

Stochastic Synaptic Dynamics Under Learning

Jakob Stubenrauch ^{1,2,*} Naomi Auer ³ Richard Kempter ^{1,3,4} and Benjamin Lindner ^{1,2}

¹*Bernstein Center for Computational Neuroscience Berlin, 10115 Berlin, Germany*

²*Physics Department of Humboldt-Universität zu Berlin, 12489 Berlin, Germany*

³*Institute for Theoretical Biology, Department of Biology, Humboldt-Universität zu Berlin, 10115 Berlin, Germany*

⁴*Einstein Center for Neurosciences Berlin, Berlin 10117, Germany*



(Received 5 September 2025; accepted 9 March 2026; published 13 April 2026)

Learning is based on synaptic plasticity, which affects and is driven by neural activity. Because pre- and postsynaptic spiking activity is shaped by randomness, the synaptic weights follow a stochastic process, requiring a probabilistic framework to capture the noisy synaptic dynamics. We consider a paradigmatic supervised learning example: a presynaptic neural population impinging in a sequence of episodes on a recurrent network of integrate-and-fire neurons through synapses undergoing spike-timing-dependent plasticity (STDP) with additive potentiation and multiplicative depression. We first analytically compute the drift and diffusion coefficients for a single synapse within a single episode (microscopic dynamics), mapping the true jump process to a Langevin and the associated Fokker-Planck equations. Leveraging recent analytical tools, we include spike-time-resolving cross-correlations between pre- and postsynaptic spikes, which corrects substantial deviations seen in standard theories purely based on firing rates. We then apply this microdynamical description to the network setup in which hetero-associations are trained over one-shot episodes into a feed-forward matrix of STDP synapses connecting to neurons of the recurrent network (macroscopic dynamics). By mapping statistically distinct synaptic populations to instances of the single-synapse process above, we self-consistently determine the joint neural and synaptic dynamics and, ultimately, the time course of memory degradation and the memory capacity. We demonstrate that specifically in the relevant case of sparse coding, our theory can quantitatively capture memory capacities which are strongly overestimated if spike-time-resolving cross-correlations are ignored. On the single-synapse level, our theory predicts a transition between acceleration and deceleration of the synaptic weight, and a possibly nonmonotonic dependence of the finite-time diffusion coefficient on the synaptic weight. On the memory level, we showcase two complementary roles of homeostatic plasticity: namely, it is both responsible for forgetting and a requirement for the ability to learn. Additionally, we show how strength of plasticity and duration of exposure to stimuli must be scaled to evade a tradeoff dilemma between memory capacity and early accuracy. We conclude with a discussion of the many directions in which our framework can be extended.

DOI: [10.1103/r21z-29nv](https://doi.org/10.1103/r21z-29nv)

I. INTRODUCTION

Synaptic plasticity constitutes the foundation of learning. Although the mechanisms responsible for plasticity are diverse and incompletely understood [1,2], the quantitative dependence of plasticity on the relative timing between pairs of pre- and postsynaptic spikes [3–6] accounts for many of the experimental observations; corrections are often extensions of this pair-based rule. The statistics of synaptic weight dynamics is thus determined by the statistics of pre- and postsynaptic spikes. Because the latter display a considerable variability, synaptic dynamics has to be described in a stochastic framework, even if the intrinsic noise sources of plasticity [7–10] are neglected.

Several aspects of synaptic dynamics have been studied before; the mean dynamics and even ensemble dynamics [11–19] have been investigated for a variety of plasticity rules and settings; see [2] for an extensive review. More recently, the conjunction of classical spike-timing-dependent plasticity (STDP) with homeostatic plasticity has received increased attention (see [20] for an overview).

Classical pair-based STDP implies that the expected synaptic weight depends on (i) pre- and postsynaptic firing rates and (ii) cross-correlations between pre- and postsynaptic spikes [11]. Since the weight change induced by STDP is sensitive to relative spike-timing differences on the order of milliseconds [1,4,5], these two contributions play fundamentally distinct roles for learning. The effect of cross-correlations on learning is typically taken into account by interpreting the post-synaptic neuron as a linear filter. This somewhat drastic simplification is either explained by model choice (conditionally Poissonian neuron) or through a daring but educated trick: to linearize nonlinear neuron models *per realization* using realization-averaged response functions [21–23]; see, e.g., [17,24,25]. Response functions can be

*Contact author: jakob.stubenrauch@rwth-aachen.de

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

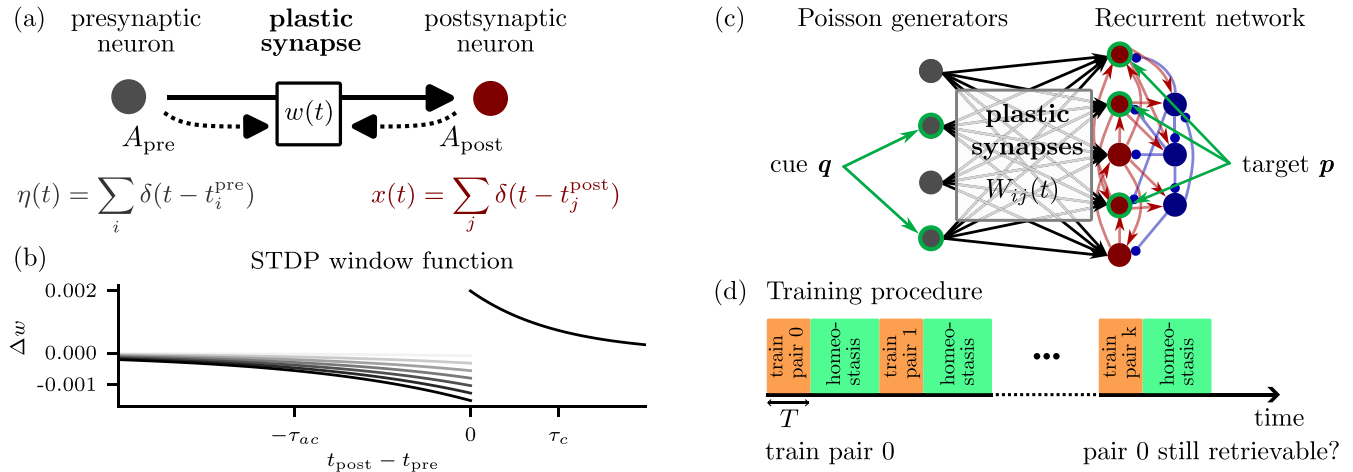


FIG. 1. We investigate memory properties by studying a single synapse dynamics that effectively captures synaptic population dynamics. (a) A Poisson process η drives the neuron x through the synapse w . The weight w is plastic, i.e., it dynamically depends on x and η . We indicated the tracers A_{pre} and A_{post} used in the embedding Eq. (3). (b) STDP window [Eq. (2)] for $w = 0.01$ (light gray) through $w = 0.2$ (black). The setup (a), (b) is studied in Sec. III. (c) A cue representation q in a layer of Poisson processes (gray) is activated and drives a recurrent neural network through a matrix of plastic synapses. Among the excitatory neurons (red) in the recurrent network, a target representation p receives an additional stimulus. The cue and target subsets are marked by the green arrows and circles. (d) Training procedure to store hetero-associations between cues and targets. The setup (c), (d) is studied in Sec. IV. Parameters $\Delta_c = 2 \times 10^{-3}$, $r_{ac} = 8 \times 10^{-3}$, $\tau_c = 16.8/20$, and $\tau_{ac} = 33.7/20$ are matched to [5].

derived with Fokker-Planck theory; for integrate-and-fire neurons see Refs. [26–31].

Despite the vast body of research on synaptic dynamics, a robust and generalizable stochastic description of synaptic dynamics is still missing. Such a description is clearly needed to develop multiscale memory models that take into account the single-neuron level and the population level. In this paper we lay the foundations to this endeavor; specifically, we analytically characterize the stochastic process of synapses endowed with STDP that feed Poisson spikes to leaky integrate-and-fire (LIF) neurons and are shaped by the spikes of both Poisson and LIF neurons. We focus on the stochasticity due to pre- and postsynaptic fluctuations and neglect the intrinsic noisiness of synapses. Instead of relying on realization-wise linearization, we incorporate cross-correlations through an exact model-independent relation between cross-correlations and response functions that two of us reported recently [32]. This approach is robust and generalizable. From the theory developed here we make several qualitative predictions about synaptic dynamics and memory.

Synapses endowed with STDP are capable of self-organized formation of diverse types of memory including (i) learning input correlations [33], (ii) reinforcement learning in recurrent networks of LIF neurons [17], or (iii) error correction of drifting assemblies [34]. Yet another type of STDP-based memory are hetero-associations, i.e., representations in one neuron population evoke representations in another one. Building on our results about stochastic synaptic dynamics, we analytically investigate hetero-associative memory. As we outline, for characterizing this type of memory, the stochastic view on synaptic dynamics is indispensable. Extending our methods to the aforementioned mechanisms (i–iii) poses exciting problems for future work.

The paper is organized as follows: In Sec. II, we first introduce the synapse and neuron model [Figs. 1(a) and 1(b)]

studied throughout the paper. Building on these models, we introduce the network model [Fig. 1(c)] and the training scheme [Fig. 1(d)], which store hetero-associations, into a plastic feed-forward matrix. In Sec. III we characterize the stochastic process of a single synaptic weight by deriving its drift and diffusion coefficients, which define a Langevin (or corresponding Fokker-Planck) equation. From this description, we derive the dynamics of the ensemble mean and variance of synaptic weights. Additionally, we discuss a transition from accelerating to decelerating synaptic weights, which entails qualitative experimental predictions. In Sec. IV we leverage the theory of single weights to study the network scenario in a mean-field theory of synaptic populations. We apply the mean-field theory to characterize forgetting and compute the memory capacity of the setup. Based on this theory, we derive a tradeoff between memory lifetime and early memory accuracy, which we resolve by an optimal training scheme. Lastly, we discuss two complementary roles of homeostatic plasticity.

II. MODELS

First, we introduce the synaptic model depicted in Figs. 1(a) and 1(b). This model depends on pre- and postsynaptic spike trains, thus, next, we define the neuron models used in this paper. Then we introduce the network model depicted in Fig. 1(c), which is composed of the synapse- and neuron models above. We are interested in how hetero-associations can be stored into this network; the necessary training scheme [see Fig. 1(d)] is explained last.

A. Synapse model

We consider a synapse with presynaptic spike train $\eta(t) = \sum_i \delta(t - t_i^{\text{pre}})$ and postsynaptic spike train $x(t) = \sum_j \delta(t -$

t_i^{post}). The evolution of the synaptic weight w in the classical STDP model [2] can be described by

$$\dot{w}(t) = \int_{-\infty}^t dt' \kappa[t - t', w(t)] \eta(t') x(t) + \int_{-\infty}^t dt' \kappa[t' - t, w(t)] \eta(t) x(t'). \quad (1)$$

Each pre-post spike pair leads to a jump in w by κ ; causal spike pairs are captured by the first line, anticausal pairs by the second line. Here $\kappa(\tau, w)$ is the STDP window function, which we assume depends on only the time difference τ between spikes and on w at the time t of update. The lower integral bounds are set to $-\infty$ so as to avoid switch-on effects of the plasticity rule. Throughout the paper, we consider an exponential window with multiplicative (i.e., w -dependent) depression and additive (i.e., w -independent) potentiation [12,35]

$$\kappa(\tau, w) = \begin{cases} \Delta_c e^{-\tau/\tau_c} & \tau \geq 0 \\ -r_{ac} w e^{\tau/\tau_{ac}} & \tau < 0 \end{cases} \quad (2)$$

[see Fig. 1(b)], with $\Delta_c, r_{ac} > 0$ determining the amplitudes and $\tau_c, \tau_{ac} > 0$ being the timescales of potentiation (index c for causal) and depression (index ac for anticausal), respectively. In addition to Eqs. (1) and (2), we demand that the considered excitatory synapse maintains a positive weight throughout, $w(t) \geq 0$, a necessary condition for Dale's law, and implemented in our simulations by a clipping boundary condition. Since clipping happens very rarely for the parameters obtained from [5,6], we can safely disregard the clipping rule in the theory. The plasticity rule (2) with its multiplicative depression and additive potentiation is chosen as it captures the experimental findings of Ref. [5] and because of its inherent stability. Generalizations of Eq. (1), e.g., triplet rules [36,37] or neuromodulatory dynamics [19], are of interest but out of the scope of our study.

For the exponential window in Eq. (2), one can rewrite Eq. (1) in terms of pre- and postsynaptic trace variables A_{pre} and A_{post} [38]:

$$\begin{aligned} \dot{w} &= \Delta_c A_{\text{pre}} x - r_{ac} w A_{\text{post}} \eta, \\ \dot{A}_{\text{pre}} &= -\tau_c^{-1} A_{\text{pre}} + \eta, \\ \dot{A}_{\text{post}} &= -\tau_{ac}^{-1} A_{\text{post}} + x. \end{aligned} \quad (3)$$

Multiple previous studies introduce their plasticity model in an embedded form like Eq. (3) (e.g., Refs. [12,18,19,37,39–41]) and interpret the tracer variables A_{pre} and A_{post} as concentrations of components relevant for plasticity [38]. Besides this potential connection of the phenomenological model (1) to the underlying biophysics, this embedding is useful for simulations as it makes the system local in time; simulation results presented in this paper are based on integrating Eq. (3); see Appendix A. To avoid switch-on effects in Eq. (3), we integrate A_{pre} and A_{post} for some time before we start integrating w ; in this way, A_{pre} and A_{post} thermalize, i.e., their statistics become stationary, which is in line with the lower integral bound at $-\infty$ in Eq. (1).

