected to PAC-Bayesian generalization error bounds [7–9]. Moreover, a recent theoretical study rigorously connects flatness of the loss surface to generalization in classification tasks under the assumption that labels are (approximately) locally constant [6]. Leveraging the great progress from deep learning community, we aim to study generalization properties of bio-plausible learning rules from a geometric perspective.

3 Results

In this section, we first describe the network and learning setup we use (Figure 1A). Next, we present a number of numerical experiments where we compute generalization gap directly on three commonly used ML and neuroscience tasks (Figure 2A-C), for truncation-based bio-plausible gradient approximations. For the same experiments, we also quantify loss landscape curvature along learning trajectories, and connect these quantities to generalization behavior (Figures 2D-F and 3). Finally, we provide theoretical arguments and a theorem that explains how gradient alignment in bio-plausible gradient approximations can affect curvature preference (Theorem 1, Figure 3) and thus, generalization. Through additional experiments, we verify the predictive power of our theory. We conclude with discussions on potential remedies used by the brain (Appendix Figure 5) and future directions.

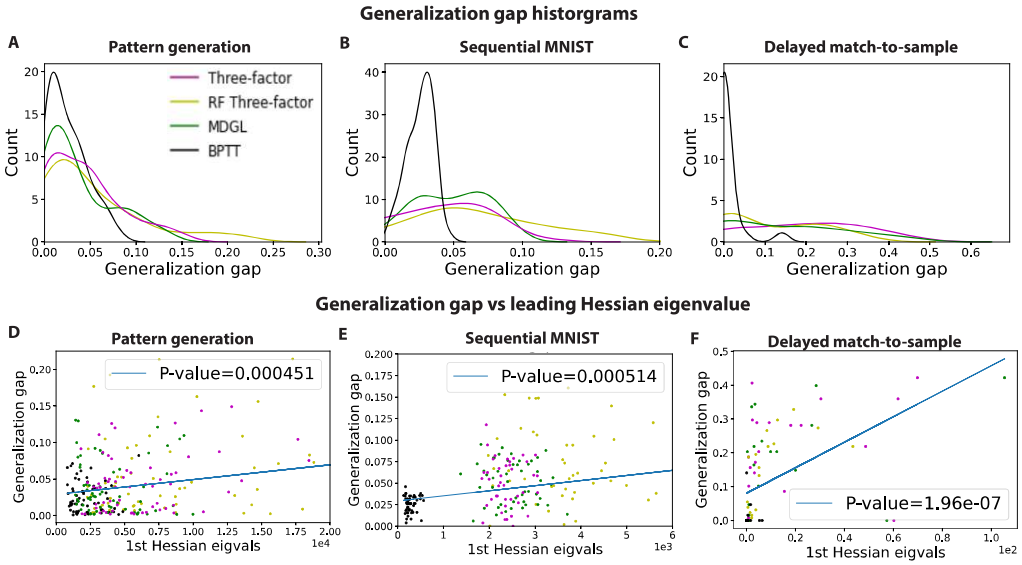

Figure 2: **Bio-plausible temporal credit assignment rules show worse and more variable generalization gap, which can be informed by loss landscape curvature**. A-C) Generalization gap distributions computed at *the end of training* across different random initializations for several well-known neuroscience and machine learning tasks. The higher the generalization gap, the worse the generalization performance. BPTT (black), bio-plausible alternatives (magenta, yellow and green). D-F) Scatter plots of generalization gap v.s. leading Loss Hessian eigenvalue. Same runs as in A-C.

3.1 Network and learning setup

The detailed governing equations of our setup can be found in Methods (Appendix A). We consider a RNN with N_{in} input units, N hidden units and N_{out} readout units (Figure 1A). We verified that trends hold for different network sizes and refer the reader to Appendix A.3 for more details. The update formula for $h_t \in \mathbb{R}^N$ (the hidden state at time t) is governed by:

$$h_{t+1} = \phi(W_h f(h_t), W_x x_t), \tag{1}$$

where $\phi(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ is the hidden state update function, $f(\cdot) : \mathbb{R}^N \to \mathbb{R}^N$ is the activation function, $W_h \in \mathbb{R}^{N\times N}$ (resp. $W_x \in \mathbb{R}^{N_{in}\times N}$) is the recurrent (resp. input) weight matrix and $x \in \mathbb{R}^{N_{in}}$ is the input. For ϕ, we consider a discrete-time implementation of a rate-based recurrent neural network (RNN) similar to the form in [134] (details in Appendix A). Readout $\hat{y} \in \mathbb{R}^{N_{out}}$, with readout weights $w \in \mathbb{R}^{N_{out}\times N}$, is defined as

$$\hat{y} = \langle w, f(h_t)\rangle. \tag{2}$$

We performed experiments on three tasks: sequential MNIST [135], pattern generation [136] and delayed match-to-sample tasks [137]. The objective is to minimize scalar loss $L \in \mathbb{R}$, which (for

regression tasks) is defined as

$$L(W_h) = \frac{1}{2TN_{out}B} \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{k=1}^{N_{out}} (\hat{y}_{k,t}^{(i)} - y_{k,t}^{(i)})^2 \tag{3}$$

given target readout $y \in \mathbb{R}^{N_{out}}$, task duration $T \in \mathbb{R}$ and batch size $B \in \mathbb{R}$. We consider cross-entropy loss for classification tasks (Methods in Appendix A).

Different learning algorithms examined in this work are BPTT (our menchmark), which update weights by computing the exact gradient ($\nabla L(W_h) \in \mathbb{R}^{N \times N}$):

$$\Delta W_h = -\eta \nabla L(W_h), \tag{4}$$

and three SoTA bio-plausible learning rules that update weights using approximate gradient:

$$\widehat{\Delta W_h} = -\eta \tilde{\nabla} L(W_h), \tag{5}$$

where $\tilde{\nabla} L(W_h) \in \mathbb{R}^{N \times N}$ denotes a gradient approximation and $\eta \in \mathbb{R}$ denotes the learning rate. These three learning rules are explained further in Appendix A (Methods) but we note that these bio-plausible learning rules are based on truncations of dependency paths — on the computational graph for the exact gradient— that are biologically implausible to compute (Figure 1A). In all figures, learning rules are labelled as "Three-factor" (symmetric e-prop), "RF Three-factor" (RFLO) and "MDGL", respectively. We remark that the focus here is on comparing artificial to bio-plausible learning rules, rather than between biological rules.

