Phase Transitions in a Sequence-Structure Channel

Abram Magner
Department of Computer Science
Purdue University
West Lafayette, IN, USA
Email: anmagner@purdue.edu

Daisuke Kihara
Dept. of Biol. Sci./Comp. Sci.
Purdue University
West Lafayette, IN, USA
Email: dkihara@purdue.edu

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

***Abstract*—We study an interesting channel which maps binary sequences to self-avoiding walks in the two-dimensional grid, inspired by a model of protein folding from statistical physics. The channel is characterized by a Boltzmann/Gibbs distribution with a free parameter corresponding to temperature. We estimate the conditional entropy between the input sequence and the output fold, giving an upper bound which exhibits an unusual phase transition with respect to temperature.**

I. Introduction

A central object of study in molecular biology is the *protein folding process* by which sequences of *amino acids* are transformed into three-dimensional structures. This process has been studied empirically via a *lattice* model [4], [5], in which amino acid sequences are represented by sequences from the alphabet $\{H, P\}$ (*H*ydrophobic and *P*olar residues). Furthermore, protein structures are represented by self-avoiding walks on a two-dimensional square lattice, which we call *folds* (see Figure 1).

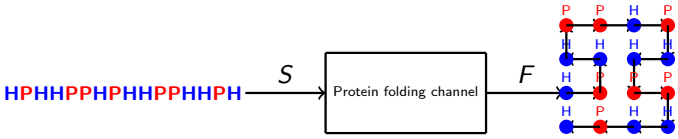

Fig. 1: A sequence passing through the channel and being paired with a fold given by a self-avoiding walk.

For each sequence s, the folds f are assigned energies $\mathcal{E}(f|s)$ depending on 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 weighted by a *scoring* matrix Q whose rows and columns are indexed by H and P. Since hydrophobic interactions are a dominant force for protein folding, it is reasonable to classify amino acids into hydrophobic (H) and polar (P). Thus, in the lattice model, contacts between H and H are more favored (lower energy) than H and P interactions [6]. In principle, nature will fold a sequence s to its lowest energy fold f, subject to some noise.

This model inspires the study of an associated channel (which we first considered in [3] from a more biological and empirical perspective; here we investigate it for its mathematical appeal) which probabilistically maps binary sequences to two-dimensional structures. More precisely, the channel is defined via the Boltzmann distribution induced by the energies. For each 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 $(\sqrt{N} - 1, \sqrt{N} - 1)$. 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 a symmetric 2×2 matrix $Q = \{Q_{ij}\}_{i,j \in \{1,2\}}$ over $\mathbb{R}$ (the *scoring matrix*). For $f \in \mathcal{F}_N$ 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 $j \in [N] = \{1, \ldots, N\}$, 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. Then we define the folding channel by the conditional probability $p_N(f|s)$ that follows the *Boltzmann* distribution. More precisely, let $\beta \geq 0$ be a real number (corresponding to an inverse temperature). Then we postulate

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

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.

This channel is interesting primarily because it exhibits several unusual mathematical properties: first, it maps sequences to structures (i..e, self-avoiding walks); 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) likely exhibit

phase transitions with respect to temperature, which seems to be an uncommon phenomenon in channels that are generally studied. 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, which, in turn, complicates the calculation of a quantity called the *free energy*, discussed below. 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.

The present paper explores the information theoretic properties of this model, asymptotically as $N \to \infty$. We focus on the conditional entropy $H(F|S)$ for a certain natural class of sequence distributions.

