Machine learning at the mesoscale: A computation-dissipation bottleneck

Alessandro Ingrosso and Emanuele Panizon

Quantitative Life Sciences, Abdus Salam International Centre for Theoretical Physics, 34151 Trieste, Italy

(Received 6 July 2023; revised 29 November 2023; accepted 5 December 2023; published 23 January 2024)

The cost of information processing in physical systems calls for a trade-off between performance and energetic expenditure. Here we formulate and study a computation-dissipation bottleneck in mesoscopic systems used as input-output devices. Using both real data sets and synthetic tasks, we show how nonequilibrium leads to enhanced performance. Our framework sheds light on a crucial compromise between information compression, input-output computation and dynamic irreversibility induced by nonreciprocal interactions.

DOI: 10.1103/PhysRevE.109.014132

I. INTRODUCTION

What does computation at the mesoscopic scale look like? To begin to answer this question, we need to bridge the formalism of computation with a theory of systems whose energy scales are close to thermal fluctuations. Stochastic thermodynamics (ST), by associating single stochastic trajectories with meaningful thermodynamic quantities [1–4], exposes the deep relation between information and dissipation. One of the fundamental results of ST is that information and time irreversibility, as measured by the rate of entropy production (EP) [5,6], are inherently related [7–11]. Thermodynamic uncertainty relations [12–15] have been derived that describe fundamental precision-dissipation trade-offs, leading to a framework successfully applied to a variety of biochemical processes, such as chemosensing [16–18], copying [19], and reaction networks [12,20,21], among others.

We will refer to computation at the mesoscopic scale as the ability of a system to react to the environment—via interactions between its parts and external heat baths—in a way that depends on some function of the environmental conditions. The space of all states reachable by these transformations, and the details of their distributions, differ in the case of equilibrium and out-of-equilibrium dynamics. In particular, the latter, larger space of transformations (which includes equilibrium relaxations) affords more refined input-output mappings and precise reactions, at the cost of maintaining the system out of equilibrium.

Encoding external signals in their entirety is one of such computations: borrowing machine learning (ML) terminology, a mesoscopic system considered as an "autoencoder" has the ability to compress information and correct errors [22,23].

Full encoding, however, may be wasteful when a computation regards a limited aspect of the environment: discarding nonrelevant information, i.e., limiting the necessary environment-system mapping in a manner dependent on the task at hand, allows one to strike a balance between performance and energy expenditure.

We recognize this task dependence of the performance-cost trade-off as the main ingredient of any physical theory of computation. Can this trade-off be framed in quantitative terms? How could it be calculated? Is it possible to predict, or understand, nonequilibrium energy expenditure from the structure of the task? These are the questions we try to address in this work.

To do so, we define a quantitative trade-off between computation and performance. We exploit such formal definition to study analytically and numerically a set of paradigmatic cases where physical systems have to "learn" different regression and classification tasks.

To set up the theoretical framework, we bridge the two extrema of the trade-off. On one side of it lies dissipation, which we will measure in terms of entropy production rate. The study of EP in many-body systems has recently started to be addressed [14,24–26]. Irreversibility of macroscopic neural dynamics is also attracting attention [27–30].

The system's computational performance, on the other side, can be formulated both in information theoretic terms and with standard error metrics employed in ML. Learning dynamics in simple classifiers has been studied using the machinery of stochastic thermodynamics [31]. A recent approach introduced a framework for irreversibility in formal models of computation [32–34], without specifying the details of physical implementations.

Here we consider generic parametrizations of mesoscopic systems whose stochastic transitions are induced by an environment, possibly out of equilibrium, so that resulting interactions may be nonreciprocal [35]. We focus on asymmetric spin models, which have been subject of intense study in the field of disordered systems [36–39] and provide a bridge to classical models of neural computation [40–44].

In line with the neural network formalism, we recognize the dynamics of our systems as producing internal representations of their inputs, the geometry of which impacts the ability to learn input-output relations. We show how entropy-producing nonreciprocal interactions [45,46] are crucial to generate effective representations, so that a fundamental trade-off emerges between expressivity and performance.

II. METHODS

A. A computation-dissipation bottleneck

Here we introduce our framework for using mesoscopic systems as input-output devices in supervised input-output tasks. We thus formulate a tradeoff between the computational performance and steady-state entropy production of such systems in terms of a computation-dissipation bottleneck.

The stochastic dynamics of mesoscopic systems, usually described using continuous-time Markov processes, results from interactions with thermal baths and external driving mechanisms. Let us consider a system $\mathcal{S}$ with discrete states s, driven by a time homogeneous input protocol x. The evolution of the probability of state $p(s,t)$ is given by a master equation,

$$\frac{d}{dt}p(s,t) = \sum_{s'}[k_{ss'}p(s',t) - k_{s's}p(s,t)], \tag{1}$$

with $k_{ss'}$ the jump rate from s' to s, generically dependent on the protocol x. To facilitate the connection to ML, we consider the jump rates to be determined by a set of parameters θ.

First, we have to associate an energy cost to the computation. To do so, we use the entropy production rate. We assume computation is performed on a timescale much longer than any initial transient. For each independent input x, the system reaches a steady-state (SS) probability $p(s|x)$, serving as an internal representation of x. At the (possibly nonequilibrium) SS, each input x is associated to an average EP rate σ. In Markovian systems with discrete states, the EP rate can be computed via the Schnakenberg formula [47,48]:

$$\sigma = \frac{1}{2}\sum_{s,s'}\mathcal{J}_{ss'}\log\frac{k_{ss'}p(s')}{k_{s's}p(s)}, \tag{2}$$

where $\mathcal{J}_{ss'} = [k_{ss'}\,p(s') - k_{s's}\,p(s)]$ are the steady state fluxes (we work in units where the Boltzmann constant $\kappa_B = 1$). Note that in our case $\sigma = \sigma(x,\theta)$ through $k_{ss'}$, $\mathcal{J}_{ss'}$ and $p(s)$.

A supervised learning task is specified by a finite set $\mathcal{D} = (x, y)$ of input-output pairs, so the EP rate averaged over the whole data set is simply $\Sigma(\theta) = \frac{1}{|\mathcal{D}|}\sum_x \sigma(x,\theta)$. Alternatively, one can define a joint input-output distribution $p(x, y)$. The (average) EP rate is similarly $\Sigma(\theta) = \sum_x p(x)\sigma(x,\theta)$.

The EP rate is a function of the dynamic process alone. How the resulting $p(s|x)$ is able to predict the output is a separate, task-specific factor. To quantify the computational performance of these processes, we have to define a measure for that.