B. Neuron model

We here introduce a neuron model that is driven (i) by Gaussian white noise and (ii) by a *single* presynaptic neuron through a single synapse. The white Gaussian noise is a simple description of unaccounted noise sources such as intrinsic channel noise and synaptic input from external neural populations that are not explicitly modeled here. The dynamics of the single synapse will be studied in Sec. III. The dynamics of *multiple* synapses providing input to a neuron, studied in Sec. IV, can be effectively mapped to the single-synapse case introduced here. The synaptic dynamics Eq. (1) is driven by the pre- and postsynaptic neural activity η and x , respectively. Thus, to specify the synaptic dynamics, we need to specify the neuron models. Throughout the paper, $\eta(t)$ is the spike train of a Poisson process with rate ν , and $x(t)$ is the spike train of a model neuron driven by η . The developed framework holds true for arbitrary neuron models x , but we give specific expressions and simulation results for the LIF neuron $x(t) = \sum_i \delta(t - t_i^{\text{post}})$, where t_i^{post} are the times at which the membrane voltage governed by

$$\tau_m \dot{v} = -v + \mu + \sqrt{2D} \xi(t) + w(t) \eta(t) \quad (4)$$

hits the threshold v_t ; additionally, at t_i^{post} , v is reset to v_r . Throughout the paper, $v_r = 0$, $v_t = 1$, and the membrane time constant is set to $\tau_m = 1$, i.e., we measure time in multiples of τ_m . The input to the neuron described by Eq. (4) consists of the mean current μ , white Gaussian noise $\sqrt{2D} \xi$ with noise intensity D and $\langle \xi(t) \xi(t') \rangle = \delta(t - t')$, and the Poissonian spikes weighted by w . Later, when we study neurons embedded in a recurrent network, the input parameters μ and D will, in addition to explicit noise, effectively capture the recurrent input. The LIF neuron is mechanistic enough to generate spike trains x that are correlated with η in a beyond-rate-based fashion. Moreover, it has been shown to reproduce experimental spike trains of pyramidal cells [42], and many of its statistical properties are available [26,27,43,44].

C. Network model

Here we introduce the network model depicted in Fig. 1(c), which can store hetero-associations, i.e., feed-forward associations between representations in two distinct populations. A layer of M Poisson processes η_i (gray disks) drives a recurrent network of N LIF neurons [45] (red and blue disks) through a plastic weight matrix W with W_{ij} being the weight by which the j th Poisson process drives the i th postsynaptic neuron. The recurrent network consists of N_E excitatory (red) and N_I inhibitory (blue) neurons, among which only the excitatory neurons are targeted by W . Additionally, each neuron has exactly C_E incoming excitatory connections from the recurrent network and C_I incoming inhibitory connections with static weights $J_{EE} = J$, $J_{EI} = J_{II} = -gJ$, $J_{IE} = hJ$, parameterized by $J > 0$, $g > 0$, and $h > 0$. Specifically, g (h) denotes the ratio of efficacies of I to E/I (E to I) and the efficacies of E to E synapses, respectively. In accordance with neuroanatomical findings, we choose sparse recurrent connectivity $C_E, C_I \ll N_E, N_I$. The computational advantage of the recurrent connectivity is to instantiate a competition between neurons that, in a

soft winner-take-all manner, support strongly driven neurons and suppress weakly driven neurons.

Most input Poisson processes have a low firing rate ν_{lo} ; however, in each training step, a subset of $f_c M$ Poisson generators (index c for cue neurons), represented by the ones in the binary vector $\mathbf{q} \in \{0, 1\}^M$ fires with increased rate ν_{hi} . Simultaneously, a subset of $f_s N_E$ excitatory neurons (index s for supervision), represented by the binary vector $\mathbf{p} \in \{0, 1\}^{N_E}$, receives direct Poissonian input γ_i with rate ν_s and static weight J_s ; the other $(1 - f_s)N_E$ excitatory neurons do not receive additional input. The neurons stimulated by cue and supervision are marked by green circles in Fig. 1. We are particularly interested in sparse representations, where the activation ratios $f_c, f_s \ll 1$.

Summarizing the above, the membrane voltages of the excitatory neurons follow

$$\begin{aligned} \dot{v}_i = & -v_i + \mu_E + \sqrt{2D_E}\xi_i(t) + J \sum_{n \in C_E(i)} x_n - gJ \sum_{n \in C_I(i)} x_n \\ & + \sum_j W_{ij}(t)\eta_j(t) + p_i J_s \gamma_i(t), \quad i = 1, \dots, N_E, \end{aligned} \quad (5)$$

and the membrane voltages of the inhibitory neurons obey

$$\begin{aligned} \dot{v}_i = & -v_i + \mu_I + \sqrt{2D_I}\xi_i(t) + hJ \sum_{n \in C_E(i)} x_n - gJ \sum_{n \in C_I(i)} x_n, \\ i = & N_E + 1, \dots, N_E + N_I, \end{aligned} \quad (6)$$

where the spike trains $x_n(t) = \sum_i \delta(t - t_{n,i})$ are given in terms of the fire-and-reset times $t_{n,i}$ of neuron n [see Eq. (4)], and μ_E, D_E (μ_I, D_I) are the baseline mean input and noise intensity of the excitatory (inhibitory) neurons. The white noise contributions aim to capture noise sources such as channel noise; when approximating Eqs. (5) and (6) with effective instances of the single neuron Eq. (4), the total noise intensity D consists of the explicit noise $D_{E/I}$ and the noise intensity of the network and Poissonian input. The recurrent network is, with $h = 1$, a variant of Model A in [45]. For $h \neq 1$, excitatory spikes have different efficacies at excitatory and inhibitory neurons, respectively. For $h > 1$, this establishes a competition between the neurons that proves useful for long memory. In a study of rat neocortex, $J_{IE} \approx 2J_{EE}$ (i.e., $h = 2$) has been reported for regular spiking excitatory neurons and fast spiking inhibitory interneurons [46,47]; throughout the paper we use $h = 2$ exclusively.

D. Training scheme

The training procedure described here mimics the situation where a presynaptic assembly (a *cue*) is activated and a supervisor drives a *target* assembly in the postsynaptic population. Then, through STDP, the matrix W learns to associate the cue with the target: after some time, activating the cue will autonomously activate the target without requiring the supervisor—the hetero-association is stored.

Specifically, the goal of training is to store associations of random (independently drawn) pattern pairs $(\mathbf{q}_k, \mathbf{p}_k)$ into the network Fig. 1(c). The training procedure is illustrated in Fig. 1(d) and proceeds as follows. The zeroth association is trained by setting $\mathbf{q} = \mathbf{q}_0$ and $\mathbf{p} = \mathbf{p}_0$. First, trace variables and membrane voltages are integrated for $T_{\text{warm}} = 20$ with

frozen W . Second, all dynamical variables, including the entries of W , are integrated for a time T ; here the synapses W_{ij} follow the dynamics (1) with the specific kernel (2); see Appendix A for implementation details. We assume that there is a pause before the next pattern pair is trained. During this pause, homeostatic plasticity occurs, which, according to experimental findings, can be modeled as a slow rescaling of weights to maintain firing rates (see [20,48] and references therein) or summed synaptic weight *per postsynaptic neuron* (approximating the experimentally observed conservation of summed synapse surface area per postsynaptic neuron over time; see [49]). We follow the latter view, and thus, after each training session, we rescale each weight as

$$W_{ij} \rightarrow W_{ij} \frac{m_0}{M^{-1} \sum_{j'=1}^M W_{ij'}}, \quad (7)$$

where we introduced the parameter m_0 that defines the average synaptic weight per postsynaptic neuron. Note that each row of W sums to Mm_0 . After homeostasis, we set $\mathbf{q} = \mathbf{q}_1$ and $\mathbf{p} = \mathbf{p}_1$, and repeat the entire procedure, then proceed with $(\mathbf{q}_2, \mathbf{p}_2)$, and so on.

III. STOCHASTIC SYNAPTIC DYNAMICS

The dynamics of a synaptic weight w [Eq. (1)] is stochastic due to the randomness of η and x . Trajectories of w can be obtained in simulations; for instance, for the window function (2), one has to integrate both stochastic equations (3) and (4). The synaptic weight performs a jump process [Fig. 2(a1)]: with each new pre- or postsynaptic spike, a new set of spike pairs is formed that leads to a finite jump, and between jumps the weight is constant. In this section we develop a simplified description in terms of a Langevin equation and use it to calculate the transient mean and the variance of an ensemble of synapses.

The rate of change of synaptic weight is given by Eq. (1). Assuming that the amplitudes κ are sufficiently small and the rate of spike pairs sufficiently high, $\dot{w}$ can be approximated by a white Gaussian process. This leads to a Langevin equation

$$\dot{w} = D^{(1)}(w) + \sqrt{2D^{(2)}(w)}\zeta(t), \quad (8)$$

which, due to causality of the synaptic updates, we interpret in the sense of Itô [50]. Here $\zeta(t)$ is white Gaussian noise with $\langle \zeta(t) \rangle = 0$ and $\langle \zeta(t)\zeta(t') \rangle = \delta(t - t')$. Samples of Eq. (8) are shown in Fig. 2(a2). Correspondingly, the transition probability $p(w, t|w_0, t_0)$ of weights follows the Fokker-Planck equation

$$\begin{aligned} \partial_t p(w, t|w_0, t_0) = & -\partial_w D^{(1)}(w)p(w, t|w_0, t_0) \\ & + \partial_w^2 D^{(2)}(w)p(w, t|w_0, t_0). \end{aligned} \quad (9)$$

In this section we derive the functions $D^{(1)}(w)$ and $D^{(2)}(w)$, which we then apply to describe the evolution of an ensemble of synaptic weights.

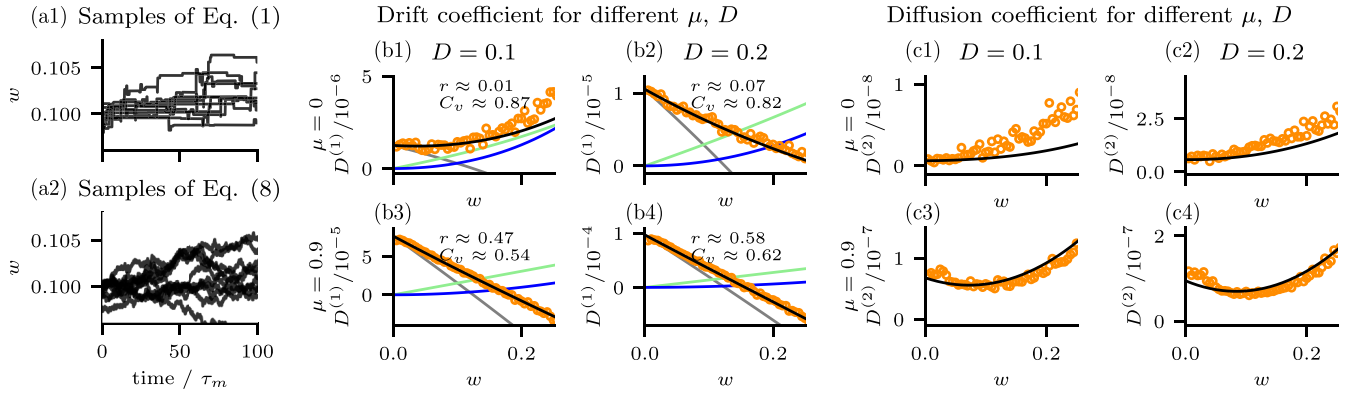


FIG. 2. Stochastic dynamics of single synapses. (a1) Sample trajectories of Eq. (1) and (a2) sample trajectories of the corresponding Langevin equation (8) for $\mu = 0.6$, $D = 0.2$, $v = 0.1$, $m_0 = 0.1$, $\sqrt{V_0} = 10^{-3}$ and STDP parameters as in Fig. 1(b). (b1)–(b4) Drift coefficient $D^{(1)}$ from simulations (orange circles) and theory [Eq. (17) (black line) and single contributions: firing-rate (gray), mean-response (green), and noise-intensity-response (blue)] for different μ and D . The instantaneous firing rate r and the coefficient of variation C_v at $w = 0.1$ are indicated in the upper center of the four panels. (c1)–(c4) Finite-time diffusion coefficient $D^{(2)}$ for $\Delta t = 10$. Simulation results (orange circles). Theory (black lines) $[V(\Delta t) + (m(\Delta t) - w)^2]/(2\Delta t)$, with m from Eq. (24) and V from Eq. (26).

