Statistical mechanics of letters in words

Greg J. Stephens[1,2,3] and William Bialek[1,2,4]
[1]*Joseph Henry Laboratories of Physics, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Center for the Study of Mind, Brain and Behavior, Princeton University, Princeton, New Jersey 08544, USA*
[4]*Princeton Center for Theoretical Science, Princeton University, Princeton, New Jersey 08544, USA*
(Received 15 May 2009; revised manuscript received 23 March 2010; published 25 June 2010)

We consider words as a network of interacting letters, and approximate the probability distribution of states taken on by this network. Despite the intuition that the rules of English spelling are highly combinatorial and arbitrary, we find that maximum entropy models consistent with pairwise correlations among letters provide a surprisingly good approximation to the full statistics of words, capturing ~92% of the multi-information in four-letter words and even "discovering" words that were not represented in the data. These maximum entropy models incorporate letter interactions through a set of pairwise potentials and thus define an energy landscape on the space of possible words. Guided by the large letter redundancy we seek a lower-dimensional encoding of the letter distribution and show that distinctions between local minima in the landscape account for ~68% of the four-letter entropy. We suggest that these states provide an effective vocabulary which is matched to the frequency of word use and much smaller than the full lexicon.

DOI: 10.1103/PhysRevE.81.066119 PACS number(s): 89.75.Fb, 05.65.+b, 64.60.De, 89.70.Cf

Many complex systems convey an impression of order not easily captured by the traditional tools of theoretical physics. Thus, it is unclear what order parameter or correlation function we should compute to detect that natural images are composed of solid objects [1], nor is it obvious what features of the amino acid sequence distinguish foldable proteins from random polymers. Recently, several groups have tried to simplify the problem of characterizing order in biological systems using the classical idea of maximum entropy [2]. Maximum entropy models consistent with pairwise correlations among neurons have proven surprisingly effective in describing the patterns of activity in real networks ranging from the retina [3,4] to the cortex [5]; these models are identical to the Ising models of statistical mechanics, which have long been explored as abstract models for neural networks [6]. Similar methods have been used to analyze biochemical [7] and genetic [8] networks, and these approaches are connected to an independent stream of work arguing that pairwise correlations among amino acids may be sufficient to define functional proteins [9]. Because of the immediate connection to statistical mechanics, this work also provides a natural path for extrapolating to the collective behavior of large networks, starting with real data [10]. Here, we test the limits of these ideas, constructing maximum entropy models for the sequence of letters in words.

As non-native speakers know well, the rules of English spelling can seem arbitrary and almost paradigmatically combinatorial (i before e except after c). In contrast, maximum entropy constructions based on pairwise correlations ignore such higher-order combinatorial effects [11] yet are rich enough to include complex systems such as spin glasses [12]. Thus, the statistics of letters in words provides an interesting test of the ability of the pairwise approach to capture the structure of natural distributions. There is a long history of statistical approaches in the analysis of language, including applications of maximum entropy ideas [13], while opposition to such statistical approaches was at the foundation of 40 years of linguistic theory [14,15]; for a recent view of these debates, see Ref. [16]. Our goal here is not to enter into these controversies about language in the broad sense, but rather to probe the power of pairwise interactions to capture seemingly complex structure. While our discussion is focused on four-letter words, similar results apply with three and five letters. Even with only four letters, there are $N=(26)^4=456\,976$ possible words, but only a tiny fraction of these are real words in English.

To analyze the interactions among letters, we sample the full joint distribution $P(\ell_1,\ell_2,\ell_3,\ell_4)$ from two large corpora [17]: a collection of novels from Jane Austen and a large compilation of American English contained in the American National Corpus (ANC). The maximum possible entropy of the full joint distribution is $S_{\rm rand}=4\log_2(26)=18.802$ bits. Letters occur with different probabilities, however, so even if we compose words by choosing letters independently and at random out of real text, the entropy will be lower than this. A more precisely defined "independent model" is the approximation

$$P(\ell_1,\ell_2,\ell_3,\ell_4)\approx P^{(1)}=\prod_{i=1}^{4}P(\ell_i), \tag{1}$$

where we note that each of $P(\ell_i)$ is different because letters are used differently at different positions in the word. In the Austen corpus, this independent model has an entropy $S_{\rm ind}=14.083\pm0.001$ bits, while the full distribution has entropy of just $S_{\rm full}=6.92\pm0.003$ bits [18]. The difference between these quantities is the multi-information, $I\equiv S_{\rm ind}-S_{\rm full}=7.163$ bits, which measures the amount of structure or correlation in the joint distribution of letters that form real words. Thus, correlations restrict the vocabulary by a factor of $2^I\sim143$ relative to the number of words that would be allowed if letters were chosen independently. Sampled from natural text, $P(\ell_1,\ell_2,\ell_3,\ell_4)$ depends both on the lexicon of four-letter words and on the frequency with which the words appear in the corpora. In the Jane Austen corpus we find that