In defining such performance measure, a natural choice is the mutual information $I(s, y)$ between the representation s and the output y: no assumption is made in this case on the additional computational burden needed to extract such information, possibly encapsulated in arbitrarily complex high-order statistics of the steady-state distribution. For many problems of relevance, and especially for those related to standard ML practice, evaluating the mutual information is unfeasible. In these case, we will follow a different path and use a subset of moments of $p(s|x)$ as representations to be then fed to a linear readout. This approach to computation, closer to ML practice, allows us to use the mean square error (MSE) or cross-entropy (CE) loss functions as approximate surrogates for $\mathcal{G}$. Both approaches and their limitations will be explored in the following.

Given a performance measure $\mathcal{G}(\theta)$, the trade-off can be encapsulated in a quantity:

$$\mathcal{L}(\theta) = \mathcal{G}(\theta) - \alpha\,\Sigma(\theta), \tag{3}$$

where α is a positive parameter with units of time. We study the trade-off by maximizing $\mathcal{L}$ over the interaction parameters θ for different values of α: increasing α, the relative cost of dissipation is enhanced, with the $\alpha \to \infty$ limit constraining the system to be at equilibrium.

In this work we first employ our framework in the context of a solvable two-spin model to show how the enhanced expressivity of nonequilibrium systems is related to the structure of the input-output tasks. Then we use computational methods to build a multispin system performing classification tasks.

B. Entropy production at steady state

In the presence of a constant-in-time protocol x, the steady state $p(s|x)$ can be obtained extracting the kernel of the matrix $R_{ss'} = k_{ss'} - \delta_{s,s'}\sum_{s''}k_{s''s}$. For systems of small size, this is viable numerically using singular value decomposition (SVD). The steady-state entropy production can thus be computed directly using Eq. (2).

In a nonlinear large-scale system, analytical calculation of the steady state is generally unfeasible. We thus simulate the stochastic dynamics of the system $\mathcal{S}$ using the Gillespie algorithm [49,50]. To do so, we generate stochastic trajectories by concatenating random jumps between states, obtained by first identifying the exit time from a given state s' and then selecting jumps according to the transition rates $k_{ss'}$.

The entropy production can be easily computed in an online fashion by accumulating the logarithmic ratio of forward and backward transition rates for each jump, which are represented by single-spin flips in our models. Further details are given in the Appendix. In analogy with ML, we consider the Gillespie simulation as a forward pass in a stochastic network, retrieving a probability over states $p(s|x)$ when presented with an input x.

C. Task performance measures

Here we define the performance measures and their quantitative evaluation used throughout the paper.

In Sec. III A we will consider a simple two-dimensional toy model. For this system, we will take $\mathcal{G}$ to be the mutual information, which can be evaluated exactly. The mutual information between the input x and the system state s at steady state can be computed using $I(s,x) = H(x) - H(s|x)$, with H the Shannon entropy. As for $I(s,y) = H(s) - H(s|y)$, the entropy term $H(s|y)$ can be obtained by exploiting the conditional independence between y and s, which implies that the joint distribution $p(s, y)$ can be written as

$$p(s,y) = \sum_{x,s,y} p(s|x)p(y|x)p(x). \tag{4}$$

The posterior distribution $p(s|y)$ is then directly calculated from Eq. (4) using Bayes' theorem.

In Sec. III B we will consider large-scale systems. For these, we resort to only using the average value for s (spin magnetization) and subsequently extract the information about the output using a linear readout. More explicitly, given an input-output pair (x^μ, y^μ) from the set $\mathcal{D} = (x, y)$, we measure task performance by first computing the vector

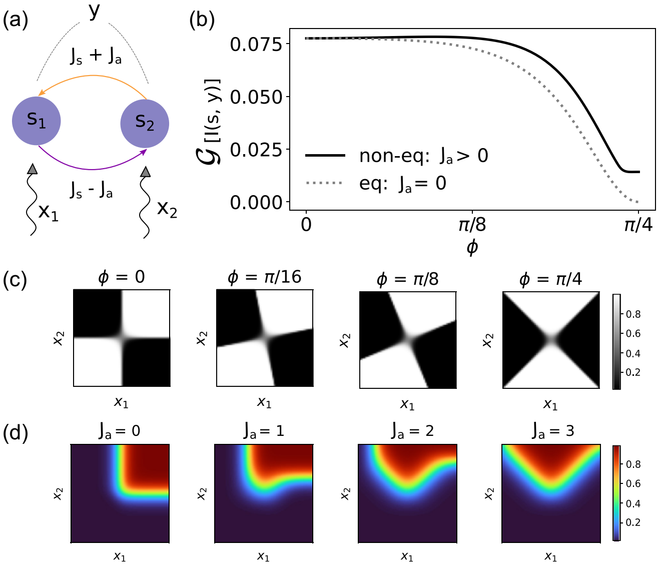

FIG. 1. (a) Schematic of a two-spin system driven by external fields x_1 and x_2. (b) Mutual information between input x and output y as a function of rotation angle ϕ. The optima for equilibrium (nonequilibrium) systems are shown with a dashed gray (continuous black) line. Parameters: $\beta = 3$, $\eta = 3$. Inputs (x_1, x_2) are Gaussian with correlations $\rho = 0.95$. (c) Conditional probability $p(y = 1|x)$ for the family of tasks described in the paper for different values of the parameter ϕ and $\eta = 2$. (d) Steady-state probability for the state $s = (+1, +1)$ for $J_s = 0$ and increasing values of the nonreciprocity strength J_a. The range is $[-2.5, 2.5]$ for both inputs x_1 and x_2.

$m_{x^\mu} = \langle s|x^\mu\rangle$ and then the error between the prediction $\hat{y}^\mu$ of the final readout and the target y^μ.

We evaluate the performance using classical loss functions employed in ML. When using the cross-entropy (CE) loss, a "logit" vector $h^\mu = W_{\text{out}} m_{x^\mu}$ is passed through a normalized exponential function (Softmax), thus getting the normalized estimated output probabilities $p_k^\mu = \frac{e^{h_k^\mu}}{\sum_{l=1}^{K} e^{h_l^\mu}}$, with K the number of output labels. The loss function then amounts at computing the cross-entropy with the targets y^μ: $L = -\frac{1}{M}\sum_{\mu=1}^{M} \log p_{y^\mu}^\mu$. The mean square error (MSE) loss, in turn, amounts at computing $L = \frac{1}{2M}\sum_{\mu=1}^{M}(y^\mu - W_{\text{out}} m_{x^\mu})^2$.

III. RESULTS

A. A tractable two-spin system

To exemplify the computation-dissipation trade-off, we first consider an analytically solvable two-spin system. Each spin s_i is subject to random flips with rates $k_s^{(i)} \propto e^{-\beta s_i (Ws+x)_i}$. The matrix W encodes the spin interactions in the asymmetric couplings $W_{12} = J_s + J_a$, $W_{21} = J_s - J_a$, and two-dimensional inputs x act as constant external fields [Fig. 1(a)]. When $J_a = 0$, the system respects detailed balance and reaches equilibrium. Nonreciprocity in the couplings leads to non-negative Σ.

