Quantifying information stored in synaptic connections rather than in firing activities of neural networks

Xinhao Fan[1,*], **Shreesh P Mysore** [2,1,3,*]
[1] The Solomon H. Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA.
[2] Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA.
[3] Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD, USA.
[*] Corresponding author. (xfan20@jhu.edu, mysore@jhu.edu)

Summary

A cornerstone of our understanding of both biological and artificial neural networks is that they store information in the strengths of synaptic connections among the neurons. However, in contrast to the well-established theory for quantifying information encoded by the firing activity of neural networks, there does not exist a framework for quantifying information stored in the network's connection distribution itself. Here, we develop a theoretical framework for synaptic information by using densely connected Hebbian networks performing autoassociative memory tasks and by modeling data patterns to be stored as log-normal distributions. Specifically, we derive analytical approximations for Shannon mutual information between the data and singletons, pairs, and arbitrary n-tuples of synaptic connections within the network. Our framework corroborates well-established insights regarding pattern storage capacity, supports the principle of distributed coding in neural firing activities, and formalizes the heterogeneity inherent in information encoding across synapses in a network. Notably, it discovers synergistic interactions among synapses, revealing that the information encoded jointly by all the synapses exceeds the 'sum of its parts'. Taken together, this study introduces a powerful, interpretable framework for quantitatively understanding information storage in the synapses of neural networks, one that illustrates the duality of synaptic connectivity and neural population activity in learning and memory.

1 Introduction

The study of neural networks through the lens of information theory has a long and rich history (Borst and Theunissen, 1999; Quian Quiroga and Panzeri, 2009; Dimitrov et al., 2011; Timme and Lapish, 2018). The analyses therein have focused typically on measuring information encoded by the firing activities of neurons, both in biological (Bialek and Rieke, 1992; Palmer

et al., 2015) and artificial neural networks (Linsker, 1988; Tishby and Zaslavsky, 2015). By contrast, the study of neural networks through the lens of learning and memory posits that information is inherently stored in the synaptic connections, with neural activity patterns serving as a manifestation of the underlying changes in connectivity. This concept was notably articulated by Hebb (Hebb, 1949), in his seminal theory of memory and learning, where he described a progression from synaptic modifications to the formation of cell assemblies, and eventually to the development of "phase sequences" representing sequential activation of assemblies. Building on this perspective, subsequent research has demonstrated that connectivity encodes essential information about the world through mechanisms such as changes in synaptic strength (Lamprecht and LeDoux, 2004), the formation of wiring motifs (Sporns and Kötter, 2004; Battiston et al., 2017), and the emergence of synaptic ensemble patterns in learning (Baeg et al., 2007; Bassett et al., 2011) and working memory (Mongillo et al., 2008; Stokes, 2015; Panichello et al., 2024).

The theoretical analysis of information stored in the distribution of synaptic connections, however, remains largely under-explored. A few studies in the field of machine learning have invoked the concept of 'information in weights' for the purpose of regularization during the training of neural networks. They have done so, typically, by introducing noise into weights (Hinton and Van Camp, 1993; Achille and Soatto, 2017). However, these analyses primarily provide an upper bound on the information stored in connections, which, while useful for practical applications in network training, do not directly address the question of how much information is actually encoded in the connections. Moreover, for the sake of theoretical simplicity, these approaches often assume that individual connections are probabilistically independent, thereby neglecting the inherently collective nature of coding by synaptic ensembles.

An alternative line of research, focused on associative neural networks such as Hopfield networks and Willshaw models, first estimates the network's overall storage capacity based on patterns retrievable in the neural firing activities, and then derives the 'information per synapse' by dividing the capacity by the total number of connections (Willshaw et al., 1969; Nadal, 1991; Palm and Sommer, 1996; Bosch and Kurfess, 1998; Knoblauch et al., 2010). However, this approach continues to rely on firing patterns to infer synaptic coding, thereby still fundamentally adopting a firing activity perspective. Moreover, division by the synapse count oversimplifies the contributions of individual connections - by assuming uniform efficacy across synapses and by failing to account for the heterogeneous roles that different connections may play in information coding. Importantly, the collective coding of information by synaptic *ensembles* remains unaddressed.

The major hurdle in quantifying the information stored in synaptic weights is the complex and opaque relationship between data and weight distributions in both biological and artificial neural networks. Unlike the more explicit mapping between data and neural firing activities, weight distributions arise implicitly from the network learning process, rather than through an explicit or closed-form transformation of the data distributions. Consequently, a theoretical analysis of information encoded in ensembles of *connections*, as opposed to ensembles of *cells*, has remained a largely open question.

In this study, we propose a foundational framework for synaptic coding based on dense Hebbian connectivity, and data patterns assumed to follow a mixture of log-normal distributions (without loss of generality). The accessibility of the weight values of Hebbian networks,

and the tractability of the selected distribution in calculating mutual information allow us to derive analytical approximations for the information encoded by individual as well as arbitrarily sized ensembles of synaptic connections, ranging from pairs and triplets to larger n-tuples. These derivations incorporate two approximations related to the log-normal distribution. The resulting expressions validate established insights about distributed coding and storage capacity from the traditional perspective of network firing activities. Additionally, they reveal intriguing new insights about information synergy among synaptic connections in neural networks. Overall, this research fills an important gap in the field by introducing a theoretical framework for characterizing information storage in neural networks, and highlights the dual significance of synaptic connectivity and neural population activity in neural coding.

2 Analytical framework: Information stored in synaptic connections

We begin by considering the N data patterns that are to be stored as independent, d-dimensional random variables. We denote them as $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$. Each of these random variables is assumed to follow a multivariate log-normal distribution, a formulation that captures the heavy-tailed statistics observed broadly in diverse neuronal populations (Buzsáki and Mizuseki, 2014). Formally, this is expressed as:

$$\ln(\mathbf{x}^k) \sim \mathcal{N}(\mu^k, \Sigma^k), \qquad \mathbf{x}^k \perp \mathbf{x}^l, \forall k \neq l \tag{1}$$

where μ^k is a d-dimensional vector representing the mean of the kth pattern in the logarithmic domain, and Σ^k is the associated $d \times d$ covariance matrix.

For our neural network model, we employ the densely connected network with synaptic weights set according to the Hebbian learning rule. The connectivity matrix of such networks is of dyadic form with continuous real values such that each element is determined entirely by the data patterns. Specifically, the synaptic connection between neuron i and neuron j is given by:

$$w_{ij} = \sum_{k=1}^{N} \mathbf{x}_i^k \mathbf{x}_j^k \tag{2}$$

With this foundation in place, we depart from the traditional approach of investigating the mutual information between the continuous firing dynamics of neurons and the data patterns, and focus instead on the Shannon mutual information between the synaptic connectivity among neurons and the data patterns. The existence of a closed-form expression for the weights in the Hebbian network (eqn. 2) facilitates this endeavor.

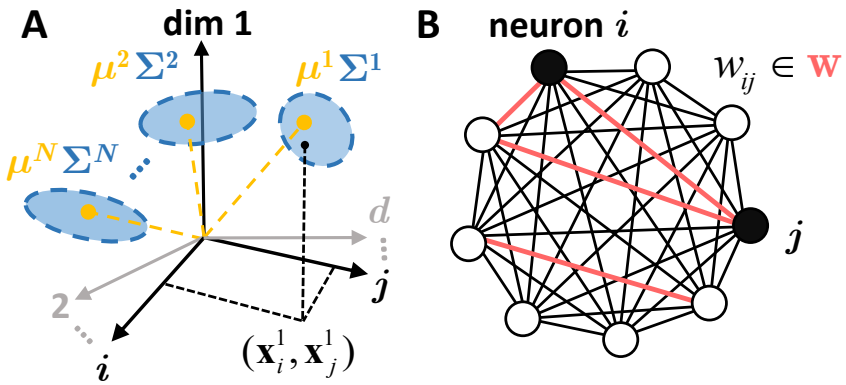

Figure 1: Model setup. (A) Input data distribution modelled as several independent patterns following multivariate log-normal distribution. (B) Densely connected Hebbian network with an example synaptic ensemble $\mathbf{w}$ (red) composed of four connections.

2.1 Information encoded by one synaptic connection

Here, we derive the approximation for the mutual information $MI(w_{ij}; \mathbf{x}^l)$ between a synaptic connection w_{ij} and a data pattern $\mathbf{x}^l$, by obtaining closed-form expressions for the two pertinent probability distributions: $p(w_{ij})$ and $p(w_{ij}|\mathbf{x}^l)$.

For a given pattern $\mathbf{x}^k$, the vector $\ln \mathbf{x}^k$ follows a multivariate Gaussian distribution. Consequently, its ith component also follows a Gaussian distribution, denoted as $\ln \mathbf{x}_i^k \sim \mathcal{N}(\mu_i^k, (\sigma_i^k)^2)$, where the mean μ_i^k is the ith component of μ^k, and the variance $(\sigma_i^k)^2$ is the ith diagonal element of the covariance matrix Σ^k. Similarly, the sum of the ith and j^{th} components follows a Gaussian distribution of the form:

$$\ln(\mathbf{x}_i^k) + \ln(\mathbf{x}_j^k) = \ln(\mathbf{x}_i^k \mathbf{x}_j^k) \sim \mathcal{N}(\mu_i^k + \mu_j^k, (\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k), \tag{3}$$

where σ_{ij}^k represents the (i, j) element of the covariance matrix Σ^k. Therefore, the random variable $\mathbf{x}_i^k \mathbf{x}_j^k$ also follows a log-normal distribution. The synaptic connection $w_{ij} = \sum_{k=1}^N \mathbf{x}_i^k \mathbf{x}_j^k$ is thus the sum of multiple independent log-normally distributed variables.

Fortunately, the problem of approximating the sum of log-normal variables has been studied extensively in engineering due to the prevalence of such noise distributions in communication systems. A widely used approach in the field, the Fenton-Wilkinson approximation (Fenton, 1960), demonstrates that the sum of log-normal variables can be reasonably well approximated by another log-normal variable using the moment-matching method. We apply this approach to derive the probability distribution $p(w_{ij})$; please see Methods (section 5.2) for proofs of all propositions.

Proposition 1.1. *For N independent log-normally distributed patterns $\{\mathbf{x}^1, ..., \mathbf{x}^N\}$ with $\ln \mathbf{x}_i^k \sim \mathcal{N}(\mu_i^k, (\sigma_i^k)^2)$, the probability distribution of the synaptic connection $w_{ij} := \sum_{k=1}^N \mathbf{x}_i^k \mathbf{x}_j^k$ in the Hebbian network can be approximated using a new log-normal distribution denoted by:*

$$\ln w_{ij} \overset{approx.}{\sim} \mathcal{N}(\mu_{w_{ij}}, (\sigma_{w_{ij}})^2) \tag{4}$$

i.e.,

$$p(w_{ij}) \approx \frac{1}{w_{ij}\sigma_{w_{ij}}\sqrt{2\pi}} \exp\left(-\frac{(\ln w_{ij} - \mu_{w_{ij}})^2}{2\sigma^2_{w_{ij}}}\right) \tag{5}$$

with the parameters $\mu_{w_{ij}}$ and $\sigma^2_{w_{ij}}$ being:

$$\mu_{w_{ij}} = \ln M_1 - \frac{1}{2}\ln(1 + \frac{M_2}{{M_1}^2}), \quad \sigma^2_{w_{ij}} = \ln(1 + \frac{M_2}{{M_1}^2}) \tag{6}$$

and with M_1 and M_2 defined as:

$$M_1 := \sum_{k=1}^{N} \exp\left(\mu_i^k + \mu_j^k + \frac{1}{2}\left[(\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right]\right) \tag{7}$$

$$M_2 := \sum_{k=1}^{N} \left[\exp\left((\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right) - 1\right] \exp\left(2(\mu_i^k + \mu_j^k) + (\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right) \tag{8}$$

Next, we derive the expression for the conditional distribution $p(w_{ij}|\mathbf{x}^l)$. Given the independence of the N terms defining a given synaptic weight, i.e., $\mathbf{x}_i^k\mathbf{x}_j^k \perp \mathbf{x}_i^l\mathbf{x}_j^l \ \forall k \neq l$, the distributions $p(w_{ij}|\mathbf{x}^l)$ and $p\left(\sum_{k\neq l}^{N} \mathbf{x}_i^k\mathbf{x}_j^k\right)$ are identical in the sense that the latter is a shifted version of the former by a constant $-\mathbf{x}_i^l\mathbf{x}_j^l$. For convenience, we denote $w_{ij/l} := \sum_{k\neq l}^{N} \mathbf{x}_i^k\mathbf{x}_j^k$. Consequently, we can derive the probability distribution for $w_{ij/l}$ instead, and do so using the same derivation steps we used in Proposition 1.1 for w_{ij}. We denote this distribution by:

$$\ln w_{ij/l} := \ln\left(\sum_{k\neq l}^{N} \mathbf{x}_i^k\mathbf{x}_j^k\right) \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_{w_{ij/l}}, (\sigma_{w_{ij/l}})^2\right), \tag{9}$$

where the parameters $\mu_{w_{ij/l}}$ and $(\sigma_{w_{ij/l}})^2$ have the same form as in Equation 6, with the lth term removed from the summation in Equations 7 and 8. For detailed expressions, see section 5.2.

With both $p(w_{ij})$ and $p(w_{ij}|\mathbf{x}^l)$ known, we can now obtain the expression for the mutual information.

Proposition 1.2. *The mutual information between a synaptic connection and a data pattern is approximated by:*

$$MI(w_{ij}; \mathbf{x}^l) = \left(\mu_{w_{ij}} - \mu_{w_{ij/l}}\right) + \left(\ln\sigma_{w_{ij}} - \ln\sigma_{w_{ij/l}}\right), \tag{10}$$

with the units in nats.

We note that the mutual information between a synaptic connection w_{ij} and a data pattern $\mathbf{x}^l$ (eqn. 10), depends on the specific connection weight w_{ij} (eqn. 5), the specific pattern $\mathbf{x}^l$, and the distribution properties of all data patterns, specifically their means and covariance matrices (eqns. 6, 7, 8).

2.2 Information encoded by a pair of connections

Since neural firing activities within a network can be correlated, the corresponding connection weights are likely to exhibit statistical dependencies rather than being independent of one another. To study the collective coding by these connections, we focus first on the mutual information between data and pairs of connections: $MI(\mathbf{w};\mathbf{x}^l)$, where $\mathbf{w} = (w_{ij}, w_{mn})^T$. This analysis requires knowledge of the distributions $p(\mathbf{w})$ and $p(\mathbf{w}|\mathbf{x}^l)$.

