

Stochastic Thermodynamics of Learning

Sebastian Goldt* and Udo Seifert

II. Institut für Theoretische Physik, Universität Stuttgart, 70550 Stuttgart, Germany

(Dated: November 30, 2016)

Virtually every organism gathers information about its noisy environment and builds models from that data, mostly using neural networks. Here, we use stochastic thermodynamics to analyse the learning of a classification rule by a neural network. We show that the information acquired by the network is bounded by the thermodynamic cost of learning and introduce a learning efficiency $\eta \leq 1$. We discuss the conditions for optimal learning and analyse Hebbian learning in the thermodynamic limit.

PACS numbers: 05.70.Ln, 05.40.-a, 84.35.+i, 87.19.lv

Introduction. – Information processing is ubiquitous in biological systems, from single cells measuring external concentration gradients to large neural networks performing complex motor control tasks. These systems are surprisingly robust, despite the fact that they are operating in noisy environments [1, 2], and they are efficient: *E. coli*, a bacterium, is near-perfect from a thermodynamic perspective in exploiting a given energy budget to adapt to its environment [3]. Thus it is important to keep energetic considerations in mind for the analysis of computations in living systems. Stochastic thermodynamics [4, 5] has emerged as an integrated framework to study the interplay of information processing and dissipation in interacting, fluctuating systems far from equilibrium. Encouraged by a number of intriguing results from its application to bacterial sensing [6–15] and biomolecular processes [16–20], here we consider a new problem: learning.

Learning is about extracting models from sensory data. In living systems, it is implemented in neural networks where vast numbers of neurons communicate with each other via action potentials, the electric pulse used universally as the basic token of communication in neural systems [21]. Action potentials are transmitted via synapses, and their strength determines whether an incoming signal will make the receiving neuron trigger an action potential of its own. Physiologically, the adaptation of these synaptic strengths is a main mechanism for memory formation.

Learning task and model. – A classic example for neurons performing associative learning are the Purkinje cells in the cerebellum [22, 23]. We model such a neuron as a single-layer neural network or perceptron [24, 25], well known from machine learning and statistical physics [26]. The neuron makes N connections to other neurons and is fully characterized by the weights or synaptic strengths $\omega \in \mathbb{R}^N$ of these connections, see figure 1. The neuron must learn whether it should fire an action potential or not for a set of P fixed input patterns or samples $\xi^\mu = (\xi_1^\mu, \dots, \xi_N^\mu)$, $\mu = 1, 2, \dots, P$. Each

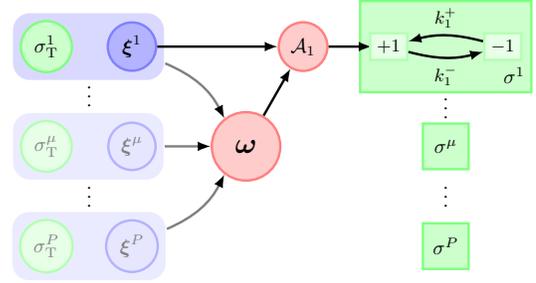


FIG. 1. **Model of a single neuron.** Given a set of inputs $\xi^\mu \in \{\pm 1\}^N$ and their true labels $\sigma_T^\mu = \pm 1$ (left), the neuron learns the mappings $\xi^\mu \rightarrow \sigma_T^\mu$ by adjusting its weights $\omega \in \mathbb{R}^N$. It processes an input by computing the activation $\mathcal{A}^\mu = \omega \cdot \xi^\mu / \sqrt{N}$ which determines the transition rates of a two-state random process $\sigma^\mu = \pm 1$ indicating the label predicted by the neuron for each sample, shown here for $\mu = 1$.

pattern describes the activity of all the other connected neurons at a point in time: if the n -th connected neuron is firing an action potential in the pattern ξ^μ , then $\xi_n^\mu = 1$. For symmetry reasons, we set $\xi_n^\mu = -1$ in case the n -th neuron is silent in the μ -th pattern. Every sample ξ^μ has a fixed true label $\sigma_T^\mu = \pm 1$, indicating whether an action potential should be fired in response to that input or not. These labels are independent of each other and equiprobable; once chosen, they remain fixed.

We model the label predicted by a neuron for each input ξ^μ with a stochastic process $\sigma^\mu = \pm 1$ (right panel in figure 1). Assuming a thermal environment at fixed temperature T , the transition rates k_μ^\pm for these processes obey the detailed balance condition

$$k_\mu^+ / k_\mu^- = \exp(\mathcal{A}^\mu / k_B T) \quad (1)$$

where k_B is Boltzmann's constant and \mathcal{A}^μ is the input-dependent activation

$$\mathcal{A}^\mu \equiv \frac{1}{\sqrt{N}} \omega \cdot \xi^\mu \quad (2)$$

where the prefactor ensures the conventional normalisation. We interpret $p(\sigma^\mu = 1 | \omega)$ with fixed ξ^μ as the probability that the μ -th input would trigger an action

* goldt@theo2.physik.uni-stuttgart.de

potential by the neuron. The goal of learning is to adjust the weights of the network ω such that the predicted labels at any one time $\sigma = (\sigma^1, \dots, \sigma^P)$ equal the true labels $\sigma_T = (\sigma_T^1, \dots, \sigma_T^P)$ for as many inputs as possible.

Let us introduce the concept of learning efficiency by considering a network with a single weight learning one sample $\xi = \pm 1$ with label σ_T , *i.e.* $N = P = 1$. Here and throughout this letter, we set $k_B = T = 1$ to render energy and entropy dimensionless. The weight $\omega(t)$ obeys an overdamped Langevin equation [27]

$$\dot{\omega}(t) = -\omega(t) + f(\omega(t), \xi, \sigma_T, t) + \zeta(t). \quad (3)$$

The total force on the weight arises from a harmonic potential $V(\omega) = \omega^2/2$, restricting the size of the weight [28], and an external force $f(\cdot)$ introducing correlations between weight and input. The exact form of this “learning force” $f(\cdot)$ depends on the learning algorithm we choose. The thermal noise $\zeta(t)$ is Gaussian with correlations $\langle \zeta(t)\zeta(t') \rangle = 2\delta(t-t')$. Here and throughout, we use angled brackets to indicate averages over noise realisations, unless stated otherwise. We assume that initially at $t_0 = 0$, the weight is in thermal equilibrium, $p(\omega) \propto \exp(-\omega^2/2)$, and the labels are equiprobable, $p(\sigma_T) = p(\sigma) = 1/2$. Choosing symmetric rates,

$$k^\pm = \gamma \exp(\pm \mathcal{A}/2), \quad (4)$$

the master equation [27] for the probability distribution $p(\sigma_T, \omega, \sigma, t)$ with given ξ reads

$$\partial_t p(\sigma_T, \omega, \sigma, t) = -\partial_\omega j_\omega(t) + j_\sigma(t), \quad (5)$$

where $\partial_t \equiv \partial/\partial t$ etc. and

$$j_\omega(t) = [-\omega + f(\omega, \xi, \sigma_T, t) - \partial_\omega] p(\sigma_T, \omega, \sigma, t), \quad (6a)$$

$$j_\sigma(t) = k^\sigma p(\sigma_T, \omega, -\sigma, t) - k^{-\sigma} p(\sigma_T, \omega, \sigma, t) \quad (6b)$$

are the probability currents for the weight and the predicted label, respectively. In splitting the total probability current for the system $(\sigma_T, \omega, \sigma)$ into the currents (6), we have used the bipartite property of the system, *i.e.* that the thermal noise in each subsystem (ω and σ), is independent of the other [29, 30]. We choose $\gamma \gg 1$, *i.e.* introduce a time-scale separation between the weights and the predicted labels, since a neuron processes a single input much faster than it learns.

