Some Inverse Problems in Statistical Physics

— Habilitationsschrift —

Pál Ruján

Fachbereich 8 Physik
Carl-von-Ossietzky Universität
D-2900 Oldenburg
Contents

1 Inverse problems in statistical mechanics 5
  1.1 A variational Ansatz ............................. 5
  1.2 Disorder and order lines ........................... 8
  1.3 Crystal growth and cellular automata ................ 11
  1.4 Combinatorial optimization ........................ 13
  1.5 Learning in neural networks ........................ 15

2 Statistical mechanics and information theory 19
  2.1 Entropy versus free energy .......................... 19
  2.2 The general setting ................................ 22
  2.3 The source coding problem ............................ 23
  2.4 Error correcting coding .............................. 27
Introduction to the Introduction

The twentieth century physics is one of the most beautiful cathedrals of science.
(L. P. Kadanoff)

No doubt. But try to visit Nôtre Dame on a sunny parisian Saturday. The building is indeed beautiful — unfinished as it will ever be. The overcrowded souvenir-shops, the overflowing restrooms, and the unpatient uproar of diesel engines makes it difficult — if not impossible — to feel or understand what the one-times builders wanted to express. What one can see is a prodigious technical performance (Wow!) before having (at last!) our Big Mac.

Is the spirit of the twenties, thirties, and sixties really gone? By opening Pandora’s box, physics made itself The instrument of military and industrial might. In many ways, the queen of natural sciences suffers today from its own success. The borderlines between basic and applied research, between science and marketing are becoming blurred and diffuse. Many researchers, especially in the interdisciplinary areas, face identity problems: how far ‘physics’ goes and what ‘real’ physics actually is. This ‘Habilitationsschrift’ summarizes my ‘story’, in the hope that others might feel encouraged to tell theirs.
Introduction

My main field of interest is statistical physics. I finished my studies in 1972, just when K. Wilson published his basic works on the renormalization group theory. The modern theory of phase transitions is a deep and beautiful achievement: it generalizes the Landau theory of second order phase transitions in a way dear to every physicist's heart. By classifying very different phenomena into universality classes depending only on the symmetries of their ground state and their dimensions, this theory comes close to the platonistic ideal of compactifying matter into elegant — and powerful — mathematical forms. Renormalization group theory has been successfully used to describe universality classes of nonlinear behaviour leading to chaos (period doubling bifurcations, for example), provides the basic groundstone for lattice gauge theories, and is widely used in signal processing applications under various names as wavelet analysis, pyramidal algorithms, multigrid methods, etc.

Although not necessarily exact from a mathematical point of view, this approach provides a very general (and thus beautiful) way of treating systems with many degrees of freedom. It also hints at the interesting technological possibility that provided appropriate symmetries are enforced, one is free to construct a model representative for its universality class. Such 'toy' models are simple but suggestive caricatures of the reality.

Now let us take an 'inverse' look: assume that we have developed some systematic, simple, and general approximation scheme: for what type of 'toy' models is the above approach exact? For example, mean field like behaviour can be obtained exactly in different limits: infinite dimensionality, infinite connectivity, long range interactions, infinite number of spin components, Bethe - lattices, etc. In renormalization group theory we can devise special lattices (hierarchical models, deterministic fractal lattices, substitution models, etc.) where the RG scheme is exact by construction. Each of these exactly soluble models is an useful starting point for perturbation
expansions or self-consistent theories.

The type of inverse problems I am going to discuss are all related to this idea of finding 'toy' models corresponding to exact limits of some approximations. In many situations of practical interest in biology, ecology, computer science, or technology in general, the problem at hand depends on a large number of strongly interacting components, is usually ill-defined (when defined at all), and no a priori Hamiltonian or dynamical description is known. Hence, we are forced to both build and solve a (static or dynamic) model. Why not a toy model? The provocative idea advocated here is that if any, then simple toy models should solve nasty real-world problems. In other words:

**Below a certain level of evidential uncertainty there is no point going beyond toy models.**

In my view, science is the search for truth (or something similar) through a stochastic gradient descent. At this particular time the gradients look rather small, which leaves us with the diffusion part. This Habilitationsschrift summarizes some of my works whose topics ranges from equilibrium statistical physics, cellular automata, to optimization theory, and neural networks. My goal is to show that despite the apparent diffusive (hopefully not diffuse!) character of this research agenda, there is a well defined logical path (drift?) underlying it.

This thesis consists basically from two parts: the first part presents briefly the main ideas and results contained in the accompanying articles. The second part deals with very recent, partly unpublished works related to the statistical physics of information theory. This part is exposed in somewhat more detail and includes some yet not published results.
Chapter 1
Inverse problems in statistical mechanics

1.1 A variational Ansatz

My interest on exactly soluble models in statistical physics stems from a 1975 preprint by my ex-colleague András Szász entitled: The exact solution of the three dimensional Ising model. This paper claimed to had expanded the Onsager—Kaufmann solution of the two-dimensional Ising model to three dimensions. As one might guess, that work was but one of the many unsuccessful attempts at solving the three dimensional Ising model. In the process of finding the errors of that particular work I became interested in the methods and the mathematical structure of exactly soluble models.

In particular, I independently rediscovered and extended[6] a variational method due to Kramers and Wannier[1] and ter Haar [2] for transfer matrices. As later it became clear, similar approaches have been used for spin Hamiltonians by Kasteleijn[3] and independently from my work by Pearson[4].

Consider a two dimensional Ising model with ferromagnetic interactions

\[-\beta E = \sum_{<i,j>} K_{ij} s_i s_j\]  

(1.1)

where $s_i = \{-1, +1\}$, $i$ denotes a vertex, and $<i,j>$ means an undirected edge of a quadratic lattice. $K_{ij} = K_x$ ($K_y$) for bonds in the $x$ ($y$) direction and $\beta = \frac{1}{kT}$. The row-to-row transfer matrix corresponding to this model is
proportional to[5]
\[ T = e^{K^x \sum_{n}^{N_x} \sigma_3(n)\sigma_3(n+1)} e^{K^z \sum_{n}^{N_z} \sigma_1(n)} \] (1.2)

with periodic boundary conditions. Here \( N_x \) is the number of spins in a row and \( \sigma_1(n), \sigma_2(n) \) are the usual Pauli matrices acting on the \( n \)-th spin, and \( \tanh K^z = e^{-2K} \). By using a series of exact (Jordan—Wigner, followed by Bogoljubov—Valatin, and finally by a Fourier) transformations, the eigenvalue problem of the transfer operator can be mapped into a massive free fermionic theory. I was interested in a more general but systematic approach which would be applicable to other spin hamiltonians than Eq. (1.1). The ground state of (1.2) can be written in general as a nodeless expansion in terms of all translational invariant even spin products:

\[ \Psi_0(\vec{\alpha}) = e^{\sum_{i=1}^{N} \sum_{i_1 \neq i_2 \neq \ldots \neq i_{2k}} a([i_2-i_1],[i_3-i_1],\ldots,[i_k-i_1])\sigma_3(i_1)\sigma_3(i_2)\ldots\sigma_3(i_k)} |0\rangle \] (1.3)

with \( k \rightarrow N/2 \) and \( \sigma_1(n)|0\rangle = |0\rangle \ \forall \ n \). Below the phase transition point, \( T_c \), the global symmetry \( (s_i \rightarrow -s_i, \forall i) \) is spontaneously broken, implying that the largest eigenvalue of the transfer matrix is degenerate. This can be taken into account by adding to (1.3) a linear combination of all odd translational invariant products of spins (the corresponding parameter vector is denoted by \( \vec{\beta} \)).

The basic observation is that the numerical value of \( \vec{\alpha}, \vec{\beta} \) parameters decreases with increasing \( |i_1 - i_k| \) and the product size both at high and at low temperatures. For example, in first order the ground state can be approximated by the variational Ansatz:

\[ \Psi_0(\alpha, \beta) = e^{\sum_{n}^{N_x} (\alpha \sigma_3(n)\sigma_3(n+1) + \beta [\sigma_3(n) + \sigma_3(n+1)])} |0\rangle \] (1.4)

Using the identity \( e^{as} = \cosh a + s \sinh a, s = \pm 1 \), this can be rewritten as a product of nearest neighbour spins, making thus the connection to product type Ansätze [4].