From the derivation in the single connection case, we obtain the marginals for each weight, which are given by:

$$\ln w_{ij} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_{w_{ij}}, (\sigma_{w_{ij}})^2\right), \quad \ln w_{mn} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_{w_{mn}}, (\sigma_{w_{mn}})^2\right), \tag{11}$$

This result indicates that the individual components of the vector $\ln \mathbf{w} = (\ln w_{ij}, \ln w_{mn})^T$ are Gaussian. While this does not necessarily imply that the joint distribution is multivariate Gaussian, we follow a common approximation for tractability by modeling the joint distribution as bivariate Gaussian here. Specifically, we express the joint distribution of $\ln \mathbf{w}$ as:

$$\ln \mathbf{w} = \begin{pmatrix} \ln w_{ij} \\ \ln w_{mn} \end{pmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}(\mu_{\mathbf{w}} = \begin{pmatrix} \mu_{w_{ij}} \\ \mu_{w_{mn}} \end{pmatrix}, \Sigma_{\mathbf{w}} = \begin{pmatrix} \sigma^2_{w_{ij}} & \sigma_{w_{ij,mn}} \\ \sigma_{w_{mn,ij}} & \sigma^2_{w_{mn}} \end{pmatrix}). \tag{12}$$

The only remaining unknown in this equation is $\sigma_{w_{ij,mn}}$, which represents the covariance between $\ln w_{ij}$ and $\ln w_{mn}$. The expression for this covariance is presented in the following proposition:

Proposition 2.1. *Under the approximation that a pair of synaptic connections* $(w_{ij}, w_{mn})^T := (\sum_k^N \mathbf{x}_i^k \mathbf{x}_j^k, \sum_k^N \mathbf{x}_m^k \mathbf{x}_n^k)^T$ *follows a 2-dimensional log-normal distribution, the covariance between* $\ln w_{ij}$ *and* $\ln w_{mn}$ *is:*

$$\sigma_{w_{ij,mn}} = \ln(1 + \frac{\sum_{k=1}^N \exp\left(\sum_\phi^{\{i,j,m,n\}} \mu_\phi^k + \frac{1}{2}\sum_\phi^{\{i,j,m,n\}}(\sigma_\phi^k)^2 + \sum_\phi^{\{ij,mn\}} \sigma_\phi^k\right)\left[\exp\left(\sum_\phi^{\{i,j\}\times\{m,n\}} \sigma_\phi^k\right) - 1\right]}{\sum_{k,l=1}^N \exp\left(\sum_\phi^{\{i,j\}} \mu_\phi^k + \sum_\phi^{\{m,n\}} \mu_\phi^l + \frac{1}{2}\sum_\phi^{\{i,j\}}(\sigma_\phi^k)^2 + \frac{1}{2}\sum_\phi^{\{m,n\}}(\sigma_\phi^l)^2 + \sigma_{ij}^k + \sigma_{mn}^l\right)}) \tag{13}$$

with the summation notation meaning:

$$\sum_\phi^{\{ij,mn\}} \sigma_\phi^k = \sigma_{ij}^k + \sigma_{mn}^k, \quad \sum_\phi^{\{i,j\}\times\{m,n\}} \sigma_\phi^k = \sigma_{im}^k + \sigma_{in}^k + \sigma_{jm}^k + \sigma_{jn}^k \tag{14}$$

With $p(\mathbf{w})$ thus specified, we next turn to $p(\mathbf{w}|\mathbf{x}^l)$. As was the case with single connections, here, $p(\mathbf{w}|\mathbf{x}^l)$ and $p\left(\sum_{k\neq l}^N \mathbf{x}_i^k \mathbf{x}_j^k, \sum_{k\neq l}^N \mathbf{x}_m^k \mathbf{x}_n^k\right)$ are identical distributions up to a location shift. For convenience, we denote $\mathbf{w}_{/l} := (\sum_{k\neq l}^N \mathbf{x}_i^k \mathbf{x}_j^k, \sum_{k\neq l}^N \mathbf{x}_m^k \mathbf{x}_n^k)^T$. Applying the same steps we

used to derive $p(\mathbf{w})$ above, we get:

$$\ln \mathbf{w}_{/l} := \begin{pmatrix} \ln w_{ij/l} \\ \ln w_{mn/l} \end{pmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}(\mu_{\mathbf{w}_{/l}} = \begin{pmatrix} \mu_{w_{ij/l}} \\ \mu_{w_{mn/l}} \end{pmatrix}, \Sigma_{\mathbf{w}_{/l}} = \begin{pmatrix} \sigma^2_{w_{ij/l}} & \sigma_{w_{ij,mn/l}} \\ \sigma_{w_{mn,ij/l}} & \sigma^2_{w_{mn/l}} \end{pmatrix}), \tag{15}$$

where the parameters with $/l$ are equal to their original corresponding parameters, with the terms related to $\mathbf{x}^l$ removed from the summation. For a more detailed expression for $\mu_{\mathbf{w}_{/l}}$ and $\Sigma_{\mathbf{w}_{/l}}$, please refer to the Methods section 5.2.

Given $p(\mathbf{w})$ and $p(\mathbf{w}|\mathbf{x}^l)$, we can then calculate the mutual information encoded by any pair of synaptic connections about a data pattern, as stated in the following proposition:

Proposition 2.2. *The mutual information between a pair of connections $\mathbf{w} = (w_{ij}, w_{mn})^T$ and a data pattern $\mathbf{x}^l$ is given by the analytical approximation:*

$$MI(\mathbf{w}; \mathbf{x}^l) = (\mu_{w_{ij}} + \mu_{w_{mn}} - \mu_{w_{ij/l}} - \mu_{w_{mn/l}}) + \frac{1}{2} \ln |\Sigma_{\mathbf{w}}| - \frac{1}{2} \ln |\Sigma_{\mathbf{w}_{/l}}|, \tag{16}$$

where $|\Sigma_{\mathbf{w}}|$ denotes the determinant of the matrix $\Sigma_{\mathbf{w}}$. The information is expressed in nats.

2.3 Information in an ensemble of multiple synaptic connections

Next, we extend the analysis to information encoded jointly in an ensemble of n synaptic connections, denoted as $\mathbf{w} = (w_{ij(1)}, w_{ij(2)}, \ldots, w_{ij(n)})^T$. Here, the set $\{ij(1), ij(2), \ldots, ij(n)\}$ represents the indices for the n synaptic weights. For example, $w_{ij(k)}$ refers to the connection between neuron $i(k)$ and neuron $j(k)$. The mutual information encoded by any n connections about the data, $MI(\mathbf{w}; \mathbf{x}^l)$, can be derived using a similar approach as before. Since all marginals for $\ln \mathbf{w}$ are normally distributed, we approximate the joint distribution as an n-dimensional log-normal distribution, leading to:

$$\ln \mathbf{w} = \begin{pmatrix} \ln w_{ij(1)} \\ \vdots \\ \ln w_{ij(n)} \end{pmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}(\mu_{\mathbf{w}} = \begin{pmatrix} \mu_{w_{ij(1)}} \\ \vdots \\ \mu_{w_{ij(n)}} \end{pmatrix}, \Sigma_{\mathbf{w}} = \begin{pmatrix} \sigma^2_{w_{ij(1)}} & \sigma_{w_{ij(1),ij(2)}} & \cdots & \sigma_{w_{ij(1),ij(n)}} \\ \sigma_{w_{ij(2),ij(1)}} & \sigma^2_{w_{ij(2)}} & \cdots & \sigma_{w_{ij(2),ij(n)}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{w_{ij(n),ij(1)}} & \sigma_{w_{ij(n),ij(2)}} & \cdots & \sigma^2_{w_{ij(n)}} \end{pmatrix}) \tag{17}$$

This expression represents $p(\mathbf{w})$ as an n-dimensional log-normal distribution. The parameters $\mu_{\mathbf{w}}$ and the diagonal elements of $\Sigma_{\mathbf{w}}$ follow the form given in Equation 6, while the off-diagonal elements of $\Sigma_{\mathbf{w}}$ follow Equation 13, specifically:

$$\sigma_{w_{ij(r),ij(s)}} = \ln(1 + \frac{\sum_{k=1}^{N} \exp\left(\sum_m^{\{r,s\}} \left[\sum_\phi^{\{i(m),j(m)\}} \mu_\phi^k + \frac{1}{2}(\sigma_\phi^k)^2\right] + \sigma^k_{ij(m)}\right) \left[\exp\left(\sum_\phi^{\{i(r),j(r)\}\times\{i(s),j(s)\}} \sigma_\phi^k\right) - 1\right]}{\sum_{k,l=1}^{N} \exp\left(\left[\sum_\phi^{\{i(r),j(r)\}} \mu_\phi^k + \frac{1}{2}(\sigma_\phi^k)^2\right] + \left[\sum_\phi^{\{i(s),j(s)\}} \mu_\phi^l + \frac{1}{2}(\sigma_\phi^l)^2\right] + \sigma^k_{ij(r)} + \sigma^l_{ij(s)}\right)}) \tag{18}$$

To obtain $p(\mathbf{w}|\mathbf{x}^l)$, as in the single and paired connection scenarios, we focus on its shifted

distribution $p(\mathbf{w}_{/l})$ with $\mathbf{w} = (w_{ij(1)/l}, w_{ij(2)/l}, \ldots, w_{ij(n)/l})^T$, which is also treated as log-normally distributed. The parameters with $/l$ are simply the original parameters with the summation terms related to $\mathbf{x}^l$ excluded:

$$\ln \mathbf{w}_{/l} = \begin{pmatrix} \ln w_{ij(1)/l} \\ \vdots \\ \ln w_{ij(n)/l} \end{pmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}(\mu_{\mathbf{w}_{/l}} = \begin{pmatrix} \mu_{w_{ij(1)/l}} \\ \vdots \\ \mu_{w_{ij(n)/l}} \end{pmatrix}, \Sigma_{\mathbf{w}_{/l}} = \begin{pmatrix} \sigma^2_{w_{ij(1)/l}} & \ldots & \sigma_{w_{ij(1),ij(n)/l}} \\ \vdots & \ddots & \vdots \\ \sigma_{w_{ij(n),ij(1)/l}} & \ldots & \sigma^2_{w_{ij(n)/l}} \end{pmatrix}) \tag{19}$$

Please refer to section 5.2 for detailed expressions for $\mu_{\mathbf{w}_{/l}}$ and $\Sigma_{\mathbf{w}_{/l}}$.

Finally, with $p(\mathbf{w})$ and $p(\mathbf{w}|\mathbf{x}^l)$, we can derive the expression for the information jointly encoded in multiple synaptic connections about a pattern:

Proposition 3. *The mutual information between a collect of n synaptic connections $\mathbf{w} = (w_{ij(1)}, w_{ij(2)}, ..., w_{ij(n)})^T$ and a data pattern $\mathbf{x}^l$ is given by the analytical approximation:*

$$MI(\mathbf{w}; \mathbf{x}^l) = \sum_{k=1}^{n} \left(\mu_{w_{ij(k)}} - \mu_{w_{ij(k)/l}} \right) + \frac{1}{2} \ln |\Sigma_{\mathbf{w}}| - \frac{1}{2} \ln |\Sigma_{\mathbf{w}_{/l}}| \tag{20}$$

Together, these derivations provide an explicit expression for the information encoded in a specified synaptic ensemble about a particular data pattern, $MI(\mathbf{w}; \mathbf{x}^l)$. Notably, it is fully determined by the parameters of the distributions (means and covariances) of the relevant components of all N data patterns. Formally, this dependence can be summarized by the set $\mathcal{S} = \{\mu_i^k, \sigma_{ij}^k \mid k \in \{1, \ldots, N\}, \forall (i,j)$ such that w_{ij} is a component of $\mathbf{w}\}$. Consequently, both the size of the ensemble, n, and the particular choice of connections included in the ensemble, influence MI, as they determine the elements included in the set $\mathcal{S}$.

Importantly, $MI(\mathbf{w}; \mathbf{x}^l)$ depends only on the parameters associated with the neurons participating in the ensemble; neurons in the network that lie outside the ensemble have no effect. This finding implies that as long as the local statistics, $\mathcal{S}$, of the specified neurons remain fixed, the value of $MI(\mathbf{w}; \mathbf{x}^l)$ will remain unchanged, regardless of the larger Hebbian connectivity in which the ensemble is embedded. In this sense, our result is context-invariant: it depends solely on the properties of the synaptic ensemble in question and is independent of the total size, d, of the network containing the ensemble.

3 Results

After obtaining the analytical expressions for the information encoded in synaptic connections, we performed a series of exploratory analyses by varying the parameter set $\mathcal{S}$ to examine its effect on MI.

We conducted three sets of simulations. The first set varied parameter values in constrained cases to reveal the dependency of MI on continuously changing parameters. The second and third sets relaxed these constraints progressively and sampled parameter combinations randomly; regression analysis was used to reveal the dependency of MI on a range of parameter combinations. Below, we describe the key differences in the constraints used for each simulation set to facilitate understanding of the subsequent results. For further details regarding the

simulation setups, please see the Methods section.

First Simulation Set: The covariance matrix Σ^k was held constant across different data patterns, and was constrained to be exchangeable: all diagonal elements were set to $\tilde{\sigma}^2$, and off-diagonal elements to $\tilde{\rho}\tilde{\sigma}^2$. Results of this analysis are shown in Figures 2A, 2B, 3A, and 3B.

Second Simulation Set: Again, Σ^k was held constant across data patterns. However, to explore more general covariance structures, a stochastic component was added to the exchangeable component (making them no longer exchangeable); it was set to be proportional to $\mathbf{A}^T\mathbf{A}$, where $\mathbf{A}$ is a stochastic matrix. Results are shown in Figures 2C and 3C.

Third Simulation Set: Here, the covariance matrix Σ^k varied between patterns. To avoid biases in matrix generation, each Σ^k was constructed by generating random eigenvalues and random orthogonal matrices. Results of the corresponding analyses are shown in Figures 2D, 3D, and 4.

We note that we use nats as the unit of mutual information in both the theoretical section and simulations for consistency.

3.1 Relationship between the information encoded by synapses and the number of data patterns

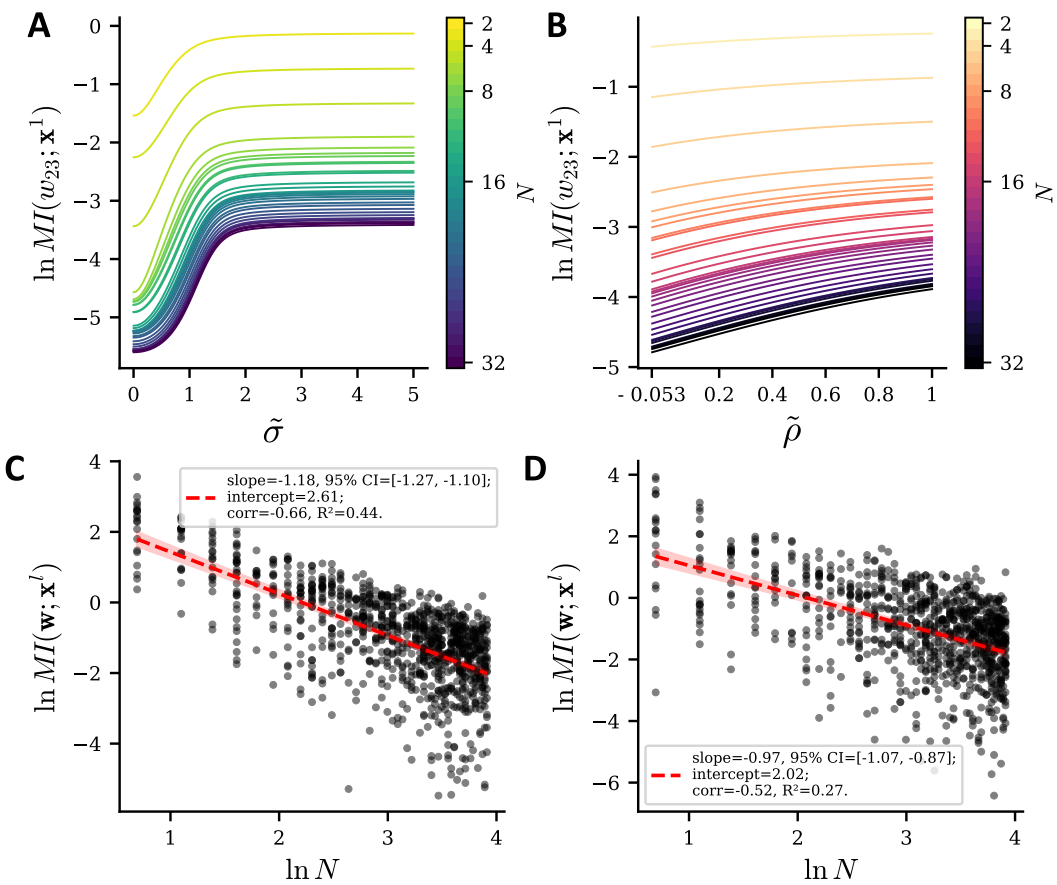

Figure 2: Encoded information as a function of the number of data patterns N. (A,B) Results from the first simulation set, and for a *single* (randomly chosen) synaptic connection. (A) Information encoded about a specific example pattern ($\mathbf{x}^1$) by a single example weight (w_{23}), plotted as a function of $\tilde{\sigma}$ (x-axis) and the number of data patterns N (different colors). $\tilde{\rho}$ is fixed at 0, indicating that the dimensions of the data patterns are independent. (B) Information encoded about $\mathbf{x}^1$ by w_{23}, plotted as a function of $\tilde{\rho}$ (x-axis) and N (colors); $\tilde{\sigma}$ is fixed at 1. In both (A) and (B), the information encoded decreases as N increases, indicated by downward shifts in the curves for higher N. (C, D) Results from the second and third simulation sets, respectively, and for synaptic ensembles $\mathbf{w}$. The second simulation set corresponds to data patterns with the same, but non-exchangeable covariance matrix, and the third simulation set to data patterns with different, random covariance matrices. Each black dot corresponds to a random choice of configuration, including the number (n) and identity of synaptic connections in the ensemble, the distribution of data patterns, and the specific data pattern $\mathbf{x}^l$. n ranges from 1 to 20. The red dashed line indicates the fitted linear regression, with the red shaded area representing the 95% confidence interval for the predicted values. The decreasing trend of MI with increasing N also persists for synaptic *ensembles* $\mathbf{w}$ with randomly sampled configurations.

