Zipf's law and criticality in multivariate data without fine-tuning

David J. Schwab*
Department of Physics and Lewis-Sigler Institute, Princeton University, Princeton, NJ 08854

Ilya Nemenman†
Departments of Physics and Biology, Emory University, Atlanta, GA 30322

Pankaj Mehta‡
Department of Physics, Boston University, Boston, MA 02215

Recently it has become possible to directly measure simultaneous collective states of many biological components, such as neural activities or antibody sequences. A striking result has been the observation that the underlying probability distributions of the collective states of these systems exhibit a feature known as Zipf's law. They appear poised near a unique critical point, where the extensive parts of the entropy and energy exactly cancel. Here we present analytical arguments and numerical simulations showing that such behavior naturally arises in systems with an unobserved random variable (e. g., input stimulus to a neural system) that affects the observed data. The mechanism does not require fine tuning and may help explain the ubiquity of Zipf's law in disparate systems.

Great advances in high throughput experimental biology now allow the joint measurement of activities of many basic components underlying collective behaviors in biological systems. These include firing patterns of many neurons responding to a movie [1–3], sequences of proteins from individual immune cells in zebrafish [4, 5], protein sequences more generally [6, 7], and even the simultaneous motion of flocking birds [8]. A remarkable result of these data and their models using, for example, Maximum Entropy (MaxEnt) techniques [9], has been the observation that these large biological systems often reside close to a critical point [10]. This is most clearly manifest directly from the data by the following striking behavior. If we order the states $\boldsymbol{\sigma}$ of a system by decreasing probability, then the frequency of the states decays as the inverse of their rank, $r(\boldsymbol{\sigma})$, to some power:

$$P(\boldsymbol{\sigma}) \propto \frac{1}{r(\boldsymbol{\sigma})^{\alpha}}. \tag{1}$$

The states of the system thus follow a power-law behavior and, consequently, there is no natural scale associated with the underlying probability distribution. It has been argued that such behavior is a model-free signature of criticality in the underlying system [10]. Many systems in fact exhibit $\alpha \simeq 1$, which is termed Zipf's law, and on which we will focus for the remainder of this paper.

The connection between such power-law behavior and criticality can be made rigorous using the language of statistical mechanics [10]. Without loss of generality, we can define the "energy" of a state $\boldsymbol{\sigma}$ to be

$$E(\boldsymbol{\sigma}) = -\log P(\boldsymbol{\sigma}) + \text{const.} \tag{2}$$

The additive constant is arbitrary, and the temperature is $k_B T = 1$. We can also define the "entropy", $S(E)$, using the density of states, $\rho(E) = \sum_{\boldsymbol{\sigma}} \delta(E - E(\boldsymbol{\sigma}))$, as

$$S(E) = \log \rho(E). \tag{3}$$

Both the energy E and the entropy $S(E)$ contain extensive terms that scale with the system size, N. An elegant argument [10] converts Eq. (1) with $\alpha = 1$ into the statement that, for a large system, $N \to \infty$, the energy and entropy are exactly equal to leading order in N. Thus in the thermodynamic limit, the probability distribution is indeed poised near an infinite order critical point where all derivatives beyond the first of the entropy with respect to energy vanish to leading order in N.

The observation of Zipf's law in myriad distributions inferred from biological data has contributed to a revival of the idea that biological systems may be poised near a phase transition [10–15]. And yet most existing mechanisms to generate Zipf's law can produce a variety of power-law exponents α (see [16, 17] and reference therein), have semi-stringent conditions [18], or require fine-tuning to a critical point, highlighting the crucial need to understand how Zipf's law can arise in data-driven models.

Here we present a mechanism that gives rise to Zipf's law and does not require fine-tuning. The observation motivating this new mechanism is that the correlations measured in biological data sets have multiple origins. Some of these are intrinsic to the system, while the others reflect extrinsic, unobserved sources of variation. For example, the distributions of activities recorded from networks of neurons in the retina reflect both the intrinsic structure of the network as well as the stimuli the neurons receive, e. g., movies of natural scenes. Similarly, firing patterns of a single neuron are controlled by refractory constraints, but also by time-varying external stimuli. Likewise, in the immune system, the pathogen environment is an external source of variation that influences the observed antibody combinations. We will show that the presence of such unobserved, hidden random variables naturally leads to Zipf's law. Unlike other

mechanisms [16, 18], our approach requires a large parameter (i. e., the system size, or the number of observations, N), with power-law behavior emerging only in the thermodynamic limit. On the other hand, our mechanism does not require fine-tuning or any special statistics of the hidden variables [19]. In other words, Zipf's law is a fairly general feature that emerges when marginalizing over relevant hidden variables.

