Breakdown of random matrix universality in Markov models

Faheem Mosam [1], Diego Vidaurre,[2,3] and Eric De Giuli [1]

[1]*Department of Physics, Ryerson University, M5B 2K3, Toronto, Canada*
[2]*Department of Psychiatry, Oxford University, OX3 7JX, United Kingdom*
[3]*Center for Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University, 8000, Denmark*

Biological systems need to react to stimuli over a broad spectrum of timescales. If and how this ability can emerge without external fine-tuning is a puzzle. We consider here this problem in discrete Markovian systems, where we can leverage results from random matrix theory. Indeed, generic *large* transition matrices are governed by universal results, which predict the absence of long timescales unless fine-tuned. We consider an ensemble of transition matrices and motivate a temperature-like variable that controls the dynamic range of matrix elements, which we show plays a crucial role in the applicability of the large matrix limit: as the dynamic range increases, a phase transition occurs whereby the random matrix theory result is avoided, and long relaxation times ensue, in the entire "ordered" phase. We furthermore show that this phase transition is accompanied by a drop in the entropy rate and a peak in complexity, as measured by predictive information [Bialek, Nemenman, and Tishby Neural Comput. **13**, 2409 (2001)]. Extending the Markov model to a Hidden Markov model (HMM), we show that observable sequences inherit properties of the hidden sequences, allowing HMMs to be understood in terms of more accessible Markov models. We then apply our findings to fMRI data from 820 human subjects scanned at wakeful rest. We show that the data can be quantitatively understood in terms of the random model, and that brain activity lies close to the phase transition when engaged in unconstrained, task-free cognition—supporting the brain criticality hypothesis in this context.

I. INTRODUCTION

A. Background and motivation

Complex systems typically display a broad spectrum of relaxation timescales [1]. This phenomenon is often linked to proposed critical behavior [1,2], in analogy with equilibrium statistical mechanics. However, equilibrium systems must be tuned to a critical point. For example, a ferromagnetic material has a phase transition between paramagnetism and ferromagnetism at the Curie temperature T_C. The system temperature T must be externally tuned to the vicinity of T_C in order to see the long-range scale-free correlations and broad relaxations characteristic of a critical point. Since there is ample evidence for critical behavior in out-of-equilibrium complex systems, we must either find a generic mechanism for self-tuning to a critical point or multiple context-dependent mechanisms for criticality, or challenge the equilibrium analogy entirely.

Bak, Tang, and Weisenfeld [3] proposed a simple mechanism for the ubiquity of critical behavior in slowly driven systems. In particular, they showed how in the limit of asymptotically slow driving compared to internal relaxation, generic systems can show scale-free behavior, thus mimicking equilibrium systems at a critical point. While indeed generic, this theory does not predict different universality classes that would distinguish various complex systems, characterized, for example, by critical exponents [4]. It is also of questionable relevance to biological systems, where driving is often not slow compared to response. Thus, even if the evidence of

criticality in complex systems is accepted, the means by which criticality is obtained is still unclear.

Recent work has instead sought context-specific mechanisms for self-organized criticality, particularly in the brain [5,6] and in glassy systems [7]. While of clear relevance in their particular domains, these works do not address the question of generic emergence of long timescales.

Here we study this problem in the simple but general setting of discrete Markov models. Recent work has focused on the eigenvalue spectra of Markov models, most often in the symmetric case [8,9], on first-passage times [10], and on general properties of Markov ensembles [11,12]. Here we define an appropriate ensemble of Markov models that spans the full dynamic regime from systems that just produce random noise, to those that are fully deterministic. Although we make use of previous results from random matrix theory [13–15], our ensemble is distinct from those considered previously [11,12].

We show that these regimes are separated by a transition with many aspects of a phase transition, whose location can be identified by a breakdown of large-N random matrix theory results. We examine the phases in terms of network- and information-theoretic quantities, and show in particular that complexity peaks near the transition. We then extend these results to Hidden Markov models (HMMs), in which a hidden Markovian system outputs "observable" sequences. Despite being more difficult to analytically characterize, the HMM inherits many properties of the original simple Markov model, which aids its interpretation. Finally, we consider neural fMRI

data from [16] that was previously fitted to a HMM. We find that the fMRI data lie very close to the phase transition as predicted analytically. This can be interpreted as support of the brain criticality hypothesis [5,6,17–23].

B. Continuous-time dynamics

To show why prediction of a broad spectrum of timescales is challenging, and to motivate our model, we consider first a generic continuous-time linear dynamics that is often applied to neural data [24]:

$$\partial_t x_i = -x_i + \tilde{M}_{ij} x_j + \xi_i, \tag{1}$$

where x_i is a dynamical variable, $\tilde{M}_{ij}$ an interaction matrix, and ξ_i noise, with $i = 1, \ldots, n$. For example, (1) is a crude model for a neural network, where the x_i could be the firing rate of the ith neuron. The first term on the right-hand side is a stabilizing term, that will send $x_i \to 0$ in the absence of any external stimuli. In the normalization of (1), this occurs over a timescale of order unity. The second term on the right-hand side represents the interactions from other neurons: if $\tilde{M}_{ij} > 0$, then neuron j is excitatory for neuron i, while in $\tilde{M}_{ij} < 0$, then neuron j is inhibitory for neuron i. The final term ξ_i is a noise term, which could represent input from the external world, for example from the part of the brain that is not explicitly modeled.

The dynamics of (1) depends largely on a balance between the stabilizing effect of the first term, and the interactions that could lead to self-sustained activity. Since (1) is linear, this balance is encoded in the spectrum of the matrix $\tilde{M}$, or more precisely the matrix $-\delta_{ij} + \tilde{M}_{ij}$ that includes the first two terms on the right-hand side of (1). (Here δ_{ij} is a Kronecker delta, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise).

The relaxation times τ_i are related to the eigenvalues $\tilde{\lambda}_i$ of $\tilde{M}_{ij}$ by [24]

$$\tau_i = \frac{1}{1 - \text{Re}[\tilde{\lambda}_i]}, \tag{2}$$

where the stability limit $\text{Re}[\tilde{\lambda}_c] = 1$ is due to the unit coefficient of the stabilizing term in (1). Generically, the matrix $\tilde{M}_{ij}$ is not symmetric and has no special symmetries. Then, in the large n limit of practical interest, its spectrum will generically tend towards the Girko law in which the eigenvalues are uniformly spread within a circle in the complex plane, say, of radius r [13–15]. Since relaxation times become large only when the real part of the eigenvalues approaches 1, the system is then either stable, if $r < 1$, with a finite maximum timescale, or unstable, if $r > 1$, except precisely at the critical point $r = 1$. Thus universality of random matrices seems to be at odds with emergence of long timescales, without fine-tuning [24].

In reality, $\tilde{M}$ is usually not fixed but rather evolves slowly. For example, in neural networks it could represent the strength of synaptic connections. Then we can ask if some dynamics will lead $\tilde{M}$ to generically show long timescales. It was recently shown in a detailed study of the symmetric case [24] that a regulation mechanism based on the overall activity can lead to an accumulation of long timescales. While such an approach can engage with problem-specific dynamical details,

it remains a challenge to identify universal mechanisms in which long timescales emerge.

The fundamental difficulty in finding a generic pathway to long timescales in (1) is that the stability limit $\text{Re}[\tilde{\lambda}_c] = 1$ is not universal, but is simply when the interactions determined by $\tilde{M}$ overwhelm the stabilizing force in (1). In the present work we propose an alternative strategy. Instead of considering the dynamics directly, as in (1), we consider the motion in phase space. That is, instead of considering the evolution of the $\{x_i(t)\}$, we consider the probability distribution over those variables, $\rho(x_1, x_2, \ldots, x_n; t)$. The latter evolves according to a master equation. Schematically this takes the form

$$\rho(y, t + dt) = \sum_x M_{yx} \rho(x, t), \tag{3}$$

where M is the transition matrix giving the probability for a transition from state x to state y, where, e.g., $x = (x_1, x_2, \ldots, x_n)$ encodes the state of the entire system. Equation (3) is written in a discrete notation, but carries over without difficulty to the fully continuous case. For a derivation of (3) from (1) see the Appendix.

The transition matrix is related to the terms in (1). If each x_i has n_x different states, then M is an $N \times N$ matrix, with $N = (n_x)^n$. Thus for a fully continuous system, M becomes an infinite matrix, an operator. To avoid technical difficulties, we stick here and in the following to the discrete case, where, however, N may be very large. We note that $\sum_y M_{yx} = 1$ and each $M_{yx} \geqslant 0$. Such matrices are called left-stochastic matrices.

The Markovian dynamics (3) is attractive because of its manifest linearity, but it has two serious disadvantages: first, it takes place in an enormous space; and second, the details of the dynamics are hidden inside the definition of M. Yet, if our interest is in extracting generic properties of complex systems, we can simply treat M as a random matrix and look for universal properties in the large N limit. This will be the perspective taken here. It is most appropriate in a regime in which the underlying dynamical system has very strong interactions and is sampled over a timescale much larger than the microscopic timescale. Indeed, under these assumptions, the phase space dynamics will have scrambled local structural correlations arising from continuity of the underlying dynamical process. In some sense, this is the opposite limit to that which favors analysis based on (1).

