Phase Transitions in a Sequence-Structure Channel

A. Magner and D. Kihara and W. Szpankowski
Purdue University
W. Lafayette, IN 47907

January 29, 2015

Center for
Science of Information
NSF Science & Technology Center

ITA, San Diego, 2015

Structural Information

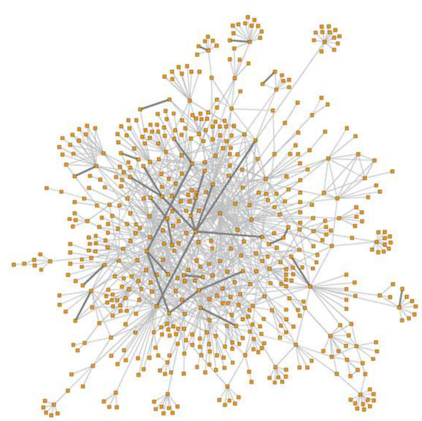
Information Theory of Data Structures: Following Ziv (1997) we propose to explore finite size **information theory** of data structures (i.e., sequences, sets, trees, graphs), that is, to develop information theory of various data structures beyond first-order asymptotics.

F. Brooks, jr. (JACM, 50, 2003, "Three Great Challenges for . . . CS"):
"We have **no theory** that gives us a metric for the Information embodied in **structure**. This is the most fundamental gap in the theoretical underpinnings of information science and of computer science."

Networks (Internet, protein-protein interactions, and collaboration network) and Matter (chemicals and proteins) have structures.

But one may also interested in structural properties of systems with local dependencies or interactions represented by Markov fields.

Another problem: flow of structural information over a noisy channel.



Sequence-Structure Channel

$$P(f|s) := \frac{e^{-\beta \mathcal{E}(s,f)}}{Z(s,\beta)} \qquad\qquad Z(s,\beta) := \sum_{f \in \mathcal{F}} e^{-\beta \mathcal{E}(s,f)}$$

Sequence-Structure Channel

$$P(f|s) := \frac{e^{-\beta\mathcal{E}(s,f)}}{Z(s,\beta)} \qquad Z(s,\beta) := \sum_{f\in\mathcal{F}} e^{-\beta\mathcal{E}(s,f)}$$

Sequences: $S = (S_1, \ldots S_N)$, i.i.d. with $P(S_i = H) = p = 1 - P(S_i = P)$.

β: a parameter that is meant to represent inverse temperature.

Folds: $\mathcal{F}_N$ denotes the set of self-avoiding walks of length N filling a square in $\mathbf{Z}^2$ of size N, starting at $(0,0)$ and ending at $(\sqrt{N}-1, \sqrt{N}-1)$.

Energy: $\mathcal{E}(s,f)$ denotes energy for a fold f computed as follows: for a given symmetric 2×2 scoring matrix $Q = \{Q_{ij}\}_{i,j\in\{1,2\}}$ define

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

where c_{xy} denotes the number of (non-adjacent) contacts in a fold.

Information Theoretic Quantities

Capacity :

$$C = \max_{P(S)} I(S;F) = \max_{P(S)} [H(F) - H(F|S)]$$

where

Conditional Entropy:

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

Information Theoretic Quantities

Capacity :

$$C = \max_{P(S)} I(S;F) = \max_{P(S)}[H(F) - H(F|S)]$$

where

Conditional Entropy:

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

Observe that

$$\mathbf{E}[\mathcal{E}(F,S)] = N \cdot \alpha$$

for some α (that can be computed).

