The Boltzmann Sequence-Structure Channel

Abram Magner
Univ. of Illinois at Urbana-Champaign
Urbana, IL, USA
Email: anmagner@illinois.edu

Daisuke Kihara
Purdue University
W. Lafayette, IN, USA
Email: dkihara@purdue.edu

Wojciech Szpankowski
Purdue University
West Lafayette, IN, USA
Email: spa@cs.purdue.edu

***Abstract*—We rigorously study a channel that maps binary sequences to self-avoiding walks in the two-dimensional grid, inspired by a model of protein statistics. This channel, which we also call the Boltzmann sequence-structure channel, is characterized by a Boltzmann/Gibbs distribution with a free parameter corresponding to temperature. In our previous work, we verified experimentally that the channel capacity has a phase transition for small temperature and decays to zero for high temperature. In this paper, we make some progress towards explaining these phenomena. We first upper bound the conditional entropy between the input sequence and the output which exhibits a phase transition with respect to temperature. Then we derive a lower bound on the conditional entropy for some specific set of parameters. This lower bound allows us to conclude that the mutual information tends to zero for high temperature.**

I. Introduction

Information theory traditionally deals with the problem of transmitting sequences over a communication channel and finding the maximum number of messages that the receiver can recover with arbitrarily small probability of error. However, databases of various sorts have come into existence in recent years that require to transmit structural data (e.g., graphs and sets). Contemporaneously, there has been significant effort focused on understanding the equilibrated states and dynamics of biomolecules. In [1], we attempted an information-theoretic explanation of the following observation previously made by biophysicists: while the number of amino acid sequences observed in nature is large, the corresponding number of dissimilar tertiary structures (folding) to which the sequences have been observed to fold is relatively small. Additionally, the frequency distribution of protein families observed in nature exhibits power law characteristics. We provided experimental evidence that explains these observations by modeling the protein folding process as a *channel*. We gave evidence in support of the hypothesis that these complex phenomena might have interesting information theoretic underpinnings.

This new channel maps binary sequences (hydrophobic, denoted by H, and polar, denoted by P) into two-dimensional self-avoiding walks (also called folds) in a square lattice (see Figure 1). A binary sequence of length N induces a labeling of the nodes of each fold of the same length, and contacting nodes in a fold induces an energy function. In particular, the channel is defined by the Boltzmann/ Gibbs distribution with a free parameter corresponding to inverse temperature. We therefore call it the *Boltzmann sequence-structure channel*.

For this channel, the key parameter is the conditional entropy between the input sequence and the output fold. In this paper, we provide a mathematically rigorous foundation to estimate this entropy and show that it may exhibit a range of interesting behaviors with respect to temperature, depending on the settings of the parameters of the model.

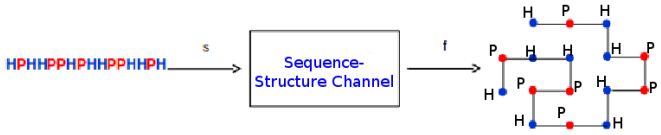

Fig. 1: A sequence passing through the channel and being paired with a fold given by a self-avoiding walk.

We now describe in more detail the construction of the channel. For each sequence s, the folds f are assigned energies $\mathcal{E}(f,s)$ by associating the ith symbol of s with the ith node of f and counting the number of different types of *contacts* between residues, that is, between neighboring, but not sequence-adjacent, nodes of the self-avoiding walk. These contact energies are given by a *scoring* matrix Q whose rows and columns are indexed by H and P. The channel is then defined by the Boltzmann distribution induced by the energies.

