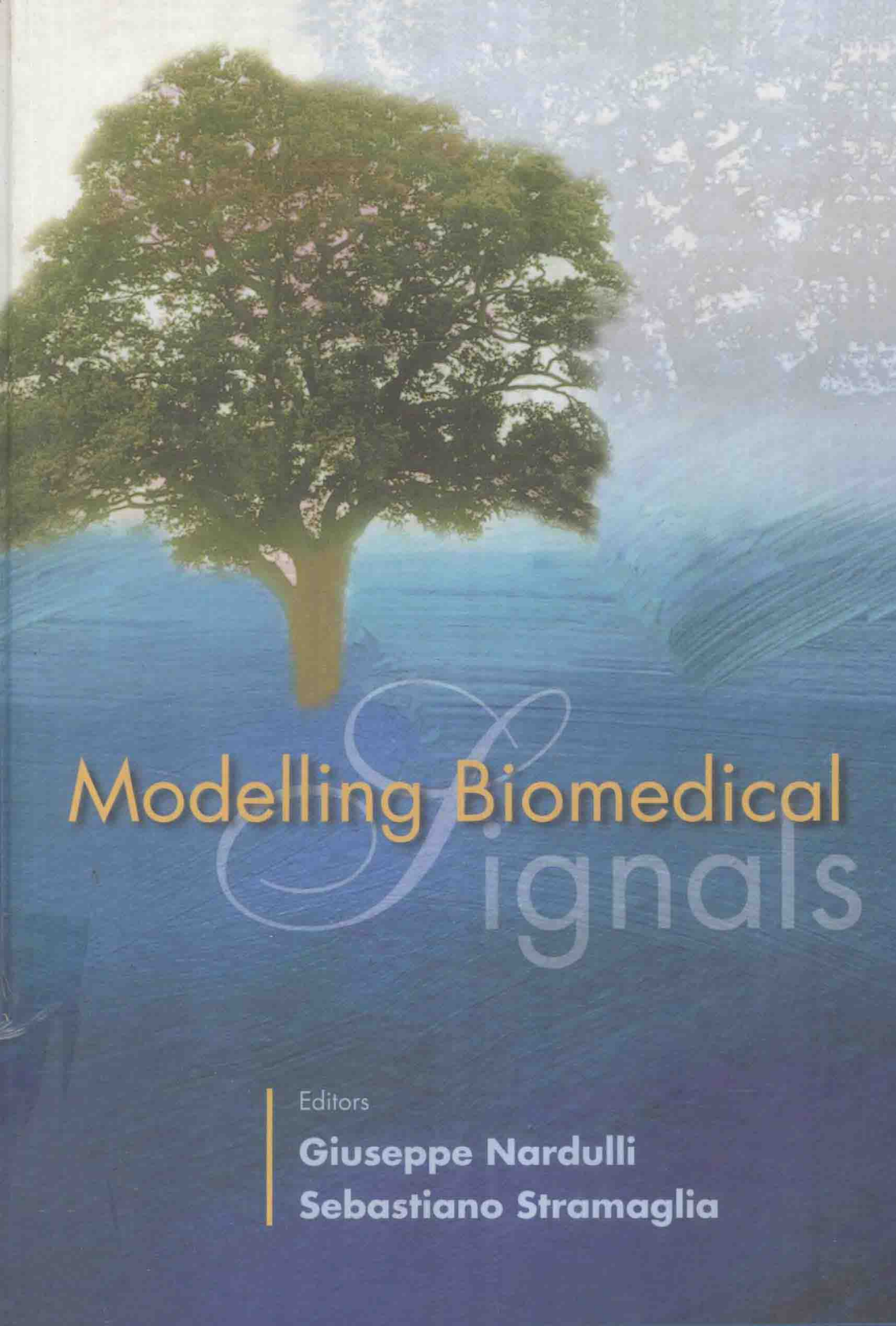
Modelling Biomedical Signals
Editors
Giuseppe Nardulli
Sebastiano Stramaglia

Modelling Biomedical Signals

Bari, Italy 19–21 September 2001

Editors

Giuseppe Nardulli

Sebastiano Stramaglia

Center of Innovative Technologies for Signal Detection and Processing
University of Bari, Italy

World Scientific
New Jersey • London • Singapore • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 912805

USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

MODELLING BIOMEDICAL SIGNALS

ISBN 981-02-4843-1

Printed in Singapore by Mainland Press

Preface

In the last few years, concepts and methodologies initially developed in theoretical physics have found high applicability in a number of very different areas. This book, a result of cross-disciplinary interaction among physicists, biologists and physicians, covers several topics where methods and approaches rooted in physics are successfully applied to analyze and to model biomedical data. The volume contains the papers presented at the International Workshop *Modelling Bio-medical Signals* held at the Physics Department of the University of Bari, Italy, on September 19-21th 2001. The workshop was gathered under the auspices of the Center of Innovative Technologies for Signal Detection and Processing of the University of Bari (TIRES Centre); the Organizing Committee of the Workshop comprised L. Angelini, R. Bellotti, A. Federici, R. Giuliani. G. Gonnella, G.Nardulli and S. Stramaglia. The workshop opened on September 19th 2001 with two colloquia given by profs. N. Accornero (University of Rome, la Sapienza), on Neural Networks and Neurosciences, and E. Marinari (University of Rome, la Sapienza) on Physics and Biology. Around 70 scientists attended the workshop, coming from different fields and disciplines. The large spectrum of competences gathered in the workshop favored an intense and fruitful exchange of scientific information and ideas. The topics discussed in the workshop include: decision support systems in medical science; several analyses of physiological rhythms and synchronization phenomena; biological neural networks; theoretical aspects of artificial neural networks and their role in neural sciences and in the analysis of EEG and Magnetic Resonance Imaging; gene expression patterns; the immune system; protein folding and protein crystallography.

For the organization of the workshop and the publication of the present volume we acknowledge financial support from the Italian Ministry of University and Scientific Research (MURST) under the project (PRIN) "Theoretical Physics pf Fundamental Interactions", from the TIRES Centre, the Physics Department of the University of Bari and from the Section of Bari of the Istituto Nazionale di Fisica Nucleare (INFN). We also thank the Secretary of the Workshop, Mrs. Fausta Cannillo and Mrs. Rosa Bitetti for their help in organizing the event.

Giuseppe Nardulli
Sebastiano Stramaglia
University of Bari

CONTENTS

---

* Italicized name indicates the author who presented the paper.

GENOMICS AND MOLECULAR BIOLOGY

ANALYSIS AND MODELS OF BIOMEDICAL DATA BY THEORETICAL PHYSICS METHODS

THE CLUSTER VARIATION METHOD FOR APPROXIMATE REASONING IN MEDICAL DIAGNOSIS

H.J. KAPPEN
Laboratory of Biophysics, University of Nijmegen,
bert@mbfys.kun.nl

In this paper, we discuss the rule based and probabilistic approaches to computer aided medical diagnosis. We conclude that the probabilistic approach is superior to the rule based approach, but due to its intractability, it requires approximations for large scale applications. Subsequently, we review the Cluster Variation Method and derive a message passing scheme that is efficient for large directed and undirected graphical models. When the method converges, it gives close to optimal results.

1 Introduction

Medical diagnosis is the a process, by which a doctor searches for the cause (disease) that best explains the symptoms of a patient. The search process is sequential, in the sense that patient symptoms suggest some initial tests to be performed. Based on the outcome of these tests, a tentative hypothesis is formulated about the possible cause(s). Based on this hypothesis, subsequent tests are ordered to confirm or reject this hypothesis. The process may proceed in several iterations until the patient is finally diagnosed with sufficient certainty and the cause of the symptoms is established.

A significant part of the diagnostic process is standardized in the form of protocols. These are sets of rules that prescribe which tests to perform and in which order, based on the patient symptoms and previous test results. These rules form a decision tree, whose nodes are intermediate stages in the diagnostic process and whose branches point to additional testing, depending on the current test results. The protocols are defined in each country by a committee of medical experts.

The use of computer programs to aid in the diagnostic process has been a long term goal of research in artificial intelligence. Arguably, it is the most typical application of artificial intelligence.