Example: Consider

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

We find

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

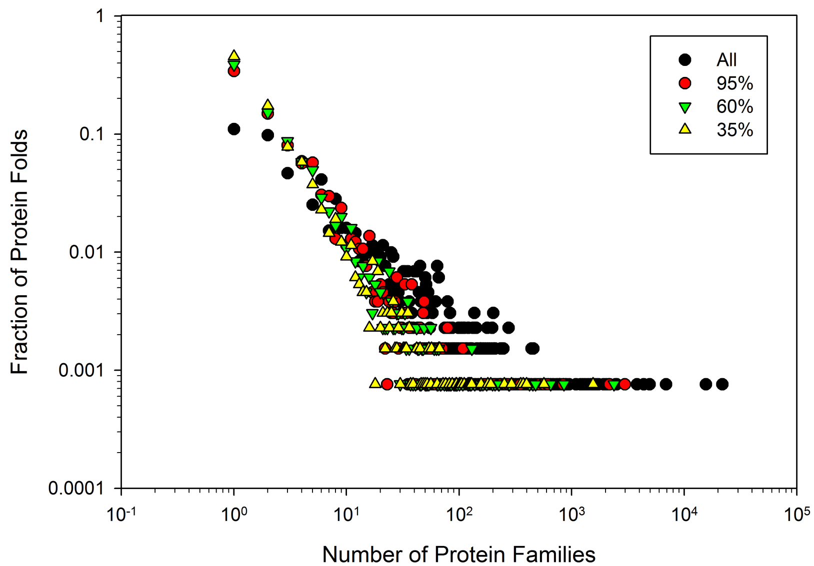
Question: What can we say about $\mathbf{E}[\log Z(S,\beta)]$?

Why to Bother?

Mathematical/Information-theoretic motivation:

- Maps sequences to structures.
- A channel with full memory.
- Several information theoretic quantities of interest exhibit unusual phase transitions with respect to temperature (=$1/\beta$).
- Capacity of the protein folding channel is conjectured to have a phase transition with repect to β.
- Probability: A nontrivial dependence structure between fold energies makes lower bounding the partition function challenging.
- Combinatorics: Quantities of interest depend crucially on the cardinality of the number of folds or number of self-avoiding walks (**open problem**). Does the limit
$$\lim_{N\to\infty} \frac{\log|\mathcal{F}_N|}{N}$$
exist? In general, what is an asymptotic behavior of $|\mathcal{F}_N|$?

Biological Motivation



Protein Folds in Nature

For each possible cardinality of protein families (x axis), count the number of protein folds (or sequences) observed in nature which are associated with that number of families. Plot on y axis the fraction of protein folds.

In nature, we observe lots of sequences with few associated folds and few sequences with lots of associated folds.

Physical/Biological motivation:

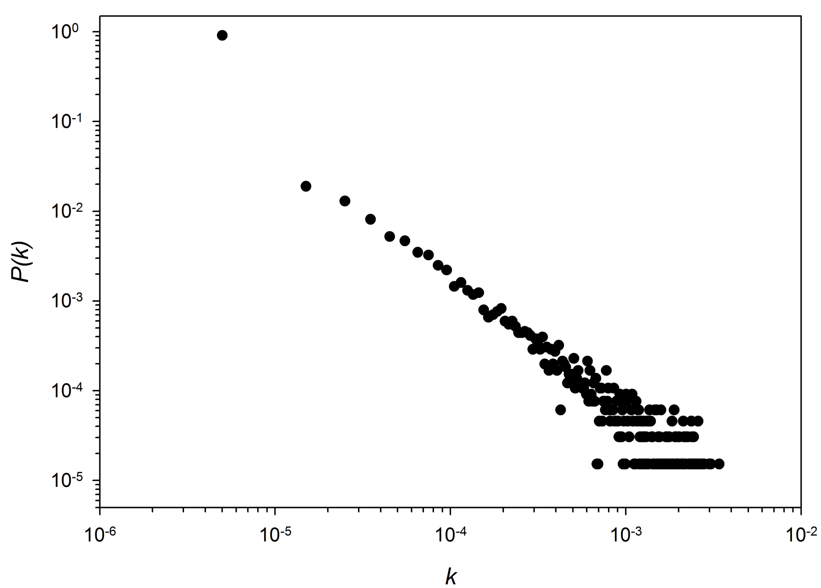
- The channel is a model of protein folding.
- Sequence distribution in nature exhibits a power law. In the channel model, such distributions (empirically) almost achieve capacity (nature prefers to avoid ambiguity!): capacity may have biological significance.

