Perspective

Minimax entropy: The statistical physics of optimal models

David P. Carcamo,[1,2] Nicholas J. Weaver,[1,2] Purushottam D. Dixit,[3,4] and Christopher W. Lynn[1,2,5,*]

[1]*Department of Physics, Yale University, New Haven, Connecticut, USA*
[2]*Quantitative Biology Institute, Yale University, New Haven, Connecticut, USA*
[3]*Department of Biomedical Engineering, Yale University, New Haven, Connecticut, USA*
[4]*Systems Biology Institute, Yale University, West Haven, Connecticut, USA*
[5]*Wu Tsai Institute, Yale University, New Haven, Connecticut, USA*

(Received 2 May 2025; published 3 December 2025)

When constructing models of the world, we aim for optimal compressions: models that include as few details as possible while remaining as accurate as possible. But which details—or features measured in data—should we choose to include in a model? Here, using the minimum description length principle, we show that the optimal features are the ones that produce the maximum entropy model with minimum entropy, thus yielding a minimax entropy principle. We review applications, which range from machine learning to optimal models of biological networks. Naive implementations, however, are limited to systems with small numbers of states and features. We therefore require new theoretical insights and computational techniques to construct optimal compressions of high-dimensional datasets arising in large-scale experiments.

DOI: 10.1103/kr9x-q59y

I. INTRODUCTION

Statistical models are inherently lossy compressions. When confronted with experimental data, one does not include every available feature. In practice, with finite samples, including too many features leads to overfitting. In theory, even without sampling errors, one should discard features that do not improve their description of a system. In this way, all models are wrong, but which are useful?

Intuitively, useful models are those that strike an optimal balance between complexity and uncertainty. Given constraints on the complexity of a model—for example, the number of parameters or features included from data—one would like to provide the most accurate description of a system, thereby minimizing uncertainty [Fig. 1(a), vertical line]. Similarly, for a given level of uncertainty, one would like to remove as many details as possible, thus minimizing complexity [Fig. 1(a), horizontal line]. Importantly, for a given system, there is not a single useful model, but rather an entire spectrum. Where one lands on this spectrum depends on the relative value they assign to accuracy versus simplicity.

Information theory provides the mathematical framework to make these intuitions concrete [1,2]. Each model—specifically, each distribution $P(x)$ over states x of a system—defines an encoding of the system [2]. The length of this encoding, known as the description length, reflects the uncertainty of the modeler. Given a set of features measured in data and nothing else, the model with the minimum description length is the one with maximum entropy [3]. Thus, the maximum entropy principle provides the most accurate mapping from experimental measurements to statistical models. However, it does not tell us which features (and therefore which model) one should select. For a given level of complexity—that is, under constraints on the number or type of features—which model minimizes our uncertainty?

Here we show that the optimal features, which minimize description length, are the ones that produce the maximum entropy model with minimum entropy. This tells us that the best compressions identify the features in the data that maximally constrain our expectations, while remaining maximally random with regard to all unobserved features. Thus, starting only with the minimum description length (MDL) principle [4,5], one can derive not only the maximum entropy principle, but also its generalization, the "minimax entropy" principle.

Existing methods in unsupervised learning, such as principal component analysis, k-means clustering, support vector machines, and independent component analysis (among others) all derive simplified descriptions by reducing dimensionality [6]. Each of these methods arises from a different choice for the objective to optimize. In the minimax entropy principle, the optimal model is the one that achieves maximal compression, that is, the lowest entropy among maximum entropy models.

Originally proposed in the context of machine learning [7,8], the minimax entropy principle has recently emerged as a data-driven approach to constructing optimal models of complex living systems [9–12]. These applications, however, are limited by our ability to solve statistical physics problems. We therefore require new theoretical and computational methods to study the high-dimensional data being generated in large-scale experiments.

Finally, even with the tools to solve the minimax entropy problem, there is no guarantee that a useful compression

*Contact author: christopher.lynn@yale.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

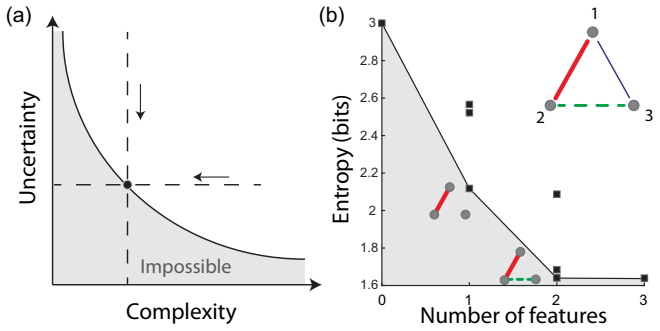


FIG. 1. Useful models as optimal compressions. (a) Illustration of uncertainty versus complexity. Optimal models minimize uncertainty for a given complexity (vertical line) or, equivalently, minimize complexity for a given uncertainty (horizontal line). (b) For the three-spin Ising model in Eq. (14), we plot the entropy $S(P)$ of maximum entropy models constructed to match different sets of pairwise correlations (square points). For each number of correlations, the optimal bound indicates the minimum entropy. Note that for maximum entropy models the entropy is equivalent to the description length $L(P)$ [Eqs. (10)–(11)].

exists. This depends critically on the system of interest. Just as images of independent pixels cannot be compressed, some systems may not admit simplified models capable of making accurate predictions. Fortunately, converging evidence across physics, biology, and neuroscience provides hope for useful models that do not rely on all of the details.

The paper is organized as follows. In Sec. II we derive the minimax entropy principle as the solution to a compression problem. In Sec. III we comment on alternative derivations, including minimizing the divergence from data and maximizing the information contained in experimental observations. In Sec. IV we survey existing applications of the minimax entropy principle. In Sec. V we highlight key outstanding challenges, open questions, and future directions. Finally, in Sec. VI we provide conclusions and outlook.

II. MINIMAX ENTROPY PRINCIPLE

A. Maximizing entropy

Consider a system of N variables with collective states $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. From experiments, we have access to features of the system, which can be represented as empirical averages $\langle f(\boldsymbol{x})\rangle_{\text{exp}}$, where $f(\boldsymbol{x})$ is an arbitrary function of the state $\boldsymbol{x}$. For example, we could build an empirical histogram by simply counting states, $P_{\text{exp}}(\boldsymbol{x}') = \langle \delta_{\boldsymbol{x},\boldsymbol{x}'}\rangle_{\text{exp}}$. Alternatively, we could measure the averages of individual variables $\langle x_i\rangle_{\text{exp}}$ or the correlations among pairs of variables $\langle x_i x_j\rangle_{\text{exp}}$, triplets $\langle x_i x_j x_k\rangle_{\text{exp}}$, or more. After measuring the averages of M functions $\{f_\mu(\boldsymbol{x})\}$, where $\mu = 1, \ldots, M$, we would like to construct a model $P(\boldsymbol{x})$ that is consistent with these features, such that

$$\langle f_\mu(\boldsymbol{x})\rangle = \langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}} \quad \text{for all } \mu, \tag{1}$$

where $\langle\cdot\rangle$ represents an average over $P(\boldsymbol{x})$. However, there are infinitely many distributions that match these features; how should we select one among them?