A simple model — In order to understand how a hidden variable can give rise to Zipf's law and concomitant criticality, we start by examining a simple case of N conditionally independent binary spins $\sigma_i = \pm 1$. The spins are influenced by a hidden variable β drawn from a probability distribution $q(\beta)$, which is smooth and independent of N. In particular, we consider the case

$$P(\boldsymbol{\sigma}|\beta) = \prod_{i=1}^{N} P(\sigma_i|\beta) = \prod_{i=1}^{N} \frac{e^{\beta\sigma_i}}{2\cosh\beta}. \tag{4}$$

We consider a scenario where the parameter β changes rapidly compared to the duration of the experiment, so that the probability of the measured data, $\boldsymbol{\sigma}$, is averaged over β:

$$P(\boldsymbol{\sigma}) = P(m) = \frac{1}{2^N}\int d\beta\, q(\beta) e^{N(\beta m - \log\cosh\beta)} \tag{5}$$

$$\equiv \frac{1}{2^N}\int d\beta\, q(\beta) e^{-N\mathcal{H}(m,\beta)}, \tag{6}$$

where we have defined the average magnetization $m = \sum_i \sigma_i/N$, and the last equation defines $\mathcal{H}(m,\beta)$. Note that the distribution $P(\boldsymbol{\sigma})$ does not factorize unlike $P(\boldsymbol{\sigma}|\beta)$. That is, the conditionally independent spins are not marginally independent. Indeed, as in [20], a sequence of spins carries information about the underlying β and hence about other spins (e. g., a prevalence of positive spins suggests $\beta > 0$, and thus subsequent spins will also likely be positive). We note that the simple model in Eq. (6) is intimately related to the MaxEnt model constructed in [3] to match the distribution of the number of simultaneously firing retinal ganglion cells.

In the limit $N \gg 1$, we can approximate the integral in Eq. (6) by Laplace's method (saddle-point approximation) to get

$$P(\boldsymbol{\sigma}) \approx 2^{-N} q(\beta^*) e^{N(\beta^* m - \log\cosh\beta^*)},\ \tanh\beta^* = m. \tag{7}$$

Here β^* is the maximum-likelihood estimate of β given the data, $\boldsymbol{\sigma}$. In deriving Eq. (7) we assumed that the distribution $q(\beta)$ has support at β^* and is sufficiently smooth, e.g. does not depend on N, so that the saddle-point over β is determined by $\mathcal{H}$, and not by the prior. In other words, we require the Fisher information $\mathcal{F}(\beta^*) \equiv -N\left.\frac{\partial^2\mathcal{H}}{\partial\beta^2}\right|_{\beta^*} = N(1-m^2) \gg 1$, and for the location and curvature of the saddle point to not be significantly modulated by $q(\beta)$. These conditions are violated at $m = \pm 1$, and there is a semi-infinite range of β that could have contributed to such states. For all nonzero values of $\mathcal{F}$, the saddle-point will eventually dominate over $q(\beta)$ as $N \to \infty$. However, the convergence is not uniform.

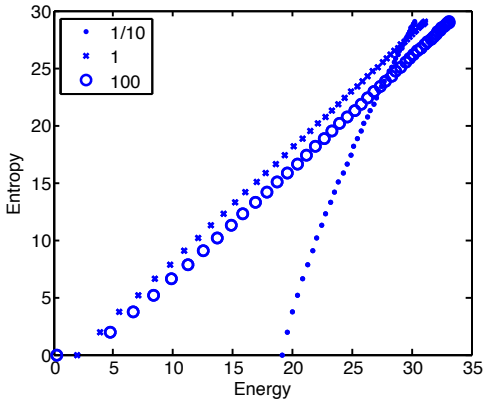

FIG. 1: Entropy, $S(m)$, vs energy, $E(m)$, for $N = 100$ identical and conditionally independent spins. Zipf's law ($E = S$) emerges as the standard deviation, $s \in \{0.1, 1, 100\}$, of the Gaussian distribution characterizing the hidden variable β is increased. Notice that there is a nearly perfect Zipf's law for 2 orders of magnitude in s. The mean of $q(\beta)$ is set to zero, and thus there is a two-fold degeneracy between states with magnetization m and $-m$.