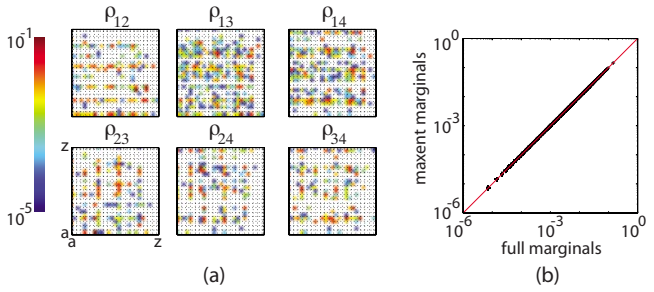

FIG. 1. (Color online) (a) The six pairwise marginal distributions of four-letter words sampled from the Jane Austen corpus. Common letter pairs such as "th" in ρ_{12} are apparent in their large marginal probability. (b) The iterative scaling algorithm solves the constrained maximization problem to high precision. All pairwise marginal components of the full distribution are compared to the marginals constructed from the computed maximum entropy distribution.

phonetic constraints reduce the number of allowable words from $N_{\text{rand}}=26^4=456\,976$ to $N=763$ while the unequal frequencies with which these words are used yield an effective vocabulary size of $N_{\text{eff}}=2^{S_{\text{full}}}\sim 121$.

Maximum entropy models based on pairwise correlations are equivalent to Boltzmann distributions with pairwise interactions among the elements of the system (see, for example, Ref. [11]); in our case this means approximating $P(\ell_1,\ell_2,\ell_3,\ell_4)\approx P^{(2)}$,

$$P^{(2)}(\ell_1,\ell_2,\ell_3,\ell_4)=\frac{1}{Z}\exp\left[-\sum_{i>j}V_{ij}(\ell_i,\ell_j)\right], \tag{2}$$

where V_{ij} are "interaction potentials" between pairs of letters and Z serves to normalize the distribution; because the order of letters matters, there are six independent potentials. Each potential is a 26×26 matrix, but the zero of energy is arbitrary, so this model has $6\times(26^2-1)=4050$ parameters, more than $100\times$ less than the number of possible states. Note that interactions extend across the full length of the word, so that the maximum entropy model built from pairwise correlations is very different from a Markovian model which only allows each letter to interact with its neighbor.

We determine the interaction potentials V_{ij} by matching to the pairwise marginal distributions, that is, by solving the six coupled sets of 26^2 equations,

$$\rho_{12}(\ell,\ell')=\sum_{\ell_3,\ell_4}P^{(2)}(\ell,\ell',\ell_3,\ell_4), \tag{3}$$

and similarly for the other five pairs. As shown in Fig. 1(a), the pairwise marginals sampled from English are highly structured; many entries in the marginal distributions are exactly zero, even in corpora with millions of words. Construction of maximum entropy models for large systems is difficult [19], but for $\sim 5\times 10^5$ states as in our problem relatively simple algorithms suffice [20]. We see from Fig. 1(b) that these methods succeed in matching the observed pairwise marginals with high precision.

As shown at left in Fig. 2, the maximum entropy model with pairwise interactions does a surprisingly good job in