What is gained from working in phase space? The key distinction between (1) and (3) is that while $\tilde{M}$ was relatively unconstrained, M is instead a stochastic matrix. For simplicity we assume that it is irreducible and aperiodic.[1] This entails that M obeys the Perron-Frobenius theorem, in particular its largest eigenvalue has unit modulus. Thus M has a natural scale, unity, induced from probability conservation.

In this work we consider a generic model for M, with a "temperature" scale ϵ that controls its heterogeneity. We will show that the properties of M change quite dramatically at a

---

[1]Periodicity complicates the analysis of long-time behavior since there is then no unique stationary state. In our random model, such behaviors are not encountered so there is no need to complicate matters here.

certain scale ϵ_c, which in a sense to be explained below, corresponds to the breakdown of large-N random matrix theory. We show that in the entire low-temperature phase, M shows large relaxation times.

In a second part, we promote (3) to a hidden Markov model, where an observable sequence $\{o(t)\}$ is produced by an additional emission matrix O_{ox} that depends on the "hidden" state $x(t)$. We will find that many properties of the observable sequences mirror those of the underlying hidden sequences. This can be useful in practice where one has direct access to only the observable sequences. We will explore in particular how the observable sequences show the signature of the transition.

II. MARKOV MODELS

Markov models arise naturally in a variety of contexts, not necessarily related to any underlying dynamical system such as (1). We consider a system with N states. We can always write M_{ab} in terms of a more primitive matrix Q_{ab} such that $M_{ab} = Q_{ab}/\sum_c Q_{ac}$; then $Q_{ab} \geq 0$ but do not require any special normalization.

It has been proved in great generality that M_{ab} has universal features in the large N limit [13–15]. Let the Q_{ab} be identically and independently distributed, with bounded density, mean μ, and finite variance σ^2. Then as $N \to \infty$ the spectrum of M_{ab} converges to the uniform law on the disk $|\lambda| < \lambda_c$ in the complex plane (Girko's law), with

$$\lambda_c = \sigma/(\mu\sqrt{N}). \tag{4}$$

In practice, this law works well for large but finite N, unless a transition intervenes, to be discussed presently.

Solving (3) in terms of the left and right eigenvectors $w_{a,\lambda}$ and $v_{a,\lambda}$ and the stationary right eigenvector $\pi_a = v_{a,1}$, we have

$$\rho_a(t) = \pi_a + \sum_{\lambda \neq 1} A_\lambda \lambda^t v_{a,\lambda}, \tag{5}$$

with $A_\lambda = \Sigma_b w_{b,\lambda} \rho_b(0)$. Since $\lambda^t = e^{t \log(\lambda)}$ we see that the relaxation times are

$$\tau_\lambda = -\frac{1}{\log |\lambda|}. \tag{6}$$

For sufficiently large N, $\lambda_c < 1$ and all relaxation times are finite. In this regime, the asymptotic large N random matrix theory (RMT) result applies. What happens when $\lambda_c > 1$? Since M is a stochastic matrix, it cannot have eigenvalues with magnitude larger than unity, thus the large-N result must break down, and the spectrum must reorganize. It is this collision of random matrix theory with Perron-Frobenius that is the main subject of this article.

A. Random matrix ensemble

Since our aim is to elucidate generic properties of Markov models, we need to identify an appropriate ensemble of the Q matrices. Each element Q_{ab} corresponds to an (unnormalized) transition rate between two states. As a simple model, suppose that the matrix elements Q_{ab} are identically and independently

formed from multiplicatively accumulating many factors.[2] Then, if the factors are sufficiently independent, Q_{ab} will have a lognormal distribution [25], viz.,

$$\mathbb{P}(Q_{ab}) \propto \frac{1}{Q_{ab}} e^{-\epsilon \log^2(Q_{ab}/q)}, \tag{7}$$

where ϵ and q are parameters of the distribution. The normalized fluctuations of $\log Q_{ab}$ have magnitude

$$\frac{\langle Q_{ab}^2 \rangle}{\langle Q_{ab} \rangle^2} - 1 = e^{1/(2\epsilon)} - 1. \tag{8}$$

We call ϵ the temperature of the matrix. The parameter q controls the overall scaling of the elements via

$$\langle \log Q_{ab} \rangle = \log q. \tag{9}$$

Since q drops out of the M matrix, it will not be varied in what follows. Note that Q_{aa} is sampled from the same distribution as Q_{ab} for $a \neq b$. Since (7) implies that $Q_{aa} \neq 0$, this also implies that the Markov chain is aperiodic.

For a matrix Q_{ab} with average $\overline{Q} = \frac{1}{N^2} \sum_{a,b} Q_{ab}$, we define its heterogeneity

$$h(Q) = \frac{1}{N^2} \sum_{a,b} \log^2(Q_{ab}/\overline{Q}), \tag{10}$$

which measures the dispersion of its elements about its mean. Notice that heterogeneity is invariant if $Q_{ab} \to Q'_{ab} = \overline{Q}^2/Q_{ab}$. It is an appropriate measure for matrices (and tensors [26]) with positive entries.

Over the ensemble defined by (7), we have $\langle \overline{Q} \rangle = q e^{1/(4\epsilon)}$. so that $\langle h \rangle = (2\epsilon)^{-1} + (16\epsilon^2)^{-1}$. This implies that "hot" matrices with $\epsilon \gtrsim 1$ have a small heterogeneity, and "cold" matrices with $\epsilon \lesssim 1$ have a large heterogeneity. Intuitively, ϵ measures how uniformly distributed the transition probabilities are: when $\epsilon \to \infty$, all $Q_{ab} \to \overline{Q}$, while at small ϵ the Q_{ab} will vary over a huge range.

Combining (4) with (8), we see that $\lambda_c = \sqrt{e^{1/(2\epsilon)} - 1}/\sqrt{N}$, thus we predict a critical temperature

$$\epsilon_c(N) = 1/[2\log(N+1)], \tag{11}$$

where $\lambda_c = 1$. The circular law cannot apply for $\epsilon < \epsilon_c$. The relevant eigenvalue scale is therefore

$$\tilde{\lambda} = \min(1, \lambda_c), \tag{12}$$

so that for $\epsilon < \epsilon_c$, $\tilde{\lambda} = 1$.

This phenomenon is illustrated in Fig. 1(a), which shows the distribution of $|\lambda|/\tilde{\lambda}$ for $N = 128$ and a large range of ϵ; qualitatively identical results hold for N down to 16. The RMT result is

$$\mathbb{P}(|\lambda|) = 2\frac{|\lambda|}{\lambda_c}, \qquad |\lambda| < \lambda_c \tag{13}$$

and holds well for $\lambda > \lambda_{\min}$, even at finite N, for $\epsilon \gtrsim \epsilon_c$. The lower cutoff $\lambda_{\min}$ simply corresponds to the value of λ for

---

[2]This is the analog of applying the central limit theorem to motivate a Gaussian distribution, but here since the $Q_{ab} \geq 0$, we must multiply factors rather than sum them.

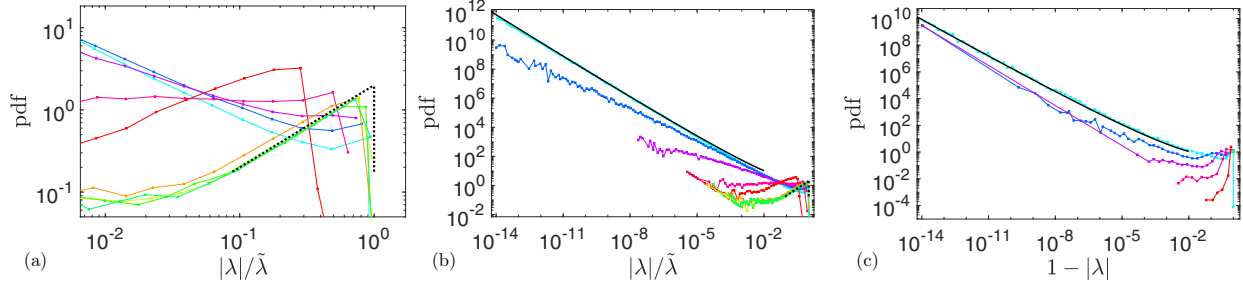FIG. 1. Transition matrix spectra for $N = 128$ as a function of ϵ, from $\epsilon/\epsilon_c = 10^{2.4}$ (green) through $\epsilon/\epsilon_c = 10^0$ (pink) to $\epsilon/\epsilon_c = 10^{-2.4}$ (cyan). ϵ varies logarithmically by factors of 4; complete labels are shown in Fig. 5. (a) Normalized spectrum of eigenvalue magnitude, compared to the random matrix theory prediction $\mathbb{P}_{RMT}(|\lambda|) = 2|\lambda|$, $|\lambda| < 1$. The latter holds for $\epsilon \gtrsim \epsilon_c$. (b) $\mathbb{P}(|\lambda|)$ over a wide range, showing a divergent tail $\mathbb{P}(|\lambda|) \sim 1/[|\lambda| \log(1/|\lambda|)^{3/2}]$ (black) at $\epsilon/\epsilon_c = 10^{-2.4}$. (c) Spectrum near $|\lambda| = 1$. In this regime, at $\epsilon/\epsilon_c = 10^{-2.4}$ we find $\mathbb{P}(|\lambda|) \sim 1/\{(1 - |\lambda|) \log[1/(1 - |\lambda|)]^{5/2}\}$ (black).

which there is typically only one eigenvalue left in the disk of radius $\lambda_{\min}$, which turns out to be $\lambda_{\min}/\lambda_c = 1/\sqrt{N}$.