We investigated the relationship between the information encoded by synapses and the number of patterns N in the dataset.

First, to develop intuition for this relationship, we began by examining how mutual information changes with N: (i) for *single* synaptic connections, here, randomly chosen to be w_{23}, and (ii) under the simplest data covariance conditions, here, using the first simulation set. We plotted $MI(w_{23}; \mathbf{x}^1)$ as a function of $\tilde{\sigma}$ (Fig. 2A), and as a function of $\tilde{\rho}$ (Fig. 2B), and in each case, obtained curves for varying numbers of data patterns (N, ranging from 2 to 32).

We found that the information encoded in a single connection increases with both $\tilde{\sigma}$ and $\tilde{\rho}$ (Fig. 2AB), appearing to reach an asymptotic limit for large $\tilde{\sigma}$. These findings are consistent with the explanation that because $\tilde{\sigma}$ represents the spread or width of the distribution, a wider distribution, typically corresponding to higher entropy in the data patterns, provides more content for the network to encode. Meanwhile, $\tilde{\rho}$ reflects the internal structure of the data: data with more structure (lower $\tilde{\rho}$) tends to be easier (lower intrinsic dimensionality) for the network to store, assuming the same level of entropy.

Notably, we found that as the number of data patterns increases from 2 to 32, the information encoded by a single connection $MI(w_{23}; \mathbf{x}^1)$ decreases independently of the specific $\tilde{\sigma}$ and $\tilde{\rho}$ values (Fig. 2AB; downward shift of curves from light green/light red to dark purple; log scale). This indicates that the information stored in each individual connection about a single data pattern decreases as the total number of patterns to be stored in the neural network increases, rather than remaining constant.

Next, with this intuition in hand, we relaxed the above constraints to consider the more general scenario of ensemble coding. Specifically, we examined how MI changes with N: (i) for synaptic *ensembles* of random sizes, here, between $n = 1$ to $n = 20$, and (ii) under more general covariance structures, i.e., using the second and third simulation sets in which the covariance matrices are not restricted to be exchangeable.

In Fig. 2C, where the (non-exchangeable) covariance matrices are identical across patterns, we observed a significant negative correlation between $MI(\mathbf{w}; \mathbf{x}^l)$ and N (slope=-1.18, 95%

CI=[-1.27, -1.10]; $R^2 = 0.44$). Similarly, in Fig. 2D, where the covariance matrices differed among patterns, we observed a significant negative correlation (slope=-0.97, 95% CI=[-1.07, -0.87]; $R^2 = 0.27$). Furthermore, by examining the slope of the linear relationship between $\ln MI(\mathbf{w}; \mathbf{x}^l)$ and $\ln N$, we find that $MI(\mathbf{w}; \mathbf{x}^l)$ follows an approximate $N^{-1.2}$ scaling in Fig. 2C and $N^{-1.0}$ in Fig. 2D (with accuracy kept to the first decimal place). Thus, our results revealed that the information encoded in synaptic ensembles, in general, decreases with the number of data patterns.

3.2 Relation between information encoded by synapses and the number of synaptic connections

Next, we investigated how mutual information changes in relation to the number of synaptic connections, rather than the number of data patterns.

Results from the first simulation set provided intuition for how the information encoded about a data pattern by a synaptic ensemble, i.e., $MI(\mathbf{w}; \mathbf{x}^1)$, changes as a function of $\tilde{\sigma}$ (Fig. 3A), and as a function of $\tilde{\rho}$ (Fig. 3B). Curves for varying numbers of synaptic connections (n, ranging from 1 to 16) are overlaid for comparison.

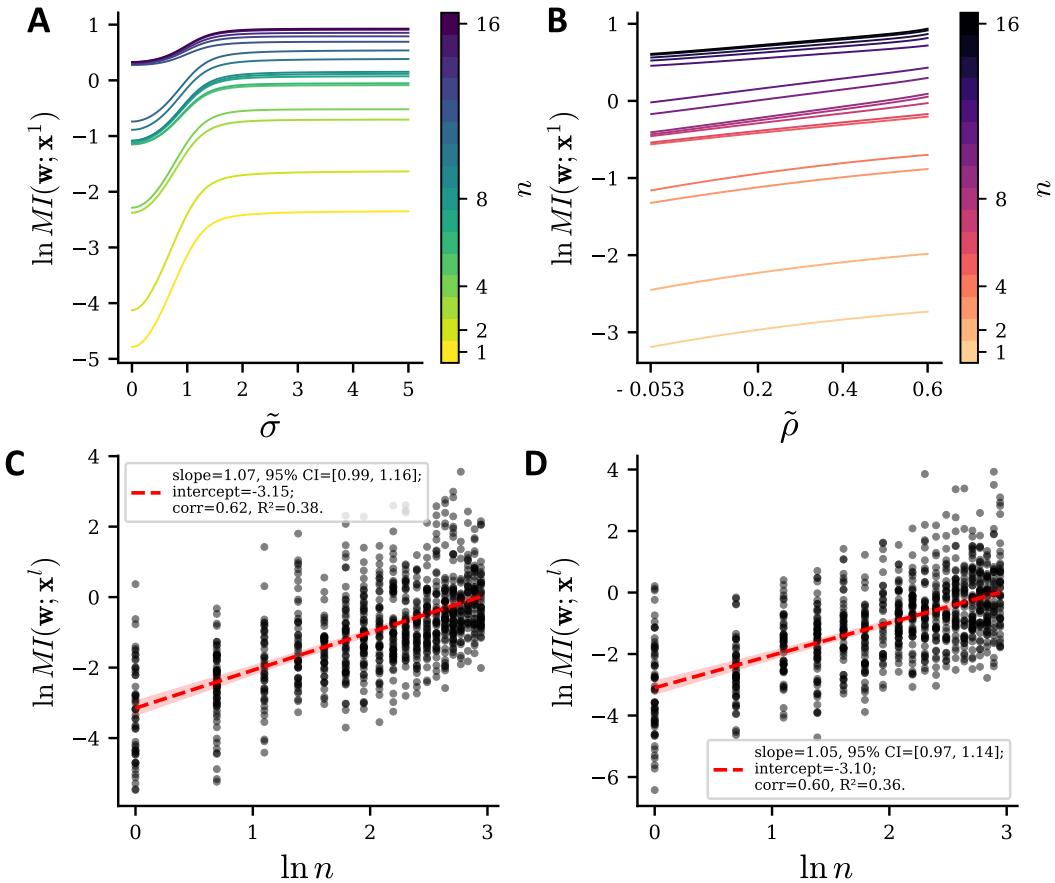

Figure 3: Encoded information as a function of the number of synaptic connections in an ensemble. (A, B) Results from the first simulation set. A randomly chosen ensemble ($\mathbf{w}$) and a randomly selected example pattern ($\mathbf{x}^1$) are used to illustrate how the mutual information depends on $\tilde{\sigma}$, $\tilde{\rho}$, and the number of synaptic connections n. (A) Information encoded about $\mathbf{x}^1$ by the example ensemble, plotted as a function of $\tilde{\sigma}$ (x-axis) for varying n (colors). Here, $\tilde{\rho} = 0$ and $N = 10$. (B) Information encoded about $\mathbf{x}^1$ by the example ensemble, plotted as a function of $\tilde{\rho}$ (x-axis) for different values of n (colors). Here, $\tilde{\sigma} = 1$ and $N = 10$. In both (A) and (B), as n increases, the curves shift upward, indicating that larger ensembles encode more information. (C, D) Results from the second and third simulation sets, respectively. Other conventions as in Figure 2CD. The increasing trend of MI as a function of n persists when both $\mathbf{w}$ and all parameter configurations are sampled randomly.

Similar to the simple case of a single synaptic connection (Fig. 2AB), the increasing relationship of $MI(\mathbf{w}; \mathbf{x}^1)$ with $\tilde{\sigma}$ and $\tilde{\rho}$ holds true for an ensemble of synaptic connections as well (Fig. 3AB).

Notably, we found that as the number of synaptic connections increases from 1 to 16, the information in the ensemble $MI(\mathbf{w}; \mathbf{x}^1)$ also increases (Fig. 3AB; upward shift of curves from light green/light red to dark purple; log scale). This result is consistent with the notion that adding more variables should allow for more of the information about the patterns to be encoded.

This increasing relationship was evident also in the second (Fig. 3C) and third (Fig. 3D) simulation sets, involving progressively more general covariance structures. In Fig. 3C, where the covariance matrices are non-exchangeable but identical across patterns, we observed a significant positive correlation between $MI(\mathbf{w}; \mathbf{x}^l)$ and n (slope=1.07, 95% CI=[0.99, 1.16]; $R^2 = 0.38$). Similarly, in Fig. 3D, where the covariance matrices differed among patterns, we observed a significant positive correlation between $MI(\mathbf{w}; \mathbf{x}^l)$ and n (slope=1.05, 95% CI=[0.97, 1.14]; $R^2 = 0.36$).

Interestingly, these results allow us to estimate the relationship between the number of data patterns, P, stored in the entire network and the size of the network (or number of neurons), d. First, linear regression between MI and the number n of synaptic connections in the ensemble reveal a power-law relationship of $MI(\mathbf{w}; \mathbf{x}^l) \sim n^{1.1}$ (with accuracy kept to the first decimal place), which is nearly identical across simulation sets (Figs. 3CD). In other words, the information capacity of synaptic ensembles scales approximately with the 1.1 power of their size. Second, since our modeled network is an undirected graph without self-connections, the size of the full synaptic ensemble and the size of the network are related quadratically, i.e., $n = d(d-1)/2 \sim d^2$. Therefore, the information capacity of synaptic ensembles scales as $MI \sim d^{2.2}$. Third, in our autoassociation setup, all neurons function both as inputs and outputs for memorization, with no hidden neurons involved. Thus, the network size equals the dimensionality of each data pattern. If we define successful memorization as the ability to accurately reconstruct each component of a pattern within the same resolution, then each component requires an equal amount of information for storage. Accordingly, the information required to store one pattern, H_{pattern}, is proportional to the number of components, i.e., $H_{\text{pattern}} \sim d$. Putting these together, we obtain the number of patterns that can be stored in the network by dividing the total information stored in the network by the information needed to store a single

pattern:

$$P \approx \frac{MI}{H_{\text{pattern}}} \sim d^{1.2},$$

This yields that the number of patterns that can be stored in the connectivity of a dense Hebbian network is proportional to the 1.2 power of the network size.

3.3 Information contributed per connection increases with the synaptic ensemble size

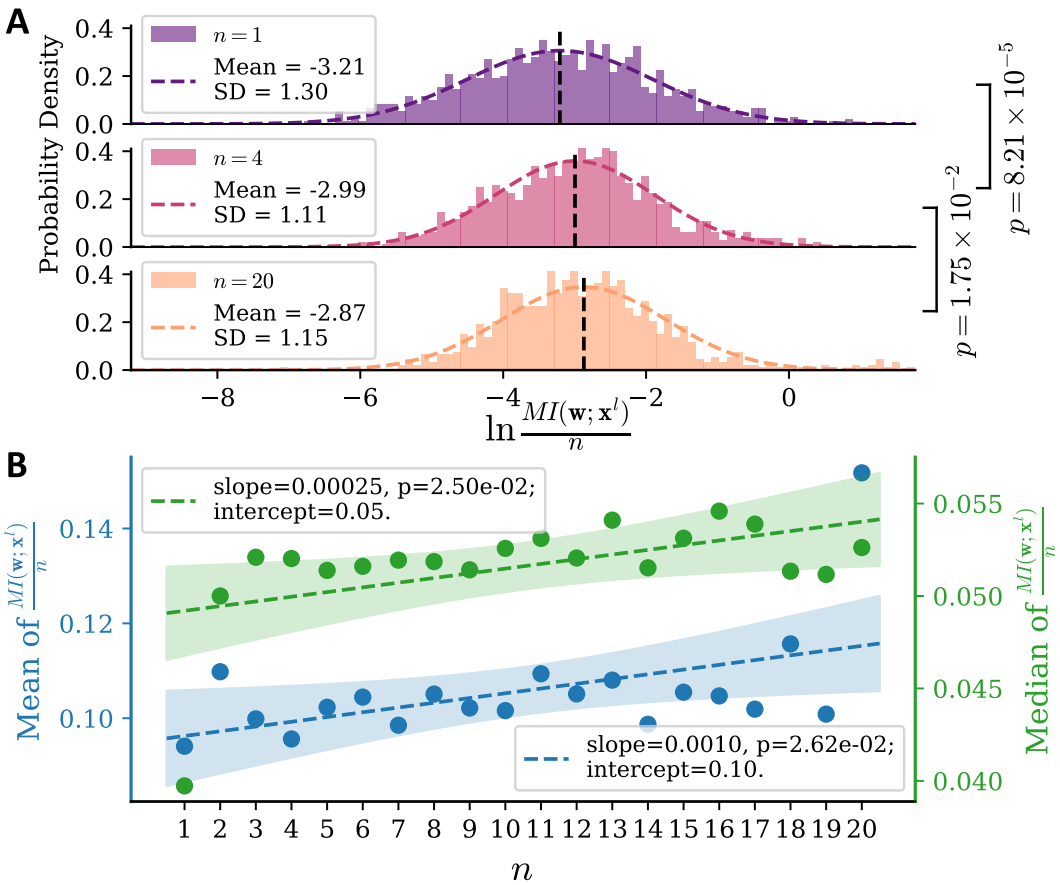

Figure 4: Information contributed per synaptic connection as a function of ensemble size. (A) Histograms showing the distribution of $\ln(MI/n)$ for $n = 1, 4, 20$. (B) Relationship between the mean (blue) and median (green) of MI/n for different values of n. Dashed lines indicate the linear regression results for these mean and median values, while shaded areas indicate the 95% confidence intervals for the predicted values.

In the previous sections, we analyzed the information encoded by a single synaptic connection, $MI(w;\mathbf{x}^l)$, and by an ensemble of synaptic connections, $MI(\mathbf{w},\mathbf{x}^l)$. However, an intriguing question that arises is: how does the contribution of a single connection to the in-

formation encoded by the entire synaptic ensemble change with the size of the ensemble? To address this question, we approximated the contribution of a each connection to the information encoded by the ensemble as $MI(\mathbf{w}, \mathbf{x}^l)/n$, where n is the size of the ensemble. This analysis sheds light on how individual connections behave when functioning as part of a collective.

Using data from simulation set three, which reflects the most general configuration, we computed the relationship between $\ln(MI(\mathbf{w}, \mathbf{x}^l)/n)$ and n. We first plotted the histograms of $\ln(MI(\mathbf{w}, \mathbf{x}^l)/n)$ for $n = 1$, 4, and 20, and found that they followed Gaussian-like distributions (Fig. 4A). Interestingly, this Gaussian-like shape suggests that the information contribution per synapse may also follow a log-normal distribution. Furthermore, these distributions shifted toward higher values as n increases; the shifts were significant between $n = 1$ and $n = 4$ ($p = 8.21 \times 10^{-5}$, two-sided t-test), and between $n = 4$ and $n = 20$ ($p = 1.75 \times 10^{-2}$, two-sided t-test) indicating notable changes in the distribution.

To explore this relationship further, we examined the values of $MI(\mathbf{w}, \mathbf{x}^l)/n$ as a function of n and found significant *positive* correlation (slope=9.6×10^{-4}, 95% CI=[$4.3\times10^{-4}, 1.5\times10^{-3}$]). Furthermore, we also examined the median and median values of $MI(\mathbf{w}, \mathbf{x}^l)/n$. There is a significant upward trend for both the mean (Fig. 4B; blue line, $p = 2.62 \times 10^{-2}$) and the median (Fig. 4B; green line, $p = 2.50 \times 10^{-2}$) as n increases. These findings reveal that the information contributed per synaptic connection increases with ensemble size.

(B) Data points showing the relationship between MI/n and n. The red dashed line represents the linear regression fit, and the red shaded area indicates the 95% confidence interval for the predicted values. Note that MI/n is shown on a logarithmic scale.