More precisely, for each even (for technical reasons explained below) perfect square integer N, we have an input set $\mathcal{S}_N$ consisting of 2^N sequences of length N over the alphabet $\{H,P\}$. The output set $\mathcal{F}_N$ consists of all directed self-avoiding walks of length N on a $\sqrt{N}\times\sqrt{N}$ integer lattice which start at $(0,0)$ and end at $(\sqrt{N}-1,\sqrt{N}-1)$. Regarding the channel, each fold is given a probability by labeling every point by a symbol (H or P) from the corresponding input sequence. Note that all but $O(\sqrt{N})$ points in the lattice have four neighbors (but only two contact points) since every walk fills the lattice completely. We endow each sequence/fold pair with an energy as follows: fix as a parameter of the model

978-1-5090-1805-5/16/$31.00 ©2016 IEEE

a symmetric 2×2 matrix $Q = \{Q_{ij}\}_{i,j\in\{1,2\}}$ over $\mathbb{R}$ (the scoring matrix). For each $f \in \mathcal{F}_N$ and $s \in \mathcal{S}_N$

$$\mathcal{E}(f,s) = 2(Q_{11}c_{HH} + Q_{22}c_{PP} + Q_{12}c_{HP}), \qquad (1)$$

where c_{xy} denotes the number of contacts $\{a,b\}$ such that $s_a = x$ and $s_b = y$ or vice-versa (throughout, for any sequence s and $j \in [N] = \{1,\dots,N\}$, we denote by s_j the jth symbol of s). Here, the multiplication by 2 is for mathematical convenience and is insignificant to the analysis.

We now define the channel by the conditional probability $p_N(f|s)$ that follows the Boltzmann distribution. More formally, let $\beta \geq 0$ be a real number (corresponding to an inverse temperature). Then we define

$$p_N(f|s) = p(f|s) = \frac{e^{-\beta\mathcal{E}(f,s)}}{Z(s,\beta)}, \quad Z(s,\beta) = \sum_{f\in\mathcal{F}_N} e^{-\beta\mathcal{E}(f,s)},$$

where the function Z is known as the *partition function*, which plays a central role in statistical mechanics models as a kind of generating function of configuration energies. Two quantities will play an especially important part in our analysis and results: the free energy $\gamma_N(\beta)$ that we define as

$$\gamma_N(\beta) = \frac{\mathbb{E}\log Z(S,\beta)}{\log|\mathcal{F}_N|} \qquad \gamma(\beta) = \limsup_{N\to\infty} \gamma_N(\beta).$$

We also denote by μ the exponential growth rate of the number of self-avoiding walks:

$$\mu_N = \frac{\log|\mathcal{F}_N|}{N}, \qquad \mu = \lim_{N\to\infty} \frac{\log|\mathcal{F}_N|}{N}.$$

Both are challenging to compute.

This channel is interesting from the information-theoretic point of view, irrespective of applications, primarily because it exhibits several unusual mathematical properties: first, it maps sequences to structures (i..e, self-avoiding walks) in a nontrivial way; second, it is a channel with full memory; and, finally, several information theoretic quantities associated with it (e.g., its capacity and conditional entropy for certain natural input distributions) exhibit phase transitions with respect to temperature for certain settings of the scoring matrix. Probabilistically, its analysis presents an interesting challenge because the nontrivial dependence structure between fold energies makes bounding the variance of the number of folds with a given maximum energy difficult. This in turn, complicates the calculation of the free energy, which plays a significant role in our calculations (and, for many models, is notoriously difficult to compute [2]). Since the exponential growth rate of the number of folds in the output alphabet appears in several quantities of interest, we also encounter combinatorial problems which are currently under active investigation.

We now review some of the relevant literature. Regarding self-avoiding walks (*SAWs*), [3] is a good general reference, including a discussion of the history of the use of SAWs as models for polymers. SAWs continue to be used as simple models for protein structures in molecular biology (see, e.g., [4], [5]). One of the fundamental problems in the theory of SAWs is the (asymptotic) enumeration of classes $\mathcal{F}_N$ of SAWs of length $N \to \infty$ with various constraints. In particular, the problem of proving the existence/determining the value of the limit $\lim_{N\to\infty} |\mathcal{F}_N|^{1/N}$ (called the *connective constant* of $\mathcal{F}_N$) is commonly studied and is quite challenging. There are a few general techniques for approaching such problems, sub/superadditivity arguments being the main ones.