A. Drift coefficient

The drift coefficient $D^{(1)}$ is the first Kramers-Moyal coefficient

$$D^{(1)}(w) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \langle w_{\text{traj}}(t + \Delta t) - w_{\text{traj}}(t) \rangle_{w_{\text{traj}}(t)=w}, \quad (10)$$

where w_{traj} is a sample of the stochastic process (1), and the subscript denotes a condition on the ensemble average. Assuming slow weight dynamics (compared to the neuron's relaxation timescale and to the width of the STDP kernel, detailed in Appendix B), Eq. (10) can be expressed by stationary statistics of the LIF neuron

$$D^{(1)}(w) = \int_{-\infty}^{\infty} d\tau \kappa(\tau, w) [vr(w) + C_{x\eta}(\tau, w)], \quad (11)$$

which is a well-known result [11]. Here $r(w) \equiv \langle x \rangle$ is the instantaneous firing rate of the postsynaptic neuron assuming weight w , and $C_{x\eta}(\tau, w) = \langle x(t + \tau)\eta(t) \rangle_w - r(w)v$ is the input-spikes-output-spikes cross-correlation. Within a diffusion approximation of the input to the LIF neuron

$$\mu + \sqrt{2D}\xi(t) + w\eta(t) \approx \mu_{\text{DA}} + \sqrt{2D_{\text{DA}}}\xi(t), \quad (12)$$

where $\mu_{\text{DA}} = \mu + wv$ and $D_{\text{DA}} = D + w^2v/2$, the instantaneous firing rate is given by [43]

$$\frac{1}{r(w)} = \int_{\frac{v_l - \mu_{\text{DA}}}{\sqrt{2D_{\text{DA}}}}}^{\frac{v_h - \mu_{\text{DA}}}{\sqrt{2D_{\text{DA}}}}} ds e^{s^2} [1 + \text{erf}(s)]; \quad (13)$$

see [51] for an efficient and stable evaluation of Eq. (13). For the STDP kernel in Eq. (2), the first, firing-rate-based, part in Eq. (11) is

$$\int_{-\infty}^{\infty} d\tau \kappa(\tau, w) vr(w) = (\Delta_c \tau_c - r_{ac} w \tau_{ac}) vr(w). \quad (14)$$

The interesting part in Eq. (11) from a spike-coding perspective is the cross-correlation $C_{x\eta}$. Since η is assumed to be Poissonian, the cross-correlation is (despite the nonlinearity of the neuron model) exactly related to the firing-rate

response [32]

$$C_{x\eta}(\tau) = v \frac{\delta}{\delta v(t)} \langle x(t + \tau) \rangle, \quad (15)$$

where $\frac{\delta}{\delta v(t)} f[v] \equiv \lim_{h \rightarrow 0} \frac{d}{dh} f[v + h\delta(t - \circ)]$ denotes a functional derivative (here $\circ$ represents the time argument of v in f). The response function of the output spikes of a LIF neuron to rate modulations is to our knowledge not known (but see [29] for such a result if the input amplitudes are exponentially distributed). However, in the diffusion approximation of the LIF neuron's input in Eq. (12), one can apply the chain rule to obtain

$$\frac{\delta \langle x(t + \tau) \rangle}{\delta v(t)} \approx w \frac{\delta \langle x(t + \tau) \rangle}{\delta \mu_{\text{DA}}(t)} + \frac{1}{2} w^2 \frac{\delta \langle x(t + \tau) \rangle}{\delta D_{\text{DA}}(t)}. \quad (16)$$

Thus, the response function in Eq. (15) can be approximated in terms of the response functions to mean- and to noise-intensity modulations. Their Fourier transforms, $\alpha(\Omega)$ (the susceptibility to mean modulations) and $\beta(\Omega)$ (the susceptibility to noise-intensity modulations), can be derived with Fokker-Planck theory and are known in terms of special functions [26,27], also presented in Appendix C. For the exponential STDP kernel in Eq. (2), the integral in Eq. (11) depends on the Laplace transform of the response functions evaluated at $1/\tau_c$ [note that due to causality, $C_{x\eta}(\tau < 0) = 0$]; since the Fourier transforms α and β are analytic, we can simply evaluate them at i/τ_c . Thus, summing up, the drift coefficient is

$$D^{(1)}(w) = (\Delta_c \tau_c - r_{ac} w \tau_{ac}) vr + \Delta_c v [\omega \alpha(i\tau_c^{-1}) + \frac{1}{2} w^2 \beta(i\tau_c^{-1})]. \quad (17)$$

$D^{(1)}$ is shown and dissected into different contributions in Fig. 2(b) and agrees with simulations of the system. The drift as a function of w may exhibit a sign change [Figs. 2(b3) and 2(b4)], in line with experimental observations in Fig. 2 of [10]. For small weights, the drift is dominated by the firing-rate contributions [first line in Eq. (17) and gray lines in Fig. 2(b)], whereas for increasingly larger weights the response

contributions dominate the drift. The mean-response contribution $\Delta_c v w \alpha (i\tau_c^{-1})$ and the noise-intensity-response contribution $\Delta_c v (1/2) w^2 \beta (i\tau_c^{-1})$ are shown in Fig. 2(b) by the green and the blue line, respectively. When computing the drift using realization-wise linearization $x \approx x_0 + \alpha * \eta$ for calculating the cross-correlation function [21], the noise-intensity response is missing. Equation (17) seems to be accurate for multiple regimes, including low firing rates and irregular spiking [note the wide range of firing rates and the large coefficients of variation C_v in the examples depicted in Fig. 2(b)]. However, for low firing rates and large weights w leading to large synaptic depression jumps, the synaptic jump process is far from a Langevin process; here Eq. (17) deviates from simulation results of the true jump process. Additionally, for very small intensity of the white Gaussian noise $D \lesssim 10^{-2}$ in Eq. (4), the diffusion approximations required for the firing rate and the response functions break down, which renders Eq. (17) inaccurate; see Appendix E for details. For other neuron models than the LIF neuron, one only needs to substitute the respective mean- and noise-intensity responses.

B. Diffusion coefficient

The diffusion coefficient is the second Kramers-Moyal coefficient,

$$D^{(2)}(w) = \frac{1}{2} \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \langle [w_{\text{traj}}(t + \Delta t) - w_{\text{traj}}(t)]^2 \rangle_{w_{\text{traj}}(t)=w}, \quad (18)$$

where the average is over sample trajectories w_{traj} of Eq. (1). Since the weight involves integrals over the product of x and η , the squared weights in Eq. (18) lead to four-point correlation functions. This makes it difficult to exactly evaluate Eq. (18) in general. However, the diffusion on timescales Δt above the inverse firing rates can be estimated by applying Wick's theorem, as an approximation since (η, x) is a non-Gaussian process, to the four-point correlation functions, which leads to (see Appendix D for details)

$$D^{(2)}(w) \approx \frac{1}{4} r v (\Delta_c^2 \tau_c + r_{ac}^2 w^2 \tau_{ac}). \quad (19)$$

Equation (19) is the infinitesimal diffusion coefficient. For noninfinitesimal Δt long enough [specifically, in Fig. 2(c), $\Delta t = 10$], the diffusion is captured by the Langevin dynamics based on Eqs. (17) and (19), as we derive in the next section.

C. Mid- and long-term evolution of an ensemble of synapses

The drift and diffusion coefficients derived above describe the process on infinitesimal times. Here we study the mean and variance of an ensemble of synapses on finite times. We provide a technically detailed analysis because the results of this section are the main building blocks of the mean-field theory of learning discussed in the next section (Sec. IV). Although the dynamics can admit a stationary solution [see, e.g., the zero of $D^{(1)}(w)$, i.e., the black solid line crosses zero in Figs. 2(b3) and 2(b4)], we are here mainly interested in transient dynamics: If the training patterns permanently (but slowly) change and the synapse is subject to slow homeostatic plasticity, the stationary state will never be reached in a learning situation. Therefore we here study the nonequilibrium dynamics of the ensemble mean and variance.

We start the discussion from a general point of view and recover that the drift- and diffusion coefficients derived above suffice for our purpose. The transition probability p of a Markov process obeys the Kramers-Moyal expansion

$$\frac{\partial}{\partial t} p(w, t | w', t') = \sum_{k=1}^{\infty} \left(-\frac{\partial}{\partial w} \right)^k [D^{(k)}(w) p(w, t | w', t')], \quad (20)$$

with the Kramers-Moyal coefficients

$$D^{(k)}(w) = \frac{1}{k!} \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \langle [w_{\text{traj}}(t + \Delta t) - w_{\text{traj}}(t)]^k \rangle_{w_{\text{traj}}(t)=w}. \quad (21)$$

From Eq. (20), we can derive the evolution of moments $\langle w^n \rangle$ in terms of Kramers-Moyal coefficients. To this end, we multiply Eq. (20) with w^n and integrate over w . On the left-hand side we have the n th moment's time derivative $\partial_t \langle w^n \rangle$ under the condition w', t' . On the right-hand side, each summand can be understood using integration by parts: the expression

$$\int w^n \left(-\frac{\partial}{\partial w} \right)^k [D^{(k)}(w) p(w, t | w', t')] dw$$

equals a sum of boundary terms plus (if $n \geq k$) $\langle n! / (n-k)! w^{n-k} D^{(k)} \rangle$. It is in our case reasonable to assume that the products $D^{(k)}(w) p(w, t | w', t')$ are vanishing and flat at the boundaries $w \in \{0, \infty\}$ and thus boundary terms do not contribute, leaving us with [52]

$$\partial_t \langle w^n \rangle = \sum_{k=1}^n \frac{n!}{(n-k)!} \langle w^{n-k} D^{(k)}(w) \rangle. \quad (22)$$

The Langevin approximation, which considers only the first two Kramers-Moyal coefficients, thus implies

$$\begin{aligned} \dot{m} &= \langle D^{(1)}(w) \rangle \\ \dot{V} &= 2[\langle w D^{(1)}(w) \rangle - \langle w \rangle \langle D^{(1)}(w) \rangle] + 2\langle D^{(2)}(w) \rangle, \end{aligned} \quad (23)$$

where $m = \langle w \rangle$ and $V = \langle (w - \langle w \rangle)^2 \rangle$ are the ensemble mean and variance. One may Taylor expand $D^{(1)}$ and $D^{(2)}$ around the self-consistent solution $m(t)$; if $D^{(1)}$ is sufficiently smooth, this leads to

$$\dot{m} = D^{(1)}(m), \quad (24)$$

$$\dot{V} = 2D^{(1)'}(m)V + \frac{r(m)v}{2} [\Delta_c^2 \tau_c + r_{ac}^2 \tau_{ac}(V + m^2)]. \quad (25)$$

We evaluate the mean dynamics Eq. (24) with a Runge-Kutta scheme; Eq. (25) can then be integrated and yields

$$\begin{aligned} V &= e^{A(t)} \left(V_0 + \int_0^t dt' 2D^{(2)}[m(t')] e^{-A(t')} \right), \\ A(t) &= \int_0^t 2D^{(1)'}[m(t')] dt' + \frac{t}{2} r v r_{ac}^2 \tau_{ac}. \end{aligned} \quad (26)$$

Equations (24) and (26) can be efficiently evaluated to obtain the mean and variance of an ensemble of synapses. These equations are thus useful formulations of the microscopic (i.e., single-synapse) model, which we leverage in the next section to derive macroscopic (i.e., population) dynamics. We also use Eqs. (24) and (26) to verify the second moment of

the stochastic increment over a finite-time window [Fig. 2(c)]; i.e., due to the invalidity of the infinitesimal diffusion coefficient, we predict the finite-time diffusion coefficient as put forward in [53]. For low mean input μ and thus low firing rates, the finite-time diffusion coefficient is monotonically growing in w [see Figs. 2(c1) and 2(c2)], in line with the infinitesimal diffusion coefficient Eq. (19). However, for larger μ and thus larger rates [Figs. 2(c3) and 2(c4)], the finite-time diffusion coefficient is nonmonotonic and minimal at a nonvanishing weight w . Experimentally, this could be tested by increasing an optogenetic stimulus (roughly, increasing μ) and observing if indeed the diffusion becomes nonmonotonic.

We finally add a remark on the truncation of the stochastic process to only two Kramers-Moyal coefficients (Fokker-Planck level). To this end, we reflect on Eq. (22). While explicitly the evolution of the n th moment depends only on the first n Kramers-Moyal coefficients, it implicitly depends on the full expansion through the probability density defining the expectation value. However, when the Kramers-Moyal coefficients are smooth enough and the spread of the ensemble small enough such that $D^{(n)}$ is well captured by its n th-order Taylor approximation, then the evolution of the n th moment depends only on moments up to order n . Consequently, in this case we achieve moment closure at order n , i.e., the system is self-consistently described by the dynamics of the first n moments.

D. Acceleration or deceleration of synaptic weights?

In Fig. 2(b) we observed acceleration [$D^{(1)}(w)$ grows with w] and deceleration [$D^{(1)}(w)$ decreases with w] of the synaptic weight. Both, acceleration and deceleration of synaptic weights seem reasonable. Acceleration seems reasonable because the potentiation of a synaptic weight increases the chance of causal spike pairs. Deceleration seems reasonable as depression scales more strongly with synaptic weight than potentiation [5], here modeled by the multiplicative nature of depression, while potentiation is additive [Eq. (2)]. Eventually, at large w , all weights must suffer a deceleration. To see this, consider the case $w \gg 1$. At this point, increasing w cannot strengthen the postsynaptic neuron's response anymore; already one presynaptic spike is guaranteed to cause a postsynaptic spike. Thus, a further increase of w does not increase the chance of causal (i.e., potentiating) spike pairs. Yet both causal (potentiating) and anticausal (depressing) spike pairs may occur by chance, the latter though being much more drastic in effect as depression scales with w . Returning to more reasonable weights $w \lesssim 0.2$ (i.e., more than five spikes needed to carry from reset voltage to threshold voltage), the drift coefficient can indeed change the sign of its slope. In Figs. 3(a) and 3(b), we show the drift coefficient for the two cases of acceleration and deceleration.