Substituting Eq. (7) into Eq. (2) and using the identities $\tanh^{-1} m = \frac{1}{2}\log\left(\frac{1+m}{1-m}\right)$ and $\cosh\left[\tanh^{-1} m\right] = (1-m^2)^{-1/2}$, we obtain the energy to leading order in N:

$$E(m) \approx -N\left[\left(\tfrac{1+m}{2}\right)\log\left(\tfrac{1+m}{2}\right) + \left(\tfrac{1-m}{2}\right)\log\left(\tfrac{1-m}{2}\right)\right]$$
$$\equiv NH(m). \tag{8}$$

Here we neglected subdominant terms that come from both the prior $q(\beta^*)$ and the fluctuations about the saddle point. It is worth noting that this energy considered as a function of the σ_i, rather than m, includes interactions of all orders, not just pairwise spin couplings.

We can also calculate the entropy $S(m)$ associated with the magnetization m. For a system of N binary spins, each state with magnetization m has $K = N\left(\frac{1+m}{2}\right)$ up spins, and there are $\binom{N}{K}$ such states. Using Stirling's approximation, one finds that the entropy takes the familiar form $S(m) = \log\binom{N}{K} \approx NH(m)$. Of course, this is the same as the energy, Eq. (8), for the system with a hidden variable β, to leading order in N.

The analytic equivalence between energy and entropy only applies when $N \to \infty$. To verify our result for a finite N, we numerically calculate $E(m)$ from Eq. (6) with $q(\beta)$ chosen from a variety of distribution families (e. g., Gaussian, exponential, uniform). For brevity, we only show plots for Gaussian distributions, but the others gave similar results. Figure 1 plots the entropy, $S(m) = \log\binom{N}{K}$, vs the energy, $E(m)$, for $N = 100$ con-

ditionally independent spins, where $q(\beta)$ has mean 0 and varying standard deviation $s \in \{0.1, 1, 100\}$. For small s, the hidden variable β is always close to zero, there is no averaging, and all states are nearly equally (im)-probable. As s increases, entropy becomes equal to energy over many decades of energies modulo an arbitrary additive constant. This holds true for two orders of magnitude of the standard deviation s, confirming that our mechanism does not require fine tuning.

The stable emergence in the thermodynamic limit, $N \to \infty$, with no fine-tuning, distinguishes our setup from a classic mechanism explaining $1/f$ noise in solids [18] and certain other biological systems [21]. We could have anticipated this result: if the extensive parts of the energy and entropy do not cancel, in thermodynamic limit, the magnetization will be sharply peaked around the m that minimizes the free-energy. Thus in order for there to be a broad distribution of magnetizations within $P(\boldsymbol{\sigma})$ the extensive part of the free-energy must be identically zero. In other words, the observation of a broad distribution of an order parameter-like quantity in data is indicative of a Zipfian distribution. One straightforward mechanism to produce a broad order parameter distribution for large N is to couple it to a hidden fluctuating variable.

A generic model — We can generalize the calculation above to show that Zipf-like criticality is a generic property of distributions with hidden variables, and is not a consequence of the specific model in Eq. (4). In particular, it does not require the observed variables to be identical or even conditionally independent. To do so, we make use of the Gärtner-Ellis theorem from Large Deviations theory [22]. Consider a real random variable A_N, parameterized by an integer N, and define the cumulant generating function by $\lambda(k) = \lim_{N\to\infty} \frac{1}{N} \log \langle e^{NkA_N} \rangle$. The theorem states that if $\lambda(k)$ exists and is differentiable for all k, then the probability, $P(A_N \in da)$, that A_N takes values in the interval $[a - da, a + da]$ is approximated by $P(A_n \in da) \approx e^{-NI(a)} da$, as $N \to \infty$, with the rate function $I(a) = \sup_k \{ka - \lambda(k)\}$.

To apply the theorem, we assume that, as above, the probability distribution for $\boldsymbol{\sigma}$ depends on an unobserved hidden variable β drawn from a distribution $q(\beta)$ so that

$$P(\boldsymbol{\sigma}) = \int d\beta q(\beta) e^{-N\beta\epsilon_N(\boldsymbol{\sigma}) - \log Z(\beta)}, \qquad (9)$$

where $Z(\beta) = \int d\boldsymbol{\sigma} e^{-N\beta\epsilon_N(\boldsymbol{\sigma})} = \langle e^{-N\beta\epsilon_N(\boldsymbol{\sigma})} \rangle$, and interactions among the spins are encoded in $\epsilon_N(\boldsymbol{\sigma})$.