4 Discussion

In this study, we adopted an under-explored perspective on neural coding, focusing on the information encoded in synaptic weights rather than in neuronal firing activities. Our objective was to derive an analytical expression for the information embedded in ensembles of synaptic connections. As a first step in addressing this gap, here, we used a network with Hebbian connectivity to derive a closed-form expression for the mutual information between log-normally distributed data patterns and ensembles of synaptic weights. Although seemingly complex, the resulting formulation represents a straightforward analytical approximation involving just the statistical parameters of the distributions of the data patterns. Using this closed-form expression, we conducted three sets of simulations that revealed interesting insights into distributed coding, network capacity, and information synergy, which are discussed in detail in the following sections. Overall, we propose that our work provides a framework for analyzing neural information through the lens of synaptic connections, offering a complementary perspective to the traditional focus on firing activities.

4.1 Distributed coding in synaptic weights

We demonstrated, here, that the mutual information about a data pattern, whether stored in a single synaptic weight, $MI(w_{ij}; \mathbf{x}^1)$, or an ensemble of synaptic weights, $MI(\mathbf{w}; \mathbf{x}^l)$, decreases as the number of stored patterns, N, increases. This suggests that introducing new patterns

for storage reduces the capacity of synaptic connections to retain information about previously stored patterns. In other words, each synaptic connection allocates part of its storage capacity to every new pattern, resulting in a scenario in which individual connections simultaneously encode partial information about all stored patterns. In contrast, one could imagine an alternative scheme, analogous to the "grandmother cell" hypothesis in neural activity coding, in which information storage relies on "grandmother connections": certain synapses specialize in encoding specific patterns, while others specialize in different patterns. In this scenario, introducing new patterns would only use capacity in previously unused connections, leaving the information stored in other connections unaffected. Our findings, however, align better with the well-known paradigm of distributed coding in the brain, in which diverse types of information—such as memory (Rissman and Wagner, 2012), sensory inputs (Stecker and Middlebrooks, 2003), or motor actions (Steinmetz et al., 2019)—are encoded collectively across groups of neurons, and further extend this principle to the perspective of synaptic connections within neural networks.

4.2 Capacity analysis for networks with Hebbian connectivity

Our work quantifies the capacity of Hebbian networks based on the information stored within the synaptic ensemble. Since Hebbian connectivity is fundamental to the Hopfield network, we use the latter as a reference for comparison. The capacity of the Hopfield network has been extensively explored, with well-known results such as $P \sim 0.14d$ (Amit et al., 1985) or $P \leq d$ (Abu-Mostafa and Jacques, 1985). Our analysis shows that a distinct storage law applies to Hebbian networks considered broadly. By assuming that each component or "digit" in a pattern requires the same amount of information, we found that the number of stored patterns (P) follows a marginally superlinear relationship with the network size (d), specifically $P \sim d^{1.2}$.

This difference can be attributed to the distinct perspectives adopted for deriving capacity. Traditionally, capacity is evaluated by calculating the number of patterns retrievable from the dynamics of neuronal firing activity. In contrast, our method focuses on the patterns directly stored within the distribution of synaptic weights.

Given that the information flow in the learning paradigm can be conceptualized as a Markov chain, where data are mapped to connectivity before being retrieved by dynamics, the data processing inequality (Cover, 1999) suggests that the mutual information between the data and connections must exceed that of the retrieval dynamics. Thus, our superlinear capacity result reflects the information loss inherent in the retrieval phase and serves as an upper bound for the practical capacity.

This echoes the perspective proposed in (Palm and Sommer, 1996), where associative memory is viewed as a communication channel consisting of two components: a storage stage, mapping data patterns to the synaptic connectivity matrix, and a retrieval stage, mapping the connectivity matrix to output patterns. These two components correspond to "storage capacity" and "retrieval capacity", respectively, each providing an upper bound on the total memory capacity when the full storage–retrieval procedure is considered. However, even with this perspective, related theoretical works (Palm, 1980, 1992) still define "storage capacity" operationally as the amount of information about the patterns that can be successfully retrieved from the synaptic weights. This is achieved by applying the stored patterns to the learned connectivity matrix, performing thresholding detection, and quantifying the information content based on the re-

trieval error. Consequently, the resulting capacity is still derived from the perspective of neural activity, rather than the information content of the connectivity itself.

Separately, our approach provides a more general perspective on network coding. Traditional studies have focused primarily on the storage capacity of the *full* $d \times d$ Hopfield network, which limits their ability to examine ensembles consisting of only a subset of neurons or connections. By contrast, our method can be applied to *any* selected ensemble of connections within the network, and reveals that the relationship between encoded information and synaptic ensemble size (n) follows $MI \sim n^{1.1}$. This generalization, obtained by adopting a complementary perspective to the traditional one, enables more detailed investigations of neural coding within different network substructures, such as by specific circuit motifs.

4.3 Synergy among synaptic connections

Our findings demonstrated that the information stored collectively by the connections in a network can exceed the sum of the information stored by individual synaptic connections; that is, $\mathbb{E}[MI(\mathbf{w};\mathbf{x}^l)] > n\,\mathbb{E}[MI(w;\mathbf{x}^l)]$. This result followed directly from the positive correlation observed between $\mathbb{E}[MI(\mathbf{w};\mathbf{x}^l)]/n$ and n (Fig. 4B), specifically: $\mathbb{E}[MI(\mathbf{w};\mathbf{x}^l)]/n > \mathbb{E}[MI(w;\mathbf{x}^l)]/1$ when $n > 1$.

This phenomenon, where the whole is greater than the sum of its parts, is known as *synergy* (Gawne and Richmond, 1993; Brenner et al., 2000; Williams and Beer, 2010). Information synergy has been a topic of long-standing interest in neuroscience, with evidence of its existence reported in ensembles of neurons (Panzeri et al., 1999; Schneidman et al., 2003) and groups of brain regions (Luppi et al., 2022). Our approach provides novel evidence of synergy at the level of synaptic connections.

Synergy in neural coding arises due to interactions among components of the system. These interactions can occur at various levels: pairwise, triplet, quadruplet, or even higher-order interactions. Classical studies of synergistic coding in neurons often focus on the role of correlations between spikes, which represent second-order interactions. More recent work has explored third-order interactions, introducing concepts such as triplet correlations (Cayco-Gajic et al., 2015). By deriving closed-form expressions for Shannon mutual information, our work offers a straightforward way to analyze interactions at all levels—2nd order and higher—by calculating $MI(\mathbf{w};\mathbf{x}^l)/n$, the contribution of a single connection to the encoding by the ensemble. In our simulations, we estimated interactions ranging from second to twentieth order using this value and observed a significant increasing trend as the number of connections n grew (Fig. 4BC). This suggests that synergistic effects extend at least to the 20th order, indicating that further insights may await discovery at higher orders. For instance, it remains unclear whether this synergistic scenario will continue indefinitely as interaction orders increase. Some studies suggest that synergistic effects dominate only at smaller neural populations, typically involving fewer than 200 neurons, while coding becomes increasingly redundant at larger scales (Kafashan et al., 2021); our analysis did not reach the scale of $n > 40,000$ synaptic connections due to computational limitations. Nonetheless, the linear regression results in Figure 4C suggest a relationship of $MI \sim 0.001n^2$, which contrasts sharply with established results for Hopfield networks, providing further credence to our expectation that the linear trend seen in Figure 4C will eventually plateau or saturate as n becomes large.

These observations leads to an intriguing question: what is the optimal size of an ensemble of neurons or synaptic connections to maximize synergy? Identifying this "most synergistic" size may point to a fundamental unit of neural coding and represents a promising direction for future research.

4.4 Potential implications for modern machine learning

Here, we highlight several potential implications of our framework for the machine learning community.

First, we note that two adjacent layers in a multilayered deep neural network can be viewed as a dense bipartite network. Furthermore, several studies have explored progressively greater equivalence between the backpropagation algorithm and Hebbian-like local learning rules (Xie and Seung, 2003; Harris, 2008; Lillicrap et al., 2016; Akrout et al., 2019; Fan and Mysore, 2025). Under this interpretation, the data pattern $\mathbf{x}$ in our framework corresponds to the concatenation of the firing activities of two adjacent layers. The mutual information between the synaptic weights and the data patterns, $MI(\mathbf{w};\mathbf{x})$, thus measures the information encoded in the weights about the joint activity of the layers. This can be decomposed into two components: information about the preceding layer, and information that facilitates the prediction of the subsequent layer given the former. From this perspective, our framework provides a way to quantify how much information synaptic weights store about cross-layer co-activity. We speculate that the phenomena revealed by our framework, such as distributed coding and synergistic information, may also emerge between adjacent layers in modern deep neural networks.

Second, our work implies that pruning individual synapses based solely on their isolated information content may be misleading. Due to synergistic interactions, the contribution of a single synapse is inherently context-dependent and determined by the ensemble in which it is embedded. Pruning strategies that rely only on the mutual information between an individual synapse and the data may therefore underestimate its functional contribution.

Finally, our framework enables comparisons of information storage across different network motifs that contain the same number of synapses but differ in connectivity patterns. This opens the door to future studies addressing which structural motifs play the most prominent roles in information storage. Understanding these relationships is a key step toward making "black box" neural networks more interpretable and could potentially yield new neuroscientific insights.

4.5 Assumptions and considerations

In our exploration of the information encoded in synaptic ensembles, we chose to employ the densely connected neural network with Hebbian connectivity owing to its mathematical tractability. More commonly used architectures, such as multi-layer fully connected networks or recurrent neural networks, were not covered here. Also not considered were the additional complexities in biological neural networks, such as sparsity and spiking-timing-dependent-plasticity, beyond those captured by the Hebbian network model. Despite these constraints, Hebbian connectivity serves as a basis for various foundational models, such as Hopfield network, in associative memory and neural information storage. Indeed, it has been successfully

applied to various brain regions, including the hippocampus (Sun et al., 2023; Kang and Toyoizumi, 2024), prefrontal cortex (Durstewitz et al., 2000), primary visual cortex (Dong and Hopfield, 1992), and olfactory bulb (Hopfield, 1991), suggesting that insights from this model may apply more generally. Consequently, our framework represents an important step forward in characterizing information storage in synaptic connections.

We also chose to model data patterns as being log-normally distributed. This assumption may appear artificial or restrictive at first glance. However, if the data patterns $\mathbf{x}$ in our framework were to be interpreted not as raw external inputs, but as the (internal) firing patterns evoked by those inputs at a preceding 'layer' of neurons, this provides stable grounding for our assumption. This is because, whereas raw environmental data may take various forms, evoked neural firing rates are governed fundamentally by log-normal distribution across many parts of the nervous system (Buzsáki and Mizuseki, 2014). Two mechanisms have previously been proposed to explain this phenomenon: First, the log-normal distribution could arise from the high nonlinearity of the neuronal transfer function (Roxin et al., 2011). Second, it could emerge from a linear transformation of log-normally distributed synaptic inputs (Wohrer et al., 2013). The latter is consistent with the heavy-tailed log-normal distribution of synaptic weights observed within the neocortex (Song et al., 2005; Ikegaya et al., 2013), and potentially mirrors the log-normal structure of many natural stimuli, such as images (Huang and Mumford, 1999), sounds (Voss RFClarke, 1975), and language (Piantadosi, 2014).

This said, our framework can, in fact, also be used to interpret information stored about raw external inputs, $\mathbf{x}_{\text{raw}}$. This is due to the invariance of Shannon mutual information under smooth, bijective transformations (Kraskov et al., 2004). As long as such a transformation exists between the raw stimulus $\mathbf{x}_{\text{raw}}$ and the resultant neural representation $\mathbf{x}$, the equality $I(\mathbf{w}; \mathbf{x}_{\text{raw}}) = I(\mathbf{w}; \mathbf{x})$ holds. Consequently, our theory can be understood to also describe how external, real-world patterns, such as images, are stored within Hebbian connectivities.

These characteristics suggest that insights from our framework go beyond the specific assumptions of the model, and may be broadly applicable to how synaptic connections encode information.

4.6 Error analysis for Fenton-Wilkinson approximation

When deriving the probability distribution for the weight ensemble, $p(\mathbf{w})$, we applied the Fenton-Wilkinson approximation to obtain a closed-form expression for subsequent derivations. This choice could introduce errors into the final mutual information estimates. Consequently, we investigated the mathematical properties of this error to provide practical insights into its control, and into heuristics regarding its dependency on other parameters. The complete mathematical derivations and proofs for the following claims are provided in the Supplementary Material.

Specifically, we prove that the error in $MI(\mathbf{w}; \mathbf{x}^l)$ vanishes as the scale of a specific random variable ϵ approaches zero, provided that the ground-truth $\ln \mathbf{w}$ has a bounded probability density function with a finite $(2+\beta)^{\text{th}}$ moment for some positive real value $\beta > 0$, and that the deviation in the probability density estimation is an L^1 function.

Based on this analysis, the primary mechanism for managing the error is to limit the scale of ϵ. Since the definition of ϵ itself is difficult to interpret, we identified a sufficient condition

for controlling its magnitude: the variance of the log-transformed input data patterns should be small along their principal direction. In other words, ensuring that $\text{Var}_{\text{principal}}[\ln \mathbf{x}^k]$ is small for all patterns serves as a practical and interpretable way to suppress the approximation error.

We also derive heuristics regarding the dependency of the error on model parameters d, n, and N. By assuming that the error magnitude scales linearly with the scale of ϵ, we conclude that the error is effectively independent of d and N, while scaling as $\sqrt{n}$ with the size of the weight ensemble. Furthermore, higher correlations across the components of $\ln \mathbf{x}^k$ increase the error magnitude by intensifying the scale of ϵ.

Finally, although our analysis primarily characterizes the limits under which the error vanishes rather than providing an exact quantitative bound, the practical utility of this analysis framework remains. Users can minimize errors by controlling $\text{Var}_{\text{principal}}[\ln \mathbf{x}^k]$ in practice. More importantly, the central conclusions of this work - distributed coding across weights, Hebbian capacity, and synergistic coding - are drawn from qualitative trends and slopes of the information curves. Therefore, these findings are robust to estimation errors and do not require perfectly precise numerical values to remain valid.

5 Methods

5.1 Parameter setup for simulation

In this section, we provide a detailed explanation of the three sets of simulations used in our analyses:

First Simulation Set: We generated N independent data patterns, $\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, where each data pattern $\mathbf{x}^k$ was a d-dimensional random variable following a multivariate log-normal distribution: $\ln(\mathbf{x}^k) \sim \mathcal{N}(\mu^k, \Sigma^k)$. The components of each μ^k were sampled from a uniform distribution between -1 and 1, and the covariance matrix was modeled as an exchangeable matrix:

$$\Sigma^1 = \Sigma^2 = ... = \Sigma^N = \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma}^2 & \ldots & \tilde{\rho}\tilde{\sigma}^2 \\ \tilde{\rho}\tilde{\sigma}^2 & \tilde{\sigma}^2 & \ldots & \tilde{\rho}\tilde{\sigma}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\rho}\tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma}^2 & \ldots & \tilde{\sigma}^2 \end{pmatrix}.$$

The dimension d was chosen to be 20.

We varied the parameters N, n, $\tilde{\sigma}$, and $\tilde{\rho}$ across different analyses. In Figure 2A, we set $n = 1$, $\tilde{\rho} = 0$, systematically varied N and $\tilde{\sigma}$, and calculated the corresponding values of $MI(\mathbf{x}^1; w_{23})$. In Figure 2B, we set $n = 1$, $\tilde{\sigma} = 1$, varied N and $\tilde{\rho}$, and again calculated $MI(\mathbf{x}^1; w_{23})$. In Figure 3A, we set $N = 10$, and $\tilde{\rho} = 0$, varied the size n by incrementally adding one connection from the set of randomly selected 15 connections to the synaptic ensemble $\mathbf{w}$, as well as varied $\tilde{\sigma}$, and computed the corresponding values of $MI(\mathbf{x}^1; \mathbf{w})$. In Figure 3B, we set $N = 10$, and $\tilde{\sigma} = 1$, varied n in the same way as above, as well as varied $\tilde{\rho}$, and computed $MI(\mathbf{x}^1; \mathbf{w})$. In all cases where $\tilde{\rho}$ was varied, it was restricted to the range $\frac{-1}{d-1} \leq \tilde{\rho} \leq 1$. This constraint arises from the requirement that the covariance matrices remain positive semi-definite. Specifically, the eigenvalues of the exchangeable covariance matrix Σ^k are $\tilde{\sigma}^2(1 - \tilde{\rho})$ with multiplicity $d - 1$, and $\tilde{\sigma}^2(1 + (d - 1)\tilde{\rho})$ with multiplicity 1; ensuring non-negativity for

all eigenvalues imposes the stated range on $\tilde{\rho}$.