How does the transition from acceleration to deceleration depend on the dynamics of the postsynaptic cell? In order to investigate this question, we vary the mean input μ and the noise intensity D of the LIF neuron. Assuming a fixed weight (here $w = 0.1$), μ and D uniquely determine the firing rate and the coefficient of variation C_v of the postsynaptic neuron [54]. In Fig. 3(c) we show the slope of $D^{(1)}(w = 0.1)$ color coded in the plane of postsynaptic rate and C_v . We observe

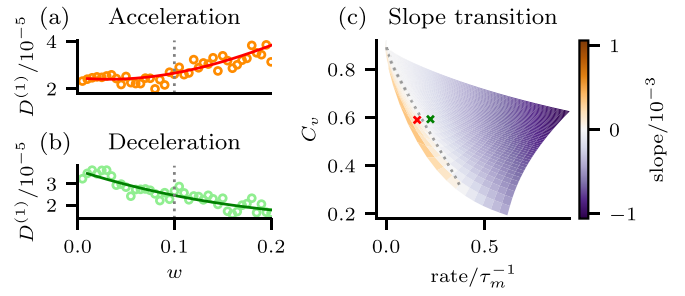


FIG. 3. Acceleration-to-deceleration transition. (a) Drift coefficient, theory (red line) and simulation (orange circles), for $\mu = 0.75$, $D = 0.028$, $r \approx 0.16$, $C_v \approx 0.58$, corresponding to the red cross in (c). (b) Drift coefficient, theory (green line) and simulation (light green circles), for $\mu = 0.71$, $D = 0.057$, $r \approx 0.23$, $C_v \approx 0.59$, corresponding to the green cross in (c). (c) Slope of $D^{(1)}(w = 0.1)$ obtained from Eq. (17) plotted against the rate and C_v of the postsynaptic neuron assuming $w = 0.1$ fixed. Other parameters as in Fig. 2.

that acceleration occurs only in the lower left corner of the rate- C_v plane (i.e., for small rates and a C_v almost linearly bound by the rate). The picture does not change qualitatively when varying the reference point w or input rate ν .

Fluorescence experiments estimating spine-volume dynamics [a proxy for $w(t)$] have found a global tendency hinting towards deceleration of synaptic weights in Fig. 2 of [10], in line with the majority of cases presented in Fig. 3(c) being decelerating. However, the results of Fig. 2 of [10] show large fluctuations with single data points in opposition to the global tendency. Our findings suggest that these fluctuations could partly be accounted for when conditioning on the rate and the C_v of the spike train. These could, for instance, be obtained using sliding averages of a length scale on which the synaptic weight does not change much. Specifically, our results suggest that in a region of low rate and low C_v , accelerating weights are overrepresented. Moreover, optogenetic stimulation of a neuron (roughly varying μ) could change a synapse from an accelerating to a decelerating regime.

IV. DYNAMICS OF LEARNING

In this section we study the network and learning scheme sketched in Figs. 1(c) and 1(d) (introduced in Secs. II C and II D). The dynamics can now be viewed on two timescales: First, on the discrete *macroscopic* timescale (pattern index k , going from one training session to the next), the data that the network is exposed to is exchanged. Since subsequent pattern pairs are uncorrelated, the sequence of matrices W_k after the k th training session including homeostasis is a Markov chain (see also [55]). Second, on the continuous *microscopic* timescale (within a training session), the present pattern pair (q_k, p_k) appears static, and the fast neural and synaptic dynamics evolve as defined in Sec. II. As we discuss in detail in this section, the system becomes stationary on the macroscopic timescale while remaining nonequilibrium on the microscopic timescale. The following analysis starts by studying the storage of a memory, namely, the association $q_k \rightarrow p_k$. This storage occurs on the microscopic timescale by exposing the system to (q_k, p_k) . The analysis proceeds

by quantifying the stationary regime on the macroscopic timescale. Assuming this regime, we finally study how memories are forgotten on the macroscopic timescale.

A. Synaptic dynamics on the microscopic timescale

Here we characterize the change of the synaptic weight matrix $W(t)$ during training session k . The starting point is the homeostatically scaled matrix from the previous session $W(0) = W_{k-1}$, where, as mentioned in the discussion of Eq. (7), each row of $W(0)$ sums to Mm_0 . The new pattern pair $(\mathbf{q}_k, \mathbf{p}_k)$ defines four synaptic populations,

$$P_{ab} = \{W_{ij} : p_{k,i} = a, q_{k,j} = b\}. \quad (27)$$

These sets sort synapses by where they point to ($a = 0$: to non-target neurons, $a = 1$: to target neurons) and where they come from ($b = 0$: from noncue neurons, $b = 1$: from cue neurons). Since $W(0)$ is still uncorrelated to $(\mathbf{q}_k, \mathbf{p}_k)$, the weights in all populations can be seen as randomly sampled from the set of weights in W_{k-1} . We describe the synaptic dynamics through the population means and variances

$$\hat{m}_{ab} = \frac{1}{|P_{ab}|} \sum_{w \in P_{ab}} w, \quad \hat{V}_{ab} = \frac{1}{|P_{ab}|} \sum_{w \in P_{ab}} (w - \hat{m}_{ab})^2, \quad (28)$$

where $|\cdot|$ here denotes the number of elements in the sets defined in Eq. (27). Strictly speaking, $\hat{m}_{ab}$ and $\hat{V}_{ab}$ are stochastic processes that depend on the realizations of the Poissonian and Gaussian processes fed into the network. Neglecting correlations, according to the central limit theorem one would estimate their fluctuations to scale as $1/\sqrt{|P_{ab}|}$; thus, because here populations are assumed to be large, it is reasonable to assume self-averaging $\hat{m}_{ab} \approx \langle \hat{m}_{ab} \rangle \equiv m_{ab}$ and $\hat{V}_{ab} \approx \langle \hat{V}_{ab} \rangle \equiv V_{ab}$, which strongly simplifies the analysis. The initial means are $m_{ab}(0) = m_0$ due to homeostasis, the initial variances $V_{ab}(0) = V_0$, where here V_0 should be understood as a parameter; later we fix it self-consistently to the stationary total variance of $W_{k \rightarrow \infty}$.

Our aim is to obtain $m_{ab}(t)$ and $V_{ab}(t)$ similarly to the single synapse case in Sec. III. To this end we first map the neurons in Eq. (5) to instances of the single neuron in Eq. (4), whose input is a mean current and Gaussian white noise with fixed intensity. Thus, we need to calculate the mean and noise intensity of the input to the neurons. The input from the Poisson layer $\sum_j W_{ij} \eta_j$ to the neurons in the target population ($a = 1$) or in the nontarget population ($a = 0$) is determined by the first- and second-order statistics of Poisson processes

$$\begin{aligned} \mu_a^{\text{cue}} &= M[m_{a1} f_c v_{\text{hi}} + m_{a0} (1 - f_c) v_{\text{lo}}], \\ D_a^{\text{cue}} &= \frac{1}{2} M[(m_{a1}^2 + V_{a1}) f_c v_{\text{hi}} + (m_{a0}^2 + V_{a0}) (1 - f_c) v_{\text{lo}}]. \end{aligned} \quad (29)$$

The total input (including cue, recurrent network, and supervision) to neurons in population a is thus by Poisson approximation as in [45]

$$\begin{aligned} \mu_a^{\text{tot}} &= \mu_E + J C_E r_E - g J C_I r_I + \mu_a^{\text{cue}} + \delta_{a1} J_s v_s, \\ D_a^{\text{tot}} &= D_E + \frac{1}{2} [J^2 C_E r_E + (gJ)^2 C_I r_I + D_a^{\text{cue}} + \delta_{a1} J_s^2 v_s], \end{aligned} \quad (30)$$

where δ_{a1} is the Kronecker symbol and r_E and r_I are the mean firing rates of the excitatory and the inhibitory neurons, respectively. These are determined by mean-field theory along the lines of [45]; see Appendix F.

As one can see from the total effective input (30), the postsynaptic neurons are, within the employed approximation, effectively decoupled apart from their common dependence on the mean fields m_{ab} , V_{ab} , r_E , and r_I , which we now determine self-consistently. Thus, assuming knowledge of the numerical values of the mean fields, the single-neuron statistics is readily determined, and we can proceed as for the single synapse in Sec. III: The drift and diffusion coefficients of the four distinct synaptic populations are, in analogy to Eqs. (17) and (19), respectively,

$$\begin{aligned} D_{ab}^{(1)}(w) &= (\Delta_c \tau_c - r_{ac} w \tau_{ac}) v_b r_a(m, V) \\ &\quad + \Delta_c v_b [w \alpha_a (i \tau_c^{-1}, m, V) + \frac{1}{2} w^2 \beta_a (i \tau_c^{-1}, m, V)], \\ D_{ab}^{(2)}(w) &= \frac{1}{4} r_a(m, V) v_b (\Delta_c^2 \tau_c + r_{ac}^2 w^2 \tau_{ac}). \end{aligned} \quad (31)$$

Note that $D^{(1)}$ and $D^{(2)}$ depend parametrically on the population-wise mean and variance of weights m and V , since these impact the rate and response of postsynaptic neurons. Lastly, we can close the self-consistency by identifying m_{ab} and V_{ab} with the ensemble average of synapses with drift and diffusion coefficients Eq. (31). Employing the method from Sec. IIIC leading to Eqs. (24) and (25), this leads to

$$\begin{aligned} \dot{m}_{ab} &= D_{ab}^{(1)}(m_{ab}; m, V), \\ \dot{V}_{ab} &= 2 D_{ab}^{(1)'}(m_{ab}; m, V) V_{ab} \\ &\quad + \frac{r_a(m, V) v_b}{2} [\Delta_c^2 \tau_c + r_{ac}^2 \tau_{ac} (V_{ab} + m_{ab}^2)], \end{aligned} \quad (32)$$

where the drift coefficient depends on m and V both as its argument and through the parametric dependence mentioned above. To summarize, the dynamics of the means and variances of the four populations in training session k is given in terms of eight coupled differential equations (32) which we integrate numerically. In Fig. 4 we show m_{ab} and $\sqrt{V_{ab}}$ as functions of time. The mean weight of all four populations grows, reflecting that their reversal points are above m_0 .

B. Stationarity on the macroscopic timescale

After evolving with STDP for the training time T , the synapses undergo homeostasis: the weights are scaled so as to fix the mean weight in W per row to m_0 . This requires rescaling of $W_{ij} \in P_{ab}$ with

$$\gamma_a = \frac{m_0}{f_c m_{a1}(T) + (1 - f_c) m_{a0}(T)}. \quad (33)$$

Thus, after the k th training session and the subsequent homeostatic process, the population means and variances of W_k are

$$\gamma_a m_{ab}(T) \text{ and } \gamma_a^2 V_{ab}(T), \quad (34)$$

respectively, and these are the separate error bars in Fig. 4.

While the total average of W_k ,

$$N_E^{-1} M^{-1} \sum_{ij} W_{k,ij} = m_0, \quad (35)$$

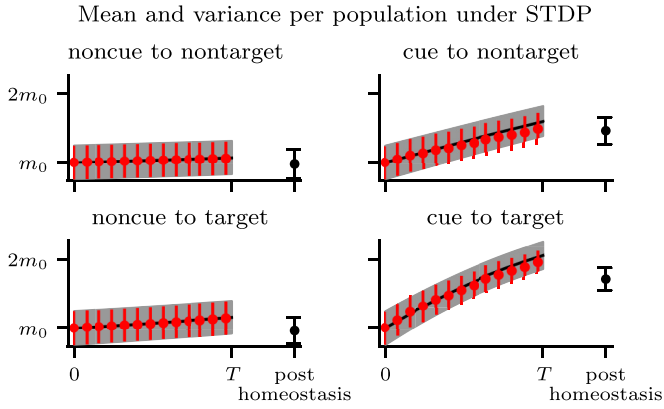


FIG. 4. Population dynamics during a training session of length $T = 50$. Mean m_{ab} and standard deviation $\sqrt{V_{ab}}$ of the four populations from Eq. (32) (black line and gray shading) and from simulations (red error bars). The separate rightmost error bars schematically illustrate the effect of homeostasis (34) (we do not consider an actual time for homeostasis). Parameters are neuron: $\mu_E = 0$, $D_E = 0.1$, $\mu_I = 0.5$, $D_I = 0.05$; network: $N_E = 4000$, $N_I = 1000$, $C_E = 200$, $C_I = 50$, $(J, g, h) = (0.01, 5, 2)$; input: $(f_c, f_s) = (0.05, 0.1)$, $M = 200$, $m_0 = 0.05$, $v_{hi} = 1$, $v_{lo} = 0.1$, $v_s = 64$, $J_s = 1/80$; and STDP: as in Fig. 1.

is fixed by construction, the same is not true for the total variance

$$V_k = N_E^{-1} M^{-1} \sum_{ij} (W_{k,ij} - m_0)^2. \quad (36)$$

However, V_k approaches a stationary value due to the interplay of STDP and homeostasis: Assuming the total variance in the beginning of the k th training session was V_* , this fixes the initial condition of all four variances in Eq. (32) to V_* . The four means and variances after STDP for time T and subsequent homeostasis, Eq. (34), thus parametrically depend on V_* . Therefore, the marginal variance of weights in W after the k th STDP session and homeostasis, V_k , depends on V_* as well. It can be expressed in terms of the results of integrating Eq. (32) as

$$V_k(V_*) = \sum_{ab}^* [\gamma_a^2 V_{ab}(T) + (\gamma_a m_{ab}(T) - m_0)^2], \quad (37)$$

where we defined the weighted population sum

$$\sum_{ab}^* X_{ab} \equiv (1 - f_s, f_s) \begin{pmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{pmatrix} \begin{pmatrix} 1 - f_c \\ f_c \end{pmatrix}. \quad (38)$$

The asymptotic variance must be stationary

$$V_* = V_k(V_*). \quad (39)$$

We solve Eq. (39) with a bisection algorithm. As tested with simulations [see Fig. 5(b)], the variance indeed approaches the result of Eq. (39) to satisfying accuracy; note that the variances shown in Fig. 5(b) are variances of subsets of synapses, all of which converge to V_* .