We can think of $\epsilon_N(\boldsymbol{\sigma})$ as a random variable through its dependence on $\boldsymbol{\sigma}$. However, henceforth we suppress the explicit dependence of $\epsilon_N(\boldsymbol{\sigma})$ on $\boldsymbol{\sigma}$. We define the intensive free-energy $g(\beta) = -\lim_{N\to\infty} \frac{1}{N} \log Z(\beta)$, which is also minus the cumulant generating function for the random variable $-\epsilon_N$. Once again we evaluate Eq. (9) using the saddle-point approximation, with the same caveats as discussed after Eq. (7). To leading order in N, one has

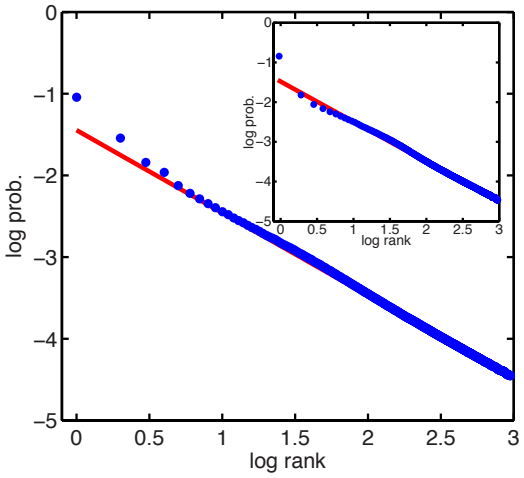

FIG. 2: Main plot: Plot of $\log_{10}$ probability vs. $\log_{10}$ rank of the most frequent 10^3 states for a system of $N = 200$ non-identical, conditionally-independent spins (model (a)). Plots are an average over 200 realizations of the quenched variables h_i that break the symmetry between spins, with 5×10^5 samples taken for each realization. Parameters: $h_i \sim \mathcal{N}(\mu = 1, s = 0.3)$, $\beta \sim \mathcal{N}(\mu = 0, s = 2)$. Red line: least-squares fit to patterns $100-1000$, slope of -1.012. Inset: Same as above, except for a model of $N = 200$ spins with quenched random interactions J_{ij} and biases h_i (model (b)). Average over 10 realizations of J_{ij} and h_i chosen from $J_{ij} \sim \mathcal{N}(\mu = 1, s = 0.5)$, $h_i \sim \mathcal{N}(\mu = 1, s = 0.85)$, $\beta \sim \mathcal{N}(\mu = 0.5, s = 0.5)$, with 3×10^5 samples taken for each realization. Red line: least-squares fit to patterns $100 - 1000$, slope of -1.011.

$P(\boldsymbol{\sigma}) \approx q(\beta^*) e^{-NH(\epsilon_N)}$ with $H(\epsilon_N) = \inf_\beta \{\beta\epsilon_N - g(\beta)\}$, and β^* is the arg min of this equation. On the one hand, we see from Eq. (2) that, to the leading order in N, the energy is equal to $E(\boldsymbol{\sigma}) = NH(\epsilon_N)$. On the other hand, from the Gärtner-Ellis theorem, we know that the density of states at a given energy is just $\rho(\epsilon_N) \approx e^{-NI(-\epsilon_N)} = e^{NH(\epsilon_N)}$, so that the entropy from Eq. (3) is $S(\epsilon_N) = \log \rho(\epsilon_N) = NH(\epsilon_N)$. Thus, quite generally, in the presence of a hidden variable, and with $N \to \infty$, the extensive parts of the energy and entropy are equal, and Zipf's law emerges, apart from the configurations of $\boldsymbol{\sigma}$ where $\mathcal{F} = 0$ or $\log q(\beta^*)$ or its derivatives are singular. It should also be noted that if the Hamiltonian is such that there exists a first-order phase transition as a function of β with other parameters held fixed, the Gärtner-Ellis theorem does not hold, and Zipf's law need not be observed. Finally, this mechanism can be generalized to K hidden variables, in which case σ_i must be conditionally independent Potts spins (i. e. categorical) or continuous variables. This setup can be interpreted as inferring the values of K parameters from N observations [20], and these relations to learning theory will be explored in the future.

