Supplemental Material

for

“Information retention by synaptic learning rules in stochastic spiking neural networks”

Rich Pang

Physical Review E

This Supplemental Material contains:

Appendices A1–A10

Supplementary Figures S1–S5

Appendices

A1. Demonstration that expectation of count of spike time intersections depends only on area

We would like to show that the expectation of the number of spike time intersections within a shaded region of a plot like Fig 1c depends only on its area. In general we will consider shapes that are sufficiently well-behaved that they can be approximated to arbitrary accuracy by dividing them into a collection of small non-overlapping rectangles of size $dsdt$ (Fig S1a). Let the pre-synaptic spike counts in each small time window of length dt be $n_j^1, n_j^2, \ldots$, and the post-synaptic spike counts within a window of length ds be $n_i^1, n_i^2, \ldots$, where the superscript indexes the window. If we consider constant pre- and post-synaptic firing rates, r_j and r_i, respectively, then the expectation of the count of intersections within one window of area $dsdt$ is equal to

$$\mathrm{E}[n_i^1 n_j^1] = \mathrm{E}[n_i^1]\mathrm{E}[n_j^1] = r_i r_j dsdt. \tag{17}$$

Integrating over the shaded shape $\mathcal{A}$ yields

$$\int_{\mathcal{A}} r_i r_j dsdt = r_i r_j \int_{\mathcal{A}} dsdt = A r_i r_j, \tag{18}$$

where A is the area of $\mathcal{A}$.

A2. Correction term for the block approximation

For $K = \mathrm{Round}[L/(2\Delta_c)]$, we want to set $K(L^{block})^2$ equal to the area of the original gray band. That is,

$$K[(1+\epsilon)2\Delta_c]^2 = L^2 - (L-\Delta_c)^2 \tag{19}$$

so

$$\epsilon = \sqrt{\frac{L^2-(L-\Delta_c)^2}{4\Delta_c^2 K}} - 1 = \sqrt{\frac{2L\Delta_c - \Delta_c^2}{4K\Delta_c^2}} - 1 = \sqrt{\frac{L-\Delta_c/2}{2K\Delta_c}} - 1. \tag{20}$$

A3. Gaussian approximation for counting coincidence events

To compute the Gaussian approximations in Fig 1f and S2b we first computed the mean and variance of the distribution estimated via the block approximation. We then computed the Gaussian probability density associated with this mean and variance, discretized it over integer support, and normalized the corresponding probability mass to sum to 1.

A4. Block approximation for biphasic STDP

For the biphasic STDP rule (Fig 1b, bottom) we assume that if a pair of pre- and post-synaptic spikes occur within Δ_c of each other, a plasticity event is triggered corresponding to a weight change of +1 if the post- follows the pre-synaptic spike and -1 if the pre- follows the post-synaptic spike [2, 3]. Geometrically, the final weight is the sum of intersections in the top half of the diagonal band shown in Fig S1b minus the sum of intersections in the bottom half of the band. To approximate this count distribution while accounting for correlations between coincidence events, we take a related approach to the monophasic block approximation (Fig 1c), except here breaking each block into a left half corresponding to potentiation events, and a right half corresponding to depression events, with the total block size L^{block} computed in the same manner as for monophasic STDP (Fig S1b).

As for the monophasic STDP analysis, the total count of plasticity events is the sum of plasticity events in each block, since each block is independent. Within each block, however, there is a dependency between the event counts in the left (potentiation) vs right (depression) halves of the block, due to the shared post-synaptic spike count $n_i^{block} \sim \mathrm{Poisson}(r_i L^{block})$. The pre-synaptic spike counts in the left and right half of the blocks are $n_{j+}^{block} \sim \mathrm{Poisson}(r_j L^{block}/2)$ and $n_{j-}^{block} \sim \mathrm{Poisson}(r_j L^{block}/2)$, respectively. The count of potentiation intersections minus depression intersections within a block is then given by

$$c^{block} = n_i^{block} n_{j+}^{block} - n_i^{block} n_{j-}^{block}. \tag{21}$$

Since n_i^{block}, n_{j+}^{block} and n_{j-}^{block} are all independent, the distribution over c^{block} is then given by

$$P(c^{block}) = \sum_{n_i^{block}} \sum_{n_{j+}^{block}} \sum_{n_{j-}^{block}} P(n_i^{block}) P(n_{j+}^{block}) P(n_{j-}^{block}) \mathbb{1}[c^{block} = n_i^{block} n_{j+}^{block} - n_i^{block} n_{j-}^{block}], \quad (22)$$

which can be directly computed if a reasonable truncation of the sum is admitted. As in the monophasic STDP rule, the distribution $P(c)$ over the total (signed) count of plasticity events across all blocks is given by the $(K-1)$-fold convolution of $P(c^{block})$ with itself.

A5. Joint entropy of weight changes of multiple synapses sharing one neuron

Here we compute the "star motif" entropies for the STDP and sum rules (Fig 3d, left). We can compute this entropy via a natural extension of the block approximation for the monophasic STDP rule. Without loss of generality, consider m synapses $w_{ij^1}, w_{ij^2}, \dots w_{ij^m}$, i.e. all pointing away from neuron i (the directionality of synapses does not matter for the entropy calculation for monophasic STDP or sum rules). Let c_{ij}^{block} be the coincidence count for synapse w_{ij}:

$$c_{ij^1}^{block} = n_i^{block} n_{j^1}^{block}, \quad c_{ij^2}^{block} = n_i^{block} n_{j^2}^{block}, \quad \dots c_{ij^m}^{block} = n_i^{block} n_{j^m}^{block}. \quad (23)$$

Recalling that $n_i^{block}, n_{j^1}^{block}, \dots, n_{j^m}^{block}$ are all independent, we find that the total joint probability distribution is

$$\begin{aligned} P(c_{ij^1}^{block}, c_{ij^2}^{block}, \dots, c_{ij^m}^{block}) = \sum_{n_i^{block}} \sum_{n_{j^1}^{block}} \cdots \sum_{n_{j^m}^{block}} P(n_i^{block}) P(n_{j^1}^{block}) \dots P(n_{j^m}^{block}) \\ \times \mathbb{1}\left[c_{ij^1}^{block} = n_i^{block} n_{j^1}^{block} \;\&\; c_{ij^2}^{block} = n_i^{block} n_{j^2}^{block} \;\&\; \dots \;\&\; c_{ij^m}^{block} = n_i^{block} n_{j^m}^{block}\right]. \end{aligned} \quad (24)$$

The final distribution summed over blocks is then computed as before via the $(K-1)$-fold m-dimensional convolution of $P(c_{ij^1}^{block}, c_{ij^2}^{block}, \dots, c_{ij^m}^{block})$ with itself and its entropy computed directly. We call this entropy h_{m*}^R, which reduces to the single-synapse entropy when $m = 1$ and the pairwise entropy of two synapses sharing one neuron when $m = 2$.