The different systems that have been developed so-far use a variety of modeling approaches which can be roughly divided into two categories: rule-based approaches with or without uncertainty and probabilistic methods. The rule-based systems can be viewed as computer implementations of the protocols, as described above. They consist of a large data base of rules of the form: $A \rightarrow B$, meaning that "if condition A is true, then perform action B"

or "if condition A is true, then condition B is also true". The rules may be deterministic, in which case they are always true, or 'fuzzy' in which case they are true to a (numerically specified) degree. Examples of such programs are Meditel[1], Quick Medical Reference (QMR)[2], DXplain[3], and Iliad[4].

In Berner et al.[5] a detailed study was reported that assesses the performance of these systems. A panel of medical experts collected 110 patient cases, and concensus was reached on the correct diagnosis for each of these patients. For each disease, there typically exists a highly specific test that will unambiguously identify the disease. Therefore, based on such complete data, diagnosis is easy. A more challenging task was defined by removing this defining test from each of the patient cases. The patient cases were presented to the above 4 systems. Each system generated its own ordered list of most likely diseases. In only 10-20 % of the cases, the correct diagnosis appeared on the top of these lists and in approximately 50 % of the cases the correct diagnosis appeared in the top 20 list. Many diagnoses that appeared in the top 20 list were considered irrelevant by the experts. It was concluded that these systems are not suitable for use in clinical practice.

There are two reasons for the poor performance of the rule based systems. One is that the rules that need to be implemented are very complex in the sense that the precondition A above is a conjunction of many factors. If each of these factors can be true or false, there is a combinatoric explosion of conditions that need to be described. It is difficult, if not impossible, to correctly describe all these conditions. The second reason is that evidence is often not deterministic (true or false) but rather probabilistic (likely or unlikely). The above systems provide no principled approach for the combination of such uncertain sources of information.

A very different approach is to use probability theory. In this case, one does not model the decision tree directly, but instead models the relations between diseases and symptoms in one large probability model. As a (too) simplified example, consider a medical domain with a number of diseases $d = (d_1, \ldots, d_n)$ and a number of symptoms or findings $f = (f_1, \ldots, f_m)$. One estimates the probability of each of the diseases $p(d_i)$ as well as the probability of each of the findings *given* a disease, $p(f_j|d_i)$. If diseases are independent, and if findings are conditionally independent given the disease, the joint probability model is given by:

$$p(d, f) = p(d)p(f|d) = \prod_i p(d_i) \prod_j p(f_j|d_i) \tag{1}$$

It is now possible to compute the probability of a disease d_i, given some

findings by using Bayes' rule:

$$p(d_i|f_t) = \frac{p(d_i, f_t)}{p(f_t)}, \tag{2}$$

where f_t is the list of findings that has been measured up to diagnostic iteration t. Computing this for different d_i gives the list of most probable diseases given the current findings f_t and provides the tentative diagnosis of the patient. Furthermore, one can compute which additional test is expected to be most informative about any one of the diagnoses, say d_i, by computing

$$I_{ij} = -\sum_{f_j} p(f_j|f_t) \sum_{d_i} p(d_i|f_t, f_j) \log p(d_i|f_t, f_j)$$

for each test j that has not been measured so-far. The test j that minimizes I_{ij} is the most informative test, since averaged over its possible outcomes, it gives the distribution over d_i with the lowest entropy.

Thus, one sees that whereas the rule based systems model the diagnostic process directly, the probabilistic approach models the relations between diseases and findings. The diagnostic decisions (which test to measure next) is then computed from this model. The advantage of this latter approach is that the model is much more transparent about the medical knowledge, which facilitates maintenance (changing probability tables, adding diseases or findings), as well as evaluation by external experts.

One of the main drawbacks of the probabilistic approach is that it is intractable for large systems. The computation of marginal probabilities requires summation over all other variables. For instance, in Eq. 2

$$p(f_t) = \sum_{d,f} \delta_{f,f_t} p(d, f)$$

and the sum over d, f contains exponentially many terms. Therefore, probabilistic models for medical diagnosis have been restricted to very small domains[6,7] or when covering a large domain, at the expense of the level of detail at which the disease areas are modeled[8].

In order to make the probabilistic approach feasible for large applications one therefore needs to make approximations. One can use Monte Carlo sampling but one finds that accurate results require very many iterations. An alternative is to use analytical approximations such as for instance mean field theory[9,10]. This approach works well for probability distributions that resemble spin systems (so-called Boltzmann Machines) but, as we will see, they perform poorly for directed probability distributions of the form Eq. 1.