The information-coding capabilities at steady state of this system have been recently analyzed [18]. We treat such a mesoscopic network as an input-output device, prescribing a stochastic rule by a known conditional distribution $p(y|x)$, with $y \in \{0, 1\}$ a binary output variable. This formulation encompasses the classic teacher-student setup [51–54] and mixture models [55,56] used in the study of feed-forward neural networks. We ask how much information $I(s, y)$ about the output y is contained in the steady state $p(s|x)$.

Let us consider a task consisting in a stochastic and continuous generalization of a parity gate, $p(y = 1|x) = \text{sigmoid}(\eta x_1^\phi x_2^\phi)$, with $x^\phi = R^\phi x$, R^ϕ a rotation of angle ϕ. This angle defines a family of tasks with a controllable degree of asymmetry in input space. Examples are shown in Fig. 1(c). The parameter η affects the sharpness in the change of the output probability as a function of x.

For $\phi = 0$, the optimal structure is an equilibrium system ($J_a^\star = 0$). As ϕ increases, the optimal two-spin network has asymmetric weights ($J_a^\star > 0$), implying a nonzero entropy production at steady state; see Fig. 1(b). Limiting the system to be at equilibrium thus results in performance degradation, down to a minimum of zero information when the rotation reaches $\phi = \pi/4$.

For a given value of ϕ and the free parameter α, one can define the computation-dissipation trade-off by maximizing Eq. (3) with $\mathcal{G} = I(s, y)$. Note the analogy with the formulation of the classic information bottleneck [57–59]. Here, instead of a compromise between input compression and retention of output information, we trade off the latter with dissipation.

We can compare the performance of an auto-encoding system with optimal couplings $\theta^{sx} = \{J_s^{sx}, J_a^{sx}\}$ ($\mathcal{G} = I(s, x\,|\,\theta)$), with that of a computing system with parameters θ^{sy} ($\mathcal{G} = I(s, y\,|\,\theta)$) for a task with $\phi = 0.5$. The optima corresponding to $\alpha = 0$ have finite nonreciprocal terms J_a [see Figs. 2(a) and 2(b)] and therefore positive, but finite, EP. We always find a maximum dissipation rate above which performance degrades [60].

Figure 2(c) shows the computation-dissipation front, each point being a different optimal compromise between input-output performance, measured by $\mathcal{G} = I(s, y\,|\,\theta)$, and Σ at steady state. We chose a parameter regime where a nonequilibrium solution is optimal also for $I(s, x)$. Crucially, a system maximizing the information on the entire input $I(s, x)$ performs worse than one tailored to maximize the output information. This is a hallmark of optimization of task-relevant information.

We now explore the relation between the nonequilibrium steady-state probability $p(s|x)$ and the task. Fixing $J_s = 0$, the effect of increasing J_a resembles a rotation by $\pi/4$ of $p(s|x)$ in the region where $|x| < J_a$; see Fig. 1(d). Increasing the nonreciprocity thus aligns the steady-state probabilities $p(s|x)$ with the rotation induced on the task by the angle parameter ϕ.

B. Multispin systems as stochastic recurrent networks

We now study the trade-off in computational tasks more akin to that of standard ML practice: we use a spin model performing an input-output computation in the form of a classification task where inputs x, schematically represented by the tape in Fig. 3, must be correctly associated with output labels y. We relax the previous requirement to control and measure the mutual information between the distribution of spins and

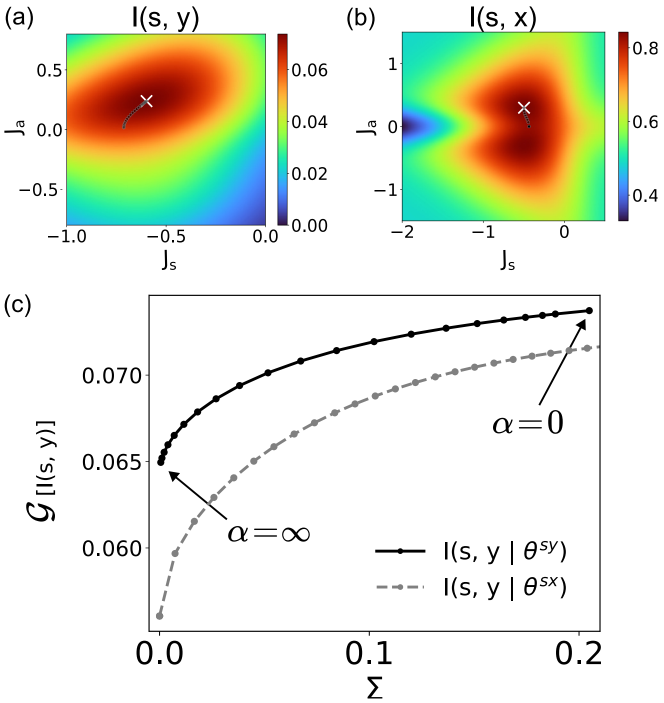

FIG. 2. (a) Color plot of mutual information $I(x, s)$ in the $\{J_s, J_a\}$ plane. The optimal parameter set θ^{sx} is shown for different values of α (white: $\alpha = 0$, black: $\alpha > 0$). (b) Same as (a) for $I(s, y)$ and the optimal parameter set θ^{sy}. (c) Mutual information $I(s, y)$ for $\phi = 0.5$ as a function of the entropy production rate at steady state Σ for both θ^{sy} (black) and θ^{sx} (gray). Inputs (x_1, x_2) are Gaussian with correlations $\rho = 0.95$. Additional parameters: $\beta = 3$, $\eta = 3$.

output labels, and, as detailed in the following, instead introduce simpler error measures to evaluate performance.

We construct a system that realizes a stochastic equivalent of a simple convolutional neural network with two small filters (kernels). Each filter is composed of a chain of N spins, interacting with possibly asymmetric couplings. Spins in the two chains are driven by the same constant external fields x_i (with a slight abuse of notation, we consider x as composed of two identical copies of the same input vector). As before, each spin s_i is subject to random flips with rates $k_s^{(i)} \propto e^{-\beta s_i (W s + x)_i}$. Interactions, encoded in the matrix W, connect spins along each chain and depend only on the relative distance between spins and not on their absolute locations.

When W is fully symmetric ($W = W^T$), the system relaxes to the equilibrium of a Hamiltonian $\mathcal{H} = -\frac{1}{2} s^T W s - x$ at inverse temperature β. Nonreciprocal interactions ($W \neq W^T$) lead to nonequilibrium and a nonzero EP rate. After a transient, the system reaches a steady state $p(s|x)$, with an average magnetization $m_x = \langle s|x \rangle$ and an entropy production rate $\sigma(x, \theta)$. For any data set $\mathcal{D}$, each W will thus be associated with both a different task performance and an average EP rate Σ.