Efficiency of learning. – The starting point to consider both the information-processing capabilities of the neuron and its non-equilibrium thermodynamics is the Shannon entropy of a random variable X with probability distribution $p(x)$,

$$S(X) \equiv - \sum_{x \in X} p(x) \ln p(x), \quad (7)$$

which is a measure of the uncertainty of X [31]. This definition carries over to continuous random variables, where the sum is replaced by an integral. For dependent

random variables X and Y , the conditional entropy of X given Y is given by $S(X|Y) \equiv - \sum_{x,y} p(x,y) \ln p(x|y)$ where $p(x|y) = p(x,y)/p(y)$. The natural quantity to measure the information learnt is the mutual information

$$I(\sigma_T : \sigma) \equiv S(\sigma_T) - S(\sigma_T|\sigma) \quad (8)$$

which measures by how much, on average, the uncertainty about σ_T is reduced by knowing σ [31]. To discuss the efficiency of learning, we need to relate this information to the thermodynamic costs of adjusting the weight during learning from $t_0 = 0$ up to a time t , which are given by the well-known total entropy production [4] of the weight,

$$\Delta S_\omega^{\text{tot}} \equiv \Delta S(\omega) + \Delta Q. \quad (9)$$

Here, ΔQ is the heat dissipated into the medium by the dynamics of the weight and $\Delta S(\omega)$ is the difference in Shannon entropy (7) of the marginalized distribution $p(\omega, t) = \sum_{\sigma_T, \sigma} p(\sigma_T, \omega, \sigma, t)$ at times t_0 and t , respectively. We will show that in feedforward neural networks with Markovian dynamics (5, 6), the information learnt is bounded by the thermodynamic costs of learning,

$$I(\sigma_T : \sigma) \leq \Delta S(\omega) + \Delta Q \quad (10)$$

for arbitrary learning algorithm $f(\omega, \xi, \sigma_T, t)$ at all times $t > t_0$. This inequality is our first result. We emphasise that while relations between changes in mutual information and total entropy production have appeared in the literature [29, 30, 32–34], they usually concern a single degree of freedom, say X , in contact with some other degree(s) of freedom Y , and relate the change in mutual information $I(X : Y)$ due to the dynamics of X to the total entropy production of X . Instead, our relation connects the entropy production in the weights with the total change in mutual information between σ_T and σ , which is key for neural networks. Our derivation [35] builds on recent work by Horowitz [30] and can be generalized to N dimensions and P samples, see eq. (16) below. Equation (10) suggests to introduce an efficiency of learning

$$\eta \equiv \frac{I(\sigma_T : \sigma)}{\Delta S(\omega) + \Delta Q} \leq 1. \quad (11)$$

Toy model. – As a first example, let us calculate the efficiency of Hebbian learning, a form of coincidence learning well known from biology [21, 36], for $N = P = 1$ in the limit $t \rightarrow \infty$. If the neuron should fire an action potential when its input neuron fires, or if they should both stay silent, *i.e.* $\xi = \sigma_T = \pm 1$, the weight of their connection increases – “fire together, wire together”. For symmetry reasons, the weight decreases if the input neuron is silent but the neuron should fire and vice versa, $\xi = -\sigma_T$. This rule yields a final weight proportional to $\mathcal{F} \equiv \sigma_T \xi$, so to minimise dissipation [37], we choose a learning force f linearly increasing with time,

$$f(\omega, \xi, \sigma_T, t) \equiv \begin{cases} \nu \mathcal{F} t / \tau & t \leq \tau \\ \nu \mathcal{F} & t > \tau, \end{cases} \quad (12)$$

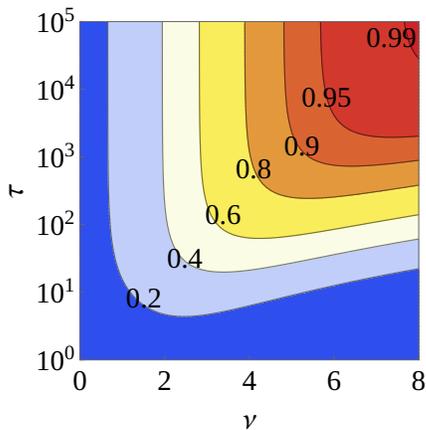


FIG. 2. **Learning efficiency of a neuron with a single weight.** We plot the efficiency η (11) for a neuron with a single weight learning a single sample as a function of the learning rate ν and learning duration τ in the limit $t \rightarrow \infty$.

where we have introduced the learning duration $\tau > 0$ and the factor $\nu > 0$ is conventionally referred to as the learning rate in the machine learning literature [24]. The total entropy production (9) can be computed from the distribution $p(\sigma_T, \omega, t)$, which is obtained by first integrating σ out of equations (5, 6) and solving the resulting Fokker-Planck equation [38]. The total heat dissipated into the medium ΔQ is given by [4]

$$\Delta Q = \int_0^\infty dt \int_{-\infty}^\infty d\omega j_\omega(t) [-\omega(t) + f(\omega(t), \xi, \sigma_T, t)] = \frac{\nu^2 \mathcal{F}^2 (e^{-\tau} + \tau - 1)}{\tau^2}. \quad (13)$$

As expected, no heat is dissipated in the limit of infinitely slow driving, $\lim_{\tau \rightarrow \infty} \Delta Q = 0$, while for a sudden potential switch $\tau \rightarrow 0$, $\lim_{\tau \rightarrow 0} \Delta Q = \nu^2 \mathcal{F}^2 / 2$. The change in Shannon entropy $\Delta S(\omega)$ is computed from the marginalized distribution $p(\omega, t) = \sum_{\sigma_T} p(\sigma_T, \omega, t)$. Finally, the mutual information (8) can be computed from the stationary solution of (5).

A plot of the efficiency (11), fig. 2, highlights the two competing requirements for maximizing η . First, all the information from the true label $S(\sigma_T) = \ln 2$ needs to be stored in the weight by increasing the learning rate ν , which leads to $\Delta S(\omega) \rightarrow \ln 2$ and a strongly biased distribution $p(\sigma|\omega)$ such that $I(\sigma_T : \sigma) \rightarrow \ln 2$. Second, we need to minimise the dissipated heat ΔQ , which increases with ν , by driving the weight slowly, $\tau \gg 1$.