We now summarize our main findings. In the next section, for sequences generated by a memoryless source, we first express the conditional entropy as

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

While it is an easy exercise to justify the linear growth of the average energy (i.e., $\mathbb{E}[\mathcal{E}(F|S)] \sim \alpha N$ for some α), the behavior of $\mathbb{E}[\log Z(S,\beta)]$ is much more sophisticated. In fact, the *free energy*, defined as the rate of growth of $\mathbb{E}[\log Z(S,\beta)]$ with respect to N, has a long history [9]. In our case, it crucially depends on the number of of self-avoiding walks $\mathcal{F}_N$. As a matter of fact, enumeration of folds with unrestricted endpoints inside a square is an open problem [2]. Denoting this set by $\mathcal{F}'_N$, we are able to prove that $\log|\mathcal{F}'_N| = \Theta(N)$ (we will include the proof in the journal version of this paper) but the question of whether or not $\lim_{N\to\infty} \frac{\log|\mathcal{F}'_N|}{N}$ exists turns out to be more challenging. Hence, we restrict our fold set to those folds which begin at one corner and end at the opposite one. For this space, the corresponding normalized logarithm limit was shown to exist and be positive in [1]. We thus define

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

We then define the *free energy* to be

$$\gamma_N(\beta) = \frac{\mathbb{E}[\log Z(S,\beta)]}{\log|\mathcal{F}_N|}, \tag{2}$$

with the asymptotic free energy given by

$$\gamma(\beta) = \limsup_{N\to\infty} \gamma_N(\beta).$$

In Theorem 1 of this paper, we find an explicit expression for the asymptotics of the properly normalized conditional entropy $H(F|S)$ as a function of α, μ, and $\gamma_N(\beta)$. More interestingly, we show an upper bound on the free energy (and, hence, the conditional entropy) which exhibits a phase transition with respect to β. We expect a similar transition to also be present in the capacity, as empirical calculations show (see Figure 2). To the best of our knowledge, this seems to be a new phenomenon in the context of channel capacity. In addition, we show in Theorem 2 that, for a natural class of sequence distributions and a sufficiently well behaved class of scoring matrices, the fold energies (after proper normalization) are asymptotically Gaussian.

As for prior work, we are not aware of any analytical results for similar channels. While the lattice model of protein folding has been used previously in computational studies [5], [4], no probabilistic analysis under assumptions on the sequence distribution is available in open literature, at least to the best of our knowledge. Self-avoiding walks with various restrictions have been studied rigorously by several authors [2], [6], but not in the context of a sequence to structure channel. We note that [6] discusses phase transitions of the free energy of a model on self-avoiding walks, but their model involves a set of walks and an energy function different from the ones we consider.

The rest of the paper is organized as follows: Section II fixes some notation and states the main results. Section III gives some elements of the proofs of the results.

II. Main Results

We now fix some useful notation, make precise our definition of the energy function, and state our main results.

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). This allows us to define the energy $\mathcal{E}(f|s)$ as in (1).

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). Then we have

$$\mathcal{E}(f|s) = \sum_{i=1}^{N} X_i(f|s).$$

Clearly,

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

for some α depending on Q (with $\alpha \neq 0$ under mild conditions on Q and the sequence distribution), where boundary conditions contribute the $O(\sqrt{N})$.

In this conference version of the paper we restrict our attention to a particular class of distributions on $\mathcal{S}_N$ that is

natural to consider and whose analysis is feasible: 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 $\mathcal{B}_N(p)$. For such a model, we illustrate below how to compute α.

Example. *Computation of* α.
For the sake of simplicity, rather than giving the general formula, we now assume that

$$Q = \begin{array}{c} \\ H \\ P \end{array} \begin{array}{c} \begin{array}{cc} H & P \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

As mentioned earlier, $N - O(\sqrt{N})$ nodes have exactly two neighbors, so calculating α reduces to an application of linearity of expectation. For node i, the energy X_i is the sum of the contributions of its contacts with its two neighbors. Taking expectations, we easily get that

$$\mathbb{E}[X_i] = 2pq.$$

Hence,

$$\mathbb{E}[\mathcal{E}(F|S)] = 2pqN + O(\sqrt{N}).$$

A. Statement of Main Results

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

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

where $\mathcal{F}_N$ denotes the set of all self-avoiding walks of length N in a lattice. The first and third equalities 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

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

It is easy to see that $\mathbb{E}[\log Z(S_N,\beta)] = O(N)$, so that

$$\gamma(\beta) < \infty.$$

We have the following theorem, which is the main finding of this paper.