Unfortunately, computing h_{m*}^R quickly becomes intractable as m increases due to the curse of dimensionality. To make this computation tractable, we used a truncated support for the Poisson distributions $P(n_i^{block}), P(n_{j^1}^{block}), \dots$ starting from 0 and capturing a fraction $1 - 10^{-4}$ of the total probability mass. Additionally, in performing the final $(K-1)$-fold convolution, we systematically reduced the support after each convolution by truncating the resulting distribution within a bounding box that retained a fraction $1 - 10^{-5}$ of the total probability mass. Using these approximations, we succeeded in computing h_{m*}^R up to $m = 5$ for the STDP rule and $m = 6$ for the sum rule.

To estimate h_{m*}^R for larger m, we employed a further approximation. We first noted the decay of Δh_{m*}^R with m, where $\Delta h_{m*}^R = h_{m*}^R - h_{m-1*}^R$ (Fig 3d) along with the observation that Δh_{m*}^R should remain above zero for large m since adding a new synapse will always introduce at least some additional entropy into the joint distribution, following from the entropy of the new neuron associated with that synapse. To capture these trends parsimoniously we fit a simple extrapolation model $\hat{\Delta h}_{m*}^R = A \exp(-m/\sigma_m) + b$ using a standard constrained optimization algorithm (scipy.optimize.curve_fit, [74]; Fig 3d). Most values of Δ_c^* produced reasonable fits (Fig S4a), but some yielded pathological parameter values (e.g. $b < 0$). We identified these pathological cases by manual inspection, removed them from the fits, then resampled b via a cubic interpolation between the minimal and maximal b for the non-pathological fits, and directly recomputed A and σ_m to retain the same value of $\hat{\Delta h}_{1*}^R$ and its first derivative $d\hat{\Delta h}_{m*}^R/dm$ evaluated at $m = 1$. Example final fits are shown in Fig 3d. We emphasize that our essential goal here was to capture the qualitatively observed decay to a plateau of Δh_{m*}^R in order to provide a tractable extrapolation of h_{m*}^R for large star motif sizes m beyond those which we were able to compute directly via Eq. 24, rather than deriving a principled expression for these quantities. When computing network entropy (A6), we used the directly computed h_{m*}^R when possible ($m = 5$ for the STDP rule and $m = 6$ for the sum rule) and the extrapolated values otherwise.

A6. Estimation of STDP and sum rule entropy in a network

We employed the following procedure to estimate the joint synaptic weight entropy produced by either STDP or the sum rule in a recurrent Poisson spiking network. Our central goal is to account

for as large a portion as possible of the correlation structure of the synapses in the network in an analytically tractable manner. We first randomly sample a set of M synapses $\{w_{ij}\}$ to be plastic from a total of $N^2 - N$ possible synapses (self-connections area excluded). We then estimate the entropy of the network, h^R_{Ntwk}, leveraging the star motif entropies computed above (Appendix A5) and two important approximations. Approximation 1, is: if a set of synapses Ω_1 share no neurons with a second set of synapses Ω_2, then the weights of Ω_1 are *conditionally independent* of Ω_2 given all intermediate synapses; where intermediate synapses are all additional synapses included in all paths from any synapse in Ω_1 to any synapse in Ω_2; and where a path from w_{ij} to w_{kl} means any undirected sequence (i.e. we ignore which direction synapses point) of synapses linking neuron i or j to neuron k or l. In on our original network model, synapses sharing no neurons are only marginally independent, hence conditional independence imposes further structure.

The algorithm is initialized by adding the first synapse to the network, which simply increases the total network entropy from 0 to h^R_1, the entropy of a single synapse given the learning rule (STDP or the sum rule), which was computed in our single-synapse analysis (Figs 1-2). The algorithm then proceeds recursively by adding in the remaining $M-1$ synapses one-by-one and updating the joint entropy in turn. Let h^R_m be the joint entropy of m synapses already included in the computation. To include synapse $m+1$ and compute h^R_{m+1}, we first define the following. Let Ω_B correspond to the "neighborhood" of synapse $m+1$, i.e. all existing synapses that share one neuron with synapse $m+1$ (Fig 3c, lower right). Let Ω_A correspond to all existing m synapses excluding Ω_B, the neighborhood of the next synapse, and let Ω_C corresponds to synapse $m+1$ itself. By Approximation 1, we assume Ω_C is conditionally independent of Ω_A given Ω_B. Leveraging the chain rule for entropy in combination with this approximation yields

$$\begin{aligned} h^R_{m+1} &\equiv H[\Omega_A, \Omega_B, \Omega_C] = H[\Omega_A, \Omega_B] + H[\Omega_C|\Omega_A, \Omega_B] \approx H[\Omega_A, \Omega_B] + H[\Omega_C|\Omega_B] \\ &= H[\Omega_A, \Omega_B] + H[\Omega_B, \Omega_C] - H[\Omega_B] = h^R_m + H[\Omega_B, \Omega_C] - H[\Omega_B], \end{aligned} \tag{25}$$

as depicted in Fig 3c.

We must thus compute $H[\Omega_B]$ and $H[\Omega_B, \Omega_C]$. To do so, let neurons i and j be the post- and pre-synaptic neurons of synapse $m+1$, respectively. Let Ω_i and Ω_j be the sets of synapses emanating (regardless of direction) from i and j, respectively, excluding synapse $m+1$. We now employ Approximation 2: assume $P(\Omega_i, \Omega_j) = P(\Omega_i)P(\Omega_j)$ and $P(\Omega_i, \Omega_j|\Omega_C) = P(\Omega_i|\Omega_C)P(\Omega_j|\Omega_C)$. I.e. we assume Ω_i and Ω_j are both marginally independent as well as conditionally independent given synapse $m+1$ (which is equivalent to Ω_C)—essentially, this ignores any shared neurons between Ω_i and Ω_j. Then

$$H[\Omega_B] \approx H[\Omega_i] + H[\Omega_j] = h^R_{m_i *} + h^R_{m_j *} \tag{26}$$

where $m_i = |\Omega_i|$ and $m_j = |\Omega_j|$ and h^R_{m*} is the star motif entropy computed above (Appendix A5). Similarly,

$$\begin{aligned} H[\Omega_B, \Omega_C] &\approx H[\Omega_i, \Omega_C, \Omega_j] = H[\Omega_i, \Omega_C] + H[\Omega_j|\Omega_i, \Omega_C] \\ &\approx H[\Omega_i, \Omega_C] + H[\Omega_j|\Omega_C] = H[\Omega_i, \Omega_C] + H[\Omega_j, \Omega_C] - H[\Omega_C] \\ &= h^R_{m_i+1*} + h^R_{m_j+1*} - h^R_1. \end{aligned} \tag{27}$$