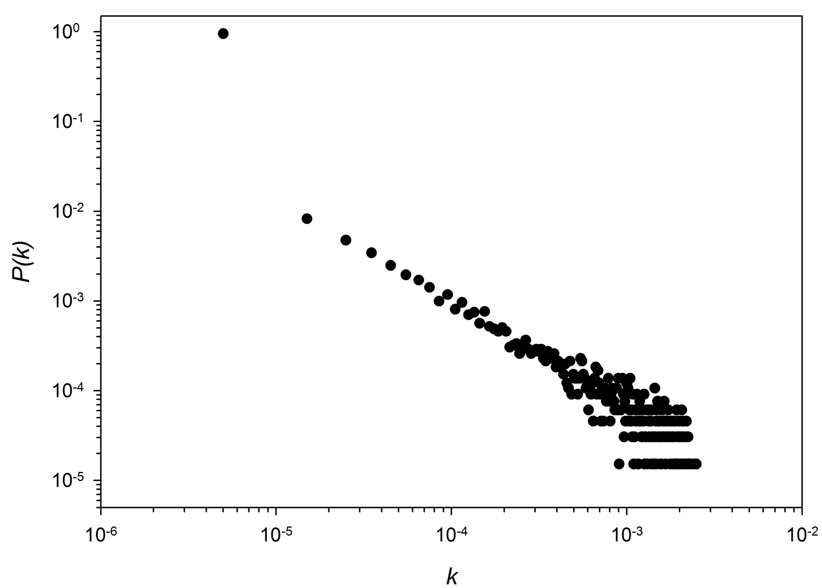
Information Theoretic Approach & Experimental Results

1. Compute the capacity achieving input distribution from Blahut-Arimoto algorithm (for lattices of size $5, 6$).

2. Partition the set of all sequences according to probability that this distribution assigns to them (sequence types).

3. Plot the sequence distribution (x axis) versus fraction (y axis) of all sequences that got that distribution. Here what we see:





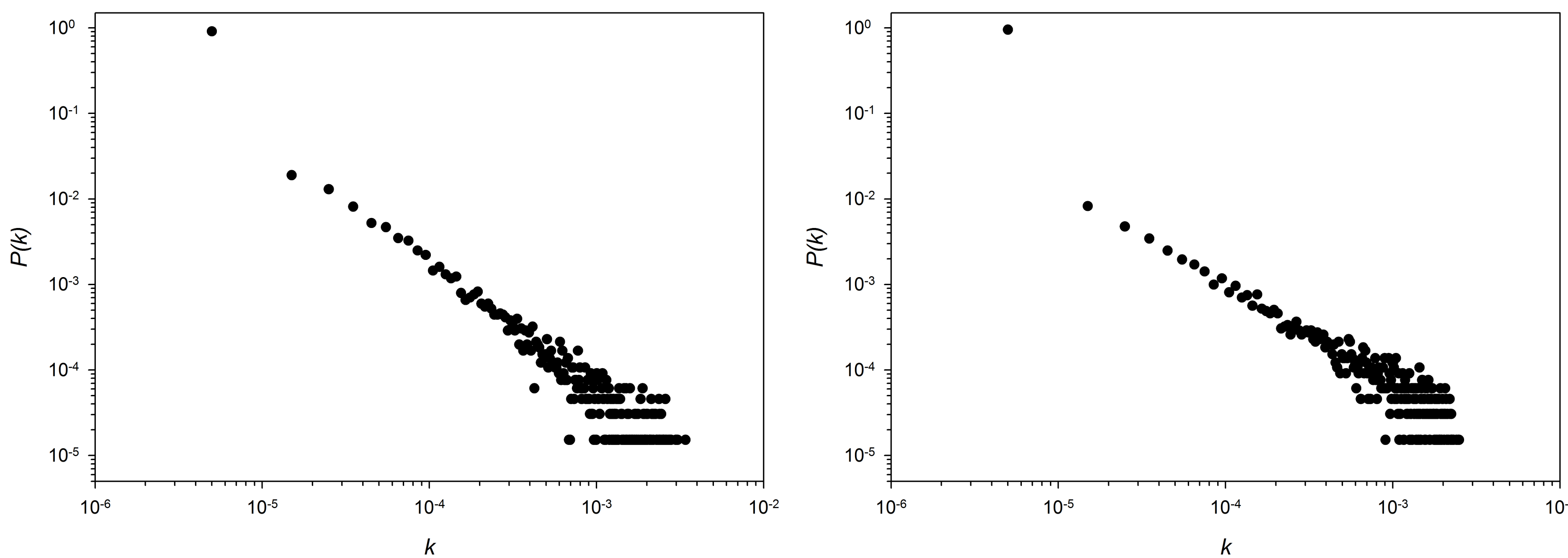
We have a power law as in nature!

