Network 1 (1990) 89–103. Printed in the UK

Statistical mechanics as the underlying theory of 'elastic' and 'neural' optimisations

Petar D Simic†
California Institute of Technology, Pasadena, CA 91125, USA

Received 6 July 1989

Abstract. There is an interesting connection between two, recently popular, methods for finding good approximate solutions to hard optimisation problems, the 'neural' approach of Hopfield and Tank and the elastic-net method of Durbin and Willshaw. They both have an underlying statistical mechanics foundation and can be derived as the leading approximation to the thermodynamic free energy of related physical models. The apparent difference in the form of the two algorithms comes from different handling of constraints when evaluating the thermodynamic partition function. If all the constraints are enforced 'softly', the 'mean-field' approximation to the thermodynamic free energy is just the neural network Lyapunov function. If, on the other hand, half of the constraints are enforced 'strongly', the leading approximation to the thermodynamic free energy is the elastic-net Lyapunov function. Our results have interesting implications for the general problem of mapping optimisation problems to 'neural' and 'elastic' networks, and suggest a natural and systematic way to generalise the elastic-net and 'neural' methods to a large class of hard optimisation problems. We derive a new algorithm of the elastic-net type based on statistical mechanics. It has some of the 'positive' ingredients of the elastic-net method, yet it does not have an intrinsic problem (discussed in this paper) of the original algorithm.

1. Introduction

One of the central problems in science today is to understand the apparent ability of the central nervous system to solve problems of enormous computational complexity. This problem is made harder by the existence of many different and potentially relevant levels of its description: quantum, macromolecular, physiological, etc. It is possible, however, that most of the interesting mental phenomena should be viewed as emergent cooperative phenomena in a large system made of many interacting networks of neurons. It is implicit in this point of view that many of the details concerning the neurochemistry of various processes are not essential, in the sense that they are reduced to the overall effect they might have on the input–output characteristic of a single neuron. Furthermore, one expects that the overall behaviour of the system is not going to be very sensitive to all details of the input–output characteristic of a single neuron, but will rest on certain essential features which should be identified and captured within a phenomenological model of neuron and neural connections. This philosophy is implicit in recent work on physics-inspired models of neural networks[1]. Such models, even if biologically not realistic, are very useful and provide a concrete

† 356–48. Division of Physics and Caltech Concurrent Computation Program, e-mail address: simic@wega.caltech.edu.

0954-898X/90/010089+15$03.50

starting point for asking definite questions about computational abilities and the underlying principles, eventually shared by many adaptive learning systems.

It has been recognised by Hopfield and Tank [2] that relaxation dynamics of simple networks of model neurons, as they settle from initial to final state, can be viewed as a parallel computation of solutions to optimisation problems. They showed how, by making an anzatz for the connection strengths, one can design a network which rapidly computes good approximate solutions to the well known, computationally hard, travelling salesman problem (TSP) [3]. Further applications of the Hopfield–Tank model to various optimisation problems [4,5] have confirmed its general applicability as an efficient model of parallel computation, but have also revealed some important limitations of the original method—most notably its failure on larger problems [6].

Optimisation problems are a 'natural' laboratory for studying computational abilities of neural models because many perceptual problems routinely 'solved' by the nervous system can be viewed as optimisations [7]. More prosaically, one can view neural methods of computation as merely algorithms for solving optimisation problems, and it is interesting to compare them with other 'physical' algorithms which are potentially implementable in parallel hardware [8]. On the other hand, the search for efficient and 'physical' algorithms for finding good approximate solutions to hard optimisation problems is of critical importance for learning in large-scale neural networks; many presently popular learning algorithms, such as back-propagation [9], learn by minimising the difference between desired results and observed results, but there are strong indications that learning by this method in larger networks is going to fail.

A problem with learning in large networks, is that the space of possible weights becomes enormously large, and the 'energy' surface acquires a complicated structure of local minima without some obvious symmetries. This strong 'inhomogeneity' is typical for complex systems and it is very difficult to invent algorithms which would have 'intelligence' to avoid being trapped into a minor local minima [10].

Recently Durbin and Willshaw [11] have proposed an interesting and very efficient algorithm for approximately solving some hard optimisation problems, in particular the TSP. Their algorithm is fundamentally geometrical. It is related to the Tea Trade model [12] for the establishment of topographically ordered projections between the neural structures with matching geometries [13]. One imagines a circular path which is gradually elongated non-uniformly under the influence of the two types of force. The first pulls it towards the nearest cities; the second pulls neighbouring points of the path together. Based on this physical picture, Durbin and Willshaw proposed a set of phenomenological, first-order difference equations and pointed out that they are the gradient descent of an appropriately defined Lyapunov function; their algorithm—they call it the 'elastic net' method—naturally lends itself to implementation in parallel hardware. Unlike the neural network algorithm which in the case of N cities requires N^2 'neural' variables and N^4 connections, the elastic net uses $O(N)$ analogue variables; simulations show very good performance even for large problems, on which the Hopfield and Tank algorithm seems to fail [6].

Since many interesting applications of neural networks can be viewed as generalisations of the Hopfield–Tank approach to the TSP, it is of considerable interest to understand if it is possible to systematically generalise the Durbin–Willshaw method. Trivial observation that the two 'natural' methods which solve the same problem (the TSP), even if superficially very different, should be somehow related, lead us to consider both methods in the more general context of physically based optimisation methods. The idea we exploit is contained in the simple observation that general physical sys-

tems just by their time-evolution 'compute' solutions to 'optimisation' problems. For example, a particle moving between two points in space spontaneously 'finds' the path of minimal length; a spin system after being disturbed by some external field will relax into the state of minimal free energy. It should be noted that in the first case the 'objective' function is the action, while in the second case it is the free energy. As one considers more complex, possibly adaptive systems or, alternatively, simple systems but non-trivially constrained, one may obtain physical models which, just by their time evolution, 'compute' solutions to hard optimisation problems. This suggests that one could find new and efficient algorithms for finding good approximate solutions to hard problems by trying to figure out the efficient ways to describe the 'computation' done by the evolving physical system. In a certain sense, both the Hopfield–Tank method and the Durbin–Willshaw algorithm are particular instances of this general idea. However, it is very important to understand better both methods: why do they work; what are their limitations; is there a way to systematically improve them; and what is the 'natural' way, if any, to generalise them?