Theorem 1. *For any distribution over* $\mathcal{S}_N$*,* $\beta > 0$*, and scoring matrix* Q*, the limit* α *exists and is finite, and*

$$\limsup_{N\to\infty} \frac{H(F|S)}{N} \le \mu \cdot \gamma(\beta) + \beta\alpha. \tag{3}$$

Furthermore, if Q *comes from a certain broad class of scoring matrices (satisfying a "niceness" condition, discussed in Section III-A), for all but finitely many choices of* p*, when* $\mathcal{S}_N \sim \mathcal{B}_N(p)$*, there exists* $\sigma^2 > 0$ *such that, uniformly over all* $f \in \mathcal{F}_N$*,*

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

Then we have the following two upper bounds:

$$\limsup_{N\to\infty} \frac{H(F|S)}{N} \le \begin{cases} \mu + \frac{1}{2}\sigma^2\beta^2 & \beta > 0 \\ \beta\sqrt{2\sigma^2\mu} & \beta \ge \beta_* = \frac{\sqrt{2\mu}}{\sigma}. \end{cases}$$

Note that the free energy appears in the asymptotic expression given for $H(F|S)$ in Theorem 1. Thus, since $\mu > 0$, a phase transition with respect to β in the free energy implies a phase transition in the conditional entropy. In passing we point out that the phase transition shown above of the conditional entropy, and most likely the channel capacity (see Figure 2), seem to be new and unexpected information theoretic phenomena.

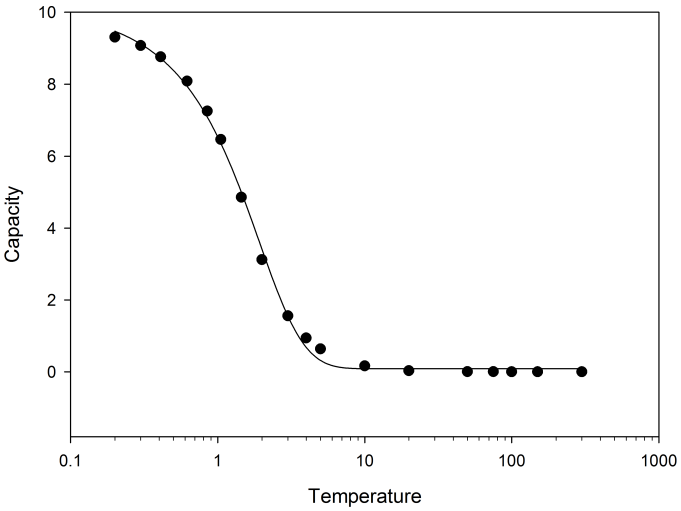

Fig. 2: Empirical evidence of a phase transition in channel capacity. Here, the capacity at various temperatures for the channel associated with 6×6 lattices is depicted. See [3] for the full figure.

The proof of Theorem 1 is based on the Central Limit Theorem (CLT) of the energy $\mathcal{E}(f|S)$ for each f, which we formulate next.

Theorem 2 (Central limit theorem for fold energies)**.** *Let* $S_N \sim \mathcal{B}_N(p)$ *for fixed* p*. Let, for any* $f \in \mathcal{F}_N$*,*

$$\hat{\mathcal{E}}_N = \frac{\mathcal{E}(f|S_N) - \mathbb{E}[\mathcal{E}(f|S_N)]}{\sqrt{N}},$$

and denote by $F_N(\cdot)$ *the distribution function of* $\hat{\mathcal{E}}_N$*. Then, provided* $\sigma^2 > 0$ *as in Theorem 1, for all but finitely many choices of* p*,*

$$\|F_N - \Phi\|_\infty = O(N^{-1/2}),$$

where the $O(\cdot)$ is uniform over all folds. Here, Φ denotes the distribution function of the normal distribution with mean 0 and variance σ^2.

The requirement that $\sigma^2 > 0$ is really a condition on the scoring matrix Q. We call it *niceness* of Q and will define and discuss it further in Section III-A.