More interesting is the inevitable failure of the RMT result for $\epsilon \lesssim \epsilon_c$. We find that when $\epsilon \lesssim \epsilon_c$, $\mathbb{P}(|\lambda|)$ develops a diverging tail at small $|\lambda|$, shown in Fig. 1(b). At the smallest ϵ, our data are consistent with a barely convergent form $\mathbb{P}(|\lambda|) \sim 1/[|\lambda| \log(1/|\lambda|)^{3/2}]$ over 10 orders of magnitude in $|\lambda|$. Any such form implies that as the matrices increase in size, arbitrarily small values of $|\lambda|$ will be found. Assuming this form, the smallest magnitude eigenvalue in a system of size N is $\sim e^{-(2N)^2}$; this corresponds to a short relaxation time $\tau_{\min} \sim 1/(2N)^2$.

Hidden in Figs. 1(a) and 1(b) is the behavior of $\mathbb{P}(|\lambda|)$ near $|\lambda| = 1$, shown in Fig. 1(c). Surprisingly, this also shows divergent behavior: at the smallest ϵ we find $\mathbb{P}(|\lambda|) \sim 1/\{(1 - |\lambda|) \log[1/(1 - |\lambda|)]^{5/2}\}$ over 10 orders of magnitude. Assuming this form, the largest magnitude eigenvalue is $|\lambda| \approx 1 - e^{-(2N/3)^{2/3}}$, aside from the Perron-Frobenius eigenvalue $|\lambda_{PF}| = 1$. Hence the system has a huge relaxation time $\tau_{\max} \sim e^{(2N/3)^{2/3}}$, diverging exponentially fast in the system size. For intermediate $\epsilon < \epsilon_c$, our data do not resolve the details of this divergence with complete confidence, but it is clear that $\log \tau_{\max} \sim N^\alpha$ for some $\alpha > 0$. This generic presence of large relaxation times in our model is the main result of this work.

We can also look at the structure of the spectra in the complex plane, shown in Fig. 2 for 1000 matrices of size $N = 64$, and several values of ϵ. Aside from the disk expected from random matrix theory, and lone outliers at $\lambda = 1$, we note a key feature: there is a bicycle spoke structure that develops at

small ϵ. The spokes are situated along roots of unity $e^{ik\pi/n}$, for $n = 2, 3, 4, \ldots$.

In the present work we will refer to ϵ_c as a phase transition. This term is used somewhat loosely, since discontinuous behavior of observables may only occur in the asymptotic large N limit (taken jointly with $\epsilon \to \epsilon_c$). However, we observe smooth changes with N even at the modest values of N that we considered.

Note that the specific forms of the eigenvalue tails near $\lambda = 0$ and $|\lambda| = 1$ are empirical fits; similar tails have been observed before in the spectra of random symmetric matrices with heavy-tailed element distributions [8,9].

B. Network interpretation

To illustrate how the transition matrices change as ϵ is varied, it is useful to consider the directed network formed by considering states as nodes and transitions as directed edges. In our model, all states are connected with some probability, thus the network is always fully connected: over a long enough timescale, transitions between any states are possible. However, the transition elements can vary over a wide range. We thus construct the networks with edges above a threshold m_0, which corresponds to choosing a maximum observation timescale. Very roughly, this corresponds to the network observed on times less than $\sim 1/m_0$. We set $m_0 = 10^{-3}$ in what follows. Typical networks for $N = 32$ are illustrated in Fig. 3, where a spring-based force algorithm was used to determine the vertex positions. Edge thicknesses and colors are determined by their weight.
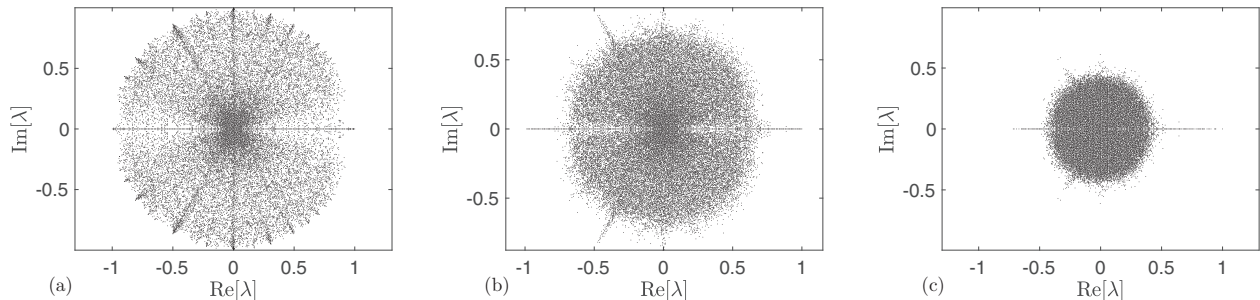

FIG. 2. Spectra of 1000 random Markov matrices in the complex λ plane, with $N = 64$, for (a) $\epsilon/\epsilon_c = 0.03$; (b) $\epsilon/\epsilon_c = 0.5$; (c) $\epsilon/\epsilon_c = 2$.
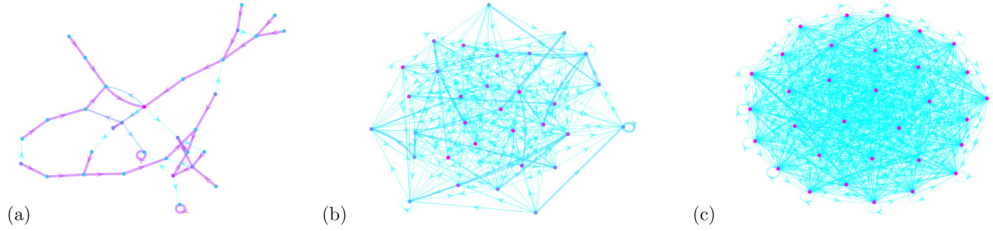
FIG. 3. Illustrative networks at with $N = 32$ at temperatures (a) $\epsilon/\epsilon_c = 10^{-2.4}$; (b) $\epsilon/\epsilon_c = 10^{-0.6}$; (c) $\epsilon/\epsilon_c = 10^{0.6}$. Edges are defined for matrix elements $M_{ab} > m_0 = 10^{-3}$ and are colored based on their weight. Nodes are colored based on their out-degree.

Several features are apparent. At high temperature, the network is nearly fully connected, and therefore has many loops. Statistically, vertices are equivalent. As the temperature is lowered, the graph becomes sparser. Slightly below the critical temperature, loops of various scales are present, thus allowing nontrivial temporal behavior on a range of scales. Finally, at the lowest temperature studied, the network is very nearly a tree, which thus will lead to a trivial dynamic behavior.

At low temperature, the network depends on the threshold. If the threshold were to be lowered, the networks would become denser, revealing rare transitions between states. At the lowest temperatures, the very wide range of timescales implies that these networks will continue to have structure as the threshold is made ever smaller.

The networks can be analyzed with standard measures [27]. For example, the out-degree of a node is the number of edges leaving it. Its distribution, shown in Fig. 4 for $N = 128$, clearly shows the transition: at temperatures above ϵ_c, it is nearly a Dirac δ function peak at its average value, since the network is nearly fully connected. It broadens continuously as ϵ decreases, giving an approximately exponential tail of well-connected nodes.

We also investigated the node centrality (not shown). For $\epsilon > \epsilon_c$, the distribution of centrality is peaked around a typical value, while for smaller ϵ it develops a tail of large centrality nodes and a weakly divergent tail of nodes with small centrality.

III. APPLICATIONS

We now consider several applications of the random Markov ensemble. First, we consider the sequences produced by the Markov model in the light of information theory. We will show that the transition identified above has a signal in information-theoretic quantities, namely, entropy and predictive information [28,29]. Then we extend the Markov ensemble to an ensemble of Hidden Markov models, which are extensively used in sequence analysis and neuroscience. Finally, we apply the Markov ensemble to neural data and show that several measured quantities can be quantitatively predicted from the random model.

