Stochastic Thermodynamics of Learning

Sebastian Goldt* and Udo Seifert
II. Institut für Theoretische Physik, Universität Stuttgart, 70550 Stuttgart, Germany
(Received 11 July 2016; revised manuscript received 12 October 2016; published 6 January 2017)

Virtually every organism gathers information about its noisy environment and builds models from those data, mostly using neural networks. Here, we use stochastic thermodynamics to analyze 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 analyze Hebbian learning in the thermodynamic limit.

DOI: 10.1103/PhysRevLett.118.010601

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 is 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 $\boldsymbol{\omega} \in \mathbb{R}^N$ of these connections, see Fig. 1. The neuron must learn whether it should fire an action potential or not for a set of P fixed input patterns or samples $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, $\mu = 1, 2, \ldots, P$. Each pattern describes the activity of all the other connected neurons at a point in time: if the nth connected neuron is firing an action potential in the pattern $\boldsymbol{\xi}^\mu$, then $\xi_n^\mu = 1$. For symmetry reasons, we set $\xi_n^\mu = -1$ in case the nth neuron is silent in the μth pattern. Every sample $\boldsymbol{\xi}^\mu$ 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 $\boldsymbol{\xi}^\mu$ with a stochastic process $\sigma^\mu = \pm 1$ (right panel in Fig. 1). Assuming a thermal environment at fixed temperature T, the transition rates $k_\mu^\pm$ for these processes obey the detailed balance condition

$$k_\mu^+ / k_\mu^- = \exp\left(\mathcal{A}^\mu / k_B T\right), \tag{1}$$

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

$$\mathcal{A}^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu, \tag{2}$$

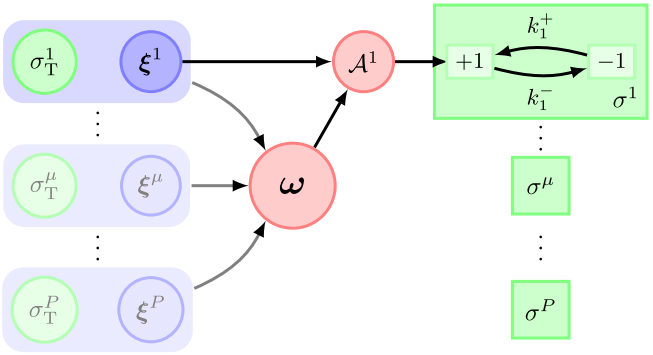

FIG. 1. Model of a single neuron. Given a set of inputs $\boldsymbol{\xi}^\mu \in \{\pm 1\}^N$ and their true labels $\sigma_T^\mu = \pm 1$ (left), the neuron learns the mappings $\boldsymbol{\xi}^\mu \to \sigma_T^\mu$ by adjusting its weights $\boldsymbol{\omega} \in \mathbb{R}^N$. It processes an input by computing the activation $\mathcal{A}^\mu = \boldsymbol{\omega} \cdot \boldsymbol{\xi}^\mu / \sqrt{N}$, 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 $\mu = 1$.

where the prefactor ensures the conventional normalization. We interpret $p(\sigma^\mu = 1|\boldsymbol{\omega})$ with fixed $\boldsymbol{\xi}^\mu$ as the probability that the μth input would trigger an action potential by the neuron. The goal of learning is to adjust the weights of the network $\boldsymbol{\omega}$ such that the predicted labels at any one time $\boldsymbol{\sigma} = (\sigma^1, \ldots, \sigma^P)$ equal the true labels $\boldsymbol{\sigma}_T = (\sigma_T^1, \ldots, \sigma_T^P)$ 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 [28]

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

The total force on the weight arises from a harmonic potential $V(\omega) = \omega^2/2$, restricting the size of the weight [29], 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 realizations, 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,

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

the master equation [28] for the probability distribution $p(\sigma_T, \omega, \sigma, t)$ with given ξ reads

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

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

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

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

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 [31,32]. 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 nonequilibrium thermodynamics is the Shannon entropy of a random variable X with probability distribution $p(x)$,

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

which is a measure of the uncertainty of X [33]. 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 learned is the mutual information

$$I(\sigma_T : \sigma) \equiv S(\sigma_T) - S(\sigma_T|\sigma), \tag{8}$$

which measures by how much, on average, the uncertainty about σ_T is reduced by knowing σ [33]. 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,

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

Here, ΔQ is the heat dissipated into the medium by the dynamics of the weight and $\Delta S(\omega)$ is the difference in the 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) and (6), the information learned is bounded by the thermodynamic costs of learning

$$I(\sigma_T : \sigma) \leq \Delta S(\omega) + \Delta Q \tag{10}$$

for arbitrary learning algorithm $f(\omega, \xi, \sigma_T, t)$ at all times $t > t_0$. This inequality is our first result. We emphasize that while relations between changes in mutual information and total entropy production have appeared in the literature [31,32,34–36], 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 [37] builds on recent work by Horowitz [32] and can be generalized to N dimensions and P samples, see Eq. (16) below. Equation (10) suggests to introduce an efficiency of learning