Finally, we note that tasks were learned with mostly comparable training accuracies for all learning rules, and that generalization gaps reflect testing departure from these values. We refer the reader to supplemental materials for more details (Appendix A.3).

3.2 Generalization gap and loss landscape curvature

To study generalization performance, our first step is to compute the generalization gap empirically. Generalization gap is defined as the accuracy on the training dataset minus that of the test set; the larger the generalization gap, the worse the generalization performance. For various learning rules, we plot the histogram of generalization gap at the end of training across runs with distinct initializations; different colors represent different learning rules (Figure 2A-C). Notice that these bio-plausible rules achieve worse and more variable generalization performance than their machine learning counterpart (BPTT in black).

We now investigate if the generalization gap behavior described above correlates well with Loss landscape curvature. We use the leading eigenvalue of the loss' Hessian (where derivatives are taken with respect to model parameters) as a measure for curvature following previous practice [119] and note that it is practical for both empirical [116] and theoretical analyses (Theorem 1). There exists other measures for flatness and due to the scale-dependence issue of Hessian spectrum [131], we also test using parameter scale-independent measures (see Appendix Figures 7 and 8). When we plot generalization gap points from Figure 2A-C against the corresponding leading Hessian eigenvalue, a statistically significant correlation is observed (Figure 2D-F). We note that we did not expect this relation to be very tight since in addition to worse generalization gap on average, bio-plausible learning rules exhibit increased variance. This is an important and consistent trend we observed but might have been overlooked in previous studies.

So far, we investigated the endpoints of optimization trajectories, where training performance has converged. Now, we visualize the whole training trajectory. This done is for two reasons: 1) account for early stopping that can halt training anywhere along the trajectory due to time constraints; 2) flatness during training could give indications of avoiding or escaping high-curvature regions. We observe that the biologically motivated gradient approximations tend to rapidly approach high-curvature regions compared to their machine learning counterpart (Figure 3). Together, these results demonstrate a clear trend and a link between generalization gap and loss landscape curvature: both being increased and more variable for bio-plausible rules. We also note that the curvature convergence behavior seems to be a shared problem of temporal truncations of the gradient (Appendix Figure 9), which is what the existing bio-plausible gradient approximations for RNNs are based on. Next, we provide a theoretical argument as to why truncated temporal credit assignment rules favor high curvature regions of the loss landscape.

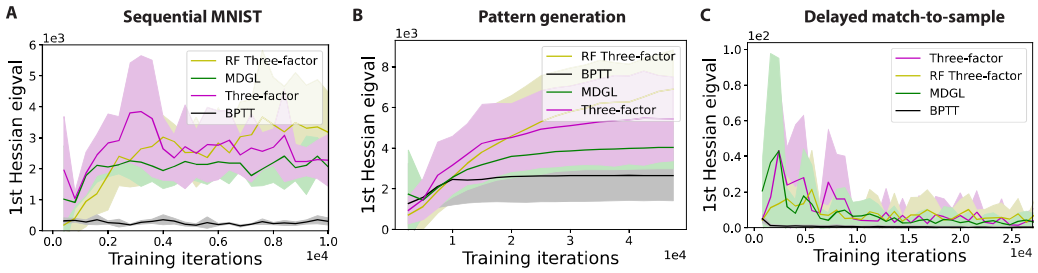

Figure 3: **Bio-plausible gradient approximations tend to approach high curvature regions in loss landscape**. Dominant Hessian eigenvalues are plotted throughout training for bio-plausible learning rules and BPTT. This analysis is done for A) sequential MNIST, B) pattern generation and C) delayed match-to-sample tasks. Solid lines/shaded regions: mean/standard deviation of dominant Hessian eigenvalue curves across five independent runs.

3.3 Theoretical analysis: link between curvature and gradient approximation error

We discussed how generalization can be linked to curvature, and we now examine the link between curvature and gradient alignment during learning dynamics. We represent approximate gradients in a *rule agnostic* manner, where an arbitrary approximation is represented in terms of its component along the gradient direction plus an arbitrary orthogonal vector (Figure 4A):

$$\vec{\hat{g}} = \rho\vec{g} + \vec{e}, \tag{6}$$

where $\vec{g} := \nabla L(W_h) \in \mathbb{R}^{N^2}$ and $\vec{\hat{g}} \in \mathbb{R}^{N^2}$ are the exact and approximate gradients, respectively (we reshaped $\nabla L(W_h)$ into a vector here); $\vec{e} \in \mathbb{R}^{N^2}$ is an arbitrary orthogonal vector to $\vec{g}$. Here, scalar $\rho \in \mathbb{R}$ represents the relative step length along the gradient direction that the approximate rule is making. As we will see in Theorem 1, ρ is an important quantity in our analysis. One can easily compute ρ from $\vec{g}$ and $\vec{\hat{g}}$ by $\rho = \frac{\vec{\hat{g}}^T\vec{g}}{\vec{g}^T\vec{g}}$.

We express weight updates as discrete dynamical systems (with weights W_h as the state variables):

$$W_h^+ \leftarrow W_h^- + F(W_h^-) = W_h^- - \eta\nabla L(W_h^-), \text{ for BPTT} \tag{7}$$

$$W_h^+ \leftarrow W_h^- + \hat{F}(W_h^-) = W_h^- - \eta\tilde{\nabla} L(W_h^-), \text{ for an approximate rule,} \tag{8}$$

where η is the learning rate, $\tilde{\nabla}$ is an approximate gradient, and $F : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2}$ (resp. $\hat{F} : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2}$) denotes the map defined by BPTT (resp. an approximate) weight update rule. We note in passing that dynamical systems view of weight updates have been used previously [138, 139].

We introduce some additional notations before presenting our main theoretical result. $J \in \mathbb{R}^{N^2\times N^2}$ (resp. $\widehat{J} \in \mathbb{R}^{N^2\times N^2}$) is the Jacobian of the dynamical system of BPTT (resp. an approximate rule) in Eq. 7 (resp. Eq. 8). $\lambda_1^J \in \mathbb{R}$ (resp. $\widehat{\lambda_1^J} \in \mathbb{R}$) is the leading eigenvalue for BPTT (resp. approximate rule) Jacobian. $\lambda_1^H \in \mathbb{R}$ is the leading eigenvalue of the loss' Hessian matrix. $W_B^* \in \mathbb{R}^{N\times N}$ (resp. $W_e^* \in \mathbb{R}^{N\times N}$) is the final fixed point for BPTT (resp. an approximate rule). We are now ready to present Theorem 1.