In addition to being interesting in its own right, the preceding central limit theorem is a key part in our derivation of the upper bounds exhibited in Theorem 1. Furthermore, it gives some indication that our model behaves asymptotically somewhat like other previously considered models (e.g., the Random Energy Model (REM) or the Sherrington-Kirkpatrick model [9]).

Finally, we remark that the existence of the limit $\lim_{N\to\infty} \frac{\log|\tilde{\mathcal{F}}_N|}{N}$ for a class of folds $\tilde{\mathcal{F}}_N$ can be demonstrated by some flavor of superadditivity argument: one shows that, for each $a \in \{1, \ldots, N-1\}$, there is an injection from $\tilde{\mathcal{F}}_a \times \tilde{\mathcal{F}}_{N-a}$ to $\tilde{\mathcal{F}}_N$, so that we have

$$\log|\tilde{\mathcal{F}}_N| \geq \log|\tilde{\mathcal{F}}_a| + \log|\tilde{\mathcal{F}}_{N-a}|.$$

In certain cases, this injection can be constructed geometrically: if $(f, g) \in \tilde{\mathcal{F}}_a \times \tilde{\mathcal{F}}_{N-a}$, then one creates a fold in $\tilde{\mathcal{F}}_N$ by concatenating f and g, possibly after some rotation, translation, or flipping. Then one applies Fekete's theorem on superadditive sequences [7] to conclude the existence of the desired limit.

III. Proof Sketches

In this section, we give proof sketches. We start with Theorem 2, as it will be used in the proof of the upper bounds.

A. Proof of Theorem 2

The central limit theorem for fold energies follows by applying a result on *m-dependent random fields* given in [8]. Slightly specifying to our case and using our notation, it can be stated as follows.

Theorem 3. *Suppose that for some $M > 0$, $\mathbb{E}[X_i^8] \leq M < \infty$ for all i and that $\{X_i(f|S)\}_{i\in[N]}$ is m-dependent, for some $m > 0$. Provided $\liminf_{N\to\infty} \frac{\mathbf{Var}\ [\mathcal{E}(f|S)]}{N} > 0$, we have*

$$\|F_N - \Phi\|_\infty = O(N^{-1/2}).$$

We first establish m-dependence. This follows easily from the fact that the local energy of a node i in a given fold can only be dependent on the local energies of those nodes j that are within a lattice-adjacency neighborhood of i of some fixed, finite radius. This, in turn, follows from the independent choice of the sequence elements. Thus, we have m-dependence with $m = 2$.

It is further required that the variance of $\mathcal{E}(f|S)$ grows at least linearly with N. We shall establish that $\mathbf{Var}\ [\mathcal{E}(f|S)] = \Theta(N)$ (subject to the niceness condition on Q). We have

$$\mathbf{Var}\ [\mathcal{E}(f|S)] = \sum_{i=1}^{N} \mathbf{Var}\ [X_i] + 2 \sum_{1\leq i<j\leq N} \mathrm{Cov}[X_i, X_j].$$

Since $N - o(N)$ nodes have exactly two contacts, the dominant contribution to the first sum comes from those nodes, all of which have the same variance $v(p)$, a polynomial in p with coefficients that are polynomials in the entries of Q.

Note, then, that if nodes i and j are not lattice-adjacent, then $\mathrm{Cov}[X_i, X_j] = 0$. Thus, any node i is involved in at most 3 nonzero covariance terms. In fact, $N - o(N)$ nodes are involved in exactly 2 such terms. All such nodes i and j have covariance equal to some fixed $r(p)$, a polynomial in p with coefficients that are polynomials in the variables Q_{HH}, Q_{HP}, Q_{PP}.

By conditioning on the symbols assigned to nodes i and j and their other two lattice neighbors, both $v(p)$ and $r(p)$ can be computed exactly. Thus, we have

$$\mathbf{Var}\ [\mathcal{E}(f|S)] = N \cdot (v(p) + 2r(p)) + o(N).$$

We call $V(p) = v(p) + 2r(p)$ the *variance polynomial* of Q. Provided it is not identically 0 (a property of Q which we call *niceness*), it has finitely many roots, at which the variance is $o(N)$. Excluding these roots, the variance is $\Theta(N)$, as claimed, and we set $\sigma^2 = V(p)$.