Second Simulation Set: The objective of this simulation set was to sample multiple distribution-information "tuples", i.e., $\big(\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}, \mathbf{w}, MI(\mathbf{w}; \mathbf{x}^l)\big)$, and to analyze potential relationships between MI and distribution parameters based on these samples. This allowed for a more general exploration of the dependence of MI on the parameters than the one with the first simulation set, which only varied the values of specific individual parameters. To this end, in this set, instead of systematically varying parameters across ranges, all distribution parameters were sampled randomly. The ensemble size n was randomly selected from 1 to 20 with equal probability. The dimension was set to $d = 20$, and the number of patterns N was randomly selected from 2 to 50 with equal probability. The parameters $\{\mu^1, \ldots, \mu^N\}$ were sampled from a uniform distribution between -1 and 1 for each component. The covariance matrices were identical for all patterns but more general than the previous set:

$$\Sigma^1 = \Sigma^2 = ... = \Sigma^N = \mathbf{A}^T\mathbf{A}/d + \begin{pmatrix} \tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma}^2 & \ldots & \tilde{\rho}\tilde{\sigma}^2 \\ \tilde{\rho}\tilde{\sigma}^2 & \tilde{\sigma}^2 & \ldots & \tilde{\rho}\tilde{\sigma}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\rho}\tilde{\sigma}^2 & \tilde{\rho}\tilde{\sigma}^2 & \ldots & \tilde{\sigma}^2 \end{pmatrix}.$$

Here, $\tilde{\sigma}$ was drawn from the absolute values of a standard normal distribution, and $\tilde{\rho}$ was sampled from a uniform distribution in the range $[\frac{-1}{d-1}, 1]$. The matrix $\mathbf{A}$ is stochastic, with each element sampled from a standard normal distribution. Both components, $\mathbf{A}^T\mathbf{A}/d$ and the exchangeable covariance matrix, are positive semi-definite, ensuring that their sum is a valid covariance matrix. The introduction of $\mathbf{A}^T\mathbf{A}/d$ adds degrees of freedom, but it tends to produce matrices with larger diagonal elements, potentially neglecting distributions with highly correlated components. In contrast, the original exchangeable matrix can capture high correlations between components due to its structure, but it lacks sufficient degrees of freedom to account for more complex variability in the data. The $1/d$ scaling factor was designed to prevent either component from dominating. Consequently, by combining the two components, the resulting covariance matrix strikes a balance between flexibly capturing a wider range of data patterns, and the ability to model correlations, thereby providing a richer and more balanced representation of the underlying data distributions.

Using this setup, we generated 1000 samples of $\big(\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}, \mathbf{w}, MI(\mathbf{w}; \mathbf{x}^l)\big)$. For Figure 2C, we analyzed the regression between MI and N. For Figure 3C, we analyzed the regression between MI and n.

Third Simulation Set: The third set of simulations followed a similar approach to the second in that we sampled numerous distribution-information "tuples" $\big(\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}, \mathbf{w}, MI(\mathbf{w}; \mathbf{x}^l)\big)$, the ensemble size n was randomly selected from 1 to 20 with equal probability, the dimension d was set to 20, the number of patterns N was randomly selected from 2 to 50 with equal probability, and the parameters $\{\mu^1, \ldots, \mu^N\}$ were sampled from a uniform distribution between -1 and 1 for each component. The key difference was that the covariance matrices varied across patterns, following:

$$\Sigma^k = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T + r\mathbf{1}\mathbf{1}^T.$$

Here, $\mathbf{Q}$ is a random orthogonal matrix generated using the common practice of performing QR decomposition on a matrix whose elements were drawn from a standard normal distribution.

$\mathbf{\Lambda}$ is a diagonal matrix with random positive eigenvalues sampled from a normal distribution (mean 0.01, standard deviation 1.2), which is slightly broader than the standard normal to enable richer representations. The scalar factor r was sampled from a uniform distribution in the range [0, 1.44] to prevent either component from dominating. Similar to the second simulation set, by combining these two components in the construction of the covariance matrices, we captured both the flexibility offered by $\mathbf{Q\Lambda Q}^T$ and the structure in $\mathbf{1}\mathbf{1}^T$ needed to model strong correlations across components.

Using this setup, we generated 1000 samples. In Figure 2D, we analyzed the regression between MI and N, while in Figure 3D, we analyzed the regression between MI and n. For the analysis of MI/n in Figure 4, we generated an additional 20,000 samples.

5.2 Proofs for propositions

Proposition 1.1. *For N independent log-normally distributed patterns $\{\mathbf{x}^1, ..., \mathbf{x}^N\}$, the distribution of the synaptic connection $w_{ij} = \sum_{k=1}^{N} \mathbf{x}_i^k \mathbf{x}_j^k$ in the Hebbian network is approximated by a new log-normal distribution:*

$$p(w_{ij}) \approx \frac{1}{w_{ij}\sigma_{w_{ij}}\sqrt{2\pi}} \exp\left(-\frac{(\ln w_{ij} - \mu_{w_{ij}})^2}{2\sigma_{w_{ij}}^2}\right) \tag{21}$$

with $\mu_{w_{ij}}, \sigma_{w_{ij}}$ defined in equation 26, 27.

Proof. Each data pattern is assumed to follow $\ln(\mathbf{x}^k) \sim \mathcal{N}(\mu^k, \Sigma^k)$ in equation 1. Since the sum of two components in a multivariate normal distribution is normal, we have:

$$\ln(\mathbf{x}_i^k \mathbf{x}_j^k) = \ln(\mathbf{x}_i^k) + \ln(\mathbf{x}_j^k) \sim \mathcal{N}(\mu_i^k + \mu_j^k, (\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k), \tag{22}$$

which means any term of the product form $\mathbf{x}_i^k \mathbf{x}_j^k$ follows a log-normal distribution. The sum of such terms from N independent data patterns, $\sum_{k=1}^{N} \mathbf{x}_i^k \mathbf{x}_j^k$, represents the connection weight between two neurons, w_{ij}, as defined as in equation 2. Therefore, studying the distribution of w_{ij} is equivalent to studying the distribution of the sum of several independent log-normal random variables. Using Fenton-Wilkinson method, we can approximate this distribution, $p(w_{ij})$, by a new log-normal distribution $p(\hat{w}_{ij}) \approx p(w_{ij})$ with the same first and second moments. In other words, we aim to derive a log-normal distribution:

$$\ln(\hat{w}_{ij}) \sim \mathcal{N}\left(\mu_{w_{ij}}, (\sigma_{w_{ij}})^2\right) \tag{23}$$

whose first two moments match those of w_{ij}:

$$\mathbb{E}[\hat{w}_{ij}] = \mathbb{E}[w_{ij}],\ \mathrm{Var}[\hat{w}_{ij}] = \mathrm{Var}[w_{ij}].$$

The values for $\mathbb{E}[w_{ij}]$ and $\mathrm{Var}[w_{ij}]$ can be derived based on independence among patterns and the properties of the log-normal distribution*. For convenience, we denote them as M_1 and M_2

*For random variable $\ln X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = e^{\mu + \frac{1}{2}\sigma^2}$, $\mathrm{Var}[X] = [e^{\sigma^2} - 1]e^{2\mu + \sigma^2}$.

from here on.

$$M_1 := \mathbb{E}[w_{ij}] = \mathbb{E}[\sum_{k=1}^{N} \mathbf{x}_i^k \mathbf{x}_j^k] = \sum_{k=1}^{N} \mathbb{E}[\mathbf{x}_i^k \mathbf{x}_j^k]$$
$$= \sum_{k=1}^{N} \exp\left(\mu_i^k + \mu_j^k + \frac{1}{2}\left[(\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right]\right) \tag{24}$$

$$M_2 := \mathrm{Var}[w_{ij}] = \mathrm{Var}[\sum_{k=1}^{N} \mathbf{x}_i^k \mathbf{x}_j^k] = \sum_{k=1}^{N} \mathrm{Var}[\mathbf{x}_i^k \mathbf{x}_j^k]$$
$$= \sum_{k=1}^{N} \left[\exp\left((\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right) - 1\right] \exp\left(2(\mu_i^k + \mu_j^k) + (\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right) \tag{25}$$

Given $\mathbb{E}[\hat{w}_{ij}]$ and $\mathrm{Var}[\hat{w}_{ij}]$, the the parameters $\mu_{w_{ij}}$ and $\sigma^2_{w_{ij}}$ in equation 23 are derived as follows:

$$\mu_{w_{ij}} = \ln M_1 - \frac{1}{2}\ln(1 + \frac{M_2}{{M_1}^2}), \tag{26}$$

$$\sigma^2_{w_{ij}} = \ln(1 + \frac{M_2}{{M_1}^2}). \tag{27}$$

In this way, we obtain a distribution for a connection weight that is fully parametrized by the known data pattern distributions. □

Proposition 1.2. *The mutual information between a synaptic connection w_{ij} and a data pattern $\mathbf{x}^l$ is given by the analytical approximation:*

$$MI(w_{ij}; \mathbf{x}^l) = \left(\mu_{w_{ij}} - \mu_{w_{ij/l}}\right) + \left(\ln \sigma_{w_{ij}} - \ln \sigma_{w_{ij/l}}\right),$$

with $\mu_{w_{ij}}, \sigma_{w_{ij}}$ defined in equation 26, 27; $\mu_{w_{ij/l}}, \sigma_{w_{ij/l}}$ defined in equation 28, 29. The unit of MI is nat.

Proof. We can calculate the mutual information $MI(w_{ij}; \mathbf{x}^l)$ as the difference of differential

entropies $h(w_{ij}) - h(w_{ij}|\mathbf{x}^l)$, which leads to:

$$
\begin{aligned}
MI(w_{ij};\mathbf{x}^l) &= h(w_{ij}) - h(w_{ij}|\mathbf{x}^l) \\
&= h(w_{ij}) + \iint p(\mathbf{x}^l)p(w_{ij}|\mathbf{x}^l)\log p(w_{ij}|\mathbf{x}^l)\,\mathrm{d}w_{ij}\mathrm{d}\mathbf{x}^l \\
&= h(w_{ij}) + \int p(\mathbf{x}^l)\left(\int p(w_{ij}|\mathbf{x}^l)\log p(w_{ij}|\mathbf{x}^l)\,\mathrm{d}w_{ij}\right)\mathrm{d}\mathbf{x}^l \\
&= h(w_{ij}) - \int p(\mathbf{x}^l)h(\sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k)\mathrm{d}\mathbf{x}^l \\
&= h(w_{ij}) - h(\sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k)
\end{aligned}
$$

In the second-to-last step, we use the fact that $p(w_{ij}|\mathbf{x}^l)$ and $p(\sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k)$ have the same differential entropy. This is because, given independence among patterns, $p(w_{ij}|\mathbf{x}^l) := p(\sum_{k}^{N}\mathbf{x}_i^k\mathbf{x}_j^k|\mathbf{x}^l)$ and $p(\sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k)$ are shifted versions of each other. For convenience, we denote $w_{ij/l} := \sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k$ in what follows. Observe that the same reasoning represented in Proposition 1.1 also extends to $w_{ij/l}$, indicating it too can be also approximated by a log-normal distribution.

The $MI(w_{ij};\mathbf{x}^l)$ is therefore the difference between the differential entropies of two log-normal variables, w_{ij} and $w_{ij/l}$. Considering the property of log-normal random variables[†], we have:

$$
\begin{aligned}
h(w_{ij}) &= \ln(\sqrt{2\pi}\sigma_{w_{ij}}e^{\mu_{w_{ij}}+\frac{1}{2}}), \\
h(w_{ij/l}) &= \ln(\sqrt{2\pi}\sigma_{w_{ij/l}}e^{\mu_{w_{ij/l}}+\frac{1}{2}}).
\end{aligned}
$$

Consequently, we obtain the expression for mutual information:

$$
\begin{aligned}
MI(w_{ij};\mathbf{x}^l) &= \ln(\sqrt{2\pi}\sigma_{w_{ij}}e^{\mu_{w_{ij}}+\frac{1}{2}}) - \ln(\sqrt{2\pi}\sigma_{w_{ij/l}}e^{\mu_{w_{ij/l}}+\frac{1}{2}}) \\
&= \left(\mu_{w_{ij}} - \mu_{w_{ij/l}}\right) + \left(\ln\sigma_{w_{ij}} - \ln\sigma_{w_{ij/l}}\right)
\end{aligned}
$$

For the sake of completeness, we explicitly write the parameters for $p(w_{ij/l})$ here. By adapting equation 23 to 27, we have:

$$
\ln w_{ij/l} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_{w_{ij/l}}, (\sigma_{w_{ij/l}})^2\right),
$$

[†]For random variable $\ln X \sim \mathcal{N}(\mu,\sigma^2)$, its entropy equals $\ln(\sqrt{2\pi}\sigma e^{\mu+\frac{1}{2}})$ in nats.

where:

$$\mu_{w_{ij/l}} = \ln M_{1/l} - \frac{1}{2}\ln(1 + \frac{M_{2/l}}{M_{1/l}^2}) \tag{28}$$

$$\sigma^2_{w_{ij/l}} = \ln(1 + \frac{M_{2/l}}{M_{1/l}^2}) \tag{29}$$

$$M_{1/l} = \sum_{k\neq l}^{N} \exp\left(\mu_i^k + \mu_j^k + \frac{1}{2}\left[(\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right]\right)$$

$$M_{2/l} = \sum_{k=1}^{N} \left[\exp\left((\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right) - 1\right] \times \exp\left(2(\mu_i^k + \mu_j^k) + (\sigma_i^k)^2 + (\sigma_j^k)^2 + 2\sigma_{ij}^k\right)$$

□

Proposition 2.1. *For a pair of synaptic connections* $\mathbf{w} = (w_{ij}, w_{mn})^T$ *following a 2-dimensional log-normal distribution, the covariance between* $\ln w_{ij}$ *and* $\ln w_{mn}$ *is:*

$$\sigma_{w_{ij,mn}} = \ln(1 + \frac{\sum_{k=1}^{N} \exp\left(\sum_{\phi}^{\{i,j,m,n\}} \mu_\phi^k + \frac{1}{2}\sum_{\phi}^{\{i,j,m,n\}}(\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}} \sigma_\phi^k\right)\left[\exp\left(\sum_{\phi}^{\{i,j\}\times\{m,n\}} \sigma_\phi^k\right) - 1\right]}{\sum_{k,l=1}^{N} \exp\left(\sum_{\phi}^{\{i,j\}} \mu_\phi^k + \sum_{\phi}^{\{m,n\}} \mu_\phi^l + \frac{1}{2}\sum_{\phi}^{\{i,j\}}(\sigma_\phi^k)^2 + \frac{1}{2}\sum_{\phi}^{\{m,n\}}(\sigma_\phi^l)^2 + \sigma_{ij}^k + \sigma_{mn}^l\right)}).$$

Proof. We have assumed that the ensemble $\mathbf{w}$ approximately follows 2-dimensional log-normal distribution, which states:

$$\ln \mathbf{w} = \begin{pmatrix} \ln w_{ij} \\ \ln w_{mn} \end{pmatrix} \overset{\text{approx.}}{\sim} \mathcal{N}(\mu_{\mathbf{w}} = \begin{pmatrix} \mu_{w_{ij}} \\ \mu_{w_{mn}} \end{pmatrix}, \Sigma_{\mathbf{w}} = \begin{pmatrix} \sigma^2_{w_{ij}} & \sigma_{w_{ij,mn}} \\ \sigma_{w_{mn,ij}} & \sigma^2_{w_{mn}} \end{pmatrix}).$$

Given the property of a multivariate log-normal random variable[‡], the covariance between two components of random variable $\mathbf{w}$, $\mathrm{Cov}[w_{ij}, w_{mn}]$, and the covariance between their logarithms, $\sigma_{w_{ij,mn}}$, are related by the following equation:

$$\mathrm{Cov}[w_{ij}, w_{mn}] = \left[\exp(\sigma_{w_{ij,mn}}) - 1\right] \exp\left(\mu_{w_{ij}} + \mu_{w_{mn}} + \frac{1}{2}(\sigma^2_{w_{ij}} + \sigma^2_{w_{mn}})\right),$$

which is equivalent to:

$$\sigma_{w_{ij,mn}} = \ln\left(1 + \frac{\mathrm{Cov}[w_{ij}, w_{mn}]}{\exp\left[\mu_{w_{ij}} + \mu_{w_{mn}} + \frac{1}{2}(\sigma^2_{w_{ij}} + \sigma^2_{w_{mn}})\right]}\right) \tag{30}$$