Theorem 1. *Consider a RNN defined in Eq. 1 with a single scalar output $\hat{y}$ and least squares loss as in Eq. 3 presented only at the last time step T, and weights are updated according to the difference equation for BPTT 7 (resp. approximate rule 8) on a single example (stochastic gradient descent) using learning rate η_B (resp. η_e). In the limit of stable fixed point convergence with zero training error, the dominant loss' Hessian eigenvalue attained by BPTT (resp. approximate rule) is bounded by $|\lambda_1^H(W_B^*)| < \frac{1}{\eta_B}$ (resp. $|\lambda_1^H(W_e^*)| < \frac{1}{|\rho|\eta_e}$).*

Proof. Please refer to **Appendix B for the full proof**. An intuition is given in Figure 4B. Here are the main steps involved in the proof:

1. Observe that the *Jacobian* of BPTT discrete time dynamical system (Eq. 7) is precisely the *loss' Hessian* scaled by a constant

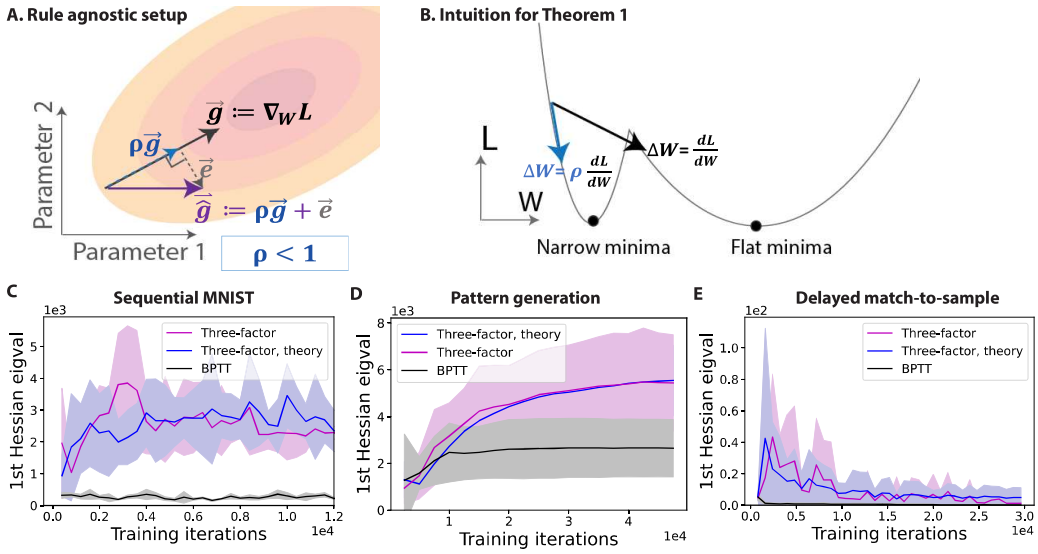

Figure 4: **Preference for high curvature regions connected to worse gradient alignment (rule agnostic analysis)**. A) An arbitrary gradient approximation, $\hat{\vec{g}}$, its component along the gradient direction, $\rho\vec{g}$, and orthogonal to the gradient $\vec{e}$ (Eq. 6). The scalar ρ represents the relative step length along the gradient direction. B) Illustration for Theorem 1: if gradient $\vec{g}$ is aligned with the sharpest directions but error $\vec{e}$ is not (see Appendix Figure 10), smaller step length along the gradient direction can make it harder to step over narrow minima. C-E) Leading loss Hessian eigenvalue v.s. training iteration for three-factor rule, BPTT, and modified BPTT (three-factor, theory). For the latter, step length along the gradient direction of BPTT was matched to three-factor rule by multiplying BPTT update with a factor of ρ and recovers curvature trends of three-factor rule.

2. Using the above relationship, we can bound $|\lambda_1^H|$ from $|\lambda_1^J| < 1$, which is the condition for discrete time dynamical system to converge to a fixed point.

3. However, the link between the Jacobian of an approximate update rule and the loss' Hessian is less obvious. Thus, we derive a link between $|\widehat{\lambda_1^J}|$ and $|\lambda_1^J|$, and then apply the step above.

□

The consequence of the upper bound derived in Theorem 1 is that truncated gradient rules can converge to minima with a higher dominant Hessian eigenvalue than BPTT, with the leading eigenvalue bound inversely proportional to $|\rho|$ ($|\rho| < 1$ usually). In practice, ρ can vary depending on task settings and in our setup, we observed it to be somewhere between 0.02 and 0.3. We remark that this higher upper-bound is consistent with the increased spread of curvature and generalization observed for bio-plausible rules in experiments. Theorem 1 highlights scalar ρ (Eq. 6), relative step length along the gradient direction, as an important factor in the curvature bound. To test that, we eliminate the factor of ρ by reducing learning rule of BPTT such that its step length is matched to that of the three-factor rule. This resulted in the blue curves in Figure 4, which is still trained using BPTT but with the update scaled by a factor of ρ. By matching the step length of BPTT and a three-factor rule along the gradient direction, similar convergence behaviors were observed. **This result then attributes the curvature preference behavior to relative step length along the gradient direction, and thereby indicating a link between curvature and gradient alignment**. Consistent with earlier results, when the step length of BPTT is matched to that of a three-factor rule, its corresponding generalization performance also worsened (Appendix Figure 6).

We make a few more remarks regarding Theorem 1. ρ is usually less than 1 because otherwise the additional orthogonal component $\vec{e}$ would imply a larger update size for the approximate rule compared to BPTT. This would make the approximate rule more prone to numerical instabilities (Eq. 9) and would require a change in solver hyperparameters such as overall learning rate which

in turn, influence the update size. Curiously, $\vec{e}$ did not seem to play a helpful role in finding flatter minima; we provide the reasoning for that in Figure 4B and Appendix Figure 10. Despite making assumptions including scalar output, MSE loss and loss available at the last step, our empirical results indicate that our conclusions also extend to other setups: vector output, cross-entropy loss and loss that accumulates over time steps (Figure 4). We discuss the generality of Theorem 1 in Appendix B.