2 The Cluster Variation Method

A very recent development is the application of the Cluster Variation method (CVM) to probabilistic inference. CVM is a method that has been developed in the physics community to approximately compute the properties of the Ising model[11]. The CVM approximates the probability distribution by a number of (overlapping) marginal distributions (clusters). The quality of the approximation is determined by the size and number of clusters. When the clusters consist of only two variables, the method is known as the Bethe approximation. Recently, the method has been introduced by Yedidia et al.[12] into the machine learning community, showing that in the Bethe approximation, the CVM solution coincides with the fixed points of the belief propagation algorithm. Belief propagation is a message passing scheme, which is known to yield exact inference in tree structured graphical models[13]. However, BP can can also give impressive results for graphs that are not trees[14].

Let $x = (x_1, \ldots, x_n)$ be a set of variables, where each x_i can take a finite number of values. Consider a probability distribution on x of the form

$$p_H(x) = \frac{1}{Z(H)} e^{-H(x)} \qquad Z = \sum_x e^{-H(x)}$$

It is well known, that p_H can be obtained as the minimum of the free energy, which is a functional over probability distributions of the following form:

$$F_H(p) = \langle H \rangle + \langle \log p \rangle \,, \tag{3}$$

where the expectation value is taken with respect to the distribution p, i.e. $\langle H \rangle = \sum_x p(x) H(x)$. When one minimizes $F_H(p)$ with respect to p under the constraint of normalization $\sum_x p(x) = 1$, one obtains p_H [a].

Computing marginals of p_H such as $p_H(x_i)$ or $p_H(x_i, x_j)$ involves sums over all states, which is intractable for large n. Therefore, one needs tractable approximations to p_H. The cluster variation method replaces the probability distribution $p_H(x)$ by a large number of (possibly overlapping) probability distributions, each describing the interaction between a small number of variables. Due to the one-to-one correspondence between a probability distribution and the minima of a free energy we can define approximate probability distributions by constructing approximate free energies and computing their minimum (or minima!). This is achieved by approximating Eq. 3 in terms of the cluster probabilities. The solution is obtained by minimizing this approximate free energy subject to normalization and consistency constraints.

[a] Minimizing the free energy can also be viewed as maximizing the entropy with an additional constraint on $\langle H \rangle$.

Define clusters as subsets of distinct variables: $x_\alpha = (x_{i_1}, \ldots, x_{i_k})$, with $1 \leq i_j \leq n$. Define a set of clusters P that contain the interactions in H and write H as a sum of these interactions:

$$H(x) = \sum_{\alpha \in P} H^\dagger_\alpha(x_\alpha)$$

For instance for Boltzmann-Gibbs distributions, $H(x) = \sum_{i>j} w_{ij} x_i x_j + \sum_i \theta_i x_i$ and P consists of all pairs and all singletons: $P = \{\alpha | \alpha = (ij), i > j \text{ or } \alpha = (i)\}$. For directed graphical models with evidence, such as Eq. 2, P is the set of clusters formed by each node i and its parent set π_i: $P = \{\alpha | \alpha = (i, \pi_i), i = 1, \ldots, n\}$. x is the set of non-evidence variables (d in this case) and $Z = p(f_t)$.

We now define a set of clusters B, that will determine our approximation in the cluster variation method. B should at least contain the interactions in $p(x)$ in the following way:

$$\forall \alpha \in P \Rightarrow \exists \alpha' \in B, \alpha \subset \alpha'.$$

In addition, we demand that no two clusters in B contain each other: $\alpha, \alpha' \in B \Rightarrow \alpha \not\subset \alpha', \alpha' \not\subset \alpha$. Clearly, the minimal choice for B is to chose clusters from P itself. The maximal choice for B is the cliques obtained when constructing the junction tree[15]. In this case, the clusters in B form a tree structure and the CVM method is exact. In general, one, can chose any set of clusters B that satisfy the above definition. Since the proposed method scales exponentially in the size of the clusters in B, the smaller the clusters in B, the faster the approximation. For a simple directed graphical model an intermediate choice of clusters is illustrated in Fig. 1.