A. Information theory and complexity

Since we study Markov models as complex systems, we can also look at measures of complexity using information theory. In this section we will consider three quantities: the Shannon entropy, whose expression for stationary Markov models we review; the predictive information of [28,29], whose construction we review; and Zipf's law.

1. Shannon entropy

The Shannon entropy $H = -\langle \log \mathbb{P}(X) \rangle$ of a probability distribution quantifies the amount of information gained upon measurement of a random variable [30]. (We define the entropy in base e.) We can apply this to the sequences of states visited by a Markov model, for example, $X = (ABAADDEABC\ldots)$, etc., which is an element in the space of sequences. Labeling the sites visited in a sequence as $x(1), x(2), \ldots$, we write $H(t)$ for the entropy of a sequence of length t. Consider a discrete and stationary Markov process. For simplicity, we assume that the initial conditions are sampled from the stationary distribution π_a, which is assumed to exist. Then we have

$$
\begin{aligned}
H(t) &= -\langle \log \mathbb{P}(x(1), x(2), \ldots, x(t)) \rangle \\
&= -\langle \log[\mathbb{P}(x(1))\mathbb{P}(x(2)|x(1)) \cdots \mathbb{P}(x(t)|x(t-1))] \rangle \\
&= -\sum_a \pi_a \log \pi_a - (t-1) \sum_{a,b} \pi_b M_{ab} \log M_{ab} \\
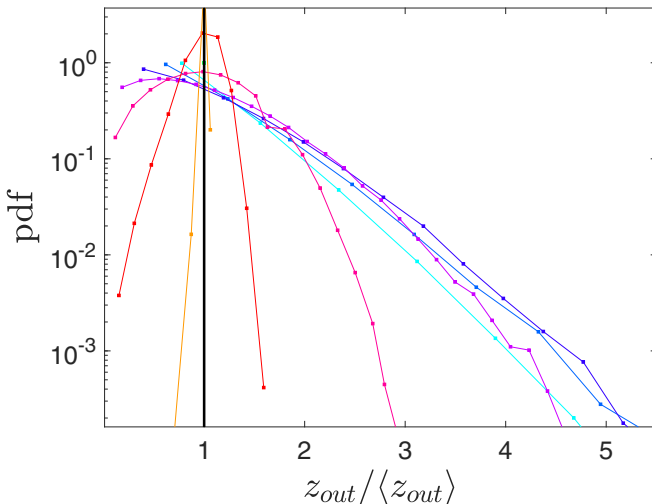&= H_\pi + (t-1)H_d,
\end{aligned}
\tag{14}
$$

FIG. 4. The distribution of node out-degree at $N = 128$ and various ϵ from $\epsilon/\epsilon_c = 10^{-2.4}$ (cyan) to $\epsilon/\epsilon_c = 10^{1.2}$ (orange). The distribution shows a transition from being pointlike for $\epsilon > \epsilon_c$ to having an exponential tail at large values at smaller ϵ.
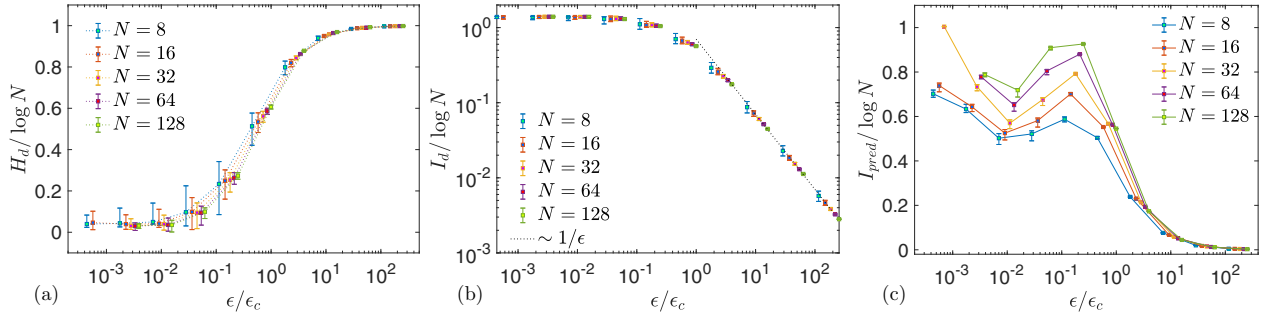
FIG. 5. Information-theoretic quantities. (a) The Shannon entropy rate of state visitation sequences in the Markov chain at indicated N; (b) information rate, showing its increase as ϵ is decreased from large values; (c) predictive information, peaking at an intermediate ϵ. Error bars in (a) and (b) correspond to 20th and 80th percentile ranges over 1000 distinct Markov chains (samples) at each parameter value, while error bars in (c) correspond to true measurement errors from 300 distinct Markov chains (samples) at each parameter value.

where we used the Markov property to expand the probability of a sequence, and we defined the entropy H_π of the stationary distribution π_a. At large times $H(t)$ is dominated by the *entropy rate*

$$H_d = \lim_{t \to \infty} \frac{1}{t} H(t). \tag{15}$$

If the Markov chain is irreducible but the initial conditions are not sampled from the stationary distribution, then there will be transient corrections to (14).

H_d is plotted in Fig. 5(a) over the full range of temperature and system size N.[3] When H_d is normalized by its maximal value $\log N$, and temperatures are normalized by $\epsilon_c(N)$, then the curves collapse at large ϵ and also at small ϵ, up to small corrections logarithmic in N (no power of $\log N$ produces a collapse at all ϵ, indicating a crossover between two functional forms). The normalized entropy rate tends to unity at large temperature, indicating that transition sequences are indistinguishable from random noise, while at the lowest temperatures it is very nearly 0, indicating that sequences are becoming nearly deterministic.

Equivalently, we can define the information rate $I_d = (\log N - H_d)/\log 2$ such that I_d is the typical amount of information contained in the probability distribution per entry of the sequence, in bits [31]. Again, this holds for stationary conditions. This is shown in Fig. 5(b) on logarithmic axes, highlighting the increase in information rate as ϵ is decreased. The dotted line shows the function $I_d/\log N \propto \epsilon_c/\epsilon$ which tracks this increase for $\epsilon \gtrsim \epsilon_c$.

The error bars shown in Figs. 5(a) and 5(b) correspond to 20th and 80th percentile ranges over 1000 distinct Markov chains (samples) at each parameter value.[4] We see that these bars decrease as N increases, so that the normalized entropy rate $H_d/\log N$ is a *self-averaging* quantity: it doesn't fluctuate from sample-to-sample, in the large size limit. This is a partial *a posteriori* justification for the lognormal distribution of matrix elements, since it implies that ϵ controls the entropy, at least for large systems. Thus, for example,

if we measure the normalized entropy rate in some system under some dynamic process, then we can infer how ϵ varies.

2. Complexity

Entropy distinguishes order from randomness, and thus captures the emergence of structure in Markov models, but it does not necessarily measure complexity. Intuitively, from Fig. 3, one expects that models at very low temperature are not complex because they are too deterministic, while models at high temperature are not complex because they are completely random. Can this intuition be quantified?

Grassberger [32] suggested that the complexity of a sequence can be measured by how quickly the entropy rate $H[x(1), x(2), \ldots, x(t)]/t$ attains its asymptotic value. This notion was formalized by Bialek, Nemenman, and Tishby [28,29] with the notion of predictive information. For a stationary process one asks how well the "past" $(x(-t), x(-t + 1), \ldots, x(-1))$ predicts the "future" $(x(0), x(1), \ldots, x(t - 1))$. This is quantified by the mutual information [30] between past and future, $I(t, t') = H(t) + H(t') - H(t + t')$ where we are using the fact that the process is stationary, so that this depends on only the length t. Bialek *et al.* then define the predictive information as

$$I_{\mathrm{pred}}(t) = \lim_{t' \to \infty} I(t, t')/\log 2, \tag{16}$$

which thus measures how well the past t states predict the entire future trajectory of the system.

Since H_d is the asymptotic extensive part of $H(t)$, the latter can be written $H(t) = tH_d + H_1(t)$ where $H_1(t)$ is the nonextensive part. Clearly dependence on H_d drops out of I_{pred}, showing the insufficiency of H_d to measure predictive power, and thus complexity, in this theory [this statement holds for general stationary processes, not necessarily Markov ones that follow (14)]. Consistent with the arguments of Grassberger, complexity is measured not by the asymptotic state but by the approach to it.

For a discrete and stationary Markov process, under the same assumptions as (14), it follows from (14) that

$$I(t, t') = H_\pi - H_d, \tag{17}$$

which is independent of both t and t', a nongeneric property [28,29]. The predictive information is plotted in Fig. 5(c). We

---

[3]For simplicity we consider only ergodic samples. Ergodicity can break down at small ϵ due to floating-point issues.