One might assume increasing learning rate of the approximate rules could compensate for the reduced step length due to ρ. However, this can raise other issues. Suppose $\Delta W = -\eta\vec{g}$ for exact gradient and we increase the learning rate of approximate update until $\widehat{\Delta W} = -\eta\vec{g} - \eta\vec{e}$ (magnitude of $\vec{e}$ is scaled accordingly), and because $\vec{g} \perp \vec{e}$:

$$\|\widehat{\Delta W}\|^2 = \eta^2\|\vec{g}\|^2 + \eta^2\|\vec{e}\|^2 > \eta^2\|g\|^2 = \|\Delta W\|^2 \tag{9}$$

In other words, if approximate updates $\widehat{\Delta W}$ make the same amount of progress as BPTT along the gradient direction, $\widehat{\Delta W}$ would have a larger magnitude due to the orthogonal component $\vec{e}$, and large update magnitude can be very problematic for numerical stability [140]. Thus, the magnitude of $\vec{e}$ limits the learning rate that can be used. To balance between this stability issue and the potential benefit of large learning rates, one can consider using a large learning rate early in training to prevent premature stabilization in sharp minima followed by gradual decay to mitigate the stability issue (see Appendix Figure 5).

Taken together, Theorem 1 and Eq. 9 connect gradient alignment with the curvature of converged solution. For an approximation that is aligned poorly with the exact gradient, large approximation error vector $\vec{e}$ limit the step length for stability reasons (Eq. 9) and small relative step length ρ correspond to a larger curvature bound (Theorem 1). This is consistent with a larger spread in curvature and generalization gap observed for bio-plausible learning rules (Figure 2, Appendix Figure 6).

4 Conclusion

While developing bio-plausible learning rules is of an interest for both answering neuroscience questions and searching for more efficient training strategies for artificial networks, generalization properties of solutions found by these rules are severely underexamined. Through various well-known machine learning and working memory tasks, we first demonstrate empirically that existing bio-plausible learning rules in RNNs attain worse generalization performance, which is consistent with their tendency to converge to high-curvature regions in loss landscape. Second, our theoretical analysis offers an explanation for this preference of high curvature regions based on worse alignment to true gradient. This regime corresponds to the situation where the step length along the gradient direction is small and the approximation error vector is large. Finally, we test this theory empirically by matching the relative step length along the gradient direction resulting in similar convergence behavior (Figure 4).

5 Discussion

Our study — a stepping stone toward understanding biological generalization using deep learning methods — raises many exciting questions for future investigations, both on the front of stronger deep learning theory and more sophisticated biological ingredients.

Deep learning theory and its implications: In this study, we investigate generalization properties using loss landscape flatness, a promising predictor with recent rigorous connection to generalization gap [6]. Yet, to what extent can flatness explain generalization is still an open question in deep learning. For instance, curvature is a local measure, which means that its informativeness of robustness against global perturbation is limited. Moreover, the theoretical association between flatness and generalization is provided in the form of upper bound [6, 9] and the bound may not always be tight, which is consistent with more variability in generalization gap for bio-plausible learning rules but offers less predictive power. Despite observing a significant correlation between generalization gap and a curvature-based measure in Figure 2, the relationship appears to be messy, suggesting other factors involved for explaining generalization. Our results are also consistent with existing findings linking learning rate to loss landscape curvature in deep networks [121]. In the case of bio-plausible

gradient approximations, small step length along the gradient direction cannot be compensated by increasing the learning rate, as that would inadvertently increase the error vector, causing numerical issues (Eq. 9). Curiously, the approximation error vector $\vec{e}$ did not seem to play a role other than restricting the learning rate. While it is well-known that stochastic gradient noise (SGN) can help with finding flat minima due to the alignment of SGN covariance and Hessian near minima [120, 141–143], that may not apply to approximation error vector $\vec{e}$ resulting primarily from temporal truncation of the gradient (see Figure 1A and Appendix Figure 10). Given that developing better predictors of generalization is still a work in progress [4], we anticipate stronger theoretical tools to be applied for studying biological generalization in the future.

Toward more detailed biological mechanisms: On the front of more sophisticated biological ingredients that may improve generalization performance, we see two lines of approaches: **1-** develop bio-plausible learning rules that align better with the gradient, as suggested by our rule agnostic analysis (Theorem 1, Figure 4); and **2-** instead of studying learning rules in isolation, consider also other neural circuit components [144–150] that could interact with the learning rule. An important component would be the architecture, including connectivity structure and neuron model, found through evolution [149] (see also [151]). To address our main question, this study focuses on the effect of various learning rules while holding data, objective function and architecture constant (see [10]). However, these different components can interact, and more sophisticated architecture can facilitate task learning [151–158]. Given the exploding parameter space resulting from such interactions, we believe it requires careful future analysis and is outside of the scope for this one paper.

Additionally, learning rate modulation [159, 160] could be one of many possible remedies employed by the brain. We conjecture that neuromodulatory mechanisms could be coupled with these learning rules to improve the convergence behavior through our scheduled learning rate experiments (Appendix Figure 5), where an initial high learning rate could prevent the learning trajectory from settling in sharp minima prematurely followed by gradual decay to avoid instabilities. One possible way to realize such learning rate modulation could be through serotonin neurons via uncertainty tracking, where the learning rate is high when the reward prediction error is high (this can happen at the beginning of learning) [161]. Since the authors of [161] showed that inhibiting serotonin led to failure in learning rate modulation, we *conjecture* such inhibition might have an impact on the generalization performance of learning outcomes. This result also connects with the finding that sensory depletion during critical periods in training deep networks, which can be related to a small learning rate early in training, can impair learning and yield convergence to sharp minima [162]. Other ingredients for future investigations could include intrinsic noise with certain structures [146, 163, 164]. Taken together, we hope to see follow up investigations — riding on the rapid advancements both at the front of deep learning theory and sophisticated biological mechanisms — into how the brain attains solutions that generalize.

References

[1] Timothy P Lillicrap and Adam Santoro. Backpropagation through time and the brain. *Current opinion in neurobiology*, 55:82–89, 2019.

[2] Luke Y Prince, Ellen Boven, Roy Henha Eyono, Arna Ghosh, Joe Pemberton, Franz Scherr, Claudia Clopath, Rui Ponte Costa, Wolfgang Maass, Blake A Richards, et al. Ccn gac workshop: Issues with learning in biological recurrent neural networks. *arXiv preprint arXiv:2105.05382*, 2021.