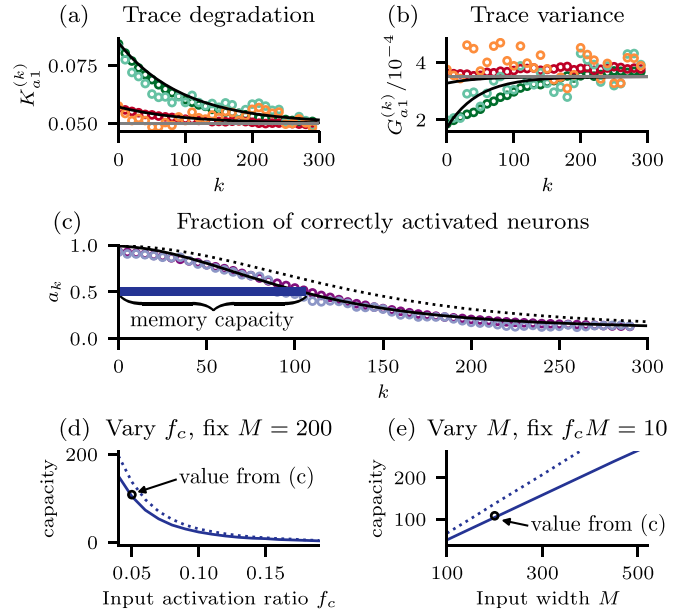


FIG. 5. Dynamics of memory traces and memory capacity. (a) Mean synaptic strength from cue to (non-)target neurons (with respect to association 0) after k subsequently stored associations. Simulation results denoted by circles, green for cue-to-target, and red for cue-to-nontarget; dark colors are averages over 100 realizations, light colors are single realizations. Theory (45) (black lines) and m_0 (gray line). (b) Variance of synaptic strength from cue to (non-)target. Colors as in (a), theory (black) from Eq. (46) and stationary variance (gray) from Eq. (39). (c) Fraction of correctly activated neurons on average over 100 realizations (dark purple circles), for a single realization (light circles) and theory (53) based on the full drift (black solid line) and neglecting cross-correlations (dotted line). (d), (e) Memory capacity (48) as a function of input sparseness f_c (d) and the input width M (e) with (solid) and without (dotted) cross-correlations. Parameters as in Fig. 4, but $v_{lo} = 0$.

C. Forgetting on the macroscopic timescale

Here we take the final steps from neuron- and synapse models to a key property of memory—the memory capacity. To this end, we compute the degradation of a memory trace from which we estimate the fraction of correctly activated neurons in an attempted recall.

1. Trace degradation

We assume that the system is in the macroscopically stationary state identified above. After the training of a specific pattern pair, say, $(\mathbf{q}_0, \mathbf{p}_0)$, the weights from $\mathbf{q}_0$ to $\mathbf{p}_0$ have mean $K_{11}^{(0)} = \gamma_1 m_{11}(T)$ and variance $G_{11}^{(0)} = \gamma_1^2 V_{11}(T)$. Analogously, the four distinct populations have mean and variance

$$K^{(0)} = \begin{pmatrix} \gamma_0 m_{00}(T) & \gamma_0 m_{01}(T) \\ \gamma_1 m_{10}(T) & \gamma_1 m_{11}(T) \end{pmatrix} \quad (40)$$

$$\text{and } G^{(0)} = \begin{pmatrix} \gamma_0^2 V_{00}(T) & \gamma_0^2 V_{01}(T) \\ \gamma_1^2 V_{10}(T) & \gamma_1^2 V_{11}(T) \end{pmatrix}. \quad (41)$$

For a successful storage of the association $\mathbf{q}_0 \rightarrow \mathbf{p}_0$ it is required that cue-to-target synapses are significantly stronger than other synapses, i.e., that $K_{11}^{(0)}$ (the mean cue-to-target

weight) exceeds the other means beyond the uncertainty captured by the variances $G^{(0)}$. This will be made rigorous below in Sec. IV C 2. Here we ask how this elevation of synapses from $\mathbf{q}_0$ to $\mathbf{p}_0$ (the reference association) decays if additional (competing) associations are trained. For instance, we track the mean $K_{11}^{(k)}$ of synapses from $\mathbf{q}_0$ to $\mathbf{p}_0$ due to the storage of k subsequent training phases. While the matrices $K^{(k)}$ and $G^{(k)}$ refer to means and variances of synapses sorted by their role in the reference association, training the k th competing association $(\mathbf{q}_k, \mathbf{p}_k)$ induces changes in all four reference populations. For instance, among the $f_s N_E \times f_c M$ synapses from $\mathbf{q}_0$ to $\mathbf{p}_0$, roughly $f_c^2 f_s^2 M N_E$ synapses are also cue-to-target synapses in association k , and $f_c^2 f_s (1 - f_s) M N_E$ synapses are cue-to-nontarget in association k . More generally, within each reference population (a, b) , four subpopulations (a', b') evolve differently. Assuming large enough neuron numbers, each subpopulation covers a fraction $[\delta_{a'0}(1 - f_s) + \delta_{a'1}f_s][\delta_{b'0}(1 - f_c) + \delta_{b'1}f_c]$ of the surrounding block (a, b) each of which consists of $[\delta_{a0}(1 - f_s) + \delta_{a1}f_s][\delta_{b0}(1 - f_c) + \delta_{b1}f_c] M N_E$ synapses. Thus, the mean and variance of each block (a, b) changes from $K_{ab}^{(k-1)}, G_{ab}^{(k-1)}$ to $K_{ab}^{(k)}, G_{ab}^{(k)}$ due to the drift and diffusion of its subpopulations (a', b') all of which are initialized at mean $m_{a'b'}(0) = K_{ab}^{(k-1)}$ and variance $V_{a'b'}(0) = G_{ab}^{(k-1)}$. Strictly speaking, the problem explodes at this point since 4×4 synaptic populations together determine their stochastic dynamics through the interaction with the heterogeneous recurrent network with different initial statistics at each training step. To study the change $K_{ab}^{(k-1)} \rightarrow K_{ab}^{(k)}$ we compute the effective shift $m_{a'b'}(T)|_{m_{a'b'}(0)} - m_{a'b'}(0)$ (where the subscript denotes the initial value) in each subpopulation. This shift depends only weakly on the initial value $m_{a'b'}(0)$ such that the precise initial value does not matter much and can be set to the mean value m_0 :

$$m_{a'b'}(T)|_{K_{ab}^{(k-1)}} - K_{ab}^{(k-1)} \approx m_{a'b'}(T)|_{m_0} - m_0.$$

We can thus express the effect of one training session on $K^{(k)}$ as a simple map:

$$K_{ab}^{(k)} = \sum_{a'b'}^* \gamma_{a'} (K_{ab}^{(k-1)} + [m_{a'b'}(T) - m_0]). \quad (42)$$

Defining

$$\begin{aligned} \phi &= 1 - \sum_{a=0}^1 [\delta_{a0}(1 - f_s) + \delta_{a1}f_s] \gamma_a, \\ c &= \sum_{ab}^* \gamma_a [m_{ab}(T) - m_0], \end{aligned} \quad (43)$$

we can recast Eq. (42) into

$$K_{ab}^{(k)} = (1 - \phi) K_{ab}^{(k-1)} + c. \quad (44)$$

This recursive equation has the solution

$$K_{ab}^{(k)} = (1 - \phi)^k K_{ab}^{(0)} + \phi^{-1} c [1 - (1 - \phi)^k] \quad (45)$$

as one can check by insertion. Thus, for small ϕ , the elevated weights relax with rate ϕ ; this relaxation rate can be fully expressed through the rescaling factors γ_a . If $\gamma_a = 1$ (i.e., no rescaling necessary to maintain m_0), the trace would not

be degraded at all. For $\gamma_a < 1$ (compensating for STDP that on average potentiates), it is rather homeostatic downscaling and less overwriting by new memories that limits the lifetime of a memory. Asymptotically, all $K_{ab}^{(k)}$ relax to the imposed population mean $m_0 = \phi^{-1}c$, as one can check by inserting Eq. (33) into Eq. (43).

An analogous argument for the variances leads to

$$\begin{aligned} G_{ab}^{(k)} &= \sum_{a'b'}^* \gamma_{a'}^2 [G_{ab}^{(k-1)} + V_{a'b'}(T) - V_* \\ &\quad + (K_{ab}^{(k-1)} + m_{a'b'}(T) - m_0)^2] - K_{ab}^{(k)2} \end{aligned} \quad (46)$$

for which we do not know an explicit solution. Still, Eq. (46) can be evaluated numerically. $K^{(k)}$ and $G^{(k)}$ are shown and compared to simulations in Figs. 5(a) and 5(b); within a few hundred patterns, $K^{(k)}$ and $G^{(k)}$ decay to their equilibrium values m_0 and V_* , respectively.

2. Recall

To get a meaningful measure of memory capacity, we investigate how well target 0 can be recalled when cue 0 is presented after k subsequent training sessions. For simplicity, we consider only the case $v_{10} = 0$, i.e., where noncue Poisson processes are silent. We follow the approach of [56] and evaluate the dendritic sums: For each postsynaptic neuron, we compute its summed synaptic weight stemming from cue 0 neurons:

$$s_i = \sum_j W_{ij} q_{0,j}. \quad (47)$$

The summed weight reflects how strongly each neuron is driven by the cue 0. If the $f_s N_E$ most strongly driven neurons are the target pattern $\mathbf{p}_0$, the association is perfectly recalled. We assume that a certain number of errors can be corrected by a downstream mechanism, e.g., by an attractor network or a perceptron. We denote neurons among the most strongly driven neurons that are indeed in $\mathbf{p}_0$ as the correctly activated neurons. Next, we derive the fraction a_k of correctly activated neurons in an attempted recall k training sessions after retrieval. The fraction a_k may be interpreted as the accuracy of an association. From the accuracy, we define the memory capacity as

$$c = \text{argmin}_k (|a_k - 0.5|), \quad (48)$$

which is the number of patterns c after which the accuracy is below 50%, $a_{k=c} \leq 0.5$, i.e., we have more false positives than correctly activated neurons. To derive the capacity we have to study the overlap between summed-weight distributions into target and nontarget neurons, respectively: If this overlap vanishes, target neurons receive significantly stronger input than nontarget neurons and can thus be clearly distinguished. If the distributions strongly overlap, it is hard to distinguish target and nontarget neurons by their activity. These considerations are detailed next.

Assuming weak correlations between elements of W we can use the central limit theorem to study Eq. (47); the summed weight to target neurons ($a = 1$) and nontarget

neurons ($a = 0$) is respectively a Gaussian random variable

$$p_a(s_i) = \mathcal{N}(s_i | f_c M K_{a1}^{(k)}, f_c M G_{a1}^{(k)}), \quad (49)$$

where the mean and variance are given by the block statistics (45) and (46). Marginalizing over all postsynaptic neurons (i.e., forgetting their identity as target or nontarget neuron with fractions f_s and $1 - f_s$, respectively), the distribution of summed weight s is

$$p_{\text{marg}}(s) = f_s \mathcal{N}(s | f_c M K_{11}^{(k)}, f_c M G_{11}^{(k)}) + (1 - f_s) \mathcal{N}(s | f_c M K_{01}^{(k)}, f_c M G_{01}^{(k)}). \quad (50)$$

The $f_s N_E$ most active neurons have $s \geq s_*$, where s_* is the $1 - f_s$ th quantile of p_{marg} :

$$\int_{s_*}^{\infty} ds p_{\text{marg}}(s) = f_s. \quad (51)$$

This integral can be evaluated to

$$\frac{f_s}{2} \text{erfc}\left(\frac{s_* - f_c M K_{11}^{(k)}}{\sqrt{2 f_c M G_{11}^{(k)}}}\right) + \frac{1 - f_s}{2} \text{erfc}\left(\frac{s_* - f_c M K_{01}^{(k)}}{\sqrt{2 f_c M G_{01}^{(k)}}}\right) = f_s, \quad (52)$$

where erfc is the complementary error function. We solve Eq. (52) numerically. Finally, the fraction of correctly activated neurons is the fraction of probability mass of Eq. (50) above s_* that is due to target neurons

$$a_k = \int_{s_*}^{\infty} ds \mathcal{N}(s | f_c M K_{11}^{(k)}, f_c M G_{11}^{(k)}) = \frac{1}{2} \text{erfc}\left(\frac{s_* - f_c M K_{11}^{(k)}}{\sqrt{2 f_c M G_{11}^{(k)}}}\right). \quad (53)$$

Equation (53) is shown in Fig. 5(c). From Eq. (53) one can compute the memory capacity. As is well known from Hopfield-like networks, training with sparse patterns is less detrimental to previous memory [57]; this observation remains valid for the spike-coding setup, as shown in Fig. 5(d). Furthermore, we observe in Figs. 5(c)–5(e) that rate-based approximations systematically and substantially overestimate the memory capacity. Our theory that incorporates spike-time-resolving cross-correlations correctly predicts memory capacities much more faithfully. Lastly, memory capacity grows linearly with the width M of the input layer [Fig. 5(e)].