For every state $\boldsymbol{x}$, each model $P(\boldsymbol{x})$ defines a code word of length

$$\ell_P(\boldsymbol{x}) = -\ln P(\boldsymbol{x}). \tag{2}$$

The lower the probability $P(\boldsymbol{x})$, the more surprised the modeler is to see the state $\boldsymbol{x}$, and therefore the longer their encoding [1,2]. To compute the total description length of the model, we must average over all hypothetical datasets that could have generated the observed features [3]. Specifically, let $\{\boldsymbol{x}^{(t)}\}$ denote a dataset consisting of T samples of the system state $\boldsymbol{x}^{(t)}$, where $t = 1, \ldots, T$. The set of all datasets consistent with the observed features is given by

$$\chi = \left\{ \{\boldsymbol{x}^{(t)}\} \,:\, \frac{1}{T}\sum_{t=1}^{T} f_\mu(\boldsymbol{x}^{(t)}) = \langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}} \;\forall\, \mu \right\}. \tag{3}$$

Averaging over these datasets, we arrive at the description length of the model,

$$L(P) = \langle \ell_P(\boldsymbol{x})\rangle_\chi. \tag{4}$$

To minimize the description length [4,5], the model should be maximally unbiased; any source of order that is not justified by the measured features will lead to increases in code word lengths. Indeed, among all models consistent with the observed features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$, one can show (in the limit of large T) that the description length $L(P)$ is minimized by the model with maximum entropy (Appendix A)

$$S(P) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \ln P(\boldsymbol{x}). \tag{5}$$

This maximum entropy model takes the form of a Boltzmann distribution [13]

$$P_{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{Z}\exp\left[\sum_{\mu=1}^{M} \lambda_\mu f_\mu(\boldsymbol{x})\right], \tag{6}$$

where

$$Z = \sum_{\boldsymbol{x}} \exp\left[\sum_{\mu=1}^{M} \lambda_\mu f_\mu(\boldsymbol{x})\right] \tag{7}$$

is the normalizing partition function (Appendix B). The parameters λ_μ are Lagrange multipliers, which must be calculated so that the model matches the measured features [Eq. (1)]. This is the canonical inverse problem in statistical physics [13–17].

The maximum entropy principle presents an exact path from measured features to statistical physics models [13,14]. For example, if each variable is binary ($x_i = \pm 1$) and we measure all of the averages $\langle x_i\rangle_{\text{exp}}$ and pairwise correlations $\langle x_i x_j\rangle_{\text{exp}}$, then we have the pairwise maximum entropy model,

$$P_{\text{pair}}(\boldsymbol{x}) = \frac{1}{Z}\exp\left[\sum_i h_i x_i + \sum_{ij} J_{ij} x_i x_j\right], \tag{8}$$

which is equivalent to an Ising model with all-to-all interactions [18]. Importantly, since all other sources of order are explicitly removed, any predictions that the model makes must arise exclusively from the features that the modeler chose to include. In this way, models like Eq. (8) allow us to determine

which of a system's collective features can be understood as emerging from simple correlations. This approach has provided effective quantitative descriptions for a wide range of collective phenomena in biology, including patterns of activity in neural networks [19–24], sequence-structure maps in proteins [25–28], expression patterns in genetic networks [29,30], folding in chromatin [31–35], dynamics in flocks of birds [36–38], and activity in social networks [39]. However, we still do not know which set of features—and therefore which model—provides the best description of a system.

B. Minimizing entropy

Each choice for the set of features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$ yields a distinct maximum entropy model $P_\mathcal{F}(\boldsymbol{x})$. Given constraints on the allowed sets of features—for example, on the number of features or their functional form—we would like to identify the optimal set $\mathcal{F}^*$ that provides the best compression of the system $P_{\mathcal{F}^*}(\boldsymbol{x})$. Due to the Boltzmann form of Eq. (6), the description length of a maximum entropy model can be expressed as

$$L(P_\mathcal{F}) = -\langle \ln P_\mathcal{F}(\boldsymbol{x})\rangle_\chi = \ln Z - \sum_{\mu=1}^{M} \lambda_\mu \langle f_\mu(\boldsymbol{x})\rangle_\chi . \quad (9)$$

Moreover, since the datasets in χ are constrained to match the experimental averages of the features in $\mathcal{F}$, we have $\langle f_\mu(\boldsymbol{x})\rangle_\chi = \langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}}$. Finally, due to the maximum entropy constraints in Eq. (1), we know that the model also matches the same averages. Together, these results reveal that the description length of any maximum entropy model is equivalent to the entropy of the model itself,

$$L(P_\mathcal{F}) = \ln Z - \sum_{\mu=1}^{M} \lambda_\mu \langle f_\mu(\boldsymbol{x})\rangle \quad (10)$$

$$= -\langle \ln P_\mathcal{F}(\boldsymbol{x})\rangle = S(P_\mathcal{F}). \quad (11)$$

We have arrived at a strikingly simple picture: minimizing the description length $L(P_\mathcal{F})$ is equivalent to minimizing the entropy $S(P_\mathcal{F})$. Thus, among all of the allowed sets of features $\mathcal{F}$, the optimal set $\mathcal{F}^*$ is the one that produces the maximum entropy model $P_{\mathcal{F}^*}(\boldsymbol{x})$ with minimum entropy $S(P_{\mathcal{F}^*})$. This is the *minimax entropy principle*, which can be stated mathematically,

$$\mathcal{F}^* = \arg\min_\mathcal{F} S(P_\mathcal{F}) \quad (12)$$

$$= \arg\min_\mathcal{F} \{\max_P S(P) : \langle f_\mu(\boldsymbol{x})\rangle = \langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}} \ \forall\ f_\mu \in \mathcal{F}\}. \quad (13)$$

This establishes a unifying perspective: by minimizing the length of our encoding, one can derive not only the optimal model for a given set of features (maximum entropy), but also the optimal features themselves (minimax entropy).

C. General solution

The form of the minimax entropy problem in Eqs. (12)–(13) immediately hints at an algorithmic solution:

(1) For a given set of features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$, one must first solve the traditional maximum entropy problem for the parameters λ_μ such that the model $P_\mathcal{F}(\boldsymbol{x})$ matches the specified features [Eq. (1)].

(2) One must then repeat step 1 for each allowed set of features $\mathcal{F}$ in order to identify the optimal set $\mathcal{F}^*$ that minimizes the model entropy $S(P_\mathcal{F})$.

This algorithm, while prohibitively inefficient in most cases (see below), provides a general solution to the minimax entropy problem.

To illustrate this solution, consider a three-spin Ising model with collective states $\boldsymbol{x} = \{x_1, x_2, x_3\}$, where $x_i = \pm 1$, defined by the Boltzmann distribution

$$P_{\text{true}}(\boldsymbol{x}) = \frac{1}{Z} \exp(3x_1x_2 - 2x_2x_3 + x_1x_3). \quad (14)$$

Without any knowledge of the system, the maximum entropy distribution $P(\boldsymbol{x})$ is uniform over the eight states $\boldsymbol{x}$, with entropy $S(P) = 3$ bits. This model of independent spins captures none of the underlying structure. By contrast, if we measure all three of the pairwise correlations $\langle x_i x_j\rangle_{\text{true}}$, then the maximum entropy model is exact $P(\boldsymbol{x}) = P_{\text{true}}(\boldsymbol{x})$, with entropy $S(P) = S(P_{\text{true}})$.

Between these two extremes, we have a choice for which correlations $\langle x_i x_j\rangle_{\text{true}}$ to include in a model. To select the single correlation that provides the best compression, we construct a different maximum entropy model $P(\boldsymbol{x}) = \frac{1}{Z}\exp(J_{ij}x_ix_j)$ for each correlation, fitting the interaction J_{ij} such that $\langle x_i x_j\rangle = \langle x_i x_j\rangle_{\text{true}}$. With partial knowledge of the system, each model yields an entropy 3 bits $\geqslant S(P) \geqslant S(P_{\text{true}})$ [Fig. 1(b)]. The optimal model, with minimum entropy, is constructed from the correlation $\langle x_1x_2\rangle$ corresponding to the strongest interaction J_{12}. Similarly, among pairs of correlations, the optimal model identifies the two strongest interactions J_{12} and J_{23} [Fig. 1(b)]; in fact, this model nearly provides an exact description of the system, with $S(P) \approx S(P_{\text{true}})$. This demonstrates how, by selecting optimal features, one does not need all of the details to construct an accurate model.

III. OPTIMIZING ACCURACY AND INFORMATION

In the previous section, we derived the minimax entropy principle as the solution to a compression problem. Here we provide two additional perspectives [7–10], which converge on the same framework for selecting optimal features from data.

