Phase Transitions in a Protein Folding 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—In 2008 P. Nurse advocated the pursuit of biological phenomena through information processing. Along these lines, we aim to explain some of the salient characteristics of the mapping between amino acid sequences to protein folds through an information theoretic lens. To this end, we propose a protein folding channel that maps protein sequences into two-dimensional self-avoiding walks (also called folds) in a square lattice. The channel itself is characterized by the Boltzmann/Gibbs distribution with a free parameter corresponding to temperature. For such a channel we estimate the conditional entropy between the input sequence and the output fold. Interestingly, this entropy 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. In particular, it is known that the number of sequences of given length observed in nature is large in comparison with the number of dissimilar tertiary structures to which the sequences have been observed to fold. Furthermore, for a given sequence, the probability distribution over its possible folds tends to be highly concentrated. In fact, the frequency distribution of protein families observed in nature exhibits the power law characteristics [5]. Sequence distributions yield a mutual information close to the capacity. Recently, Magner et al. [2] used information theoretic tools to explain this observation by viewing the protein folding process as a channel and noting that skewed sequence distributions yield a mutual information close to the capacity. Moreover, a sequence distribution that comes close to achieving the capacity of the channel has a natural interpretation as a thermodynamically efficient encoding of protein structures.

Specifically, the authors of [2] observed that evolution provides a mechanism that leads to pressure to achieve channel capacity. An organism must produce certain quantities of a certain set of protein structures in order to survive to reproduce. Furthermore, the proteins must be encoded in the form of amino acid sequences, which fold according to some probability distribution over structures. It is energy-efficient to choose sequences with the most concentrated conditional structure distributions (i.e., sequences that are, relatively speaking, unambiguous), because relatively few “executions” of those sequences lead to the desired folds. Minimizing the ambiguity in the encoding has the effect of maximizing the mutual information between the sequence and structure distributions, and, thus, achieving the capacity of the sequence-structure channel. To summarize, it seems plausible that choosing a distribution over sequences that achieves the channel capacity minimizes energy expended in the protein production process, which can be evolutionarily advantageous.

In order to provide theoretical evidence of this hypothesis, the authors of [2] used a lattice model of protein folding [4], [6], in which amino acid sequences are represented by sequences from the alphabet \{H, P\} (hydrophobic and polar residues). Furthermore, protein structures are represented by self-avoiding walks on a two-dimensional square lattice (see Figure 1). For each sequence \(s\), the folds \(f\) are assigned energies \(E(f|s)\) 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 represented 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 [7]. In principle, Nature will fold a sequence \(s\) to its lowest energy fold \(f\).

More precisely, the protein folding channel is then 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 (which we call folds) of length \( N \) on a \( \sqrt{N} \times \sqrt{N} \) integer lattice. Note that all but \( o(N) \) 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 \( 2 \times 2 \) matrix \( Q = \{Q_{ij}\}_{i,j \in \{1,2\}} \) over \( \mathbb{R} \) (the scoring matrix). For \( f \in \mathcal{F}_N \) and \( s \in \mathcal{S}_N \)

\[
E(f|s) = 2(Q_{11}c_{HH} + Q_{22}c_{PP} + Q_{12}c_{HP}),
\]

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 \( j \in \mathcal{N} = \{1, \ldots, N\} \), we denote by \( s_j \) the \( j \)th 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

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

where the function \( Z \) is known as the partition function.

The present paper investigates this model from a mathematically rigorous perspective. In particular, we consider asymptotic properties of the channel for large \( N \). We focus on the minimum conditional entropy

\[
\min_S H(F|S),
\]

where the minimum is taken over all sequence distributions. This objective function is well motivated by our biological consideration. Indeed, by minimizing the conditional entropy we (Nature) select sequences that are as unambiguous as possible [2, 6].