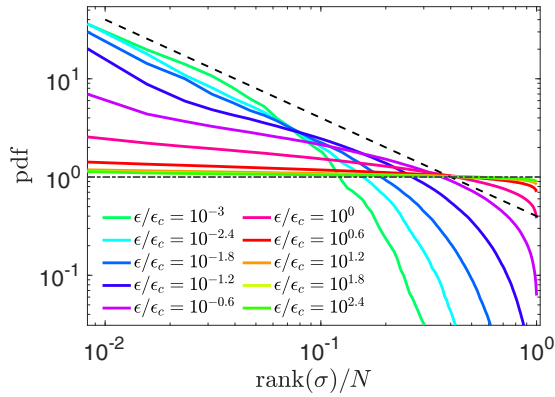[4]The true error bars in the measurements are then smaller by a factor of $\sim 1/\sqrt{1000} \sim 1/32$.

FIG. 6. Probability distribution of the rank of states (denoted by A in the axis label) for $N = 128$. Zipf's law is a power-law relation $\mathbb{P}(r) \sim 1/r$, as indicated by the dashed line.

confirm that it has a peak at intermediate $\epsilon \sim 0.1\epsilon_c$, although this precise location is not resolved in our data. This furthermore supports the notion that the high and low-temperature phases are separated by a phase transition, rather than just a cross-over.

Curiously, we find that I_{pred} also increases as ϵ becomes very small. The interpretation of this result we leave for future work.

3. Zipf's law

Finally, often discussed in the context of sequence complexity is Zipf's law [33–37]. Consider the sequence of states visited by the Markov model. Zipf's law states that when states A are ranked in terms of decreasing frequency of appearance, their probability is inversely proportional to their rank: $\mathbb{P}(r(A)) \sim 1/r(A)$, where $r(A)$ is the rank of the state A. Zipf's law is found to hold to a good approximation in human language texts (when "states" are words), as explained by various theories [33–37]. This quantity is shown for our $N = 128$ data in Fig. 6 on logarithmic axes. For $\epsilon > \epsilon_c$, the curve is nearly flat: all sites are visited with the same frequency. Below ϵ_c, the curve develops a nontrivial shape, including approximate power-law regions, and for very small ϵ one finds Zipf's law, indicated by the dashed line.

B. Hidden Markov models

In most dynamical systems, we have only incomplete knowledge of the detailed configuration of the system components. Often we can consider that the system undergoes a Markovian evolution, as above, but we have access to only a sequence of observables, which are outputted as the internal system changes state. This leads to the definition of a Hidden Markov model [38,39]: we have (1) a set of internal "hidden" states, $\{1, 2, \ldots, N\}$ with an internal transition matrix M_{ab}, as above; and (2) in addition, a set of observable states $\{1, 2, \ldots, T\}$ (distinct from the hidden states), and an emission matrix O_{aB} giving the probability that when the system is in hidden state a it outputs observable B. It is supposed that one observable symbol is outputted between each transition of hidden states.

For example, while an infant may be able to produce only a small number of distinguishable noises, its hidden mental state may be significantly richer. Hidden Markov models (HMMs) have found application in various domains, particularly in bioinformatics [39,40], natural language processing [41], and, most recently, neuroscience [16,42–45]. HMMs are also the simplest level of the Chomsky hierarchy when applied to probabilistic grammars, and thus serve as toy models for syntactic structure [26,46,47].

All of our earlier considerations apply to the sequence of hidden states visited by a HMM. When we add an emission matrix, then we can also investigate quantities that depend only on the observed states. The same arguments that led us to consider a lognormal distribution for the elements of the Markov matrix also apply to the emission matrix. We introduce a primitive matrix P and write $O_{aB} = P_{aB}/\sum_C P_{aC}$. We consider the matrix P to be drawn from a lognormal distribution, now with a "surface" temperature ϵ_s, which we fix to $\epsilon_s = 0.01$. We also fix $T = 12$.

We will investigate the sequences produced by the HMM both in terms of their entropy, and in terms of their complexity.

1. Entropy

We consider first the Shannon entropy rate of observed sequences, H_s. There is no simple expression for the entropy rate in terms of M and O, although rigorous algorithms exist for its estimation [48]. Here we estimate H_s by sampling, using the method of Grassberger to eliminate sampling biases [49]. First we define the k-gram surface entropy rate $H_s(k) = \frac{1}{k}H(o(1), o(2), \ldots, o(k))$ where the $o(j)$ are the observable symbols. The asymptotic entropy rate can be estimated directly by the limit $H_s(k)$ as $k \to \infty$, but it is also estimated by the differential entropy rate

$$\delta H_s(k) = (k + 1)H_s(k + 1) - kH_s(k), \qquad (18)$$

and the latter converges faster [50]. We estimate these entropy rates by constructing 300 different Markov models at each value of N and ϵ, and sampling a sequence of total length $\sim 10^5$ for each sample, at each parameter value. This allows us to estimate $H_s(k)$ for $k \leqslant 4$.[5]

The resulting $\delta H_s(3)$ and $\delta H_s(2)$ are plotted in Fig. 7 as solid and dotted lines, respectively. First, we notice that for $N \leqslant 16$ the two differential entropy rates are very nearly equal, and thus close to their asymptotic value everywhere. For larger N, the rates are nearly converged for $\epsilon > \epsilon_c$ but display some finite-k effects at smaller ϵ. These effects in fact indicate correlations in the sequences [28], studied below. Here we focus on the systematic trends with ϵ and N.

Consider a sequence outputted by any given HMM. It has two sources of randomness: that in M, and that in O. In our case, since $\epsilon_s = 0.01$ is relatively small, the O matrix is very sparse: given the hidden states, the observed sequence is nearly deterministic. Variation in δH_s is thus due to changes in the physics of the hidden states.

Consider first $\epsilon > \epsilon_c$. If the outputted sequence is completely random noise, then we will have $\delta H_s \approx \log T$, as

---

[5]As a crude estimate, since the kth entropy rate has a phase space of size T^k, we are limited to $T^k \ll 10^5$ or $k < 4.6$.
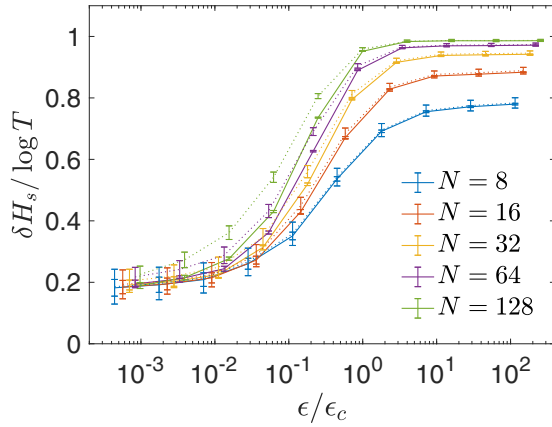
FIG. 7. Differential surface Shannon entropy rates $\delta H_s(k)$ for the Hidden Markov model with $k = 3$ (solid) and $k = 2$ (dotted), at indicated N. Error bars correspond to 20th and 80th percentile ranges over 300 distinct Markov chains (samples) at each parameter value. Here $T = 12$ and $\epsilon_s = 0.01$.
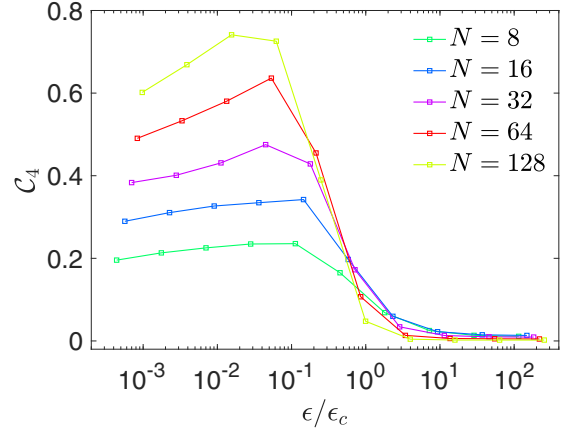

FIG. 8. Complexity measure for observable sequences in the Hidden Markov model, with $k = 3$. Consistent with the predictive information measure for the hidden states, this peaks at an intermediate temperature below ϵ_c.

indeed we find for large enough N. As N is lowered, δH_s also decreases. The interpretation of this regime is as follows: even if the hidden state physics is completely random, outputted sequences can have structure simply because O need not output each observed symbol with the same frequency. For example, if the symbol A is over-represented in the output of O, then this will impact the observed entropy rate, even for uniformly random hidden states. This is why δH_s can be small even when $\epsilon > \epsilon_c$. However, there is a competing averaging effect: when the hidden states are sampled uniformly, each O element only contributes in the form $\frac{1}{N} \sum_a O_{aB}$, which is averaged over the hidden states. When N becomes large, this is subject to the law of large numbers and this becomes equal to $\langle O_{aB} \rangle$, which is then independent of B (in our model). Thus, finally, for large N and $\epsilon > \epsilon_c$, the entropy becomes again that of a uniform random variable and tends to $\log T$ as $N \to \infty$, as observed.