A. Minimizing divergence

Consider a system defined by a true underlying distribution $P_{\text{true}}(\boldsymbol{x})$, and suppose we have access to an experiment that provides accurate measurements of system features, such that $\langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}} = \langle f_\mu(\boldsymbol{x})\rangle_{\text{true}}$. As discussed above, for any set of features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$, the maximum entropy model $P_\mathcal{F}(\boldsymbol{x})$ in Eq. (6) provides the most unbiased description of the system [2,13]. Generally, the features $\mathcal{F}$ provide only partial information about the system, yielding a model with entropy $S(P_\mathcal{F}) \geqslant S(P_{\text{true}})$. As we increase the number of features, this maximum entropy decreases until, in the limit that $\mathcal{F}$ fully defines the system, the model is exact, and $S(P_\mathcal{F}) = S(P_{\text{true}})$.

These observations leave us with the intuition that the difference $S(P_\mathcal{F}) - S(P_{\text{true}}) \geqslant 0$ reflects the accuracy of the

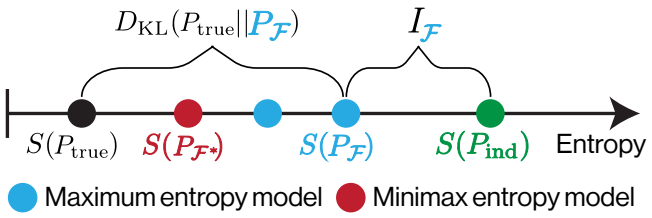


FIG. 2. Minimizing divergence and maximizing information. Among an allowed set of maximum entropy models (blue), the minimax entropy model has minimum entropy (red). This optimal model minimizes the KL divergence $D_{\mathrm{KL}}(P_{\mathrm{true}}||P_{\mathcal{F}}) = S(P_{\mathcal{F}}) - S(P_{\mathrm{true}})$ with the true distribution over states (black). Beginning from an independent model (green), the optimal model also maximizes the information $I_{\mathcal{F}} = S(P_{\mathrm{ind}}) - S(P_{\mathcal{F}})$ contained in correlations.

model. Indeed, this difference in entropy is precisely the Kullback-Leibler (KL) divergence between the system and the model (Fig. 2),

$$\begin{aligned} D_{\mathrm{KL}}(P_{\mathrm{true}}||P_{\mathcal{F}}) &= \left\langle \ln \frac{P_{\mathrm{true}}(\boldsymbol{x})}{P_{\mathcal{F}}(\boldsymbol{x})} \right\rangle_{\mathrm{true}} \\ &= \ln Z - \sum_{\mu} \lambda_{\mu} \langle f_{\mu}(\boldsymbol{x}) \rangle_{\mathrm{true}} - S(P_{\mathrm{true}}) \\ &= \ln Z - \sum_{\mu} \lambda_{\mu} \langle f_{\mu}(\boldsymbol{x}) \rangle - S(P_{\mathrm{true}}) \\ &= S(P_{\mathcal{F}}) - S(P_{\mathrm{true}}), \end{aligned} \tag{15}$$

where the third equality follows from the maximum entropy constraints $\langle f_{\mu}(\boldsymbol{x}) \rangle = \langle f_{\mu}(\boldsymbol{x}) \rangle_{\mathrm{exp}} = \langle f_{\mu}(\boldsymbol{x}) \rangle_{\mathrm{true}}$. Thus, by minimizing the entropy $S(P_{\mathcal{F}})$, the optimal features $\mathcal{F}^*$ also minimize the KL divergence with the true distribution (Fig. 2),

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} D_{\mathrm{KL}}(P_{\mathrm{true}}||P_{\mathcal{F}}). \tag{16}$$

In this way, the minimax entropy principle generates the most accurate description of the system $P_{\mathcal{F}^*}$ [7].

B. Maximizing information

Each feature that we measure in experiments not only improves the accuracy of our model; it also provides information about the underlying system. For example, consider a system of binary variables $\boldsymbol{x} = \{x_1, \ldots, x_N\}$. By measuring the averages $\langle x_i \rangle_{\mathrm{exp}}$, the maximum entropy model is equivalent to the independent model

$$P_{\mathrm{ind}}(\boldsymbol{x}) = \prod_i P_i(x_i), \tag{17}$$

where $P_i(x_i)$ is the marginal distribution of x_i. If, in addition, we measure a single correlation $\langle x_i x_j \rangle_{\mathrm{exp}}$, then the maximum entropy model takes the form

$$P(\boldsymbol{x}) = P_{ij}(x_i, x_j) \prod_{k \neq i,j} P_k(x_k), \tag{18}$$

where $P_{ij}(x_i, x_j)$ is the marginal of x_i and x_j. Thus, by measuring $\langle x_i x_j \rangle_{\mathrm{exp}}$, our uncertainty about the system decreases by an amount,

$$S(P_{\mathrm{ind}}) - S(P) = S(P_i) + S(P_j) - S(P_{ij}), \tag{19}$$

equal to the mutual information $I(x_i; x_j)$ between x_i and x_j. If we measure all of the correlations, of all orders, then the maximum entropy model is exact. In this limit, we gain an amount of information $I_{\mathrm{tot}} = S(P_{\mathrm{ind}}) - S(P_{\mathrm{true}})$, which is known as the total correlation or multi-information [40].

Between these extremes, we could measure a set of features $\mathcal{F}$ that includes the averages $\langle x_i \rangle_{\mathrm{exp}}$ and a subset of correlations. Each selection of correlations provides information about the underlying system in the amount

$$I_{\mathcal{F}} = S(P_{\mathrm{ind}}) - S(P_{\mathcal{F}}). \tag{20}$$

Thus, by minimizing the entropy $S(P_{\mathcal{F}})$, the optimal correlations $\mathcal{F}^*$ also provide the maximum information about the system (Fig. 2) [9,10],

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} I_{\mathcal{F}}. \tag{21}$$

Together, these results demonstrate that the minimax entropy principle, in addition to producing optimal compressions, also yields models that are as accurate and informative as possible.

IV. APPLICATIONS, SOLUTIONS, AND APPROXIMATIONS

A. Greedy algorithm

The minimax entropy principle was originally proposed over 25 years ago in the context of machine learning [7,8]. Yet, despite decades of advances in maximum entropy modeling, the minimax entropy principle has only recently found applications in the study of complex living systems [9–12]. The primary limitation is practical: While the algorithm in Sec. II provides a general solution to the minimax entropy problem, a naïve implementation is prohibitively inefficient for all but the smallest systems. For each set of features $\mathcal{F} = \{f_{\mu}(\boldsymbol{x})\}$, our ability to calculate the maximum entropy parameters λ_{μ} is fundamentally limited by our ability to compute model averages $\langle f_{\mu}(\boldsymbol{x}) \rangle$. Without simplifying assumptions, this often requires computationally intensive Monte Carlo simulations [15,18,41,42]. Then we must repeat this process for every allowed set of features $\mathcal{F}$; however, this number can explode combinatorially. For example, if our goal is to find the M optimal features out of $M' > M$ possibilities, then the number of allowed sets $\mathcal{F}$ is $\binom{M'}{M}$, which grows super-exponentially with M. Thus, a brute-force search over sets of features $\mathcal{F}$ is often infeasible [43].

Instead, we propose a greedy algorithm that grows the set of features by iteratively selecting the optimal feature at each step:

(1) To begin, select the single allowed feature $f_1(\boldsymbol{x})$ whose empirical average $\langle f_1(\boldsymbol{x}) \rangle_{\mathrm{exp}}$ yields the maximum entropy model with minimum entropy.

(2) After selecting $m < M$ features $f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$, search over the remaining allowed features to find the one $f_{m+1}(\boldsymbol{x})$ that yields the lowest entropy $S(P_{\mathcal{F}})$, where $\mathcal{F} = \{f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}), f_{m+1}(\boldsymbol{x})\}$.

(3) After M steps, we have converged on the locally optimal set of features $\mathcal{F}^* = \{f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\}$.

Although this greedy algorithm is not generally guaranteed to converge to the globally optimal features [43,44], it provides an approximate solution with a significant increase in

efficiency [7,8,11]. In turn, this opens the door for practical applications of the minimax entropy principle.

