Chapter 3. Statistical Mechanical Networks

3.1 Statistical Mechanics

Statistical mechanics deals with systems containing a large number of particles. A collection of identical systems is called an ensemble, and is characterized by the average of its component systems. For example, if \( P_r \) is the probability that a system has an energy \( E_r \), then the average energy of the ensemble of such systems is

\[
\langle E \rangle = \sum_r P_r E_r
\]

The probability that the system is at a certain energy \( E_r \) is proportional to an exponential factor

\[
P_r = \frac{e^{-\beta E_r}}{Z}
\]

where \( \beta \) is a parameter that depends on the temperature. The sum of all such probabilities must be 1, \( \sum_r P_r = 1 \), thus the normalization constant is given as

\[
Z = \sum_r e^{-\beta E_r}
\]

\( Z \) is called the partition function in statistical physics and contains the known information about the system under study. The exponential factor \( -\beta E_r \) is called the Boltzmann factor and the probability distribution \( P_r \) is called the Boltzmann distribution. Ensembles whose properties follow the Boltzmann distribution are called canonical ensembles. The factor \( \beta \) is related to the absolute temperature \( T \)

\[
\beta = (k_B T)^{-1}
\]

where \( k_B \) is a constant known as the Boltzmann constant.

In a physical system, entropy is defined as

\[
S = -k_B \sum_r P_r \ln P_r
\]

This definition of entropy differs from the information-theoretic entropy by only a constant multiplier.

3.2 Ising Models

Definition. The Ising model is a prototypical model of cooperative phenomena. Consider a one-dimensional array of atoms on a regular lattice. The spin \( x_i \) on site \( i \) can assume one of two values, either +1 or -1, depending on whether it is aligned parallel or antiparallel to an external magnetic field \( H \).
The energy of a state $x$ is

$$E(x; J, H) = -\frac{1}{2} \sum_{i,j} J_{ij} x_i x_j - H \sum_i x_i$$

where $H$ is the external magnetic field and $J_{ij}$ are the strengths of interaction between electronic spins $i$ and $j$. $J_{ij} = J$ for $(i, j) \in N$, and $J_{ij} = J$, otherwise.

$J > 0$ then ferromagnetic, $J < 0$ then antiferromagnetic.

At equilibrium at temperature $T$, the probability that the state is $x$ is

$$P(x|\beta, J, H) = \frac{1}{Z(\beta, J, H)} \exp \left[-\beta E(x; J, H)\right]$$

where $\beta = 1/k_B T$ and

$$Z(\beta, J, H) = \sum_x \exp \left[-\beta E(x; J, H)\right]$$

### 3.3 Hopfield Networks

A Hopfield network is a fully connected recurrent network. It can be used as an associative memory. There are two different types of Hopfield network: binary (discrete) and continuous.
Figure 2. Architecture of a Hopfield Network

Binary Hopfield Network

- Activation function
  \[ a_i = \sum_j w_{ij} x_j \]
  \[ x_i = \theta(a_i) = \begin{cases} 1 & a_i \geq 0 \\ -1 & a_i < 0 \end{cases} \]

- Learning rule
  \[ w_{ij} = \rho \sum_n x_i^{(n)} x_j^{(n)} \]
  using the training data \( D = \{ x^{(n)} \mid n = 1, \ldots, N \} \)

Continuous Hopfield Network

Activation function

\[ a_i = \sum_j w_{ij} x_j \]
\[ x_i = \tanh(a_i) = \frac{1}{1 + e^{-2a_i}} \]

Convergence of Hopfield Network

Energy Function

\[ E(x; J, H) = \frac{1}{2} \sum_{ij} f_{ij} x_i x_j - \sum_i h_i x_i \]

**Homework:** Show that the energy \( E(x; J, H) \) of the Hopfield network monotonically decreases as the
activations of the units are updated. In other words, show that the energy function of the Hopfield network is a Liapunov function.

Figure 3. Hopfield network as an associative memory. The stored, original pattern can be reconstructed from a noisy pattern by recall.