The purpose of this paper is to clarify the relation between the two, and to answer some of these questions. It turns out that both algorithms are in a deeper sense an example of what one may call a 'physical computation'; we show that there is a rather interesting connection between them and that they are closely related to the thermodynamics of certain simple physical systems. Both algorithms are essentially derivable in the course of the evaluation of the thermodynamic partition function of related physical systems. The apparent difference between the two comes from a slightly different handling of the constraints when evaluating the partition function. If all the constraints are enforced 'softly', the leading approximation to the thermodynamic free energy is just the Hopfield–Tank Lyapunov function, and the 'neural' formulation follows. If, on the other hand, half of the constraints are enforced 'strongly', the leading approximation to the thermodynamic free energy is essentially the elastic-net Lyapunov function, and the elastic-net algorithm follows.

One virtue of the algorithms which have an underlying statistical mechanical foundation is that they incorporate in a consistent way the noise —the scale of the noise being the temperature parameter.This naturally invites incorporating the ideas of the simulated annealing [14] and,in section 2, we identify this as an important ingredient of the elastic-net method.Another virtue is that one can, at least in principle, correct those algorithms by calculating systematically the effects of fluctuations [15], extending in this way the applicability and improving the accuracy of solutions, although there is usually a cost for this in terms of increased computational complexity.

We also point out an intrinsic problem with the elastic-net method: it solves a 'wrong' problem, unless the number, M, of variables that it uses is made as large as necessary for any particular instance of the problem; this, generally, may require $M \gg N$ (N = number of cities). By generalising the statistical mechanics considerations which lead to the 'derivation' of the elastic net (section 2) we derive a new algorithm of the 'elastic-net type' (section 3) which solves the 'right' problem and requires exactly N (vector) variables and equations.

In section 4 we discuss some implications of our results and point out some interesting extensions.

2. 'Derivation' of the elastic-net algorithm from statistical mechanics

In this section we will show that the elastic-net algorithm can be derived by considering

the thermodynamic limit of certain simple physical systems. Along the way we will acquire a better understanding of the two important questions: how can the elastic-net method be generalised and what are its intrinsic limitations?

We start with a simple intuitive idea which may help motivate the derivation which follows. A point particle of unit mass moving in an ordinary Euclidean plane follows the path of minimal length.The action of a particle is

$$L = \tfrac{1}{2} \int \mathrm{d}t\, \dot{\boldsymbol{x}}^2 \tag{1}$$

where $\boldsymbol{x}$ is the position vector of the particle. $L = L[\boldsymbol{x}]$ is the path functional and its stationary value (minimum) is the actual particle trajectory. If we now impose certain global constraints which restrict the set of allowed trajectories of our particle such that it puts them in correspondence with valid tours in the TSP, we obtain a physical model which just by its evolution 'computes' a good solution to the TSP, the action functional, $L = L[\boldsymbol{x}]$ playing the role of the cost function.

The direct way to find the particle trajectory, $\boldsymbol{x}_0$, is by solving the set of 'evolution' equations which follow from the variational principle associated with $L[\boldsymbol{x}]$. In the presence of the non-trivial global constraints, this method is impractical. We explore here an alternative approach which also leads to deterministic, 'relaxation' equations for the optimal trajectory, and is based on certain thermodynamical considerations.

2.1. Partition function formulation and 'cell' variables

Regarding particle trajectories as an 'ensemble', a 'statistical mechanics' can be defined by considering the following partition function,

$$Z_\beta = \sum_{\text{configurations}} \exp(-\beta L) \tag{2}$$

where the summation goes over all 'legal' trajectories, i.e. the paths which obey certain global constraints. Imagine now that the space is a discrete, irregular grid consisting of N points, $\{(x_p, y_p),\ p = 1, 2, .., N\}$, embeded in an ordinary Euclidean plane; the particle trajectories are closed and pass once through all the N points. This configuration space is in an obvious correspondence with the space of valid tours in the TSP.

An important property of the partition function is that, in the zero-temperature limit, it is dominated by the term of smallest 'cost'

$$Z_\beta \underset{\beta \to 0+}{\approx} \exp(-\beta L[\boldsymbol{x}_0]). \tag{3}$$

As the consequence, the ensemble-averaged particle trajectory, $\langle \boldsymbol{x} \rangle_\beta$, becomes

$$\langle \boldsymbol{x} \rangle_\beta = \frac{1}{Z_\beta} \sum_{\forall \boldsymbol{x}} \boldsymbol{x} \exp(-\beta L[\boldsymbol{x}]) \underset{\beta \to 0+}{\approx} \boldsymbol{x}_0 + \mathrm{O}(\exp(-\beta |\delta L|)) \tag{4}$$

We see that, in the zero-temperature limit, the average particle trajectory approaches the optimal trajectory—the system is in its ground state. This suggests that instead of minimising the 'cost' function ($L[\boldsymbol{x}]$) directly, one could attempt to evaluate statistical averages, such as $\langle \boldsymbol{x} \rangle_\beta$ and $\langle L \rangle_\beta$, at a given level of the noise (read 'temperature'), and

then try to follow how the averages change as one decreases the level of the noise towards zero [14].

It is quite interesting, yet well known, that this procedure is equivalent to minimising an effective energy function called the thermodynamic free energy ($F_\beta[\boldsymbol{x}]$ in the following), which can be constructed once the partition function of the problem is known. Since the free energy 'knows' about the noise—it depends in a non-trivial way on the temperature—its energy surface is much smoother than the original surface associated with $L[\boldsymbol{x}]$. Typically, features (such as a minor local minima) whose 'size' is smaller then the scale characterising the noise (T) are averaged over, and only gross features are present; on lowering the temperature fine details gradually appear and the effective 'cost' function ($F_\beta[\boldsymbol{x}]$) goes over to the original cost function, $L[\boldsymbol{x}]$.