B. Minimax entropy Gaussians

Perhaps the simplest application of the minimax entropy principle is to systems of N continuous variables $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ with experimental covariances $\Sigma_{ij} = \langle x_i x_j \rangle_{\text{exp}} - \langle x_i \rangle_{\text{exp}} \langle x_j \rangle_{\text{exp}}$. A subset of these covariances can be visualized as a network G with nodes representing the different variables and edges defining the covariances that we include in our model. Given a network of covariances G, the maximum entropy model takes the form of a Gaussian,

$$P_G(\boldsymbol{x}) = \sqrt{\frac{|J|}{(2\pi)^N}} \exp\left[-\sum_{(ij)\in G} J_{ij} x_i x_j \right], \tag{22}$$

where the sum runs over all of the edges in G, and $|\cdot|$ is the determinant. Since the entropy of a Gaussian does not depend on the means of the variables, these can be subtracted from the data without loss of generality. If we include all of the covariances in our model, yielding an all-to-all network G, then the maximum entropy parameters are given directly by $J = \Sigma^{-1}$. More generally, for each covariance that we include in the model [that is, for each edge (ij) in G], the parameter J_{ij} must be calculated so that the model covariance $(J^{-1})_{ij}$ matches the measured value Σ_{ij}. For covariances not constrained in the model, $J_{ij} = 0$. Each subset of covariances (equivalently, each network G) therefore yields a Gaussian graphical model, which have found widespread applications across biology, psychology, and economics [45–49].

But given covariances Σ measured in data, which should we include in a model? The minimax entropy principle tells us that we should select the network G that yields the model with the lowest entropy

$$S(P_G) = -\frac{1}{2} \ln |J| + \frac{N}{2} \ln(2\pi e), \tag{23}$$

or, equivalently, the largest determinant $|J|$. Given a desired number of covariances M, we can use the greedy algorithm described above to grow a locally optimal network G with M edges [Fig. 3(a)]. In fact, this is a common strategy for model selection in Gaussian graphical models [45].

To illustrate this approach, consider neural activity in human subjects measured, for example, using functional magnetic resonance imaging (fMRI). In these recordings, the collective state of the brain is defined by a vector $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, where x_i represents the activity of region i. By studying the covariances between brain regions Σ, one can understand their functional relationships and how these relationships are altered over the course of development, between cognitive tasks, and in diseases such as schizophrenia and Alzheimer's [51,52].

Using the minimax entropy principle, one can investigate which covariances are most important for constraining the distribution over neural states. In a large cohort of 100 subjects [12], recent work has shown that one can achieve a large reduction in entropy with a relatively small number of covariances [Fig. 3(b)]. With only 10% of the available covariances, the optimal model captures over 90% of the information contained in the data, and accurately predicts most of the correlations between brain regions [Figs. 3(c)–3(d)]. By contrast, a random selection of covariances yields a much smaller drop in entropy [Fig. 3(b)] and fails to predict many correlations in the data [Fig. 3(d)]. These results demonstrate that, by optimizing the features to include in a model, the brain may be amenable to compressed descriptions.

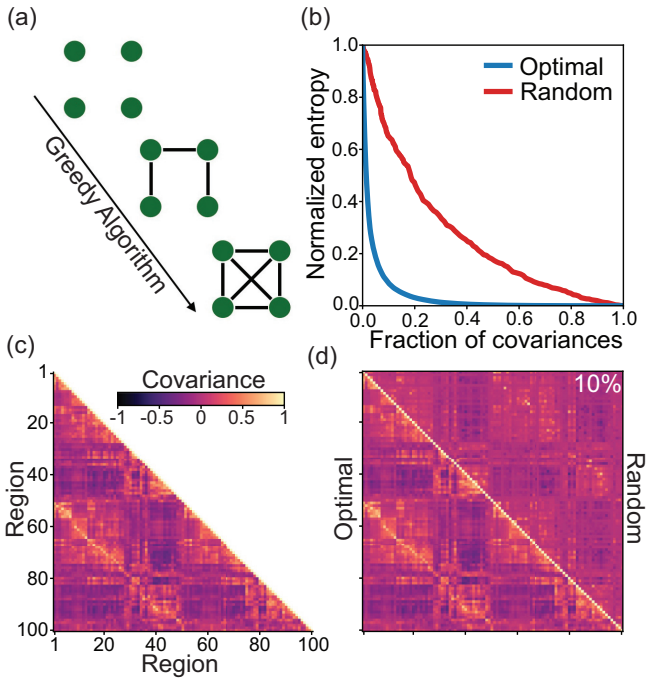


FIG. 3. Minimax entropy models of human neural activity. (a) To select an optimal set of covariances, the greedy algorithm iteratively adds edges to a network G so as to minimize the entropy $S(P_G)$. (b) Normalized entropy $[S(P_G) - S(P_{\text{tot}})]/[S(P_{\text{ind}}) - S(P_{\text{tot}})]$, where P_{tot} is the full Gaussian consistent with all of the covariances Σ, versus the number of covariances included in the model for networks constructed using the greedy algorithm (blue) and for random networks (red). Neural activity is recorded using fMRI from the cortex of 100 healthy human subjects as part of the Human Connectome Project [50]. (c) Measured covariances between brain regions Σ. (d) Covariances predicted by models constructed from the optimal network (left) and a random network (right), each containing 10% of the experimental covariances. Adapted from [12].

C. Minimax entropy Ising models

Many systems are described by discrete rather than continuous variables. For example, decreasing in scale from the whole brain to a population of neurons, within a short window of time each nerve cell i is either active ($x_i = 1$) or silent ($x_i = 0$). Just as in the continuous case, a subset of the correlations $\langle x_i x_j \rangle_{\text{exp}}$ between pairs of neurons can be represented as a network G. If our set of features includes all of the average activities $\langle x_i \rangle_{\text{exp}}$ and a network G of correlations, then the maximum entropy model takes the form

$$P_G(\boldsymbol{x}) = \frac{1}{Z} \exp\left[\sum_i h_i x_i + \sum_{(ij)\in G} J_{ij} x_i x_j \right]. \tag{24}$$

This is equivalent to an Ising model with allowed interactions defined by the network G. If we include all of the pairwise correlations, yielding an all-to-all network G, then we arrive at the pairwise model in Eq. (8). While this approach has provided key insights into a wide range of collective biological and physical phenomena [14,18–39,53], one might hope for a more compressed description. Indeed, as the number of variables N increases in large-scale experiments, the number of pairwise correlations grows as N^2, leading to a quadratic increase in the number of features (and thus the number of parameters) included in the model.

To compress our description of the system, the minimax entropy principle tells us that we should focus on the subset of correlations G that yields the Ising model P_G with the minimum entropy $S(P_G)$. However, for a given set of correlations, calculating the parameters h_i and J_{ij} (known as the inverse Ising problem) is notoriously difficult [15]; and even if we can infer the parameters, we still need to search for the model with the lowest entropy.

Calculations in statistical physics are difficult in part due to feedback loops in the interactions J_{ij}; eliminate these loops, and calculations often become tractable [55]. In fact, by focusing on networks G without loops (known as trees), the information contained in a set of correlations can be decomposed into a sum of mutual information [56]

$$I_G = S(P_{\text{ind}}) - S(P_G) = \sum_{(ij)\in G} I(x_i; x_j). \tag{25}$$

This decomposition reveals that the optimal tree [which minimizes $S(P_G)$ and maximizes I_G] is the one with the largest total mutual information. Finding this optimal network is a minimum spanning tree problem, which admits a number of efficient solutions [57], including the greedy algorithm described previously. Moreover, once the optimal tree is identified, one can analytically calculate the corresponding parameters h_i and J_{ij}, thus solving the inverse Ising problem [15,56]. Together, these results present an exact and efficient solution to the minimax entropy problem, enabling optimal Ising models of very high-dimensional systems [9,10].