When possible we use the directly computed values of h^R_{m*} ($m \leq 5$ for STDP and $m \leq 6$ for the sum rule), otherwise we use the extrapolation of these quantities described in Appendix A5. As expected, Approximation 2 is used rarely for sparse networks and increasingly frequently as the density of plastic synapses increases (Fig S4b).

A7. Estimation of entropy from samples

To estimate entropy of count distributions from simulations with finite samples, we expand the entropy in $1/N_{samp}$, fit a first-order model,

$$H_{samp}\left(\frac{1}{N_{samp}}\right) = H_{true} + \gamma\left(\frac{1}{N_{samp}}\right) + \cdots \tag{28}$$

then extrapolate the model prediction to $1/N_{samp} \to 0$ [75, 76]. In our simulations we first generate 30,000 independent samples of the total plasticity event counts. Then, using $N_{samp} \in [500, 250, 100, 50]$ we resample N_{samp} samples from the total collection of samples and compute the entropy averaged over 2000 trials of this resampling process. We then use least-squares regression to estimate H_{true}.

A8. Physically constrainted optimal information retention simulations

Illustrative simulations of deterministic gating of plasticity in Fig 4a were performed by assuming each synapse was independent, i.e. no synapses shared any pre- or post-synaptic neurons, and sampling each synapse's pre- and post-synaptic spike trains from a Poisson process, with equal rates $r = 0.5$ Hz across all neurons. The parameters of each simulation were chosen such that the total information retention (entropy of the final joint synaptic weight changes) was approximately equal. Specifically, Simulation A used $L = 3$s and an STDP rule with $\Delta_c^* = .109$, Simulation B used $L = 1$s and the 1-factor rule, and Simulation C used $L = 2.2$s and an STDP rule with $\Delta_c^* = 0.8515$.

A9. Information retention with stochastic plasticity gating

Here we compute the information stored in synaptic weights about spike trains when the plasticity gating signals are stochastic. We divide the learning session of length T into Q periods, each of length L, the duration of an individual plasticity period. Let e_{ij}^q be the eligibility trace at synapse (i, j) in period q, determined from the spikes in that period via plasticity rule R. Let $g_{ij}^q \in \{0, 1\}$ be the gating signal at synapse (i, j) in period q. If $g_{ij}^q = 1$, the synapse is plastic during period q, and the eligibility trace e_{ij}^q is converted to a weight change $w_{ij}^q \leftarrow e_{ij}^q$; if $g_{ij}^q = 0$ then the synapse does not change during period q, regardless of e_{ij}^q. We will also assume $p_g \equiv p(g_{ij}^q = 1) \ll 1$, i.e. the probability of a non-zero gating signal is small, such that the chance of two gating non-zero signals occurring the in the same synapse over the Q periods is negligible (sparse plasticity). Our precise goal is then to compute:

$$\begin{aligned} \mathrm{MI}[\{w_{ij}\}; \{t_k^l\}] &= H[\{t_k^l\}] - H[\{t_k^l\}|\{w_{ij}\}] \\ &= H[\{t_k^l\}] - \sum_{\{v_{ij}\}} p(\{w_{ij}\} = \{v_{ij}\}) H[\{t_k^l\}|\{w_{ij}\} = \{v_{ij}\}] \end{aligned} \tag{29}$$

where $\{v_{ij}\}$ represents a specific instance of the weights. Let there be $M_{total} = |\{w_{ij}\}| = |\{v_{ij}\}|$ synapses in total (which are not all necessary made plastic during the learning session).

For simplicity and to illustrate our key scientific points, in all of the analysis in these section we use binary eligibility traces $e_{ij}^q \in \{0, 1\}$, with $p_e \equiv p(e_{ij}^q = 1)$ and $p(e_{ij}^q = 0) = 1 - p_e$. Further, let $h^e = -p_e \log p_e - (1 - p_e) \log(1 - p_e)$ be the entropy of e_{ij}^q. We first assume that all synapses are independent, such that e_{ij}^q are independent across all periods and synapses. Since the eligibility traces are a deterministic function of the spikes, and $\{w_{ij}\}$ is independent of the spikes given the eligibility traces, we have

$$\begin{aligned} \mathrm{MI}[\{w_{ij}\}; \{t_k^l\}] &= \mathrm{MI}[\{w_{ij}\}; \{e_{ij}^q\}] \\ &= H[\{e_{ij}^q\}] - \sum_{\{v_{ij}\}} P(\{w_{ij}\} = \{v_{ij}\}) H[\{e_{ij}^q\}|\{w_{ij}\} = \{v_{ij}\}]. \end{aligned} \tag{30}$$

That is, once the eligiblity traces are known, the precise spike timing that gave rise to them can have no further influence on the weights. As all eligibility traces are independent, $H[\{e_{ij}^q\}] = QM_{total}h^e$. Finally, assume that $P(\{w_{ij}\})$ is sufficiently dominated by typical values of $\{w_{ij}\}$ that we can approximate

$$\sum_{\{v_{ij}\}} P(\{w_{ij}\} = \{v_{ij}\}) H[\{e_{ij}^q\}|\{w_{ij}\} = \{v_{ij}\}] \approx H[\{e_{ij}^q\}|\{w_{ij}\} = \{v_{ij}\}], \tag{31}$$

where we assume $\{v_{ij}\}$ on the right-hand side is a *typical* final set of weights at the end of the learning session, i.e. the number of potentiated synapses is very close to its mean. Thus, our remaining task is to compute the right hand side of Eq. 31, which is a functional of $p(\{e_{ij}^q\}|\{v_{ij}\})$. The central challenge in computing the latter is marginalizing over the gating signals $\{g_{ij}^q\}$.