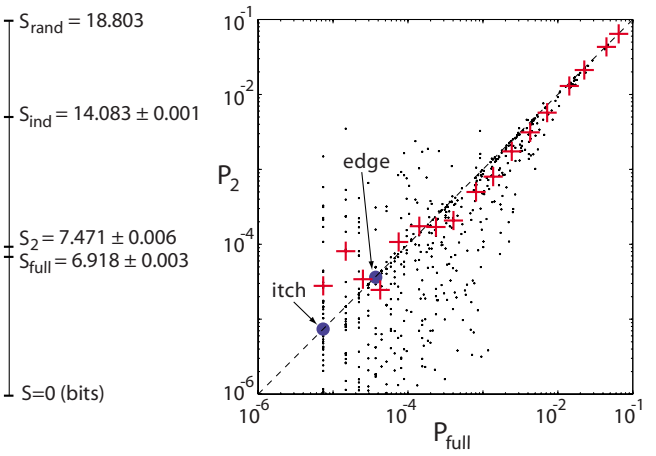

FIG. 2. (Color online) (left) The pairwise maximum entropy model provides an excellent approximation to the full distribution of four-letter words, capturing 92% of the multi-information. (right, dots) Scatter plot of the four-letter word probabilities in the full distribution P_{full} vs the corresponding probabilities in the maximum entropy distribution P_2. (right, red crosses) To facilitate the comparison we divided the full probability into 20 equally logarithmic-spaced bins and computed the mean maximum entropy probability conditioned on the states in the full distribution within each bin. The dashed line marks the identity. (right, blue circles) We singled out two words, edge and itch, whose small probability is well captured by the pairwise model yet whose sounds, "dge" and "tch," could have resulted from a triplet interaction.

capturing the structure of the full distribution. In the Austen corpus the model predicts an entropy $S_2=7.48$ bits, which means that it captures 92% of the multi-information, and similar results are found with the ANC, where we capture 89% of the multi-information. Pairwise interactions thus restrict the vocabulary by a factor of $2^{S_{\text{ind}}-S_2}\sim 100$ relative to the words which are possible by choosing letters independently.

While very close, the maximum entropy model is not perfect, as shown at right in Fig. 2. There is good agreement on average, especially for the more common words, but with substantial scatter. On the other hand, there are particular words with low probability, whose frequency of use is predicted with high accuracy. We have singled out two of these, "edge" and "itch," which contain sounds composed of three letters in sequence. While we might have expected that these combinations require three-body interactions among the letters, it seems that pairwise couplings alone are sufficient.

Another way of looking at structure in the distribution of words is the Zipf plot [21], the probability of a word's use as a function of its rank (Fig. 3). If we look at all words, we recover the approximate power law which initially intrigued Zipf; when we confine our attention to four-letter words, the long tail is cut off [22]. The maximum entropy model does a good job of reproducing the observed Zipf plot, but removes some weight from the bulk of the distribution and reassigns it to words which do not occur in the corpus, repopulating the tail. Importantly, as shown in the inset to Fig. 3, many of these "non-words" are perfectly good English words which happen not to have been used by Jane Austen. Quantitatively, the maximum entropy models assign 15% of the probability

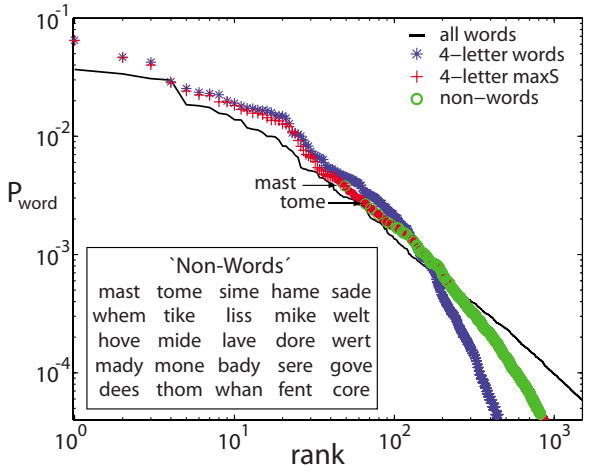

FIG. 3. (Color online) The Zipf plot for all words in the corpus (solid black line), four-letter words in the corpus (blue asterisks), and four-letter words in the maximum entropy model (red crosses). Green circles denote non-words, states in the maximum entropy model that did not appear in the corpus. The 25 most likely non-words are shown in the text inset (ordered in decreasing probability from left to right and top to bottom). Some of these are recognizable as real words that did not appear in the corpus, and even the others have plausible spelling.