Despite the simplicity of the 'energy' function, L, formula (2) is impractical; its complexity is hidden in the requirement that the summation should be performed over the non-trivially constrained configurations. These constraints can be made explicit by introducing a redundant set of binary variables ('cell' variables in the following), $\{\eta_p^i\}$, associated with possible events along the particle trajectory. For example, $\eta_p^i = 1$ means that the particle at time i occupies space-point p. In terms of the new variables our particle trajectory can be decomposed as

$$x(i) = \sum_p x_p \eta_p^i \tag{5a}$$

$$y(i) = \sum_p y_p \eta_p^i \tag{5b}$$

where i is the time parameter, $i = 1, 2, ..., M\ (= N)$; it is a count of the clock attached to the particle clicking in regular time intervals, $\mathrm{d}t = 1$; M is in all our formulae assumed equal to N, yet it is important to keep in mind the distinction between these two numbers: N determines the 'size' of the problem and we expect that as $N \to \infty$, L scales as $\sqrt{N}$; M determines how fine the discretisation is of the particle trajectory, and $M \to \infty$ is the continuum limit.

The 'cell' variables, η_p^i , are similar to the 'neural' variables of Hopfield and Tank. They are either 1 or 0 depending on whether a particular event with which they are associated occurred; if the particle at time i occupies point p, then $\eta_p^i = 1$ and $\eta_p^j = \eta_q^i = 0$ for all q and j different from p and i, respectively. This decomposition is implicit in the work of Hopfield and Tank [2,4]. A similar trick was used by Fox and Furmanski in their generalisation of the Hopfield–Tank approach to a large class of problems in parallel computation. [5]

The remaining constraints of our problem are written compactly in terms of new variables:

$$\sum_i \eta_p^i \eta_q^i = \delta_{pq} \qquad \forall i < j \tag{6a}$$

$$\sum_p \eta_p^i \eta_p^j = 0. \tag{6b}$$

The constraints (6a) and (6b) ensure that the particle cannot visit the two points at the same time, and, that it visits all the points once and only once. The action, L, in the new variables becomes

$$L = \sum_i \sum_{p,q} \boldsymbol{x}_p \boldsymbol{x}_q (\eta_p^i \eta_q^i - \eta_p^{i+1} \eta_q^i) \tag{7}$$

where we have introduced vector variables, $\boldsymbol{x} = (x, y)$. Using part of the constraints, L can be conveniently rewritten as

$$L = \frac{1}{4}\sum_{i}\sum_{p,q} d_{pq}^2 \eta_p^i(\eta_q^{i+1} + \eta_q^{i-1}) \tag{8}$$

where d_{pq} is the distance between points p and q. When the constraints are obeyed, this term is numerically equal to the sum of distance-squares along the particle trajectory. Clearly, minimising this term is only under certain conditions equivalent to minimising the length of the particle trajectory.We will return to this point later.

Having changed variables, the sum over particle trajectories in (2) translates into a sum over all η-configurations subject to the constraints (6*a*) and (6*b*). There are basically two methods to do this summation, depending on the way one deals with the constraints. The first method is to enforce the constraints 'softly', by adding penalty terms to the 'energy' function, L:

$$L = \frac{1}{4}\sum_{i} d_{pq}^2 \eta_p^i(\eta_q^{i+1} + \eta_q^{i-1}) + \frac{\alpha}{2}\sum_{i}\sum_{p,q} d_{pq}^2 \eta_p^i \eta_q^i + \frac{1}{4}\sum_{i,j}\sum_{p} \gamma_{ij}\eta_p^i \eta_p^j + \sum_{p} \gamma_p \left(\sum_{i} \eta_p^i - 1\right)^2 \tag{9}$$

where α, γ_{ij} $(\gamma_{ii} = 0)$, and γ_p are arbitrary parameters, penalities for violation of the constraints, (6*a*) and (6*b*), respectively. Then one evaluates the partition function (2) by the unrestricted summation over all 2^{N^2} possible configurations of N^2 'cell' variables. The second method is to enforce 'strongly' all, or a significant part of the constraints, and to perform the summation in (2) over 'legal' configurations only. The approximate evaluation of the partition function is usually easier by employing the first method; however, the second method is what one is really instructed to do according to (2).

The second method is employed in this section. We enforce part of the constraints 'strongly' by summing only over those configurations which obey the constraints (6*b*), and the diagonal part of the constraints (6*a*); in this case the third and the fourth term in equation (9) can be dropped. The rest of the constraints , in (6*a*), are enforced 'softly', but we eventually correlate the strength of the penalty term , α, with the inverse temperature so that $\alpha \propto O(\beta)$; in the annealing mode the temperature is decreased towards zero and these constraints are also obeyed.

The central object we want to calculate is the thermodynamic partition function, $Z_\beta[\boldsymbol{H}]$, defined as

$$Z_\beta[\boldsymbol{H}] = \sum_{\{\eta\}} \exp\left[-\beta\left(L[\boldsymbol{\eta}] + \sum_{i,p} H_p^i \eta_p^i\right)\right] \tag{10}$$

where $L[\boldsymbol{\eta}]$ is given by the first two terms in (9) and the summation is restricted to the subspace made of η-configurations satisfying the constraints (6*b*) and the diagonal part of (6*a*); $\boldsymbol{H}$ is the matrix of external 'forces'. Once the partition function is calculated—it depends on the temperature, the set of coupling strengths, and, the external 'forces' ($\boldsymbol{H}$)—all the relevant statistical information about the system can be deduced from it. Differentiating it with respect to external 'forces' about $\boldsymbol{H} = 0$, one generates all η-correlation functions. For example, the one-point correlation function$\langle \eta_p^i \rangle$, is given by

$$m_p^i = \langle \eta_p^i \rangle = -\frac{1}{\beta}\left(\frac{\delta \ln Z_\beta[\boldsymbol{H}]}{\delta H_p^i}\right)_{\boldsymbol{H}\to 0^+} \tag{11}$$

and is interpreted as the probability of the event η_p^i; that is, the probability that a particle visits point p at time i. Knowing $\langle \eta_p^i \rangle$ one can calculate the ensemble average particle trajectory simply as,

$$\langle \boldsymbol{x}(i) \rangle = \sum_{\forall p} \boldsymbol{x}_p \langle \eta_p^i \rangle. \tag{12}$$