More samples, higher dimensions. – Moving on to a neuron with N weights ω learning P samples with true labels $\sigma_T \equiv (\sigma_T^1, \dots, \sigma_T^\mu, \dots, \sigma_T^P)$, we have a Langevin equation for each weight ω_n with independent thermal noise sources $\zeta_n(t)$ such that $\langle \zeta_n(t) \zeta_m(t') \rangle = 2\delta_{nm} \delta(t-t')$ for $n, m = 1, \dots, N$. Two learning scenarios are possible: *batch learning*, where the learning force is a function of

all samples and their labels,

$$\dot{\omega}_n(t) = -\omega_n(t) + f(\omega_n(t), \{\xi_n^\mu, \sigma_T^\mu\}, t) + \zeta_n(t). \quad (14)$$

A more realistic scenario from a biological perspective is *online learning*, where the learning force is a function of only one sample and its label at a time,

$$\dot{\omega}_n(t) = -\omega_n(t) + f(\omega_n(t), \xi_n^{\mu(t)}, \sigma_T^{\mu(t)}, t) + \zeta_n(t). \quad (15)$$

The sample and label which enter this force are given by $\mu(t) \in \{1, \dots, P\}$, which might be a deterministic function or a random process. Either way, the weights ω determine the transition rates of the P independent two-state processes for the predicted labels $\sigma \equiv (\sigma^1, \dots, \sigma^\mu, \dots, \sigma^P)$ via (1) and (2). Again, we assume that the thermal noise in each subsystem, ω_n or σ^μ , is independent of all the others, and choose initial conditions at $t_0 = 0$ to be $p(\omega) \propto \exp(-\omega \cdot \omega / 2)$ and $p(\sigma_T^\mu) = p(\sigma^\mu) = 1/2$. The natural quantity to measure the amount of learning after a time t in both scenarios is the sum of $I(\sigma_T^\mu : \sigma^\mu)$ over all inputs. We can show [35] that this information is bounded by the total entropy production of all the weights,

$$\sum_{\mu=1}^P I(\sigma_T^\mu : \sigma^\mu) \leq \sum_{n=1}^N [\Delta S(\omega_n) + \Delta Q_n] = \sum_{n=1}^N \Delta S_n^{\text{tot}} \quad (16)$$

where ΔQ_n is the heat dissipated into the medium by the n -th weight and $\Delta S(\omega_n)$ is the change from t_0 to t in Shannon entropy (7) of the marginalized distribution $p(\omega_n, t)$. This is our main result.

Let us now compute the efficiency of online Hebbian learning in the limit $t \rightarrow \infty$. Since a typical neuron will connect to ~ 1000 other neurons [21], we take the thermodynamic limit by letting the number of samples P and the number of dimensions N both go to infinity while simultaneously keeping the ratio

$$\alpha \equiv P/N \quad (17)$$

on the order of one. The samples ξ^μ are drawn at random from $p(\xi_n^\mu = 1) = p(\xi_n^\mu = -1) = 1/2$ and remain fixed [39]. We choose a learning force on the n -th weight of the form (12) with $\mathcal{F} \rightarrow \mathcal{F}_n$ and assume that the process $\mu(t)$ is a random walk over the integers $1, \dots, P$ changing on a timescale much shorter than the relaxation time of the weights. Since f^2 is finite, the learning force is effectively constant with

$$\mathcal{F}_n = \frac{1}{\sqrt{N}} \sum_{\mu=1}^P \xi_n^\mu \sigma_T^\mu, \quad (18)$$

where the prefactor ensures the conventional normalisation [24]. Hence all the weights ω_n are independent of each other and statistically equivalent. Averaging first over the noise with fixed σ_T , we find that ω_n is normally distributed with mean $\langle \omega_n \rangle = \nu \mathcal{F}_n$ and variance

1 [40]. The average with respect to the quenched disorder σ_T , which we shall indicate by an overline, is taken second by noting that \mathcal{F}_n is normally distributed by the central limit theorem with $\overline{\mathcal{F}_n} = 0$ and $\overline{\mathcal{F}_n^2} = \alpha$, hence $\overline{\langle \omega_n \rangle} = 0$ and $\overline{\langle \omega_n^2 \rangle} = 1 + \alpha\nu^2$. The change in Shannon entropy of the marginalized distribution $p(\omega_n)$ is hence $\Delta S(\omega_n) = \ln(1 + \alpha\nu^2)$. Likewise, the heat dissipated by the n -th weight ΔQ_n is obtained by averaging eq. (13) over $\mathcal{F} \rightarrow \mathcal{F}_n$.

The mutual information $I(\sigma_T^\mu : \sigma^\mu)$ is a functional of the marginalized distribution $p(\sigma_T^\mu, \sigma^\mu)$ which can be obtained by direct integration of $p(\sigma_T, \omega, \sigma)$ [35]. Here we will take a simpler route starting from the *stability* of the μ -th sample [41]

$$\Delta^\mu \equiv \frac{1}{\sqrt{N}} \omega \cdot \xi^\mu \sigma_T^\mu = \mathcal{A}^\mu \sigma_T^\mu. \quad (19)$$

Its role can be appreciated by considering the limit $T \rightarrow 0$, where it is easily verified using the detailed balance condition (1) that the neuron predicts the correct label if and only if $\Delta^\mu > 0$. For $T = 1$, the neuron predicts the μ -th label correctly with probability

$$p_C^\mu \equiv p(\sigma^\mu = \sigma_T^\mu) = \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \frac{e^{\Delta^\mu}}{e^{\Delta^\mu} + 1} \quad (20)$$

where $p(\Delta^\mu)$ is the distribution generated by thermal noise and quenched disorder, yielding a Gaussian with mean ν and variance $1 + \alpha\nu^2$ [35]. The mutual information follows as

$$I(\sigma_T^\mu : \sigma^\mu) = \ln 2 - S(p_C^\mu) \quad (21)$$

with the shorthand for the entropy of a binary random variable $S(p) = -p \ln p - (1-p) \ln(1-p)$ [31]. It is plotted in fig. 3 together with the mutual information obtained by Monte Carlo integration of $p(\sigma_T, \omega, \sigma)$ with $N = 10000$. For a vanishing learning rate $\nu \rightarrow 0$ or infinitely many samples $\alpha \rightarrow \infty$, $p_C^\mu \rightarrow 1/2$ and hence $I(\sigma_T^\mu : \sigma^\mu) \rightarrow 0$. The maximum value $I(\sigma_T^\mu : \sigma^\mu) = \ln 2$ is only reached for small α and decreases rapidly with increasing α , even for values of α where it is possible to construct a weight vector that classifies all the samples correctly [25]. This is a consequence of both the thermal noise in the system and the well-known failure of Hebbian learning to use the information in the samples perfectly [24]. We note that while the integral in eq. (20) has to be evaluated numerically, p_C^μ can be closely approximated analytically by $p(\Delta^\mu > 0)$ with the replacement $\nu \rightarrow \nu/2$ [35] (dashed lines in fig. 3).

Together, these results allow us to define the efficiency $\tilde{\eta}$ of Hebbian learning as a function of just α and ν ,

$$\tilde{\eta} \equiv \alpha \frac{I(\sigma_T^\mu : \sigma^\mu)}{\Delta S(\omega_n) + \Delta Q_n}, \quad (22)$$

where we have taken the mutual information per sample and the total entropy production per weight, multiplied

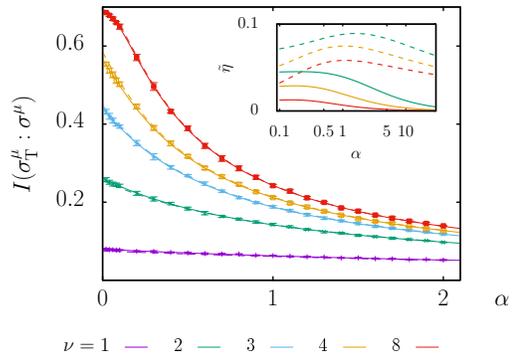


FIG. 3. **Hebbian learning in the thermodynamic limit.**

We plot the mutual information between the true and predicted label of a randomly chosen sample (21) in the limit $t \rightarrow \infty$ with $N, P \rightarrow \infty$ as a function of $\alpha \equiv P/N$, computing p_C^μ from (20) (solid lines) and by Monte Carlo integration of $p(\sigma_T, \omega, \sigma)$ (crosses, error bars indicate one standard deviation). The inset shows the learning efficiency (22) in the limits $\tau \rightarrow 0$ (solid) and $\tau \rightarrow \infty$ (dashed). In both plots, ν increases from bottom to top.

by the number of samples and weights, respectively. Plotted in the inset of figure 3, this efficiency never reaches the optimal value 1, even in the limit of vanishing dissipation $\tau \rightarrow \infty$ (solid lines in fig. 3).