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

Toy model.—As a first example, let us calculate the efficiency of Hebbian learning, a form of coincidence learning well known from biology [21,39], for $N = P = 1$ in the limit $t \to \infty$. 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 minimize dissipation [40], we choose a learning force f linearly increasing with time

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

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 Eqs. (5) and (6) and solving the resulting Fokker-Planck equation [41]. The total heat dissipated into the medium ΔQ is given by [4]

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

As expected, no heat is dissipated in the limit of infinitely slow driving, $\lim_{\tau \to \infty} \Delta Q = 0$, while for a sudden potential switch $\tau \to 0$, $\lim_{\tau \to 0} \Delta Q = \nu^2 \mathcal{F}^2/2$. The change in the 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 Eq. (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 $\Delta S(\omega) \to \ln 2$ and a strongly biased distribution $p(\sigma|\omega)$ such that $I(\sigma_T : \sigma) \to \ln 2$. Second, we need to minimize 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 $\boldsymbol{\omega}$ learning P samples with true labels $\boldsymbol{\sigma}_T \equiv (\sigma_T^1, \ldots, \sigma_T^\mu, \ldots, \sigma_T^P)$, 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 $n, m = 1, \ldots, N$. Two learning scenarios are possible. In batch learning, the learning force is a function of all samples and their labels,

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

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

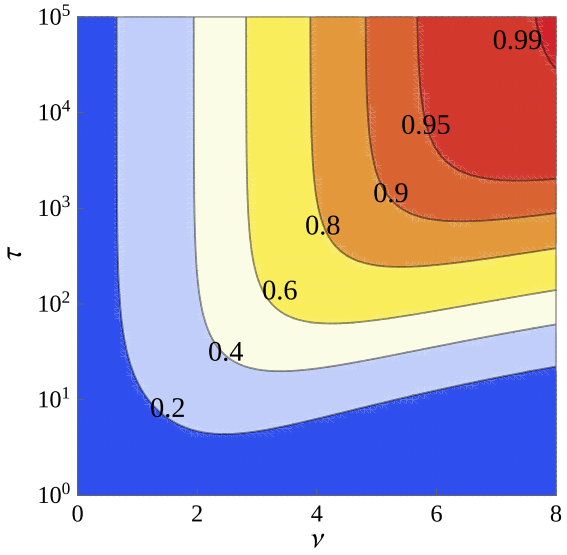

FIG. 2. Learning efficiency of a neuron with a single weight. We plot the efficiency η (11) for a neuron with a single weight learning a single sample as a function of the learning rate ν and learning duration τ in the limit $t \to \infty$.

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

The sample and label that enter this force are given by $\mu(t) \in \{1, \ldots, P\}$, which might be a deterministic function or a random process. Either way, the weights $\boldsymbol{\omega}$ determine the transition rates of the P independent two-state processes for the predicted labels $\boldsymbol{\sigma} \equiv (\sigma^1, \ldots, \sigma^\mu, \ldots, \sigma^P)$ via Eqs. (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 $p(\boldsymbol{\omega}) \propto \exp(-\boldsymbol{\omega} \cdot \boldsymbol{\omega}/2)$ 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 [37] that this information is bounded by the total entropy production of all the weights,

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

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

Let us now compute the efficiency of on-line Hebbian learning in the limit $t \to \infty$. 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

$$\alpha \equiv P/N \tag{17}$$

on the order of 1. The samples $\boldsymbol{\xi}^\mu$ are drawn at random from $p(\xi_n^\mu = 1) = p(\xi_n^\mu = -1) = 1/2$ and remain fixed [42]. We choose a learning force on the nth weight of the form (12) with $\mathcal{F} \to \mathcal{F}_n$ and assume that the process $\mu(t)$ is a random walk over the integers $1, \ldots, P$ changing on a time scale much shorter than the relaxation time of the weights. Since f^2 is finite, the learning force is effectively constant with

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

where the prefactor ensures the conventional normalization [24]. Hence, all the weights ω_n are independent of each other and statistically equivalent. Averaging first over the noise with fixed $\boldsymbol{\sigma}_T$, we find that ω_n is normally distributed with mean $\langle \omega_n \rangle = \nu \mathcal{F}_n$ and variance 1 [43]. The average with respect to the quenched disorder $\boldsymbol{\sigma}_T$, 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 the 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 nth weight $\overline{\Delta Q_n}$ is obtained by averaging Eq. (13) over $\mathcal{F} \to \mathcal{F}_n$.

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 $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$ [37]. Here, we will take a simpler route starting from the stability of the μth sample [44]

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

Its role can be appreciated by considering the limit $T \to 0$, 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 μth label correctly with probability

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

where $p(\Delta^\mu)$ is the distribution generated by thermal noise and quenched disorder, yielding a Gaussian with mean ν and variance $1 + \alpha\nu^2$ [37]. The mutual information follows as

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

with the shorthand for the entropy of a binary random variable $S(p) = -p \ln p - (1-p)\ln(1-p)$ [33]. It is plotted in Fig. 3 together with the mutual information obtained by Monte Carlo integration of $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$ with $N = 10000$. For a vanishing learning rate $\nu \to 0$ or infinitely many samples $\alpha \to \infty$, $p_C^\mu \to 1/2$ and hence $I(\sigma_T^\mu : \sigma^\mu) \to 0$. 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 $\nu \to \nu/2$ [37] (dashed lines in Fig. 3).