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

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

While it is an easy exercise to justify the linear growth of the average energy \( \mathbb{E}[E(F|S)] \sim N\alpha \) for some \( \alpha \), 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 \( \mathbb{E}[\log Z(S, \beta)] \), has a long history [10]. In our case, it crucially depends on the number of self-avoiding walks \( \mathcal{F}_N \) in a lattice (square) of \( N \) points. As a matter of fact, enumeration of \( \mathcal{F}_N \) is an open problem [1]; however, we can prove that \( \log |\mathcal{F}_N| = \Theta(N) \). The question of whether or not \( \lim_{N \to \infty} \frac{\log |\mathcal{F}_N|}{N} \) exists turns out to be more challenging, so we work with \( \mu := \limsup_{N \to \infty} \frac{\log |\mathcal{F}_N|}{N} \). We then define the free energy to be

\[
\gamma(\beta) = \limsup_{N \to \infty} \frac{\mathbb{E}[\log Z(S, \beta)]}{\log |\mathcal{F}_N|},
\]

In Theorem 1 of this paper, we find an explicit expression for the limit superior of the properly normalized conditional entropy \( H(F|S) \) as a function of \( \alpha, \mu \) and \( \gamma(\beta) \). More interestingly, we show that the conditional entropy (the free energy) exhibits a phase transition with respect to \( \beta \). We expect that this phase transition will also be present in the channel capacity. 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 already mentioned, the folding channel was proposed for the first time in [2], and we are not aware of any analytical results for similar channels. We should point out that this is an interesting channel for many reasons: first, it maps sequences into two-dimensional structures (self-avoiding walks); second, it is a channel with full memory; and finally, for some natural input distributions, it exhibits a surprising phase transition with respect to temperature. Furthermore, while the lattice model of protein folding has been used previously in computational studies [6], [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 [1], [7], but not the one we discuss in this paper. In fact, there is no analytic expression for \( \mu \). We note that [7] discusses phase transitions of the free energy, 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 \( j \)th node in \( F \) by \( \pi_F(j) \). For any \( j, k \in \mathcal{N} \), we say that \( j \) and \( k \) are sequence-adjacent if \( |j - k| = 1 \) (here, \( \mathcal{N} = \{1, 2, \ldots, N\} \)). 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 \( \ell_1 \) norm). This allows us to define the energy \( E(F|s) \) as in (1).

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

\[
X_i = Q_{11}c_{HH}(i) + Q_{22}c_{PP}(i) + Q_{12}c_{HP}.
\]
where \( c_{xy}(i) \), discussed above, denotes the number of contacts \( \{i, j\} \) whose sequence elements are \( x \) and \( y \) or vice-versa. Then we have
\[
E(f|s) = \sum_{i=1}^{N} X_i(f|s).
\]

Clearly,
\[
E[E(f|S)] = \sum_{i} E[X_i(f|s)] = N\alpha + o(N)
\]
for some \( \alpha \geq 0 \) (with \( \alpha > 0 \) under mild conditions on \( Q \) and the sequence distribution), where boundary conditions contribute the \( o(N) \).

In this conference version of the paper we restrict our attention to a particular class of distributions on \( S_N \) 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 \( q = 1 - p \). We shall call this the Bernoulli product measure with parameter \( p \) and denote it by \( B_N(p) \). For such a model, we illustrate below how to compute \( \alpha \).

**Example. Computation of \( \alpha \).**

For the sake of simplicity, rather than giving the general formula, we now assume that
\[
Q = \begin{pmatrix} H & P \\ P & 1 - P \end{pmatrix}
\]

As mentioned earlier, \( N - o(N) \) nodes have exactly two neighbors, so calculating \( \alpha \) 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
\[
E[X_i] = 2pq.
\]

**A. Statement of Main Results**

We start with an expression for the conditional entropy. We have
\[
H(F|S) = \sum_{s \in S_N} p(s) \sum_{f \in F_S} p(f|s) \log p(f|s)
= E[\log Z(S, \beta)] + \beta \sum_{s, f} p(f, s) E(f|s)
= E[\log Z(S, \beta)] + \beta E[E(F|S)]
\]
where \( F_S \) 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
\[
\frac{H(F|S)}{N} = \frac{\log |F_N|}{N} \cdot \frac{\gamma(S, \beta)}{N} + \beta \frac{E[E(F|S)]}{N}.
\]