[3] Owen Marschall, Kyunghyun Cho, and Cristina Savin. A unified framework of online learning algorithms for training recurrent neural networks. *Journal of Machine Learning Research*, 21(135):1–34, 2020.

[4] Yiding Jiang, Pierre Foret, Scott Yak, Daniel M Roy, Hossein Mobahi, Gintare Karolina Dziugaite, Samy Bengio, Suriya Gunasekar, Isabelle Guyon, and Behnam Neyshabur. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020.

[5] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[6] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[8] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[9] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pages 9636–9647. PMLR, 2020.

[10] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[11] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11(1):1–15, dec 2020.

[12] James M Murray. Local online learning in recurrent networks with random feedback. *ELife*, 8:e43299, 2019.

[13] Yuhan Helena Liu, Stephen Smith, Stefan Mihalas, Eric Shea-Brown, and Uygar Sümbül. Cell-type–specific neuromodulation guides synaptic credit assignment in a spiking neural network. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[14] Colin Bredenberg, Benjamin Lyo, Eero Simoncelli, and Cristina Savin. Impression learning: Online representation learning with synaptic plasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Yanis Bahroun, Dmitri Chklovskii, and Anirvan Sengupta. A normative and biologically plausible algorithm for independent component analysis. *Advances in Neural Information Processing Systems*, 34, 2021.

[16] Basile Confavreux, Friedemann Zenke, Everton Agnes, Timothy Lillicrap, and Tim Vogels. A meta-learning approach to (re) discover plasticity rules that carve a desired function into a neural network. *Advances in Neural Information Processing Systems*, 33:16398–16408, 2020.

[17] Gabriel Ocker and Michael Buice. Tensor decompositions of higher-order correlations by nonlinear hebbian plasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang. Biological key-value memory networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Alexander Meulemans, Matilde Tristany Farinha, Javier Garcia Ordonez, Pau Vilimelis Aceituno, João Sacramento, and Benjamin F Grewe. Credit assignment in neural networks through deep feedback control. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Brandon McMahan, Michael Kleinman, and Jonathan Kao. Learning rule influences recurrent network representations but not attractor structure in decision-making tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

[21] Paul Haider, Benjamin Ellenberger, Laura Kriener, Jakob Jordan, Walter Senn, and Mihai A Petrovici. Latent equilibrium: A unified learning theory for arbitrarily fast computation with arbitrarily slow neurons. *Advances in Neural Information Processing Systems*, 34:17839–17851, 2021.

[22] Roman Pogodin, Yash Mehta, Timothy Lillicrap, and Peter E Latham. Towards biologically plausible convolutional networks. *Advances in Neural Information Processing Systems*, 34:13924–13936, 2021.

[23] Johannes Friedrich, Siavash Golkar, Shiva Farashahi, Alexander Genkin, Anirvan Sengupta, and Dmitri Chklovskii. Neural optimal feedback control with local learning rules. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] Colin Bredenberg, Eero Simoncelli, and Cristina Savin. Learning efficient task-dependent representations with synaptic plasticity. *Advances in Neural Information Processing Systems*, 33:15714–15724, 2020.

[25] Elias Najarro and Sebastian Risi. Meta-learning through hebbian plasticity in random networks. *Advances in Neural Information Processing Systems*, 33:20719–20731, 2020.

[26] Pablo Tano, Peter Dayan, and Alexandre Pouget. A local temporal difference code for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33:13662–13673, 2020.

[27] Lea Duncker, Laura Driscoll, Krishna V Shenoy, Maneesh Sahani, and David Sussillo. Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in neural information processing systems*, 33:14387–14397, 2020.

[28] Aran Nayebi, Sanjana Srivastava, Surya Ganguli, and Daniel L Yamins. Identifying learning rules from neural network observables. *Advances in Neural Information Processing Systems*, 33:2639–2650, 2020.

[29] Zoe Ashwood, Nicholas A Roy, Ji Hyun Bak, and Jonathan W Pillow. Inferring learning rules from animal decision-making. *Advances in Neural Information Processing Systems*, 33:3442–3453, 2020.

[30] Siavash Golkar, David Lipshutz, Yanis Bahroun, Anirvan Sengupta, and Dmitri Chklovskii. A simple normative network approximates local non-hebbian learning in the cortex. *Advances in neural information processing systems*, 33:7283–7295, 2020.

[31] Daniel R Kepple, Rainer Engelken, and Kanaka Rajan. Curriculum learning as a tool to uncover learning principles in the brain. In *International Conference on Learning Representations*, 2021.

[32] Danil Tyulmankov, Guangyu Robert Yang, and LF Abbott. Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. *Neuron*, 110(3):544–557, 2022.

[33] Eli Moore and Rishidev Chaudhuri. Using noise to probe recurrent neural network structure and prune synapses. *Advances in Neural Information Processing Systems*, 33:14046–14057, 2020.

[34] Grace Wan Yu Ang, Clara S Tang, Y Audrey Hay, Sara Zannone, Ole Paulsen, and Claudia Clopath. The functional role of sequentially neuromodulated synaptic plasticity in behavioural learning. *PLoS Computational Biology*, 17(6):e1009017, 2021.

[35] Eren Sezener, Agnieszka Grabska-Barwińska, Dimitar Kostadinov, Maxime Beau, Sanjukta Krishnagopal, David Budden, Marcus Hutter, Joel Veness, Matthew Botvinick, Claudia Clopath, et al. A rapid and efficient learning rule for biological neural circuits. *BioRxiv*, 2021.

[36] Owen Marschall, Kyunghyun Cho, and Cristina Savin. Using local plasticity rules to train recurrent neural networks. *arXiv preprint arXiv:1905.12100*, 2019.

[37] David Clark, LF Abbott, and SueYeon Chung. Credit assignment through broadcasting a global error vector. *Advances in Neural Information Processing Systems*, 34:10053–10066, 2021.

[38] Maxence Ernoult, Fabrice Normandin, Abhinav Moudgil, Sean Spinney, Eugene Belilovsky, Irina Rish, Blake Richards, and Yoshua Bengio. Towards scaling difference target propagation by learning backprop targets. *arXiv preprint arXiv:2201.13415*, 2022.

[39] Jack Lindsey and Ashok Litwin-Kumar. Learning to learn with feedback and local plasticity. *Advances in Neural Information Processing Systems*, 33:21213–21223, 2020.