The second observation [6] is that the Rayleigh quotient

\[ E_0(\alpha, \beta) = \frac{\langle \Psi_0(\alpha, \beta)|T|\Psi_0(\alpha, \beta)\rangle}{\langle \Psi_0(\alpha, \beta)|\Psi_0(\alpha, \beta)\rangle} = \frac{Z_2}{Z_1} \] (1.5)

can be reinterpreted as the ratio of two simpler problems. In fact, \( Z_2 \) is the partition function of a two-row Ising problem with nearest neighbor couplings including the variational parameters, while \( Z_1 \) is the partition
function of an Ising chain whose couplings are given by twice the value of the variational parameters. Thus, this approximation reduces a fully twodimensional problem to a problem of lower dimension. It is easy to show[6] that the variational ground state energy is minimal when the thermodynamic average of a product corresponding to a particular coupling in $Z_2$ equals to the same correlation evaluated according to the effective Boltzmann factor of $Z_1$.

Thus, for the example above one has two equations:

$$ < s_n s_{n+1} >_2 = < s_n s_{n+1} >_1 \quad \text{from} \quad \frac{\partial E_0(\alpha, \beta)}{\partial \alpha} = 0 \quad (1.6) $$

$$ < s_n >_2 = < s_n >_1 \quad \text{from} \quad \frac{\partial E_0(\alpha, \beta)}{\partial \beta} = 0 \quad (1.7) $$

Instead of the full transfer matrix (1.2) we can use its Hamiltonian limit[7] $K_x \sim K_y^* \to 0$

$$ -\hat{H} = K_x \sum_{n}^{N_x} \sigma_3(n) \sigma_3(n+1) + K_y^* \sum_{n}^{N_x} \sigma_1(n) \quad (1.8) $$

which is known as an Ising chain in a transverse field. Working with the same (1.4) Ansatz requires the calculation of certain (local) correlations in the ‘effective’ one-dimensional variational Ising chain with the partition function $Z_1$.

This approximation can be systematically extended to include all possible interactions between spins in a one-dimensional chain for a given interaction range. The free energy is very accurate at both low and high temperatures, as it can be shown by comparing the variational results with series expansions results of the original problem.

Hence, all thermodynamic quantities are approximated to a high degree of accuracy with much less numerical effort than required by other methods. Since the number of minimization equations (1.6) is usually finite, the variational ground state energy is an analytical function of the variational parameters. In the variational Ansatz the order parameter is related to $\beta$. The equation determining $\beta$ is algebraic, hence the critical exponents are always of mean field type. In fact, the Ansatz does not represent well the long range ‘tail’ of the ground state, as expected from the use of one-dimensional ‘effective’ approximations of finite range. Thus, if one is interested primarily on the values of the critical exponents, one must develop a long ranged
variational Ansatz for approximating the first excitation gap (the inverse of the correlation length)[8].

In higher dimensions, the above method leads to a recursive reduction of the dimensionality. This method has been successfully applied to different statistical models and to three and four dimensional lattice gauge field theoretical models with different symmetry groups. It provides accurate values of different thermodynamic quantities with almost no computational load when compared to Monte Carlo simulations[9] of similar quality.

1.2 Disorder and order lines

The next act of the ‘Ansatz’ story happened at the 14-th IUPAP Statistical Physical Conference in Edmonton (1980), where Walter Selke presented his simulations on an intriguing extension of the Ising model, called nowadays the Axial-Next-Nearest-Neighbour Ising model (ANNNI)[10]. The Hamiltonian limit of this model is similar to (1.8), except that a second neighbour antiferromagnetic interaction is also present:

\[-\hat{H} = K_x \sum_{n=1}^{N_x} \sigma_3(n)\sigma_3(n+1) - K_2 \sum_{n=1}^{N_x} \sigma_3(n)\sigma_3(n+2) + K_y \sum_{n=1}^{N_x} \sigma_1(n) \]  

Using the Ansatz (1.4) with \( \beta = 0 \) it is easy to show that if

\[ 1 + \tau^2 = (\kappa - 1)^2 \]  

with \( \tau = \frac{K_y}{K_x} \) and \( \kappa = -\frac{K_2}{2K_x} > 0 \) the Ansatz is actually the exact ground state of the Hamiltonian (1.9)! This was first suggested by the puzzling result of Peschel and Emery [12], who have shown that on the particular path defined by (1.10) the Hamiltonian (1.9) could be mapped into the time evolution operator of the one-dimensional Ising model (kinetic Ising model) with nearest neighbor interactions only.

If one remembers that the stationary distribution (alias ground state) of the kinetic Ising model is a Boltzmann distribution whose form is identical to the Ansatz distribution in \( Z_1 \) this mapping is after all not that surprising. Hence, the variational Ansatz provides an exact solution for some parameter values on a significant ‘toy’ model!

These findings suggests a first class of inverse problems defined as the following:
Given a simple Ansatz of form (1.4) and a 'kinetic energy' term $\sum_n \sigma_1(n)$ or $e^{K_1} \sum_n \sigma_1(n)$, what kind of competing interactions a quantum spin Hamiltonian or a transfer matrix must have so that the Ansatz (1.3) is an exact ground state in some subspace of the interaction parameters' space?

This question was the main starting point for the works which then followed [13, 14, 15, 16]. The name disorder (order) line has historical reasons, since in two dimensional models the parametrized line (1.10) lies entirely either on a disordered (paramagnetic) or ordered (ferromagnetic) phase. The ground state is either nondegenerate along the whole trajectory except the end point or it is twofold degenerate. In higher dimensional systems this is not true any longer, since a two dimensional variational effective model might itself undergo a phase transitions. Nevertheless, I will call this type of solution a disorder line, even if strictly speaking this is incorrect.

The work on disorder-order lines is summarized already in [17]. However, for the sake of completeness I quote here a few of the new results contained in [13, 14]:

- Disorder solutions appear only in systems with competing interactions,
- Disorder trajectories end in the interaction parameter spaces on special Lifshitz tricritical points, the meeting point of a modulated, an ordered and a disordered phase.
- At these critical points one obtains different critical exponents in the direction with competition compared to the direction without competition. The critical exponents are related to the static and dynamic critical exponents of a lower dimensional model (without competition).
- The disorder (order) line divides two regions of the disordered (ordered) phase: a part with a monotonic decay of correlations from a part with an incommensurably modulated decay of the correlations. Although this is NOT a phase transition, it has been later shown to play an important role in understanding colloidal structures.
- The correlation length shows only weak singularities on the neighbourhood of disorder (order) lines.

A rather exhaustive classification of disorder lines in two-dimensional models is given in [15], including the Potts model on a triangular lattice. In
collaboration with A. Patkös similar techniques have been applied also for a lattice formulation of QED[16].

Are there non-integrable disorder lines, that is models exhibiting similar physical properties as the ones described above but where the simple Ansatz (1.4) is not exact? The answer is affirmative. The existence of special lines in the parameter space where spin-spin correlations change from a monotonic decay to a modulatory one is supported by numerical simulations, exact finite transfer matrix calculations, and Monte Carlo simulations on a number of models for surface adsorption on metal surfaces [18].

Another possible extension of these ideas would be the construction of a quantum spin Hamiltonian where both the ground state and the first excited state are given. This natural extension of the inverse scattering method[11] to many body systems remains up to date unsolved. The product-type Ansätze of form (1.4) can be generalized to include matrix operator products and were applied successfully to a large class of quantum antiferromagnets [19].

My attempts at understanding the structure of disorder solutions put the whole problematic of integrable systems into a constructive perspective. In fully integrable systems the disorder solutions correspond to special (but not trivial) solutions of the Yang—Baxter equations. In the same vein, one can look for some simple sufficient conditions for the integrability of spin models with short range interactions. The Yang—Baxter (star-triangle) equations require the commutator of two transfer matrices to vanish. This results in a set of overdetermined nonlinear equations, which due to some ‘magical’ conditions can be parametrized appropriately and lead then to non trivial solutions. The difficulty is that there are no given recipes for finding the right parametrization.