D. Tradeoff between early accuracy and capacity

Neural systems are often forced to trade one benefit for another; for instance, speed-accuracy tradeoffs have been reported both experimentally and theoretically, e.g., in decision making [58] and in memory consolidation [59]. We here discuss a tradeoff between early accuracy and memory lifetime.

Naively, one could assume that the higher the early accuracy (here defined as the fraction of correctly activated neurons right after training, a_0), the higher the memory capacity. This is not always the case. In Fig. 6(a) we vary different parameters of the plasticity rule and the training scheme and find that the optimal capacity does not coincide

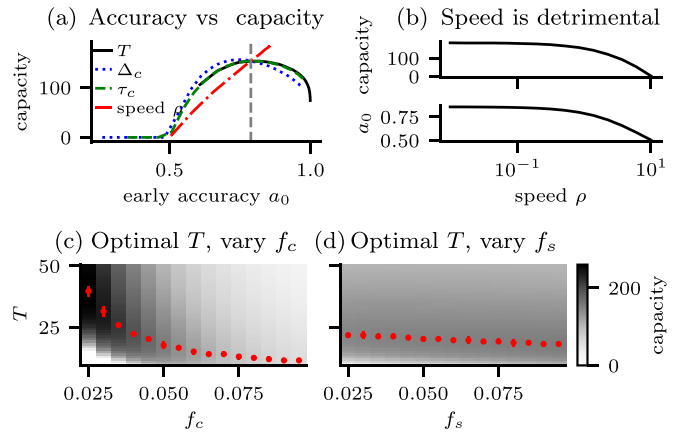


FIG. 6. How to get high capacity? (a) Memory capacity against early accuracy a_0 when varying T (black solid), Δ_c (blue dotted), τ_c (green dashed), and speed (red dash-dotted). To understand how the parameters are varied, note that $a_0(T)$, $a_0(\Delta_c)$, and $a_0(\tau_c)$ are all increasing functions of their argument. Speed here means upscaling of $\kappa(\tau)$ while downscaling T . (b) Capacity and early accuracy saturate for low speed and decrease with large speed. In (a) and (b), unless explicitly varied, $T = 17$, which optimizes the capacity at parameters from Fig. 5. (c), (d) Memory capacity gray coded for various f_c (c) or f_s (d) and T . The red bars show the regime of optimal capacity for a given sparseness f_c or f_s . Parameters that are not varied are as in Fig. 5.

with the maximum early accuracy, i.e., the maximal capacity is not at $a_0 = 1$. Specifically, when varying the amplitude of potentiation Δ_c , the timescale of potentiation τ_c , or the length of a training session T , the capacity is optimal for $a_0 < 1$. It thus seems as if a system must decide to either optimize capacity or early accuracy.

A way to avoid the tradeoff altogether is to simultaneously change both the STDP amplitude $\kappa(\tau) \rightarrow \rho \kappa(\tau)$ and the training time $T \rightarrow T/\rho$. The scaling parameter ρ determines the learning speed but keeps the mean effect of a training session on cue-to-target synapses constant. The dilemma is in fact resolved: lowering the speed ρ increases both capacity and early accuracy thus evading the capacity-accuracy tradeoff; see Fig. 6(b). If speed itself is considered a desired benefit, this means we have a speed-accuracy-capacity triple tradeoff and maximizing capacity and accuracy comes at the expense of a low learning speed.

E. Optimal exposure time

The heteroassociative network studied in this paper captures only a single stage in memory formation, which likely involves several stages and parallel pathways. Yet, assuming that the performance of single stages is related to the performance of the whole system, we can even derive qualitative predictions about behavior: The duration T of a training session can likely be behaviorally controlled by changing the time a proband is exposed to two stimuli they are supposed to associate (e.g., a sound and an image). Is there an optimal duration of exposure or is longer always better? To answer this question we follow the idea of [56] and determine the time T (corresponding to the transition probability in [56])

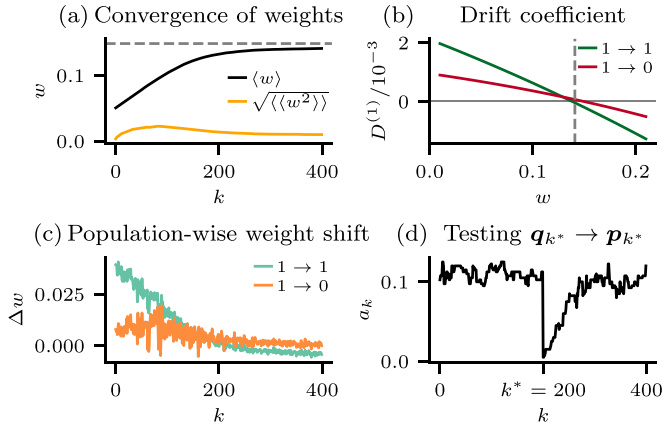


FIG. 7. Training without homeostasis. (a) Transient population mean and standard deviation of the synaptic weights over several training steps k . The dashed line is the weighted sum $f_s w_{11} + (1 - f_s) w_{01}$ of the zeros w_{11} (w_{01}) of the drift coefficient of cue-to-target (cue-to-nontarget) synapses, respectively; see (b). (b) Drift coefficient (31) of cue-to-target (green) and cue-to-nontarget (red) synapses assuming the final values in (a) for the statistics of the recurrent network. Horizontal line at zero. Vertical dashed line at asymptotic mean weight from (a). (c) Difference of synaptic weight before and after training step k , averaged over cue-to-target synapses in pattern pair k (turquoise) and over cue-to-nontarget synapses in pattern pair k (orange). (d) Fraction of correctly activated neurons when attempting to retrieve an association that is trained at step $k^* = 200$. Parameters as in Fig. 5 but with homeostasis deactivated.

that optimizes capacity. As shown in Figs. 6(c) and 6(d), the memory capacity is optimized at a finite exposure time. The quantitative value of the optimal time is quite sensitive to the presynaptic sparseness [Fig. 6(c)], yet quite insensitive to the postsynaptic sparseness [Fig. 6(d)].

F. Two complementary roles of homeostasis

Homeostasis, i.e., the slow scaling of weights (during the breaks between learning sessions), modeled by Eq. (7) is both responsible for forgetting and a requirement for learning. The responsibility for forgetting is apparent from Eq. (45): The elevated mean weight of cue-to-target synapses degrades with rate ϕ , an important ingredient of which are the homeostatic scaling factors γ_a ; see Eq. (43). However, as discussed now, without homeostasis, in the long run, new associations cannot even be learned.

If we omit the homeostatic scaling, then the overall mean synaptic weight m_0 is not enforced anymore. Instead, the mean weight converges to a stable fixed point [black line in Fig. 7(a)] with the standard deviation (orange line) being bounded, i.e., the weights do not diverge. The fixed point of the mean weight is around the zeros of the drift coefficients of cue-to-target and cue-to-nontarget synapses [Fig. 7(b)]. In the stationary regime, is it still possible to train patterns? To train pattern k , we require that the synapses from $\mathbf{q}_k$ to $\mathbf{p}_k$ grow significantly, i.e., that the difference of cue-to-target synaptic weight Δw over the k th training session is positive and large. Specifically, the Δw of cue-to-nontarget synapses must be smaller. These two competing shifts are compared in

Fig. 7(c) and display a surprising behavior: at low k where we start the weights below the fixed point, training works; however, towards the stationary regime (large k), i.e., around the fixed point, the picture is reversed and we observe a kind of antilearning. This is demonstrated by tracking the retrievability of an association, namely, for the pattern pair applied at session $k^* = 200$ (close to the stationary regime). Indeed, after training this pattern pair, it is *less likely than chance* to have correctly activated neurons in recall, reflected by the downstroke in Fig. 7(d).

We conclude that while the STDP rule by [5] is inherently stable, it does not allow for Hebbian association learning without a homeostasis mechanism shifting it far from the STDP rule's inherent stable fixed point. Moreover, we hypothesize that deactivation of homeostatic plasticity not only decreases the memory capacity, but also can even lead to antilearning.

V. DISCUSSION

In this paper we analytically described the stochastic process of the weight of synapses endowed with STDP and driven by stochastic spike trains of pre- and postsynaptic neurons. Specifically, we derived a concise description in form of a Langevin equation that captures the first two Kramers-Moyal coefficients of the true jump process. From this description we computed the dynamics of the mean and variance of an ensemble of synaptic weights. We next studied a training setup where hetero-associations are stored into a feed-forward matrix of synapses endowed with STDP. Through a mean-field theory, we mapped this setup to the single-synapse case above, which led to a quantitative description of the memory lifetime. Training and retrieval in this setting are successful (with memory capacity scaling linearly with the input width) for a range of parameters, thus the procedure is robust. We made qualitative predictions about the drift and diffusion coefficients of single synapses, about a tradeoff between early accuracy and memory capacity, and about the role of homeostatic plasticity in memory.

We included the effect of pre-post cross-correlations on the synaptic dynamics through an exact link between the cross-correlation and the neuron's response functions [32]. We found that both for the single-synapse process and for the network process, cross-correlations have a significant impact if synaptic weights are strong enough. The discrepancy between the full solution and a rate-based approximation is particularly prominent at sparse input patterns. Thus, especially in sparse-coding situations, the theory developed here is an important advancement over rate-based approaches.

Two research perspectives on memory and learning have been united in this paper. On the one hand, in the context of STDP, research often focuses on the transient (i.e., nonstationary) dynamics of ensembles of synaptic weights [11–19]. These dynamics can, e.g., be expressed by continuous-time differential equations of moments of the synaptic weights. The process described in our theory corresponds to this type of dynamics during each training session, which we referred to above as the microscopic timescale. On the other hand, memory is often described as a discrete update process of a weight matrix happening at a macroscopic timescale [55,56,60–64]. Thus, with each discrete time step, a new association is stored

such that it can partially overwrite and interfere with previous associations (termed palimpsest [61]). If such a learning scheme incorporates homeostasis, this discrete-time process approaches stationary dynamics. In our process, this second perspective is covered too: when subsampling the process once after each training session including homeostasis, the weight matrices W_k approach a stationary Markov chain.

Interestingly, we found that the degradation of the memory cue must first be attributed to homeostasis [see Eq. (43) and Eq. (45)] and thus only indirectly occurs due to the storage of new memory. Effectively, the memory degradation with homeostatic plasticity is thus relaxational as opposed to the palimpsest-like forgetting due to overwriting [61]; apparently the representations are large enough for overwriting effects to average out for many training steps and the delimiting factor is homeostatic relaxation. It would be interesting to clarify experimentally under which circumstances forgetting is rather relaxational or palimpsest-like.

The hetero-associative setting with training induced by exposure [Fig. 1(c)] can model memory dynamics on multiple stages. Assuming that the cue and target patterns are sensory representations, our setting models initial retrieval of associations. Additionally, the cue and/or target patterns could be set by a different synaptic pathway, a teacher. In the context of memory consolidation, the teacher may be a synaptic pathway via the hippocampus, which consolidates memory by transferring the association to more stable pathways, as discussed in [65].

The success of the approach at hand opens a vast set of intriguing directions: Previous theory on recurrent plasticity (e.g., Refs. [17,34]) should be revisited from the present point of view to justify approximations and find corrections. In such recurrent settings, one might need to go beyond the Poissonian-input approximation by applying the results on colored shot noise in [32] and on self-consistent power spectra in [66]. When adding recurrent plasticity to the postsynaptic population in our setup, attracting rate states can arise which certainly impact memory, especially if one considers correlated training patterns.

It has been experimentally reported that synaptic weights fluctuate even in the absence of neural activity [7,10]; consequently, one could also study an extension of our model in which the intrinsic synaptic noise is taken into account. While the change of sign of the drift coefficient [see Fig. 2(b)] has been experimentally reported, e.g., in [10], the nonmonotonic dependence of the diffusion coefficient [see Fig. 2(c)] has to our knowledge not been reported and should be investigated experimentally.

Another interesting direction concerns the recent finding that different transfer functions in multilayer perceptrons give rise to qualitatively different representation (or coding) schemes in the feature layer [67]; correspondingly, it would be quite interesting to study multistage feed-forward and locally recurrent networks of spiking neurons under STDP. When exposing the input and readout layer to data, as in the present paper, representations in the hidden layer will arise. It would be interesting to study the statistics of these representations.

Lastly, while for rate-based neural networks the joint neural and synaptic dynamics have been comprehensively described [68], a corresponding theory for spiking

(integrate-and-fire) neurons is still missing; the approach presented here is a step in that direction.

ACKNOWLEDGMENTS

We acknowledge helpful discussions with Holger Kantz and Moritz Helias. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), SFB1315, project-ID 327654276 to B.L. and R.K.

DATA AVAILABILITY

The data that support the findings of this article are openly available [69].