[40] Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg, and Nicolas Brunel. Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience*, 18(12):1804–1810, 2015.

[41] Artur Luczak, Bruce L. McNaughton, and Yoshimasa Kubo. Neurons learn by predicting future activity. *Nature Machine Intelligence*, 4(1):62–72, January 2022.

[42] James C.R. Whittington and Rafal Bogacz. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, 2019.

[43] Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

[44] Matthew G Perich, Charlotte Arlt, Sofia Soares, Megan E Young, Clayton P Mosher, Juri Minxha, Eugene Carter, Ueli Rutishauser, Peter H Rudebeck, Christopher D Harvey, et al. Inferring brain-wide interactions using data-constrained recurrent neural network models. *bioRxiv*, pages 2020–12, 2021.

[45] Jimmy Smith, Scott Linderman, and David Sussillo. Reverse engineering recurrent neural networks with jacobian switching linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] Michael Kleinman, Chandramouli Chandrasekaran, and Jonathan Kao. A mechanistic multi-area recurrent network model of decision-making. *Advances in Neural Information Processing Systems*, 34, 2021.

[47] Friedrich Schuessler, Francesca Mastrogiuseppe, Alexis Dubreuil, Srdjan Ostojic, and Omri Barak. The interplay between randomness and structure during learning in rnns. *Advances in neural information processing systems*, 33:13352–13362, 2020.

[48] Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Ila Rani Fiete, et al. Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice. *bioRxiv*, 2020.

[49] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306, 2019.

[50] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[51] Joshua Glaser, Matthew Whiteway, John P Cunningham, Liam Paninski, and Scott Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in neural information processing systems*, 33:14867–14878, 2020.

[52] Jonathan Kadmon, Jonathan Timcheck, and Surya Ganguli. Predictive coding in balanced neural networks with noise, chaos and delays. *Advances in neural information processing systems*, 33:16677–16688, 2020.

[53] Jonathan Dong, Ruben Ohana, Mushegh Rafayelyan, and Florent Krzakala. Reservoir computing meets recurrent kernels and structured transforms. *Advances in Neural Information Processing Systems*, 33:16785–16796, 2020.

[54] Elia Turner, Kabir V Dabholkar, and Omri Barak. Charting and navigating the space of solutions for recurrent neural networks. *Advances in Neural Information Processing Systems*, 34:25320–25333, 2021.

[55] Sandra Nestler, Christian Keup, David Dahmen, Matthieu Gilson, Holger Rauhut, and Moritz Helias. Unfolding recurrence by green's functions for optimized reservoir computing. *Advances in Neural Information Processing Systems*, 33:17380–17390, 2020.

[56] Robert Kim and Terrence J Sejnowski. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nature Neuroscience*, 24(1):129–139, 2021.

[57] Adrian Valente, Srdjan Ostojic, and Jonathan Pillow. Probing the relationship between linear dynamical systems and low-rank recurrent neural network models. *arXiv preprint arXiv:2110.09804*, 2021.

[58] Tom Macpherson, Anne Churchland, Terry Sejnowski, James DiCarlo, Yukiyasu Kamitani, Hidehiko Takahashi, and Takatoshi Hikida. Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, 144:603–613, 2021.

[59] Ben Tsuda, Stefan C Pate, Kay M Tye, Hava T Siegelmann, and Terrence J Sejnowski. Neuromodulators generate multiple context-relevant behaviors in a recurrent neural network by shifting activity hypertubes. *bioRxiv*, pages 2021–05, 2022.

[60] Nuttida Rungratsameetaweemana, Robert Kim, John T Serences, and Terrence J Sejnowski. Probabilistic visual processing in humans and recurrent neural networks. *Journal of Vision*, 22(3):24–24, 2022.

[61] Brian DePasquale, Christopher J Cueva, Kanaka Rajan, G Sean Escola, and LF Abbott. full-force: A target-based method for training recurrent networks. *PloS one*, 13(2):e0191527, 2018.

[62] Christopher Langdon and Tatiana A Engel. Latent circuit inference from heterogeneous neural responses during cognitive tasks. *bioRxiv*, 2022.

[63] R. J. Williams and D. Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin and D. E. Rumelhart, editors, *Back-propagation: Theory, Architectures and Applications*, chapter 13, pages 433–486. Hillsdale, NJ: Erlbaum, 1995.

[64] Drew Linsley, Alekh Karkada Ashok, Lakshmi Narasimhan Govindarajan, Rex Liu, and Thomas Serre. Stable and expressive recurrent vision models. *arXiv preprint arXiv:2005.11362*, 2020.

[65] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, pages 1–12, 2020.

[66] Isabella Pozzi, Sander Bohté, and Pieter Roelfsema. A biologically plausible learning rule for deep learning in the brain. *arXiv preprint arXiv:1811.01768*, 2018.

[67] Axel Laborieux, Maxence Ernoult, Benjamin Scellier, Yoshua Bengio, Julie Grollier, and Damien Querlioz. Scaling equilibrium propagation to deep convnets by drastically reducing its gradient estimator bias. *Frontiers in neuroscience*, 15:129, 2021.

[68] Beren Millidge, Alexander Tschantz, Anil K Seth, and Christopher L Buckley. Activation relaxation: A local dynamical approximation to backpropagation in the brain. *arXiv preprint arXiv:2009.05359*, 2020.

[69] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *arXiv preprint arXiv:1810.11393*, 2018.

[70] Yali Amit. Deep learning with asymmetric connections and hebbian updates. *Frontiers in computational neuroscience*, 13:18, 2019.

[71] Jonathan E Rubin, Catalina Vich, Matthew Clapp, Kendra Noneman, and Timothy Verstynen. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *European Journal of Neuroscience*, 53(7):2234–2253, 2021.

[72] Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, pages 1–10, 2021.

[73] Pieter R. Roelfsema and Anthony Holtmaat. Control of synaptic plasticity in deep cortical networks. *Nature Reviews Neuroscience*, 19(3):166–180, feb 2018.

[74] 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.

[75] Jeffrey C. Magee and Christine Grienberger. Synaptic Plasticity Forms and Functions. *Annual Review of Neuroscience*, 43(1):95–117, jul 2020.

[76] Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 12:53, jul 2018.

[77] Magdalena Sanhueza and John Lisman. The CaMKII/NMDAR complex as a molecular memory. *Molecular Brain*, 6(1):10, feb 2013.