A weaker condition for integrability is the requirement that the transfer matrix commutes with a Hamiltonian derived from the same transfer operator. This Hamiltonian can be thought as the first element of an infinite series of motion invariant charges (strings). For self-dual models it can be shown that this condition (the existence of a local Hamiltonian) is sufficient for integrability[20]. The computation of this commutator involves straightforward but long calculations, leading to a set of overdetermined linear equations. Once the compatibility conditions are met, the subspace of interactions parameters where the model is fully integrable is easily found. I developed a REDUCE program in order to perform by computer the lengthy algebraic manipulations. As an example, I applied this approach to the case of the Ashkin—Teller model[21]. All known exactly soluble subspaces of the
1.3 Crystal growth and cellular automata

As already mentioned, the Glauber type dynamics of an Ising model enters 'unexpectedly' into the calculations involving disorder lines[12, 15]. The square of the Ansatz approximating the ground state of a quantum spin Hamiltonian can be reinterpreted as the Boltzmann distribution of a lower dimensional system. While this makes the relation between a quantum spin Hamiltonian and a kinetic master equation operator easy to understand, it is interesting to ask what happens if the same idea is applied to a transfer matrix instead of a quantum spin Hamiltonian. Thus, the second class of inverse problems can be formulated as following.

Given a certain Ansatz of form (1.3) what is the dynamics corresponding to a transfer matrix constructed such that the Ansatz is exact?

As already mentioned for quantum Hamiltonians, along a disorder (order) line the transfer matrix can be identified with the time developing operator of a lower dimensional system. Mathematically, this means that for special parametrizations of the coupling constants the transfer matrix is stochastic (preserves a probability distribution). Such stochastic matrices have been used as crystal growth models: this is the already mentioned link between disorder lines and the dynamics of stochastic crystal growth[15]. One advantage of this formulation is its extreme simplicity. It allowed finding disorder lines in a wide variety of models, including the Potts model on a triangular lattice, on a bcc lattice, as well as for other, more complicated \( \mathbb{Z}_q \) models[15].

In contrast to the time developing operator of the kinetic Ising model, which has a sparse structure due to the sequential updating rule (each \( \Delta t \) time interval only one spin is updated), the stochastic growth models are defined by full matrices, corresponding to parallel updating (each \( \Delta t \) a macroscopic number of spins is updated). Perhaps not surprisingly, many cellular automata fall into this class[22].

This elementary observation implies that many probabilistic (and as a special case, deterministic) cellular automata correspond to a 'disorder' line type solution of a higher dimensional statistical physical model with competing interactions! This link allows for the classification and discussion of cellular automata from a statistical physical point of view[22]. In many cases...
one can map the dynamics of a $d-$dimensional dynamics into the statics of a $d+1$-dimensional anisotropic spin model with competing interactions [22]. Thus, methods developed for calculating static properties of spin systems are available for treating the dynamics of cellular automata models.

Among the results I have obtained in this way I mention here the identification of exactly soluble cellular automata, as well as the derivation of exact duality relations between them [22]. Even more surprising from a theoretical point of view, dynamic phase transitions in cellular automata are related to the structure of partition function singularities in $d+1$-dimensional spin systems with competing interactions and complex valued temperatures, leading to some interesting extensions of the Lee—Yang mechanism.

The general mapping described above was used to construct dynamical systems with time dependent updating rules exhibiting strange attractors. Starting with one-dimensional Ising chains with quenched couplings or fields, simple single spin automata whose dynamics shows similar chaotic behaviour were constructed and analyzed.

The field of cellular automata has many beautiful examples coming from biology, pure mathematics, computer science, ecology, epidemiology, economy, etc. In many complex systems we do not have a full knowledge about the system’s underlying mechanisms, nor do we have enough data to decide between competing theoretical models (if such models have been developed at all). Then, it seems reasonable to work within a kind of simple atomistic model where the dynamical rules are directly related to measurable properties of the system. Cellular automata modeling is a first (and more often than not: the last) step towards a systematic description of the encountered processes. Cellular automata have parallel dynamics, are easy to develop, and numerically efficient.

Hence, the third inverse problem, which I call cellular automata engineering is

Find the dynamic rules for a cellular automaton such that it displays a set of desirable properties or it runs into some given stationary state(s).

In view of the relation to higher dimensional models, we can reformulate this inverse problem in a different form:

Find the (anisotropic, competing) interactions in a classical spin system such that a given set of states are ground states.

This type of inverse problems comes close to the problem of memorization and learning discussed later. Before entering into that field, however, we make first a small excursion into optimization theory.
1.4 Combinatorial optimization

Although many problems in pattern recognition are mathematically ill-defined, a good first step is to formulate them as optimization problems. In statistical mechanics this is trivial, since we know the Hamiltonian from the beginning. In general, however, the definition of a cost (energy) function requires a deep understanding of the characteristic features important for a given problem. Even when a cost function is available, it might be subjected to many additional and contradictory constraints. Thus, finding the ground state (the optimal solution) could require a lot of numerics, especially when scaling up the problem size. This is the typically the case for the class of problems known as \( \mathcal{NP} \)-hard\[24].

One of the most spectacular contribution of statistical physics to the praxis of combinatorial optimization is the method of simulated annealing[25]. This method is basically a Monte Carlo simulation performed at slowly decreasing temperatures (in the hope of maintaining thermal equilibrium). The term annealing is used in metalurgy for a slow, controlled cooling of a metal object. The goal is to keep the defects and dislocations in thermal equilibrium, so that they have time to diffuse to the surface of the cooled object as the temperature is lowered. Similarly, simulated annealing allows a given solution to escape from local minima via thermal activation. This method is particularly helpful in difficult optimization problems, where the number of low lying metastable states might be exponentially high.

Although simulated annealing helped M. Marchand and myself to speed up the so-called back-propagation learning algorithm [36] for feedforward networks up to a factor five to ten, this remained unacceptably slow. Thus, a basic question was to see where does the simulated annealing strategy fail and whether one can devise some other general methods which perform better. The main idea of my approach[27] was to use a classical diffusion equation analogue to a density matrix Schrödinger equation for implementing a method I called simulated tunneling[27]. In this approach a wave packet is represented by a population of individual solutions, as in the Green Function Monte Carlo method[28].

Using as toy-problem a one-dimensional tight binding model with a hierarchic potential, I have shown that the average motion of the wave packet towards the lowest minimum can at best proceed in simulated annealing according to a diffusion-like law \( <x> \sim \sqrt{t} \), while in simulated tunneling the wave packet has a ballistic behavior with logarithmic corrections, \( <x> \sim t/\ln\alpha t \). This dramatic improvement is due to two main factors:
• Cooperation of individuals in searching for the best optimum,
• Restriction of search to suboptimal solutions only.

Perhaps it is worthwhile to expose both points in more details. In the 'quantum mechanical' approach each member of the population (wavefunction) interacts with other members of the same population. Thus, in contrast to the Monte Carlo-type approach, the individual solutions exchange useful ‘information’ about the underlying energy landscape. Mathematically, this is related to the fact that the Monte Carlo dynamics is generated by a stochastic matrix. Hence, each time-trajectory is independent and carries an equal weight. The Schrödinger (or transfer matrix) operator is not stochastic and the wave packet must be reweighed after each iteration. Since we wish to keep the population finite, we need to introduce branching (reproduction) of the paths with a high weight together with a simultaneous selection (removal) of paths. While in the usual GFMC method the uneven weight of different paths leads unfortunately to a divergence of the variance[29], in our case it is exactly this feature which reflects the cooperative interaction between different members of the population.

The search efficiency can be improved further by introducing variational guidance (a good ground state Ansatz, for example). Such a guidance can force the search away from already explored territories, or can prevent different members of the population from falling into already known metastable states. Likewise, any a priori information about the system can be easily incorporated in such a form.

The observed tunneling-like behavior can be explained as following: once a member of the population has discovered a new, better potential minimum, its weight grows exponentially fast with the difference in potential energy. Within a short time interval the center of mass of the whole population moves practically to the new location. For a more detailed discussion in terms of spectral properties and wave function localization see the enclosed paper[27].

The second improvement is obtained by restricting the search to a subspace of almost optimal configurations. This is achieved by starting always the search process from the best known solution. This is in sharp contrast to simulated annealing, where the search proceeds from very high temperatures (with a randomly chosen configuration). Hence, a lot of uninteresting states are sampled before finally reaching a suboptimal solution.

In the simulated tunneling approach a localized population of solutions is created around the local minimum by increasing appropriately the kinetic
energy (temperature) term. Then, a (not necessarily slow) cooling process is applied to the whole population until a better local minimum is possibly found. The search is repeated again from this new location. This process simulates the wave reduction clearly seen in the exact numerical simulations of the tunneling process.

This complicated journey of leaving a local minimum and wandering downhills until finding another, possibly deeper local minimum can be thought as one single 'big' Monte Carlo step. Using this 'big step MC' strategy alone leads already to substantial improvements for the TSP [30].