Finally, it is required that, for all i, $\mathbb{E}[X_i^8] < \infty$. Since X_i is bounded between two constants with probability 1, all moments exist, and the proof is complete.

B. Proof of Theorem 1: Upper bounds

The overall structure of the proof is similar to that given by Talagrand for the Random Energy Model in [9]. The main challenge comes from the fact that, whereas, in the REM, all energies are Gaussian distributed, our fold energies are only *asymptotically* Gaussian.

In particular, we need a lemma describing the asymptotic behavior of the moment-generating function $\phi_N(t\sqrt{N})$ of $\hat{\mathcal{E}}_N$.

Lemma 1 (Asymptotics of the MGF of $\hat{\mathcal{E}}_N$)**.** *We have, for arbitrary fixed $t \in \mathbb{R}$,*

$$\lim_{N\to\infty} \frac{\log \phi_N(t\sqrt{N})}{N} = \log \phi(t) = \frac{1}{2}\sigma^2 t^2.$$

Here, $\phi(t)$ denotes the MGF of the normal distribution with mean 0 and variance σ^2.

Proof: The strategy is to show that the tails of the integral are negligible, leaving a central region that can be handled via Theorem 2. Using Hoeffding's inequality applied to a Doob martingale with respect to the local energies, we can show the following:

Lemma 2 (Large deviations of $\mathcal{E}(f|S)$)**.** *For any $t > 0$ and $f \in \mathcal{F}_N$,*

$$\Pr[|\mathcal{E}(f|S) - \mathbb{E}[\mathcal{E}(f|S)]| \geq tN] \leq 2\exp\left(-\frac{t^2 N}{C}\right),$$

for some constant $C > 0$.

The proof uses the fact that each node energy is dependent on at most a constant number of others to bound the martingale differences.

Now, let $F_N(x)$ be the distribution function of $\hat{\mathcal{E}}_N$. Taking the tail at $\theta\sqrt{N}$ of the MGF integral yields

$$\int_{\theta\sqrt{N}}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}F_N(x),$$

and integration by parts and application of Lemma 2 shows that the integral is $o(1)$, so negligible, provided $\theta > 2Ct$.

This leaves the central region:

$$\begin{aligned}\int_{-\theta\sqrt{N}}^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}F_N(x) &= \Theta(1)\int_{-\theta\sqrt{N}}^{\theta\sqrt{N}} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x)\\ &\sim \Theta(1)\int_{-\infty}^{\infty} e^{t\sqrt{N}x}\,\mathrm{d}\Phi(x)\\ &= \Theta(1)e^{\frac{1}{2}t^2\sigma^2 N}.\end{aligned}$$

Here, the first equality is by Theorem 2, and the asymptotic equivalence follows from the fact that the tails of the Gaussian distribution are negligible. ∎

Now, for the first upper bound, we proceed as follows.

$$\begin{aligned}\mathbb{E}[\log Z(S,\beta)] &\leq \log \mathbb{E}[Z(S,\beta)]\\ &= \log \sum_{f\in\mathcal{F}_N} e^{-\beta\mathbb{E}[\mathcal{E}(f|S)]}\mathbb{E}\left[e^{-\beta\sqrt{N}\frac{\mathcal{E}(f|S)-\mathbb{E}[\mathcal{E}(f|S)]}{\sqrt{N}}}\right]\\ &= \log \sum_{f\in\mathcal{F}_N} e^{-\beta\mathbb{E}[\mathcal{E}(f|S)]}\mathbb{E}\left[e^{-\beta\sqrt{N}\hat{\mathcal{E}}_N}\right]\\ &= \log \sum_{f\in\mathcal{F}_N} e^{-\beta\alpha N(1+o(1))}\cdot e^{\frac{1}{2}\sigma^2\beta^2 N(1+o(1))}\\ &= N\left(\frac{\log|\mathcal{F}_N|}{N}\right.\\ &\quad\left.-\beta\alpha(1+o(1)) + \frac{1}{2}\sigma^2\beta^2(1+o(1))\right).\end{aligned}$$