[78] Stijn Cassenaer and Gilles Laurent. Conditional modulation of spike-timing-dependent plasticity for olfactory learning. *Nature*, 482(7383):47–51, feb 2012.

[79] Sho Yagishita, Akiko Hayashi-Takagi, Graham C.R. Ellis-Davies, Hidetoshi Urakubo, Shin Ishii, and Haruo Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620, sep 2014.

[80] Aparna Suvrathan. Beyond STDP — towards diverse and functionally relevant plasticity rules. *Current Opinion in Neurobiology*, 54:12–19, feb 2019.

[81] Wolfram Schultz. Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, 17(3):183, 2016.

[82] Johnatan Aljadeff, James D'amour, Rachel E Field, Robert C Froemke, and Claudia Clopath. Cortical credit assignment by hebbian, neuromodulatory and inhibitory plasticity. *arXiv preprint arXiv:1911.00307*, 2019.

[83] Răzvan V Florian. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural computation*, 19(6):1468–1502, 2007.

[84] Michael A. Farries and Adrienne L. Fairhall. Reinforcement Learning With Modulated Spike Timing–Dependent Synaptic Plasticity. *Journal of Neurophysiology*, 98(6):3648–3665, dec 2007.

[85] Robert Legenstein, Dejan Pecevski, and Wolfgang Maass. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS computational biology*, 4(10):e1000180, 2008.

[86] Zuzanna Brzosko, Susanna B Mierau, and Ole Paulsen. Neuromodulation of spike-timing-dependent plasticity: past, present, and future. *Neuron*, 103(4):563–581, 2019.

[87] Verena Pawlak, Jeffery R Wickens, Alfredo Kirkwood, and Jason ND Kerr. Timing is not everything: neuromodulation opens the stdp gate. *Frontiers in synaptic neuroscience*, 2:146, 2010.

[88] Stephen J Smith, Uygar Sümbül, Lucas T Graybuck, Forrest Collman, Sharmishtaa Seshamani, Rohan Gala, Olga Gliko, Leila Elabbady, Jeremy A Miller, Trygve E Bakken, et al. Single-cell transcriptomic evidence for dense intracortical neuropeptide networks. *Elife*, 8:e47889, 2019.

[89] Stephen J Smith, Michael Hawrylycz, Jean Rossier, and Uygar Sümbül. New light on cortical neuropeptides and synaptic network plasticity. *Current Opinion in Neurobiology*, 63:176–188, 2020.

[90] Asier Mujika, Florian Meier, and Angelika Steger. Approximating real-time recurrent learning with random kronecker factors. *Advances in Neural Information Processing Systems*, 31, 2018.

[91] Corentin Tallec and Yann Ollivier. Unbiased online recurrent optimization. *arXiv preprint arXiv:1702.05043*, 2017.

[92] Tim Cooijmans and James Martens. On the variance of unbiased online recurrent optimization. *arXiv preprint arXiv:1902.02405*, 2019.

[93] Christopher Roth, Ingmar Kanitscheider, and Ila Fiete. Kernel rnn learning (kernl). In *International Conference on Learning Representations*, 2018.

[94] Jacob Menick, Erich Elsen, Utku Evci, Simon Osindero, Karen Simonyan, and Alex Graves. A practical sparse approximation for real time recurrent learning. *arXiv preprint arXiv:2006.07232*, 2020.

[95] Friedemann Zenke and Emre O Neftci. Brain-inspired learning on neuromorphic substrates. *arXiv preprint arXiv:2010.11931*, 2020.

[96] Yinan Cao, Christopher Summerfield, and Andrew Saxe. Characterizing emergent representations in a space of candidate learning rules for deep networks. *Advances in Neural Information Processing Systems*, 33:8660–8670, 2020.

[97] Ganlin Song, Ruitu Xu, and John Lafferty. Convergence and alignment of gradient descent with random backpropagation weights. *Advances in Neural Information Processing Systems*, 34, 2021.

[98] Manuela Girotti, Ioannis Mitliagkas, and Gauthier Gidel. Convergence analysis and implicit regularization of feedback alignment for deep linear networks. *arXiv preprint arXiv:2110.10815*, 2021.

[99] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[100] Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? *Advances in Neural Information Processing Systems*, 32, 2019.

[101] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[102] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[103] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.

[104] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[105] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[106] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[107] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[108] Xu Sun, Zhiyuan Zhang, Xuancheng Ren, Ruixuan Luo, and Liangyou Li. Exploring the vulnerability of deep neural networks: A study of parameter corruption. *arXiv preprint arXiv:2006.05620*, 2020.

[109] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[110] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

[111] Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[112] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

[113] Huan Wang, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. Identifying generalization properties in neural networks. *arXiv preprint arXiv:1809.07402*, 2018.

[114] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[115] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[116] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

[117] Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34, 2021.

[118] Zeke Xie, Qian-Yuan Tang, Yunfeng Cai, Mingming Sun, and Ping Li. On the power-law spectrum in deep learning: A bridge to protein science. *arXiv preprint arXiv:2201.13011*, 2022.

[119] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[120] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

[121] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Finding flatter minima with sgd. 2018.

[122] Chiyuan Zhang, Qianli Liao, Alexander Rakhlin, Brando Miranda, Noah Golowich, and Tomaso Poggio. Theory of deep learning iib: Optimization properties of sgd. *arXiv preprint arXiv:1801.02254*, 2018.

[123] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[124] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[125] Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2020.

[126] Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. Questions for flat-minima optimization of modern neural networks. *arXiv preprint arXiv:2202.00661*, 2022.

[127] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8156–8165, 2021.

[128] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[129] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. *Advances in neural information processing systems*, 32, 2019.

[130] Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. *arXiv preprint arXiv:2012.03801*, 2020.

[131] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[132] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pages 888–896. PMLR, 2019.

[133] Akshay Rangamani, Nam H Nguyen, Abhishek Kumar, Dzung Phan, Sang H Chin, and Trac D Tran. A scale invariant flatness measure for deep network minima. *arXiv preprint arXiv:1902.02434*, 2019.

[134] Daniel B Ehrlich, Jasmine T Stone, David Brandfonbrener, Alexander Atanasov, and John D Murray. Psychrnn: An accessible and flexible python package for training recurrent neural network models on cognitive tasks. *Eneuro*, 8(1), 2021.

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

[136] Wilten Nicola and Claudia Clopath. Supervised learning in spiking neural networks with force training. *Nature communications*, 8(1):1–15, 2017.