I tested extensively the simulated tunneling method for solving the euclidean traveling salesman problem, a standard test problem for difficult optimization problems[27]. I obtained significant improvements over the best known heuristic methods (like the Kernighan-Lin method, for example). Later, I compared my results with the largest known exactly solved problem of 532 cities[31] known at that time: I obtained a solution very close to the exact global minimum and better than any other heuristic result known to me at the time (see Figure 1.4).

1.5 Learning in neural networks

As remarked above, not all cellular automata have simple stationary states. Although the map into a higher dimensional statical model is always possible, only special classes of cellular automata with an irreducible, primitive time developing operator display a stationary Boltzmann-like state. Among those, there is a simple class of cellular automata, where only two body interactions and external fields are allowed[22]. Such automata are better known as (artificial) neural networks, are direct descendents of the McCulloch and Pitts[32] model, and were use by Little and Shaw [33] to model a dynamic neuronal autoassociative memory.

The dynamics of such a network is generated by running the automata with fixed interactions (rules). The inverse problem, or learning, is the search for the values of the two-body interactions (known as synapses in analogy to synaptic connections between neurons) and of the external fields (biases) in order to adapt the behaviour of the automata to some external conditions. Although still a controversial subject in neurobiology, it is generally believed that learning and memory are related to the so-called Long Term Potentiation (LTP) process of synaptic connections. This mechanism implies that synchronous activation of a pre- and a postsynaptic neuron
The optimal tour length is: 27741

The QM5 tour length is: 27798

Figure 1.1: a) The optimal solution for the AT&T 532 American cities tour. b) The best simulated tunneling solution (after 5 iterations on a population of 100 tours = 10 CPU at 6 MFLOPS). The best Lin-Kernighan solution had a length of 27949 miles for 30 min CPU.
leads to the creation and/or facilitation of a persistent link between the two neurons.

Thus, while cellular automata are usually short ranged and homogeneous in both rules and structure, neural networks have 'individual' rules and long range interactions. As an example, let us consider one single neuron, modeled as a simple linear threshold unit. The task is here to determine the synaptic (interaction) strengths of the underlying spin model so as to perform a given task. Typical tasks performed by a neural network are function representation and hetero- and autoassociation tasks. The main analogy used by physicists is to think of the associations to be learned as some kind of energy minima of an appropriately defined energy or cost function, or in other cases, as an attractor of a certain dynamic process. As I have already explained, within a stochastic dynamics these seemingly different ideas can be embedded into a single, unified formalism.

From the beginning, I considered learning as the crucial task of the theory. From a mathematical point of view, learning amounts to finding a parametric energy model (dynamical process) whose ground (stationary) states are a priori given. Strictly speaking, this defines only a memorization process. However, due to the statistical character and the distributed storage of information characteristic for such systems, the resulting basins of attraction should represent well the (explicitely unknown) statistical properties of the example set (the associations known by the system). The main task of the theory is to characterize the generalization probability and related properties of adaptive systems.

Even with sophisticated search techniques, the learning problem in neural networks remained puzzling. For a fixed feedforward network architecture with hidden units, the problem of deciding whether the network can memorize all presented examples without any mistake is \( \mathcal{NP} \)-hard\[35\]. This does not seem to be the case for human learning, especially when considering the amount of things we can memorize\[34\]. The only possible conclusion is that the learning model itself must be flawed.

In collaboration with Mario Marchand, we considered the problem of learning in feedforward neural networks from a different perspective. We based our approach on a geometrical view of how networks perform information processing. In this language, a given architecture corresponds to a particular hyperplane partitioning of the configuration space of the input (example) vectors \[37\]. We developed a powerful heuristic algorithm partitioning the input space with a possibly minimal number of hyperplanes (neurons). Thus, instead of fitting a given architecture to the examples, we
actually build from scratch an architecture which is in some well defined sense optimal for the problem under consideration. For example, a minimal architecture is expected to represent well the symmetries and correlations hidden in the example set. If the number of processors is the bottleneck, then a hierarchic tree-type architecture is best suited, while for minimal execution time one needs a relatively large single layer of hidden units, etc.

A year later, we further improved on the mathematical conditions allowing us to build an one hidden layer feedforward architecture[38]. Further optimization of the search for single neurons (Perceptron type algorithms - notable [41]) led to a fast and powerful program for building feedforward neural networks[39]. Although I consider today the learning problem in feedforward networks satisfactorily solved, a number of open questions are still under investigation.

One such problem is the optimization of the generalization probability of the network: for the simple case of a Perceptron, one can show that a protocol allowing the machine to ask intelligent questions can help tremendously[40]. For a Perceptron with maximal stability this procedure changes the asymptotic behavior of how the average error tends to zero from $O(1/N)$ to an exponentially fast decay. In other problems one can show that by allowing queries an otherwise $NP$-hard problem learning problem is becoming a $P$-problem.

More details can be found in a talk I gave in 1990 in Clausthal[42] and reproduced below. Allowing for optimal queries into the learning protocol is therefore a step in the right direction. However, further developments in setting up a realistic learning model are hardly needed. This leads us to consider more general problems, related to the coding, communication, and robust storage of information.
Chapter 2

Statistical mechanics and information theory

2.1 Entropy versus free energy

The framework of information theory is based on the works of Claude Shannon [43, 44], who used extensively statistical mechanical concepts in proving basic theorems on storage and transmission capacities. Shannon introduced information entropy in full analogy to the Boltzmann-entropy in statistical mechanics but gave it a new, different interpretation. Shannon's works shaped almost entirely the foundations of information theory, which in its present form deals mainly with information entropy and is thus based on Gibbs' microcanonical ensemble. In statistical physics, however, it is often more convenient to work in the canonical (or the grand) canonical ensemble. In this representation the thermodynamic potential of interest is the Gibbs free energy $F(T) = U - T S(\mu N)$ instead of the entropy $S(U)$. The use of the free energy requires the introduction of temperature, a concept relatively unknown in information theory. However, as we will prove below, temperature will play an important role in optimal coding and decoding.

Consider a set of $N$ discrete variables with $q$ possible states $l_i = \{0, 1, \ldots, q-1\}$, $i = 1, 2, \ldots, N$. In statistical physics we call such variables Potts-spins, while in information theory this is an alphabet. The group properties (spin—algebra) of the discrete set will be specified in more detailed where needed. In what follows I will use the statistical physical notation $\beta = 1/T$, where $T$ is the temperature (in $k_B$ units). The Boltzmann entropy is then defined
as
\[ S = - \sum_{\alpha} p_{\alpha} \ln p_{\alpha} \]  
(2.1)

where \( \alpha \) is indexing a disjoint covering of the whole configuration space defined by \( \{v^{(\nu)}\}, \nu = 1,2,\ldots,q^N \). \( \vec{r} \) is a given configuration of spins with components \( \vec{r} = (r_1, r_2, \ldots, r_N) \). \( p_{\alpha} \) is the probability that a given event belongs to the subspace indexed by \( \alpha \). The entropy is a convex function and is maximal for a homogeneous distribution \( p_1 = p_2 = \ldots = p_M = 1/M \). Its minimal value is obtained when the probability distribution is concentrated in a single (or a few) event(s) \( \alpha, p_{\beta} = \delta_{\alpha,\beta} \).

The average number of questions with outcome \{0, 1, \ldots, q - 1\} needed to decide \( \alpha \) for a given event \( A = \vec{r} \) is given by the Hartley-Shannon-Wiener information entropy
\[ H = - \sum_{\alpha} p_{\alpha} \log_q p_{\alpha} \]  
(2.2)

where \( H \) is measured in quinary digits (\textit{quits}?).

The above definitions of the entropy depend only on the existence of the probability distribution \( p_{\alpha} \) and can otherwise describe equilibrium as well as non-equilibrium situations [57]. This is the case, for example, when the covering is time dependent or when it defines a stationary distribution different from the Boltzmann distribution (we deal here only with classical systems). In many practical applications we are given a set of data without a priori information on their possible probability distribution. In such situations, one must be careful not to forget that we are dealing with frequencies instead of probabilities and that our predictive power is bounded by the possible occurrence of rare events. The definition of the entropy can be generalized to describe quantum mechanical systems (via the density matrix) or the stationary behaviour of dynamic processes (Kolmogorov-Sinai and topological entropy).