The above solution hinges on a method known as exact renormalization, which is typically applied to systems without loops, such as one-dimensional Ising models or Bethe lattices [55,58]. The key insight is that, when computing the partition function, one can remove a node with one connection and arrive at an equivalent model with $N-1$ variables. If there are no loops in the network—that is, if the network forms a tree—then one can iteratively remove all of the nodes, thus computing the partition function and all other statistics of interest [9,10,15]. Recently, this exact renormalization method has been extended to remove nodes with two connections [11]. In turn, this extension exactly solves the maximum entropy problem in a broader class of networks known as generalized series-parallel (GSP) networks, which contain loops of interactions. Similarly to trees, GSP networks are defined by recursive construction. Yet when adding a new node, rather than connecting it to one existing node (to form a tree), one connects it to two existing nodes that are already neighbors. In fact, this is the maximal class of Ising models that can be solved using exact renormalization. With GSP networks, one can include $\sim 2N$ correlations from data, twice as many as a tree but still a vanishingly small fraction of the $\sim N^2/2$ possible correlations. As N increases in high-dimensional data, can such sparse models capture anything meaningful about a system?

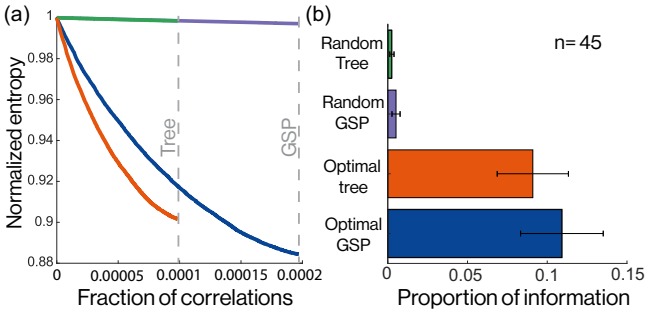


FIG. 4. Minimax entropy models of large populations of neurons. (a) Entropy [normalized by the independent entropy $S(P_{\text{ind}})$] versus the fraction of possible correlations included in different models of collective activity in a population of $N \approx 10\,000$ neurons in the mouse visual cortex [11,54]. Using the greedy algorithm, we select the optimal tree of correlations (orange) and the optimal GSP network (blue). We compare against a random tree (green) and a random GSP network (purple). Dashed lines indicate the maximum number of correlations allowed in trees and GSP networks. (b) Information I_G [normalized by $S(P_{\text{ind}})$] captured by different networks of correlations. Values and error bars represent averages and standard deviations across 45 different recordings [11,54].

To answer this question, we study 45 recordings of $\sim 10\,000$ neurons in the mouse visual cortex [54,59]. In Fig. 4, we compare the minimax entropy models constructed from optimal trees and optimal GSP networks of correlations with maximum entropy models constructed from random trees and random GSP networks. When the optimal tree and GSP models are saturated with edges, they contain 0.02% and 0.04% of the available pairwise correlations, respectively. Despite this sparsity, by focusing on the most important correlations, these optimal models capture 9% and 11% of the independent entropy $S(P_{\text{ind}})$ in the populations [Fig. 4(b)]. For comparison, random models with the same numbers of correlations provide almost no benefit over independent models. Together, these results demonstrate that large amounts of information can be packed into sparse networks of correlations, providing hope for highly compressed descriptions of neural activity.

D. Minimax entropy texture modeling

The minimax entropy principle was first developed for texture modeling in computer vision [7,8]. The task of modeling texture is to construct a distribution over images that is constrained to some of their characteristic features, which are defined by filters. For example, some filters detect vertical edges, while others detect horizontal edges. Additionally, the overall intensity of the pixel values can be used as a filter, effectively normalizing the pixel values across the image. Given a set of filters, the maximum entropy principle provides the most unbiased distribution over images; yet the modeler is still tasked with choosing the set of filters. The minimax entropy principle reveals that the optimal filters, which provide

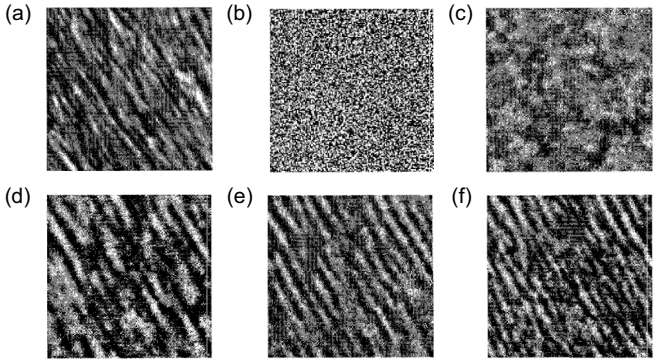


FIG. 5. Minimax entropy model of texture. (a) Target image of fur. (b)–(f) Images generated from models with zero (b), one (c), two (d), three (e), and six (f) filters. Adapted from [8].

the most accurate distribution over images, are the ones that produce the maximum entropy model with minimum entropy [7,8].

This method is illustrated in Fig. 5 for an image of fur [Fig. 5(a)]. With no filters, the model contains no information, and we are left with an image of stochastic noise [Fig. 5(b)]. Using the greedy algorithm described above, one proceeds to iteratively select filters that produce the largest reduction in entropy. As more filters are added, the original image quickly comes into focus [Figs. 5(c)–5(e)], until, with only six filters, the model accurately reproduces the desired texture [Fig. 5(f)]. Thus, just as in the neural systems described above, only a small subset of features are needed to capture large amounts of variability in the data, thereby providing a good compression.

E. Minimax entropy for multiple distributions

While the minimax entropy principle provides a systematic approach to identify the optimal subset of features from a larger set, the modeler is still tasked with defining this larger set of allowed features. In full generality, one could allow any function $f(\boldsymbol{x})$ over states. Given this flexibility, the minimum entropy is achieved by selecting $f(\boldsymbol{x})$ so that the model matches the empirical distribution over states, $P(\boldsymbol{x}) = P_{\rm exp}(\boldsymbol{x})$. However, in high-dimensional data, this empirical distribution can be vastly undersampled; and even if the states are well sampled, simply matching the empirical distribution does not provide a compressed description of the system.

Here we consider a generalization of the minimax entropy principle, where the modeler seeks to select a set of features that apply across multiple distributions. This approach is most relevant for modeling normalized abundance counts, which arise across biology and engineering, including gene expression, microbial ecology, and image pixel intensities [30,60–63]. For concreteness, we consider A different models $P_\alpha(\boldsymbol{x})$, where $\alpha = 1, \ldots, A$, each sharing the same M features $f_\mu(\boldsymbol{x})$, where $\mu = 1, \ldots, M$. For a given choice of features, each model is defined by a maximum entropy distribution [64]

$$P_\alpha(\boldsymbol{x}) = \frac{1}{Z_\alpha} \exp\left[\sum_{\mu=1}^{M} \lambda_{\alpha\mu} f_\mu(\boldsymbol{x})\right]. \tag{26}$$

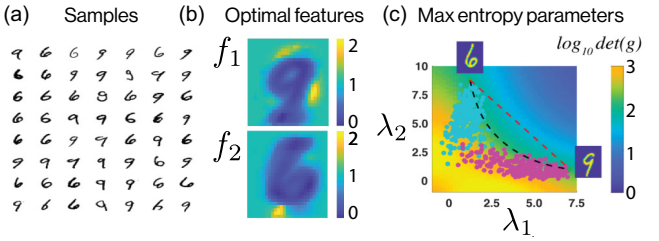


FIG. 6. Minimax entropy models of handwritten digits. (a) Digitized images of handwritten sixes and nines from the MNIST database [63]. (b) The first two optimal features capture the salient structure in the data. Color bar represents the values of the constraint for each pixel. Blue indicates a lower value of the constraint, and yellow indicates a higher value. (c) Each image defines a unique point in the two-dimensional space of maximum entropy parameters. Colors indicate the volume element of the Fisher-Rao metric g, which defines a natural metric for parameterized probability distribution. The dashed black line illustrates the geodesic between two points in this metric, which passes through data-rich regions of the space. For comparison, the dashed red line represents the straight (Euclidean) line between the same two points.