We now demonstrate that
\[
0 < \limsup_{N \to \infty} \frac{\log |F_N|}{N} = \mu < \infty.
\]

Let \( F_n \) denote the set of self-avoiding walks on the two-dimensional integer lattice (\( F_n \) is the unrestricted version of \( F_n \)). For any \( n, m > 0 \), we have that
\[
|F_n| / |F_m| = \frac{|F_n|}{|F_m|} \leq \frac{|F_n|}{|F_m|}
\]

since any fold \( f_n \in F_{n+m} \) can be constructed by choosing an initial part \( f_n \) of length \( n \) and a final part \( f_m \) of length \( m \). There are at most \( |F_n| \) ways of choosing the first part and at most \( |F_m| \) ways of choosing the second. Thus,
\[
\gamma \leq \gamma + \gamma \mu \leq \gamma + \gamma \mu
\]

(i.e., \( \gamma \) is subadditive). By Fekete’s lemma (see [8]), and since \( |F_n| \leq |F_m| \) for all \( n, m \),
\[
\limsup_{N \to \infty} \frac{\log |F_N|}{N} \leq \limsup_{N \to \infty} \frac{\log |F_N|}{N} < \infty.
\]

Clearly, \( \mu \geq 0 \), since \( |F_N| \geq 0 \). To demonstrate that this is a strict inequality, we sketch how to generate a number of walks that is exponential in \( N \), leaving the details to the journal version of this paper. For simplicity, we assume that \( |S_N| = 5 \), but the case where this does not hold is easily handled. We subdivide the \( \sqrt{N} \times \sqrt{N} \) square into subququares of size \( 5 \times 5 \). We will fill each subququare with a walk in such a way that the result is a valid walk for the entire square. In particular, the connections between the subququares will form a zig-zag pattern. This necessitates the construction of two types of 25-node walks: those which begin on one side of the subququare and exit on the opposite side, which we call forward walks; and those which enter on one side and exit on an orthogonal side, which we call turning walks. Furthermore, there exist at least two constructions for each type of walk, so that there are at least two valid ways to fill each subququare. Thus, there are at least \( 2^{N/25} \) walks of length \( N \) on a square with \( N \) nodes. Thus, we have shown that
\[
\mu > 0.
\]

We will prove in Section III-B that \( E[\log Z(S_N, \beta)] = O(N) \), so that the free energy is finite.

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

**Theorem 1.** For any distribution over \( S_N, \beta > 0 \), and scoring matrix \( Q \), the limit \( \alpha \) exists and is finite, the limit superior \( \mu \) is finite and positive, and
\[
\limsup_{N \to \infty} \frac{H(F|S)}{N} \leq \mu \cdot \gamma(\beta) + \beta \alpha.
\]

In particular, this holds for any distribution over \( S_N \) which minimizes \( H(F|S) \). Furthermore, for all but finitely many
Then we have the following two upper bounds:

- For all $\beta > 0$,
  \[
  \mu \cdot \gamma(\beta) \leq \mu - \beta \alpha + \frac{1}{2} \sigma^2 \beta^2.
  \]
- For $\beta = \beta_* = \frac{\sqrt{2 \mu}}{\sigma}$,
  \[
  \mu \cdot \gamma(\beta) \leq \beta \sqrt{2 \sigma^2 \mu} - \beta \alpha.
  \]

Then we have the following two upper bounds:

- For all $\beta > 0$,
  \[
  \mu \cdot \gamma(\beta) \leq \mu - \beta \alpha + \frac{1}{2} \sigma^2 \beta^2.
  \]
- For $\beta = \beta_* = \frac{\sqrt{2 \mu}}{\sigma}$,
  \[
  \mu \cdot \gamma(\beta) \leq \beta \sqrt{2 \sigma^2 \mu} - \beta \alpha.
  \]

Then we have the following two upper bounds:

- For all $\beta > 0$,
  \[
  \mu \cdot \gamma(\beta) \leq \mu - \beta \alpha + \frac{1}{2} \sigma^2 \beta^2.
  \]
- For $\beta = \beta_* = \frac{\sqrt{2 \mu}}{\sigma}$,
  \[
  \mu \cdot \gamma(\beta) \leq \beta \sqrt{2 \sigma^2 \mu} - \beta \alpha.
  \]

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 [10]).