As we decrease ϵ, the observed entropy rate always drops substantially from its value at ϵ_c: the physics of hidden states always plays an important role. We notice that at sufficiently small ϵ, the curves for different N collapse on top of one another. The asymptotic small ϵ entropy rate depends only on T and ϵ_s. In this regime, the hidden state evolution is effectively bottlenecked by O: even if it is completely deterministic, the observed sequences have randomness simply due to that in O.

2. Complexity

We can also investigate the complexity of observable sequences. The full I_{pred} is numerically challenging to obtain, but following the above discussion we can define an observable that measures the approach to the asymptotic entropy rate. As a simple measure, we define a complexity $\mathcal{C}_k = -H_s(k) + \frac{k}{k-1} H_s(k-1)$, plotted in Fig. 8. This combination of entropies eliminates the contribution to $H_s(k)$ from the asymptotic entropy rate and thus depends only on the rate at which this asymptotic rate is approached. Consistent with I_{pred} defined above, we find that $\mathcal{C}_k$ is nonmonotonic, with a peak approximately at $\epsilon \sim 0.1\epsilon_c$.

A consequence of these analyses is that for a HMM, many properties of the observable sequences mirror those of the underlying hidden sequences. This is useful, because the hidden sequence is Markovian and therefore easier to analyze *a priori*. In the usual scenario where only observable sequences are known, this is tempered by the fact that the hidden sequence must be reconstructed from the observations. Fortunately, efficient algorithms and toolboxes exist for this process [38,43].

C. Hidden Markov model applied to neural data

The explosion in availability of neural fMRI data requires tools that can reduce the dimensionality of the data and connect the data to functionally significant "modes" of neural activity [51]. Hidden Markov models are well suited to this task and have been extensively applied in neuroscience [16,42–45,52].

To test the extent to which the random HMM discussed above applies to these experimental data, we obtained the HMM transition matrix M and numerous hidden-state sequences from [16], which itself uses neural data from the Human Connectome project [53]. The whole-brain fMRI data are taken from the resting-state activity of 820 human subjects. $N = 12$ states were inferred at the group level, using multivariate Gaussian distributions applied to the raw fMRI data. After determining the states, the transition matrices M were inferred at the subject level. Pseudocounts were added to the state transition counts, following Laplace's rule of succession [54]. Approximately 80% of subjects used all 12 states, while 20% used 11 states, and a handful used 8–10 states. In our subsequent analysis, we defined and used an N value for each subject.

A typical state-by-state network from this data is shown in Fig. 9 (no threshold was used). Comparing with the networks in Fig. 3, a conspicuous feature here is that the self-transitions $A \to A$ are much more probable than transitions between states. This possibility was not taken into account in the random model, but its consequences will be deduced below.

Note that although these data were fitted to a Hidden Markov model, we will analyze only properties of the hidden
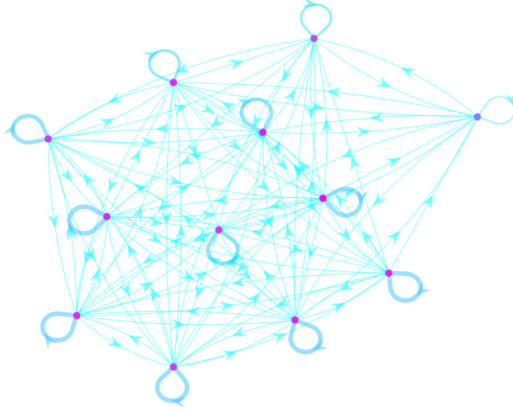
FIG. 9. Representative network for the Hidden Markov model applied to neural data.

states, so that we work in the context of the original Markov ensemble.

First, we naively analyze the HMMs in the context of the random model. For each subject, we measure the heterogeneity h of its M matrix and infer a temperature $\epsilon \equiv 1/(4[-1 + \sqrt{1 + h}])$, the entropy rate H_d, using (14), and I_{pred}, also derived from (14). The probability distributions of these quantities, rescaled as in earlier sections, are shown in Fig. 10. All of these quantities are sharply peaked around typical values, compared to the vast region explored, for example, in Fig. 5, which extends over five orders of magnitude in ϵ/ϵ_c. In particular, the most likely temperature is very near the critical one, a striking result given its status as a phase transition, and the peak in complexity near ϵ_c [Fig. 5(c)]. The mean values are $\epsilon/\epsilon_c \approx 0.97$, $H_d/\log N \approx 0.375$, and $I_{\text{pred}}/\log N \approx 0.84$. However, if we naively try to match these values with the random model, the comparison clearly fails: for example, $\epsilon \approx \epsilon_c$ would imply that $H_d/\log N \approx 0.6$, significantly higher than the observed value. How can we interpret this?

The key is the feature already noticed in Fig. 9, namely, that self-transitions are much more probable than transitions between states. Quantitatively, we find that the probability of a self-transition is $p \approx 0.75$, with fluctuations $\sim \pm 0.06$. For simplicity consider that all self-transitions have this same probability p. Then we can decompose the entropy rate as

follows:

$$
\begin{aligned}
H_d &= -\sum_{a,b} \pi_b M_{ab} \log M_{ab} \\
&= -\sum_b \pi_b p \log p - \sum_b \sum_{a \neq b} \pi_b M_{ab} \log M_{ab} \\
&= -p \log p - \sum_b \sum_{a \neq b} \pi_b (1-p) \tilde{M}_{ab} \log[(1-p)\tilde{M}_{ab}] \\
&= [-p \log p - (1-p)\log(1-p)] \\
&\quad - (1-p) \sum_b \sum_{a \neq b} \pi_b \tilde{M}_{ab} \log \tilde{M}_{ab} \\
&= H_{\text{Ber}}(p) + (1-p)\tilde{H}_d,
\end{aligned}
\tag{19}
$$

where $H_{\text{Ber}}(p)$ is the entropy of a Bernoulli process with probability p and $\tilde{M}_{ab} = M_{ab}/(1-p)$ is the conditional probability of the transition $b \to a$, given that this is not a self-transition. Thus the entropy rate is decomposed into the entropy for the decision whether or not to transition between states, and the remaining entropy of the conditional distribution described by $\tilde{M}_{ab}$. The crucial point is that the latter contribution is weighted by the small factor $1 - p \approx 0.25$. Thus if the conditional process is considered a random process with $\tilde{H}_d/\log N \approx 0.6$, we expect $H_d/\log N \approx [H_{\text{Ber}}(0.75)/\log 12] + 0.25 \times 0.6 = 0.375$, which is exactly what we find.

We can apply a similar analysis to I_{pred}. If $\epsilon = \epsilon_c$, then from Fig. 5(c) we estimate $\tilde{I}_{\text{pred}}/\log N \approx 0.4$ for the random model. Using (17) this implies $H_\pi/\log N = \log(2) \times 0.4 + 0.6 = 0.88$, which, again using (17), then implies $I_{\text{pred}}/\log N = (0.88 - 0.26)/\log(2) = 0.89$. This is remarkably close to the measured mean value 0.84.

Altogether, if we know the value of p, and assume $\epsilon = \epsilon_c$, then we can quantitatively predict the values of the entropy rate and the predictive information. This was unexpected, because our random model contains only two parameters, namely, the number of hidden states N and the temperature ϵ. Naively, one might imagine that many parameters would be needed to capture the properties of the HMM. This is especially so in the present context since it was shown in [16] that the HMM contains information on functional modes of neural activity (such as motor network versus language network *vs*
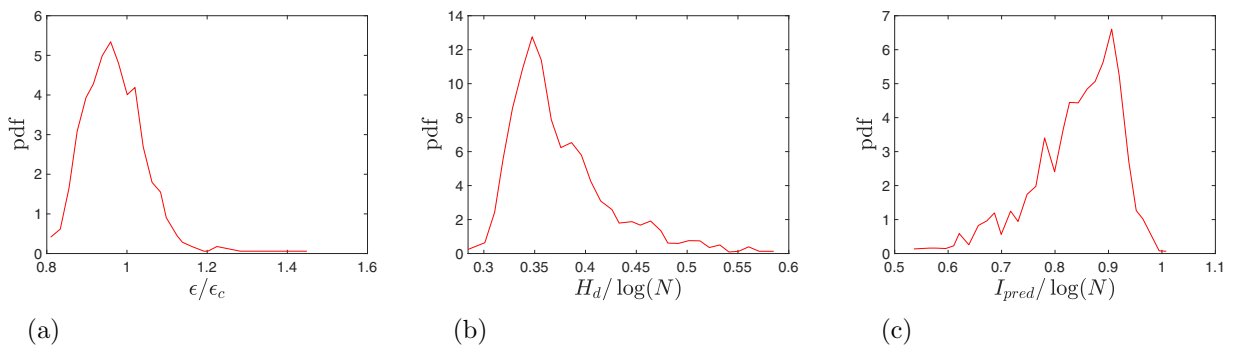

FIG. 10. Hidden Markov model applied to neural data. Probability distributions of (a) temperature, (b) Shannon entropy rate of hidden sequences, and (c) predictive information of hidden sequences.
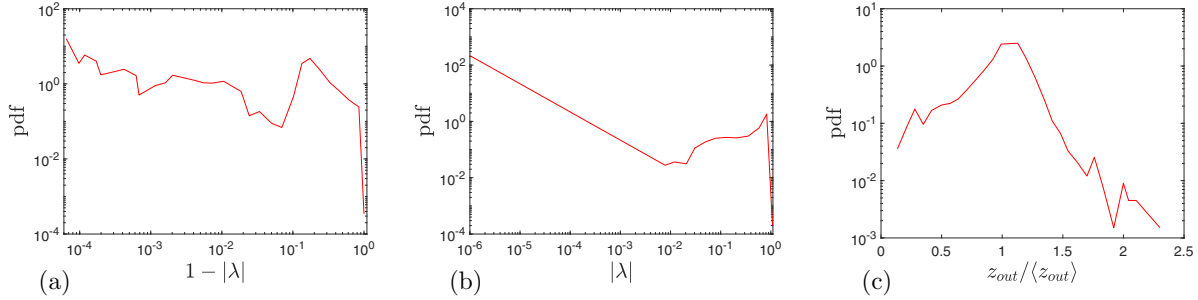
FIG. 11. Hidden Markov model applied to neural data. (a, b) Eigenvalue spectra, (c) out-degree distribution.

visual network), which are known to exhibit considerable spatial and temporal complexity.