For each distribution α, the parameters $\lambda_{\alpha\mu}$ must be calculated so that the model averages match the empirical averages,

$$\sum_{\boldsymbol{x}} f_\mu(\boldsymbol{x}) P_\alpha(\boldsymbol{x}) = \langle f_\mu(\boldsymbol{x}) \rangle_{\alpha,\rm exp} \quad \text{for all } \mu. \tag{27}$$

In Fig. 6 we illustrate how optimal features can be inferred for multiple distributions using digitized images of handwritten sixes and nines from the MNIST database [63]. Each image in Fig. 6(a) is defined by normalized pixel intensities, which can be treated as an empirical histogram. Using the minimax entropy principle, the first two optimal features $f_\mu(\boldsymbol{x})$ clearly represent images of sixes and nines [Fig. 6(b)]. Thus, even without access to the labels "6" and "9," minimax entropy extracts this structure from the data. For each image α, we also infer the maximum entropy parameters $\lambda_{\alpha\mu}$, thus defining a point in a two-dimensional space spanning the two features [Fig. 6(c)]. Once again, this example demonstrates how a large amount of information can be compressed into a small number of features.

V. FUTURE DIRECTIONS AND OPEN CHALLENGES

A. Applications

Maximum entropy models have become increasingly applied as a principled way to map experimental data to quantitative predictions across biology, neuroscience, physics, and engineering [13,14,19–38,53,65,66]. In each of these contexts, the minimax entropy principle has the potential to provide immediate improvements by identifying the optimal set of features—and therefore the optimal model—to describe the underlying system.

For example, gene regulatory networks are often inferred indirectly from correlations in gene expressions [29]. Rather than including all of the correlations or selecting networks based on known interactions, one can use the minimax entropy principle as a data-based method for identifying the correlations—and therefore the underlying interactions—that

provide the most accurate predictions for gene coexpression. As discussed above, restricting attention to a sparse tree of dependencies can highlight the most informative structure in the data and previous work shows that tree representations can preserve cell-cell distances in gene expression with low distortion [67]. Similarly, DNA folds into compact structures known as chromatin, yielding correlations in the physical locations of genetic sites; yet the mechanisms are not fully understood. While maximum entropy approaches have been used to infer all-to-all interactions [31–34], one might suspect that the underlying network is sparse. To test this hypothesis, one needs to infer the optimal network of interactions for a given level of sparsity. In precisely this way, the minimax entropy principle provides the tools to infer compressed descriptions and sparse networks of interactions not only in gene expression and chromatin structure, but also in the brain [19–24], animal populations [36–38], social networks [39], proteins [25–28], and other complex living systems.

B. Challenges

The primary challenge in applying the minimax entropy principle is practical: For each set of features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$, one must compute the parameters λ_μ to define the maximum entropy model $P_\mathcal{F}(\boldsymbol{x})$; then one must repeat this process many times to search for the optimal features $\mathcal{F}^*$ that minimize the entropy $S(P_\mathcal{F})$. To solve the maximum entropy problem for the parameters λ_μ, one must first be able calculate the model averages $\langle f_\mu(\boldsymbol{x})\rangle$. For a Gaussian model, computing covariances amounts to inverting the interaction (or precision) matrix, so the maximum entropy problem can be solved efficiently [45,46]. But for systems with discrete states, such as the Ising model, computing model averages is generally difficult, requiring computationally intensive Monte Carlo simulations [15,18,41,42]. As discussed above, one solution is to focus on networks of interactions that allow for exact solutions [9,10,55]; yet these methods can impose severe restrictions on network topology. Another potential solution lies in mean-field approximations, which replace stochastic averages with deterministic self-consistency equations [41,68–70]. However, increasing evidence suggests that some biological systems may be poised near criticality [10,38,71–73], which can lead to catastrophic failures in mean-field approximations [74]. Pseudolikelihood methods have also found success in approximating the collective activity of biological systems and efficiently inferring model parameters [75–77]. These approximations, in addition to improving the efficiency of maximum entropy inference, can also be used to evaluate changes in entropy at each step of the greedy algorithm, thus improving the inference of optimal features. Yet future work is still needed to develop scalable strategies for solving both the maximum and minimax entropy problems [15,78], particularly in high-dimensional systems and parameter regimes that are relevant for biology.

Even if we can compute the model $P_\mathcal{F}(\boldsymbol{x})$ for a given set of features $\mathcal{F}$, one still needs to search for the optimal set of features $\mathcal{F}^*$. As discussed previously, searching for the optimal features by brute force is intractable for all but the smallest systems [43]. Instead, one typically relies on a greedy algorithm, which converges to a locally optimal solution, but is not guaranteed to find the global optimum [7,8,11,44]. In some cases, however, it is possible to derive performance guarantees. For example, for trees of pairwise correlations, the minimax entropy problem can be reduced to a minimum spanning tree problem, and the greedy algorithm identifies the globally optimal correlations (Sec. IV). More generally, if including more features in the set $\mathcal{F}$ provides decreasing reductions in entropy $S(P_\mathcal{F})$, then the optimization problem is known as submodular [79]. For submodular functions (the discrete analog of convex functions), the greedy algorithm is guaranteed to converge to a solution that is within $1 - 1/e$ of the global optimum [80]. More generally, in systems that are sufficiently small, one can identify the globally optimal features via brute force search. In small groups of neurons, this method has confirmed that the greedy algorithm is strikingly close to globally optimal [11]. Additionally, rather than greedily growing the optimal set of features, one can instead begin with all allowed features and iteratively remove those that produce the smallest increases in entropy. By comparing these "bottom-up" and "top-down" solutions, one can test the consistency of the greedy algorithm [81]. Developing performance guarantees and improvements to the greedy algorithm remains a clear open challenge.

C. Parameterized features

Thus far, we have focused on selecting among discrete features $\langle f_\mu(\boldsymbol{x})\rangle_{\text{exp}}$, such as average states $\langle x_i\rangle_{\text{exp}}$ or pairwise correlations $\langle x_i x_j\rangle_{\text{exp}}$. Instead, one could imagine optimizing over a parameterized family of functions $f_{\boldsymbol{w}}(\boldsymbol{x})$, where $\boldsymbol{w}$ is a vector of parameters that can be tuned continuously. For example, consider the family of features $f_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_i w_i x_i$, which yields the maximum entropy model

$$P_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{Z}\exp\left[\lambda \sum_i w_i x_i\right], \tag{28}$$

where λ must be calculated so that

$$\sum_i w_i \langle x_i\rangle = \sum_i w_i \langle x_i\rangle_{\text{exp}}. \tag{29}$$

Now, rather than selecting between discrete features, we want to find the values for $\boldsymbol{w}$ that produce the maximum entropy model with minimum entropy $S(P_{\boldsymbol{w}})$. For binary variables, one can show that these optimal parameters $\boldsymbol{w}^*$ are the ones that give the correct independent model, such that $P_{\boldsymbol{w}^*}(\boldsymbol{x}) = P_{\text{ind}}(\boldsymbol{x})$.

More generally, by focusing on parameterized features, one can take advantage of methods from continuous optimization. For example, one can construct a gradient descent algorithm by computing the the gradients $\nabla_{\boldsymbol{w}} S(P_{\boldsymbol{w}})$. This continuous version of the minimax entropy problem may provide complementary insights to the original discrete version, which have yet to be explored.

VI. CONCLUSIONS

When constructing models, there exists an inherent tradeoff between accuracy and simplicity. A good model is a good compression, providing an accurate description with as few details as possible. From an information-theoretic

perspective, accurate models define efficient encodings of data, thus minimizing description length. Given a set of details—or features measured in experiments—the description length is minimized by the model with maximum entropy [3]. While this has long been understood, it does not tell us which features we should choose to begin with.

Here we show that the optimal features, which minimize description length, are the ones that yield the maximum entropy model with minimum entropy. Thus, for a given number of features, the minimax entropy principle provides the optimal compression.

Although the minimax entropy principle was proposed decades ago [7,8], it has only recently been applied in the study of complex living systems [9–12]. Indeed, applications are limited by our ability to solve statistical physics problems and then search over many solutions. These practical challenges become even more daunting as experiments grow to record from larger systems. Future success therefore depends on the alignment of advanced techniques from theoretical physics with their appropriate contexts in biology, neuroscience, and engineering.

ACKNOWLEDGMENTS

We thank W. Bialek, S. E. Palmer, F. Mignacco, L. Di Carlo, Q. Yu, M. P. Leighton, B. Machta, and C. Smith for helpful discussions. D.P.C., N.J.W., and C.W.L. acknowledge support from the National Institutes of Health (NIH/NIGMS R35GM160188). P.D.D. acknowledges support from the National Institutes of Health (NIH/NIGMS R35GM142547).