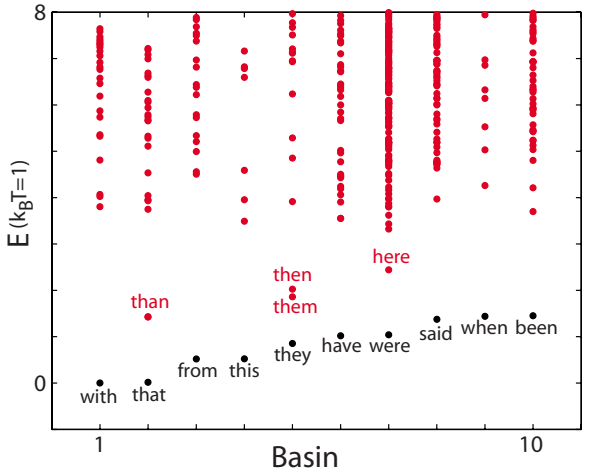

FIG. 4. (Color online) The energy landscape of the maximum entropy model estimated from the ANC corpus. To find the local minima we take each word in the corpus, compute the change in energy for all single-letter replacements, substitute the letter replacement which is lowest in energy, and repeat until letter replacements only increase the energy. This dynamics assigns each word to unique local minima (black filled circle with the lowest energy in each basin) and distinctions between these local minima account for ~68% of the full entropy. We order the basins by decreasing probability of their ground states and show only the lower-energy excitations in each basin and the first ten ground states. The energy landscape suggests that an effective vocabulary is built by first populating distinct basin states (which are error correcting by construction). Finer distinctions are then contained in the low-lying excitations. The arbitrary energy offset is chosen so the most likely stable word has zero energy.

to words which do not appear in the corpus, but of these 1/5 are words that can be found in the dictionary, and the same factor for the "correct discovery" of new words is found with the ANC. We also note that even words not found in the dictionary are speakable combinations of letters, not obviously violating known spelling rules.

The maximum entropy models discussed above were introduced to model the full joint distribution of letters $P(\ell_1,\ell_2,\ell_3,\ell_4)$ which includes the effects of both word frequencies and word occurrence. To separate these aspects we also considered a "dictionary" distribution that is sampled from the same sources but in which we assign equal probability to each word appearing in the corpus. In the dictionary distribution the correlations between letters arise entirely from phonetic constraints. For the Jane Austen corpus, we find $S_{\rm full}^{\rm dict}=\log 763=9.57$ bits while the pairwise maximum entropy model recovers $S_2^{\rm dict}=11.27$ bits. Since the independent entropy is $S_{\rm ind}^{\rm dict}=17.58$ bits, the pairwise model captures 79% of the multi-information and thus 79% of the correlations induced solely by phonetic constraints. Interestingly, we note that the pairwise models do better when modeling the statistics of the natural letter distribution. The ability of these models to account for phonetic rules is also seen in the reasonable spellings of the non-words shown in Fig. 3.

If we take Eq. (2) seriously as a statistical-mechanics model, then we have constructed an energy landscape $E(\ell_1,\ell_2,\ell_3,\ell_4)=\Sigma_{i>j}V_{ij}(\ell_i,\ell_j)$ on the space of words, much in the spirit of Hopfield for the states of neural networks [6]. In this landscape there are local minima, combinations of letters for which any single-letter change will result in an increase in the energy (Fig. 4). In the maximum entropy model for the ANC, there are 136 of these local minima, of which 118 are real English words, capturing nearly two thirds (63.5%) of the full distribution; similar results are obtained in the Austen corpus. We note that such "stable words" have the property that any single-letter spelling error can always be corrected by relaxing to the nearest local minimum of the energy. It is tempting to suggest that if we could construct the energy landscape for sentences (rather than for words), then almost all legal sentences would be locally stable.

In English we use only a very small fraction (~1/3700) of the roughly half million possible four-letter combinations. The hierarchy of maximum entropy constructions [11] allows us to decompose these spelling rules into contributions from interactions of different order. In particular, the size of our (reduced) vocabulary can be written as $N=\epsilon N_{\rm rand}$, where

$$\epsilon = 2^{-(S_{\rm rand}-S_{\rm ind})}2^{-(S_{\rm ind}-S_2)}2^{-(S_2-S_{\rm full})}. \quad (4)$$

A significant factor $(2^{-(S_{\rm rand}-S_{\rm ind})}\sim 1/26)$ comes from the unequal probabilities with which individual letters are used, a larger factor $(2^{-(S_{\rm ind}-S_2)}\sim 1/100)$ comes from the pairwise interactions among letters, and higher-order interactions contribute only a small factor $(2^{-(S_2-S_{\rm full})}\sim 1/1.5)$. Similar results hold for three- and five-letter words where—in the ANC, for example—we, respectively, capture 94% and 84% of the multi-information, suggesting that these principles are a general property of English spelling. The pairwise model represents an enormous simplification, which nonetheless has the

power to capture most of the structure in the distribution of letters and even to discover combinations of letters that are legal but unused in the corpora from which we have learned. The analogy to statistical mechanics also invites us to think about the way in which combinations of competing interactions enforce a complex landscape, singling out words which can be transmitted stably even in the presence of errors. Although our primary interest has been to test the power of the maximum entropy models, these ideas of generalization and error correction seem relevant to understanding the cognitive processing of text [16,23].