Similarly, one can evaluate the average value of any other quantity, $A(\boldsymbol{x})$ which is a function of the system state, by using the following result:

$$\langle A(\boldsymbol{x}) \rangle_\beta = \frac{1}{Z_\beta} \sum_{\forall \boldsymbol{x}} A(\boldsymbol{x}) \exp(-\beta L[\boldsymbol{x}]) = A\left(-\frac{1}{\beta}\frac{\delta}{\delta \boldsymbol{H}}\right) \ln Z_\beta[\boldsymbol{H}]. \tag{13}$$

As we have already mentioned, the thermodynamic free energy, can also be calculated from the partition function; it is given by

$$F_\beta[\boldsymbol{m}] = -\frac{1}{\beta} \ln Z_\beta[\boldsymbol{H}] - \sum_{i,p} H_p^i m_p^i. \tag{14}$$

It is our main goal, in the rest of this section, to derive an explicit expression for $F_\beta[\boldsymbol{m}]$ and to show how it relates to the 'elastic net' Lyapunov function [11].

2.2. *Derivation of the elastic net*

Using the well known trick

$$\exp(-\tfrac{1}{2}\beta\eta^2) = \frac{\int \mathrm{d}\phi \exp(\phi^2/2\beta - \eta\phi)}{\int \mathrm{d}\phi \exp(\phi^2/2\beta)} \tag{15}$$

we can rewrite the partition function (10) as

$$Z_\beta[\boldsymbol{H}] = \sum_{\{\eta\}} \int \mathrm{d}\phi \exp\left(\frac{\beta}{2} \sum_{i,j} \sum_{p,q} \phi_p^i (J^{-1})_{pq}^{ij} \phi_q^j + \sum_{i,p} \eta_p^i (\phi_p^i + H_p^i)\right)$$

$$J_{pq}^{ij} = \tfrac{1}{2} d_{pq}^2 (\delta_{j-1}^i + \delta_{j+1}^i + \alpha \delta_j^i) \tag{16}$$

where $\mathrm{d}\phi$ is the scale invariant measure, defined up to a multiplicative factor as

$$\mathrm{d}\phi \propto \prod_{i,n} \mathrm{d}\phi_n^i \left[\int \prod_{j,m} \mathrm{d}\phi_m^j \exp\left(\frac{\beta}{2} \sum_{p,q} \sum_{k,l} \phi_p^k (J^{-1})_{pq}^{kl} \phi_q^l\right)\right]^{-1}. \tag{17}$$

One can easily prove, from (16) , the relation

$$\langle \phi_p^i \rangle = \sum_{j,q} J_{pq}^{ij} m_q^j \tag{18}$$

where $\langle \phi_p^i \rangle$ is the average defined with respect to the measure (17).

Let us now perform the summation over all η-configurations which obey (6*b*) and the diagonal part of (6*a*).

$$\sum_{\{\eta\}} \exp\left(-\beta \sum_{i,p} \eta_p^i(\phi_p^i + H_p^i)\right) = \prod_p \sum_j \exp[-\beta(\phi_p^j + H_p^j)]$$
$$= \exp\left(\sum_p \ln \sum_j \exp[-\beta(\phi_p^j + H_p^j)]\right). \tag{19}$$

The partition function (16) becomes

$$Z_\beta[\boldsymbol{H}] = \int \mathrm{d}\phi \exp(-\beta E(\boldsymbol{\phi})) \tag{20}$$

$$E(\boldsymbol{\phi}) = -\frac{1}{2} \sum_{i,j} \sum_{p,q} \phi_p^i (J^{-1})_{pq}^{ij} \phi_q^j - \frac{1}{\beta} \sum_p \ln \sum_i \exp[-\beta(\phi_p^i + H_p^i)]. \tag{21}$$

An important feature, apparent in (16), is that the integral over $\boldsymbol{\phi}$ can be exactly evaluated by expanding it around its saddle-point. The same holds true, in the sense of the 'mean-field' theory, for the integral (20). To see this, consider its saddle-point equation

$$\left(\frac{\delta E}{\delta \phi_p^i}\right)_{\phi=\tilde{\phi}} \propto \tilde{\phi}_p^i - \sum_{j,q} J_{pq}^{ij} \frac{\exp(-\beta \tilde{\phi}_q^j)}{\sum_l \exp(-\beta \tilde{\phi}_q^l)} = 0. \tag{22}$$

On the other hand, differentiating Z_β with respect to the external fields, around $\boldsymbol{H} = 0$, we obtain

$$m_p^i = \left\langle \frac{\exp(-\beta \phi_p^i)}{\sum_j \exp(-\beta \phi_p^j)} \right\rangle \simeq \frac{\exp(-\beta \langle \phi_p^i \rangle)}{\sum_j \exp(-\beta \langle \phi_p^j \rangle)} + \mathrm{O}\left(\frac{1}{N^\delta}\right) \tag{23}$$

where the final expression is obtained in the usual 'mean-field' approximation, which amounts to the neglect of fluctuations, and is expected to be valid in the thermodynamic limit ($N \to \infty$). If we now replace (23) in the exact formula (18) we obtain just the saddle-point equation (22). This completes our argument, showing that the average field, $\langle \phi_p^i \rangle$ is indeed the saddle-point of the integral over ϕ in (20)

$$\langle \phi_p^i \rangle \simeq \tilde{\phi}_p^i + \mathrm{O}\left(\frac{1}{N^\delta}\right) \qquad \delta > 0.$$