In close analogy with ML, we use a final linear readout $W_{\text{out}} m_x$ of the average magnetization, with a learnable matrix W_{out}. This allows us to separately consider the system's computation as a two-step process: (1) a nonlinear deformation of the input space x into m_x induced by the dynamics, akin to what occurs in the hidden layers of a neural network (due to the choice of translational symmetry for W, the steady-state magnetizations in each chain are equivariant to translation in the input x, so that the system is a stochastic, mesoscopic version of an implicit convolutional layer [61,62], see the Appendix) and (2) a separation in the m_x space producing the output y.

To maximize $\mathcal{L}$ over $\theta = \{W, W_{\text{out}}\}$, we couple a standard Gillespie algorithm for the simulation of the system's evolution with each input x, to a gradient-based optimization. Due to the stochastic nature of the trajectory, the use of standard back-propagation is not possible in our context, and finite-difference approximations for the gradients are required. In order to deal with the high dimension of the W parameter space, we adopted an efficient method called the simultaneous perturbation stochastic approximation (SPSA) [63] to compute an estimate of the gradient (see the Appendix for details). The solutions at each value of α allow us to construct an optimal front $\mathcal{G}^*(\Sigma)$, where the asterisk denotes that optimal values of Eq. (3), as shown in Figs. 4(a) and 4(c).

We showcase our approach with two different tasks. The first is MNIST-1D [64], a one-dimensional version of the classic digit-classification MNIST task. Each element, with input dimension $N = 40$, belongs to one of 10 different classes. See an example of the input configurations in Fig. 4(b). To enable multilabel classification, we apply a Softmax function SM to the output, thus getting a 10-dimensional probability vector $\hat{y} = \text{SM}(W_{\text{out}} m_x)$, and measure task performance with the negative cross-entropy $\mathcal{G} = -\text{CE}(\hat{y}, y)$ between the labels and $\hat{y}$.

Our results show a direct relation between task performance and entropy production at steady state [Fig. 4(a)]. Enforcing the system to be at equilibrium ($\alpha \to \infty$) reduces performance by ≈5% and accuracy—defined as the percentage of labels correctly identified—by 7%. This highlights how nonreciprocal interactions enhance the complexity of internal representations needed for learning (see Fig. 5 in the Appendix), at the cost of higher dissipation.

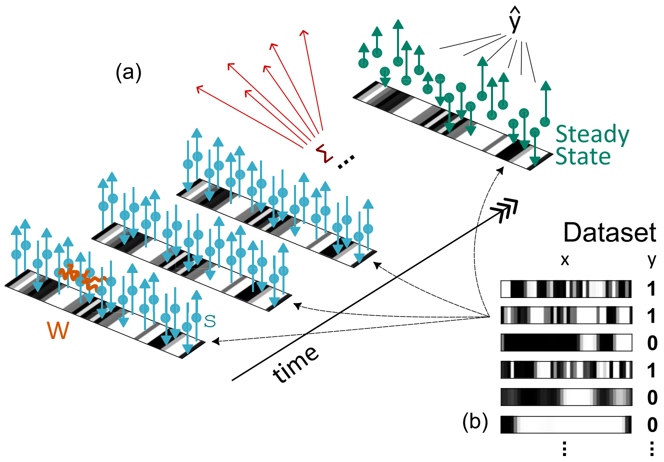

FIG. 3. Schematic of a multispin system processing inputs in a classification task. (a) The system $\mathcal{S}$ evolves in time in the presence of constant inputs (external fields) x and couplings W, until a nonequilibrium steady state $p(s|x)$ is reached. Time evolution is associated with an entropy production rate Σ. Information about the output label is extracted from $p(s|x)$, e.g., by a linear readout W_{out} on the averages m_x. (b) A subset of an input-output data set $\mathcal{D}$.

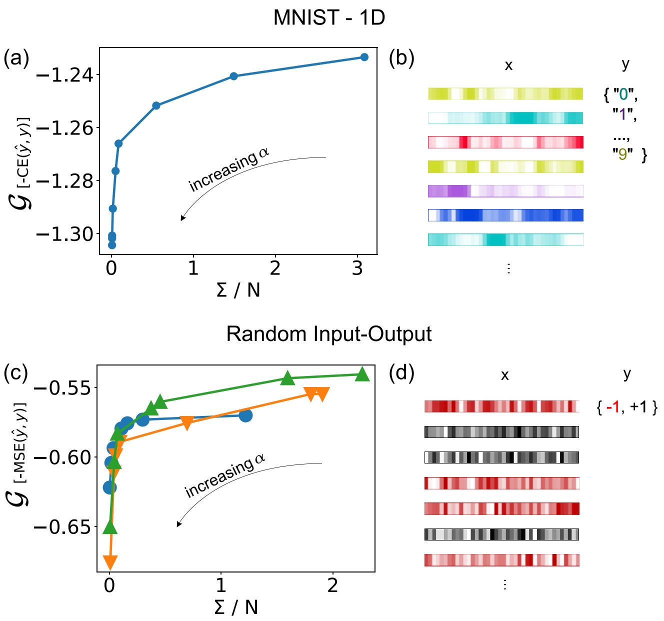

FIG. 4. Computation-dissipation bottleneck in a system solving a classification task at steady state. (a) Cross-entropy error (CE) vs normalized entropy production rate Σ/N on the MNIST-1D data set ($M = 4000$ data points, 10 labels, $N = 40$) of a system composed of two spin chains with interactions up to the second nearest neighbor. (b) Schematic of the MNIST-1D task. (c) Mean square error (MSE) vs entropy production rate Σ/N normalized by the number of spins on the random input-output task with $M = 100$ input patterns and two labels. Each curve is the minimum over 10 initialization seeds in a Gillespie-based optimization in three different realizations of the task. (d) Example realizations of the random data set.

The second task is a classic random input-output association [65–67], where input components x_i^μ of each pattern $\mu = 1, \dots, M$ are drawn i.i.d. from a Normal distribution, and labels are random $y \in \{-1, +1\}$ with probability 1/2 [Fig. 4(d)]. We measure the performance by the mean squared error (MSE): $\mathcal{G} = -\mathrm{MSE}(\hat{y}, y)$, where $\hat{y} = W_{\mathrm{out}} m_x$. For all random instances of this task, we reproduce the front between entropy production and performance [Fig. 4(c)]. While quantitative details differ slightly among instances, the performance consistently increases with the amount of nonreciprocity in the optimal coupling matrix W, and therefore with dissipation in the system.