We now review what is known about some relevant models from statistical physics. For general references, see [2], [6]. For a set Γ_N of configurations, each configuration $\xi \in \Gamma_N$ is endowed with its own (possibly random) *energy* $\mathcal{E}(\xi)$. The set Γ_N is then endowed with a probability distribution governed by this energy (chosen so as to have maximum entropy under the constraint that the system has a given energy density), known as the *Boltzmann/Gibbs* measure:

$$p(\xi) = \frac{e^{-\beta\mathcal{E}(\xi)}}{Z(\beta)},$$

where $\beta \in [0,\infty)$ is a free parameter which intuitively behaves like an inverse temperature, and Z above is the *partition function*, given by $Z(\beta) = \sum_{\xi\in\Gamma_N} e^{-\beta\mathcal{E}(\xi)}$. The main problem is to establish the existence/estimate the asymptotic value of the *free energy*:

$$\lim_{N\to\infty} \frac{\mathbb{E}[\log Z(\beta)]}{\log|\Gamma_N|}. \qquad (2)$$

This quantity is studied because other important parameters, such as the *entropy density* and *energy density* can be written in terms of it (see [6] for details). One of the simplest interesting models is the *random energy model* (REM), in which the configuration space has size 2^N, and the configurations are i.i.d. (exactly) Gaussian random variables: $\mathcal{E}(\xi) \sim \mathcal{N}(0, N/2)$. The free energy for this model is exactly solvable (which is unusual for these sorts of models):

$$\lim_{N\to\infty} \frac{\mathbb{E}[\log Z(\beta)]}{N} = \begin{cases} \beta^2/4 + \log 2 & \beta \leq 2\sqrt{\log 2} \\ \beta\sqrt{\log 2} & \beta \geq 2\sqrt{\log 2}. \end{cases}$$

Note that the free energy exhibits a phase transition with respect to temperature, since, for small β, it grows quadratically, while it grows linearly when $\beta \geq 2\sqrt{\log 2}$. This sort of phenomenon is quite common (though not universal) in statistical physics, and we will encounter it in our analysis in this paper. The situation becomes significantly more complicated when correlations between configurations are introduced as in our model.

We now move on to discuss our contributions. First, though the self-avoiding walk model and associated energy function for proteins has been considered empirically before [4], [5],

we appear to be the first to define the channel [1] that we consider here and study its information theoretic quantities. Of particular interest is the *capacity* of the channel: $C = \max_{p(S)}[H(F) - H(F|S)]$, where S and F are the input and output of the channel, respectively, and the maximum is taken over all probability distributions on the set of sequences; see [8]. In our previous work [1], we studied this quantity numerically. Specifically, using a specific scoring matrix taken from the biology literature, we computed the conditional probabilities constituting the channel for $N = 36$ (due to computational limitations, we could not do the same for much larger N). We then computed the capacity for various temperatures using the Blahut-Arimoto algorithm ([8]), resulting in Figure 2. We note two phenomena illustrated by the plot: first, there is a phase transition with respect to temperature in the capacity. Second, the capacity tends to 0 as temperature tends to infinity (for fixed N, this is simple to prove, but significantly more interesting when $N \to \infty$).

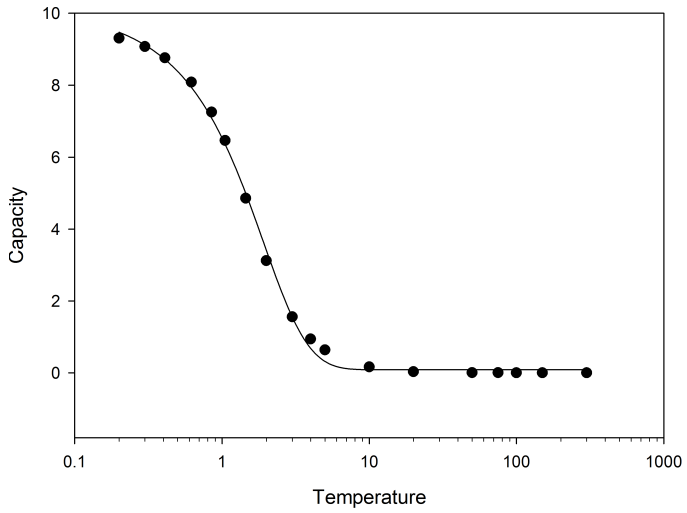

Fig. 2: Empirical evidence of a phase transition in channel capacity associated with 6×6 lattices (see [1]).