Conclusion and perspectives. – We have introduced neural networks as models for studying the thermodynamic efficiency of learning. For the paradigmatic case of learning arbitrary binary labels for given inputs, we showed that the information acquired is bounded by the thermodynamic cost of learning. This is true for learning an arbitrary number of samples in an arbitrary number of dimensions for any learning algorithm without feedback for both batch and online learning.

Our framework opens up numerous avenues for further work. It will be interesting to analyse the efficiency of learning algorithms that employ feedback or use an auxiliary memory [42]. Furthermore, synaptic weight distributions are experimentally accessible [43, 44], offering the exciting possibility to test predictions on learning algorithms by looking at neural weight distributions. The inverse problem, *i.e.* deducing features of learning algorithms or the neural hardware that implements them by optimising some functional like the efficiency, looks like a formidable challenge, despite some encouraging progress in related fields [45, 46].

ACKNOWLEDGMENTS

We thank David Hartich for stimulating discussions and careful reading of the manuscript.

-
- [1] S. Leibler and N. Barkai, *Nature* **387**, 913 (1997).
- [2] W. Bialek, *Biophysics : Searching for Principles*, Princeton University Press, 2011.
- [3] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, *Nat. Phys.* **8**, 422 (2012).
- [4] U. Seifert, *Rep. Prog. Phys.* **75**, 126001 (2012).
- [5] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, *Nat. Phys.* **11**, 131 (2015).
- [6] H. Qian and T. C. Reluga, *Phys. Rev. Lett.* **94**, 028101 (2005).
- [7] Y. Tu, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11737 (2008).
- [8] P. Mehta and D. J. Schwab, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17978 (2012).
- [9] G. De Palo and R. G. Endres, *PLoS Comput. Biol.* **9**, e1003300 (2013).
- [10] C. C. Govern and P. R. ten Wolde, *Phys. Rev. Lett.* **113**, 258102 (2014).
- [11] C. C. Govern and P. R. ten Wolde, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17486 (2014).
- [12] A. C. Barato, D. Hartich, and U. Seifert, *New J. Phys.* **16**, 103024 (2014).
- [13] A. H. Lang, C. K. Fisher, T. Mora, and P. Mehta, *Phys. Rev. Lett.* **113**, 14 (2014).
- [14] P. Sartori, L. Granger, C. F. Lee, and J. M. Horowitz, *PLoS Comput. Biol.* **10**, e1003974 (2014).
- [15] S. Ito and T. Sagawa, *Nat. Commun.* **6**, 7498 (2015).
- [16] D. Andrieux and P. Gaspard, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9516 (2008).
- [17] A. Murugan, D. A. Huse, and S. Leibler, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12034 (2012).
- [18] D. Hartich, A. C. Barato, and U. Seifert, *New J. Phys.* **17**, 055026 (2015).
- [19] S. Lahiri, Y. Wang, M. Esposito, and D. Lacoste, *New J. Phys.* **17**, 085008 (2015).
- [20] A. C. Barato and U. Seifert, *Phys. Rev. Lett.* **114**, 158101 (2015).
- [21] E. R. Kandel, J. H. Schwartz, T. M. Jessell, and Others, *Principles of Neural Science*, McGraw-Hill New York, 2000.
- [22] D. Marr, *J. Physiol.* **202**, 437 (1969).
- [23] J. S. Albus, *Math. Biosci.* **10**, 25 (1971).
- [24] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning*, Cambridge University Press, 2001.
- [25] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.
- [26] Experimental justification for focusing on a single neuron comes from studies on psychophysical judgements in monkeys, which have been shown to depend on very few neurons [47].
- [27] N. van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, 1992.
- [28] Restricting the size of the weights reflects experimental evidence suggesting the existence of an upper bound on synaptic strength in diverse nervous systems [48].
- [29] D. Hartich, A. C. Barato, and U. Seifert, *J. Stat. Mech.* **2014**, P02016 (2014).
- [30] J. M. Horowitz, *J. Stat. Mech.* **2015**, P03006 (2015).
- [31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2006.
- [32] A. E. Allahverdyan, D. Janzing, and G. Mahler, *J. Stat. Mech.* **2009**, P09011 (2009).
- [33] T. Sagawa and M. Ueda, *Phys. Rev. Lett.* **104**, 090602 (2010).
- [34] J. M. Horowitz and M. Esposito, *Phys. Rev. X* **4**, 031015 (2014).
- [35] See Supplemental Material at ..., which includes Ref. [49], for a detailed derivation.
- [36] D. O. Hebb, *The organization of behavior: A neuropsychological approach*, John Wiley & Sons, 1949.
- [37] D. Abreu and U. Seifert, *Europhys. Lett.* **94**, 10001 (2011).
- [38] H. Risken, *The Fokker-Planck Equation*, Springer, 1996.
- [39] In the limit of large N , only the first two moments of the distribution will matter, making this choice equivalent to sampling ξ^μ from the surface of a hypersphere in N dimensions in that limit.
- [40] ω_n is normally distributed since the Langevin equation (15) defines an Ornstein–Uhlenbeck process ω_n which for a Gaussian initial condition as we have chosen remains normally distributed [27].
- [41] E. Gardner, *Europhys. Lett.* **4**, 481 (1987).
- [42] D. Hartich, A. C. Barato, and U. Seifert, *Phys. Rev. E* **93**, 022116 (2016).
- [43] N. Brunel, V. Hakim, P. Isope, J.-P. Nadal, and B. Barbour, *Neuron* **43**, 745 (2004).
- [44] B. Barbour, N. Brunel, V. Hakim, and J.-P. Nadal, *Trends Neurosci.* **30**, 622 (2007).
- [45] G. Tkačik, A. M. Walczak, and W. Bialek, *Phys. Rev. E* **80**, 031920 (2009).
- [46] T. R. Sokolowski and G. Tkačik, *Phys. Rev. E* **91**, 062710 (2015).
- [47] W. T. Newsome, K. H. Britten, and J. A. Movshon, *Nature* **341**, 52 (1989).
- [48] P. Dayan and L. F. Abbott, *Theoretical Neuroscience*, MIT Press, 2001.
- [49] J. M. Horowitz and H. Sandberg, *New J. Phys.* **16**, 125007 (2014).
- [50] If we restricted ourselves to online learning, where the learning force is a local force with only one sample and its label acting on the weights, we could consider this as an upper bound on the amount of information that the weights can acquire during learning, yielding the same result for the efficiency.

SUPPLEMENTAL MATERIAL STOCHASTIC THERMODYNAMICS OF LEARNING

Sebastian Goldt and Udo Seifert

II. Institut für Theoretische Physik, Universität Stuttgart, 70550 Stuttgart, Germany

(Dated: November 30, 2016)

In this supplemental material, we discuss the stochastic thermodynamics of neural networks in detail in section I and derive our main result, eq. (16) of the main text, in section II. Furthermore, we complement our discussion Hebbian learning in the thermodynamic limit with additional analytical calculations in section III.