Here we consider two specific cases. In Case 1, Random Gating of Plasticity, we will assume all gating signals g_{ij}^q are independent. In Case 2, Clustered Gating of Plasticity, we will assume that the gating signals are clustered; that is, we organize the synapses into clusters and assume that in any given period all synapses in a cluster receive the same gating signal.

Case 1: Random Gating of Plasticity

Given $\{v_{ij}\}$, the eligibility trace sequences $\{\mathbf{e}_{ij}\}$, where $\mathbf{e}_{ij} \equiv [e_{ij}^1, \ldots, e_{ij}^Q]^T$, are independent across synapses: $p(\{e_{ij}^q\}|\{v_{ij}\}) = \prod_{(i,j)} p(\mathbf{e}_{ij}|v_{ij})$, allowing us to simply add the entropies of each term in the

product (Fig S5a). Letting $\mathbf{g}_{ij} \equiv [g_{ij}^1, \dots g_{ij}^Q]^T$ be the gating sequence of synapse (i,j), we have

$$p(\mathbf{e}_{ij}|v_{ij}) = \sum_{\mathbf{g}_{ij}} p(\mathbf{e}_{ij}, \mathbf{g}_{ij}|v_{ij}) = \sum_{\mathbf{g}_{ij}} \frac{p(\mathbf{e}_{ij}, \mathbf{g}_{ij}, v_{ij})}{p(v_{ij})} = \sum_{\mathbf{g}_{ij}} \frac{p(\mathbf{e}_{ij})p(\mathbf{g}_{ij})\mathbb{1}[v_{ij} = \mathbf{e}_{ij}^T \mathbf{g}_{ij}]}{p(v_{ij})}, \tag{32}$$

where the final equality follows from the fact that $v_{ij} = \mathbf{e}_{ij}^T \mathbf{g}_{ij}$ is a deterministic function of $\mathbf{e}_{ij}$ and $\mathbf{g}_{ij}$. Define

$$f_{\alpha\beta\omega} = p(\mathbf{e}_{ij} = [1, 1, \dots, 0]^T) p(\mathbf{g}_{ij} = [1, 0, \dots, 0]^T) / p(v_{ij} = \omega) \tag{33}$$

where $\mathbf{e}_{ij}$ contains α 1's, and $\mathbf{g}_{ij}$ contains β 1's; $f_{\alpha\beta\omega}$ is equal for all arrangements of 1's in $\mathbf{e}_{ij}$ and $\mathbf{g}_{ij}$. Each term in the right hand side of Eq. 33 can be computed directly: $p(\mathbf{e}_{ij}) = p_e^\alpha (1-p_e)^{Q-\alpha}$, $p(\mathbf{g}_{ij}) = p_g^\beta (1-p_g)^{Q-\beta}$,

$$p(v_{ij} = 0) = \sum_{\mathbf{g}_{ij}} p(v_{ij} = 0|\mathbf{g}_{ij}) p(\mathbf{g}_{ij}) = (1-p_g)^Q + Q p_g (1-p_g)^{Q-1}(1-p_e), \tag{34}$$

and similarly

$$p(v_{ij} = 1) = Q p_g (1-p_g)^{Q-1} p_e, \tag{35}$$

where we have assumed $p(\mathbf{g}_{ij})$ is negligible for any $\beta > 1$, since $p_g \ll 1$. Then, using Eq. 32 and Fig S5b as a guide to sum over $\mathbf{g}_{ij}$, we have that

$$\begin{aligned} p(\mathbf{e}_{ij}|v_{ij} = 0) &= \sum_{\mathbf{g}_{ij}} f_{\alpha\beta 0} \mathbb{1}[v_{ij} = \mathbf{e}_{ij}^T \mathbf{g}_{ij}] = f_{\alpha 00} + (Q-\alpha) f_{\alpha 10} \\ p(\mathbf{e}_{ij}|v_{ij} = 1) &= \sum_{\mathbf{g}_{ij}} f_{\alpha\beta 1} \mathbb{1}[v_{ij} = \mathbf{e}_{ij}^T \mathbf{g}_{ij}] = \alpha f_{\alpha 11} \end{aligned} \tag{36}$$

where we recall that α is a function of $\mathbf{e}_{ij}$.

The entropies are

$$h_0^{rand} \equiv H[\mathbf{e}_{ij}|v_{ij} = 0] = \sum_{\mathbf{e}_{ij}} p(\mathbf{e}_{ij}|v_{ij} = 0)(-\log p(\mathbf{e}_{ij}|v_{ij} = 0)) \tag{37}$$

and

$$h_1^{rand} \equiv H[\mathbf{e}_{ij}|v_{ij} = 1] = \sum_{\mathbf{e}_{ij}} p(\mathbf{e}_{ij}|v_{ij} = 1)(-\log p(\mathbf{e}_{ij}|v_{ij} = 1)), \tag{38}$$

where h_0^{rand} (h_1^{rand}) is the entropy of the eligibility trace sequence for an unpotentiated (potentiated) synapse, i.e. $v_{ij} = 0$ ($v_{ij} = 1$).

Using Fig S5b again as a guide, but now to sum over $\mathbf{e}_{ij}$, we find:

$$h_0^{rand} = \sum_{\alpha=0}^{Q} \binom{Q}{\alpha} (f_{\alpha 00} + (Q-\alpha) f_{\alpha 10})(-\log[f_{\alpha 00} + (Q-\alpha) f_{\alpha 10}]) \tag{39}$$

and

$$h_1^{rand} = \sum_{\alpha=0}^{Q} \binom{Q}{\alpha} \alpha f_{\alpha 11}(-\log \alpha f_{\alpha 11}), \tag{40}$$

where we have grouped terms by α. The final entropy is therefore

$$H[\{e_{ij}^q\}|\{w_{ij}\} = \{v_{ij}\}] = m_0^{syn} h_0^{rand} + m_1^{syn} h_1^{rand} \tag{41}$$

where m_0^{syn} and m_1^{syn} are the number of non-potentiated and potentiated synapses. For a typical sparse $\{v_{ij}\}$ where the number of potentiated synapses is very close to its mean, and assuming different synapses have been potentiated at each timestep (sparse gating of plasticity) we have $m_1^{syn} = Q M_{total} p_e p_g$, and $m_0^{syn} = M_{total} - m_1^{syn}$.