APPENDIX A: NUMERICAL METHODS

1. Implementation

The software used to produce the figures is published [69]. Simulations of synapses and neurons are implemented in Cython [70]. For the LIF neurons [Eq. (4)], membrane voltages are integrated with the Euler-Maruyama scheme. Numerical Kramers-Moyal coefficients in Fig. 2 are obtained from simulations that are initialized at $w(0)$, uniformly distributed in $[0,0.31]$, and thermalizing for a period of $T_{\text{warm}} = 50$. In the network [Eqs. (5) and (6)], spike detection in one time step results in voltage reset and spike delivery in the same time step. The synaptic model is simulated by integrating Eq. (3) with the Euler scheme. Additionally, whenever the weight is set to a negative value, it is clipped to zero. For the experimentally inspired parameters used in the present paper, this almost never happens. For both Eq. (4) and Eq. (3), the Euler time step is $\Delta t = 10^{-4} \tau_m$.

Equations (24) and (32) are integrated numerically using a Runge-Kutta scheme of order 5 (4).

2. Runtime

To exemplify the runtime needed for evaluation of the theory and the simulations, respectively, we here comment on the production of Fig. 5. On an AMD Ryzen 5 PRO 4650U, the evaluation of all theoretical lines took ≈ 18.4 s, and the simulation took per realization ≈ 4.5 h. The 100 realizations used in Fig. 5 have been performed in parallel on a high-performance computing cluster.

APPENDIX B: EXPRESSING SYNAPTIC DRIFT BY CROSS-CORRELATIONS

Here we provide details of the steps from the synapse model (1) to the synaptic drift in terms of cross-correlations, Eq. (11). The drift coefficient is defined by Eq. (10). As opposed to the diffusion coefficient, the limit in Eq. (10) can be written as a time derivative, since the expression in the expectation value is linear in the trajectory w_{traj} . Thus, the model can be directly plugged in such that

$$D^{(1)}(w) = \int_{-\infty}^t dt' \langle \kappa[t - t', w_{\text{traj}}(t)] \eta(t') x(t) \rangle + \int_{-\infty}^t dt' \langle \kappa[t' - t, w_{\text{traj}}(t)] \eta(t) x(t') \rangle, \quad (\text{B1})$$

where the average is over trajectories of the weight w_{traj} under the condition that $w_{\text{traj}}(t) = w$. Due to this condition, the multiplicativity of the synaptic depression (i.e., the dependence of κ on w_{traj}) does not complicate the analysis, as

$$\kappa[t' - t, w_{\text{traj}}(t)] \equiv \kappa(t' - t, w). \quad (\text{B2})$$

Thus, under the expectation value we can join the two integrals in Eq. (1):

$$D^{(1)}(w) = \int_{-\infty}^{\infty} d\tau \kappa(\tau, w) [\Theta(\tau) \langle x(t) \eta(t - \tau) \rangle + \Theta(-\tau) \langle x(t + \tau) \eta(t) \rangle]. \quad (\text{B3})$$

The final step to Eq. (11) requires one to assume stationarity of the process x , since then the term in the square brackets of Eq. (B3) becomes $\langle x \rangle \langle \eta \rangle + C_{x\eta}(\tau)$. Since the synaptic weight w may drift, x is strictly speaking nonstationary. However, on the small duration covered by the STDP window (here ≈ 40 ms), the process appears stationary since changes of the weights (and thus of the moments of x) are slow for the biophysically motivated parameters used. Additionally, since transient times of the LIF neuron are typically quick (only a few interspike intervals if the coefficient of variation is big enough [71]), these stationary statistics may be evaluated at the instantaneous value of w .

APPENDIX C: RESPONSE FUNCTIONS OF THE LIF NEURON

The susceptibilities required in Eq. (17) are the susceptibility of the LIF neuron (4) to mean- and noise-intensity modulations, respectively. These have been derived using Fokker-Planck theory in [27],

$$\alpha(\Omega) = \frac{ri\Omega/\sqrt{D_{\text{DA}}}}{i\Omega - 1} \frac{\mathcal{D}_{i\Omega-1}\left(\frac{\mu_{\text{DA}} - v_t}{\sqrt{D_{\text{DA}}}}\right) - e^{\Delta} \mathcal{D}_{i\Omega-1}\left(\frac{\mu_{\text{DA}} - v_r}{\sqrt{D_{\text{DA}}}}\right)}{\mathcal{D}_{i\Omega}\left(\frac{\mu_{\text{DA}} - v_t}{\sqrt{D_{\text{DA}}}}\right) - e^{\Delta} \mathcal{D}_{i\Omega}\left(\frac{\mu_{\text{DA}} - v_r}{\sqrt{D_{\text{DA}}}}\right)}, \quad (\text{C1})$$

$$\beta(\Omega) = \frac{ri\Omega(i\Omega - 1)}{D(2 - i\Omega)} \frac{\mathcal{D}_{i\Omega-2}\left(\frac{\mu_{\text{DA}} - v_t}{\sqrt{D_{\text{DA}}}}\right) - e^{\Delta} \mathcal{D}_{i\Omega-2}\left(\frac{\mu_{\text{DA}} - v_r}{\sqrt{D_{\text{DA}}}}\right)}{\mathcal{D}_{i\Omega}\left(\frac{\mu_{\text{DA}} - v_t}{\sqrt{D_{\text{DA}}}}\right) - e^{\Delta} \mathcal{D}_{i\Omega}\left(\frac{\mu_{\text{DA}} - v_r}{\sqrt{D_{\text{DA}}}}\right)}, \quad (\text{C2})$$

where $\mathcal{D}_z(x)$ is Whittaker's parabolic cylinder function. For the drift coefficient we need to evaluate Eqs. (C1) and (C2) at imaginary frequencies. This is convenient, as it requires one only to evaluate $\mathcal{D}_z(x)$ at $z \in \mathbb{R}$. For the figures, we use the implementation `scipy.special.pbdv(z, x)`.

APPENDIX D: APPROXIMATION OF $D^{(2)}$

The diffusion coefficient is the second Kramers-Moyal coefficient

$$D^{(2)}(w) = \frac{1}{2} \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \langle [w_{\text{traj}}(t + \Delta t) - w_{\text{traj}}(t)]^2 \rangle_{w_{\text{traj}}(t)=w}, \quad (\text{D1})$$

where w_{traj} are realizations of Eq. (1). We can express the increment as

$$w_{\text{traj}}(t + \Delta t) - w_{\text{traj}}(t) = \int_t^{t+\Delta t} dt' \dot{w}(t') dt'. \quad (\text{D2})$$

Thus,

$$D^{(2)}(w) = \frac{1}{2} \lim_{\tau \rightarrow 0} \frac{1}{\Delta t} \int_t^{t+\Delta t} dt' \times \int_t^{t+\Delta t} dt'' \langle \dot{w}(t') \dot{w}(t'') \rangle_{w_{\text{traj}}(t)=w}. \quad (\text{D3})$$

The integration domain is an area of size Δt^2 , and we have a prefactor $1/\Delta t$; thus the only terms in the integrand that contribute after taking the limit $\lim_{\Delta t \rightarrow 0}$ are those carrying a $\delta(t' - t'')$. Plugging Eq. (1) into the second moment, we get

$$\begin{aligned} & \langle \dot{w}(t') \dot{w}(t'') \rangle_{w_{\text{traj}}(t)=w} \\ &= \left\langle \left[\int_{-\infty}^0 d\tau \kappa(\tau, w) \eta(t') x(t' + \tau) + \int_{-\infty}^0 d\tau \kappa(-\tau, w) \eta(t' + \tau) x(t') \right] \right. \\ & \quad \times \left[\int_{-\infty}^0 d\tau \kappa(\tau, w) \eta(t'') x(t'' + \tau) + \int_{-\infty}^0 d\tau \kappa(-\tau, w) \eta(t'' + \tau) x(t'') \right] \Bigg\rangle. \end{aligned} \quad (\text{D4})$$

Multiplying out the product in Eq. (D4) yields four double integrals. Each integral contains a four-point correlator $\langle \eta(t_a) \eta(t_b) x(t_c) x(t_d) \rangle$, which is in general difficult to evaluate. As an approximation, we assume that we can apply Wick's theorem for the treatment of this four-point correlator even though (η, x) is not a Gaussian process. Proceeding with this assumption, we next note that in Eq. (D3), only terms carrying a $\delta(t' - t'')$ contribute. Cross-correlations between $x(t_a)$ and $\eta(t_b)$ can contain instantaneous parts $\propto \delta(t_a - t_b)$; however, these will eliminate *both* integrals in Eq. (D4) and leave no Dirac delta for the integral in Eq. (D3); such contributions thus vanish at $\Delta t \rightarrow 0$. Furthermore, as argued above, constant parts do not contribute. This leaves us with

$$\langle \eta(t_a) \eta(t_b) x(t_c) x(t_d) \rangle \approx vr(w) \delta(t_a - t_b) \delta(t_c - t_d) \quad (\text{D5})$$

as the only nonvanishing contribution. The cross products in Eq. (D4) produce constant parts and thus do not contribute. Finally, the diagonal products yield

$$\langle \dot{w}(t') \dot{w}(t'') \rangle_{w_{\text{traj}}(t)=w} = vr(w) \delta(t' - t'') \int_{-\infty}^{\infty} d\tau \kappa(\tau, w)^2 \quad (\text{D6})$$

such that

$$\begin{aligned} D^{(2)}(w) &= \frac{1}{2} r(w) v \int_{-\infty}^{\infty} d\tau \kappa(\tau, w)^2 \\ &= \frac{1}{4} r(w) v (\Delta_c^2 \tau_c + r_{ac}^2 w^2 \tau_{ac}). \end{aligned} \quad (\text{D7})$$

APPENDIX E: BREAKDOWN OF EQ. (17) FOR LOW NOISE

The LIF neuron (4) studied throughout the paper is driven by white Gaussian noise with intensity D , which aims to model internal and external noise sources. In addition, in the recurrent network (5), the noise intensity also captures the recurrent input: it is now a sum of non-negative contributions

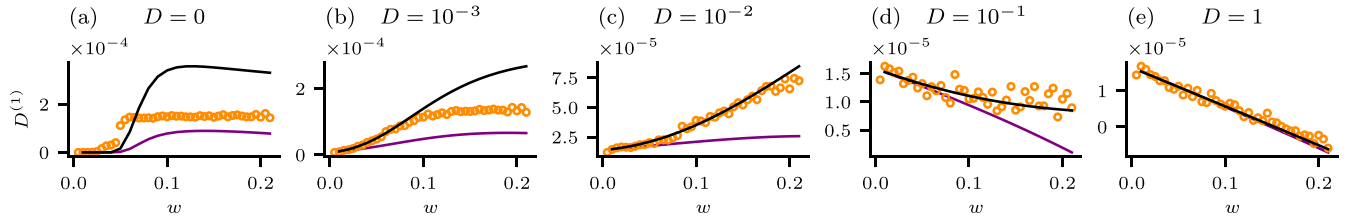


FIG. 8. Breakdown of Eq. (17) at low noise intensity. Drift coefficient (17) with (black line) and without (purple line) the impact of the noise response β , and simulation results (orange circles). The parameters correspond to the diffusion coefficient in Fig. 2(b), but the noise intensity D is varied as indicated in the title and μ is adjusted such that for $w = 0.1$, the firing rate is $r = 0.1$. The integration time step in the simulation is $dt = 10^{-4}$ and the time step to evaluate the drift coefficient is $\Delta t = 10^{-2}$, the drift coefficient is averaged over 10 000 trials.

(30) which is typically substantially large. However, in order to explore the limit case, we now consider the input current to the neuron to be almost deterministic (small D). In this case the approximations leading to Eq. (17) and Eq. (19) become problematic since the expressions for the instantaneous firing rate (13) and for the response functions (C1) and (C2) rely on the diffusion approximation. In Fig. 8 we demonstrate this breakdown: We study a synapse with weight w driving a neuron [Eq. (4)] with input $\mu + \sqrt{2D}\xi(t) + w\eta(t)$. We vary the noise intensity D and adjust μ , such that, for $w = 0.1$, the firing rate (13) of the neuron is 0.1 (corresponding to a rate of 5 Hz for a membrane time constant of 20 ms). The drift coefficient of the synaptic weight [Eq. (17)] is well captured by the theory for $D > 10^{-2}$ but breaks down below. In Fig. 8(b) we observe that for small enough synaptic weight $w \lesssim 0.1$, the theory works even below $D = 10^{-2}$; likely this reflects that for small amplitudes w , the Poissonian input is closer to its diffusion approximation.

APPENDIX F: MEAN-FIELD THEORY OF THE RECURRENT NETWORK

The drift and diffusion coefficients of feed-forward synapses in Eq. (31) depend on the input from the recurrent

network. To determine its statistics, we follow [45]; for more sophisticated approaches to population dynamics see, e.g., [72]. The main difference to [45] is that we have three populations, target, nontarget, and inhibitory neurons, and that for $h \neq 1$, inhibitory neurons receive stronger excitation than excitatory neurons. The total input to excitatory target and nontarget neurons is presented in Eq. (30). The total input to inhibitory neurons is

$$\begin{aligned}\mu_I^{\text{tot}} &= \mu_I + hJC_E r_E - gJC_I r_I, \\ D_I^{\text{tot}} &= D_I + \frac{1}{2}(hJ)^2 C_E r_E + \frac{1}{2}(gJ)^2 C_I r_I.\end{aligned}\quad (\text{F1})$$

Here the excitatory firing rate is the weighted sum of the target and the nontarget neuron's firing rate

$$r_E = f_s r_1 + (1 - f_s) r_0. \quad (\text{F2})$$

Thus, the input to all neurons is determined by the firing rate r_1 , r_0 , and r_I , which in turn are determined by Eq. (13). This self-consistent set of equations is solved using a damped saddle-point iteration.