Information Theoretic Approach & Experimental Results

1. Compute the capacity achieving input distribution from Blahut-Arimoto algorithm (for lattices of size $5, 6$).

2. Partition the set of all sequences according to probability that this distribution assigns to them (sequence types).

3. Plot the sequence distribution (x axis) versus fraction (y axis) of all sequences that got that distribution. Here what we see:



We have a power law as in nature!

Magner, Szpankowski, Kihara, "On the Origin of Protein Superfamilies and Superfolds", Scientific Reports, 2015.

Back to Theory ...

Recall that

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

Define free energy $\gamma(\beta, S)$ as

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

By submultiplicativity property of F_N we conclude

$$\lim_{N\to\infty} \frac{\log |F_N|}{N} = \log \mu.$$

for $\mu > 1$.

Then

$$\mathbf{E} \log Z(S,\beta) \sim \log |\mathcal{F}_N| \cdot \gamma(\beta, S) \sim N \log \mu \cdot \gamma(\beta, S)$$

leading to

$$\boxed{H(F|S) \sim N[\gamma(\beta, S) \log \mu + \beta\alpha]}$$

Main Results

Theorem 1. *For any distribution over $\mathcal{S}_N$, $\beta > 0$, and scoring matrix Q we have*

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

Furthermore, if Q comes from a certain broad class of scoring matrices (satisfying a "niceness" condition), there exists $\sigma^2 > 0$ such that

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

Then we have the following **phase transition***:*

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

The conditional entropy phase transition is a consequence of the free energy phase transition:

$$\log\mu \cdot \gamma(\beta, S) \leq \begin{cases} \log\mu - \beta\alpha + \frac{1}{2}\sigma^2\beta^2 & \beta < \frac{\sqrt{2\log\mu}}{\sigma} \\ \beta(\sqrt{2\sigma^2\log\mu} - \alpha) & \beta \geq \frac{\sqrt{2\log\mu}}{\sigma} \end{cases}$$

Current/Future Work

Lower bound: We conjecture that a matching lower bound on the free energy holds, which would give

$$\log \mu \cdot \gamma(\beta, S) = \begin{cases} \log \mu - \beta\alpha + \frac{1}{2}\sigma^2\beta^2 & \beta < \frac{\sqrt{2\log\mu}}{\sigma} \\ \beta\sqrt{2\sigma^2 \log \mu} - \beta\alpha & \beta \geq \frac{\sqrt{2\log\mu}}{\sigma} \end{cases}$$