We numerically test the validity of our analytic result for finite N in two systems more complex than Eq. (4): (a) a collection of non-identical but conditionally inde-

pendent spins, and (b) an Ising model with random interactions and fields. The main graph of Fig. 2 shows a Zipf plot for system (a), so that

$$P(\boldsymbol{\sigma}|\beta) = \prod_{i=1}^{N} P(\sigma_i|\beta) = \prod_{i=1}^{N} \frac{e^{-\beta h_i \sigma_i}}{2\cosh \beta h_i}, \qquad (10)$$

where h_i are quenched, Gaussian distributed random variables unique for each spin. In the simulations, the hidden variable β was drawn from a Gaussian distribution, but similar results were found for other distributions. The quenched fields h_i break the symmetry between spins. In agreement with our derivation, on a log-log plot, the states generated from simulations fall on a line with slope very close to -1 (Fig. 2), the signature of Zipf's law. The first few states deviate slightly from our bulk prediction, but the overall agreement is remarkable.

To verify that conditional independence is not required for this mechanism, we studied system (b) that generalizes the model in Eq. (10) to include random exchange interactions between spins:

$$P(\boldsymbol{\sigma}|\beta) \propto e^{-\beta\left(\frac{1}{N}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right)}, \qquad (11)$$

where the J_{ij} and h_i are quenched Gaussian distributed interactions and fields, and β is as above. As shown in Fig. 2 (inset), the data again fall on a line with slope nearly equal to -1. In Eq. (11), β multiplies all terms in the exponent. If interactions are present that don't couple to β, the energy and entropy will differ by precisely those terms. However, if such interactions are weak, as is often observed in inferred MaxEnt distributions [2, 23], Zipf's law may still approximately hold.

To see our mechanism at work in data, consider a neural spike train from a single blowfly motion-sensitive neuron H1 stimulated by a time-varying motion stimulus, $\beta(t)$ (see [24, 25] for experimental details). We can discretize time with a resolution of τ and interpret the spike train as an ordered sequence of N spins, such that $\sigma_i = \pm 1$ corresponds to the absence/presence of a spike in a time window $t \in [\tau(i-1), \tau i)$. The probability of a spike in a time window depends on β. However, neural refractoriness prevents two spikes from being close to each other, irrespective of the stimulus, resulting in a repulsion that does not couple to β. The rank-ordered plot of spike patterns produced by the neuron is remarkably close to the Zipf behavior (Fig. 3). We also simulated a refractory Poisson spike train using the same values of β. We match the mean firing rate and the refractory period to the data, so that $P(\sigma_i|\beta)$ is given by Eq. (4), with a hard repulsive constraint between positive spins extending over the refractory period. The rank-ordered plot for this model that manifestly includes interactions uncoupled from the hidden stimulus β still exhibits Zipf's law (Fig. 3).

Discussion — It is possible that evolution has tuned biological systems or exploited natural mechanisms of self-

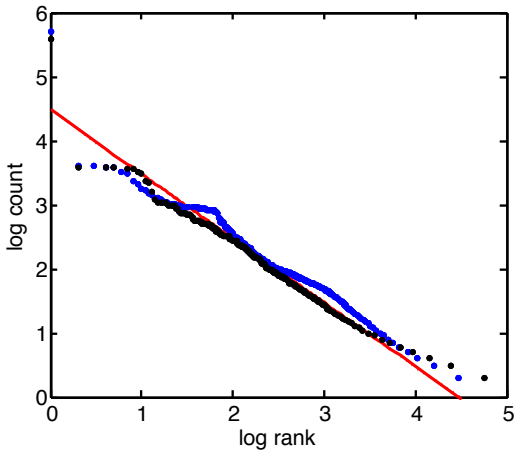

FIG. 3: Rank-count plot from a motion-sensitive blowfly neuron, logs base 10; discretization is $\tau = 1$ ms, and $N = 40$. Black: empirical rank-ordered counts. Blue: rank-ordered counts from a refractory Poisson spike train with the same input stimulus, same mean firing rate, and the same refractory period of 2 ms. Red: slope of -1 guide to the eye.

organization [11] to arrive at Zipf's law. Alternatively, informative data-driven models may lie close to a critical point due to the high density of distinguishable models there [26]. Our work suggests another possibility: Zipf's law can robustly emerge due to the effects of unobserved hidden variables. While our approach is biologically motivated, it is likely to be relevant to other systems where Zipf's law has been observed, and it will be interesting to unearth the dominant mechanisms in particular systems. For this, if a candidate extrinsic variable can be identified, such as the input stimulus to a network of neurons, its variance could be modulated experimentally as in Fig. 1. Our mechanism would expect Zipf's law to appear only for a broad distribution of the extrinsic variable, and for $N \gg 1$ observed variables.