To confirm that these results are not a coincidence, we can check that other properties of the random model hold for the neural data. To this end we measure the spectra of the transition matrices and the degree distribution of the networks. As earlier, we plot the eigenvalue density versus $1 - |\lambda|$ and versus $|\lambda|$, and the distribution of $z_{\text{out}}/\langle z_{\text{out}} \rangle$, in Figs. 11(a), 11(b), and 11(c), respectively. Although noisy, these distributions have key features of Fig. 1(c), Fig. 1(b), and Fig. 4, respectively. In particular: $\mathbb{P}(|\lambda|)$ has a tail of small values, corresponding to short relaxation times, as well as a tail of values close to 1, corresponding to long relaxation times, and $\mathbb{P}(z_{\text{out}})$ is peaked around unity with the beginnings of an exponential tail at large values.

Let us note that the parameter p is related to the sampling rate of the data: if the decision whether or not to change state is a Bernoulli process, then the number of time steps Δn before a true transition is a geometric random variable, with mean $1/(1 - p)$. These parameters are thus equivalent. If the sampling rate is f (in Hz), then the typical time (in s) between transitions is $\Delta t = \langle \Delta n \rangle / f = 1/[f(1 - p)]$.

Finally, let us comment on the surprising finding $\epsilon \approx \epsilon_c$, that is, subjects lie at the phase transition at which the large-N random matrix theory result breaks down. Since long relaxation times exist in the entire low-temperature phase, there is no simple *a priori* reason to expect neural activity to be right at this transition. However, since complexity peaks near the transition [Fig. 5(c)], this could be explained if brains are driven towards a state of maximal complexity.

These results lend support, at the whole-brain level, to the hypothesis that the brain operates in a critical regime, adding mathematical evidence to previous studies on small populations of individual neurons and EEG [17–23].

IV. CONCLUSION

The analysis of complex systems is often limited by the lack of a nontrivial null model to which one can compare data. One approach to this problem is to consider the ensemble of all models of a given class, for example, discrete Markov models of a given size. The ensemble is generally vast, including all possible models to which the data could be fitted, and cannot be completely characterized if the system under consideration has a high complexity. However, if one can identify a few key parameters that control variation of interesting observables in the ensemble, then one can consider

this *random model* as a null model. One can then look for phase transitions in the ensemble, with the aim of mapping out the general behaviors available within the class of model. This program has been advocated as a general strategy for the study of complex systems [55] and has been attempted for constraint satisfaction problems [56], for language syntax [26,57], and for ecosystems [58], to give a few examples.

In this work we have proposed a simple ensemble of discrete Markov models and a straightforward extension to Hidden Markov models. As a control parameter we introduced ϵ, which we call a temperature. The physical interpretation of ϵ is that it controls the dynamic range of transition matrix elements: large ϵ corresponds to a small dynamic range. In other words, when ϵ is large the various transitions have similar probabilities, while when ϵ is small, these probabilities can be vastly different. We have shown that the ϵ plays a key role in the behavior of the Markov models: there is a critical value ϵ_c such that large-N random matrix theory results hold for $\epsilon > \epsilon_c$, but not otherwise. In the low-temperature phase, the Markov models generically have a broad spectrum of relaxation times, without any fine-tuning. The critical point ϵ_c thus acts as a phase transition between a simple high-temperature phase and a more ordered low-temperature phase. In our ensemble the value $\epsilon_c = 1/[2 \log(N + 1)]$ depends only on the number of states in the Markov model. We have furthermore shown that ϵ controls the entropy of state-transition sequences.

The utility of this ensemble was explicitly shown using human fMRI data, where we could quantitatively characterize the values of entropy and predictive information found in a HMM fit to fMRI data. The latter measures how informative observations of the "past" are with respect to predicting the future and is an essentially unique measure of complexity [29].

Owing to the generality of Markov models, this work fits into a number of larger problems. First, Hidden Markov models are the simplest class of probabilistic grammars, which are organized by the Chomsky hierarchy [26,46,47]. The next level of complexity in grammars are the context-free grammars, considered as an ensemble in [26,57,59]. These are related to syntactic structure in both natural and computer languages. A similar phase transition to that discussed here was found in [26,57], but without any interpretation in terms of random matrix theory. As discussed therein, the phase transition can be interpreted in terms of breaking of permutation symmetry, an interpretation that also holds in the present case. Briefly, at high temperature, all the symbols are statistically

equivalent: the permutation symmetry among them is preserved. As temperature is lowered through ϵ_c, the transition rates span a broad enough dynamic range that the symbols can then be distinguished. This is precisely what is measured by the Zipf plot, Fig. 6. This connects to the standard Landau paradigm of phase transitions in terms of symmetry breaking.

Second, the low-temperature regime of the Markov ensemble has matrices that are effectively sparse; thus the phase transition discussed here can be interpreted as a full-to-sparse transition of random matrices. The behavior of sparse random matrices, while important, is understood much more poorly than that of their dense cousins. Recent work has built theoretical tools for such study [8,9,60] and could be useful to theoretically derive the behavior of the eigenvalue tails we found near $|\lambda| = 0$ and $|\lambda| = 1$. Eigenvalue tails have previously been predicted in models of complex networks [61] but in the symmetric case. We found that the eigenvalue spectra in the low-temperature regime show a bicycle wheel structure in the complex plane, with spokes at the roots of unity. We are not aware of any previous mention of this phenomenon.

Third, we have considered Markov models in discrete time, but a natural generalization would be to continuous time. This would also allow a connection to stochastic thermodynamics [62].

Finally, the presence of a phase transition allows one to place stochastic matrices in the phase diagram and evaluate how near or far they are from the transition. Applied to fMRI data, we found that the data are very near the transition, supporting the brain criticality hypothesis. A natural question is whether stochastic matrices from other complex systems are also near this transition. Since the ensemble includes all discrete Markov models, it can unify the study of disparate systems, in the goals of seeking universal patterns. This may shed light on the origin and possible universality of criticality in biological systems.

APPENDIX: FROM CONTINUOUS-TIME DYNAMICS TO THE MASTER EQUATION

Here we discuss how the transition matrix M is related to the interaction matrix $\tilde{M}$. It is simplest to discretize time, x, and ξ. A continuum limit can be taken later, if necessary.

For example, if we discretize time, x, and ξ, then (1) becomes

$$x_i(t + dt) = x_i(t)(1 - dt) + dt\tilde{M}_{ij}x_j(t) + d\xi_i(t). \quad (A1)$$

This induces a motion in phase space with a transition matrix between states $x = (x_1(t), x_2(t), \ldots, x_n(t))$ and $y = (x_1(t + 1), x_2(t + 1), \ldots, x_n(t + 1))$ as

$$M_{yx} = \sum_{\{d\xi\}} \mathbb{P}(d\xi) \prod_i \delta_{y_i - x_i(1-dt) - dt\tilde{M}_{ij}x_j - d\xi_i}, \quad (A2)$$

where $\mathbb{P}(d\xi)$ is the probability of noise increments $d\xi = (d\xi_1, \ldots, d\xi_n)$, and the Kronecker δ's impose the dynamics, which trivially generalizes to any nonlinear extension of (1). Notice that $\sum_y M_{yx} = 1$ and each $M_{yx} \geqslant 0$. The phase space dynamics is given by the master equation

$$\rho(y, t + dt) = \sum_x M_{yx}\rho(x, t), \quad (A3)$$

where $\rho(x, t)$ is the probability of state x at time t. If each x_i has n_x different states, then M is an $N \times N$ matrix, with $N = (n_x)^n$.