We thank D. Chigirev, T. Mora, S. E. Palmer, E. Schneidman, and G. Tkačik for helpful discussions. This work was supported in part by the National Science Foundation Grants No. IIS-0613435 and No. PHY-0650617, by the National Institutes of Health Grants No. P50 GM071508 and No. T32 MH065214, and by the Swartz Foundation.

[1] See, for example, D. Mumford, in *New Directions in Statistical Signal Processing: From Systems to Brain*, edited by S. Haykin *et al.* (MIT Press, Cambridge, MA, 2005), pp. 3–34.
[2] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[3] E. Schneidman, M. J. Berry II, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).
[4] J. Shlens *et al.*, J. Neurosci. **26**, 8254 (2006).
[5] See the presentations at the Society for Neuroscience (http://www.sfn.org/am2007/); I. E. Ohiorhenuan and J. D. Victor, Annual Meeting of the Society for Neuroscience (unpublished), 615.8/O01; S. Yu, D. Huang, W. Singer, and D. Nikolić, Annual Meeting of the Society for Neuroscience (unpublished), 615.14/O07; M. A. Sacek *et al.*, Annual Meeting of the Society for Neuroscience (unpublished), 790.1/J12; A. Tang *et al.*, Annual Meeting of the Society for Neuroscience (unpublished), 792.4/K27.
[6] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[7] G. Tkačik, Ph.D. thesis, Princeton University, 2007.
[8] T. R. Lezon *et al.*, Proc. Natl. Acad. Sci. U.S.A. **103**, 19033 (2006).
[9] W. Bialek and R. Ranganathan, e-print arXiv:0712.4397.
[10] G. Tkačik, E. Schneidman, M. J. Berry II, and W. Bialek, e-print arXiv:q-bio.NC/0611072.
[11] E. Schneidman, S. Still, M. J. Berry II, and W. Bialek, Phys. Rev. Lett. **91**, 238701 (2003).
[12] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[13] A. L. Berger, S. Della Pietra, and V. J. Della Pietra, Comput. Linguist. **22**, 39 (1996).
[14] N. Chomsky, IRE Trans. Inf. Theory **2**, 113 (1956).
[15] "In one's introductory linguistics course, one learns that Chomsky disabused the field once and for all of the notion that there was anything of interest to statistical models of language. But one usually comes away a little fuzzy on the question of what, precisely, he proved;" S. Abney, in *The Balancing Act: Combining Statistical and Symbolic Approaches to Language*, edited by J. L. Klavans and P. Resnik (MIT Press, Cambridge, MA, 1996), pp. 1–26.
[16] F. Pereira, Philos. Trans. R. Soc. London, Ser. A **358**, 1239 (2000).
[17] The Austen word corpus was created via Project Guttenberg (www.gutenberg.org), combining text from *Emma*, *Lady Susan*, *Love and Friendship*, *Mansfield Park*, *Northhanger Abbey*, *Persuassion*, *Pride and Prejudice*, and *Sense and Sensibility*. Out of 676 302 total words in our Austen corpus there were 7114 unique words, 763 of which were four-letter words; the four-letter words occurred in the corpus a total of 135 441 times. We used the second release of the ANC (www.americannationalcorpus.org) with $\sim 2 \times 10^7$ words and restricted ourselves to words used more than 100 times, providing 798 unique four-letter words occurring 2 179 108 times. These numbers indicate that we can sample the distribution of four-letter words with reasonable confidence. To control for potential typographic errors, words were also checked against a large dictionary database (http://wordlist.sourceforge.net/12dicts-readme.html).
[18] For technical points about finite sample sizes and error bars, see N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, e-print arXiv:cs/0502017v1.
[19] T. Broderick, M. Dudik, G. Tkačik, R. Schapire, and W. Bialek, e-print arXiv:0712.2437.
[20] J. N. Darroch and D. Ratcliff, Ann. Math. Stat. **43**, 1470 (1972).
[21] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language* (Harvard University Press, Cambridge, MA, 1932).
[22] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).
[23] S. Dehaene, L. Cohen, M. Sigman, and F. Vinckier, Trends Cogn. Sci. **9**, 335 (2005).