Case 2: Clustered Gating of Plasticity

Here we organize the M_{total} synapses into clusters of Z synapses each, yielding $Y = M_{total}/Z$ clusters. We will assume clusters are large: $Z \gg 1$. In any period, all synapses in the same cluster receive the same gating signal. Let $Z_0 = (1-p_e)Z$ and $Z_1 = p_e Z$ be the mean number of synapses experiencing an eligibility trace of 0 and 1, respectively, in a cluster. Let p_g^{clust} be the probability that a cluster is plastic during one period. We would like the expected total number of plastic synapses in each period to be equal to the Random Gating case, so we need $p_g M_{total} = p_g^{clust} YZ$, hence p_g^{clust} equals p_g.

Given $\{v_{ij}\}$, the eligibility traces $E_y \in \{0,1\}^{Z\times Q}$ in cluster y are independent from all other clusters. Thus, it suffices to compute the entropy associated with each cluster individually (Fig S5c). Assuming $p_g \ll 1$, each cluster can be plastic in at most one period. For nonzero p_e and large Z, it will also be extremely unlikely that a cluster receiving a gating signal experiences no potentiation events. Hence, clusters in which any weight $v_{ij} = 1$ almost certainly experienced exactly one gating signal, and clusters in which all $v_{ij} = 0$ almost certainly experienced no gating signals. We will also assume $p_e Z = Z_1$ is sufficiently large that the number of potentiated synapses in a cluster that experienced a gating signal is highly concentrated around the mean Z_1. Let $\mathbf{w}_y$ be the weights in cluster y and $\mathbf{v}_y$ be a typical set of potentiated weights in the cluster, i.e. containing about Z_1 1's and Z_0 0's. Let $G_y \in \{0,1\}^{Z\times Q}$ be the gating signals in cluster y. If a cluster received no gating signal, no information is retained about the eligibility traces, hence their entropy is equal to their prior entropy

$$h_0^{clust} \equiv H[E_y|\mathbf{w}_y = \mathbf{0}] = ZQh^e. \tag{42}$$

To determine the entropy of a cluster that received a gating signal, we compute

$$p(E_y|\mathbf{v}_y) = \sum_{G_y} p(E_y, G_y|\mathbf{v}_y) = \sum_{G_y} \frac{p(E_y, G_y, \mathbf{v}_y)}{p(\mathbf{v}_y)} = \sum_{G_y} \frac{p(E_y)p(G_y)\mathbb{1}[\mathbf{v}_y = \sum_q E_y^q \odot G_y^q]}{p(\mathbf{v}_y)} \tag{43}$$

where $\odot$ is the element-wise product and E_y^q and G_y^q are the q-th columns of E_y and G_y, respectively (Fig S5d). We will also assume that $p(E_y|\mathbf{v}_y)$ is dominated by typical E_y (containing about Z_1 1's in each column), such that it will suffice to assume most E_y are typical when computing the entropy. Define

$$f^* = p(E_y^*, G_y^*|\mathbf{v}_y) = p(E_y^*)p(G_y^*)/p(\mathbf{v}_y), \tag{44}$$

where here E_y^* contains at least one column equal to $\mathbf{v}_y$ and all others columns containing Z_1 1's, G_y^* contains one column of 1's with all other columns equal to $\mathbf{0}$, and $\mathbf{v}_y$ is a representative typical vector of weights containing Z_1 1's. We are allowed to ignore E_y containing zero copies of $\mathbf{v}_y$ because these terms vanish in the right-hand side of Eq. 43 (since there is no way such eligibility traces could have been converted into weights $\mathbf{v}_y$). Then,

$$\begin{aligned} p(E_y^*) = p_e^{Z_1 Q}(1-p_e)^{Z_0 Q} \qquad p(G_y^*) = p_g(1-p_g)^{Q-1} \\ p(\mathbf{v}_y) = \sum_{G_y} p(\mathbf{v}_y|G_y)p(G_y) = Qp(G_y^*)p(\mathbf{v}_y|G_y^*) = Qp_g(1-p_g)^{Q-1}p_e^{Z_1}(1-p_e)^{Z_0}, \end{aligned} \tag{45}$$

where G_y^* has one column of 1's (the cluster was plastic at exactly one timestep). Using Fig S5d as a guide to sum over G_y, we arrive at

$$h_1^{clust} \equiv H[E_y|\mathbf{w}_y = \mathbf{v}_y] \approx \sum_{\eta=1}^{Q} \binom{Q}{\eta} \left[\binom{Z}{Z_1} - 1\right]^{Q-\eta} \eta f^*(-\log \eta f^*), \tag{46}$$

where η is the number of copies of $\mathbf{v}_y$ contained in the columns of E_y and the two multiplicity terms correspond to (1) the number of ways to arrange η copies of $\mathbf{v}_y$ and (2) the number of ways to arrange the nonzero eligibility traces in each remaining column (the -1 excludes the arrangement equal to $\mathbf{v}_y$), respectively. For large Z and Z_1 the probability mass is concentrated in the first term of the sum, hence we can further simplify

$$h_1^{clust} \approx -\log f^*. \tag{47}$$

Finally, we have

$$H[\{e_{ij}^q\}|\{w_{ij}\} = \{v_{ij}\}] \approx y_0^{clust} h_0^{clust} + y_1^{clust} h_1^{clust}, \tag{48}$$

where $y_1^{clust} \approx QYp_g$ clusters are typically written to over the learning session and $y_0^{clust} = Y - y_1^{clust}$ are left undisturbed.

Case 2.1: Post-synaptically Clustered Gating of Plasticity

When gating is clustered but eligibility traces are independent, then at low firing rates the sum rule stores the most information, followed by the 1-factor and then the STDP rule (Fig 2e). When synapses in a cluster share a post-synaptic neuron, however, information storage using the sum rule degrades, approaching that of a pre-synaptic 1-factor rule. This is because across Z synapses in a cluster, the information stored by the pre-synaptic 1-factor rule in one plasticity period is Zh_1^{1f} and including the post-synaptic activity can increase this entropy by at most $\sim O(1)$ bits, which is negligible for large Z. Thus, post-synaptically clustering effectively reduces the information stored by the sum rule to that of the 1-factor rule.