Define a set of clusters M that consist of any intersection of a number of clusters of B: $M = \{\beta | \beta = \cap_k \alpha_k, \alpha_k \in B\}$, and define $U = B \cup M$. Once U is given, we define numbers a_β recursively by the Moebius formula

$$1 = \sum_{\alpha \in U, \alpha \supset \beta} a_\alpha, \quad \forall \beta \in U$$

In particular, this shows that $a_\alpha = 1$ for $\alpha \in B$.

The Moebius formula allows us to rewrite interactions on potentials in P in terms of interactions on clusters in U:

$$H(x) = \sum_{\beta \in P} H^\dagger_\beta(x_\beta) = \sum_{\beta \in P} \sum_{\alpha \in U, \alpha \supset \beta} a_\alpha H^\dagger_\beta(x_\beta) = \sum_{\alpha \in U} a_\alpha H_\alpha,$$

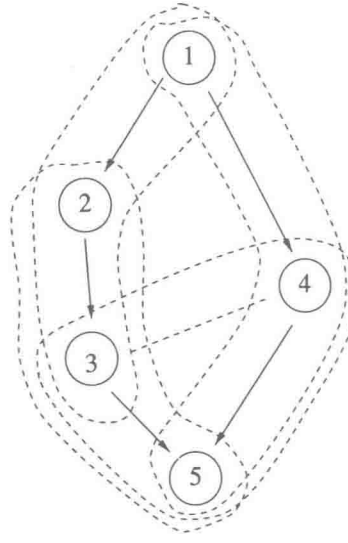

Figure 1. Directed graphical model consisting of 5 variables. Interactions are defined on clusters in $P = \{(1), (1,2), (2,3), (1,4), (3,4,5)\}$. The clusters in B are depicted by the dashed lines ($B = \{(1,2,3), (2,3,5), (1,4,5), (3,4,5)\}$. The set $M = \{(1), (2,3), (3), (5), (3,5)\}$

where we have defined H_α as the sum of all interactions in $\beta \in P$ that are contained in cluster $\alpha \in U$:

$$H_\alpha(x_\alpha) = \sum_{\beta \in P, \beta \subset \alpha} H_\beta^\dagger(x_\beta)$$

Since interactions may appear in multiple clusters, the constants a_α ensure that double counting is compensated for. [b] Thus, we can express $\langle H \rangle$ in Eq. 3 explicitly in terms of the cluster probabilities p_α as

$$\langle H \rangle = \sum_{\alpha \in U} a_\alpha \langle H_\alpha \rangle = \sum_{\alpha \in U} a_\alpha \sum_{x_\alpha} H_\alpha(x_\alpha) p_\alpha(x_\alpha) \tag{4}$$

[b] In the case of the Boltzmann distribution

$$\begin{aligned} H_i^\dagger &= H_i = \theta_i x_i \\ H_{ij}^\dagger &= w_{ij} x_i x_j \\ H_{ij} &= w_{ij} x_i x_j + \theta_i x_i + \theta_j x_j \end{aligned}$$

and $a_{(ij)} = 1$ and $a_{(i)} = 2 - n$.

Whereas $\langle H \rangle$ can be written exactly in terms of p_α, this is not the case for the entropy term in Eq. 3. The approach is to decompose the entropy of a cluster α in terms of 'connected entropies' in the following way: [c]

$$S_\alpha = -\sum_{x_\alpha} p_\alpha(x_\alpha) \log p_\alpha(x_\alpha) = \sum_{\beta \subset \alpha} S^\dagger_\beta. \tag{5}$$

Such a decomposition can be made for any cluster. In particular it can be made for the 'cluster' consisting of all variables, so that we obtain

$$S = -\sum_{x} p(x) \log p(x) = \sum_{\beta} S^\dagger_\beta \tag{6}$$

where β runs over all subsets of variables [d]. The cluster variation method approximates the total entropy by restricting this sum to only clusters in U and re-expressing $S^\dagger_\beta$ in terms of S_α, using the Moebius formula and the definition Eq. 5.

$$S \approx \sum_{\beta \in U} S^\dagger_\beta = \sum_{\beta \in U} \sum_{\alpha \supset \beta} a_\alpha S^\dagger_\beta = \sum_{\alpha \in U} a_\alpha S_\alpha \tag{7}$$

Since S_α is a function of p_α (Eq. 5) we have expressed the entropy in terms of cluster probabilities p_α.

