NEURAL NETWORK FUNDAMENTALS WITH GRAPHS, ALGORITHMS, AND APPLICATIONS

N. K. Bose
HRB-Systems Professor of Electrical Engineering
The Pennsylvania State University, University Park

P. Liang
Associate Professor of Electrical Engineering
University of California, Riverside

McGraw-Hill, Inc.
New York St. Louis San Francisco Auckland Bogotá
Caracas Lisbon London Madrid Mexico City Milan Montreal
New Delhi San Juan Singapore Sydney Tokyo Toronto

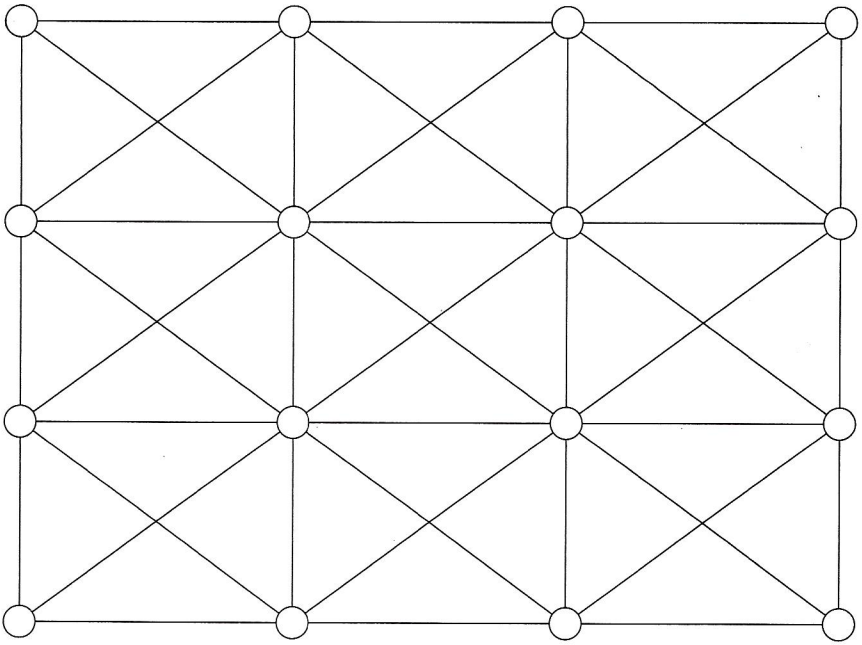

FIGURE 8.9
A two-dimensional cellular neural network structure.

to respond in any way until the association is propagated to units in its neighborhood. The speed of associative recall will be much slower than in the completely connected network. Since a CNN is a strongly diluted Hopfield network, its memory capacity is severely limited.

The biggest advantage of the CNN is that its interconnections are local, making it well suited for VLSI implementation of large networks. A two-dimensional CNN can be viewed as a parallel nonlinear two-dimensional filter, which is highly suited for image processing applications. Therefore, the most popular application for cellular neural networks has been in image processing, essentially because of their analog features and sparse connections, which are conducive to real-time processing. Cellular neural networks for noise removal, shape extraction, edge detection, and Chinese character recognition have also been demonstrated [80].

Although only symmetric cellular neural networks have been investigated, asymmetric CNNs may also be possible for applications involving temporal variations of spatial patterns. Analysis of the dynamic behavior would become more difficult in that case (see Section 8.5).

8.3 SEEKING THE GLOBAL MINIMUM: SIMULATED ANNEALING

In the Hopfield network, many local minima are used to store information. Its ability to function as an associative memory depends on reliable recall of the local minima.

However, in many cases, the network is required to reach the global minimum, such as in optimization tasks. A Hopfield network follows a gradient descent rule. Once it reaches a local minimum, it is stuck there. To find the global minimum of a function using only local information, some randomness must be added to the gradient descent rule to increase the chance of hitting the global minimum. Next, we introduce the simulated annealing method for optimization.