In our analysis we therefore assumed that when gating signals were post-synaptically clustered, the pre-synaptic 1-factor rule was used to convert spike trains into plasticity events. We then sought to compare the information stored in this scenario to that stored in independent synapes obeying the sum rule but with fully random gating. For a given amount of total information stored in the post-synaptically clustered case (y-axis in Fig 4j), we used the curves in Fig 4h (right) to determine p_e, converted this to an entropy $h^e = -p_e \log p_e - (1-p_e)\log(1-p_e)$, then converted this entropy to λ using the 1-factor curve in Fig 2e. The energy was then determined via $L = r\lambda$ and Eq. 15, using $M = 10^7$ synapses, $\gamma = 1$, and $r = 1$ Hz (note that the value of r does not affect the results). To compute the equivalent information for Random Gating we looked up the entropy for the sum rule at the same λ (Fig 2e), converted this to a new p_e, then used this p_e to compute the information for the Random Gating case (Fig 4j, black).

Curiously, information storage is not maximized for $p_e = .5$ for the Random Gating case. In particular, for a given fixed total energy, decreasing λ and in turn p_e can (1) increase information stored per synapse, as well as (2) increase the total number of synapses that can be used, since decreasing λ decreases L, allowing M to be increased as per Eq. 15. For the Random+ case (Fig 4j, teal), we therefore decreased λ and increased M, holding the energy (originally determined from the post-synaptically clustered case) fixed, until information storage was maximized. We then reported this total information, divided by the original number of synapses used in the post-synaptically clustered case, such that the curves in Fig 4j compare in the total information stored in the three cases in the same units.

A10. Estimation of plateau potential rate

We estimated the rate of spontaneous plateau potentials in CA3 based on experimental data reported in [8]. In this data only 14 out of 185 CA3 pyramidal neurons recorded exhibited at least one plateau potential during the recording. Assuming Poissonian plateaus, we can estimate the mean rate of plateaus, $\lambda_{plateau}$, per recording as

$$P_{poiss}(n_{plateau} = 0|\lambda_{plateau}) = e^{-\lambda_{plateau}} = 171/185 \quad \text{so} \quad \lambda_{plateau} \approx .08 \tag{49}$$

Although durations of individual neural recordings were not reported, mice were trained in 1-hour behavioral sessions prior to neural recordings. Using a range of 20 minutes to 1 hour as a stand-in recording length produces an overall plateau rate in the range of 10^{-5} to 10^{-4} Hz (which we additionally note is less than the rate of plateau potentials produced in CA1 [27, 28, 29]).

Supplementary figures

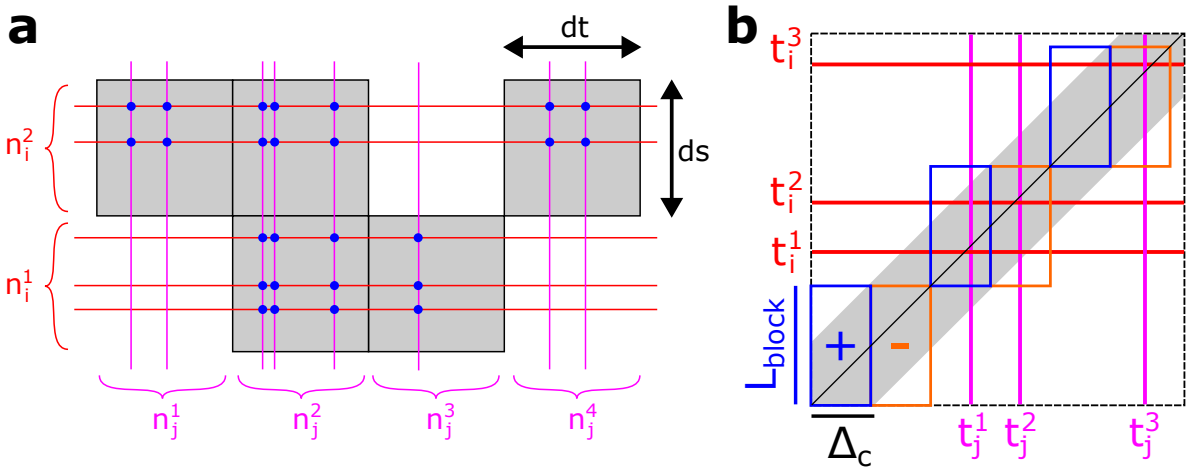

Figure S1: **Auxiliary diagrams for block approximations.** **a**. Schematic demonstrating how expected number of pre- and post-synaptic intersections depends only on shaded area (Appendix A1). **b**. Schematic of block approximation for anti-symmetric biphasic STDP rule (Appendix A4).