To a first approximation, the partition function is simply

$$Z_\beta \simeq \exp(-\beta E(\tilde{\boldsymbol{\phi}})) \tag{24}$$

and the thermodynamic free energy, according to (14),becomes

$$F[\boldsymbol{m}] \simeq -\frac{1}{2} \sum_{i,j} \sum_{p,q} J_{pq}^{ij} m_p^i m_q^j - \frac{1}{\beta} \sum_p \ln \sum_i \exp\left[-\beta\left(\sum_{j,q} J_{pq}^{ij} m_q^j + H_p^i\right)\right]$$
$$- \sum_{i,p} H_p^i m_p^i + \mathrm{O}\left(\frac{1}{N}\right). \tag{25}$$

Expanding in powers of the external fields, H_p^i, and using (23), one obtains

$$F[\boldsymbol{m}] \simeq -\frac{1}{2}\sum_{i,j}\sum_{p,q} J_{pq}^{ij} m_p^i m_q^j - \frac{1}{\beta}\sum_p \ln \sum_i \exp\left(-\beta \sum_{j,q} J_{pq}^{ij} m_q^j\right) + \mathrm{O}(H^2) \tag{26}$$

where $\mathrm{O}(H^2)$ and higher powers of the external fields can be dropped since we are interested in 'spontaneous' time evolution of our particle system, i.e. the $H \to 0$ limit.

The equilibrium states of a thermodynamical system are those configurations which minimise its thermodynamic free energy. Expression (26) is still not in the standard form expected from thermodynamics; in the present case, this is

$$F = L - TS = \frac{1}{2}\sum_i |\boldsymbol{x}^{i+1} - \boldsymbol{x}^i|^2 - TS \tag{27}$$

where S is the entropy (note the 'wrong' sign of the first term in (26)). To obtain the free energy in the standard form, we expand the exponent in (26) in powers of $\delta\phi_p^i = \frac{1}{2}\sum_q \mathrm{d}_{pq}^2(m_q^{i+1} + m_q^{i-1})$ and obtain

$$\begin{aligned} F[\boldsymbol{m}] \simeq{}& \frac{1}{4}\sum_i\sum_{p,q} d_{pq}^2 m_p^i (m_q^{i+1} + m_q^{i-1}) + \frac{\alpha}{2}\sum_i\sum_{p,q} d_{pq}^2 m_p^i m_q^i \\ & - \frac{1}{\beta}\sum_p \ln \sum_i \exp\left(-\beta\frac{\alpha}{2}\sum_q d_{pq}^2 m_q^i\right) \\ & + \frac{\beta}{8}\sum_{i,j}\sum_{p,p',q} (\delta^{ij} m_p^i - m_p^i m_p^j) d_{pq}^2 d_{pp'}^2 (m_q^{j+1} + m_q^{j-1})(m_{p'}^{i+1} + m_{p'}^{i-1}) \\ & + \cdots . \end{aligned} \tag{28}$$

When the constraints are obeyed, the first term in (28) is equal to $L = \frac{1}{2}\sum_i |\boldsymbol{x}^{i+1} - \boldsymbol{x}^i|^2$, now with the correct sign; the second term corresponds to the entropy S. All the other terms vanish. Using the decomposition of the average particle trajectory

$$\boldsymbol{x}^i = \langle \boldsymbol{x}(i) \rangle = \sum_p \boldsymbol{x}_p m_p^i \tag{29}$$

one can rewrite $F[\boldsymbol{m}]$ as a function of the average particle trajectory only; this interpretation of $F[\boldsymbol{m}]$ reduces the dimensionality of the energy surface associated with it, from N^2 (in m_p^i space) to $2M$ (in $\boldsymbol{x}$ space). We obtain

$$F[\boldsymbol{m}] \simeq \frac{1}{2}\sum_i |\boldsymbol{x}^{i+1} - \boldsymbol{x}^i|^2 - \frac{1}{\beta}\sum_p \ln \sum_j \exp\left(\frac{\beta}{2}\alpha |\boldsymbol{x}_p - \boldsymbol{x}^j|^2\right) + \mathrm{O}\left(\frac{1}{M^2}\right). \tag{30}$$

The first term, $L = \frac{1}{2}\sum_i |\boldsymbol{x}^{i+1} - \boldsymbol{x}^i|^2$, is of the order $\mathrm{O}(1/M)$, where M is the measure of how finely we have discretised the particle trajectory; the continuum limit corresponds to $M \to \infty$. On the other hand, this is precisely the relevant limit for the TSP because only in this limit is minimising L equivalent to minimising the length of the particle trajectory. We drop therefore $\mathrm{O}(1/M^2)$ and higher-order terms in (30). Remarkably, if one now choses $\alpha = \beta$, $F[\boldsymbol{m}]$ becomes precisely the Durbin–Willshaw–Yuille [16] energy function associated with the elastic net.

The elastic-net equations are just the relaxation equations associated with the thermodynamic system described by $F[\boldsymbol{m}]$,

$$\delta x^i = -\delta t \frac{1}{\beta} \frac{\delta F}{\delta x^i} \qquad i = 1, ..., M \tag{31}$$

where the time-step δt is measured in units of temperature, $T = 1/\beta = K$; K is a free parameter appearing in the original proposal of the elastic-net algorithm, and we see that it corresponds to the temperature of our system. Finally, the elastic ring ('snake') of Durbin and Willshaw has rather interesting interpretation. It is just the expectation value of the particle trajectory, $\boldsymbol{x} = \langle \boldsymbol{x}(i) \rangle$ at a given level of the thermal noise, $T = K$.

This completes our first look at the thermodynamical interpretation of the elastic-net algorithm.