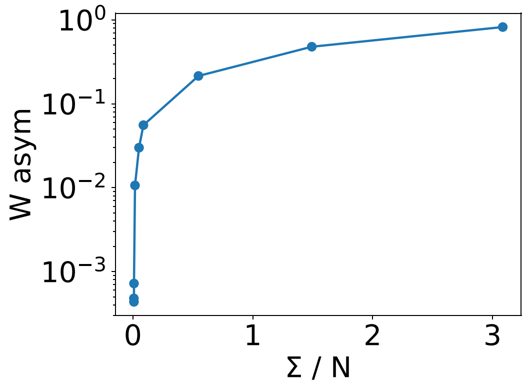

FIG. 5. Normalized measure of asymmetry of interaction matrix W found by SPSA as a function of entropy production Σ for the MNIST-1D task [Fig. 4(a)].

IV. CONCLUSIONS

We introduce a framework to characterize a trade-off between computational capabilities and dissipation in mesoscopic systems. We showcase how such systems can be used in supervised learning tasks with real data sets and how limiting entropy production degrades their performance.

Our results point to the general necessity to gauge encoding and task relevance while considering energetic trade-offs. In a tractable two-spin system, we show how nonreciprocal interactions affect the capability of the system to solve tasks optimally. A simple modulation of the task switches its optimal configuration from an equilibrium to a highly nonequilibrium one.

Linear stochastic systems are another case where analytical expressions can be derived for the computation-dissipation trade-off (see the Appendix), again controlled by the asymmetry of the task in input space.

In this study, we concentrated on one-time statistics at steady state, leaving aside interesting properties of time correlations. The study of both transient behavior and nonstationary protocols—where special care must be used in distinguishing between housekeeping and excess entropy production [68]—opens an interesting avenue to investigate general speed-dissipation-computation trade-offs within our framework.

Studying the impact of hidden units is an important avenue for future work. Marginalization over hidden states is the main ingredient to induce higher-order interactions in generative models. This forms the basis for the attention mechanism in transformers [69], arguably the most powerful ML models to date [70,71], as the recent works on modern Hopfield networks have shown [72–76].

Drawing a bridge between ML and ST can prove fruitful in elucidating how representations depend on the cost. Rate-distortion approaches have been used to study the impact of information compression on classification accuracy and maximal attainable rewards [77–82], but a general theory is lacking. Our perspective is complementary: energetic costs have a strong impact on the complexity of internal representations, leading to different mechanisms for information processing.

ACKNOWLEDGMENTS

We wish to thank Antonio Celani, Roman Belousov, and Edgar Roldan for fruitful discussions and for reading a preliminary version of this manuscript.

APPENDIX

1. Steady state in the two-spin system

Following [18], the stationary state can be computed by imposing the stationary condition in Eq. (1) and the normalization of $p(s|x)$, thus getting

$$p(s|x) = e^{-\beta(F+\delta F)}/Z_x, \tag{A1}$$

where

$$F = -J_s s_1 s_2 - x_1 s_1 - x_2 s_2 \tag{A2}$$

and

$$\delta F = -\beta^{-1} \log \left[e^{\beta J_a s_1 s_2} \frac{\cosh [\beta (x_1 - 2J_a s_2)]}{\cosh \beta x_1 + \cosh \beta x_2} \right. \tag{A3}$$

$$\left. + e^{-\beta J_a s_1 s_2} \frac{\cosh [\beta (x_2 + 2J_a s_1)]}{\cosh \beta x_1 + \cosh \beta x_2} \right]. \tag{A4}$$

2. Training of a multispin system

a. Details on the system

We consider a system composed of two chains of size N. Interactions connect spins up to the kth neighbors, where we use $k = 2$. Self-interactions are set to zero. If we identify a spin by (m, n) where $1 \leqslant m \leqslant N$ is the position in the chain and $n = 1, 2$ the chain index, two spins (m_i, n_i) and (m_j, n_j) are connected if $|m_i - m_j| \leqslant k$. The interaction parameter W_{ij} depends only on $m_i - m_j$ and $n_i - n_j$, so that the number of nonzero, fully independent parameters of W is $4k - 2$. The external input x is repeated such that it is the same for both chains. Such a spin system at steady state implements a stochastic version of an implicit convolutional layer with two channels. Implicit layers are building blocks of deep equilibrium models [61,62].

b. Data sets

MNIST-1D is a one-dimensional version of size $N = 40$ of the classic MNIST handwritten digits data set [64]. We used 4000 training samples, organized in 10 different classes, each containing roughly 400 samples. Data are available at [83], where a description of its generation from the original MNSIT data set is given.

We generated instances of the random task by drawing $M = 100$ patterns x^μ in dimension $N = 10$, with components x_i^μ independently drawn from a Normal distribution. The corresponding labels y^μ, drawn from $\{-1, +1\}$ with probability $1/2$, were randomly associated with each pattern.

c. Details of Gillespie simulations

Let us consider a system with a discrete number of states s and transition rates $k_{ss'}$, which are constant in time. Given a current state s_{start}, the Gillespie algorithm identifies both the time τ and the final state s_{end} of the following jump.

As a first step, the total rate $k_{\text{out}} = \sum_s k_{ss_{\text{start}}}$ of leaving state s_{start} is computed. The time τ until the following jump is then drawn from an exponential distribution with mean $1/k_{\text{out}}$. The landing state is selected with probability $p(s) = k_{ss_{\text{start}}}/k_{\text{out}}$. The trajectory is thus constructed concatenating jumps.

First, the initial state s_0 is chosen (in our case, at random) at time $t = 0$. A first jump (τ_1, s_1) is selected starting from s_0, and then a second (τ_2, s_2) starting from s_1. The process is repeated until one of two criteria is met, either a total time or a maximum number of steps. Average occupations can be computed considering that the system occupies state s_i exactly for a time τ_i between jumps i and $i + 1$.

In our system, s is a vector of $2N$ individual spins s_i taking values in $\{-1, +1\}$. We will restrict the jumps to single spin flips. Given a state s, an input x (external field) and a interaction matrix W, the transition where the ith spin flips has a rate $k_s^{(i)} \propto e^{-\beta s_i h_i}$, with $h_i = (Ws + x)_i$. The actual proportionality term (identical for all spins), which determines the timescale of the jumps, is not relevant since we are interested only in steady-state properties and average occupancy.

To measure the average magnetization m_x for each input x, we first select a random state s_0 and proceed to construct a trajectory up to a final time $T_{\max} = 5000$ or, alternatively, a maximum number of jumps $N_{\max} = 10\,000$. The average magnetization of individual spins m_x for that input is calculated after an initial transient time of $T_{\text{transient}} = 200$ is removed.

Since we consider only the steady state, we can evaluate the entropy production rate by summing the quantity $\Delta\sigma_n \equiv \log \frac{k_{s_{n+1}}^{(i)}}{k_{s_n}^{(i)}} = -2\beta s_{n,i} h_{n,i}$ for each jump $s_n \rightarrow s_{n+1}$, consisting of a single spin flip, and dividing by the total time [84].

d. Parameter optimization

The minimization of a loss $-\mathcal{G}$ with respect to W_{out} was performed either via a linear solver (for MSE) or a multinomial classifier solver (for CE), using standard libraries in julia, which retrieve optimal W_{out}^* at fixed W, for the full input set. We used MSE loss for the binary classification in the random task, whereas we employed the CE loss for multilabel classification in the MNIST-1D task.