Given that the expressions for $\mu_{w_{ij}}, \mu_{w_{mn}}$ and $\sigma^2_{w_{ij}}, \sigma^2_{w_{mn}}$ have been derived in equation 26

[‡]For multivariate random variable $\ln X \sim \mathcal{N}(\mu, \Sigma)$, $\mathrm{Var}[X]_{ij} = (e^{\Sigma_{ij}} - 1)e^{\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj})}$

and 27, the problem reduces to finding $\mathrm{Cov}[w_{ij}, w_{mn}]$. From its definition, we have:

$$
\begin{aligned}
&\mathrm{Cov}[w_{ij}, w_{mn}] = \mathbb{E}[w_{ij}w_{mn}] - \mathbb{E}[w_{ij}]\mathbb{E}[w_{mn}] \\
&= \mathbb{E}\left[(\sum_{k=1}^{N}\mathbf{x}_i^k\mathbf{x}_j^k)(\sum_{l=1}^{N}\mathbf{x}_m^l\mathbf{x}_n^l)\right] - \mathbb{E}\left[\sum_{k=1}^{N}\mathbf{x}_i^k\mathbf{x}_j^k\right]\mathbb{E}\left[\sum_{l=1}^{N}\mathbf{x}_m^l\mathbf{x}_n^l\right] \\
&= \sum_{k,l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^l\mathbf{x}_n^l\right] - \sum_{k=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\right]\sum_{l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^l\mathbf{x}_j^l\right] \\
&= \left(\sum_{k\neq l\ k,l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^l\mathbf{x}_n^l\right] + \sum_{k=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k\right]\right) - \sum_{k,l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\right]\mathbb{E}\left[\mathbf{x}_i^l\mathbf{x}_j^l\right] \\
&= \sum_{k\neq l\ k,l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\right]\mathbb{E}\left[\mathbf{x}_m^l\mathbf{x}_n^l\right] + \sum_{k=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k\right] - \sum_{k,l=1}^{N}\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\right]\mathbb{E}\left[\mathbf{x}_i^l\mathbf{x}_j^l\right] \\
&= \sum_{k=1}^{N}\left(\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k\right] - \mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\right]\mathbb{E}\left[\mathbf{x}_m^k\mathbf{x}_n^k\right]\right). \qquad (31)
\end{aligned}
$$

The second-to-last step uses the independence condition between data different patterns.

To derive the term $\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k\right]$ in the last line, recall that $\mathbf{x}_i^k, \mathbf{x}_j^k, \mathbf{x}_m^k, \mathbf{x}_n^k$ are four components of the same multivariate log-normal random vector $\mathbf{x}^k$. Thus, the product $\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k$ is log-normally distributed. Therefore, by extending the same reasoning in equation 22, we get:

$$
\ln(\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k) \sim \mathcal{N}\left(\sum_{\phi}^{\{i,j,m,n\}}\mu_\phi^k, \sum_{\phi}^{\{i,j,m,n\}}(\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}}2\sigma_\phi^k + \sum_{\phi}^{\{i,j\}\times\{m,n\}}2\sigma_\phi^k\right),
$$

where the different summations are defined as follows:

$$
\begin{aligned}
&\sum_{\phi}^{\{i,j,m,n\}}\mu_\phi^k = \mu_i^k + \mu_j^k + \mu_m^k + \mu_n^k, \\
&\sum_{\phi}^{\{i,j,m,n\}}(\sigma_\phi^k)^2 = (\sigma_i^k)^2 + (\sigma_j^k)^2 + (\sigma_m^k)^2 + (\sigma_n^k)^2, \\
&\sum_{\phi}^{\{ij,mn\}}2\sigma_\phi^k = 2\sigma_{ij}^k + 2\sigma_{mn}^k\ , \quad \sum_{\phi}^{\{i,j\}\times\{m,n\}}2\sigma_\phi^k = 2\sigma_{im}^k + 2\sigma_{in}^k + 2\sigma_{jm}^k + 2\sigma_{jn}^k.
\end{aligned}
$$

Consequently, the mean of this quadruple product term is:

$$
\mathbb{E}\left[\mathbf{x}_i^k\mathbf{x}_j^k\mathbf{x}_m^k\mathbf{x}_n^k\right] = \exp\left(\sum_{\phi}^{\{i,j,m,n\}}\mu_\phi^k + \frac{1}{2}\sum_{\phi}^{\{i,j,m,n\}}(\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}}\sigma_\phi^k + \sum_{\phi}^{\{i,j\}\times\{m,n\}}\sigma_\phi^k\right). \qquad (32)
$$

For the term $\mathbb{E}\left[\mathbf{x}_i^k \mathbf{x}_j^k\right] \mathbb{E}\left[\mathbf{x}_m^k \mathbf{x}_n^k\right]$ in the equation 31, given that $\mathbb{E}\left[\mathbf{x}_i^k \mathbf{x}_j^k\right]$ and $\mathbb{E}\left[\mathbf{x}_m^k \mathbf{x}_n^k\right]$ were obtained as components in the summation in equation 24, we get:

$$\mathbb{E}\left[\mathbf{x}_i^k \mathbf{x}_j^k\right] \mathbb{E}\left[\mathbf{x}_m^k \mathbf{x}_n^k\right] = \exp\left(\sum_{\phi}^{\{i,j,m,n\}} \mu_\phi^k + \frac{1}{2} \sum_{\phi}^{\{i,j,m,n\}} (\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}} \sigma_\phi^k\right). \tag{33}$$

Substituting the results from equations 32 and 33 into equation 31, we obtain the covariance between two synaptic weights:

$$\begin{aligned}\mathrm{Cov}[w_{ij}, w_{mn}] = \sum_{k=1}^{N} \exp\left(\sum_{\phi}^{\{i,j,m,n\}} \mu_\phi^k + \frac{1}{2} \sum_{\phi}^{\{i,j,m,n\}} (\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}} \sigma_\phi^k\right) \times \\ \left[\exp\left(\sum_{\phi}^{\{i,j\}\times\{m,n\}} \sigma_\phi^k\right) - 1\right].\end{aligned} \tag{34}$$

Finally, by combining the covariance $\mathrm{Cov}[w_{ij}, w_{mn}]$ with equations 26, 27, 30, we obtain an expression for $\sigma_{w_{ij,mn}}$ as a function of the parameters characterizing the data pattern distributions:

$$\begin{aligned}&\sigma_{w_{ij,mn}} = \ln(1 + \\ &\frac{\sum_{k=1}^{N} \exp\left(\sum_{\phi}^{\{i,j,m,n\}} \mu_\phi^k + \frac{1}{2}\sum_{\phi}^{\{i,j,m,n\}} (\sigma_\phi^k)^2 + \sum_{\phi}^{\{ij,mn\}} \sigma_\phi^k\right)\left[\exp\left(\sum_{\phi}^{\{i,j\}\times\{m,n\}} \sigma_\phi^k\right) - 1\right]}{\sum_{k,l=1}^{N} \exp\left(\sum_{\phi}^{\{i,j\}} \mu_\phi^k + \sum_{\phi}^{\{m,n\}} \mu_\phi^l + \frac{1}{2}\sum_{\phi}^{\{i,j\}} (\sigma_\phi^k)^2 + \frac{1}{2}\sum_{\phi}^{\{m,n\}} (\sigma_\phi^l)^2 + \sigma_{ij}^k + \sigma_{mn}^l\right)}).\end{aligned} \tag{35}$$

□

Proposition 2.2. *Under the approximation that both* $\mathbf{w} := (w_{ij}, w_{mn})^T$ *and* $\mathbf{w}_{/l} := (w_{ij/l}, w_{mn/l})^T = (\sum_{k\neq l}^{N} \mathbf{x}_i^k \mathbf{x}_j^k,\ \sum_{k\neq l}^{N} \mathbf{x}_m^k \mathbf{x}_n^k)^T$ *follow a multivariate log-normal distribution, i.e.* $\ln \mathbf{w} \sim \mathcal{N}(\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}})$ *and* $\ln \mathbf{w}_{/l} \sim \mathcal{N}(\mu_{\mathbf{w}_{/l}}, \Sigma_{\mathbf{w}_{/l}})$*, the mutual information between the pair of connections* $\mathbf{w}$ *and a data pattern* $\mathbf{x}^l$ *can be expressed as:*

$$MI(\mathbf{w}; \mathbf{x}^l) = (\mu_{w_{ij}} + \mu_{w_{mn}} - \mu_{w_{ij/l}} - \mu_{w_{mn/l}}) + \frac{1}{2}\ln|\Sigma_{\mathbf{w}}| - \frac{1}{2}\ln|\Sigma_{\mathbf{w}_{/l}}|,$$

where the expressions for $\mu_{w_{ij}}$*,* $\mu_{w_{mn}}$ *are given in equation 26, and for* $\mu_{w_{ij/l}}$*,* $\mu_{w_{mn/l}}$ *in equation 28. For matrix* $\Sigma_{\mathbf{w}}$*, the diagonal elements are given by equation 27, and the off-diagonal elements by equation 35; for* $\Sigma_{\mathbf{w}_{/l}}$*, the corresponding expressions are provided by equations 29 and 36. Here,* $|\Sigma_{\mathbf{w}}|$ *denotes the determinant of the matrix* $\Sigma_{\mathbf{w}}$*. All information values are expressed in nats.*

Proof. Given that $\mathbf{w}$ follows a 2-dimensional log-normal variable, its differential entropy (in

nats) can be written as[§]:

$$h(\mathbf{w}) = \frac{1}{2}\ln\left((2\pi e)^2|\Sigma_{\mathbf{w}}|\right) + \mu_{w_{ij}} + \mu_{w_{mn}}.$$

For $h(\mathbf{w}|\mathbf{x}^l)$, observe that $p(\mathbf{w}|\mathbf{x}^l)$ is a shifted version of $p(\mathbf{w}_{/l}) := p(\sum_{k\neq l}^{N}\mathbf{x}_i^k\mathbf{x}_j^k,\ \sum_{k\neq l}^{N}\mathbf{x}_m^k\mathbf{x}_n^k)$. Therefore, the two distributions have the same entropy. This leads to:

$$h(\mathbf{w}|\mathbf{x}^l) = h(\mathbf{w}_{/l}) = \frac{1}{2}\ln\left((2\pi e)^2|\Sigma_{\mathbf{w}_{/l}}|\right) + \mu_{w_{ij/l}} + \mu_{w_{mn/l}}.$$

The expressions for $\mu_{w_{ij/l}}, \mu_{w_{mn/l}}$, and for the diagonal elements of the covariance matrix, $\sigma^2_{w_{ij/l}}, \sigma^2_{w_{mn/l}}$, are given in the proof of Proposition 1.2. For the off-diagonal element $\sigma_{w_{ij,mn}/l}$, its derivation mirrors that of $\sigma_{w_{ij,mn}}$ in Proposition 2.1. For clarity, we present the explicit formula here:

$$\sigma_{w_{ij,mn/l}} = \ln(1 + \frac{\sum_{k\neq l}^{N}\exp\left(\sum_{\phi}^{\{i,j,m,n\}}\mu_{\phi}^k + \frac{1}{2}\sum_{\phi}^{\{i,j,m,n\}}(\sigma_{\phi}^k)^2 + \sum_{\phi}^{\{ij,mn\}}\sigma_{\phi}^k\right)\left[\exp\left(\sum_{\phi}^{\{i,j\}\times\{m,n\}}\sigma_{\phi}^k\right) - 1\right]}{\sum_{k,g\neq l}^{N}\exp\left(\sum_{\phi}^{\{i,j\}}\mu_{\phi}^k + \sum_{\phi}^{\{m,n\}}\mu_{\phi}^g + \frac{1}{2}\sum_{\phi}^{\{i,j\}}(\sigma_{\phi}^k)^2 + \frac{1}{2}\sum_{\phi}^{\{m,n\}}(\sigma_{\phi}^g)^2 + \sigma_{ij}^k + \sigma_{mn}^g\right)}). \tag{36}$$

With $h(\mathbf{w})$ and $h(\mathbf{w}|\mathbf{x}^l)$ computed, we can now obtain the mutual information between any pair of synaptic interconnections and a data pattern:

$$MI(\mathbf{w};\mathbf{x}^l) = h(\mathbf{w}) - h(\mathbf{w}|\mathbf{x}^l) = (\mu_{w_{ij}} + \mu_{w_{mn}} - \mu_{w_{ij/l}} - \mu_{w_{mn/l}}) + \frac{1}{2}\ln|\Sigma_{\mathbf{w}}| - \frac{1}{2}\ln|\Sigma_{\mathbf{w}_{/l}}|.$$

□

Proposition 3. *For a synaptic ensemble of n connections, $\mathbf{w} := (w_{ij(1)}, w_{ij(2)}, ..., w_{ij(n)})^T$, where each $w_{ij(k)}$ denotes the connection between neuron $i(k)$ and $j(k)$, and under the approximation that both $\mathbf{w}$ and $\mathbf{w}_{/l} := (w_{ij(1)/l}, w_{ij(2)/l}, ..., w_{ij(n)/l})^T$ follow a multivariate log-normal distribution, i.e. $\ln\mathbf{w} \sim \mathcal{N}(\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}})$ and $\ln\mathbf{w}_{/l} \sim \mathcal{N}(\mu_{\mathbf{w}_{/l}}, \Sigma_{\mathbf{w}_{/l}})$, the mutual information between the joint activity of these n synaptic connections $\mathbf{w}$ and a data pattern $\mathbf{x}^l$ is given by the analytical approximation:*

$$MI(\mathbf{w};\mathbf{x}^l) = \sum_{k=1}^{n}\left(\mu_{w_{ij(k)}} - \mu_{w_{ij(k)/l}}\right) + \frac{1}{2}\ln|\Sigma_{\mathbf{w}}| - \frac{1}{2}\ln|\Sigma_{\mathbf{w}_{/l}}|.$$

where the expressions for $\mu_{w_{ij(k)}}$ are given in equation 26, and for $\mu_{w_{ij(k)/l}}$ in equation 28. The diagonal and off-diagonal elements of $\Sigma_{\mathbf{w}}$ are given by equations 27 and 35, respectively; for $\Sigma_{\mathbf{w}_{/l}}$, by equations 29 and 36, respectively. All information values are expressed in nats.

Proof. Based on the properties of an n-dimensional log-normal variable, the differential entropy

[§]For n-dimensional multivariate random variable $\ln X \sim \mathcal{N}(\mu, \Sigma)$, its differential entropy $h(X) = \frac{1}{2}\ln((2\pi e)^n|\Sigma|) + \sum_i \mu_i$ in nats.

of $\mathbf{w}$ is given by:

$$h(\mathbf{w}) = \frac{1}{2}\ln\left((2\pi e)^n|\Sigma_{\mathbf{w}}|\right) + \sum_{k=1}^{n}\mu_{w_{ij(k)}}.$$

Meanwhile, $p(\mathbf{w}|\mathbf{x}^l)$ and $p(\mathbf{w}_{/l})$ have the same entropy owing to their shifting relation. Therefore, we have:

$$h(\mathbf{w}|\mathbf{x}^l) = h(\mathbf{w}_{/l}) = \frac{1}{2}\ln\left((2\pi e)^n|\Sigma_{\mathbf{w}_{/l}}|\right) + \sum_{k=1}^{n}\mu_{w_{ij(k)/l}}.$$

Consequently, the information stored in ensemble $\mathbf{w}$ about pattern $\mathbf{x}^l$ is:

$$MI(\mathbf{w};\mathbf{x}^l) = h(\mathbf{w}) - h(\mathbf{w}|\mathbf{x}^l) = \sum_{k=1}^{n}\left(\mu_{w_{ij(k)}} - \mu_{w_{ij(k)/l}}\right) + \frac{1}{2}\ln|\Sigma_{\mathbf{w}}| - \frac{1}{2}\ln|\Sigma_{\mathbf{w}_{/l}}|.$$

Since the derivation in Proposition 1.1, 1.2 is valid for any single connection $w_{ij(k)}$, and that in Proposition 2.1 for any pair, the expressions derived earlier can be directly applied to the quantities in this equation. □

Acknowledgments

This work was supported by funding from a JHU Discovery Award co-funded with the One Neuro Initiative (SPM), and from NIH grant 2R01EY027718 (SPM).