2.3. *Why does the elastic net work and what are its limitations?*

The elastic-net method has several ingredients which conspire to make it an efficient calculation. We have seen that it incorporates effectively the thermal noise, characterised by the temperature parameter (K). Complex optimisation problems are often characterised by many locally stable states many of which are much higher in energy than the ground state. Noise is obviously a good ingredient since it enables the system to fluctuate over some energy barriers, and to escape being trapped in local minima.

Furthermore, the level of the noise is controlled during the calculation. Indeed, the prescription for computation with the elastic-net equations (31) given by Durbin and Willshaw is that one pick an initial value $K = Ki$, and then gradually decrease the value of K towards zero. In the same time, the time-step is decreased with K. Having identified K with the temperature and $F[m]$ with the thermodynamic free energy of our physical system, we understand that this algorithm actually performs simulated annealing [14]. It is a fast way[1,2,17] of doing simulated annealing: instead of randomly generating configurations until the system equilibrates at given T, one searches for the equilibrium configurations directly by minimising the thermodynamic free energy of the system. This is done by integrating, at each T, a deterministic set of first-order relaxation equations, an essentially parallel operation.

We will see that the noise and the possibility for annealing are common ingredients for both the elastic-net algorithm and the neural network approach. What is unique to the elastic net is the way constraints are built into the functional form of the algorithm. We have seen that some of the constraints are enforced exactly. The penalty, for violation of those which are enforced 'softly' is correlated with β and becomes infinite as the temperature decreases towards zero; as the calculation progresses all the constraints become exactly satisfied.

Finally, we point out an intrinsic problem of the elastic-net algorithm. It follows from the observation that the basic quantity that the algorithm tries to minimise is not the length of the particle trajectory but the sum over distance-squares along the trajectory. It is only in the limit of very fine discretisation of the particle trajectory that the sum over distance-squares is proportional to the length (square) of the trajectory. This means that the number of equations, M, needed for good performance of the algorithm depends (for fixed N) strongly on a particular distribution of the cities and it can be made very large by, for example, increasing the distance of some cities from the others.

3. Connection with neural networks and a novel physical algorithm for solving the TSP

In this section we clarify the relation between the elastic net and the Hopfield–Tank neural network approach, by considering the statistical mechanics of a slightly different but related physical model. We show that if all the constraints are enforced 'softly' when evaluating the thermodynamic free energy of this model, one obtains the Hopfield–Tank 'neural' formulation of the TSP; we only sketch the derivation of this result since the connection between Hopfield's equations and mean-field equations of the spin-glass is well known [1,2,5,17]. If, on the other hand, half of the constraints are enforced 'strongly', one arrives at the new algorithm of the 'elastic-net type'. The difference is that since the particle action we consider now is precisely equal to the length of the particle trajectory, the resulting 'elastic-net type' algorithm is expected to require exactly $M = N$ (vector) variables; it has the positive 'thermodynamic' ingredients of the elastic net, yet it does not have its basic problem—the strong dependence of the number of variables, M, required for its good performance on a particular instance of the problem—the $M \gg N$ requirement.

Consider a two-dimensional Riemannian space described by metric tensor $g_{\mu\nu}$ of Euclidean signature. The length of a path in this space is

$$L = \int \mathrm{d}\tau \left(g_{\mu\nu} \frac{\mathrm{d}x^{\mu}}{\mathrm{d}\tau} \frac{\mathrm{d}x^{\nu}}{\mathrm{d}\tau} \right)^{1/2} \tag{32}$$

where τ is an arbitrary parameter that labels the points along the path. The L can be interpreted as the action of a point particle moving in the spacetime described by $g_{\mu\nu}$. Alternatively, L can be interpreted as the static energy of a closed string embedded in the two-dimensional space described by $g_{\mu\nu}$. Finally, from the point of view of the TSP, L is just the cost function (the length of the tour) and one wants to find the tour of minimal cost.

3.1. Enforcing the constraints 'softly'—the neural network approach

Specialising, as before, to the case of an ordinary Euclidean plane with N points (cities) embedded in it, we consider a 'statistical mechanics' defined over the same set of configurations as before but L now becomes

$$L = \frac{1}{4} \sum_i d_{pq} \eta_p^i (\eta_q^{i+1} + \eta_q^{i-1}) + \frac{\alpha}{2} \sum_i \sum_{p,q} d_{pq} \eta_p^i \eta_q^i + \frac{1}{4} \sum_p \sum_{i,j} \gamma_{ij} \eta_p^i \eta_p^j + \sum_p \gamma_p \left(\sum_i \eta_p^i - 1 \right)^2 . \tag{33}$$

The difference from the case of the non-relativistic particle in (2+1)-dimensional spacetime, studied previously, is the appearance of d_{pq} instead of d_{pq}^2 in (33). When the constraints are obeyed the second and the third terms are zero, while the first term is equal to the length of the particle trajectory. We introduce now the 'spin' variables, $s_p^i = 2\eta_p^i - 1$, and consider the 'entropy' term

$$S = \ln \sum_{\{s\}} \exp\left(-\beta \sum_{i,p} s_p^i (\phi_p^i + H_p^i) \right). \tag{34}$$

Since we have included the penalty terms for 'soft' violation of the constraint in L, (33), we can proceed and evaluate the partition function by summing over all possible configurations of N^2 'neural' variables. The result is

$$S \approx \sum_{i,p} \ln 2\cosh[\beta(\phi_p^i + H_p^i)]. \tag{35}$$

Omitting details of the derivation, which is similar to that presented in the previous section,the thermodynamic free energy in this case is

$$F[\boldsymbol{m}] \approx \frac{1}{4}\sum_i \sum_{p,q} d_{pq} m_p^i (m_q^{i+1} + m_q^{i-1}) + \frac{\alpha}{4}\sum_i \sum_{p,q} d_{pq} m_p^i m_q^i + \frac{1}{4}\sum_p \sum_{i,j} \gamma_{ij} m_p^i m_q^j + \frac{1}{\beta}\sum_{i,q}[(1-m_q^i)\ln(1-m_q^i) + m_q^i \ln m_q^i]. \tag{36}$$

Although this expression looks different from the neural network Lyapunov function written originally by Hopfield [1], it is in fact equivalent to it (up to some improvement, suggested by our derivation, and reflected in the form of the penalty terms); to see this is a matter of a simple partial integration of Hopfield's original expression.