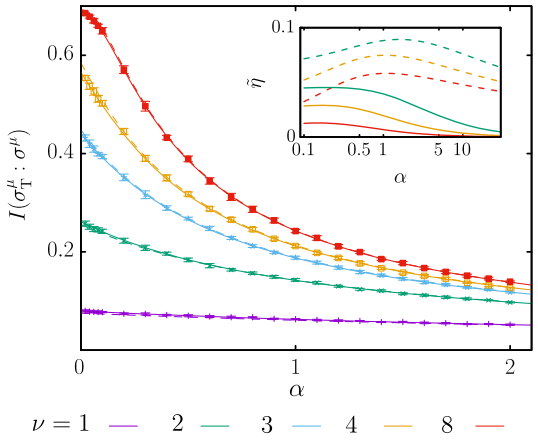

FIG. 3. Hebbian learning in the thermodynamic limit. We plot the mutual information between the true and predicted label of a randomly chosen sample (21) in the limit $t \to \infty$ with $N, P \to \infty$ as a function of $\alpha \equiv P/N$, computing p_C^μ from Eq. (20) (solid lines) and by Monte Carlo integration of $p(\boldsymbol{\sigma}_T, \boldsymbol{\omega}, \boldsymbol{\sigma})$ (crosses; the error bars indicate 1 standard deviation). The inset shows the learning efficiency (22) in the limits $\tau \to 0$ (solid) and $\tau \to \infty$ (dashed). In both plots, ν increases from bottom to top.

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

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

where we have taken the mutual information per sample and the total entropy production per weight, multiplied by the number of samples and weights, respectively. Plotted in the inset of Fig. 3, this efficiency never reaches the optimal value 1, even in the limit of vanishing dissipation $\tau \to \infty$ (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 on-line learning.

Our framework opens up numerous avenues for further work. It will be interesting to analyze the efficiency of

learning algorithms that employ feedback or use an auxiliary memory [45]. Furthermore, synaptic weight distributions are experimentally accessible [46,47], 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 optimizing some functional like the efficiency, looks like a formidable challenge, despite some encouraging progress in related fields [48,49].

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

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

[1] S. Leibler and N. Barkai, Nature (London) **387**, 913 (1997).
[2] W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, NJ, 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**, 148103 (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, Cambridge, England, 2001).
[25] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 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 [27].
[27] W. T. Newsome, K. H. Britten, and J. A. Movshon, Nature (London) **341**, 52 (1989).
[28] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, New York, 1992).
[29] Restricting the size of the weights reflects experimental evidence suggesting the existence of an upper bound on synaptic strength in diverse nervous systems [30].
[30] P. Dayan and L. F. Abbott, *Theoretical Neuroscience* (MIT Press, Cambridge, MA, 2001).
[31] D. Hartich, A. C. Barato, and U. Seifert, J. Stat. Mech. (2014) P02016.
[32] J. M. Horowitz, J. Stat. Mech. (2015) P03006.
[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2006).
[34] A. E. Allahverdyan, D. Janzing, and G. Mahler, J. Stat. Mech. (2009) P09011.
[35] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).
[36] J. M. Horowitz and M. Esposito, Phys. Rev. X **4**, 031015 (2014).
[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.118.010601, which includes Ref. [38], for a detailed derivation.
[38] J. M. Horowitz and H. Sandberg, New J. Phys. **16**, 125007 (2014).
[39] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Approach* (John Wiley & Sons, New York, 1949).
[40] D. Abreu and U. Seifert, Europhys. Lett. **94**, 10001 (2011).
[41] H. Risken, *The Fokker-Planck Equation* (Springer, New York, 1996).
[42] In the limit of large N, only the first two moments of the distribution will matter, making this choice equivalent to sampling $\boldsymbol{\xi}^{\mu}$ from the surface of a hypersphere in N dimensions in that limit.
[43] 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 [28].
[44] E. Gardner, Europhys. Lett. **4**, 481 (1987).
[45] D. Hartich, A. C. Barato, and U. Seifert, Phys. Rev. E **93**, 022116 (2016).
[46] N. Brunel, V. Hakim, P. Isope, J.-P. Nadal, and B. Barbour, Neuron **43**, 745 (2004).
[47] B. Barbour, N. Brunel, V. Hakim, and J.-P. Nadal, Trends Neurosci. **30**, 622 (2007).
[48] G. Tkačik, A. M. Walczak, and W. Bialek, Phys. Rev. E **80**, 031920 (2009).
[49] T. R. Sokolowski and G. Tkačik, Phys. Rev. E **91**, 062710 (2015).