I. STOCHASTIC THERMODYNAMICS OF NEURAL NETWORKS

We now give a detailed account of the stochastic thermodynamics of neural networks. For simplicity, here we will focus on batch learning; the generalisation to online learning is straightforward. For a network with N weights $\omega_n \in \mathbb{R}^N$ learning P samples $\boldsymbol{\xi}^\mu \in \{\pm 1\}^N$ with their labels $\sigma_T^\mu = \pm 1$, $\mu = 1, 2, \dots, P$, we have N Langevin equations [S1]

$$\dot{\omega}_n(t) = -\omega_n(t) + f(\omega_n(t), \{\xi_n^\mu, \sigma_T^\mu\}, t) + \zeta_n(t). \quad (\text{S1})$$

The Gaussian noise $\zeta_n(t)$ has correlations $\langle \zeta_n(t) \zeta_m(t') \rangle = 2T \delta_{nm} \delta(t-t')$ for $n, m = 1, \dots, N$ where T is the temperature of the surrounding medium and we have set Boltzmann's constant to unity to render entropy dimensionless. The weights $\boldsymbol{\omega}$ determine the transition rates of the P independent two-state processes for the predicted labels σ^μ via

$$k_\mu^+ / k_\mu^- = \exp(\mathcal{A}^\mu / T) \quad (\text{S2})$$

where \mathcal{A}^μ is the input-dependent activation

$$\mathcal{A}^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu \quad (\text{S3})$$

For the remainder of this supplemental material, we set $T = 1$, rendering energy dimensionless. We assume that the thermal noise in each subsystem, like ω_n or σ^μ , is independent of all the others. This multipartite assumption [S2] allows us to write the master equation for the distribution $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}, t)$ with $\boldsymbol{\sigma}_T \equiv (\sigma_T^1, \dots, \sigma_T^P)$ and $\boldsymbol{\sigma} \equiv (\sigma^1, \dots, \sigma^P)$ as

$$\partial_t p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}, t) = - \sum_{n=1}^N \partial_n j_n(t) + \sum_{\mu=1}^P j_\mu(t), \quad (\text{S4})$$

where $\partial_t \equiv \partial/\partial t$, $\partial_n \equiv \partial/\partial \omega_n$ and the probability currents for the n -th weight ω_n and the μ -th predicted label σ^μ are given by

$$j_n(t) = \left[-\omega_n + f(\omega_n, \boldsymbol{\xi}^{\mu(t)}, \sigma_T^{\mu(t)}, t) - \partial_n \right] p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}, t), \quad (\text{S5a})$$

$$j_\mu(t) = k^+ p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \sigma^1, \dots, -\sigma^\mu, \dots, \sigma^P, t) - k^- p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}, t). \quad (\text{S5b})$$

We choose symmetric rates $k_\mu^\pm = \gamma \exp(\pm \mathcal{A}^\mu / 2)$ with $\gamma \gg 1$. Initially, the true labels $\boldsymbol{\sigma}_T$, weights $\boldsymbol{\omega}$ and predicted labels are all uncorrelated with

$$p_0(\sigma_T^\mu) = 1/2, \quad (\text{S6})$$

$$p_0(\sigma^\mu) = 1/2, \quad \text{and} \quad (\text{S7})$$

$$p_0(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{N/2}} \exp(-\boldsymbol{\omega} \cdot \boldsymbol{\omega} / 2). \quad (\text{S8})$$

Since the following discussion applies to the time-dependent dynamics (S4), we understand that all quantities that will be introduced in the remainder of this section have an implicit time-dependence via the distribution $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}, t)$ or the currents (S5).

Our starting point for the stochastic thermodynamics of this system is the well-known total entropy production \dot{S}^{tot} of the network which obeys the following second-law like inequality [S3]

$$\dot{S}^{\text{tot}} = \partial_t S(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) + \dot{S}^{\text{m}} \geq 0 \quad (\text{S9})$$

with equality in equilibrium only. Here, we have the Shannon entropy [S5] of the system,

$$S(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) = - \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) \ln p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}). \quad (\text{S10})$$

Here, we include the variables $\boldsymbol{\sigma}_T$, $\boldsymbol{\omega}$ and $\boldsymbol{\sigma}$ as arguments of the function S in a slight abuse of notation to emphasise that we consider the Shannon entropy of the full distribution $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$. \dot{S}^{m} gives the rate of entropy production in the medium. For a system at constant temperature $T = 1$, $\dot{S}^{\text{m}} \equiv \dot{Q}$, the rate of heat dissipation into the medium [S3]. Let us first focus on the change in Shannon entropy by differentiating (S10) with respect to time,

$$\partial_t S(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) = - \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} \dot{p}(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) \ln p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}), \quad (\text{S11})$$

where we have used that $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$ is, of course, normalised. Using the master equation (S4), we find that

$$\partial_t S(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) = \sum_{n=1}^N \dot{S}_n + \sum_{\mu=1}^P \dot{S}_\mu \quad (\text{S12})$$

where

$$\dot{S}_n \equiv \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} \partial_n j_n(t) \ln p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}), \quad (\text{S13})$$

$$\dot{S}_\mu \equiv - \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} j_\mu \ln p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}), \quad (\text{S14})$$

are the rate of change of the Shannon entropy $S(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$ due to the dynamics of ω_n and σ^μ , respectively. The key point here is that multipartite dynamics, a consequence of the uncorrelated noise across subsystems, lead to a linear splitting of the probability currents and hence to a linear splitting of all quantities which are functions of the total probability current. Similarly, for the rate of heat dissipation \dot{Q} , we can write

$$\dot{Q} = \sum_{n=1}^N \dot{Q}_n + \sum_{\mu=1}^P \dot{Q}_\mu \quad (\text{S15})$$

where

$$\dot{Q}_n = \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} j_n(t) F_n(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) \quad (\text{S16})$$

with the total force on the n -th weight $F_n = -\omega_n(t) + f(\omega_n(t), \{\xi_n^\mu, \sigma_T^\mu\}, t)$, while

$$\dot{Q}_\mu = \sum_{\boldsymbol{\sigma}_T, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} d\boldsymbol{\omega} j_\mu(t) \sigma^\mu \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu / 2. \quad (\text{S17})$$

Finally, total entropy production \dot{S}^{tot} can also be split,

$$\dot{S}^{\text{tot}} = \sum_{n=1}^N \dot{S}_n^{\text{tot}} + \sum_{\mu=1}^P \dot{S}_\mu^{\text{tot}}. \quad (\text{S18})$$

It can easily be shown that each of these total entropy productions of a subsystem obeys a separate second-law like inequality, *e.g.*

$$\dot{S}_n^{\text{tot}} = \dot{S}_n(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) + \dot{Q}_n \geq 0 \quad (\text{S19})$$

for the n -th weight.