The equilibrium configurations, m_α, are found by solving the following equations:

$$\left(\frac{\delta F[\boldsymbol{m}]}{\delta m_p^i}\right)_{m_\alpha} = 0. \tag{37}$$

Hopfield's neural network equations are basically a relaxation method of solving these equations.

3.2. *Enforcing the constraints 'strongly'—a novel elastic-net type algorithm*

The essential ingredient in our statistical mechanics 'derivation' of the elastic-net method was the exact ('strong') enforcement of the constraints, when evaluating the partition function; similarly, the 'soft' way of enforcing the constraints was essential in obtaining the neural network formulation (36) and (37). A 'strong' enforcement of the constraints has an obvious advantage from the point of view of statistical mechanics since it leads to a more accurate expression for the free energy of the thermodynamic system. To generalise the elastic-net algorithm to the present case,we approximately evaluate (34) by summing only over the configurations of 'neural' variables which obey the constraints (6*b*) and the diagonal part of (6*a*). We obtain

$$S \approx \sum_p \ln \sum_j \exp[\beta(\phi_p^j + H_p^j)]. \tag{38}$$

Omitting the details of the derivation, the leading approximation to the thermodynamic free energy in the present case is

$$F[\boldsymbol{x}] \approx -\frac{1}{2}\sum_i |\boldsymbol{x}^{i+1} - \boldsymbol{x}^i| - \frac{1}{\beta}\sum_p \ln \sum_j \exp\left(-\frac{\beta}{2}(\alpha|\boldsymbol{x}_p - \boldsymbol{x}^j| + |\boldsymbol{x}_p - \boldsymbol{x}^{j+1}| + |\boldsymbol{x}_p - \boldsymbol{x}^{j-1}|)\right). \tag{39}$$

This function resembles the elastic-net Lyapunov function although there are some significant differences. This is an example of what we mean by 'generalisation' of the elastic net. The associated relaxation equations are

$$\delta x^i = -\delta t K \frac{\delta F}{\delta x^i} \qquad i = 1, ..., N \tag{40}$$

where K and α (see (39)) are free parameters and the appropriate choice for them may or may not be the same as in the original elastic-net algorithm [18].

When used in the annealing mode, in the way similar to that prescribed by Durbin and Willshaw, those equations hold promise to represent an efficient algorithm for the TSP. They scale excellently with the size of the problem; since the associated free energy (39) in the low-temperature limit becomes exactly the length of the particle trajectory, one needs exactly $2N$ variables for this algorithm. This should be compared with the elastic-net algorithm which, as we discussed previously, depending on a particular instance of the problem, may need $M \gg N$ variables and with the neural network algorithm which needs N^2 variables.

4. Concluding remarks

One of the virtues of the deterministic algorithms which have an underlying statistical mechanics foundation is that they incorporate in a consistent way the noise—the scale of the noise being the temperature parameter. Complex systems are often characterised by many locally stable states much higher in energy than the ground state. Noise is obviously a 'good' ingredient since it enables the system to fluctuate over some energy barriers and escape of being trapped into a minor local minima. This, together with the appropriate control of the noise level, amounts of introducing the resolution scale in the search: at 'temperature' T the algorithm distinguishes only those features of the system whose 'size' is bigger then T; as the temperature is lowered, fine details of the system are gradually accounted for.

Another virtue is that one can view such algorithms as just the leading approximation to the much more powerful computations done by the underlying physical system, and there is a way ('loop-expansion'), at least in principle, to correct them by calculating systematically, order-by-order, the effect of fluctuations. These corrections, although often computationally expensive, may in certain cases extend the applicability and improve the quality of solutions. In fact, we have recently shown [15] that some of the terms which are usually introduced by hand as heuristics to improve the convergence of neural optimisations appear, even if one does not introduce them, spontaneously as the result of correlations induced by thermal fluctuations.

In most natural physical systems noise is present and to some extent it is generated by the elements of the system itself. Local controlling of this noise—something analogueous to simulated annealing—may be an important part of the adaptive strategy of natural systems and may contribute much to their 'computational' power. It is important to have both, the noise and the procedure for its control (simulated annealing), incorporated into the algorithms for solving complex problems. We have seen that the elastic-net algorithm incorporates both of these ingredients in a consistent way and we have seen that the neural network algorithm could, by incorporating controlled decrease of the temperature, do the same. Yet the essential difference in the form of the two algorithms—by the 'form' we mean the form of the associated

Lyapunov functions, character of the variables and dimensionality of the representation space, etc —comes from the way one handles the constraints when evaluating the thermodynamic free energy of the corresponding physical systems. Roughly speaking, the 'soft' enforcement of the constraints leads to the neural network formulation of the problem with N^2 'neural' variables and the same number of equations, while the 'strong' enforcement of the constraints leads to the 'elastic-net type' algorithm with $O(N)$ variables and equations.

Taking the way the constraints are enforced as the basic characteristic of the elastic net, we have derived a novel algorithm of the elastic-net type. This is an example of what we mean by 'generalising' the elastic net. The new algorithm has the positive thermodynamic ingredients of the elastic net, but avoids an intrinsic problem of the original algorithm. It has excellent scaling properties requiring just N (vector) variables for any instance of the problem. Its time complexity, the correct choice of parameters K and α, and the overall performance remains to be investigated computationally [18]. At the same time it would be interesting to study more seriously the scaling properties of the original elastic net algorithm, in particular as one, for fixed N, deforms the initial distribution of the cities by separating groups of the cities from the rest. This should give some more quantitative idea about limitations of the original elastic-net algorithm.