As a long-term goal, we would like to rigorously establish the asymptotic behavior of the capacity of this channel for all temperatures and suitable scoring matrices. Our focus in this work is more modest: we mainly study here the behavior of the conditional entropy for a memoryless source in the high-temperature regime (i.e., $\beta \xrightarrow{N\to\infty} 0$).

First, in Theorem 1 we give upper bounds on the free energy (hence the conditional entropy) whose behavior depends on the difference between the expected energies of a Boltzmann-distributed fold and one chosen uniformly at random. We then show how a series representation, involving the higher moments of the partition function, may be derived for the free energy via Taylor's expansion. Next, in Theorem 2 we present a class of scoring matrices for which the covariance between any two fold energies depends on the number of shared contacts between the two folds. For such matrices, we derive a formula for the variance of the partition function in terms of the number of contacts shared between two random folds, which implies a lower bound on the free energy. As an application of the lower bound, we give a sufficient condition on the temperature under which the mutual information between the channel input and output tends to 0.

The model presents several mathematical challenges: due to geometric constraints (e.g., Hamiltonicity), the configurations (folds) cannot easily be decomposed into subconfigurations. Thus, techniques which are useful for other models do not appear to be easily adapted to our case. Probabilistically, the correlation structure between fold energies does not appear to be captured by other existing models (e.g., the REM). Moreover, while many models are defined so that configuration energies are normally distributed, the fold energies are only *asymptotically* normally distributed. Finally, our analysis leads to some classic open questions about enumerating self-avoiding walks, including proving the existence of the connective constant for geometrically constrained walk sets and determining distributional information about the number of shared contacts between two randomly chosen folds.

II. Main results

A. Description of the model

Throughout, we use F to denote a *random* fold generated by choosing a random sequence according to some distribution and passing it through the channel. We generally use f to denote an arbitrary fixed fold. For any fold $f \in \mathcal{F}_N$, we denote the two-dimensional position of the jth node in f by $\pi_f(j)$. For any $j, k \in [N]$, we say that j and k are *sequence-adjacent* if $|j - k| = 1$ (here, $[N] = \{1, 2, \ldots, N\}$). We say that they are *lattice-adjacent* and that they form a *contact* if they are not sequence-adjacent and $\|\pi_f(j) - \pi_f(k)\|_1 = 1$ (here, $\|\cdot\|_1$ denotes the ℓ_1 norm on $\mathbb{Z}^2$). This allows us to define the energy $\mathcal{E}(f, s)$ as in (1). We also define $\mathcal{E}_{\beta,S}(F)$ (see (4)) to be the energy of the (random) fold at the output of the channel at inverse temperature β with the sequence S on its input.

We can also express the $\mathcal{E}(f, s)$ as a sum of *local energies*: for each $i \in [N]$, define $X_i = X_i(f, s)$ to be

$$X_i = Q_{11}c_{HH}(i) + Q_{22}c_{PP}(i) + Q_{12}c_{HP}(i),$$

where $c_{xy}(i)$, discussed above, denotes the number of contacts $\{i, j\}$ whose sequence elements are x and y or vice-versa (we note that the multiplication by 2 in (1) is because, by summing over all X_i, we count each contact twice).

We restrict our attention to a particular class of distributions on $\mathcal{S}_N$ that is natural to consider: the symbols are i.i.d. random variables, taking the value H with probability $p \in (0, 1)$ and P with probability $q = 1 - p$. That is, we take a binary

memoryless source with parameter p, which we denote by $\mathcal{B}_N(p)$. Then we can also write $\mathcal{E}(f,s) = \sum_{i=1}^{N} X_i(f,s)$. Clearly,

$$\mathbb{E}[\mathcal{E}(f,S)] = \sum_i \mathbb{E}[X_i(f,S)] = N\alpha + O(\sqrt{N})$$

with $\alpha/2 = p^2 Q_{HH} + 2pqQ_{HP} + q^2 Q_{PP}$ (with $\alpha \neq 0$ under mild conditions on Q and the sequence distribution), where boundary conditions contribute the $O(\sqrt{N})$. In contrast, $\mathbb{E}[\mathcal{E}_{\beta,S}(F)]$, the expected energy of a Boltzmann fold, is more difficult to compute. We discuss some of its properties below.