Due to the stochastic nature of the dynamics, the optimization of the interaction parameters W cannot be performed with standard gradient-based methods. Additionally, typical gradient evaluation through finite difference quickly becomes prohibitive as the number of independent parameters in W grows. To overcome this issue, we employ simultaneous perturbation stochastic approximation (SPSA) [85,86], where the gradient is approximated via a single finite difference in a random direction of the parameter space.

To evaluate the gradient $\nabla \mathcal{L}|_W$, a random vector δW is constructed at every update step. Two symmetrical parameters configurations are constructed: $W^\pm = W \pm \delta W$. Independent dynamics are simulated to produce the average spin magnetizations $m^\pm$ and measure entropy production rates $\Sigma^\pm$. The average magnetizations $m^\pm$ are thus used to compute the performances $\mathcal{G}^\pm$. Finally, the gradient approximation reads $\nabla \mathcal{L}|_W \approx [\mathcal{G}^+ - \mathcal{G}^- - \alpha(\Sigma^+ - \Sigma^-)] \frac{\delta W}{2|\delta W|}$. To avoid being trapped into local maxima, we performed several initializations for each value of α.

We performed preliminary checks in systems of small size—where the computation of the steady-state distributions is viable via SVD—and confirmed that SPSA converges to a global optima of $\mathcal{L}$, obtained by explicit parameter enumeration.

e. Relation between nonreciprocity and entropy production

We report in Fig. 5 an example of the relation between asymmetry of interactions and entropy production at steady state in the spin system from Sec. III B. The normalized asymmetry A is computed as $A = \frac{\sum_{ij} (W_{ij} - W_{ji})^2}{\sum_{ij} W_{ij}^2}$.

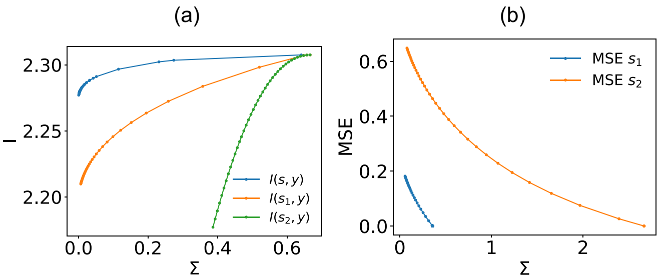

FIG. 6. (a) Optimal mutual information $I(s, y)$ between the system state s and the output y (blue curve) as a function of the entropy production rate at steady state Σ in a two-particle linear system described in Sec. III A. We also show the single-particle mutual information $I(s_1, y)$ and $I(s_2, y)$ in the orange and green curve, respectively. (b) Same as in (a) for the optimal squared deviation MSE when the output y is linearly read out at each unit s_i, namely, $\langle (y - s_i)^2 \rangle$.

3. Computation-dissipation bottleneck in linear systems

Let us consider a system whose dynamics, in the presence of a constant input x, is described by a multidimensional Ornstein-Uhlenbeck process:

$$\dot{s} = Ws + x + \sigma_s \xi \tag{A5}$$

with $\langle \xi \xi^T \rangle = \delta(t - t')\mathcal{I}$ and $\mathcal{I}$ the identity matrix. The (generally nonequilibrium) steady state distribution $p(s|x)$ is a Gaussian with mean $m_x = W^{-1}x$ and whose covariance C solves the Lyapunov equation:

$$WC + CW^T + \sigma_s^2 \mathcal{I} = 0. \tag{A6}$$

Let us consider a noisy linear function $y = w_0^T x + \xi_y$, with $\langle \xi_y \rangle = 0$ and $\sigma_y^2 = \langle \xi_y^2 \rangle$. Assuming x is a Gaussian with mean zero and covariance C_x, one has $C_y = \langle y^2 \rangle = w_0^T C_x w_0 + \sigma_y^2$ and $C_{sy} = \langle sy \rangle = -W^{-1}\langle xy \rangle$.

To compute the mutual information, we use

$$I(s, y) = H(s) - H(s|y) \tag{A7}$$

and the relation for the entropy of a zero-mean, d dimensional Gaussian variable z with covariance C_z, $H(z) = \frac{1}{2}\log((2\pi e)^d \det C_z)$, to get

$$\begin{aligned} I(s, y) = {} & \frac{1}{2}\log\det(W^{-1}C_x W^{-T} + C) \\ & - \frac{1}{2}\log\det\left(C_s - C_{sy}C_y^{-1}C_{ys}\right), \end{aligned} \tag{A8}$$

where we used the fact that the covariance matrix $C_s = \langle ss^T \rangle$, averaged over the entire input distribution, equals $C_s = W^{-1}C_x W^{-T} + C$ and that the conditional covariance matrix of s given y is $C_{s|y} = C_s - C_{sy}C_y^{-1}C_{ys}$.

As shown in [87], the entropy production can be computed in terms of an integral

$$\sigma = \int_{-\infty}^{+\infty} \frac{d\omega}{2\pi}\mathcal{E}(\omega), \tag{A9}$$

where the density $\mathcal{E}(\omega)$ is given by

$$\mathcal{E}(\omega) = \tfrac{1}{2}\mathrm{Tr}\{C(\omega)[C^{-1}(-\omega) - C^{-1}(\omega)]\}, \tag{A10}$$

with $C(\omega)$ the Fourier transform of the steady state autocorrelation $C(t - t') = \langle s(t)s^T(t') \rangle$.

The expressions derived thus far can be used to obtain the computation-dissipation bottleneck by optimization over any physically consistent parametrization of the coupling matrix W for a stable system, with different values of the tradeoff parameter α. To exemplify the approach, the next section treats a two-dimensional case where simple analytical expressions can be derived and a full enumeration of the parameter space is viable.

a. An example of a computation-dissipation bottleneck in a two-dimensional linear case

Let us then consider the case of a two-particle linear system with an interaction matrix of the form

$$W = \begin{pmatrix} -1 & J_s + J_a \\ J_s - J_a & -1 \end{pmatrix}. \tag{A11}$$

Stability is guaranteed for $\Delta = 1 + J_a^2 - J_s^2 > 0$. The solution of the Lyapunov equation (A6) for an input noise with covariance $\sigma_s^2 \mathcal{I}$ is

$$C = \frac{\sigma_s^2}{2\Delta}\begin{pmatrix} 1 + J_s J_a + J_a^2 & J_s \\ J_s & 1 - J_s J_a + J_a^2 \end{pmatrix}. \tag{A12}$$