Competing Interests

The author declare no competing interests.

Author Contributions

Conceptualization: X.F. and S.P.M.; Theoretical construction: X.F.; Formal analysis: X.F.; Funding acquisition: S.P.M.; Supervision: S.P.M.; Visualization: X.F. and S.P.M.; Writing and Editing: X.F. and S.P.M.

Code Availability

The code used to generate the simulations and figures in this paper can be found on Github at https://github.com/PixelsForest/weight-information .

References

Abu-Mostafa, Y. and Jacques, J. S. (1985). Information capacity of the hopfield model. *IEEE Transactions on Information Theory*, 31(4):461–464.

Achille, A. and Soatto, S. (2017). On the emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 125(126-127):14.

Akrout, M., Wilson, C., Humphreys, P., Lillicrap, T., and Tweed, D. B. (2019). Deep learning without weight transport. *Advances in neural information processing systems*, 32.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530.

Baeg, E. H., Kim, Y. B., Kim, J., Ghim, J.-W., Kim, J. J., and Jung, M. W. (2007). Learning-induced enduring changes in functional connectivity among prefrontal cortical neurons. *Journal of Neuroscience*, 27(4):909–918.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646.

Battiston, F., Nicosia, V., Chavez, M., and Latora, V. (2017). Multilayer motif analysis of brain networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4).

Bialek, W. and Rieke, F. (1992). Reliability and information transmission in spiking neurons. *Trends in neurosciences*, 15(11):428–434.

Borst, A. and Theunissen, F. E. (1999). Information theory and neural coding. *Nature neuroscience*, 2(11):947–957.

Bosch, H. and Kurfess, F. J. (1998). Information storage capacity of incompletely connected associative memories. *Neural Networks*, 11(5):869–876.

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., and Steveninck, R. R. d. R. v. (2000). Synergy in a neural code. *Neural computation*, 12(7):1531–1552.

Buzsáki, G. and Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nature reviews neuroscience*, 15(4):264–278.

Cayco-Gajic, N. A., Zylberberg, J., and Shea-Brown, E. (2015). Triplet correlations among similarly tuned cells impact population coding. *Frontiers in computational neuroscience*, 9:57.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Dimitrov, A. G., Lazar, A. A., and Victor, J. D. (2011). Information theory in neuroscience. *Journal of computational neuroscience*, 30:1–5.

Dong, D. W. and Hopfield, J. J. (1992). Dynamic properties of neural networks with adapting synapses. *Network: Computation in Neural Systems*, 3(3):267.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature neuroscience*, 3(11):1184–1191.

Fan, X. and Mysore, S. P. (2025). Contribute to balance, wire in accordance: Emergence of backpropagation from a simple, bio-plausible neuroplasticity rule. *ArXiv*, pages arXiv–2405.

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on communications systems*, 8(1):57–67.

Gawne, T. J. and Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771.

Harris, K. D. (2008). Stability of the fittest: organizing learning through retroaxonal signals. *Trends in neurosciences*, 31(3):130–136.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley, New York.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

Hopfield, J. (1991). Olfactory computation and object perception. *Proceedings of the National Academy of Sciences*, 88(15):6462–6466.

Huang, J. and Mumford, D. (1999). Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 541–547. IEEE.

Ikegaya, Y., Sasaki, T., Ishikawa, D., Honma, N., Tao, K., Takahashi, N., Minamisawa, G., Ujita, S., and Matsuki, N. (2013). Interpyramid spike transmission stabilizes the sparseness of recurrent network activity. *Cerebral Cortex*, 23(2):293–304.

Kafashan, M., Jaffe, A. W., Chettih, S. N., Nogueira, R., Arandia-Romero, I., Harvey, C. D., Moreno-Bote, R., and Drugowitsch, J. (2021). Scaling of sensory information in large neural populations shows signatures of information-limiting correlations. *Nature communications*, 12(1):473.

Kang, L. and Toyoizumi, T. (2024). Distinguishing examples while building concepts in hippocampal and artificial networks. *Nature Communications*, 15(1):647.

Knoblauch, A., Palm, G., and Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289–341.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138.

Lamprecht, R. and LeDoux, J. (2004). Structural plasticity and memory. *Nature Reviews Neuroscience*, 5(1):45–54.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):13276.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Luppi, A. I., Mediano, P. A., Rosas, F. E., Holland, N., Fryer, T. D., O'Brien, J. T., Rowe, J. B., Menon, D. K., Bor, D., and Stamatakis, E. A. (2022). A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, 25(6):771–782.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, 319(5869):1543–1546.

Nadal, J.-P. (1991). Associative memory: on the (puzzling) sparse coding limit. *Journal of Physics A: Mathematical and General*, 24(5):1093.

Palm, G. (1980). On associative memory. *Biological cybernetics*, 36(1):19–31.

Palm, G. (1992). On the information storage capacity of local learning rules. *Neural Computation*, 4(5):703–711.

Palm, G. and Sommer, F. T. (1996). Associative data storage and retrieval in neural networks. In *Models of Neural Networks III: Association, Generalization, and Representation*, pages 79–118. Springer.

Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913.

Panichello, M. F., Jonikaitis, D., Oh, Y. J., Zhu, S., Trepka, E. B., and Moore, T. (2024). Intermittent rate coding and cue-specific ensembles support working memory. *Nature*, pages 1–8.

Panzeri, S., Schultz, S. R., Treves, A., and Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1423):1001–1012.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130.

Quian Quiroga, R. and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185.

Rissman, J. and Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual review of psychology*, 63(1):101–128.

Roxin, A., Brunel, N., Hansel, D., Mongillo, G., and Van Vreeswijk, C. (2011). On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45):16217–16226.

Schneidman, E., Bialek, W., and Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology*, 3(3):e68.

Sporns, O. and Kötter, R. (2004). Motifs in brain networks. *PLoS biology*, 2(11):e369.

Stecker, G. C. and Middlebrooks, J. C. (2003). Distributed coding of sound locations in the auditory cortex. *Biological cybernetics*, 89(5):341–349.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273.

Stokes, M. G. (2015). 'activity-silent'working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, 19(7):394–405.

Sun, W., Advani, M., Spruston, N., Saxe, A., and Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, 26(8):1438–1448.

Timme, N. M. and Lapish, C. (2018). A tutorial for information theory in neuroscience. *eneuro*, 5(3).

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Voss RFClarke, J. (1975). 1/f noise'in music and speech. *Nature*, 258:317318.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222(5197):960–962.

Wohrer, A., Humphries, M. D., and Machens, C. K. (2013). Population-wide distributions of neural activity during perceptual decision-making. *Progress in neurobiology*, 103:156–193.

Xie, X. and Seung, H. S. (2003). Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15(2):441–454.

Error Analysis of the Fenton–Wilkinson Approximation in Differential Entropy Estimation

In the main text, we derived an expression for $MI(\mathbf{w};\mathbf{x}^l)$, which quantifies the information stored in an arbitrary ensemble $\mathbf{w}$ about a selected data pattern $\mathbf{x}^l$. This quantity is given by the difference between two differential entropy terms, $h(\mathbf{w}) - h(\mathbf{w}|\mathbf{x}^l)$, where each term depends on the probability distributions $p(\mathbf{w})$ and $p(\mathbf{w}_{/l})$, respectively.

As a reminder, we defined $\mathbf{w} := (w_{ij(1)}, w_{ij(2)}, \ldots, w_{ij(n)})^T$ and $\mathbf{w}_{/l} := (w_{ij(1)/l}, w_{ij(2)/l}, ..., w_{ij(n)/l})^T$, with $w_{ij(h)} = \sum_{k=1}^{N} \mathbf{x}^k_{i(h)}\mathbf{x}^k_{j(h)}$ and $w_{ij(h)/l} = \sum_{k\neq l}^{N} \mathbf{x}^k_{i(h)}\mathbf{x}^k_{j(h)}$. Consequently, evaluating these entropy terms reduces to characterizing the distribution of a sum of multiple independent n-dimensional log-normal random variables. In the main text, we employed the Fenton–Wilkinson approximation, which approximates this sum by a single log-normal distribution through matching its first two moments.

The purpose of this supplementary material is to analyze the factors that control the error introduced by the Fenton–Wilkinson approximation, and to determine the order at which this error scales with respect to these factors. Since the derivations of $h(\mathbf{w})$ and $h(\mathbf{w}|\mathbf{x}^l)$ are structurally symmetric—in essence, both correspond to the differential entropy of a sum of independent log-normal variables, differing only in that one involves a sum over n terms while the other involves $n-1$ terms—we focus our analysis on the approximation error in $h(\mathbf{w})$. The resulting error in $MI(\mathbf{w};\mathbf{x}^l)$ follows the same scaling order.

In the following, we provide four propositions to demonstrate how the approximation error can be controlled. We focus on the logarithm of the synaptic ensemble, $\log \mathbf{w}$, as the primary variable for our analysis.

First, we prove that this target random variable can be decomposed into a normal random variable and a residual term, the magnitude of which is controlled by a quantity ϵ (Proposition S1). Next, we show that if the magnitude of this residual term is sufficiently small, the resulting deviation in the probability density function is also small (Proposition S2). We then demonstrate that, under specific assumptions for the target variable, a small deviation in the probability density function implies a small deviation in the entropy (Proposition S3). Finally, we prove that if the entropy deviation for the log-transformed variable is small, the error in the quantity of interest, $h(\mathbf{w})$, is similarly controlled (Proposition S4). Following these proofs, we provide an interpretation of the quantity ϵ and discuss how it can be applied in practical contexts.

1 Notations

Before presenting the first proposition, we introduce the notation used throughout the following analysis. We define

$$Q^k_h := \ln \mathbf{x}^k_{i(h)}\mathbf{x}^k_{j(h)}, \quad Q^k := (Q^k_1, Q^k_2, \ldots, Q^k_n)^T. \tag{1}$$

Since Q^k is constructed from different components of the same data pattern, it is an n-dimensional random vector that follows a multivariate normal distribution.

For notational consistency in the derivations below, we use W rather than $\mathbf{w}$ to denote the synaptic ensemble. Accordingly, we write

$$W := \mathbf{w} = \sum_k (e^{Q_1^k}, e^{Q_2^k}, \ldots, e^{Q_n^k})^T = \sum_k e^{Q^k}. \tag{2}$$

We further introduce the notation $T := \ln W$.

Unless stated otherwise (e.g., in the case of matrix multiplication), all operations applied to vector-valued quantities in the following derivations—such as division between vectors—are understood to be performed component-wise.

2 Propositions for error analysis

Proposition S1. *The random variable T can be decomposed into the sum of a normal random variable and a residual term. The residual term is $O(\epsilon)$, where*

$$\epsilon := \frac{\sum_k e^{Q^k - \overline{Q}} - \mathbb{E}[e^{Q^k - \overline{Q}}]}{\sum_k \mathbb{E}[e^{Q^k - \overline{Q}}]},$$

and $\overline{Q} = \frac{1}{N} \sum_k Q^k$.

Proof.

By definition of T, we have

$$T = \ln \sum_k e^{Q^k} = \ln \sum_k e^{\overline{Q} + (Q^k - \overline{Q})} = \ln \left(e^{\overline{Q}} \sum_k e^{Q^k - \overline{Q}} \right) \tag{3}$$

$$= \overline{Q} + \ln \sum_k e^{Q^k - \overline{Q}}, \tag{4}$$

where $\overline{Q} := \frac{1}{N} \sum_k Q^k$ is the average of N independent normal random vectors. Consequently, $\overline{Q}$ itself follows an n-dimensional multivariate normal distribution.

The remaining term can be decomposed as

$$\ln \sum_k e^{Q^k - \overline{Q}} = \ln \left(\mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}} \right] + \sum_k e^{Q^k - \overline{Q}} - \mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}} \right] \right) \tag{5}$$

$$= \ln \left(\mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}} \right] \right) + \ln \left(\mathbf{1} + \frac{\sum_k e^{Q^k - \overline{Q}} - \mathbb{E}[e^{Q^k - \overline{Q}}]}{\sum_k \mathbb{E}[e^{Q^k - \overline{Q}}]} \right), \tag{6}$$

where $\mathbf{1}$ denotes a vector of ones, and we define

$$\epsilon := \frac{\sum_k e^{Q^k - \overline{Q}} - \mathbb{E}[e^{Q^k - \overline{Q}}]}{\sum_k \mathbb{E}[e^{Q^k - \overline{Q}}]}. \tag{7}$$

Combining the above expressions, we obtain

$$T = \overline{Q} + \ln\left(\mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}}\right]\right) + \ln(\mathbf{1} + \epsilon) \tag{8}$$

$$= \overline{Q} + \ln\left(\mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}}\right]\right) + O(\epsilon). \tag{9}$$

Since the first term is Gaussian and the second term is a deterministic vector, all non-Gaussian contributions to T arise from the residual term $O(\epsilon)$. Here, the notation $O(\epsilon)$ is understood in the sense of scaling with respect to the magnitude of the random vector ϵ.

This decomposition suggests that if $\|\epsilon\|_\infty$ is sufficiently small, the deviation of T from a Gaussian random variable is also small. While this condition may be stronger than necessary, a weaker and more practical sufficient condition is that $\|\epsilon\|_\infty < 1$ and $\|\epsilon\|_2 \ll \|\overline{Q}\|_2$. When $\|\epsilon\|_\infty < 1$, the term $\ln(\mathbf{1} + \epsilon)$ is dominated by its first-order expansion, and when $\|\epsilon\|_2 \ll \|\overline{Q}\|_2$, its overall magnitude is negligible compared to the leading Gaussian term. □

Proposition S2. *Let δ be a scalar that quantifies the magnitude of the difference between T and the logarithm of the random variable obtained from the Fenton–Wilkinson approximation, denoted by T_{FW}. Then $p_T(z) = p_{T_{FW}}(z) + O(\delta)$.*

Proof.

Working in logarithmic space, let T_{FW} denote the random variable obtained from the Fenton–Wilkinson approximation. We assume that the true random variable deviates from this approximation by a residual term, so that

$$T = T_{\mathrm{FW}} + \delta F,$$

where δ is a scalar quantifying the magnitude of the residual, and F is a random vector normalized to unit scale (e.g., under the largest principal variance), capturing the shape of the residual distribution. We emphasize that F is not assumed to be independent of T_{FW}.

To this end, we study the relationship between the distributions p_T and $p_{T_{\mathrm{FW}}}$. Since T depends jointly on T_{FW} and F, its density can be expressed by marginalizing over F:

$$p_T(z) = \int_{\mathbb{R}^n} p_{T_{\mathrm{FW}},F}(z - \delta y, y)\, dy \tag{10}$$

$$= \int p_{T_{\mathrm{FW}}|F}(z - \delta y \mid y)\, p_F(y)\, dy. \tag{11}$$

For notational convenience, define $f(x_1, x_2) := p_{T_{\mathrm{FW}}|F}(x_1|x_2)$, where the first argument corresponds to the value of T_{FW} and the second to the value of F. A Taylor

expansion with respect to the first argument yields

$$f(z-\delta y, y) = f(z,y) - \delta y^T \nabla_z f(z,y) + O(\delta^2). \tag{12}$$

Substituting this expansion back into the expression for p_T, we obtain

$$\begin{aligned}
p_T(z) &= \int \left(p_{T_{\text{FW}}|F}(z|y) - \delta y^T \nabla_z p_{T_{\text{FW}}|F}(z|y) + O(\delta^2)\right) p_F(y) dy && (13)\\
&= p_{T_{\text{FW}}}(z) - \delta \int y^T \nabla_z p_{T_{\text{FW}},F}(z,y) dy + O(\delta^2) && (14)\\
&= p_{T_{\text{FW}}}(z) - \delta \int \sum_i y_i \, \partial_{z_i} p_{T_{\text{FW}},F}(z,y) dy + O(\delta^2) && (15)\\
&= p_{T_{\text{FW}}}(z) - \delta \sum_i \partial_{z_i} \int y_i \, p_{T_{\text{FW}},F}(z,y) dy + O(\delta^2) && (16)\\
&= p_{T_{\text{FW}}}(z) - \delta \nabla_z \cdot \int y \, p_{T_{\text{FW}},F}(z,y) dy + O(\delta^2) && (17)\\
&= p_{T_{\text{FW}}}(z) - \delta \nabla_z \cdot \int y \, p_{F|T_{\text{FW}}}(y|z) p_{T_{\text{FW}}}(z) dy + O(\delta^2) && (18)\\
&= p_{T_{\text{FW}}}(z) - \delta \nabla_z \cdot (\mathbb{E}[F|T_{\text{FW}} = z] p_{T_{\text{FW}}}(z)) + O(\delta^2) && (19)\\
&= p_{T_{\text{FW}}}(z) + O(\delta). && (20)
\end{aligned}$$

□

Proposition S3. *Assume p_T is a bounded probability density function with a finite $(2+\beta)^{\text{th}}$ moment for some $\beta > 0$. If $p_T(z) = p_{T_{FW}}(z) + \delta r(z)$ where $r(z) \in L^1(\mathbb{R}^n)$, then as $\delta \to 0$, $h(T_{\text{FW}}) \to h(T)$.*

Proof.