B. Statement of main results

We start with an expression for the conditional entropy. We have

$$\begin{aligned} H(F|S) &= -\sum_{s\in\mathcal{S}_N} p(s) \sum_{f\in\mathcal{F}_N} p(f|s)\log p(f|s) \\ &= \mathbb{E}[\log Z(S,\beta)] + \beta \sum_{s,f} p(f,s)\mathcal{E}(f,s) \\ &= \mathbb{E}[\log Z(S,\beta)] + \beta\mathbb{E}[\mathcal{E}_{\beta,S}(F)] \end{aligned} \tag{3}$$

where $\mathcal{F}_N$ denotes the set of self-avoiding walks of length N that we consider and we explicitly write

$$\mathbb{E}[\mathcal{E}_{\beta,S}(F)] = \sum_{s,f} p(f,s)\mathcal{E}(f,s). \tag{4}$$

The first and third equalities of (3) are elementary, and the second is by substitution of the definition of the channel into the right-hand side. Dividing by N on both sides, we have

$$\frac{H(F|S)}{N} = \frac{\log|\mathcal{F}_N|}{N} \cdot \frac{\mathbb{E}[\log Z(S,\beta)]}{\log|\mathcal{F}_N|} + \beta\frac{\mathbb{E}[\mathcal{E}_{\beta,S}(F)]}{N}.$$

It is easy to see that $\mathbb{E}[\log Z(S_N,\beta)] = O(N)$, so that the free energy $\gamma(\beta) < \infty$. Moreover, defining

$$\alpha_*(\beta,N) = \alpha_*(\beta) = \frac{\mathbb{E}[\mathcal{E}_{\beta,S}(F)]}{N},$$

it can be shown that, for $\beta = O(1)$ with respect to N, $\alpha_*(\beta)$ is bounded above as $N \to \infty$.

We note an important property of $\mathbb{E}[\mathcal{E}_{\beta,S}(F)]$ expressed in (4) for an arbitrary fold f (equivalently, a uniformly distributed fold f, since both have the same expected energy when labeled by a sequence from a memoryless source) $\mathbb{E}[\mathcal{E}_{\beta,S}(F)] \leq \mathbb{E}[\mathcal{E}(f,S)]$. This follows from an easy inductive proof, using the fact that the Boltzmann energy distribution is monotone decreasing (i.e., the Boltzmann distribution gives higher probability to lower energy folds).

We have the following upper bound on the free energy, and, hence, the conditional entropy (see [9] for a proof).

Theorem 1 (Upper bound on the conditional entropy for memoryless sources)**.** *For any distribution over $\mathcal{S}_N$, $\beta > 0$, and scoring matrix Q,*

$$\frac{H(F|S)}{N} = \mu\cdot\gamma_N(\beta) + \beta\alpha_*(\beta) + o(1). \tag{5}$$

Furthermore, when $S \sim \mathcal{B}_N(p)$, if the scoring matrix Q is such that, uniformly over all $f \in \mathcal{F}_N$,

$$\mathbf{Var}\ [\mathcal{E}(f,S)] \sim N\sigma^2, \qquad \sigma > 0,$$

then we have the following upper bound: for all $\beta > 0$,

$$\frac{H(F|S)}{N} \leq \mu_N - \beta(\alpha - \alpha_*(\beta)) + \frac{1}{2}\sigma^2\beta^2 - O(\beta N^{-1/2}), \tag{6}$$

and for bounded $\beta \geq \beta_* = \frac{\sqrt{2\mu}}{\sigma}$,

$$H(F|S) \leq \beta N(\sqrt{2\sigma^2\mu_N} - (\alpha - \alpha_*(\beta)) + O(N^{-1/2})). \tag{7}$$

The condition on the scoring matrix is quite general. It is equivalent to requiring that Q_{HH}, Q_{HP}, and Q_{PP} are not all equal (in this case, a typical contact energy has positive variance).