8.3.1 Simulated Annealing in Optimization

Many problems in science and engineering can be formulated as optimization problems. There are two basic strategies for heuristic search for a minimum of a function: divide-and-conquer and iterative improvement. The first strategy is problem-specific, because an effective decomposition of the problem depends on the nature of the problem itself. Iterative improvement, commonly known as gradient descent, is a general method. Its difficulty is that it often gets stuck at a local minimum. *Simulated annealing* (SA) [189] is a method that introduces randomness to allow the system to jump out of a local minimum. The method draws an analogy between optimization of a cost function in a large number of variables with minimization of an energy function in statistical mechanics.

Statistical mechanics is the central discipline of condensed-matter physics. It offers a body of methods for analyzing aggregate properties of physical systems with a very large number of particles. For a given physical system with a set of possible states or configurations $\{\alpha_i\}$, define an energy function E. The total energy of the system at configuration α_i is denoted as E_{α_i}. If the system is at a temperature T greater than absolute zero (zero Kelvin, i.e., -273 degrees Celsius), there will be *thermal fluctuations,* which are state transitions that cause the configuration energy to either increase or decrease slightly. When the system only fluctuates around constant average values without drifting far away, it is said to have reached *thermal equilibrium.*

A fundamental result from statistical mechanics shows that in thermal equilibrium the probability P_{α_i} of finding the system at a configuration α_i is given by the *Boltzmann-Gibbs distribution*

$$P_{\alpha_i} = \frac{1}{Z} \exp\left\{\frac{-E_{\alpha_i}}{k_B T}\right\}, \tag{8.55}$$

where T is the temperature in Kelvin and k_B is the *Boltzmann constant.* Since the "temperature" in simulated annealing and neural networks is not related to the physical temperature at all, we will normalize the Boltzmann constant to 1 in this book. The denominator Z in the preceding equation is called the *partition function* and is given as

$$Z = \sum_i \exp\left\{\frac{-E_{\alpha_i}}{T}\right\}. \tag{8.56}$$

The Boltzmann-Gibbs distribution holds for systems with a very large number of particles, unless the system is constrained so that it cannot explore all its possible states.

Suppose there are two configurations α_i and α_j. Then the ratio of the probabilities of the two configurations is given by

$$\frac{P_{\alpha_i}}{P_{\alpha_j}} = \exp\left\{-\frac{(E_{\alpha_i} - E_{\alpha_j})}{T}\right\}. \tag{8.57}$$

If the temperature is very high, the probabilities of finding the system at high and low energy levels are roughly the same, because the large denominator T makes the difference in probabilities between different configurations small. That is, due to thermal agitation, the system is almost equally likely to transit to states that increase its energy as to states that decrease the energy. When the temperature is reduced, the probability of finding the system at a high energy state decreases. The system is more likely to assume a configuration with low energy. When the temperature is very low, the probability of the system assuming a configuration with an energy higher than the lowest level becomes negligibly small, and the system is said to be *frozen* at the configuration with the lowest energy. However, the system should not be set at a low temperature for finding the minimum energy configuration. Experiments that seek the low-temperature state of a material are done by careful *annealing*. In annealing, the substance is first melted at high temperature. The temperature is then lowered slowly. The substance is kept for a long time at temperatures in the vicinity of the freezing point. An example is the growing of a single crystal from a melt. If the temperature drops very rapidly, the resulting crystal will have many defects, or the substance may form a *glass,* which has only metastable, locally optimal structures and no crystalline order.

Iterative optimization is much like the microscopic transitions of particles in statistical mechanics, which seeks out the system configuration with the minimum energy. The cost function plays the role of the energy function. The many variables play the roles of the particles. Accepting only transitions that lower the cost function is much like extremely rapid quenching from high temperature to $T = 0$. It is not surprising that the resulting solutions are often local minima.