The quality of this approximation is illustrated in Fig. 2. Note, that the both the Bethe and Kikuchi approximation strongly deteriorate around $J = 1$, which is where the spin-glass phase starts. For $J < 1$, the Kikuchi approximation is superior to the Bethe approximation. Note, however, that this figure only illustrates the quality of the truncations in Eq. 7 assuming that the exact marginals are known. It does not say anything about the accuracy of the approximate marginals using the approximate free energy.

Substituting Eqs. 4 and 7 into the free energy Eq. 3 we obtain the approximate free energy of the Cluster Variation method. This free energy must be minimized subject to normalization constraints $\sum_{x_\alpha} p_\alpha(x_\alpha) = 1$ and consistency constraints

$$p_\alpha(x_\beta) = p_\beta(x_\beta), \quad \beta \in M, \alpha \in B, \beta \subset \alpha. \tag{8}$$

Note, that we have excluded constraints between clusters in M. This is sufficient because when $\beta, \beta' \in M$, $\beta \subset \beta'$ and $\beta' \subset \alpha \in B$: $p_\alpha(x_{\beta'}) = p_{\beta'}(x_{\beta'})$

[c] This decomposition is similar to writing a correlation in terms of means and covariance. For instance when $\alpha = (i)$, $S_{(i)} = S^\dagger_{(i)}$ is the usual mean field entropy and $S_{(ij)} = S^\dagger_{(i)} + S^\dagger_{(j)} + S^\dagger_{(ij)}$ defines two node correction.

[d] On n variables this sum contains 2^n terms.

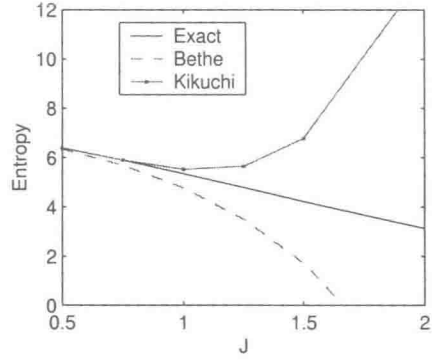

Figure 2. Exact and approximate entropies for the fully connected Boltzmann-Gibbs distribution on $n = 10$ variables with random couplings (SK model) as a function of mean coupling strength. Couplings w_{ij} are chosen from a normal Gaussian distribution with mean zero and standard deviation $J/\sqrt{n}$. External fields θ_i are chosen from a normal Gaussian distribution with mean zero and standard deviation 0.1. The exact entropy is computed from Eq. 6. The Bethe and Kikuchi approximations are computed using the approximate entropy expression Eq. 7 with exact marginals and by choosing B as the set of all pairs and all triplets, respectively.

and $p_\alpha(x_\beta) = p_\beta(x_\beta)$ implies $p_{\beta'}(x_\beta) = p_\beta(x_\beta)$. In the following, α and β will be from B and M respectively, unless otherwise stated [e].

Adding Lagrange multipliers for the constraints we obtain the Cluster Variation free energy:

$$F_{\text{cvm}}(\{p_\alpha(x_\alpha)\}, \{\lambda_\alpha\}, \{\lambda_{\alpha\beta}(x_\beta)\}) = \sum_{\alpha\in U} a_\alpha \sum_{x_\alpha} p_\alpha(x_\alpha)\left(H_\alpha(x_\alpha) + \log p_\alpha(x_\alpha)\right)$$
$$- \sum_{\alpha\in U} \lambda_\alpha \left(\sum_{x_\alpha} p_\alpha(x_\alpha) - 1\right) - \sum_{\alpha\in U}\sum_{\beta\subset\alpha}\sum_{x_\beta} \lambda_{\alpha\beta}(x_\beta)\left(p_\alpha(x_\beta) - p_\beta(x_\beta)\right) \tag{9}$$

3 Iterating Lagrange multipliers

Since the Moebius numbers can have arbitrary sign, Eq. 9 consists of a sum of convex and concave terms, and therefore is a non-convex optimization problem. One can separate F_{cvm} in a convex and concave term and derive an

[e] In fact, additional constraints can be removed, when clusters in M contain subclusters in M. See Kappen and Wiegerinck[16].