While our mechanism does not require fine-tuning, it nonetheless suggests that biological systems operate in a special regime. Indeed, the system size N required to exhibit Zipf's law depends on the sensitivity of the observed $\boldsymbol{\sigma}$ to the variations of the hidden variable. If the system is poorly adapted to the distribution of β, e. g. the mean of $q(\beta)$ is very large or its width is too small to cause substantial variability in $\boldsymbol{\sigma}$ (as in Fig. 1, $s = 0.1$), a very large N will be required. In other words, a biological system must be sufficiently adapted to the statistics of β for Zipf's law to be observed at modest system sizes. Indeed, this type of adaptation is well established in both neural and molecular systems [27–30].

Acknowledgments We would like to thank Bill Bialek, Justin Kinney, H.G.E. Hentschel, Thierry Mora, and Martin Tchernookov for useful conversations. We thank Robert de Ruyter van Steveninck and Geoff Lewen for providing the data in Fig. 3. PM was partially sup-

ported by an Alfred Sloan Fellowship. IN was partially supported by the James S. McDonnell Foundation. DJS was supported by National Institute of Health Grant K25 GM098875-02.

* Electronic address: dschwab@princeton.edu
† Electronic address: ilya.nemenman@emory.edu
‡ Electronic address: pankajm@bu.edu

[1] S. Cocco, S. Leibler, and R. Monasson, Proc Natl Acad Sci (USA) **106**, 14058 (2009).
[2] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[3] G. Tkačik, O. Marre, T. Mora, D. Amodei, M. Berry II, and W. Bialek, J Stat Mech **2013**, P03011 (2013).
[4] T. Mora, A. Walczak, W. Bialek, and C. Callan, Proc Natl Acad Sci (USA) **107**, 5405 (2010).
[5] A. Murugan, T. Mora, A. Walczak, and C. Callan, Proc Natl Acad Sci (USA) **109**, 16161 (2012).
[6] M. Weigt, R. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc Natl Acad Sci (USA) **106**, 67 (2009).
[7] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[8] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak, Proc Natl Acad Sci (USA) **109**, 4786 (2012).
[9] E. Jaynes, Phys Rev **106**, 620 (1957).
[10] T. Mora and W. Bialek, J Stat Phys **144**, 268 (2011).
[11] P. Bak, C. Tang, and K. Wiesenfeld, Phys Rev A **38**, 364 (1988).
[12] J. Beggs and D. Plenz, J Neurosci **23**, 11167 (2003).
[13] J. Beggs, Phil Trans R Soc A **366**, 329 (2008).
[14] M. G. Kitzbichler, M. L. Smith, S. R. Christensen, and E. Bullmore, PLoS Comput Biol **5**, e1000314 (2009).
[15] D. R. Chialvo, Nature Phys. **6**, 744750 (2010).
[16] M. E. Newman, Contemporary Phys **46**, 323 (2005).
[17] A. Clauset, C. Shalizi, and M. Newman, SIAM Rev **51**, 661 (2009).
[18] P. Dutta and P. Horn, Rev Mod Phys **53**, 497 (1981).
[19] J. Tyrcha, Y. Roudi, M. Marsili, and J. Hertz, J Stat Mech p. 03005 (2013).
[20] W. Bialek, I. Nemenman, and N. Tishby, Neural Comp. **13**, 2409 (2001).
[21] Y. Tu and G. Grinstein, Phys Rev Lett **94**, 208101 (2005).
[22] H. Touchette, Phys Rep **478**, 1 (2009).
[23] A. Tang, D. Jackson, J. Hobbs, W. Chen, J. Smith, H. Patel, A. Prieto, D. Petrusca, M. Grivich, A. Sher, et al., J Neurosci **28**, 505 (2008).
[24] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, Phys Rev E **69**, 056111 (2004).
[25] I. Nemenman, G. Lewen, W. Bialek, and R. de Ruyter van Steveninck, PLoS Comput Biol **4**, e1000025 (2008).
[26] I. Mastromatteo and M. Marsili, J Stat Mech p. 10012 (2011).
[27] S. Laughlin, Z Naturforsch, C, Biosci **36**, 910 (1981).
[28] N. Brenner, W. Bialek, and R. de Ruyter van Steveninck, Neuron **26**, 695 (2000).
[29] H. Berg, E. Coli *in Motion* (Springer, 2004).
[30] I. Nemenman, in *Quantitative Biology: From Molecular to Cellular Systems*, edited by M. Wall (CRC Press, 2012), p. 73.