Writing

$$p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma}) = p(\omega_n) p(\boldsymbol{\sigma}_T, \bar{\boldsymbol{\omega}}, \boldsymbol{\sigma} | \omega_n) \quad (\text{S20})$$

with $\bar{\omega} \equiv (\dots, \omega_{n-1}, \omega_{n+1}, \dots)$, we can split $\dot{S}_n(\sigma_T, \omega, \sigma)$ into two parts: first, the change of Shannon entropy of the marginalized distribution $p(\omega_n)$,

$$\dot{S}_n(\omega_n) = \sum_{\sigma_T, \sigma} \int_{-\infty}^{\infty} d\omega \partial_n j_n(t) \ln p(\omega_n) = \partial_t S(\omega_n), \quad (\text{S21})$$

where the last equality follows from the fact that an entropy change of the marginalized distribution $p(\omega_n)$ can only come from the dynamics of ω_n . The second part is called the learning rate [S4]

$$l_n(\omega_n; \sigma_T, \sigma, \bar{\omega}) = - \sum_{\sigma_T, \sigma} \int_{-\infty}^{\infty} d\omega \partial_n j_n(t) \ln p(\sigma_T, \sigma, \bar{\omega} | \omega_n) \quad (\text{S22})$$

or information flow [S6, S7]. We emphasise that this learning rate l_n is thermodynamic and has nothing to do with the learning rate ν that goes into the definition of the learning algorithms, see for example eq. (12) of the main text. To avoid confusion, we will refer to l_n as the thermodynamic learning rate for the remainder of this supplemental material. The second law (S19) for the n -th weight hence becomes

$$\dot{S}_n^{\text{tot}} = \partial_t S(\omega_n) + \dot{Q}_n - l_n(\omega_n; \sigma_T, \sigma, \bar{\omega}) \geq 0 \quad (\text{S23})$$

The thermodynamic learning rate is a thermodynamically consistent measure of how much the dynamics of ω_n change the mutual information $I(\omega_n : \sigma_T, \bar{\omega}, \sigma)$, in particular for a system that continuously rewrites a single memory [S8].

We can further refine the second law (S23) by exploiting the causal structure of the dynamics, as was recently suggested by Horowitz [S2]. The subsystem ω_n directly interacts only with those degrees of freedom that appear in its probability current $j_n(t)$ (S5). From inspection of the current $j_n(t)$, we see that ω_n is directly influenced only by itself and the given labels σ_T . Keeping this in mind, we use the chain rule for mutual information [S5] to write

$$I(\omega_n : \sigma_T, \bar{\omega}, \sigma) = I(\omega_n : \sigma_T) + I(\omega_n : \bar{\omega}, \sigma | \sigma_T), \quad (\text{S24})$$

where we use the conditional mutual information

$$I(\omega_n : \bar{\omega}, \sigma | \sigma_T) = S(\omega_n | \sigma_T) - S(\omega_n | \bar{\omega}, \sigma, \sigma_T) \quad (\text{S25})$$

$$= - \sum_{\sigma, \sigma_T} \int_{-\infty}^{\infty} d\omega p(\sigma_T, \omega, \sigma) \ln \frac{p(\sigma_T, \omega, \sigma) p(\sigma_T)}{p(\omega_n, \sigma_T) p(\bar{\omega}, \sigma, \sigma_T)}. \quad (\text{S26})$$

Accordingly, we split the thermodynamic learning rate (S22) into a thermodynamic learning rate of the n -th weight with the degrees of freedom that it directly interacts with, *i.e.* the true labels σ_T ,

$$l_n(\omega_n; \sigma_T) = - \sum_{\sigma, \sigma_T} \int_{-\infty}^{\infty} d\omega \partial_n j_n(t) \ln p(\sigma_T | \omega_n), \quad (\text{S27})$$

and a thermodynamic learning rate with the other subsystems given the true labels,

$$l_n(\omega_n; \bar{\omega}, \sigma | \sigma_T) = - \sum_{\sigma, \sigma_T} \int_{-\infty}^{\infty} d\omega \partial_n j_n(t) \ln \left(\frac{p(\omega_n, \bar{\omega}, \sigma | \sigma_T)}{p(\omega_n | \sigma_T) p(\bar{\omega}, \sigma | \sigma_T)} \right). \quad (\text{S28})$$

Horowitz proved [S2] the following second-law like inequality including the refined thermodynamic learning rate (S27),

$$\partial_t S(\omega_n) + \dot{Q}_n - l_n(\omega_n; \sigma_T) \geq 0. \quad (\text{S29})$$

which is the basis for our proof of the main inequality, equation (16) of the main text.

II. DERIVATION OF INEQUALITY (16) OF THE MAIN TEXT

The stochastic thermodynamics of neural networks yields N inequalities of the form (S29). Integrating over time and summing over all the weights, we find

$$\sum_{n=1}^N [\Delta S(\omega_n) + \Delta Q_n] \geq \sum_{n=1}^N \int_0^{\infty} dt l_n(\omega_n; \sigma_T) = \sum_{n=1}^N \Delta I(\omega_n : \sigma_T) \quad (\text{S30})$$

The precise definition of all the terms are discussed in the main text and in section I of this supplemental material. The crucial point for the last equality is that the labels σ_T are static, so that the mutual information $I(\omega_n : \sigma_T)$ changes only due to the dynamics of ω_n and hence $\partial_t I(\omega_n : \sigma_T) = l_n(\omega_n; \sigma_T)$ [S50]. To make progress towards our main result, inequality (16) of the main text, we need to show that

$$\sum_{n=1}^N \Delta I(\omega_n : \sigma_T) \geq \sum_{\mu=1}^P \Delta I(\sigma_T^\mu : \sigma^\mu). \quad (\text{S31})$$

First, we note that from the chain rule of mutual information [S5], we have

$$\Delta I(\omega : \sigma_T) = \Delta I(\omega_1, \dots, \omega_n : \sigma_T) = \sum_{n=1}^N \Delta I(\omega_n : \sigma_T | \omega_{n-1}, \dots, \omega_1) \quad (\text{S32})$$

with the conditional mutual information [S5]

$$I(\omega_n : \sigma_T | \omega_{n-1}, \dots, \omega_1) \equiv S(\omega_n | \omega_{n-1}, \dots, \omega_1) - S(\omega_n | \sigma_T, \omega_{n-1}, \dots, \omega_1). \quad (\text{S33})$$

Due to the form of the Langevin equation for the single weight, eq. (S1), individual weights are uncorrelated, and hence the conditional mutual information simplifies to

$$\Delta I(\omega_n : \sigma_T | \omega_{n-1}, \dots, \omega_1) = \Delta S(\omega_n | \omega_{n-1}, \dots, \omega_1) - \Delta S(\omega_n | \sigma_T, \omega_{n-1}, \dots, \omega_1) \quad (\text{S34})$$

$$= \Delta S(\omega_n) - \Delta S(\omega_n | \sigma_T) \quad (\text{S35})$$

$$= \Delta I(\omega_n : \sigma_T) \quad (\text{S36})$$

such that

$$\sum_{n=1}^N \Delta I(\omega_n : \sigma_T) = \Delta I(\omega : \sigma_T). \quad (\text{S37})$$

Next, we show that

$$\Delta I(\omega : \sigma_T) = \sum_{\mu=1}^P \Delta I(\omega : \sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) \stackrel{!}{\geq} \sum_{\mu=1}^P \Delta I(\omega : \sigma_T^\mu). \quad (\text{S38})$$

using the independence of the given labels σ_T . We first note that

$$\Delta I(\omega : \sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) = \Delta S(\sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) - \Delta S(\sigma_T^\mu | \omega, \sigma_T^{\mu-1}, \dots, \sigma_T^1) \quad (\text{S39})$$

$$= \Delta S(\sigma_T^\mu) - \Delta S(\sigma_T^\mu | \omega, \sigma_T^{\mu-1}, \dots, \sigma_T^1) \quad (\text{S40})$$

while

$$\Delta I(\omega : \sigma_T^\mu) = \Delta S(\sigma_T^\mu) - \Delta S(\sigma_T^\mu | \omega) \quad (\text{S41})$$

Hence for $\Delta I(\omega : \sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) \stackrel{!}{\geq} \Delta I(\omega : \sigma_T^\mu)$, we need

$$\Delta I(\omega : \sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) - \Delta I(\omega : \sigma_T^\mu) \quad (\text{S42})$$

$$= \Delta S(\sigma_T^\mu | \omega) - \Delta S(\sigma_T^\mu | \omega, \sigma_T^{\mu-1}, \dots, \sigma_T^1) \quad (\text{S43})$$

$$= \Delta I(\sigma_T^\mu : \sigma_T^{\mu-1}, \dots, \sigma_T^1 | \omega) \quad (\text{S44})$$

$$\geq 0 \quad (\text{S45})$$

where we first used that the σ_T^μ are independent and identically distributed. The last inequality follows since any mutual information, conditional or not, is always greater than or equal to zero [S5]. We have thus shown that $\Delta I(\omega : \sigma_T^\mu | \sigma_T^{\mu-1}, \dots, \sigma_T^1) \geq \Delta I(\omega : \sigma_T^\mu)$ and hence (S38) is true.