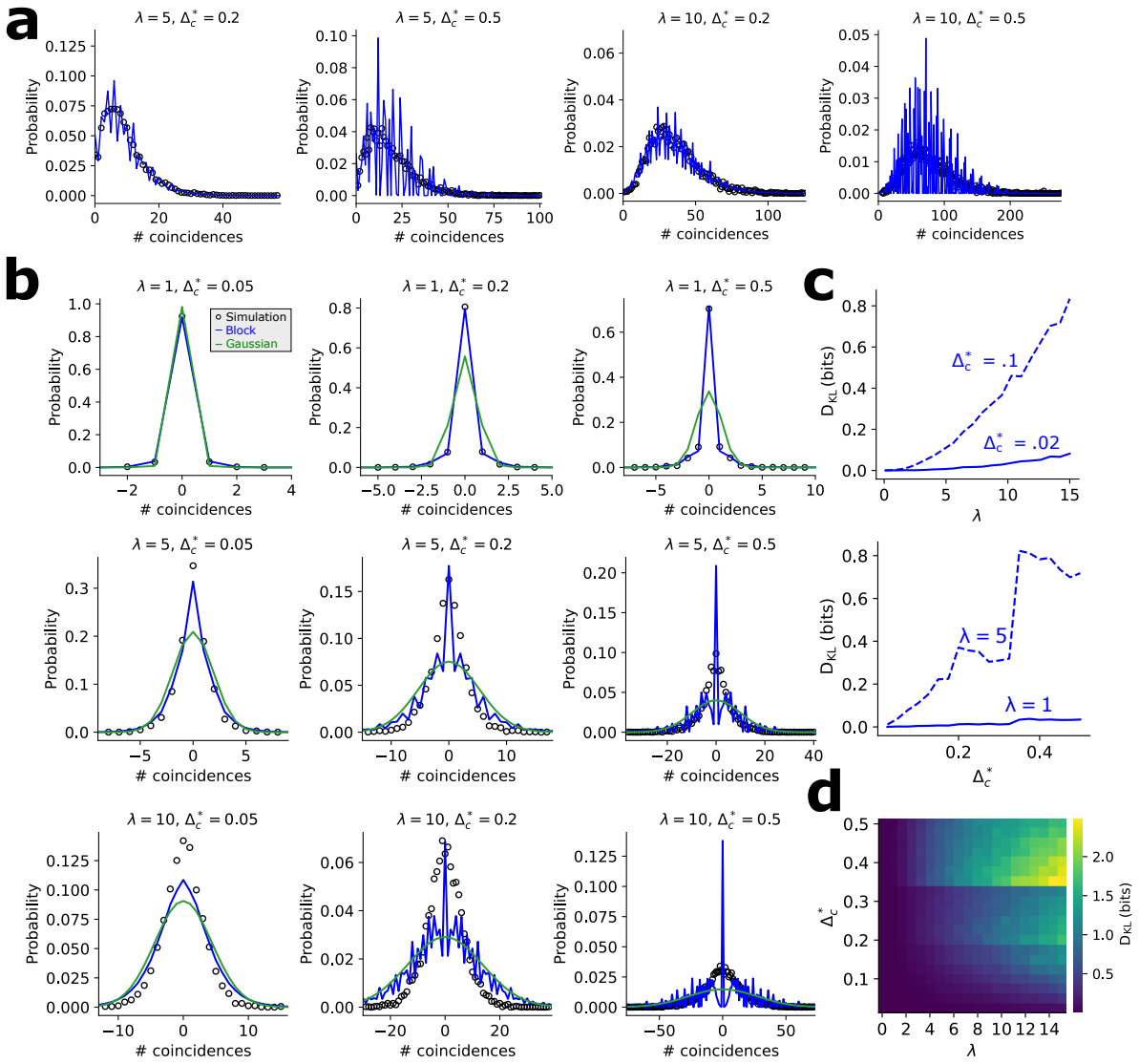

Figure S2: **Block approximations for STDP rules.** **a**. Example coincidence event distributions using the block approximation (blue) vs simulations (black circles) at high λ and Δ_c^* for the monophasic STDP rule. **b**. Example block approximations (blue), simulations, and Gaussian approximation (green) for the biphasic STDP rule for several parameters. **c**. Kullback-Leibler divergences (D_{KL}) between coincidence count distributions estimated via the block approximation and simulations vs λ for example Δ_c^* (top) or vs Δ_c^* for example λ (bottom) for the biphasic STDP rule. **d**. Heatmap of the D_{KL} computed in **c**, for a range of λ and Δ_c^*.

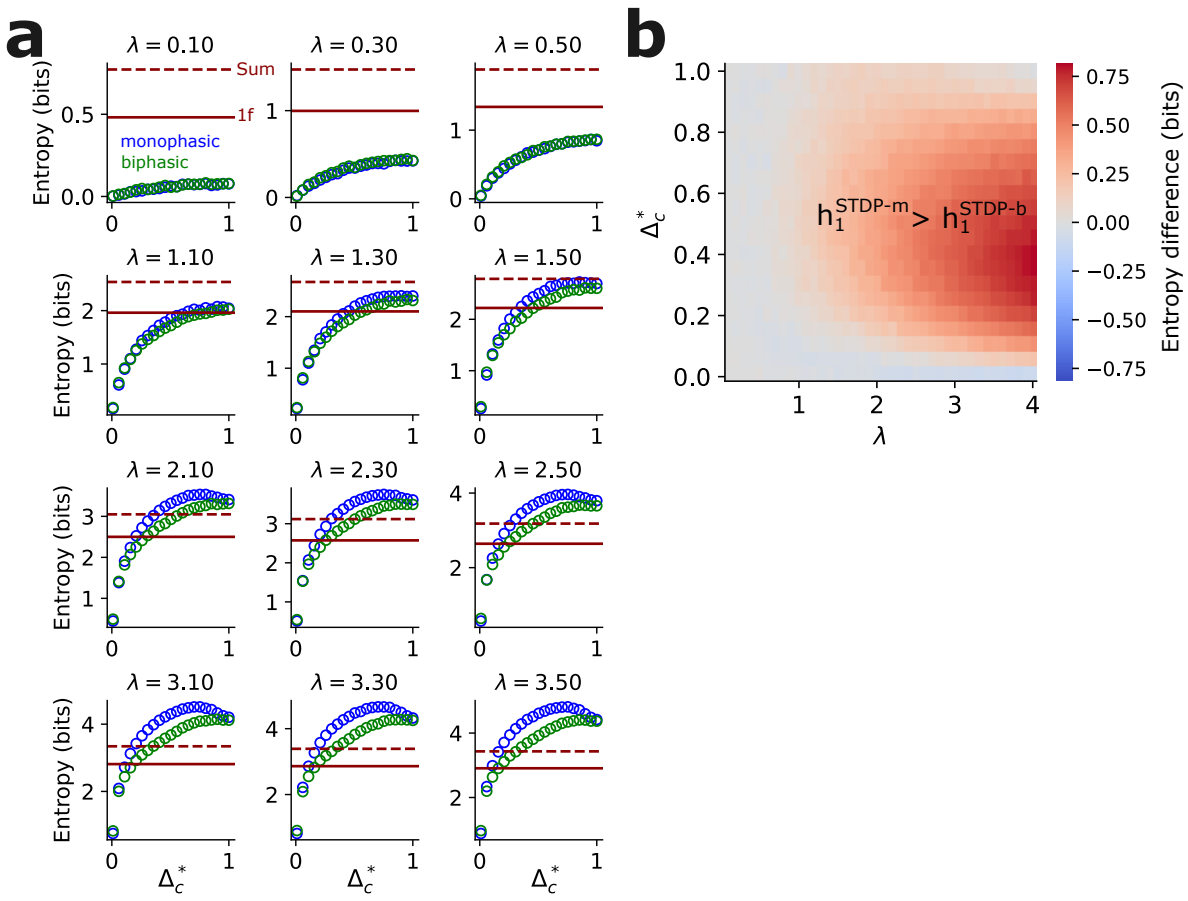

Figure S3: **Information retention by mono- vs biphasic STDP rules. a**. Single-synapse weight entropies produced by mono- vs biphasic STDP rules, estimated from simulations, vs Δ_c^* and λ. Weight entropies produced by 1-factor and sum rules are shown in the horizontal lines for each λ plot. **b**. Heatmap representation of entropy difference between the mono- (STDP-m) and bi-phasic (STDP-b) rules.