The SA method introduces a stochastic ingredient into the gradient descent rule using the parameter T and an annealing procedure. The resulting algorithm allows both downhill and uphill state transitions. The probability of an uphill motion is controlled by T. This procedure was first introduced by Metropolis et al. [256] in 1953 for computer simulation of a collection of atoms in equilibrium at a given temperature. The stochastic state transitions can be described by a set of probabilities $P(\alpha_i \rightarrow \alpha_j)$. However, such a set of transition probabilities may not lead to thermal equilibrium. Instead, it may lead to a limit cycle or chaotic behavior. We are mainly interested in thermal equilibrium, because analysis is much simplified at equilibrium. A sufficient condition on $P(\alpha_i \rightarrow \alpha_j)$ that guarantees equilibrium is that the probability of transitions from α_i to α_j equals that from α_j to α_i on average; i.e.,

$$P_{\alpha_i} P(\alpha_i \rightarrow \alpha_j) = P_{\alpha_j} P(\alpha_j \rightarrow \alpha_i). \tag{8.58}$$

If this condition is satisfied, the system will reach equilibrium according to the Boltzmann-Gibbs distribution, and

$$\frac{P(\alpha_i \rightarrow \alpha_j)}{P(\alpha_j \rightarrow \alpha_i)} = \frac{P_{\alpha_j}}{P_{\alpha_i}} = \exp\left(-\frac{E_{\alpha_j} - E_{\alpha_i}}{T}\right) = \exp\left(-\frac{\Delta E}{T}\right), \tag{8.59}$$

where $\Delta E = E_{\alpha_j} - E_{\alpha_i}$.

The Metropolis algorithm adopted in ref. [189] satisfies the condition in Eq. (8.58). It uses the following transition probability:

$$P(\alpha_i \rightarrow \alpha_j) = \begin{cases} 1 & \text{if } \Delta E < 0; \\ \exp(-\Delta E/T) & \text{otherwise.} \end{cases} \tag{8.60}$$

At each step a variable is given a small random displacement; in other words, the state is changed from α_i to α_j. The resulting change in energy ΔE is computed. If the energy is reduced (i.e., $\Delta E < 0$), the transition is accepted, and the new configuration is used as the starting point of the next step. If $\Delta E \geq 0$, the transition from a lower to a higher energy state is accepted with a probability $P(\Delta E) \triangleq \exp(-\Delta E/T)$. This can be realized by generating a random number uniformly distributed in the interval (0, 1). If the random number is less than $P(\Delta E)$, the transition is accepted; otherwise, the original state is used as the starting point for the next step. Repeating the procedure many times simulates the thermal motion of atoms in thermal contact with a heat bath at temperature T, leading to a Boltzmann-Gibbs distribution. Note that the Metropolis simulation can be carried out in parallel; i.e., many transitions can occur simultaneously.

In summary, the SA process consists of the following four components:

1. A concise description of the configurations (the states) of the system.
2. A random number generator, based on which random state transitions are selected.
3. A quantitative cost function, capturing the criteria and constraints of the optimization problem.
4. An annealing schedule, which consists of a sequence of temperatures and the number of steps at each temperature. An annealing procedure normally consists of melting the system at high temperature and then lowering the temperature by slow stages to the freezing temperature. At each temperature the Metropolis simulation must proceed long enough for the system to reach a steady state (i.e., thermal equilibrium). There is no fixed annealing schedule that will work for any problem. A trial-and-error process is required to identify a satisfactory annealing procedure.

The basic principle of SA is to assign any state a nonzero probability at nonzero temperature. Arrival at the global minimum is not guaranteed in *finite time*. Simulated annealing actually employs an adaptive divide-and-conquer strategy. At high temperatures the energy differences among the states are reduced. The fine details of the energy differences among the states start to have an effect on the state transitions only at low temperature. This corresponds to a coarse search at high temperatures in the global topography and a fine search at low temperatures in the local terrain

around the state into which the system finally settles. An important point made by Kirkpatrick et al. [189] is that an average-behavior analysis is more useful in assessing the value of a heuristic for optimization than is the traditional worst-case analysis. A worst-case analysis simply assigns a blanket categorization of NP-completeness to many optimization problems. It appears that in a large number of difficult problems that arise in engineering, the most probable behavior of the solution to an optimization problem is more useful than the worst-case performance evaluation.