-
- [1] L. F. Abbott and S. B. Nelson, Synaptic plasticity: Taming the beast, *Nat. Neurosci.* **3**, 1178 (2000).
 - [2] J. Sjöström and W. Gerstner, Spike-timing dependent plasticity, *Scholarpedia* **5**, 1362 (2010).
 - [3] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, A neuronal learning rule for sub-millisecond temporal coding, *Nature (London)* **383**, 76 (1996).
 - [4] H. Markram, J. Lübke, M. Frotscher, A. Roth, and B. Sakmann, Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex, *J. Physiol.* **500**, 409 (1997).
 - [5] G. Bi and M. Poo, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type, *J. Neurosci.* **18**, 10464 (1998).
 - [6] G. Bi and M. Poo, Synaptic modification by correlated activity: Hebb's postulate revisited, *Annu. Rev. Neurosci.* **24**, 139 (2001).
 - [7] N. Yasumatsu, M. Matsuzaki, T. Miyazaki, J. Noguchi, and H. Kasai, Principles of long-term dynamics of dendritic spines, *J. Neurosci.* **28**, 13592 (2008).
 - [8] R. Rosenbaum, J. E. Rubin, and B. Doiron, Short term synaptic depression imposes a frequency dependent filter on synaptic information transfer, *PLoS Comput. Biol.* **8**, e1002557 (2012).
 - [9] R. Rosenbaum, J. E. Rubin, and B. Doiron, Short-term synaptic depression and stochastic vesicle dynamics reduce and shape neuronal correlations, *J. Neurophysiol.* **109**, 475 (2013).
 - [10] A. Statman, M. Kaufman, A. Minerbi, N. E. Ziv, and N. Brenner, Synaptic size dynamics as an effectively stochastic process, *PLoS Comput. Biol.* **10**, e1003846 (2014).
 - [11] R. Kempter, W. Gerstner, and J. L. van Hemmen, Hebbian learning and spiking neurons, *Phys. Rev. E* **59**, 4498 (1999).
 - [12] M. C. W. van Rossum, G. Q. Bi, and G. G. Turrigiano, Stable Hebbian learning from spike timing-dependent plasticity, *J. Neurosci.* **20**, 8812 (2000).

- [13] A. N. Burkitt, H. Meffin, and D. B. Grayden, Spike-timing-dependent plasticity: The relationship to rate-based learning for models with weight dynamics determined by a stable fixed point, *Neural Comput.* **16**, 885 (2004).
- [14] H. Meffin, J. Besson, A. N. Burkitt, and D. B. Grayden, Learning the structure of correlated synaptic subgroups using stable and competitive spike-timing-dependent plasticity, *Phys. Rev. E* **73**, 041911 (2006).
- [15] M. Gilson, A. N. Burkitt, D. B. Grayden, D. A. Thomas, and J. L. van Hemmen, Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks. I. Input selectivity–strengthening correlated input pathways, *Biol. Cybern.* **101**, 81 (2009).
- [16] N. Frémaux and W. Gerstner, Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules, *Front. Neural Circuits* **9**, 85 (2016).
- [17] G. K. Ocker and B. Doiron, Training and spontaneous reinforcement of neuronal assemblies by spike timing plasticity, *Cereb. Cortex* **29**, 937 (2019).
- [18] A. E. Akil, R. Rosenbaum, and K. Josić, Balanced networks under spike-time dependent plasticity, *PLoS Comput. Biol.* **17**, e1008958 (2021).
- [19] B. Sosis and J. E. Rubin, Distinct dopaminergic spike-timing-dependent plasticity rules are suited to different functional roles, *bioRxiv* (2024).
- [20] T. Keck, T. Toyozumi, L. Chen, B. Doiron, D. E. Feldman, K. Fox, W. Gerstner, P. G. Haydon, M. Hübener, H.-K. Lee, J. E. Lisman, T. Rose, F. Sengpiel, D. Stellwagen, M. P. Stryker, G. G. Turrigiano, and M. C. van Rossum, Integrating Hebbian and homeostatic plasticity: The current state of the field and future research directions, *Phil. Trans. R. Soc. B* **372**, 20160158 (2017).
- [21] B. Lindner, B. Doiron, and A. Longtin, Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback, *Phys. Rev. E* **72**, 061919 (2005).
- [22] D. Grytskyy, T. Tetzlaff, M. Diesmann, and M. Helias, A unified view on weakly correlated recurrent networks, *Front. Comput. Neurosci.* **7**, 131 (2013).
- [23] M. Layer, M. Helias, and D. Dahmen, Effect of synaptic heterogeneity on neuronal coordination, *PRX Life* **2**, 013013 (2024).
- [24] W. Gerstner, Coding properties of spiking neurons: Reverse and cross-correlations, *Neural Netw.* **14**, 599 (2001).
- [25] G. K. Ocker, A. Litwin-Kumar, and B. Doiron, Self-organization of microcircuits in networks of spiking neurons with plastic synapses, *PLoS Comput. Biol.* **11**, e1004458 (2015).
- [26] N. Brunel, F. S. Chance, N. Fourcaud, and L. F. Abbott, Effects of synaptic noise and filtering on the frequency response of spiking neurons, *Phys. Rev. Lett.* **86**, 2186 (2001).
- [27] B. Lindner and L. Schimansky-Geier, Transmission of noise coded versus additive signals through a neuronal ensemble, *Phys. Rev. Lett.* **86**, 2934 (2001).
- [28] M. J. E. Richardson, Spike-train spectra and network response functions for non-linear integrate-and-fire neurons, *Biol. Cybern.* **99**, 381 (2008).
- [29] M. J. E. Richardson and R. Swarbrick, Firing-rate response of a neuron receiving excitatory and inhibitory synaptic shot noise, *Phys. Rev. Lett.* **105**, 178102 (2010).
- [30] F. Droste and B. Lindner, Exact analytical results for integrate-and-fire neurons driven by excitatory shot noise, *J. Comput. Neurosci.* **43**, 81 (2017).
- [31] F. Droste and B. Lindner, Exact results for power spectrum and susceptibility of a leaky integrate-and-fire neuron with two-state noise, *Phys. Rev. E* **95**, 012411 (2017).
- [32] J. Stubenrauch and B. Lindner, Furutsu-Novikov–like cross-correlation–response relations for systems driven by shot noise, *Phys. Rev. X* **14**, 041047 (2024).
- [33] R. Gütiğ, R. Aharonov, S. Rotter, and H. Sompolinsky, Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity, *J. Neurosci.* **23**, 3697 (2003).
- [34] Y. F. K. Kossio, S. Goedeke, C. Klos, and R.-M. Memmesheimer, Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation, *Proc. Natl. Acad. Sci. USA* **118**, e2023832118 (2021).
- [35] A. Morrison, M. Diesmann, and W. Gerstner, Phenomenological models of synaptic plasticity based on spike timing, *Biol. Cybern.* **98**, 459 (2008).
- [36] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, Rate, timing, and cooperativity jointly determine cortical synaptic plasticity, *Neuron* **32**, 1149 (2001).
- [37] J.-P. Pfister and W. Gerstner, Triplets of spikes in a model of spike timing-dependent plasticity, *J. Neurosci.* **26**, 9673 (2006).
- [38] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, Hebbian learning of pulse timing in the barn owl auditory system, in *Pulsed Neural Networks*, edited by W. Maass and C. M. Bishop (MIT Press, Cambridge, MA, 1998), pp. 353–377.
- [39] R. Kempter, W. Gerstner, and J. L. v. Hemmen, Intrinsic stabilization of output rates by spike-based Hebbian learning, *Neural Comput.* **13**, 2709 (2001).
- [40] C. Clopath, L. Büsing, E. Vasilaki, and W. Gerstner, Connectivity reflects coding: A model of voltage-based STDP with homeostasis, *Nat. Neurosci.* **13**, 344 (2010).
- [41] F. Zenke, E. J. Agnes, and W. Gerstner, Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks, *Nat. Commun.* **6**, 6922 (2015).
- [42] A. Rauch, G. La Camera, H.-R. Lüscher, W. Senn, and S. Fusi, Neocortical pyramidal cells respond as integrate-and-fire neurons to in vivo-like input currents, *J. Neurophysiol.* **90**, 1598 (2003).
- [43] N. Fourcaud and N. Brunel, Dynamics of the firing probability of noisy integrate-and-fire neurons, *Neural Comput.* **14**, 2057 (2002).
- [44] A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk, On the distribution of firing rates in networks of cortical neurons, *J. Neurosci.* **31**, 16217 (2011).
- [45] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, *J. Comput. Neurosci.* **8**, 183 (2000).
- [46] M. Beierlein, J. R. Gibson, and B. W. Connors, Two dynamically distinct inhibitory networks in layer 4 of the neocortex, *J. Neurophysiol.* **90**, 2987 (2003).
- [47] D. Bernardi, G. Doron, M. Brecht, and B. Lindner, A network model of the barrel cortex combined with a differentiator detector reproduces features of the behavioral response to single-neuron stimulation, *PLoS Comput. Biol.* **17**, e1007831 (2021).

- [48] F. Zenke and W. Gerstner, Hebbian plasticity requires compensatory processes on multiple timescales, *Phil. Trans. R. Soc. B* **372**, 20160259 (2017).
- [49] J. N. Bourne and K. M. Harris, Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP, *Hippocampus* **21**, 354 (2011).
- [50] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, Proceedings in Life Sciences (Springer, Berlin, 1985).
- [51] M. Layer, J. Senk, S. Essink, A. van Meegen, H. Bos, and M. Helias, NNMT: Mean-field based analysis tools for neuronal network models, *Front. Neuroinform.* **16**, 835657 (2022).
- [52] M. R. Rahimi Tabar, *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*, 2019th ed., Understanding Complex Systems (Springer Nature, Cham, 2019).
- [53] M. Ragwitz and H. Kantz, Indispensable finite time corrections for Fokker-Planck equations from time series data, *Phys. Rev. Lett.* **87**, 254501 (2001).
- [54] R. D. Vilela and B. Lindner, Are the input parameters of white-noise-driven integrate & fire neurons uniquely determined by rate and CV? *J. Theor. Biol.* **257**, 90 (2009).
- [55] D. J. Amit and S. Fusi, Learning in neural networks with material synapses, *Neural Comput.* **6**, 957 (1994).
- [56] N. Auer, L. Chen, J. Stubenrauch, B. Lindner, and R. Kempter, Population sparseness determines strength of Hebbian plasticity for maximal memory lifetime in associative networks, *bioRxiv* (2025).
- [57] G. Palm, Neural associative memories and sparse coding, *Neural Netw.* **37**, 165 (2013).
- [58] R. Heitz and J. Schall, Neural mechanisms of speed-accuracy tradeoff, *Neuron* **76**, 616 (2012).
- [59] B. J. Bhasin, J. L. Raymond, and M. S. Goldman, Synaptic weight dynamics underlying memory consolidation: Implications for learning rules, circuit organization, and circuit function, *Proc. Natl. Acad. Sci. USA* **121**, e2406010121 (2024).
- [60] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982).
- [61] J. P. Nadal, G. Toulouse, J. P. Changeux, and S. Dehaene, Networks of formal neurons and memory palimpsests, *Europhys. Lett.* **1**, 535 (1986).
- [62] J. T. Buckingham, Delicate Nets, Faint recollections: A study of partially connected associative network memories, Ph.D. thesis, University of Edinburgh, 1992.
- [63] D. Tyulmankov, C. Fang, A. Vadaparty, and G. R. Yang, Biological learning in key-value memory networks, *Adv. Neural Inf. Process. Syst.* **34**, 22247 (2021).
- [64] G. Iatropoulos, J. Brea, and W. Gerstner, Kernel memory networks: A unifying framework for memory modeling, *Adv. Neural Inf. Process. Syst.* **35**, 35326 (2022).
- [65] M. W. H. Remme, U. Bergmann, D. Alevi, S. Schreiber, H. Sprekeler, and R. Kempter, Hebbian plasticity in parallel synaptic pathways: A circuit mechanism for systems memory consolidation, *PLoS Comput. Biol.* **17**, e1009681 (2021).
- [66] B. Dummer, S. Wieland, and B. Lindner, Self-consistent determination of the spike-train power spectrum in a neural network with sparse connectivity, *Front. Comput. Neurosci.* **8**, 104 (2014).
- [67] A. van Meegen and H. Sompolinsky, Coding schemes in neural networks learning classification tasks, *Nat. Commun.* **16**, 3354 (2025).
- [68] D. G. Clark and L. F. Abbott, Theory of coupled neuronal-synaptic dynamics, *Phys. Rev. X* **14**, 021001 (2024).
- [69] J. Stubenrauch, N. Auer, R. Kempter, and B. Lindner, Stochastic synaptic dynamics under learning—Scripts for simulation and illustration, Zenodo (2026), <https://doi.org/10.5281/zenodo.19019985>.
- [70] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. Seljebotn, and K. Smith, Cython: The best of both worlds, *Comput. Sci. Eng.* **13**, 31 (2011).
- [71] B. Lindner, *Coherence and Stochastic Resonance in Nonlinear Dynamical Systems* (Logos Verlag, Berlin, 2002).
- [72] T. Schwalger, M. Deger, and W. Gerstner, Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size, *PLoS Comput. Biol.* **13**, e1005507 (2017).