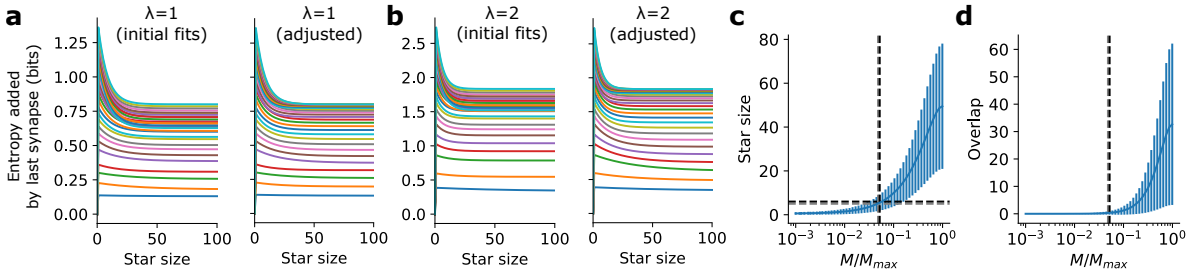

Figure S4: **Meta-analysis of network entropy computations. a-b**. Initial and adjusted fits of the entropies of different star motif sizes for the STDP rule for Δ_c^* ranging from $0.01, 0.02, \ldots 0.3$ (colored lines), with $\Delta_c^* = 0.01$ corresponding to the bottom line (Appendix A5). **c**. Mean and standard deviation (error bars) of star motif sizes accessed during the computation of the network entropy. Gray and black horizontal dashed lines indicate star sizes of 5 (STDP) and 6 (sum rule) synapses, and the vertical lines the corresponding network sizes at which these occur. **d**. Mean and standard deviations of overlap sizes (number of shared neurons between Ω_i and Ω_j; Fig 3c) encountered during the computation of the network entropy for STDP and the sum rule. Vertical lines are as in **c**.

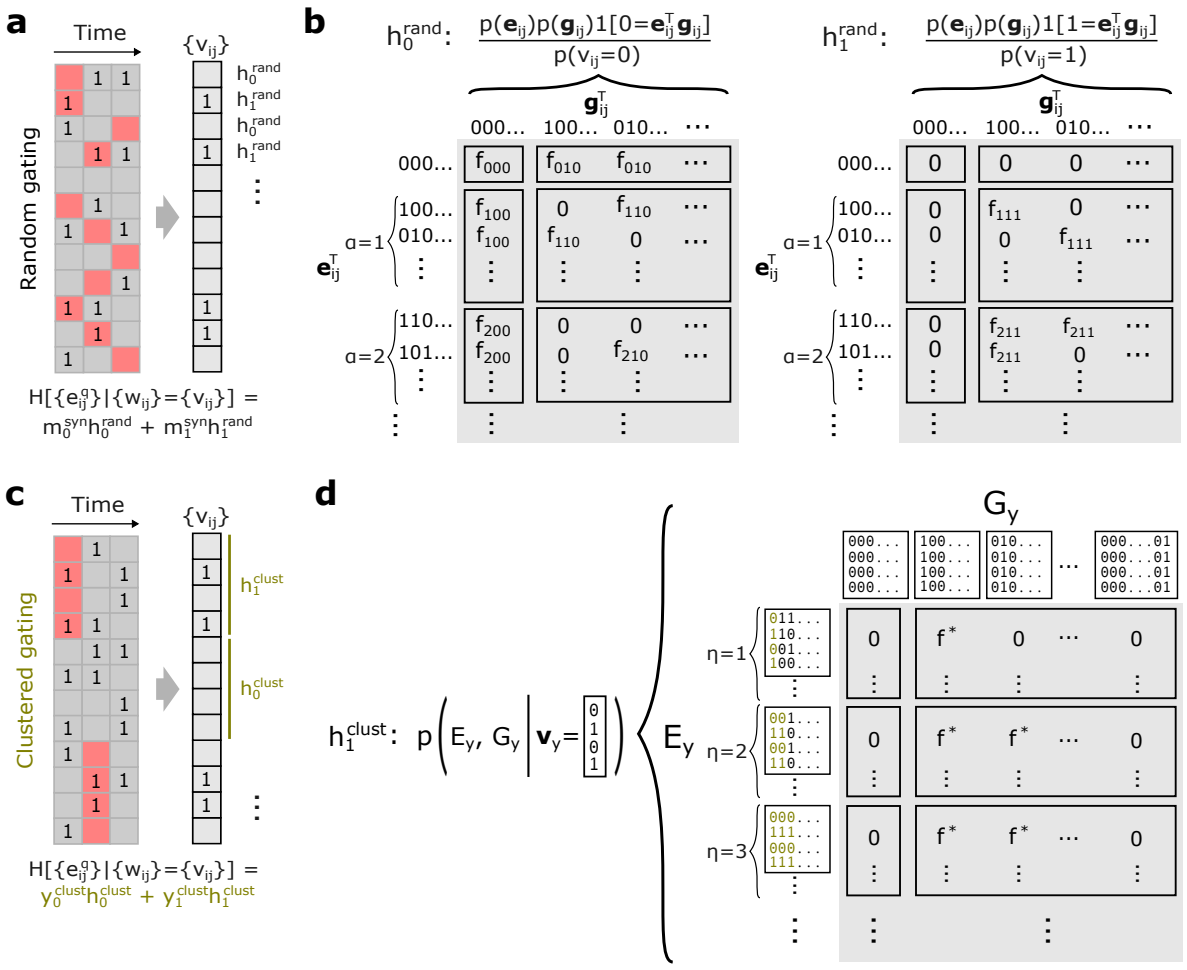

Figure S5: **Computing information retention with stochastic plasticity gating. a**. Schematic of Random Gating of Plasticity (Appendix A9, Case 1). Inscribed 1's indicate positive eligibility traces and pink shading positive gating signals. **b**. Diagram indicating how the eligibility trace entropies conditional on single-synapse weights are computed. First the joint distribution $p(\mathbf{e}_{ij}, \mathbf{g}_{ij}|v_{ij})$ is determined, then summing over $\mathbf{g}_{ij}$ yields $p(\mathbf{e}_{ij}|v_{ij})$, whose entropy is determined using Shannon's formula [32]. **c**. Schematic of Clustered Gating of Plasticity (Appendix A9, Case 2). **d**. Diagram indicating how the eligibility trace entropies given a typical potentiated cluster are computed. First the joint distribution $p(E_y, G_y|\mathbf{v}_y)$ is determined, then we sum over G_y to obtain the $p(E_y|\mathbf{v}_y)$ and compute its entropy.