Statistical mechanics provides methods for computing macroscopic properties from microscopic averages. An ensemble average can be obtained using the partition function Z. The average thermal energy $\langle x \rangle$ of a variable x is given by

$$\langle x \rangle = \sum_i x_{\alpha_i} P_{\alpha_i}, \tag{8.61}$$

where P_{α_i} is the probability of finding x at x_{α_i}. For example, the *average thermal energy* of a system is defined as

$$\langle E(T) \rangle = \sum_i E_{\alpha_i} P_{\alpha_i}. \tag{8.62}$$

Define the *free energy* of a system as

$$F(T) = -T \ln Z. \tag{8.63}$$

It can be easily shown that

$$F(T) = \langle E(T) \rangle - TS, \tag{8.64}$$

where S is the *entropy,* defined as

$$S = -\frac{\partial F}{\partial T} = -\sum_i P_{\alpha_i} \ln P_{\alpha_i}. \tag{8.65}$$

Then the *average energy* can be expressed as

$$\langle E(T) \rangle = \frac{\partial (F(T)/T)}{\partial (1/T)}. \tag{8.66}$$

The rate of change of the average energy with respect to the temperature is given by

$$C(T) = \frac{d\langle E(T) \rangle}{dT} = \frac{[\langle E^2(T) \rangle - \langle E(T) \rangle^2]}{T^2}. \tag{8.67}$$

In statistical mechanics $C(T)$ is called the *specific heat.* A large value of C signals a change in the state of the order in a system. It can be used in the simulated annealing process to indicate that freezing has begun, implying that very slow cooling is required.

The analogy between the processes of cooling a fluid and optimization may fail in one important respect. In ideal fluids all the atoms are similar and the ground state is a regular crystal. An optimization problem may contain many distinct, noninterchangeable elements, and optimization is subject to many conflicting constraints. Re-

search in condensed-matter physics on systems with quenched-in randomness (i.e., not all atoms are alike) may lend some insight. A feature of such systems, termed *frustration,* is that interactions favoring different and incompatible kinds of ordering may be simultaneously present. The magnetic alloys known as *spin glasses* are among the best-understood examples of the frustration phenomenon [353, 369]. The physical properties of spin glasses at low temperatures may provide some guide to the understanding of optimization problems subject to conflicting constraints [189].

8.3.2 Stochastic Networks: Applying Simulated Annealing to Hopfield Networks

The idea of simulated annealing in searching for the global minimum is generally applicable. It can be applied to search for the global minimum in a discrete Hopfield network. To apply simulated annealing, the transition rule of the units, Eq. (8.4), is made stochastic according to a probability distribution. The transition probability should be a function of the energy change ΔE and the "temperature" parameter T.

Any transition probability that satisfies the condition in Eq. (8.58) can be applied; it will lead to a Boltzmann-Gibbs distribution of the states at thermal equilibrium. A network of TLUs whose state transition rule is probabilistic, leading to Boltzmann-Gibbs distribution at equilibrium, is sometimes called a Boltzmann machine (BM) [4]. However, the term *Boltzmann machine* is more often used to denote a class of learning networks to be discussed in the next section.

To allow parallel computation of the stochastic transition rules, ΔE should be computable locally. It should be computed using only states of the presynaptic and postsynaptic units. Because the connection weight matrix is symmetric, the energy function in Eq. (8.5) for the discrete Hopfield network can be used. When y_i changes from -1 to $+1$ and $w_{ii} = 0$ in Eq. (8.7), the energy change is

$$\Delta E_i = E(y_i = +1) - E(y_i = -1) = -2\left(\sum_j w_{ij} y_j - \theta_i\right).$$

This quantity can be computed locally. The following transition probabilities may be used.

$$P(y_i = +1) = \frac{1}{1 + \exp(\Delta E_i / T)} = \frac{1}{1 + \exp[-2(\sum_j w_{ij} y_j - \theta_i)/T]} \tag{8.68}$$

and

$$P(y_i = -1) = 1 - \frac{1}{1 + \exp(\Delta E_i / T)}. \tag{8.69}$$

It can be verified that this choice of the transition probabilities satisfies the condition in Eq. (8.59). Therefore, the global state will follow the Boltzmann-Gibbs distribution at thermal equilibrium. An effective annealing schedule is usually determined empirically.