Finally, to prove that $\Delta I(\omega : \sigma_T^\mu) > \Delta I(\sigma_T^\mu : \sigma^\mu)$, we consider the full probability distribution $p(\sigma_T, \omega, \sigma)$. From the master equation, eq. (S4), we can write this distribution as

$$p(\sigma_T, \omega, \sigma) = p(\sigma_T) p(\omega | \sigma_T) \left[p^{(0)}(\sigma | \omega) + \frac{1}{\gamma} p^{(1)}(\sigma | \omega) + \mathcal{O}(1/\gamma^2) \right] \quad (\text{S46})$$

with $\gamma \gg 1$ for physiological reasons as described in the text – it takes the neuron longer to learn than to generate an action potential. Hence to first order, $\sigma_{\text{T}} \rightarrow \omega \rightarrow \sigma$ is by definition a Markov chain [S5]. Integrating out all the labels, true and predicted, except for the μ -th one, we have the Markov chain $\sigma_{\text{T}}^{\mu} \rightarrow \omega \rightarrow \sigma^{\mu}$. For such a Markov chain, it is easy to show the following data processing inequality [S5],

$$\Delta I(\sigma_{\text{T}}^{\mu} : \omega) \geq \Delta I(\sigma_{\text{T}}^{\mu} : \sigma^{\mu}), \quad (\text{S47})$$

which completes our derivation.

III. HEBBIAN LEARNING IN THE THERMODYNAMIC LIMIT

In this section, we provide additional analytical calculations for Hebbian learning in the thermodynamic limit for long times $t \rightarrow \infty$.

A. Direct integration of the full distribution $p(\sigma_{\text{T}}, \omega, \sigma)$

To compute the mutual information between the true and predicted label of a given sample, $I(\sigma_{\text{T}}^{\mu} : \sigma^{\mu})$, we need the distribution $p(\sigma_{\text{T}}^{\mu}, \sigma^{\mu})$ or, since both σ_{T}^{μ} and σ^{μ} are symmetric binary random variables, the probability that $\sigma_{\text{T}}^{\mu} = \sigma^{\mu}$. Our aim in this section is to obtain this probability for Hebbian learning in the thermodynamic limit with $t \rightarrow \infty$ by direct integration of the full distribution over the true labels, weights and predicted labels for a given set of samples $\{\xi^{\mu}\}$, which will also give additional motivation for introducing the stability Δ^{μ} of a sample.

We start with the full probability distribution

$$p(\sigma_{\text{T}}, \omega, \sigma) = \left(\frac{1}{2}\right)^P \left(\prod_{n=1}^N \frac{e^{-(\omega_n - \nu \mathcal{F}_n)^2/2}}{\sqrt{2\pi}}\right) \left(\prod_{\mu=1}^P \frac{e^{\sigma^{\mu} \omega \cdot \xi^{\mu}/2\sqrt{N}}}{e^{-\omega \cdot \xi^{\mu}/2\sqrt{N}} + e^{\omega \cdot \xi^{\mu}/2\sqrt{N}}}\right), \quad (\text{S48})$$

where ν is the learning rate and \mathcal{F}_n is a suitably scaled average over the samples and labels,

$$\mathcal{F}_n = \frac{1}{\sqrt{N}} \sum_{\rho=1}^P \sigma_{\text{T}}^{\rho} \xi_n^{\rho} \quad (\text{S49})$$

While the sum over the predicted labels $\sigma^{\rho \neq \mu} = \pm 1$ is trivial, we can integrate over the true labels by noting that we can rewrite the exponent as

$$p(\sigma_{\text{T}}, \omega, \sigma^{\mu}) = \left(\frac{1}{2}\right)^P \left(\prod_{n=1}^N \frac{e^{-(\omega_n - \nu \sigma_{\text{T}}^{\mu} \xi_n^{\mu}/\sqrt{N} - \nu \mathcal{F}_n^{\bar{\mu}})^2/2}}{\sqrt{2\pi}}\right) \frac{e^{\sigma^{\mu} \omega \cdot \xi^{\mu}/2\sqrt{N}}}{e^{-\omega \cdot \xi^{\mu}/2\sqrt{N}} + e^{\omega \cdot \xi^{\mu}/2\sqrt{N}}} \quad (\text{S50})$$

where the only dependence of the weight distribution on the true labels $\sigma_{\text{T}}^{\rho \neq \mu}$ is now confined to the sum

$$\mathcal{F}_n^{\bar{\mu}} \equiv \frac{1}{\sqrt{N}} \sum_{\rho \neq \mu}^P \sigma_{\text{T}}^{\rho} \xi_n^{\rho}. \quad (\text{S51})$$

In the thermodynamic limit, this allows us to replace the sum over all $\sigma_{\text{T}}^{\mu \neq \rho}$ by an integral over the stochastic variable $\mathcal{F}_n^{\bar{\mu}}$, which is normally distributed by the central limit theorem and has mean 0 and variance α . Carrying out the integral, we find

$$p(\sigma_{\text{T}}^{\mu}, \omega, \sigma^{\mu}) = \left(\prod_{n=1}^N \frac{e^{-(\omega_n - \nu \sigma_{\text{T}}^{\mu} \xi_n^{\mu}/\sqrt{N})^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}}\right) \frac{e^{\sigma^{\mu} \omega \cdot \xi^{\mu}/2\sqrt{N}}}{e^{-\omega \cdot \xi^{\mu}/2\sqrt{N}} + e^{\omega \cdot \xi^{\mu}/2\sqrt{N}}} \quad (\text{S52})$$

Since both σ_{T}^{μ} and σ^{μ} are binary random variables and $\sigma_{\text{T}}^{\mu} = \pm 1$ with equal probabilities, the mutual information between the true and predicted label can be written as

$$I(\sigma_{\text{T}}^{\mu} : \sigma^{\mu}) = \ln 2 - S[p(\sigma_{\text{T}}^{\mu} = \sigma^{\mu})] \quad (\text{S53})$$

with the shorthand for the binary entropy $S[p] = -p \ln p - (1-p) \ln(1-p)$ [S5]. With $\sigma^\mu = \sigma_T^\mu$ in the exponential term of eq. (S52) and noting that $(\sigma_T^\mu \xi_n^\mu)^2 = 1$ for all σ_T^μ, ξ_n^μ , we then have

$$p(\sigma_T^\mu = \sigma^\mu, \boldsymbol{\omega}) = \left(\prod_{n=1}^N \frac{e^{-(\omega_n \sigma_T^\mu \xi_n^\mu - \nu/\sqrt{N})^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}} \right) \frac{e^{\sigma_T^\mu \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu/2\sqrt{N}}}{e^{-\boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu/2\sqrt{N}} + e^{\boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu/2\sqrt{N}}} \quad (\text{S54})$$