Many problems in early vision—shape from shading, edge detection, motion analysis, surface interpolation, etc—are conveniently formulated in terms of minimising a cost function which usually has an underlying physical interpretation, and depends on certain set of general ('analogue') coordinates [19,20]. One of the crucial issues there, is the enormous computational complexity of the problems and the need to 'solve' them in real time; another is the reduction of the noisy visual data to stable descriptions. The approach presented in this paper may have important applications to these problems. On the one hand, it may lead to better theoretical foundation—in the sense of a physical theory—of some of the algorithms used in early vision; on the other, it may lead to 'discovery' of new and powerful computational models for early vision, which would be manifestly 'neural' or manifestly 'analogue' , depending on the representation used to describe the same physical computation [21].

Acknowledgments

I thank G Fox, W Furmanski, and J J Hopfield for helpful conversations and for a critical reading of the manuscript. I also thank P Hipes for proofreading of the manuscript and K Rose for checking some of the equations. This work is supported in part by the DOE under Grant no DE-AC03–81ER40050 and DE-FG03–85ER25009, the Office of the Program Manager of the Joint Tactical Fusion Office, and the NSF under Grant no EET-8700064.

References

[1] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci. USA* **79** 2554–8; 1984 Neurons with graded response have collective computational properties like those of two-state neurons *Proc. Natl Acad. Sci. USA* **81** 3088–92; 1987 Learning algorithms and probability distributions in feed-forward and feed-back networks *Proc. Natl Acad. Sci. USA* **84** 8429–33

Peretto P 1984 Collective properties of neural networks: A statistical physics approach *Biol. Cybern.* **50** 51–62

Amit D J, Gutfreund H and Sompolinsky H 1985 Spin-glass models of neural networks *Phys. Rev.* A **32** 1007

[2] Hopfield J J and Tank D W 1985 Neural computation of decisions in optimisation problems *Biol. Cybern.* **52** 141–52

[3] Lawler E L, Lenstra J K, Rinnooy Kan A H G and Shmoys D B (eds) 1984 *Travelling Salesman Problem* (New York: Wiley)

Garey M R and Johnson D S 1987 *Computers and Intractability* (San Francisco: Freeman)

Johnston D 1987 More approaches to the travelling salesman guide *Nature* **330** 525

[4] Hopfield J J and Tank D W 1986 Simple 'neural' optimisation networks: An A/D converter, signal decision circuit, and a linear programming circuit *IEEE Trans. Circuits Syst.* **CS-33** 533

For some applications in early vision see, for example:

Koch C, Marroquin J and Yuille A 1986 Analogue 'neural' networks in early vision *Proc. Natl Acad. Sci. USA* **83** 4263–7

Grzywacz N M and Yuille A 1988 Massively parallel implementations of theories for apparent motion *Spatial Vision* **3** 15–44

For other applications see, for example:

Anderson D (ed) 1987 *Neural Information Processing Systems* (New York: American Institute of Physics)

[5] Fox G and Furmanski W 1988 The physical structure of concurrent problems and concurrent computers *Phil. Trans. R. Soc.* A **326** 411–44; 1987 *Report* CalTech C3P-493, California Institute of Technology; 1988 *The Third Conference on Hypercube Concurrent Computers and Applications, vol* **1** ed G C Fox pp 241–78, 285–305 (New York: ACM Press)

Fox G, Furmanski W, Ho A, Koller J, Simic P and Wong I 1989 Neural networks and dynamic complex systems *Report* CalTech C3P-695 (presented at SCS Eastern Conference, Florida, 1989)

[6] Wilson G V and Pawley G S 1988 On the stability of the travelling salesman problem algorithm of Hopfield and Tank *Biol. Cybern.* **58** 63–70

[7] Hinton G E and Sejnowski T J 1983 Optimal perceptual inference *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, Washington, DC, June 1983* (Piscataway, NJ: IEEE) pp 448–53

Bertero M, Poggio T and Torre V 1988 Ill-posed problems in early vision *Proc. IEEE, August 1988* ed H Li and J R Kender (Piscataway, NJ: IEEE)

Terzopoulos D 1984 *Multi-resolution Image Processing and Analysis* ed A Rosenfeld (Berlin: Springer)

[8] Mead C 1988 *Analogue VLSI and Neural Systems* (Reading, MA: Addison-Wesley)

[9] Rumelhart D, Hinton G and Williams R 1986 Learning internal representations by error propagation *Parallel Distributed Processing* vol **1**: *Foundations* (Cambridge, MA: MIT Press)

Le Cun Y 19895 A learning procedure for asymmetric threshold network *Proc. Cognitiva* **85** 599–604

[10] Minski M and Papert S 1988 *Perceptrons* (expanded edition) (Cambridge, MA: MIT Press)

[11] Durbin R and Willshaw D 1987 An analogue approach to the travelling salesman problem using an elastic net method *Nature* **326** 689–91

[12] Willshaw D J and Von der Malsburg Ch 1979 A marker induction mechanism for the establishment of ordered neural mappings: Its application to the retinotectal problem *Phil. Trans. R. Soc.* B **287** 203–43

[13] Gaze R M 1970 *The Formation of Nerve Connections* (New York: Academic)

Cowan W M and Hunt R K 1986 *The Molecular Basis of Neural Development* eds G M Edelman, W E Gall and W M Cowan (New York: Wiley) pp 389–428

[14] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 Optimization by simulated annealing *Science* **220** 671

[15] Simic P D 1989 *Preprint* CalTech

[16] This particular form of the second term in equation (25) was suggested by A Yuille; see the comment in [11].

[17] Peterson C and Anderson J 1987 A mean field theory learning algorithm for neural networks *Complex Systems* **1** 995–1019

[18] Simic P 1989 Preprint CalTech

[19] For a recent survey see papers and references in:

Li H and Kender J R (eds) 1988 *Proc. IEEE, August 1988* (Piscataway, NJ: IEEE)

Yuille A, Energy functions for early vision and analog networks *A I Memo* 987 MIT

[20] Blake A and Zisserman A 1987 *Visual Reconstruction* (Cambridge, MA: MIT Press)

[21] Simic P D 1989 *Preprint* CalTech