The difference between the two entropies can be rewritten as the following two integrals, for which we will prove its convergence separately.

$$h(T_{\text{FW}}) - h(T) = \int_{\mathbb{R}^n} (p_T - p_{T_{\text{FW}}}) \log p_{T_{\text{FW}}} dz + \int_{\mathbb{R}^n} p_T \log \frac{p_T}{p_{T_{\text{FW}}}} dz \tag{21}$$

For the first term, $\int_{\mathbb{R}^n} (p_T - p_{T_{\text{FW}}}) \log p_{T_{\text{FW}}} dz$, since $p_{T_{\text{FW}}}$ is multivariate normal with finite variances, we can find a quadratic bound for $|\log p_{T_{\text{FW}}}|$:

$$|\log p_{T_{\text{FW}}}(z)| \leq c_1 + c_2 \|z\|^2$$

Therefore, the size of the first term can be controlled via:

$$|\int_{\mathbb{R}^n} \delta r(z) \log p_{T_{\text{FW}}} dz| \leq \int_{\mathbb{R}^n} |\delta r(z) \log p_{T_{\text{FW}}}| dz \leq \delta \int_{\mathbb{R}^n} |r(z)|(c_1 + c_2\|z\|^2) dz \tag{22}$$

Since $r(z) \in L^1$, $\delta c_1 \int |r(z)| dz$ converges to zero as $\delta \to 0$. As for the remaining

component, $\delta c_2 \int |r(z)|\|z\|^2 dz$, by applying Hölder's inequality, we obtain:

$$c_2 \int_{\mathbb{R}^n} |\delta r(z)|\|z\|^2 dz = c_2 \int_{\mathbb{R}^n} \left(|\delta r(z)|^{\frac{\beta}{2+\beta}}\right) \left(|\delta r(z)|^{\frac{2}{2+\beta}} \|z\|^2\right) dz \tag{23}$$

$$\leq c_2 \left(\int_{\mathbb{R}^n} |\delta r(z)| dz\right)^{\frac{\beta}{2+\beta}} \left(\int_{\mathbb{R}^n} |\delta r(z)|\|z\|^{2+\beta}\right)^{\frac{2}{2+\beta}} \tag{24}$$

$$\leq c_2 \left(\delta \int_{\mathbb{R}^n} |r(z)| dz\right)^{\frac{\beta}{2+\beta}} \left(\mathbb{E}[\|T\|^{2+\beta}] + \mathbb{E}[\|T_{\text{FW}}\|^{2+\beta}]\right)^{\frac{2}{2+\beta}} \tag{25}$$

The first factor is the L^1 norm of $\delta r(z)$ raised to a positive power. Since $r(z) \in L^1$, this norm approaches zero. The second factor is controlled by the $(2+\beta)^{\text{th}}$ moments of the distributions. Given our assumptions on the ground truth T and the multivariate normal nature of T_{FW}, these moments are uniformly bounded. Therefore, this remaining component also converges to zero as $\delta \to 0$.

Consequently, the whole first term vanishes with $\delta \to 0$.

For the second term, $\int_{\mathbb{R}^n} p_T \log \frac{p_T}{p_{T_{\text{FW}}}} dz$, in order to prove its convergence to zero, *the first step* is to prove the uniform integrability of the function $p_T \log \frac{p_T}{p_{T_{\text{FW}}}}$.

To prove the uniform integrability, we investigate both the local and tail behavior of this function.

Let's start with the local behavior where the Euclidean distance of z from the origin is bounded by M. For any measurable set within this zone, i.e. $E \subset \{z : \|z\| \leq M\}$, the integral over E is controlled by:

$$\int_E |p_T \log p_T - p_T \log p_{T_{\text{FW}}}| dz \leq \int_E |p_T \log p_T| dz + \int_E |p_T \log p_{T_{\text{FW}}}| dz \tag{26}$$

For $|p_T \log p_T|$, the function is bounded by its maximum value on $\{z : \|z\| \leq M\}$. Therefore, $\int_E |p_T \log p_T| dz$ is bounded by the product of this maximum value and the Lebesgue measure of E. Consequently, as long as the measure of E is sufficiently small, $\int_E |p_T \log p_T| dz$ can be made arbitrarily small.

For $|p_T \log p_{T_{\text{FW}}}|$, the Gaussian nature of $p_{T_{\text{FW}}}$ ensures that the factor $|\log p_{T_{\text{FW}}}| \leq c_1 + c_2 M^2$ within this zone. Therefore, the integral $\int_E |p_T \log p_{T_{\text{FW}}}| dz$ is bounded by $(c_1 + c_2 M^2) \int_E p_T dz$. Since p_T is a probability density function, this value also becomes arbitrarily small as the measure of E decreases.

Thus, for any $\epsilon' > 0$, we can find a $\delta' > 0$ such that for all subsets $E \subset \{z : \|z\| \leq M\}$ with measure smaller than δ', the integral $\int_E |p_T \log \frac{p_T}{p_{T_{\text{FW}}}}| dz < \epsilon'$.

Let's now look at the tail behavior of the function $p_T \log \frac{p_T}{p_{T_{\text{FW}}}}$ on $\{z : \|z\| > M\}$.

Its integration over this whole zone is controlled by:

$$\int_{\|z\|>M} |p_T \log p_T - p_T \log p_{T_{\text{FW}}}| dz \leq \int_{\|z\|>M} |p_T \log p_T| dz + \int_{\|z\|>M} |p_T \log p_{T_{\text{FW}}}| dz \tag{27}$$

For $|p_T \log p_T|$, to study its integration over the tail zone, let's start by investigating the integral over the entire domain, $\int_{\mathbb{R}^n} |p_T \log p_T| dz$. This can be divided into two zones: $\{z : p_T(z) \geq 1\}$ and $\{z : 0 \leq p_T(z) < 1\}$. The boundedness of p_T ensures integrability on the set $\{z : p_T(z) \geq 1\}$, as it is bounded by the product $\log(\max(p_T)) \int_{p_T \geq 1} p_T dz$, and the $\int_{p_T \geq 1} p_T dz$ is finite because p_T is a probability density function.

The finite moment $\mathbb{E}[\|T\|^{2+\beta}] < \infty$ ensures the density decays faster than $\|z\|^{-(n+2+\beta)}$, which is sufficient to guarantee integrability over the set $\{z : 0 \leq p_T(z) < 1\}$. This is because, due to the sub-polynomial growth of the logarithm, we have the inequality $|p_T \log p_T| \leq C_{\delta'} p_T^{1-\delta'}$ for any $\delta' \in (0, 1)$. As a result, after splitting $\int_{p_T<1} |p_T \log p_T| dz$ into integrals over $\|z\| < 1$ and $\|z\| \geq 1$, we have $\int_{p_T<1} |p_T \log p_T| dz \leq \int_{p_T<1, \|z\|<1} |p_T \log p_T| dz + \int_{p_T<1, \|z\|\geq 1} \frac{C'}{\|z\|^{(1-\delta')(n+2+\beta)}} dz$. The former is finite because the integrand is locally bounded on a compact domain. The latter term, in n-dimensional spherical coordinates, scales as $\int_1^\infty \frac{r^{n-1}}{r^{(1-\delta)(n+2+\beta)}} dr = \int_1^\infty \frac{1}{r^{3+\beta-\delta'(n+2+\beta)}} dr$. Since δ' can be chosen arbitrarily small, the exponent of radius can be made greater than 1, leading to a finite integral.

Therefore, $\int_{\mathbb{R}^n} |p_T \log p_T| dz$ is finite. Consequently, the tail integral $\int_{\|z\|>M} |p_T \log p_T| dz$ vanishes as $M \to \infty$.

For $|p_T \log p_{T_{\text{FW}}}|$, we have $\int_{\mathbb{R}^n} |p_T \log p_{T_{\text{FW}}}| dz \leq \int_{\mathbb{R}^n} p_T(c_1 + c_2\|z\|^2) dz$. Since p_T is a probability density function with finite second moments, the right-hand side is finite. Therefore, $\int_{\mathbb{R}^n} |p_T \log p_{T_{\text{FW}}}| dz$ is finite. Consequently, the tail integral $\int_{\|z\|>M} |p_T \log p_{T_{\text{FW}}}| dz$ vanishes as $M \to \infty$.

Therefore, for any arbitrary positive threshold, we can always find a sufficiently large M such that the integral $\int_{\|z\|>M} |p_T \log \frac{p_T}{p_{T_{\text{FW}}}}| dz$ is smaller than that threshold.

Combining the results from both the local and tail behavior, we establish the uniform integrability as follows:

For any arbitrary $\epsilon' > 0$, we first select a radius M large enough such that $\int_{\|z\|>M} |p_T \log \frac{p_T}{p_{T_{\text{FW}}}}| dz \leq \frac{\epsilon'}{2}$. Within the local zone $\|z\| \leq M$, we can find a $\delta' > 0$ such that any set E with measure smaller than δ' satisfies $\int_{E \cap \{\|z\| \leq M\}} |p_T \log \frac{p_T}{p_{T_{\text{FW}}}}| dz \leq \frac{\epsilon'}{2}$.

It follows that for any set E in $\mathbb{R}^n$ with measure smaller than δ', the total integral $\int_E |p_T \log \frac{p_T}{p_{T_{\text{FW}}}}| dz$ is bounded by the sum of the contributions from the tail and the local zone, totaling no more than ϵ'. This confirms that the function $p_T \log \frac{p_T}{p_{T_{\text{FW}}}}$ is uniformly integrable.

Given the uniform integrability, *the final step* is to prove the convergence of the second term, $\int_{\mathbb{R}^n} p_T \log \frac{p_T}{p_{T_{\text{FW}}}} dz$. As $\delta \to 0$, $p_T \log \frac{p_T}{p_{T_{\text{FW}}}} \to 0$ in a pointwise manner. Therefore, by Vitali's Convergence Theorem, the uniform integrability implies the con-

vergence of the integral:

$$\lim_{\delta\to 0}\int_{\mathbb{R}^n} p_T \log\frac{p_T}{p_{T_{\text{FW}}}}dz = \int_{\mathbb{R}^n}\lim_{\delta\to 0}\left(p_T\log\frac{p_T}{p_{T_{\text{FW}}}}\right)dz = \int_{\mathbb{R}^n} 0dz = 0 \tag{28}$$

Taken together, these results demonstrated that the difference in differential entropies vanishes in the limit, i.e., $\lim_{\delta\to 0} h(T_{\text{FW}}) - h(T) = 0$. □

Proposition S4. *As $\delta \to 0$, $h(W_{\text{FW}}) \to h(W)$, where $W_{\text{FW}} := e^{T_{\text{FW}}}$ denotes the approximated weight ensemble.*

Proof.

By the transformation property of differential entropy, we have

$$h(W) = h(e^T) = h(T) + \sum_i \mathbb{E}[T_i]. \tag{29}$$

Consequently, the error in entropy estimation can be written as

$$\begin{aligned} h(W) - h(W_{\text{FW}}) &= h(e^T) - h(e^{T_{\text{FW}}}) && (30)\\ &= h(T) + \sum_i \mathbb{E}[T_{\text{FW},i} + \delta F_i] - h(T_{\text{FW}}) - \sum_i \mathbb{E}[T_{\text{FW},i}] && (31)\\ &= h(T) - h(T_{\text{FW}}) + \sum_i \mathbb{E}[\delta F_i], && (32) \end{aligned}$$

As $\delta \to 0$, the latter term $\sum_i \mathbb{E}[\delta F_i] \to 0$. By proposition S3, $h(T_{\text{FW}}) - h(T) \to 0$. As a result, $h(W) - h(W_{\text{FW}}) \to 0$. □

Following the proposed propositions, we conclude that when ϵ (defined in equation 7) is sufficiently small, T admits only a small additive deviation from a multivariate normal random variable. Moreover, this deviation induces only a small error in the estimated probability density function $p_{T_{\text{FW}}}$. Under the assumption of certain regularity properties for T, this translates to a small error in the entropy estimates of $h(\mathbf{w})$ and $h(\mathbf{w}_{/l})$, ultimately resulting in a correspondingly small error in the mutual information $MI(\mathbf{w}; \mathbf{x}^l)$.

3 Practical interpretation of error control

The result above indicates that controlling $\|\epsilon\|_2$ is sufficient for controlling the magnitude of the deviation from a multivariate log-normal random variable. However, ϵ itself is not directly interpretable or convenient for practical analysis. To obtain quantities that are both interpretable and practically useful, we use a sequence of intuitive arguments to motivate a set of sufficient conditions leading to a tractable proxy for controlling ϵ.

Recall that ϵ is defined as

$$\epsilon = \frac{\left(\sum_k e^{Q^k - \overline{Q}}\right) - \mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}}\right]}{\mathbb{E}\left[\sum_k e^{Q^k - \overline{Q}}\right]}. \tag{33}$$

This expression corresponds to an element-wise division between two n-dimensional vectors. Since both the numerator and the denominator are sums over N independent data patterns, the L_2 norm of this ratio can be controlled by controlling the L_2 norm of the corresponding ratio at the level of individual patterns. In particular, it is sufficient that

$$\left\| \frac{e^{Q^k - \overline{Q}} - \mathbb{E}[e^{Q^k - \overline{Q}}]}{\mathbb{E}[e^{Q^k - \overline{Q}}]} \right\|_2 \tag{34}$$

is small for every pattern k.

This condition implies that the log-normal random vector $e^{Q^k - \overline{Q}}$ exhibits only small fluctuations around its mean. Such behavior is ensured if its logarithm, $Q^k - \overline{Q}$, itself has small fluctuations.

By definition, both Q^k and $\overline{Q}$ are linear combinations of the logarithms of the original data patterns. Consequently, a sufficient condition for controlling the fluctuations of $Q^k - \overline{Q}$ is that $\ln \mathbf{x}^k$ exhibits small fluctuations for all data patterns $\mathbf{x}^k$.

Since $\ln \mathbf{x}^k$ is a d-dimensional multivariate normal random variable, controlling its fluctuation can be achieved by controlling the variance along its principal directions—equivalently, by controlling the largest eigenvalue of its covariance matrix, i.e. $\text{Var}_{\text{principal}}[\ln \mathbf{x}^k]$.

4 Heuristic analysis of error scaling across parameter regimes

Based on the previous analysis, we have shown that the error introduced by the Fenton-Wilkinson approximation can be controlled by limiting the scale of the random variable ϵ. Although the exact functional dependency between the error magnitude and ϵ is unknown, we provide here a heuristic derivation of the dependency between the error and other model parameters to give the reader a better practical understanding. Specifically, by assuming the error magnitude scales linearly with the scale of ϵ, we explain how the error scales with respect to the number of neurons (d), the size of the weight ensemble (n), the number of data patterns (N), and the correlation structure of the data.

For d (number of neurons): Since ϵ is defined by a ratio of n-dimensional vectors, where each component depends on the individual input activity of a neuron, the scale of ϵ is effectively independent of d.

For n (size of the weight ensemble): Because ϵ is an n-dimensional vector, we expect its overall magnitude to scale as $\sqrt{n}$, consistent with the scaling of the Euclidean norm for high-dimensional vectors.

For N (number of data patterns): Since both the numerator and denominator of ϵ

are sums over N independent patterns, the magnitude of ϵ is primarily governed by the extremal ratios within the set. Consequently, the scale of ϵ does not exhibit an explicit, direct dependence on the value of N.

For the correlation structure: As the components of $\ln \mathbf{x}^k$ become more correlated, the components of the deviation $Q^k - \overline{Q}$ likewise show increased correlation. This results in an ϵ with a more structured covariance, leading to a larger variance along the principal direction and thus a larger overall scale.