The entropy production can be evaluated using Eq. (A10) and the Fourier transform of the system's Green's function $G(\omega) = (i\omega - W)^{-1}$:

$$G(\omega) = \frac{1}{\Delta - \omega^2 + 2i\omega}\begin{pmatrix} 1 + i\omega & J_s + J_a \\ J_s - J_a & 1 + i\omega \end{pmatrix}. \tag{A13}$$

From the Fourier transform of the steady-state autocorrelation $C(\omega) = G(\omega)G^{\dagger}(\omega)$ we get for the entropy production density

$$\mathcal{E}(\omega) = \frac{8\omega^2 J_a^2}{\left|(1 + i\omega)^2 + J_a^2 - J_s^2\right|^2}. \tag{A14}$$

After integration in Eq. (A9), and noting that C doesn't depend on x, we get for a stable system

$$\Sigma = 2J_a^2. \tag{A15}$$

We show in Fig. 6 the results for a system with $\sigma_s = 0.1$ tasked to compute a linear function $y = w_0^T x + \xi_y$ with $w_0 = (\cos\phi, \sin\phi)$, with $\phi = \frac{\pi}{6}$ and ξ_y a zero-mean Gaussian variable with standard deviation $\sigma_y = 0.1$. In a similar vein, each particle s_i can be used as a direct readout for the output y. In such a case, the average squared deviation $\mathrm{MSE}_i = \langle (y - s_i)^2 \rangle$ at steady state again shows a characteristic front with respect to entropy production. As discussed in the main text, the trade-off between entropy production and output information is again controlled by the angle ϕ, that sets the degree of asymmetry in the input space.

[1] U. Seifert, Entropy production along a stochastic trajectory and an integral fluctuation theorem, Phys. Rev. Lett. **95**, 040602 (2005).

[2] U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines, Rep. Prog. Phys. **75**, 126001 (2012).

[3] C. Van den Broeck and M. Esposito, Ensemble and trajectory thermodynamics: A brief introduction, Physica A **418**, 6 (2015).

[4] L. Peliti and S. Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, Princeton, 2021).

[5] D. Andrieux, P. Gaspard, S. Ciliberto, N. Garnier, S. Joubaud, and A. Petrosyan, Entropy production and time asymmetry in nonequilibrium fluctuations, Phys. Rev. Lett. **98**, 150601 (2007).

[6] J. M. R. Parrondo, C. Van den Broeck, and R. Kawai, Entropy production and the arrow of time, New J. Phys. **11**, 073008 (2009).

[7] R. Landauer, Irreversibility and heat generation in the computing process, IBM J. Res. Dev. **5**, 183 (1961).

[8] C. H. Bennett, Notes on Landauer's principle, reversible computation, and Maxwell's demon, Studies History Philos. Sci. B **34**, 501 (2003).

[9] M. Esposito and C. Van den Broeck, Second law and Landauer principle far from equilibrium, EPL (Europhys. Lett.) **95**, 40004 (2011).

[10] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics, Nature (London) **483**, 187 (2012).

[11] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, Thermodynamics of information, Nat. Phys. **11**, 131 (2015).

[12] A. C. Barato and U. Seifert, Thermodynamic uncertainty relation for biomolecular processes, Phys. Rev. Lett. **114**, 158101 (2015).

[13] U. Seifert, Stochastic thermodynamics: From principles to the cost of precision, Physica A **504**, 176 (2018).

[14] T. Koyuk and U. Seifert, Thermodynamic uncertainty relation for time-dependent driving, Phys. Rev. Lett. **125**, 260604 (2020).

[15] J. M. Horowitz and T. R. Gingrich, Thermodynamic uncertainty relations constrain nonequilibrium fluctuations, Nat. Phys. **16**, 15 (2020).

[16] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, The energy–speed–accuracy trade-off in sensory adaptation, Nat. Phys. **8**, 422 (2012).

[17] P. Sartori, L. Granger, C. F. Lee, and J. M. Horowitz, Thermodynamic costs of information processing in sensory adaptation, PLoS Comput. Biol. **10**, e1003974 (2014).

[18] V. Ngampruetikorn, D. J. Schwab, and G. J. Stephens, Energy consumption and cooperation for optimal sensing, Nat. Commun. **11**, 975 (2020).

[19] P. Sartori and S. Pigolotti, Thermodynamics of error correction, Phys. Rev. X **5**, 041039 (2015).

[20] R. Rao and M. Esposito, Nonequilibrium thermodynamics of chemical reaction networks: Wisdom from stochastic thermodynamics, Phys. Rev. X **6**, 041064 (2016).

[21] J. H. Fritz, B. Nguyen, and U. Seifert, Stochastic thermodynamics of chemical reactions coupled to finite reservoirs: A case study for the Brusselator, J. Chem. Phys. **152**, 235101 (2020).

[22] A. C. Barato, D. Hartich, and U. Seifert, Information-theoretic versus thermodynamic entropy production in autonomous sensory networks, Phys. Rev. E **87**, 042104 (2013).

[23] A. C. Barato, D. Hartich, and U. Seifert, Efficiency of cellular information processing, New J. Phys. **16**, 103024 (2014).

[24] T. Herpich, J. Thingna, and M. Esposito, Collective power: Minimal model for thermodynamics of nonequilibrium phase transitions, Phys. Rev. X **8**, 031056 (2018).

[25] M. Sune and A. Imparato, Out-of-equilibrium clock model at the verge of criticality, Phys. Rev. Lett. **123**, 070601 (2019).

[26] T. Herpich, T. Cossetto, G. Falasco, and M. Esposito, Stochastic thermodynamics of all-to-all interacting many-body systems, New J. Phys. **22**, 063005 (2020).

[27] R. Cofré and C. Maldonado, Information entropy production of maximum entropy Markov chains from spike trains, Entropy **20**, 34 (2018).

[28] R. Cofré, L. Videla, and F. Rosas, An introduction to the nonequilibrium steady states of maximum entropy spike trains, Entropy **21**, 884 (2019).

[29] C. W. Lynn, C. M. Holmes, W. Bialek, and D. J. Schwab, Emergence of local irreversibility in complex interacting systems, Phys. Rev. E **106**, 034102 (2022).

[30] C. W. Lynn, C. M. Holmes, W. Bialek, and D. J. Schwab, Decomposing the local arrow of time in interacting systems, Phys. Rev. Lett. **129**, 118101 (2022).

[31] S. Goldt and U. Seifert, Stochastic thermodynamics of learning, Phys. Rev. Lett. **118**, 010601 (2017).

[32] D. H. Wolpert, A. Kolchinsky, and J. A. Owen, A space–time tradeoff for implementing a function with master equation dynamics, Nat. Commun. **10**, 1727 (2019).

[33] D. H. Wolpert, The stochastic thermodynamics of computation, J. Phys. A: Math. Theor. **52**, 193001 (2019).