Remark Note that there is a trivial information-theoretic upper bound on $H(F|S)$:

$$H(F|S) \leq H(F) \leq \log|\mathcal{F}_N| = N\mu_N$$

Provided that $\beta = o(1)$ and $\alpha - \alpha_*(\beta) = \Theta(1)$, the first upper bound given above in ((6) beats this one. Similarly, if $\alpha - \alpha_*(\beta)$ is sufficiently large for any fixed β, the second upper bound (7) is nontrivial. Moreover, the proof of the second upper bound (7) implies that a refinement of the first upper bound (6) for all β yields a corresponding refinement in the second.

Our next theorem gives, for each $p \in (0,1)$, a natural class of scoring matrices that endows the set of fold energies with a correlation structure similar to that arising in several models associated with combinatorial optimization problems. In particular, the covariance between the energies of two folds f and g varies linearly with a measure of *overlap* between them: namely, the number of shared contacts between f and g (denoted by $k_{f,g}$). For such matrices, we establish a lower bound which holds for sufficiently small β, depending on the behavior of the moment generating function (MGF) of the random variable K which represents the number of shared contacts between two folds chosen uniformly at random with replacement.

Theorem 2 (Free energy lower bound for high temperature)**.** *Let $S \sim \mathcal{B}_N(p)$ for fixed $p \in (0,1)$. Let K denote the number of shared contacts between two folds $f, g \in \mathcal{F}_N$ chosen uniformly at random with replacement. There exists a scoring matrix for which, provided that*

$$\mathbb{E}_K[e^{3\beta_N^2\sigma^2 K}] = 1 + o(1), \tag{8}$$

and $\beta = \beta_N = o(1)$, we have

$$\frac{H(F|S)}{N} \geq \mu_N - \beta(\alpha - \alpha_*(\beta)) + \frac{1}{2}\beta^2\sigma^2 - o(1), \tag{9}$$

where the $o(1)$ is expressible in terms of $\mathbb{E}_K[e^{3\beta_N^2\sigma^2 K}]$.

We remark that while essentially nothing is known about K in the condition (8), we do know that $K \leq N+O(\sqrt{N})$, since that is the total number of contacts in a fold. Thus, a sufficient condition for (8) to hold is that $\beta = o(N^{-1/2})$. However, we conjecture that $K = O(1)$ with high probability. Note that the lower bound (9) matches the upper bound (6) up to the β term if $\alpha - \alpha_*(\beta) = \Theta(1)$ and the $o(1)$ term is $o(\beta)$.

Also, note that one cannot expect such a lower bound for general scoring matrices. This is because, for "most" matrices, the covariance of the energies of two contacts (i.e., unordered pairs of distinct sequence indices) that share exactly one node is positive, which implies that the covariance between two node energies is positive. This, in turn, implies that the covariance between *any* two fold energies is linear in N; that is, the dependence between fold energies is quite strong, in contrast with the situation in the REM. The scoring matrices considered in Theorem 2 are chosen so that the covariance between energies of nonidentical contacts is 0, so that the covariance between folds is only linear in the number of shared contacts (for details see [9]).

Remark For $\beta \xrightarrow{N\to\infty} 0$ and the class of scoring matrices considered in Theorem 2, we may be able to refine our estimate of the coefficient of β^2 in the expansion of the free energy by Taylor expanding the function $\log Z$ around $Z = \mathbb{E}[Z]$ and then taking expectations :

$$\mathbb{E}[\log Z] = \log \mathbb{E}[Z] - \frac{\mathbf{Var}\ [Z]}{2(\mathbb{E}[Z])^2} \tag{10}$$

$$+ \sum_{m=3}^{\infty} \frac{(-1)^{m+1}}{m} \cdot \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^m]}{(\mathbb{E}[Z])^m}. \tag{11}$$

This boils the problem down to the estimation of the centered moments of the partition function. For example, we can prove that the first two terms of the expansion (10) are

$$\log|\mathcal{F}_N| - \beta\alpha N + \frac{1}{2}\beta^2\sigma^2 N$$
$$- (1 + O(N^{-1/2}))(\mathbb{E}_K[e^{3\beta^2\sigma^2 K}] - 1)/2 + O(N^{-1/2}).$$

Provided that $\beta = o(N^{-1/2})$, the contribution of the variance term becomes asymptotically equivalent to $-3\beta^2\sigma^2\mathbb{E}[K]/2$. In particular, note that both the expected value and variance terms of (10) contribute to the coefficient of β^2. More generally, the mth moment may be written in terms of the MGFs of the random variables $K_{m,j}$, for $j = 1, \ldots, m$, defined to be the number of contacts shared among exactly j folds among m folds chosen uniformly at random with replacement. The random variable K is a special case: $K = K_{2,2}$. Provided that $K_{m,j}$ are sufficiently well-behaved, the series (10) above converges, and this gives a series representation for the coefficient of β^2, which may be bounded.