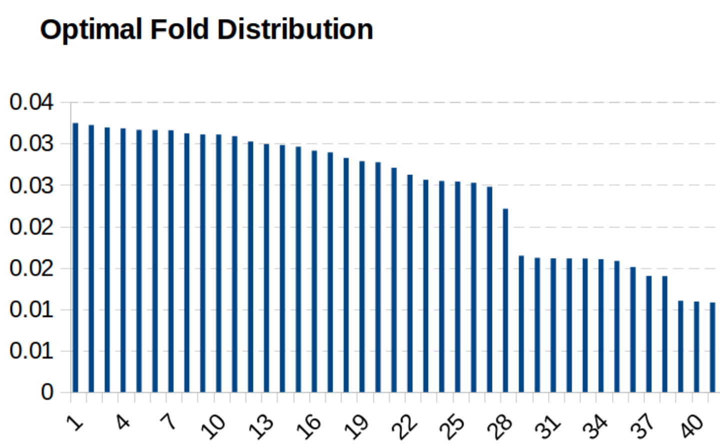
The lower bounds requires to understand dependencies between folds.

More general source models: We considered here sequences generated by a memoryless source, but more general models (e.g., Markov, mixing) are more realistic and probably mathematically tractable.

Extensions to k-dimensional self-avoiding walks and other structures are possible.

Capacity Conjecture

The optimal output distribution observed in experiments seems to be uniform, as shown below:



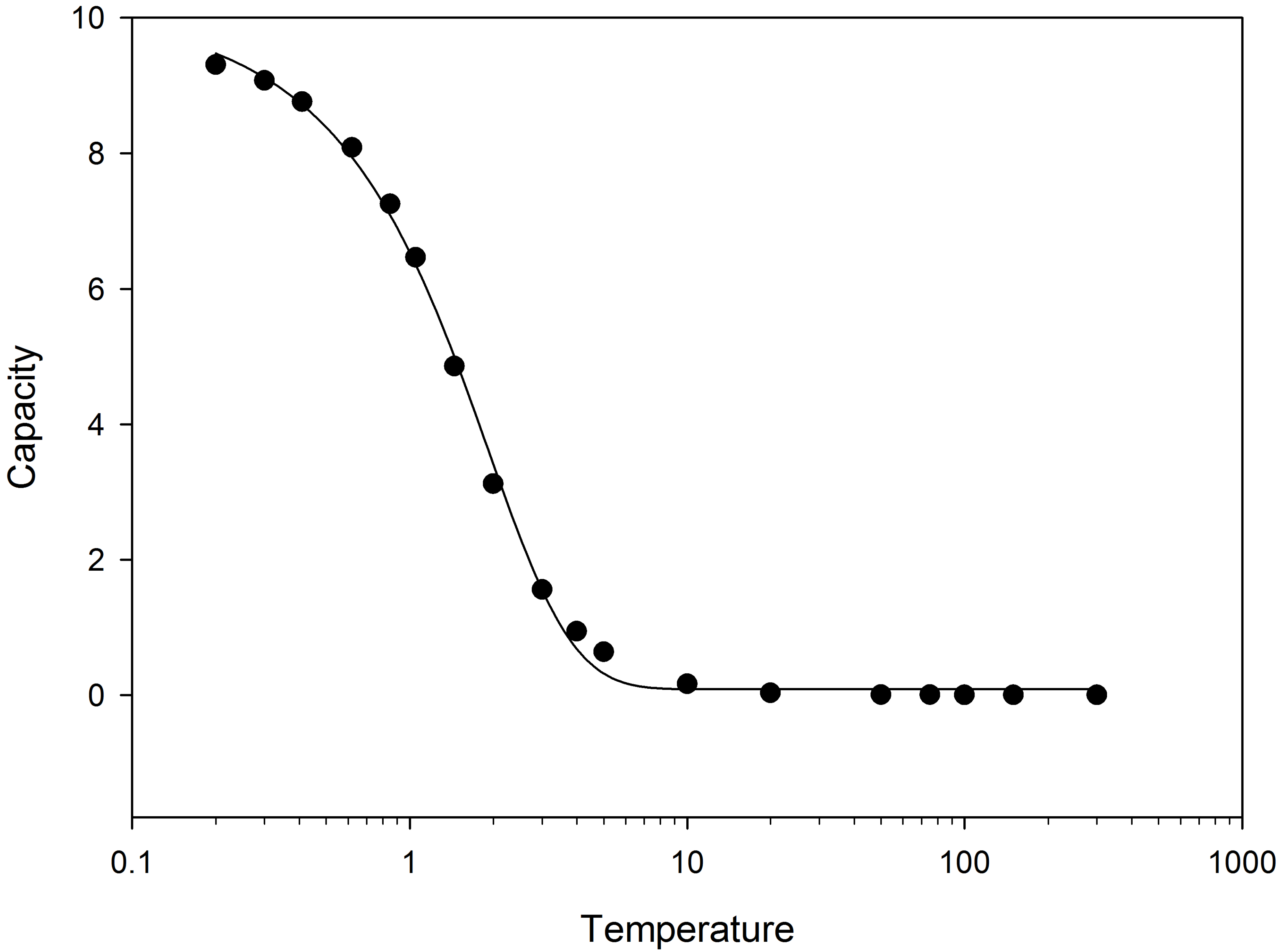
Conjecture. We conjecture that

$$C \sim \log |\mathcal{F}_N| - \min_{P(S)} H(F|S)$$

where $H(F|S)$ we just computed. The minimization over p (for the memoryless case) is easy to perform.

Thus one should expect a phase transition, with respect to β, of the capacity. Experiments do confirm it.

Experimental Confirmation of Phase Transition in the Capacity



Upper Bound Proof Sketch

1. First upper bound:

$$\mathbf{E}[\log Z(S,\beta)] \leq \log \mathbf{E}[Z(S,\beta)] = \log \sum_{f \in \mathcal{F}_N} \mathbf{E}[e^{-\beta \mathcal{E}(f|S)}],$$

because $Z(S,\beta)$ is a convex function.

Upper Bound Proof Sketch

1. First upper bound:

$$\mathbf{E}[\log Z(S,\beta)] \leq \log \mathbf{E}[Z(S,\beta)] = \log \sum_{f\in\mathcal{F}_N} \mathbf{E}[e^{-\beta\mathcal{E}(f|S)}],$$

because $Z(S,\beta)$ is a convex function.

2. Denote by $F_N(x)$ the CDF of

$$\hat{\mathcal{E}}(f|S) = \frac{(\mathcal{E}(f|S) - \mathbf{E}[\mathcal{E}(f|S)])}{\sqrt{N}}.$$

Let $\Phi(x)$ be the CDF of $\mathcal{N}(0,\sigma^2)$. Then it can be proved

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

by results on m-dependent random fields, that is, $\hat{\mathcal{E}}(f|S) \sim N(0,\sigma^2)$.

Upper Bound Proof Sketch

1. First upper bound:

$$\mathbf{E}[\log Z(S,\beta)] \leq \log \mathbf{E}[Z(S,\beta)] = \log \sum_{f \in \mathcal{F}_N} \mathbf{E}[e^{-\beta \mathcal{E}(f|S)}],$$

because $Z(S,\beta)$ is a convex function.

2. Denote by $F_N(x)$ the CDF of

$$\hat{\mathcal{E}}(f|S) = \frac{(\mathcal{E}(f|S) - \mathbf{E}[\mathcal{E}(f|S)])}{\sqrt{N}}.$$

Let $\Phi(x)$ be the CDF of $\mathcal{N}(0,\sigma^2)$. Then it can be proved

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

by results on m-dependent random fields, that is, $\hat{\mathcal{E}}(f|S) \sim N(0,\sigma^2)$.

3. Each energy is a sum of local energies: denoting by $X_i(f|S)$ the contact energy of the ith residue,

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

and each residue has a contact with at most 3 others (so each term of the sum is dependent on at most 3 others).