[137] Travis Meyer, Xue-Lian Qi, Terrence R Stanford, and Christos Constantinidis. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *Journal of neuroscience*, 31(17):6266–6276, 2011.

[138] Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[139] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[140] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[141] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

[142] Mingwei Wei and David J Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *arXiv preprint arXiv:1910.00195*, 2019.

[143] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[144] Vladimir Ivanov and Konstantinos Michmizos. Increasing liquid state machine performance with edge-of-chaos dynamics organized by astrocyte-modulated plasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[145] Lukas Braun and Tim Vogels. Online learning of neural computations from sparse temporal feedback. *Advances in Neural Information Processing Systems*, 34, 2021.

[146] Joel Dapello, Jenelle Feather, Hang Le, Tiago Marques, David Cox, Josh McDermott, James J DiCarlo, and SueYeon Chung. Neural population geometry reveals the role of stochasticity in robust perception. *Advances in Neural Information Processing Systems*, 34, 2021.

[147] Maurizio De Pittà, Nicolas Brunel, and Andrea Volterra. Astrocytes: Orchestrating synaptic plasticity? *Neuroscience*, 323:43–61, 2016.

[148] James M Murray and G Sean Escola. Remembrance of things practiced with fast and slow learning in cortical and subcortical pathways. *Nature Communications*, 11(1):1–12, 2020.

[149] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[150] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

[151] Victor Geadah, Stefan Horoi, Giancarlo Kerg, Guy Wolf, and Guillaume Lajoie. Goal-driven optimization of single-neuron properties in artificial networks reveals regularization role of neural diversity and adaptation. *bioRxiv*, 2022.

[152] Ilenna Simone Jones and Konrad Paul Kording. Can single neurons solve mnist? the computational power of biological dendritic trees. *arXiv preprint arXiv:2009.01269*, 2020.

[153] Nicolas Perez-Nieves, Vincent CH Leung, Pier Luigi Dragotti, and Dan FM Goodman. Neural heterogeneity promotes robust learning. *Nature communications*, 12(1):1–9, 2021.

[154] Laureline Logiaco, LF Abbott, and Sean Escola. Thalamic control of cortical dynamics in a model of flexible motor sequencing. *Cell reports*, 35(9):109090, 2021.

[155] Guangyu Robert Yang and Manuel Molano Mazon. Next-generation of recurrent neural network models for cognition. 2021.

[156] Corinne Teeter, Ramakrishnan Iyer, Vilas Menon, Nathan Gouwens, David Feng, Jim Berg, Aaron Szafer, Nicholas Cain, Hongkui Zeng, Michael Hawrylycz, et al. Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature communications*, 9(1):1–15, 2018.

[157] Chloe Winston, Dana Mastrovito, Eric Shea-Brown, and Stefan Mihalas. Heterogeneity in neuronal dynamics is learned by gradient descent for temporal processing tasks. *bioRxiv*, 2022.

[158] Christoph Stöckl, Dominik Lang, and Wolfgang Maass. Probabilistic skeletons endow brain-like neural networks with innate computing capabilities. *bioRxiv*, 2021.

[159] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[160] Niru Maheswaranathan, David Sussillo, Luke Metz, Ruoxi Sun, and Jascha Sohl-Dickstein. Reverse engineering learned optimizers reveals known and novel mechanisms. *Advances in Neural Information Processing Systems*, 34, 2021.

[161] Cooper D Grossman, Bilal A Bari, and Jeremiah Y Cohen. Serotonin neurons modulate learning rate through uncertainty. *Current Biology*, 2021.

[162] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.

[163] Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anti-correlated noise injection for improved generalization. *arXiv preprint arXiv:2202.02831*, 2022.

[164] Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pages 7594–7602. PMLR, 2019.

[165] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[166] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

[167] Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.

[168] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

[169] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.

[170] Benjamin Cramer, Sebastian Billaudelle, Simeon Kanya, Aron Leibfried, Andreas Grübl, Vitali Karasenko, Christian Pehle, Korbinian Schreiber, Yannik Stradmann, Johannes Weis, et al. Surrogate gradients for analog neuromorphic computing. *Proceedings of the National Academy of Sciences*, 119(4), 2022.

[171] Dongsung Huh and Terrence J Sejnowski. Gradient descent for spiking neural networks. *Advances in neural information processing systems*, 31, 2018.

[172] Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):1514–1541, 2018.

[173] Wenrui Zhang and Peng Li. Temporal spike sequence learning via backpropagation for deep spiking neural networks. *Advances in Neural Information Processing Systems*, 33:12022–12033, 2020.

[174] Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Yisen Wang, and Zhouchen Lin. Training feedback spiking neural networks by implicit differentiation on the equilibrium state. *Advances in Neural Information Processing Systems*, 34, 2021.

[175] Nicolas Perez-Nieves and Dan Goodman. Sparse spiking gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

[176] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[177] Nicolas Skatchkovsky, Osvaldo Simeone, and Hyeryung Jang. Learning to time-decode in spiking neural networks through the information bottleneck. *arXiv preprint arXiv:2106.01177*, 2021.

[178] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: A system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[179] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[180] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] To make this easier for the readers, we have referred to the pertinent Figures and Theorem to support our summary of contributions in Introduction.
   (b) Did you describe the limitations of your work? [Yes] Limitations and future work can be found in our Discussion section.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work contributes to the theoretical understanding of biologically-plausible learning rules for recurrent neural networks. As such, we do not anticipate any direct ethical or societal impact. In the long-term, our work can have impact on related research communities such as neuroscience and deep learning, with societal impact depending on the development of these fields.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have carefully read the ethics review guidelines and confirm that this paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state all assumptions precisely when we first presented the Theorem in Results, We further details, discussions and proofs in the appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs can be found in Appendix B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Anonymized code link is provided in Appendix A.3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All training details are provided in Appendix A.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We tried to provide this information in all applicable figures. This is stated as "Solid lines/shaded regions: mean/standard deviation obtained at the end of training across five independent runs" in the legends of several figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Information pertaining to computing resources and simulation time can be found in Appendix A.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] For dataset, we cited MNIST as well as representative neuroscience papers for the two neuroscience motivated tasks. We have also cited existing coding assets in Appendix A.3.

(b) Did you mention the license of the assets? [Yes] We mentioned that in Appendix A.3.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]