In what follows, I will work mainly in binary systems with spin variables \( s_i = 2 \times l_i - 1 = \{\pm 1\}, q = 2 \). In thermodynamic equilibrium with a heat bath at temperature \( T \) the probability of a given configuration \( \vec{s} \) is given by
\[ p_{\vec{s}} = e^{-\beta E(\vec{s})}/Z \]  
(2.3)

where
\[ Z = \sum_{\vec{s}} e^{-\beta E(\vec{s})} \]  
(2.4)
is the canonical partition function. Substituting (2.3) in (2.1) leads to the well known formula

\[ F = \langle E \rangle - TS = U - TS \quad \text{where} \quad F = -T \ln Z \] (2.5)

This relation is true in thermodynamic equilibrium once \( E(\bar{s}) \) is a priori known.

In most inverse problems Eq.(2.3) will be interpreted as the definition of the energy functional \( E(\bar{s}) \). Such a parametrization can be constructed by expanding \( E(\bar{s}) \) in a full basis of the configuration space. Of course, this form might be useful or useless depending on the number of parameters characterizing the energy function. Minimizing the number of parameters corresponds to finding a basis with low disorder (in contrast to a basis where all coefficients are roughly equal). Therefore, a good data representation is obtained by minimizing the entropy.

Usually, the formation of patterns by self-organization corresponds to 'cooling' (lowering \( T \)) a system with very complex interactions. At sufficiently low temperatures the energy term takes over and is responsible for the 'order' emerging from 'chaos'. In pattern recognition one often uses the principle of entropy minimization (instead maximization) [45]. This can be seen as an attempt of effectively lowering the temperature and thus uncovering the low energy states.

The entropy is a thermodynamical potential and contains a lot of information about the interactions between different parts of a system. In information theory one often wants to make more general statements, which are independent from the actual form of the probability distribution. Such a quantity is the 'capacity' of a device, where one maximizes some form of the entropy over all possible probability distributions. When no other constraints are imposed, the entropy is maximal for independent, equally probable variables.

The capacity of a physical device like a flip-flop is 1 bit if the transition probabilities from one state to the other are equal. In physical units this means that the maximal amount of entropy storable in a binary variable is \( \ln 2 \) (or \( \ln q \) for \( q \)-state alphabets). Most of the information theory deals with theorems regarding the capacity (maximal amount of information) of storage devices, transmission channels, and other type of information transforms.

I find it challenging to move beyond existence theorems towards constructive methods. In other words, by using the experience accumulated in statistical mechanics and nonlinear dynamics, one should be able to develop
2.2 The general setting

Assume that a source $S$ produces a message $A$ consisting of $N$ bits, sampled uniformly and independently from the source probability distribution $P_S(A)$, $A = (s_1, s_2, \ldots, s_N)$, $s_i = \pm 1$, $i = 1, 2, \ldots, N$. Such a source can be a friend sending over bitnet a long series of snapshots depicting Ising bits obtained during a Monte Carlo simulation. Unfortunately he forgets to tell us which lattice and which Hamiltonian he simulated. The only thing we know is that the snapshots show configurations of spins equilibrated at the same temperature. This is the typical problem one faces in compression and interpretation of visual data. Another example can be a coded radio message sent by a satellite from deep space, or electric signals sent by retinal ganglion cells towards the visual cortex. A schematic picture of such a system is shown in Fig. 2.2.

Very similar problems arise in a variety of common situations: from speech transmission and FAXing to copying a diskette into the main core of a computer. Living beings are particularly good at this kind of information processing, from the single cell level up to the level of human societies.

It is clear that the task of data compression (called source coding) and the
task of building in redundance against information loss during transmission are in conflict with each other, so that at best a kind of compromise is sought. Viewed from the physical perspective we can try to model the source as a stationary Markov process, a Gibbs-distribution, or as the attractor of some unknown dynamical process. The basic task of the source coding is to find the most appropriate model (transition probabilities, energy functional, recursion relations) for the data at hand.

2.3 The source coding problem

In what follows I will discuss only how to code the source with a Gibbs distribution. This idea was introduced in computer science for the analysis of contours in an image by Geman and Geman[46] under the name Random Markov Fields. Assume that we have a two dimensional image represented by a set of pixels with a given depth of gray levels. The task is to change from pixel representation to a ‘cartoon’ representation by identifying the location of edges (discontinuities in pixel field). One possible solution is to look at the gray-level gradient (or optical flow gradient), similarly to what some Gabor-type filter cells are doing in the visual cortex. Geman and Geman introduced a model rather similar to a combination of the 8-vertex model in both vertex and spin representation, as shown in Fig. 2.3.

Thus, one has ‘spin’-variables accounting for the gray level of a pixel and
Figure 2.3: Boltzmann weights for interaction faces: a) Boltzmann factors $\epsilon_i$ for different configurations of vertices discourage configuration 2, 5, and to a lesser extent, 6; b) interactions between nearest neighbor lines discourage close parallel lines (1-2) and add kinetic energy for a kink as shown in $\omega_3$.

some discontinuity (line) variables defined on the dual lattice, describing the probability of finding there an edge. This system of variables (spin + line) has short range interactions. The corresponding Boltzmann factors are shown in Fig. 2.3.

Hence, we made an hypothesis about the possible structure of our energy functional. This Hamiltonian describes the local interactions between the spin-spin, vertex-vertex, and vertex-spin variables. We get thus a two-dimensional classical statistical mechanics problem with local but slightly unusual interactions.

The next step is to compute the numerical values of the interaction parameters. Now that we have reduced the problem of approximating the contours by a parametric distribution (the Boltzmann-distribution with the given energy functional) we can fit the parameters to the data. This is done as following: take a set of representative images, cut them into pieces (windows) of the right size and fit the coupling values to the data. In order to do this fit we need first a measure of similarity between the Gibbs distribution generated by the data and the one generated by the model. A good candidate for such a distance is the Kullback—Leibler distance between
two probability distributions:

\[ D(P(\mathbf{s}), Q(\mathbf{s})) = \sum_{\{\mathbf{s}\}} P(\mathbf{s}) \ln \frac{P(\mathbf{s})}{Q(\mathbf{s})} \]  

(2.6)

Without loss of generality both distributions can be parametrized as:

\[ P(\mathbf{s}) = \frac{\exp[-\beta E_P(\mathbf{s})]}{\sum_{(\mathbf{s})} \exp[-\beta E_P(\mathbf{s})]} = \frac{1}{Z_P} \exp[-\beta E_P(\mathbf{s})] \]  

(2.7)

where

\[ E(\mathbf{s}) = \sum_{\alpha} J_{\alpha} \prod_{i \in \alpha} s_i \]  

(2.8)

and similarly for \( Q(\mathbf{s}) \). Here \( \alpha \) denotes some tuplet indexing a given set of lattice points. Using this form one obtains

\[ D(P(\mathbf{s}), Q(\mathbf{s})) = \beta (E_P - E_Q) + \langle \ln \frac{Z_Q}{Z_P} \rangle \]  

(2.9)

where \( \langle A \rangle \) is the average value of \( A \) according to \( P(\mathbf{s}) \).

This distance is not unknown to physicists: consider a system consisting of bulk \( \mathbf{s} \) and boundary \( \mathbf{\sigma} \) spins, \( \mathbf{s} = (\mathbf{\sigma}, \mathbf{\delta}) \), such that \( E_P = E_P(\mathbf{\sigma}) + E_0(\mathbf{\delta}) \) and \( E_Q = E_Q(\mathbf{\sigma}) + E_0(\mathbf{\delta}) \). The sum in \( D \) reduces then to a sum over the possible boundary spin configurations \( \mathbf{\sigma} \) only. For a ferromagnet, for example, we can fix the boundary spins to the value \( +1 \) in one case and to a value forcing a wall into the system (upper half of boundary spins up, lower half of boundary spins down). In the ferromagnetic phase the interface free energy associated with the wall separating two ferromagnetic domains of opposite magnetization is simply

\[ \Sigma = -k_B T \ln \frac{Z_{1,1}}{Z_{1,1}} = -k_B T 2 \ln 2 D(1, 1; 1, 1) \]  

(2.10)

Therefore, we can think of the K-L distance as being a generalized interface energy. The Boltzmann-machine[47], for example, is a system of spins with two group of variables: visible (input-output) variables and hidden variables. The machine is presented with different input-output patterns enforced on the visible inputs. The visible units can be thought as some kind of special external boundary. The learning algorithm consists then of fitting the energy couplings between spins in such a way, as to minimize the (generalized) interface free energy for all presented input-output relations.
Hence, learning in the Boltzmann-machines is a special case of the type of source coding considered here.