Depending on the asymptotics of the difference $\alpha - \alpha_*(\beta)$, Theorem 2 yields an interesting result about the mutual information $I(F;S) = H(F) - H(F|S)$ as the temperature tends to ∞. When α and $\alpha_*(\beta)$ are asymptotically equivalent and β is sufficiently small, the lower bound of Theorem 2 implies that $H(F|S) = \log|\mathcal{F}_N| - o(1)$. Thus, $I(F;S) = o(1)$.

Corollary 1. *With p and the scoring matrix Q as in Theorem 2, if β_N is such that $\alpha = \alpha_*(\beta_N) + \psi(N)$, where $\psi(N) = o(1)$ and $\beta_N\psi(N)N = o(1)$, and $\beta_N = o(N^{-2/3})$, then $I(F;S) = o(1)$.*

Note that one naturally expects that the mutual information tends to 0 when the temperature tends to infinity quickly enough (because then the Boltzmann distribution converges to uniformity), but this only becomes trivial when $\beta_N = O(1/N)$. The corollary, being a statement about the decay of the mutual information, is a small step in the direction of our stated goal of characterizing the capacity of the channel, in particular determining when it tends to 0.

Finally, we give an example scoring matrix which exhibits a rather different behavior from the ones in Theorem 2.

Theorem 3. *Let Q be the scoring matrix which maps $HH \mapsto -1/2$, $HP/PH \mapsto -1/4$, $PP \mapsto 0$. Then, for arbitrary sequence distributions, the free energy is given by*

$$\gamma(\beta) = 1 + \beta \limsup_{N\to\infty} \frac{\mathbb{E}[D_S(H)]}{\log|\mathcal{F}_N|},$$

where $D_S(H)$ is the number of i for which $S_i = H$. In the case of $S \sim \mathcal{B}_N(p)$, this becomes $\gamma(\beta) = 1 - \beta\alpha/\mu$.

Acknowledgment

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, by NSF Grant CCF-1524312, NIH Grant 1U01CA198941-01, and NCN grant UMO-2013/09/B/ST6/02258. W. Szpankowski is also with the Faculty of ETI, Gdańsk University of Technology, Poland.

References

[1] A. Magner, W. Szpankowski, and D. Kihara, "On the origin of protein superfamilies and superfolds," *Scientific Reports*, vol. 5, no. 8166, 2015.
[2] M. Talagrand, *Spin Glasses: A Challenge for Mathematicians*. New York, NY, USA: Springer, 2003.
[3] N. Madras and G. Slade, *The Self-Avoiding Walk*. Birkhäuser Basel, 2013.
[4] A. Sali, E. Shakhnovich, and M. Karplus, "How does a protein fold?" *Nature*, vol. 369, no. 6477, pp. 248–251, May 1994. [Online]. Available: http://dx.doi.org/10.1038/369248a0
[5] H. K. Nakamura and M. Sasai, "Population analyses of kinetic partitioning in protein folding," *Proteins Structure, Function, and Genetics*, vol. 43, pp. 280–291, 2001.
[6] M. Mézard and A. Montanari, *Information, Physics, and Computation*. New York, NY, USA: Oxford University Press, Inc., 2009.
[7] G. Parisi, "A sequence of approximate solutions to the s-k model for spin glasses," *Journal of Physics A*, vol. 13, pp. L–115, 1980.
[8] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2006.
[9] A. Magner, W. Szpankowski, and D. Kihara, "A study of the Boltzmann sequence-structure channel," 2015 (see http://www.cs.purdue.edu/homes/spa/papers/boltzmann15.pdf).