It thus becomes clear that $\boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu \sigma_T^\mu$ is the sum of N random variables with mean ν/\sqrt{N} and variance $1 + \alpha\nu^2$. We are then motivated to introduce the stability of a sample,

$$\Delta^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu \sigma_T^\mu = \mathcal{A}^\mu \sigma_T^\mu. \quad (\text{S55})$$

which, from eq. (S54), is normally distributed with mean ν and variance $1 + \alpha\nu^2$. Introducing the stability allows us to replace the integral over all the weights by an integral over the stability,

$$p(\sigma_T^\mu = \sigma^\mu) = \int_{-\infty}^{\infty} d\Delta^\mu \frac{e^{-(\Delta^\mu - \nu)^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}} \frac{e^{\Delta^\mu}}{1 + e^{\Delta^\mu}} = \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \frac{e^{\Delta^\mu}}{1 + e^{\Delta^\mu}} \quad (\text{S56})$$

which is the distribution obtained as eq. (20) of the main text.

B. Direct derivation of the distribution of stabilities

Let us quickly show how the distribution of stabilities

$$\Delta^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu \sigma_T^\mu, \quad (\text{S57})$$

$\mu = 1, \dots, P$, is obtained directly from its definition. The weights are given by

$$\boldsymbol{\omega} = \frac{1}{\sqrt{N}} \nu \sum_{\rho=1}^P \boldsymbol{\xi}^\rho \sigma_T^\rho + \mathbf{y} \quad (\text{S58})$$

with $\mathbf{y} = (y_1, y_2, \dots, y_N)$ where y_n are normally distributed random variables with mean 0 and variance 1 arising from the thermal fluctuations in equilibrium. Substituting eq. (S58) into (S57), we have

$$\Delta^\mu = \frac{1}{N} \nu \sum_{\rho=1}^P \sigma_T^\rho \sigma_T^\mu \boldsymbol{\xi}^\rho \cdot \boldsymbol{\xi}^\mu + \frac{1}{\sqrt{N}} \sigma_T^\mu \boldsymbol{\xi}^\mu \cdot \mathbf{y} \quad (\text{S59})$$

$$= \nu + \frac{1}{N} \nu \sum_{\rho \neq \mu}^P \sigma_T^\rho \sigma_T^\mu \boldsymbol{\xi}^\rho \cdot \boldsymbol{\xi}^\mu + \frac{1}{\sqrt{N}} \sigma_T^\mu \boldsymbol{\xi}^\mu \cdot \mathbf{y} \quad (\text{S60})$$

where going to the last line we have used the fact that $\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu = N$. By inspection, we see that the second term is the sum of $N(P-1) \approx NP$ random numbers $\pm\nu/N$ and the last term is the sum of N random numbers y_n/\sqrt{N} . By the central limit theorem, Δ^μ is hence normally distributed with mean $\overline{\langle \Delta^\mu \rangle} = \nu$ and variance

$$\overline{\langle (\Delta^\mu)^2 \rangle} - \overline{\langle \Delta^\mu \rangle}^2 = \nu^2 + NP \frac{\nu^2}{N^2} + N \frac{1}{N} - \nu^2 = 1 + \alpha\nu^2. \quad (\text{S61})$$

C. Analytical approximation for $I(\sigma_T : \sigma)$

We quantify the success of learning using the mutual information per sample,

$$I(\sigma_T^\mu : \sigma^\mu) = \ln 2 - S(p_C^\mu) \quad (\text{S62})$$

where $S(p) = -[p \ln p + (1-p) \ln(1-p)]$ is the binary Shannon entropy and p_C^μ is defined as

$$p_C^\mu \equiv p(\sigma^\mu = \sigma_T^\mu) = \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \frac{e^{\Delta^\mu}}{e^{\Delta^\mu} + 1} \quad (\text{S63})$$

The stabilities Δ^μ are normally distributed with mean ν and variance $1 + \alpha\nu^2$ (see section III B). This integral does not have a closed-form analytical solution, but here we will demonstrate a very good analytical approximation.

To that end, we first rewrite the sigmoid function in the integrand in terms of the hyperbolic tangent and exploit the similarity of the latter to the error function:

$$p_C^\mu = \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \frac{e^{\Delta^\mu/2}}{e^{\Delta^\mu/2} + e^{-\Delta^\mu/2}} \quad (\text{S64})$$

$$= \frac{1}{2} + \frac{1}{2} \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \tanh(\Delta^\mu/2) \quad (\text{S65})$$

$$\simeq \frac{1}{2} + \frac{1}{2} \int_{-\infty}^{\infty} d\Delta^\mu p(\Delta^\mu) \operatorname{erf}(\gamma\Delta^\mu/2) \quad (\text{S66})$$

where we choose $\gamma = 4/5$ by inspection of the graphs of the two functions. Now the convolution of a normal distribution and an error function has an exact solution,

$$\frac{1}{\sqrt{2\pi d^2}} \int_{-\infty}^{\infty} dx \operatorname{erf}(ax + b) \exp\left(-\frac{(x - c)^2}{2d^2}\right) = \operatorname{erf}\left(\frac{b + ac}{\sqrt{1 + 2a^2 d^2}}\right). \quad (\text{S67})$$

Setting $a = \gamma/2$, $b = 0$, $c = \nu$ and $d^2 = 1 + \alpha\nu^2$, we find that

$$p_C^\mu(\alpha, \nu) \simeq \frac{1}{2} + \frac{1}{2} \operatorname{erf} \frac{\gamma\nu/2}{\sqrt{1 + \gamma^2(1 + \alpha\nu^2)}/2} \quad (\text{S68})$$

$$= \frac{1}{2} + \frac{1}{2} \operatorname{erf} \frac{\nu/2}{\sqrt{25/16 + 1/2 + \alpha\nu^2/2}} \quad (\text{S69})$$

$$\simeq \frac{1}{2} + \frac{1}{2} \operatorname{erf} \frac{\nu/2}{\sqrt{2(1 + \alpha\nu^2/4)}} \quad (\text{S70})$$

$$= p(\Delta^\mu > 0 | \alpha, \nu/2) \quad (\text{S71})$$

where in the last line we recognise by inspection that our result is nothing but the integral over the distribution of stabilities $p(\Delta^\mu | \alpha, \nu/2)$ from 0 to ∞ . The probability that the neuron predicts the correct label is hence given by the probability that the neuron learned the label correctly, $\Delta^\mu > 0$, with *half the learning rate*.

[S1] N. van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, 1992.

[S2] J. M. Horowitz, *J. Stat. Mech.* **2015**, P03006 (2015).

[S3] U. Seifert, *Rep. Prog. Phys.* **75**, 126001 (2012).

[S4] D. Hartich, A. C. Barato, and U. Seifert, *J. Stat. Mech.* **2014**, P02016 (2014).

[S5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2006.

[S6] A. E. Allahverdyan, D. Janzing, and G. Mahler, *J. Stat. Mech.* **2009**, P09011 (2009).

[S7] J. M. Horowitz and M. Esposito, *Phys. Rev. X* **4**, 031015 (2014).

[S8] J. M. Horowitz and H. Sandberg, *New J. Phys.* **16**, 125007 (2014).