This is again an inverse problem:

**Given a set of boundary conditions and their appearance probability, find a set of interaction parameters such that the interface free energy is minimal.**

A simple calculation shows that the necessary condition for such a minimum is given by

\[
(\prod_{i \in a} s_i)(P) = (\prod_{i \in a} s_i)(Q) \quad \forall \alpha
\]

for all products \( \alpha \) multiplied by the coupling parameters \( J_\alpha \).

A technical problem encountered when working with this equation is the evaluation of different correlations in a rather complicated two dimensional model. In all known applications a Monte Carlo simulation is used for the computation of the correlation: this requires a rather long time and is not very accurate. A simpler way of obtaining this type of information is by using a variational approach[6] for the transfer matrix corresponding to the 2D Hamiltonian. The variational Ansatz introduces further free parameters into the problem but we are left with a quasi one-dimensional problem whose numerical evaluation is relatively simple (a more detailed analysis will be published elsewhere). This variational approach delivers an accurate estimate of the desired correlations except for the immediate vicinity of a phase transition point. There is no point in striving for a better accuracy, since the averages obtained over the data have surely more noise.

Even more interesting is the possibility of using the maps explained in Chapter I. Once we have a good estimate of the transfer matrix ground state vector, we can construct a dynamic system (a stochastic matrix) which delivers the same correlations[22]. Hence, in one shot we can use instead a Random Markov Field an equivalent Markov Chain (probabilistic cellular automata). This allows for a very fast parallel hardware implementation of such an algorithm, which can be then used to interpret unknown images.

As always the case for such inverse problems, we have the freedom to use any parametrizing model we might think appropriate: hence, we might look for source parametrizations within the classes of exactly solvable statistical mechanical models. Two candidates seem rather promising: one is the Bethe lattice (or other mean-field) method, the other one is the use of lattices on which the renormalization group becomes exact: hierarchical and Berker-lattices. In both cases the evaluation of correlations is more or less straightforward.
I think that the most promising approach is the one involving hierarchical lattices. Here one can systematically introduce longer and longer range interactions so as to represent the data on many length scales. In statistics this is known as the principle of minimal structural risk [48]. The idea is to start with a relatively simple hypothesis (simple interactions) and to increase their complexity (longer range interactions) only when the prediction ability of the more complicated structure is better. Thus, we come here close to the goal of having a flexible, nonparametric algorithm which can automatically develop the most appropriate hypothesis (Hamiltonian structure) from the available data.

Another aspect of the source coding procedure is related to what should happen with the coded data. If we only want to use the code for data compression (for example by storing only the line variables) we are free to concentrate on this part of the problem alone, as explained above. If we want to send the data over a noisy transmission channel, however, we have to bias our energy functional as to include possible products of many spins, possibly with few common spins. Why this is necessary will become clear in the next section.

2.4 Error correcting coding

Having build a good parametrization of our source code, we must face the next problem, error-correcting codes. I include here a paper which appeared recently in Phys. Rev. Lett. [49].

Consider a message \( \bar{s} = (s_1, s_2, \ldots, s_N) \) consisting of \( N \) bits \( s_i = \pm 1 \), sampled independently and identically from the source probability distribution \( P_S(\bar{s}) \). The message is sent to a receiver \( \mathcal{R} \) through a noisy, memoryless communication channel \( \mathcal{C} \), which can take as input one binary variable per \( \tau \) time interval. In practice, the input variable is sent as a physical signal \( \pm v \). During transmission this signal is corrupted by white noise with zero mean and \( \sigma^2 \) variance. If the receiver \( \mathcal{R} \) accepts only binary inputs one can simply assume that every single bit passing through the channel is flipped independently with the same probability \( p < 1/2 \) (binary symmetric channel = BSC). If the receiver can deal with the whole perturbed signal with mean \( \mu s_i \) and variance \( \sigma^2 \), the channel is said to be gaussian (GC). The received message will be denoted by \( \bar{r} = (r_1, \ldots, r_N) \) and the average error probability per bit by \( p_e(\bar{s}, \bar{r}) \).

Encoding (decoding) is a procedure introducing (reducing) redundancy
so as to minimize the average error. The rate of the code is defined as \( R = \frac{N}{M} \), where \( M \) is the length of the message after coding. The capacity of the channel is the maximal mutual information obtainable from all possible source distributions \( P_S(\hat{\sigma}) \)

\[
C = \lim_{N \to \infty} \frac{1}{N} \max_{\hat{\sigma}} I(\hat{\sigma}, \hat{\sigma}) = \lim_{N \to \infty} \frac{1}{N} \max_{\hat{\sigma}} [H(\hat{\sigma}) + H(\hat{\sigma}) - H(\hat{\sigma}, \hat{\sigma})] \tag{2.12}
\]

where \( H \) is the Shannon entropy. \( C \) gives the maximal amount of information per bit which can pass through the channel for a given noise type and strength. For the simple channel models considered here

\[
C_{BSC} = 1 - h(p); \quad C_{GC} = \frac{1}{2} \log_2 \left( 1 + \frac{v^2}{w^2} \right) \tag{2.13}
\]

where \( h(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \). The famous channel coding theorem states that in the thermodynamic limit \( N \to \infty \) there are codes which will saturate the channel capacity \( (R \to C_-) \) with a vanishing average error \( p_e \to 0 \) [43]. Unfortunately, Shannon’s proof is not constructive, nor does it consider the algorithmic complexity of the encoding and decoding process. All known codes with computationally tractable coding-decoding algorithms do not saturate the channel capacity [52].

Recently, Sourlas [50] suggested a family of codes based on gauge invariant spin-glass models. The original message is stored as the \( T = 0 \) ground state of a spin-glass gauge-invariant Hamiltonian and only the (binary) coupling constants are transmitted over the noisy channel. The decoded message is defined as the ground state of a similarly structured Hamiltonian but with a set of coupling constants perturbed by the channel noise. It turns out [51] that the widely used convolution codes correspond to one-dimensional spin-glass models with complicated interactions. The maximum likelihood Viterbi decoding algorithm [52] is equivalent to a transfer matrix method for computing the \( T = 0 \) ground state.

**Encoding** consists thus of forming a set of coupling constants \( \gamma_\alpha = \pm 1 \) as

\[
\gamma_\alpha = \prod_{i \in \alpha} s_i \tag{2.14}
\]

\( i \) refers to \( \vec{r}_i \), a lattice vector in \( \mathbb{R}^d \), \( \alpha \) is a \( k \)-tuplet of indices \( (\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_k) \). Obviously, \( \gamma_\alpha \) is a binary variable lying on the vertices of a \( N/R \)-dimensional hypercube, the coupling space. A vector \( \{ \gamma \} \) constructed according to (2.14)
is called a codeword. The Hamming distance between two messages \( \bar{s} \) and \( \bar{\delta} \) is defined as

\[
d(\bar{s},\bar{\delta}) = \frac{1}{2} \sum_{i=1}^{N} (s_i - \delta_i)^2 = N - \sum_{i=1}^{N} s_i \delta_i
\]  

(2.15)

The same definition applies to the distance between two points in the coupling space, \( D(\{\gamma\}, \{K\}) \). Hence, the distance between a codeword generated by the message \( \bar{s} \) via Eq. (2.14) and an arbitrary set of couplings \( \{K\} \) is given by

\[
D(\{K\}, \bar{s}) = \frac{1}{2} \sum_{\alpha} (K_{\alpha} - v \prod_{i \in \alpha} s_i)^2 = \frac{1}{2} \sum_{\alpha} K_{\alpha}^2 + \frac{N}{2R} v^2 + E
\]  

(2.16)

where \( v = 1 \) for a binary channel and \( E \) is the Hamiltonian

\[
E(\{K\}, \bar{s}) = -\sum_{\alpha} K_{\alpha} v \prod_{i \in \alpha} s_i
\]  

(2.17)

Therefore, the energy of a configuration measures its codeword's distance from a set of coupling constants \( \{K\} \). The distance between a configuration \( \bar{s} \) and its own codeword \( \{v,\gamma\} \) is zero, so

\[
E(\bar{s}) = E_0 = \min_{\bar{s}} E(\{v,\gamma\}, \bar{s}) = -\frac{N}{R} v^2
\]  

(2.18)

Again, \( v = 1 \) for a binary channel. By carefully eliminating spurious symmetries, the ground state \( E_0 \) can be made unique. In the absence of noise, Eq. (2.14) maps a message into a codeword and the minimization of the energy (2.17) maps back a codeword into a spin configuration. This one-to-one mapping is shown in Fig. 2.4a.