[34] D. H. Wolpert and A. Kolchinsky, Thermodynamics of computing with circuits, New J. Phys. **22**, 063047 (2020).

[35] A. V. Ivlev, J. Bartnick, M. Heinen, C.-R. Du, V. Nosenko, and H. Löwen, Statistical mechanics where Newton's third law is broken, Phys. Rev. X **5**, 011035 (2015).

[36] A. Crisanti and H. Sompolinsky, Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model, Phys. Rev. A **36**, 4922 (1987).

[37] A. Crisanti and H. Sompolinsky, Dynamics of spin systems with randomly asymmetric bonds: Ising spins and Glauber dynamics, Phys. Rev. A **37**, 4865 (1988).

[38] M. Aguilera, S. A. Moosavi, and H. Shimazaki, A unifying framework for mean-field theories of asymmetric kinetic Ising systems, Nat. Commun. **12**, 1197 (2021).

[39] M. Aguilera, M. Igarashi, and H. Shimazaki, Nonequilibrium thermodynamics of the asymmetric Sherrington-Kirkpatrick model, Nat. Commun. **14**, 3685 (2023).

[40] I. Ginzburg and H. Sompolinsky, Theory of correlations in stochastic neural networks, Phys. Rev. E **50**, 3171 (1994).

[41] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, The asynchronous state in cortical circuits, Science **327**, 587 (2010).

[42] Y. Roudi, B. Dunn, and J. Hertz, Multi-neuronal activity and functional connectivity in cell assemblies, Curr. Opin. Neurobiol. **32**, 38 (2015).

[43] B. Dunn, M. Mørreaunet, and Y. Roudi, Correlations and functional connections in a population of grid cells, PLoS Comput. Biol. **11**, e1004052 (2015).

[44] Y.-L. Shi, R. Zeraati, A. Levina, and T. A. Engel, Spatial and temporal correlations in neural networks with structured connectivity, Phys. Rev. Res. **5**, 013005 (2023).

[45] S. A. M. Loos and S. H. L. Klapp, Irreversibility, heat and information flows induced by non-reciprocal interactions, New J. Phys. **22**, 123051 (2020).

[46] S. A. M. Loos, S. Arabha, A. Rajabpour, A. Hassanali, and É. Roldán, Nonreciprocal forces enable cold-to-hot heat transfer between nanoparticles, Sci. Rep. **13**, 4517 (2023).

[47] J. Schnakenberg, Network theory of microscopic and macroscopic behavior of master equation systems, Rev. Mod. Phys. **48**, 571 (1976).

[48] E. Roldán and J. M. R. Parrondo, Estimating dissipation from single stationary trajectories, Phys. Rev. Lett. **105**, 150607 (2010).

[49] D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. **22**, 403 (1976).

[50] D. T. Gillespie, Stochastic simulation of chemical kinetics, Annu. Rev. Phys. Chem. **58**, 35 (2007).

[51] H. Schwarze and J. Hertz, Generalization in a large committee machine, Europhys. Lett. **20**, 375 (1992).

[52] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

[53] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[54] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[55] B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová, Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions, in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, 2021), Vol. 34, pp. 10144–10157.

[56] M. Refinetti, S. Goldt, F. Krzakala, and L. Zdeborova, Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed, in *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research*, edited by M. Meila and T. Zhang (PMLR, 2021), Vol. 139, pp. 8936–8947.

[57] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (Monticello, IL, USA, 1999), pp. 368–377.

[58] D. Strouse and D. J. Schwab, The deterministic information bottleneck, Neural Comput. **29**, 1611 (2017).

[59] M. Chalk, O. Marre, and G. Tkacik, Relevant sparse codes with variational information bottleneck, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, 2016), Vol. 29.

[60] M. Baiesi and C. Maes, Life efficiency does not always increase with the dissipation rate, J. Phys. Commun. **2**, 045017 (2018).

[61] S. Bai, J. Z. Kolter, and V. Koltun, Deep equilibrium models, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, 2019), Vol. 32.

[62] S. Bai, V. Koltun, and J. Z. Kolter, Multiscale deep equilibrium models, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, 2020), Vol. 33, pp. 5238–5250.

[63] J. C. Spall, An overview of the simultaneous perturbation method for efficient optimization, Johns Hopkins APL Tech. Dig. **19**, 482 (1998).

[64] S. Greydanus, Scaling down deep learning, arXiv:2011.14439 (2020).

[65] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A: Math. Gen. **21**, 271 (1988).

[66] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, J. Phys. A: Math. Gen. **22**, 1983 (1989).

[67] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[68] T. Hatano and S.-I. Sasa, Steady-state thermodynamics of Langevin systems, Phys. Rev. Lett. **86**, 3463 (2001).

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (NIPS, 2017).

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018).

[71] OpenAI: J. Achiam *et al.*, GPT-4 technical report, arXiv:2303.08774.

[72] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities., Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[73] D. Krotov and J. J. Hopfield, Dense associative memory for pattern recognition, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, 2016), Vol. 29.

[74] D. Krotov and J. Hopfield, Large associative memory problem in neurobiology and machine learning, arXiv:2008.06996 (2021).

[75] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve *et al.*, Hopfield networks is all you need, arXiv:2008.02217 (2021).

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, 2017), Vol. 30.

[77] D. Russo and B. V. Roy, An information-theoretic analysis of Thompson sampling, J. Machine Learn. Res. **17**, 1 (2016).

[78] A. Xu and M. Raginsky, Information-theoretic analysis of generalization capability of learning algorithms, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, and R. Garnett (Curran Associates, 2017), Vol. 30.

[79] D. Russo and B. Van Roy, Satisficing in time-sensitive bandit learning, Math. Operat. Res. **47**, 2547 (2022).

[80] V. Pacelli and A. Majumdar, Task-driven estimation and control via information bottlenecks, in *2019 International Conference on Robotics and Automation (ICRA)* (2019), pp. 2061–2067.

[81] V. Pacelli and A. Majumdar, Learning task-driven control policies via information bottlenecks, in *Robotics: Science and Systems XVI* (Robotics: Science and Systems Foundation, 2020).

[82] A. Majumdar and V. Pacelli, Fundamental performance limits for sensor-based robot control and policy learning, arXiv:2202.00129 (2022).

[83] https://github.com/greydanus/mnist1d.

[84] T. Martynec, S. H. L. Klapp, and S. A. M. Loos, Entropy production at criticality in a nonequilibrium Potts model, New J. Phys. **22**, 093069 (2020).

[85] J. C. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE Trans. Auto. Control **37**, 332 (1992).

[86] J. C. Spall, Implementation of the simultaneous perturbation algorithm for stochastic optimization, IEEE Trans. Aerospace Elect. Syst. **34**, 817 (1998).

[87] D. S. Seara, B. B. Machta, and M. P. Murrell, Irreversibility in dynamical phases and transitions, Nat. Commun. **12**, 392 (2021).