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: MAXIMUM ENTROPY FROM MINIMUM DESCRIPTION LENGTH

Consider a system with states $\boldsymbol{x}$ and a set of M functions $\{f_\mu(\boldsymbol{x})\}$, where $\mu = 1, \ldots, M$. From experiments, we have access to the empirical averages $\langle f_\mu(\boldsymbol{x})\rangle_{\rm exp}$. These features could have been generated by one of many datasets $\{\boldsymbol{x}^{(t)}\}$ consisting of T samples of the system state $\boldsymbol{x}^{(t)}$, where $t = 1, \ldots, T$. The set of all datasets consistent with the observed features is denoted χ [Eq. (3)].

For every state $\boldsymbol{x}$, a model $P(\boldsymbol{x})$ defines a code word of length $\ell_P(\boldsymbol{x}) = -\log P(\boldsymbol{x})$ [1,2]. By averaging over all of the possible datasets in χ, we arrive at the description length of the model $L(P)$ [Eq. (4)]. To compute this average, we note that each dataset $\{\boldsymbol{x}^{(t)}\}$ in χ defines a histogram

$$Q(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T}\delta_{\boldsymbol{x},\boldsymbol{x}^{(t)}}. \tag{A1}$$

The description length can therefore be written

$$L(P) = \langle \ell_P(\boldsymbol{x})\rangle_\chi = \frac{1}{|\chi|}\sum_{\{\boldsymbol{x}^{(t)}\}\in\chi}\langle \ell_P(\boldsymbol{x})\rangle_Q. \tag{A2}$$

For each dataset, $TQ(\boldsymbol{x})$ counts the number of appearances of the state $\boldsymbol{x}$. Thus, $\frac{T!}{\prod_{\boldsymbol{x}}(TQ(\boldsymbol{x}))!}$ counts the number of datasets $\{\boldsymbol{x}^{(t)}\}$ corresponding to a given histogram $Q(\boldsymbol{x})$. We can therefore rewrite the average over datasets as a weighted average over histograms,

$$L(P) = \frac{1}{|\chi|}\sum_{Q\in\mathcal{Q}}\frac{T!}{\prod_{\boldsymbol{x}}[TQ(\boldsymbol{x})]!}\langle \ell_P(\boldsymbol{x})\rangle_Q, \tag{A3}$$

where $\mathcal{Q}$ is the set of histograms generated by datasets in χ.

In the limit of large T, one can use Stirling's approximation to show that

$$\frac{T!}{\prod_{\boldsymbol{x}}[TQ(\boldsymbol{x})]!} \approx e^{TS(Q)}, \tag{A4}$$

where $S(Q)$ is the entropy of $Q(\boldsymbol{x})$ [Eq. (5)]. Thus, in the same limit, the sum in Eq. (A3) becomes dominated by the histogram $Q^*(\boldsymbol{x})$ with the largest entropy, such that

$$L(P) \approx \langle \ell_P(\boldsymbol{x})\rangle_{Q^*}. \tag{A5}$$

We recognize this as the cross-entropy between the maximum entropy distribution $Q^*(\boldsymbol{x})$ and the model $P(\boldsymbol{x})$, which can be decomposed into two terms

$$L(P) = S(Q^*) + D_{\rm KL}(Q^*||P). \tag{A6}$$

Thus, among all of the distributions $P(\boldsymbol{x})$ consistent with the measured features, the one that minimizes the description length $L(P)$ is the one with maximum entropy $Q^*(\boldsymbol{x})$ [3].

APPENDIX B: MAXIMUM ENTROPY DISTRIBUTION

Given a set of M features $\mathcal{F} = \{f_\mu(\boldsymbol{x})\}$, the maximum entropy model $P_{\mathcal{F}}(\boldsymbol{x})$ is defined by the Boltzmann distribution [2,13]. To derive this functional form, one begins with the Lagrangian

$$\mathcal{L}(P) = S(P) + \sum_{\mu=1}^{M}\lambda_\mu\left(\langle f_\mu(\boldsymbol{x})\rangle_{\rm exp} - \sum_{\boldsymbol{x}} f_\mu(\boldsymbol{x})P(\boldsymbol{x})\right) + C\left(1 - \sum_{\boldsymbol{x}} P(\boldsymbol{x})\right), \tag{B1}$$

where the parameters λ_μ are Lagrange multipliers that enforce the constraints $\langle f_\mu(\boldsymbol{x})\rangle = \langle f_\mu(\boldsymbol{x})\rangle_{\rm exp}$, and C ensures normalization. Setting the derivative $\frac{\partial\mathcal{L}}{\partial P(\boldsymbol{x})}$ equal to zero, one can solve for the functional form for the maximum entropy model in Eq. (6).

[1] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, New York, 2006).

[3] M. Feder, Maximum entropy as a special case of the minimum description length criterion (Corresp.), IEEE Trans. Inf. Theory **32**, 847 (1986).

[4] P. D. Grünwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, 2007).

[5] M. H. Hansen and B. Yu, Model selection and the principle of minimum description length, J. Am. Stat. Assoc. **96**, 746 (2001).

[6] T. Hastie, J. Friedman, and R. Tisbshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics (Springer, New York, NY, 2017).

[7] S. C. Zhu, Y. N. Wu, and D. Mumford, Minimax entropy principle and its application to texture modeling, Neural Comput. **9**, 1627 (1997).

[8] S. C. Zhu, Y. Wu, and D. Mumford, FRAME: filters, random fields, and minimax entropy towards a unified theory for texture modeling, in *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 1996), pp. 686–693.

[9] C. W. Lynn, Q. Yu, R. Pang, W. Bialek, and S. E. Palmer, Exactly solvable statistical physics models for large neuronal populations, Phys. Rev. Res. **7**, L022039 (2025).

[10] C. W. Lynn, Q. Yu, R. Pang, S. E. Palmer, and W. Bialek, Exact minimax entropy models of large-scale neuronal activity, Phys. Rev. E **111**, 054411 (2025).

[11] D. P. Carcamo and C. W. Lynn, Statistical physics of large-scale neural activity with loops, Proc. Natl. Acad. Sci. USA **122**, e2426926122 (2025).

[12] N. J. Weaver, J. I. Faskowitz, R. F. Betzel, and C. W. Lynn, Quantifying the compressibility of the human brain, arXiv:2510.16327.

[13] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[14] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics, Rev. Mod. Phys. **85**, 1115 (2013).

[15] H. C. Nguyen, R. Zecchina, and J. Berg, Inverse statistical problems: From the inverse Ising problem to data science, Adv. Phys. **66**, 197 (2017).

[16] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).

[17] J. Chayes, L. Chayes, and E. H. Lieb, The inverse problem in classical statistical mechanics, Commun. Math. Phys. **93**, 57 (1984).

[18] S. G. Brush, History of the Lenz-Ising model, Rev. Mod. Phys. **39**, 883 (1967).

[19] E. Schneidman, M. J. Berry II, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, Nature (London) **440**, 1007 (2006).

[20] A. Ashourvan, P. Shah, A. Pines, S. Gu, C. W. Lynn, D. S. Bassett, K. A. Davis, and B. Litt, Pairwise maximum entropy model explains the role of white matter structure in shaping emergent co-activation states, Commun. Biol. **4**, 210 (2021).

[21] R. E. Rosch, D. R. W. Burrows, C. W. Lynn, and A. Ashourvan, Spontaneous brain activity emerges from pairwise interactions in the larval zebrafish brain, Phys. Rev. X **14**, 031050 (2024).

[22] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, Collective behavior of place and non-place neurons in the hippocampal network, Neuron **96**, 1178 (2017).

[23] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry II, Searching for collective behavior in a large network of sensory neurons, PLoS Comput. Biol. **10**, e1003408 (2014).

[24] O. Marre, S. El Boustani, Y. Frégnac, and A. Destexhe, Prediction of spatiotemporal patterns of neural activity from pairwise correlations, Phys. Rev. Lett. **102**, 138101 (2009).