Now switch on the noise. At the receiving end of the channel one obtains a set of coefficients \( K_{\alpha} \), all identically distributed according to

\[
P(K_{\alpha}) = (1 - p)\delta(K_{\alpha} - \gamma_{\alpha}) + p\delta(K_{\alpha} + \gamma_{\alpha})
\]  

(2.19)

for a binary channel and

\[
P(K_{\alpha}) = \frac{1}{\sqrt{2\pi} w} \exp \left[ -\frac{(K_{\alpha} - v\gamma_{\alpha})^2}{2w^2} \right]
\]  

(2.20)

for a gaussian channel. The effect of noise is thus to move at random the original coupling set \( \{\gamma\} \) into a new set of variables \( \{K\} \). The maximal
Figure 2.4: a) The message (left) and the coupling space (right). There is a one-to-one map between messages and codewords (crosses) but not every coupling vector (dots) is a codeword. \(d\) measures the distance in message, \(D\) in the coupling space. b) After transmission the message \(\tilde{s}\) 'lands' at a coupling vector \(\{K\}\). The usual decoding strategy is to look for the closest codeword. \(D_m\) corresponds to the minimal energy, the ground state is \(\tilde{\sigma}\). The maximal entropy decoding suggested in this Letter considers all codewords contained in the spherical shell of \(D_f\) radius and \(dE\) width (shaded area).

Likelihood decoding procedure[52] is then to find the codeword closest to \(\{K\}\), which via (2.16) corresponds to minimizing the energy (2.17). This is illustrated in Fig. 2.4b. Our main result is that the maximal information about the original message can be obtained from a different type of decoding geometry.

The energy functional \(E(\{K\}, \tilde{\sigma}) (2.17)\) defines a generalized spin-glass model invariant under the gauge transformations

\[
\sigma_k \rightarrow \epsilon_k \sigma_k, \quad k, \quad K_\alpha \rightarrow K_\alpha \prod_{i \in \alpha} \epsilon_i, \quad \forall \alpha
\]

(2.21)

where \(\{\epsilon_i = \pm 1\}\) is an arbitrary configuration of spins. All configurations \(\tilde{s}\) obtained from the original message by a gauge transformation (2.21) have the same energy. One can fix the gauge variables \(\{\epsilon_i\}\) by transforming an arbitrary message \(\tilde{s}\) into a ferromagnetically ordered configuration \(s_1 = 1, \forall i\), implying \(\gamma_{1\alpha} = 1, \forall \alpha\). In this ferromagnetic gauge the average error...
per bit (2.15) is simply

$$p_e = \frac{1}{2^N} (N - \sum_{i=1}^{N} s_i \sigma_i) = \frac{1 - m}{2}$$  \hspace{1cm} (2.22)

where $m$ is the magnetization per spin of the decoded configuration $\tilde{\sigma}$. Likewise, after transmission the average energy of the original ferromagnetic message will be:

$$E_f^{BSC} = -N \frac{(1 - 2p)}{R}, \quad E_f^{GC} = -N \frac{v^2}{R}$$  \hspace{1cm} (2.23)

implying

$$D_f^{BSC} = 2p \frac{N}{R}, \quad D_f^{GC} = w^2 \frac{N}{2R}$$  \hspace{1cm} (2.24)

for the binary and for the gaussian channel, respectively.

$D_f$ has a simple geometrical meaning (see Fig. 2.4: it is the average distance between the original codeword (message) and the perturbed coupling vector $\{K\}$. This distance is gauge invariant, self-averaging, and depends only on the noise strength and the coding rate. This suggests that instead of choosing the codeword closest to $\{K\}$ ($E = E_0$) one should search for codewords whose energy is between $E_f$ and $E_f + dE$. All configurations $\tilde{\sigma}$ whose energy matches that of the original message, (2.23) are equally probable candidates for representing the original message. This type of decoding geometry corresponds to the principle of maximal entropy [53].

In information theory the decoding procedure must deliver an unique message. This requirement is not always feasible. For the BSC, for example, the $T = 0$ ground state of (2.17) is macroscopically degenerated due to frustration effects. Hence, one has a large number of possible ground states, none being 'better' than the others. In such cases one might resort to the Bayes criterion, which suggests a majority decision for each component $\sigma_i$ (or any particular 'word' formed by $\sigma_i$'s). Likewise, in order to implement the maximal entropy decoding strategy, one must construct for a given $\{K\}$ all configurations of spins with a fixed energy $E_f$ and then apply a Bayesian decision. This implies the use of the microcanonical ensemble. In the canonical ensemble formalism the corresponding procedure is to determine the temperature at which the average energy equals $E_f$. Fortunately, the solution to this problem is already known.

Several years ago, Nishimori[54, 55] made the remarkable observation that spin-glass models which are invariant under the gauge transformations
(2.21) can be solved analytically at special temperatures. More precisely, all gauge-invariant physical quantities can be expressed as averages over the quenched coupling distribution (2.19,2.20) at the following temperatures:

\[ \beta_N = \frac{1}{2} \ln \frac{1 - p}{p} \quad (2.25) \]

for the binary distribution (BSC) and

\[ \beta_N = \frac{v}{w^2} \quad (2.26) \]

for the gaussian distribution (GC), where \( \beta = \frac{1}{kT} \) is the inverse temperature. At the Nishimori temperature the internal energy (which is gauge invariant) is thus exactly equal \([54]\) to \( E_I \), (2.23). Therefore, the loss of information due to channel noise can be simulated by heating up the spin-glass system to the Nishimori temperature. This observation suggests the following strategy for maximal entropy decoding:

a) Compute the local spin averages

\[ m_i({\{K\}}, \beta_N) = \langle \sigma_i \rangle = \frac{\sum_{\{\sigma_k\}} \sigma_i \exp(-\beta_N E)}{\sum_{\{\sigma_k\}} \exp(-\beta_N E)} \quad (2.27) \]

at the Nishimori temperature \( \beta_N \) with \( E \) given by (2.17).

b) Apply the Bayes criterion

\[ \sigma_{i, \text{decoded}} = \text{sgn}(m_i) \quad (2.28) \]

As an example, consider the following \( R = \frac{1}{2} \) code:

\[ \gamma_3(k + 1) = s_k s_{k+1} s_{k+2}, \quad \gamma_2(k + 1) = s_k s_{k+2} \quad (2.29) \]

where the variables \( s_k, \ k = 1, 2, \ldots, N \) form a one dimensional chain with free boundary conditions. The passage of the couplings through the channel is simulated by flipping independently the variables \( \gamma_{2,3}(k) \) with probability \( p \) (BSC). This results on a set of couplings \( K_{2,3}(k) \). The energy functional is now:

\[ E = - \sum_{k=2}^{N-1} E_k; \quad E_k = K_3(k)s_{k-1}s_k s_{k+1} + K_2(k)s_{k-1}s_{k+1} \quad (2.30) \]

Note that each bulk spin is 'anchored' in place by five different couplings. Flipping one spin requires a \( \sim p^3 \) order process at low \( p \). A cluster of two
neighboring spins has, however, a local field of six coupling constants. If three of those are flipped, the cluster is free to choose its orientation at random (frustration effect). Hence, a macroscopic number of spins will have a vanishing local magnetization at $T = 0$. A more detailed low temperature analysis will be published elsewhere.

In order to compute effectively the local magnetizations I use a simple variant of the transfer formalism [56]. First, assume that the spin variables are summed successively from left to right. In step $n$ one has to perform the sum

$$
\sum_{s_n} \Psi_n^>(s_n, s_{n+1}) e^{-\beta E_{n+1}(s_n, s_{n+1}, s_{n+2})} = \lambda_{n+1}^> \Psi_{n+1}^>(s_{n+1}, s_{n+2})
$$

(2.31)

where $-\ln \lambda_{n+1}^>$ contributes to the free energy. A possible parameterization of the vector $\Psi$ is

$$
\Psi_n^>(s_n, s_{n+1}) = \exp \left[ h_1^>(n)s_n + h_2^>(n)s_{n+1} + h_{12}^>(n)s_ns_{n+1} \right]
$$

(2.32)

with $\Psi_0^> = 1$. Similar expressions are used for summing up successively the spins when starting from the right end. The parameters $h_{1,2,12}^>(n)$ (2.32) defining the right $\Psi_0^<$ and left vectors $\Psi_0^>$ are stored during the left and right iterations corresponding to (2.31). The local magnetizations $\{m_i\}$ (2.27) (or other local correlations) can be now computed by summing up the chain as far as possible from both the left and the right end. This leads to the expression:

$$
m_i = \frac{\sum_{s_i-1,s_i,s_i+1} \Psi_{i-1}^> s_i e^{-\beta E_i} \Psi_{i+1}^<}{\sum_{s_i-1,s_i,s_i+1} \Psi_{i-1}^< e^{-\beta E_i} \Psi_{i+1}^>}
$$

(2.33)

Finally, the decoded spin is assigned the sign of the local magnetization.