Let us now generalize to the continuum limit in x and ξ. M becomes a function of two arguments

$$M(y, x) = \int dP(\xi) \prod_i \delta[y_i - x_i(1 - dt) - dt\tilde{M}_{ij}x_j - d\xi_i]. \quad (A4)$$

For simplicity we assume ξ is Gaussian with $\langle d\xi_i \rangle = 0$ and correlation matrix $C_{ij} dt = \langle d\xi_i d\xi_j \rangle$. Then

$$
\begin{aligned}
\rho(y, t + dt) &= \int dx M_{yx}\rho(x, t) \\
&= \int dx\rho(x, t) \int dP(\xi) \prod_i \delta[y_i - x_i(1 - dt) - dt\tilde{M}_{ij}x_j - \boldsymbol{d\xi}_i] \\
&= |1 - dt + dt\tilde{M}|^{-1} \int dP(\xi)\rho([1 - dt + dt\tilde{M}]^{-1}(y - \boldsymbol{d\xi}), t) \\
&= e^{-dt(-n + \text{tr}\tilde{M}) + \mathcal{O}(dt^2)} \int dP(\xi)\rho(y - d\xi + dty - dt\tilde{M}y + \mathcal{O}(dt^2), t) \\
&= [1 + dt(n - \text{tr}\tilde{M}) + \mathcal{O}(dt^2)] \int dP(\xi)[\rho(y, t) + (dty - dt\tilde{M}y - \boldsymbol{d\xi})\nabla\rho(y, t) + \tfrac{1}{2}\boldsymbol{d\xi}\boldsymbol{d\xi}:\nabla\nabla\rho(y, t) + \mathcal{O}(dt^2)] \\
&= \rho(y, t) + dt(n - \text{tr}\tilde{M})\rho(y, t) + dt(y - \tilde{M}y)\nabla\rho(y, t) + \tfrac{1}{2}dtC:\nabla\nabla\rho(y, t) + \mathcal{O}(dt^2) \\
&= \rho(y, t) + dt\boldsymbol{\nabla} \cdot [(y - \tilde{M} \cdot y)\rho(y, t)] + \tfrac{1}{2}dt\boldsymbol{\nabla}\boldsymbol{\nabla}:[C\rho(y, t)] + \mathcal{O}(dt^2)
\end{aligned}
$$

or

$$\partial_t \rho(y,t) = \nabla \cdot [(\boldsymbol{y} - \tilde{M} \cdot \boldsymbol{y})\rho(y,t)] + \tfrac{1}{2}\nabla \nabla :[C\rho(y,t)], \tag{A5}$$

which is the Fokker-Planck equation, as expected.

[1] M. A. Munoz, Rev. Mod. Phys. **90**, 031001 (2018).

[2] T. Mora and W. Bialek, J. Stat. Phys. **144**, 268 (2011).

[3] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987).

[4] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 2000).

[5] A. Levina, J. M. Herrmann, and T. Geisel, Nat. Phys. **3**, 857 (2007).

[6] R. Zeraati, V. Priesemann, and A. Levina, Front. Phys. **9**, 619661 (2021).

[7] M. Müller and M. Wyart, Annu. Rev. Condens. Matter Phys. **6**, 177 (2015).

[8] R. Kühn, Acta Phys. Pol. B **46**, 1653 (2015).

[9] R. Kühn, Europhys. Lett. **109**, 60003 (2015).

[10] S. Bartolucci, F. Caccioli, F. Caravelli, and P. Vivo, arXiv:2106.02730.

[11] M. Horvat, J. Stat. Mech.: Theory Exp. (2009) P07005.

[12] L. H. Oliveira and M. Novaes, J. Math. Phys. **60**, 123508 (2019).

[13] V. L. Girko, Theory Probab. Appl. **29**, 694 (1985).

[14] T. Tao and V. Vu, Commun. Contemp. Math. **10**, 261 (2008).

[15] C. Bordenave, P. Caputo, and D. Chafaï, Probab. Theory Relat. Fields **152**, 751 (2012).

[16] D. Vidaurre, S. M. Smith, and M. W. Woolrich, Proc. Natl. Acad. Sci. USA **114**, 12827 (2017).

[17] L. de Arcangelis, C. Perrone-Capano, and H. J. Herrmann, Phys. Rev. Lett. **96**, 028107 (2006).

[18] J. M. Beggs, Philos. Trans. R. Soc. A **366**, 329 (2008).

[19] G. Deco and V. K. Jirsa, J. Neurosci. **32**, 3366 (2012).

[20] J. Hesse and T. Gross, Front. Syst. Neurosci. **8**, 166 (2014).

[21] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry, and W. Bialek, Proc. Natl. Acad. Sci. USA **112**, 11508 (2015).

[22] A. J. Fontenele, N. A. P. deVasconcelos, T. Feliciano, L. A. A. Aguiar, C. Soares-Cunha, B. Coimbra, L. DallaPorta, S. Ribeiro, A. J. Rodrigues, N. Sousa, P. V. Carelli, and M. Copelli, Phys. Rev. Lett. **122**, 208101 (2019).

[23] L. J. Fosque, R. V. Williams-García, J. M. Beggs, and G. Ortiz, Phys. Rev. Lett. **126**, 098101 (2021).

[24] X. Chen and W. Bialek, arXiv:2008.11674.

[25] D. Sornette and R. Cont, J. Phys. I **7**, 431 (1997).

[26] E. DeGiuli, Phys. Rev. Lett. **122**, 128301 (2019).

[27] M. Newman, *Networks* (Oxford University Press, Oxford, 2018).

[28] W. Bialek, I. Nemenman, and N. Tishby, Phys. A **302**, 89 (2001).

[29] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001).

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).

[31] G. Parisi, *Statistical Field Theory* (Addison-Wesley, New York, 1988).

[32] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986).

[33] R. F. i Cancho and R. V. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2003).

[34] B. Corominas-Murtra and R. V. Solé, Phys. Rev. E **82**, 011102 (2010).

[35] A. Corral, G. Boleda, and R. Ferrer-i-Cancho, PloS ONE **10**, e0129031 (2015).

[36] D. J. Schwab, I. Nemenman, and P. Mehta, Phys. Rev. Lett. **113**, 068102 (2014).

[37] G. K. Zipf, *The Psycho-biology of Language: An Introduction to Dynamic Philology* (Routledge, Milton Park, UK, 2013).

[38] L. Rabiner and B. Juang, IEEE ASSP Mag. **3**, 4 (1986).

[39] S. R. Eddy, Nat. Biotechnol. **22**, 1315 (2004).

[40] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998).

[41] B. H. Juang and L. R. Rabiner, Technometrics **33**, 251 (1991).

[42] J. Ou, L. Xie, C. Jin, X. Li, D. Zhu, R. Jiang, Y. Chen, J. Zhang, L. Li, and T. Liu, Brain Topogr. **28**, 666 (2015).

[43] D. Vidaurre, R. Abeysuriya, R. Becker, A. J. Quinn, F. Alfaro-Almagro, S. M. Smith, and M. W. Woolrich, Neuroimage **180**, 646 (2018).

[44] A. Stevner, D. Vidaurre, J. Cabral, K. Rapuano, S. F. V. Nielsen, E. Tagliazucchi, H. Laufs, P. Vuust, G. Deco, and M. W. Woolrich, Nat. Commun. **10**, 1035 (2019).

[45] K. Goucher-Lambert and C. McComb, in *Proceedings of the Design Society: International Conference on Engineering Design*, vol. 1 (Cambridge University Press, Cambridge, 2019), pp. 1873–1882.

[46] N. Chomsky, *Syntactic Structures* (Walter de Gruyter, Berlin, 2002).

[47] A. Carnie, *Syntax: A Generative Introduction* (John Wiley and Sons, New York, 2013).

[48] A. M. Jurgens and J. P. Crutchfield, J. Stat. Phys. **183**, 32 (2021).

[49] P. Grassberger, arXiv:physics/0307138.

[50] T. Schürmann and P. Grassberger, Chaos **6**, 414 (1996).

[51] D. J. Lurie, D. Kessler, D. S. Bassett, R. F. Betzel, M. Breakspear, S. Kheilholz, A. Kucyi, R. Liégeois, M. A. Lindquist, and A. R. McIntosh, Netw. Neurosci. **4**, 30 (2020).

[52] D. Vidaurre, PLoS Comput. Biol. **17**, e1008580 (2021).

[53] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, and M. P. Harms, Neuroimage **80**, 144 (2013).

[54] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson, Phys. Rev. E **90**, 012132 (2014).

[55] G. Parisi, Physica A **263**, 557 (1999).

[56] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).

[57] E. De Giuli, J. Phys. A: Math. Theor. **52**, 504001 (2019).

[58] G. Biroli, G. Bunin, and C. Cammarota, New J. Phys. **20**, 083051 (2018).

[59] K. Nakaishi and K. Hukushima, arXiv:2105.00832.

[60] V. A. Susca, P. Vivo, and R. Kühn, arXiv:2101.08029.

[61] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **68**, 046109 (2003).

[62] C. Van den Broeck, Phys. Complex Colloids **184**, 155 (2013).