where we used Jensen's inequality to bring the expectation into the logarithm, and we used the fact that all of the relative errors are uniform over the set of folds. We thus have

$$\gamma(\beta) \leq \mu - \beta\alpha + \frac{1}{2}\sigma^2\beta^2.$$

For the second upper bound, the strategy is to find an upper bound on the derivative with respect to β of the function $\phi(\beta) = \mathbb{E}[\log Z(S,\beta)]$.

We have

$$\begin{aligned}&-\beta\min_{f\in\mathcal{F}_N}\mathcal{E}(f|S) \leq \log\left(\sum_{f\in\mathcal{F}_N} e^{-\beta\mathcal{E}(f|S)}\right)\\ &\implies \limsup_{N\to\infty}\frac{\mathbb{E}[-\min_{f\in\mathcal{F}_N}\mathcal{E}(f|S)]}{N} \leq \beta^{-1}\mu - \alpha + \frac{1}{2}\beta\sigma^2,\end{aligned}$$

where the first inequality is elementary, and the second is due to the first upper bound. We find that setting $\beta = \beta_* = \frac{\sqrt{2\mu}}{\sigma}$ minimizes the upper bound, yielding

$$\limsup_{N\to\infty}\frac{\mathbb{E}[-\min_{f\in\mathcal{F}_N}\mathcal{E}(f|S)]}{N} \leq \sqrt{2\sigma^2\mu} - \alpha.$$

Furthermore, for arbitrary β,

$$\begin{aligned}\phi'(\beta) &= \mathbb{E}\left[-\frac{\sum_{f\in\mathcal{F}_N}\mathcal{E}(f|S)e^{-\beta\mathcal{E}(f|S)}}{\sum_{f\in\mathcal{F}_N}e^{-\beta\mathcal{E}(f|S)}}\right]\\ &\leq \mathbb{E}\left[\left(-\min_{f\in\mathcal{F}_N}\mathcal{E}(f|S)\right)\frac{Z(S,\beta)}{Z(S,\beta)}\right]\\ &\leq N(\beta^{-1}\mu - \alpha + \frac{1}{2}\beta\sigma^2)\end{aligned}$$

Now, for $\beta > \beta_*$,

$$\phi(\beta) \leq \phi(\beta_*) + \phi'(\beta_*)(\beta - \beta_*),$$

since $\phi(\beta)$ is known to be convex. Applying the upper bound for $\phi'(\beta)$ yields the second upper bound in the theorem.

ACKNOWLEDGMENT

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSA Grant 130923, NSF Grant DMS-0800568, and and the MNSW grant DEC-2013/09/B/ST6/02258. W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland.

REFERENCES

[1] H. L. Abbott and D. Hanson. A lattice path problem. *Ars Combinatoria*, 6:163–178, 1978.

[2] M. Bousquet-Mélou, A. J. Guttmann, and I. Jensen. Self-avoiding walks crossing a square. *Journal of Physics A: Mathematical and General*, 38(42), 2005.

[3] Abram Magner, Wojciech Szpankowski, and Daisuke Kihara. On the origin of protein superfamilies and superfolds. *Scientific Reports*, 2015.

[4] H. K. Nakamura and M. Sasai. Population analyses of kinetic partitioning in protein folding. *Proteins Structure, Function, and Genetics*, 43:280–291, 2001.

[5] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369(6477):248–251, May 1994.

[6] C. E. Soteros and S. G. Whittington. Contacts in self-avoiding walks and polygons. *Journal of Physics A: Mathematical and General*, 34(19), 2001.

[7] Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[8] Hiroshi Takahata. On the rates in the central limit theorem for weakly dependent random fields. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64, 1983.

[9] Michel Talagrand. *Spin Glasses: A Challenge for Mathematicians*. Springer, New York, NY, USA, 2003.