Table I contains numerical values of the average overlap at different temperatures and noise strengths. The numerical values have been obtained by averaging 10 different transmissions of a message consisting of $10^8$ bits. The finite temperature decoding procedure delivers systematically better results (large overlap) than the $T = 0$ maximal likelihood method. The most remarkable improvement is observed at low noise levels. This can be understood as following: while at $T = 0$ a finite concentration of spins is free to flip due to frustration effects, the finite temperature entropy provides an effective stabilizing field. The results obtained at low $p$ suggest that the optimal decoding temperature should be higher than $T_N$. However, for
Table I

The average overlap $o = \sum_{i=1}^{N} s_i^{\text{input}} s_i^{\text{output}}$, $N = 10^5$, and error $p_e$ as a function of noise and inverse temperature ($\beta_N = \frac{1}{2} \ln \frac{1-p}{p}$).

The average is taken over ten independent transmissions. Each single run provides the same temperature dependence.

<table>
<thead>
<tr>
<th></th>
<th>10$\beta_N$</th>
<th>2.0$\beta_N$</th>
<th>1.5$\beta_N$</th>
<th>$\beta_N$</th>
<th>0.75$\beta_N$</th>
<th>0.5$\beta_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 0.025$</td>
<td>$\beta$</td>
<td>18.32</td>
<td>3.66</td>
<td>2.75</td>
<td>1.83</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>99945 ± 5.68</td>
<td>99963 ± 2.38</td>
<td>99963 ± 2.38</td>
<td>99963 ± 2.38</td>
<td>99963 ± 2.38</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.275e-3</td>
<td>0.185e-3</td>
<td>0.185e-3</td>
<td>0.185e-3</td>
<td>0.185e-3</td>
</tr>
<tr>
<td>$p = 0.05$</td>
<td>$\beta$</td>
<td>14.722</td>
<td>2.94</td>
<td>2.21</td>
<td>1.47</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>98844.2 ± 9.57</td>
<td>98824.6 ± 9.42</td>
<td>98824.6 ± 9.42</td>
<td>988834.2 ± 9.68</td>
<td>98880.6 ± 8.98</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.577e-2</td>
<td>0.588e-2</td>
<td>0.588e-2</td>
<td>0.583e-2</td>
<td>0.56e-2</td>
</tr>
<tr>
<td>$p = 0.075$</td>
<td>$\beta$</td>
<td>12.56</td>
<td>2.51</td>
<td>1.88</td>
<td>1.26</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>94429.4 ± 33.2</td>
<td>94437.4 ± 32.0</td>
<td>94477.55 ± 30.9</td>
<td>94519.4 ± 32.3</td>
<td>94481.8 ± 32.3</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.0279</td>
<td>0.0278</td>
<td>0.0276</td>
<td>0.0274</td>
<td>0.0276</td>
</tr>
<tr>
<td>$p = 0.100$</td>
<td>$\beta$</td>
<td>10.99</td>
<td>2.20</td>
<td>1.65</td>
<td>1.10</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>87512.8 ± 48.5</td>
<td>87517.6 ± 48.58</td>
<td>87567.4 ± 50.3</td>
<td>87590.8 ± 48.3</td>
<td>87309.6 ± 40.52</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.0624</td>
<td>0.0624</td>
<td>0.0622</td>
<td>0.0620</td>
<td>0.0635</td>
</tr>
<tr>
<td>$p = 0.125$</td>
<td>$\beta$</td>
<td>9.73</td>
<td>1.95</td>
<td>1.46</td>
<td>0.97</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>78609.8 ± 89.3</td>
<td>78732.4 ± 93.18</td>
<td>78852 ± 96.86</td>
<td>78875.4 ± 109.3</td>
<td>78329.8 ± 92.43</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.107</td>
<td>0.106</td>
<td>0.106</td>
<td>0.1056</td>
<td>0.108</td>
</tr>
<tr>
<td>$p = 0.150$</td>
<td>$\beta$</td>
<td>8.67</td>
<td>1.73</td>
<td>1.30</td>
<td>0.87</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>$o$</td>
<td>66208.6 ± 112.5</td>
<td>66364.2 ± 109.8</td>
<td>66676.4 ± 105.6</td>
<td>66951 ± 93.7</td>
<td>66178.2 ± 78.9</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0.169</td>
<td>0.168</td>
<td>0.167</td>
<td>0.165</td>
<td>0.169</td>
</tr>
</tbody>
</table>
values of \( p \) below 0.05 (\( p^3 < 1.25 \times 10^{-4} \)) the chain is too short for generating a statistically reliable number of events. The results obtained for \( p > 0.05 \) clearly show that the best overlaps are obtained at the Nishimori point and that the error increases steeply at higher temperatures.

It is worthwhile remarking that around \( p = 0.13 \) the code (2.29) looses its error correcting ability. The convolution codes correct errors by spreading the local spin information over the range of the couplings. As \( p \) increase, so does the typical correlation length associated with the disorder. When the mean cluster size of the flipped couplings becomes of the same order of magnitude as the maximal interaction range, the coding fails. This suggests the presence of a geometric phase transition in the coding ability of convolution codes.

The convolution code (2.29) does not have a particularly good performance, especially not in the BSC setup. It was used here only as an example substantiating the claim that the transmission error is systematically reduced by decoding at finite temperature. Since the Viterbi decoding algorithm is itself equivalent to an one-dimensional transfer matrix method, this performance improvement comes at no additional computational costs.

Sourlas [51] has developed coding schemes based on large non-abelian alphabets and suggested simulated annealing as a possible decoding procedure. The results presented in this Letter indicate that a simple Monte Carlo simulation at the Nishimori temperature combined with a Boltzmann factor weighed Bayes majority rule might function rather well as decoding algorithm. The above theory applies also to many signal processing applications, like image reconstruction with the Random Markov Fields method.

Thus, switching from the information entropic view to the free energy has helped us to understand why decoding at the Nishimori temperature decreases the average error. Now, assume that we have a good source code as explained in the previous Section: then we can simply add the parametrized part of the energy \( E_S \) to the code energy \( E \) when computing the magnetizations of the spins! Hence, the information we know a priori about the source can be used directly to influence our decoding decision. In order to have effective error correcting codes it is useful to code with large, independent products of spins. This might in turn influence our choice of source parametrization: when possible use large products of spins with few common elements!
Outlook

About two years ago I was visiting the ZFE Siemens AG. Somebody asked my profession. When told that I am a statistical physicist, he could not stop laughing: ‘That’s the last thing we need: a statistical chip’. In spite of this typical engineering reaction, the application of statistical physical methods to information technology seems to me inevitable. The modern statistical mechanics has developed a formalism in which most statements of information theory are conceptually simple. The wealth of knowledge accumulated during the study of collective phenomena, both from a static and a dynamic point of view, allows a statistical physicist to search for valid analogies on a huge ‘data-base’. It seems to me that in the next few years we shall witness a breakthrough of such methods in many areas of information theory, including data compression, source and channel coding, error correcting codes, and the theory of learning. I also hope that in the next few years this approach will lead to new, efficient constructive algorithms and analogue machines for many problems of practical interest. The engineers do not particularly love us and this is their playfield. We got to prove ourselves according to their rules.

For a scientist, however, the most challenging task of all remains understanding how Nature has solved similar problems: how biological information processing systems actually work. The challenge is here and the time is ripe. We, physicists, have to make our choice: either get involved or not.

Eppur si muove: with or without us.
Bibliography

[3] W. Kasteleijn Physica (Utrecht) 18 (1952) 104
[34] L. Standing Quar. J. Exp. Psychol. 25 (1973) 207

37