### 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 [9]. 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 $\lim \inf_{N \to \infty} \frac{\operatorname{Var}[\mathcal{E}(f|S)]}{N} > 0$, 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 $\operatorname{Var}[\mathcal{E}(f|S)] = \Theta(N)$ (subject to the niceness condition on $Q$). We have

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

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 $\operatorname{Cov}[X_i, X_j] = 0$. 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

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

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 \( i \), \( \mathbb{E}[X_i^2] < \infty \). 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 [10]. 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 \( t \in \mathbb{R} \),

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

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

**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

\[
\int_{\theta \sqrt{N}}^{\infty} e^{\theta \sqrt{N}x} \, dF_n(x),
\]

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:

\[
\int_{-\theta \sqrt{N}}^{\theta \sqrt{N}} e^{\theta \sqrt{N}x} \, dF_n(x) = \Theta(1) \int_{-\theta \sqrt{N}}^{\theta \sqrt{N}} e^{\theta \sqrt{N}x} \, d\Phi(x)
\]

\[
\sim \Theta(1) \int_{-\infty}^{\infty} e^{\theta \sqrt{N}x} \, d\Phi(x)
\]

\[
= \Theta(1) e^{\frac{1}{2} \theta^2 \sigma^2 N}.
\]

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:

\[
\mathbb{E}[\log Z(S, \beta)] \leq \log \mathbb{E}[Z(S, \beta)]
\]

\[
= \log \sum_{f \in \mathcal{F}_N} e^{-\beta \mathbb{E}[\mathcal{E}(f|S)]] \mathbb{E} \left[ e^{-\beta \sqrt{N} \mathcal{E}(f|S)] - \mathbb{E}[\mathcal{E}(f|S))] \right]
\]

\[
= \log \sum_{f \in \mathcal{F}_N} e^{-\beta \mathbb{E}[\mathcal{E}(f|S)]] \mathbb{E} \left[ e^{-\beta \sqrt{N} \mathcal{E}(f|S)]} \right]
\]

\[
= \log \sum_{f \in \mathcal{F}_N} e^{-\beta \alpha N(1 + o(1)) \cdot e^{\frac{1}{2} \sigma^2 \beta^2 N(1 + o(1))}}
\]

\[
= N \frac{\log |\mathcal{F}_N|}{N} - \beta \alpha (1 + o(1)) + \frac{1}{2} \sigma^2 \beta^2 (1 + o(1)).
\]

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

\[
\limsup_{N \to \infty} \frac{\mathbb{E}[\log Z(S, \beta)]}{N} \leq \mu - \beta \alpha + \frac{1}{2} \sigma^2 \beta^2.
\]

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

We have

\[
- \beta \min_{f \in \mathcal{F}_N} \mathcal{E}(f|S) \leq \log \left( \sum_{f \in \mathcal{F}_N} e^{-\beta \mathcal{E}(f|S)]} \right)
\]

\[
\implies \limsup_{N \to \infty} \frac{\mathbb{E}[\min_{f \in \mathcal{F}_N} \mathcal{E}(f|S)]}{N} \leq \beta^{-1} \mu - \alpha + \frac{1}{2} \beta \sigma^2,
\]

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

\[ \phi'(\beta) = \mathbb{E} \left[ -\sum_{f \in \mathcal{F}_N} \mathcal{E}(f|S)e^{-\beta \mathcal{E}(f|S)} \right] \]

\[ \leq \mathbb{E} \left[ -\min_{f \in \mathcal{F}_N} \mathcal{E}(f|S) \right] \frac{Z(S, \beta)}{Z(S, \beta^*)} \]

\[ \leq N(\beta^{-1}\mu - \alpha + \frac{1}{2}\beta\sigma^2) \]

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 H98230-11-1-0141, and NSF Grants DMS-0800568, and CCF-0830140. W. Szpankowski is also a Visiting Professor at ETI, Gdańsk University of Technology, Poland.

**REFERENCES**