[25] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE **6**, e28766 (2011).

[26] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).

[27] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt *et al.*, An evolution-based model for designing chorismate mutase enzymes, Science **369**, 440 (2020).

[28] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).

[29] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, Proc. Natl. Acad. Sci. USA **103**, 19033 (2006).

[30] P. D. Dixit, Quantifying extrinsic noise in gene expression using the maximum entropy framework, Biophys. J. **104**, 2743 (2013).

[31] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, Transferable model for chromosome architecture, Proc. Natl. Acad. Sci. USA **113**, 12168 (2016).

[32] X. Lin, Y. Qi, A. P. Latham, and B. Zhang, Multiscale modeling of genome organization with maximum entropy optimization, J. Chem. Phys. **155**, 010901 (2021).

[33] G. Shi and D. Thirumalai, A maximum-entropy model to predict 3D structural ensembles of chromatin from pairwise distances with applications to interphase chromosomes and structural variants, Nat. Commun. **14**, 1150 (2023).

[34] J. J. Messelink, M. C. F. van Teeseling, J. Janssen, M. Thanbichler, and C. P. Broedersz, Learning the distribution of single-cell chromosome conformations in bacteria reveals emergent order across genomic scales, Nat. Commun. **12**, 1963 (2021).

[35] P. Farré and E. Emberly, A maximum-entropy model for predicting chromatin contacts, PLoS Comput. Biol. **14**, e1005956 (2018).

[36] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).

[37] A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, and A. M. Walczak, Dynamical maximum entropy approach to flocking, Phys. Rev. E **89**, 042707 (2014).

[38] W. Bialek, A. Cavagna, I. Giardina, T. Mora, O. Pohl, E. Silvestri, M. Viale, and A. M. Walczak, Social interactions dominate speed control in poising natural flocks near criticality, Proc. Natl. Acad. Sci. USA **111**, 7212 (2014).

[39] C. W. Lynn, L. Papadopoulos, D. D. Lee, and D. S. Bassett, Surges of collective human activity emerge from simple pairwise correlations, Phys. Rev. X **9**, 011022 (2019).

[40] E. Schneidman, S. Still, M. J. Berry, and W. Bialek, Network information and connected correlations, Phys. Rev. Lett. **91**, 238701 (2003).

[41] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, USA, 2021), Vol. 14.

[42] A. M. Ferrenberg and D. P. Landau, Critical behavior of the three-dimensional Ising model: A high-resolution Monte Carlo study, Phys. Rev. B **44**, 5081 (1991).

[43] R. G. Parker and R. L. Rardin, *Discrete Optimization* (Elsevier, Amsterdam, 2014).

[44] D. Jungnickel, *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics (Springer, Berlin, Heidelberg, 2005), Vol. 3.

[45] D. Edwards, *Introduction to Graphical Modelling*, Springer Texts in Statistics (Springer, New York, NY, 2012).

[46] M. Yuan and Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika **94**, 19 (2007).

[47] S. Epskamp, L. J. Waldorp, R. Mõttus, and D. Borsboom, The Gaussian graphical model in cross-sectional and time-series data, Multivar. Behav. Res. **53**, 453 (2018).

[48] T. Wang, Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen, FastGGM: An efficient algorithm for the inference of Gaussian graphical model in biological networks, PLoS Comput. Biol. **12**, e1004755 (2016).

[49] C. Uhler, Gaussian graphical models: An algebraic and geometric perspective, arXiv:1707.04345.

[50] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, The WU-Minn Human Connectome Project: An overview, Neuroimage **80**, 62 (2013).

[51] D. S. Bassett and O. Sporns, Network neuroscience, Nat. Neurosci. **20**, 353 (2017).

[52] C. W. Lynn and D. S. Bassett, The physics of brain network structure, function and control, Nat. Rev. Phys. **1**, 318 (2019).

[53] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan Jr., Maximum entropy models for antibody diversity, Proc. Natl. Acad. Sci. USA **107**, 5405 (2010).

[54] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, High-dimensional geometry of population responses in visual cortex, Nature (London) **571**, 361 (2019).

[55] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Elsevier, Amsterdam, 2016).

[56] C. Chow and C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inf. Theory **14**, 462 (1968).

[57] C. Moore and S. Mertens, *The Nature of Computation* (Oxford University Press, Oxford, 2011).

[58] O. J. Rosten, Fundamentals of the exact renormalization group, Phys. Rep. **511**, 177 (2012).

[59] C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris, Spontaneous behaviors drive multidimensional, brainwide activity, Science **364**, eaav7893 (2019).

[60] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, Host lifestyle affects human microbiota on daily timescales, Genome Biol. **15**, R89 (2014).

[61] H. Rowley, S. Baluja, and T. Kanade, Neural network-based face detection, IEEE Trans. Pattern Anal. Machine Intell. **20**, 23 (1998).

[62] Y. Zhang, R. Jin, and Z.-H. Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. **1**, 43 (2010).

[63] L. Deng, The MNIST database of handwritten digit images for machine learning research, IEEE Signal Proc. Mag. **29**, 141 (2012).

[64] P. D. Dixit, Thermodynamic inference of data manifolds, Phys. Rev. Res. **2**, 023201 (2020).

[65] S. J. Phillips, M. Dudík, and R. E. Schapire, A maximum entropy approach to species distribution modeling, in *Proceedings of the Twenty-First International Conference on Machine Learning* (ACM, New York, 2004), p. 83.

[66] A. Ratnaparkhi, Learning to parse natural language with maximum entropy models, Mach. Learn. **34**, 151 (1999).

[67] D. H. de Groot, S. X. M. Leonardo, M. Pachkov, and E. van Nimwegen, *Bonsai*: Tree representations for distortion-free visualization and exploratory analysis of single-cell omics data, bioRxiv:2025.05.08.652944 (2025).

[68] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Solution of "Solvable model of a spin glass", Philos. Mag.: J. Theor. Expt. Appl. Phys. **35**, 593 (1977).

[69] T. Tanaka, Mean-field theory of Boltzmann machine learning, Phys. Rev. E **58**, 2302 (1998).

[70] Y. Roudi, J. Tyrcha, and J. Hertz, Ising model for neural data: Model quality and approximate methods for extracting functional connectivity, Phys. Rev. E **79**, 051915 (2009).

[71] J. M. Beggs and D. Plenz, Neuronal avalanches in neocortical circuits, J. Neurosci. **23**, 11167 (2003).

[72] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry, and W. Bialek, Thermodynamics and signatures of criticality in a network of neurons, Proc. Natl. Acad. Sci. USA **112**, 11508 (2015).

[73] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, Coarse graining, fixed points, and scaling in a large population of neurons, Phys. Rev. Lett. **123**, 178103 (2019).

[74] L. Di Carlo, F. Mignacco, C. W. Lynn, and W. Bialek, Extended mean-field theories for networks of real neurons, arXiv:2504.15197.

[75] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, High-dimensional Ising model selection using ℓ_1-regularized logistic regression, Ann. Stat. **38**, 1287 (2010).

[76] E. Aurell and M. Ekeberg, Inverse Ising inference using all the data, Phys. Rev. Lett. **108**, 090201 (2012).

[77] V. K. Olsen, J. R. Whitlock, and Y. Roudi, The quality and complexity of pairwise maximum entropy models for large cortical populations, PLoS Comput. Biol. **20**, e1012074 (2024).

[78] S. Cocco and R. Monasson, Adaptive cluster expansion for the inverse Ising problem: Convergence, algorithm and tests, J. Stat. Phys. **147**, 252 (2012).

[79] S. Fujishige, *Submodular Functions and Optimization*, Annals of Discrete Mathematics (Elsevier, Amsterdam, 2005), Vol. 58.

[80] A. Krause and D. Golovin, Submodular function maximization, Tractability **3**, 3 (2014).

[81] M. Bauer, M. D. Petkova, T. Gregor, E. F. Wieschaus, and W. Bialek, Trading bits in the readout from a genetic network, Proc. Natl. Acad. Sci. USA **118**, e2109011118 (2021).