Upper Bound Proof Sketch (continued)

4. Let $\varphi_N(x) = \mathbf{E}[e^{x\hat{\mathcal{E}}(f|S)}]$ and $\varphi(x) = \mathbf{E}[e^{xN(0,\sigma^2)}] = \exp(\frac{1}{2}x^2\sigma^2)$.
Then large deviations via martingale inequalities, integration by parts of the fold MGF integral, and fold energy CLT give

$$\lim_{N\to\infty} \frac{\log \varphi_N(t\sqrt{N})}{N} = \log \varphi(t) = \frac{1}{2}\sigma^2 t^2,$$

so that we conclude

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

which leads to the first upper bound.

Second Upper Bound

5. To derive the second upper bound, we observe

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

leading to

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

which is minimized at $\beta = \beta_* = \frac{\sqrt{2\mu}}{\sigma}$. Hence we find

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

Second Upper Bound: Continuation

6. Let $\psi(\beta) = \mathbf{E}[\log Z(S,\beta)]$. By concavity for $\beta > \beta_*$, we have

$$\psi(\beta) \leq \psi(\beta_*) + \psi'(\beta_*)(\beta - \beta_*).$$

where

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

Applying the upper bound on $\psi(\beta_*)$ gives the second upper bound.

Lower Bound Intuition

CLT for fold energies suggests that our problem looks somewhat like Derrida's *Random Energy Model* (configuration energies are i.i.d. standard Gaussians). How far does this go?

For two folds f and g,

$$\mathrm{Cov}[\mathcal{E}(f|S), \mathcal{E}(g|S)] = O(\sqrt{N}) = o(\mathbf{E}[\mathcal{E}(f|S)]),$$

by limited dependence structure of local energies. So $\mathcal{E}(f|S)$ and $\mathcal{E}(g|S)$ are asymptotically not too correlated.

We can apply the Crámer-Wold theorem to show that

$$(\hat{\mathcal{E}}(f|S), \hat{\mathcal{E}}(g|S)) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \mathbb{I}_2).$$

So negligible correlation implies negligible dependence.

Conclusion: Our model looks a lot like the Random Energy Model. Adapt the lower bound proof technique in that case.

That's It

THANK YOU

New Book

How do you distinguish a cat from a dog by their DNA? Did Shakespeare really write all of his plays?

Pattern matching techniques can offer answers to these questions and to many others, from molecular biology, to telecommunications, to classifying Twitter content.

This book for researchers and graduate students demonstrates the probabilistic approach to pattern matching, which predicts the performance of pattern matching algorithms with very high precision using analytic combinatorics and analytic information theory. Part I compiles known results of pattern matching problems via analytic methods. Part II focuses on applications to various data structures on words, such as digital trees, suffix trees, string complexity and string-based data compression. The authors use results and techniques from Part I and also introduce new methodology such as the Mellin transform and analytic depoissonization.

More than 100 end-of-chapter problems help the reader to make the link between theory and practice.

Philippe Jacquet is a research director at INRIA, a major public research lab in Computer Science in France. He has been a major contributor to the Internet OLSR protocol for mobile networks. His research interests involve information theory, probability theory, quantum telecommunication, protocol design, performance evaluation and optimization, and the analysis of algorithms. Since 2012 he has been with Alcatel-Lucent Bell Labs as head of the department of Mathematics of Dynamic Networks and Information. Jacquet is a member of the prestigious French Corps des Mines, known for excellence in French industry, with the rank of "Ingenieur General". He is also a member of ACM and IEEE.

Wojciech Szpankowski is Saul Rosen Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University, where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information. Szpankowski is a Fellow of IEEE and an Erskine Fellow. He received the Humboldt Research Award in 2010.

Cover design: Andrew Ward

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-0-521-87608-7

9 780521 876087 >

9780521876087: Jacquet & Szpankowski: PPC: C M Y K

Jacquet and Szpankowski Analytic Pattern Matching CAMBRIDGE

Philippe Jacquet and
Wojciech Szpankowski

Analytic Pattern Matching

From DNA to Twitter