3.4 Boltzmann Machines

Definition. The Boltzmann machines are stochastic Hopfield networks. The energy of a state $x$ is
Figure 4. Architecture of a Boltzmann Machine. A fully connected recurrent network with hidden units. The units between the visible and hidden units are also fully connected.

Activation function

\[ a_i = \sum_j w_{ij} x_j \]
\[ x_i = \begin{cases} 1 & \text{with prob } p_i = \frac{1}{1 + e^{-a_i/T}} \\ 0 & \text{otherwise} \end{cases} \]

Energy function

\[ E = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} w_{ij} x_i x_j \]

Learning rule

\[ \Delta w_{ij} = \epsilon (p_{ij}^+ - p_{ij}^-) \]

where \( p_{ij}^+ = \sum_{a,b} P^+ (V_a \land H_b) x_i^{ab} x_j^{ab} \) and \( p_{ij}^- = \sum_{a,b} P^- (V_a \land H_b) x_i^{ab} x_j^{ab} \)

(definitions of the symbols are given below)

Deriving the learning rule for Boltzmann machines

Objective function to minimize

\[ G = \sum_{i=1}^{q} P_i \log_2 \frac{P_{1i}}{P_{2i}} \]

\( P^+ (V_a \land H_b) \): probability that vector \( V_a \) is clamped to the visible units and that vector \( H_b \) appears on
The hidden units

\[ P^+(V_a) = \sum_b P^+(V_a \land H_b) \]

The total energy of the system with \( V_a \) on the visible units and \( H_b \) on the hidden units is

\[ E_{ab} = -\sum_{i<j} w_{ij} x_i^{ab} x_j^{ab} \quad (3.2) \]

where \( x_i^{ab} \) can refer to either a visible unit or a hidden unit.

With none of the visible units clamped, the probability that \( V_a \) will appear on the visible units is given by

\[ P^-(V_a) = \sum_b P^-(V_a \land H_b) \]

\[ P^-(V_a \land H_b) = \frac{e^{-E_{ab}/T}}{\sum_{m,n} e^{-E_{mn}/T}} \quad (3.3) \]

\[ = \frac{e^{-E_{ab}/T}}{Z} \]

Then,

\[ P^-(V_a) = \frac{\sum_b e^{-E_{ab}/T}}{\sum_{m,n} e^{-E_{mn}/T}} \quad (3.6) \]

The explicit functional form of \( G \) now becomes

\[ G = \sum_a P^+(V_a) \ln \frac{P^+(V_a)}{P^-(V_a)} \]

Differentiating \( G \) gives

\[ \frac{\partial G}{\partial w_{ij}} = -\sum_a \frac{P^+(V_a)}{P^-(V_a)} \frac{\partial P^-(V_a)}{\partial w_{ij}} \quad (3.7) \]

Notice that the \( P^+(V_a) \) are independent of \( w_{ij} \) because the visible units are clamped to and do not vary with changes in the \( w_{ij} \).

From Eq. (3.6),

\[ \frac{\partial P^-(V_a)}{w_{ij}} = -\frac{1}{T} \sum_b e^{-E_{ab}/T} \frac{\partial E_{ab}}{w_{ij}} - \sum_b \frac{e^{-E_{ab}/T}}{Z^2} \frac{\partial Z}{w_{ij}} \quad (3.8) \]

The derivative of the energy function is

\[ \frac{\partial E_{ab}}{w_{ij}} = -x_i^{ab} x_j^{ab} \quad (3.9) \]

and the derivative of the partition function is

\[ \frac{\partial Z}{w_{ij}} = \sum_{m,n} \left( -\frac{1}{T^2} \frac{\partial E_{mn}}{w_{ij}} e^{-E_{mn}/T} \right) \]

\[ = \frac{1}{T} \sum_{m,n} x_i^{mn} x_j^{mn} e^{-E_{mn}/T} \quad (3.10) \]

Substituting Eqs. (3.9) and (3.10) into Eq. (3.8) yields

\[ \frac{\partial P^-(V_a)}{w_{ij}} = \frac{1}{T} \sum_b P^-(V_a \land H_b) x_i^{ab} x_j^{ab} - \frac{P^-(V_a)}{T} \frac{1}{T} \sum_{m,n} P^-(V_m \land H_n) x_i^{mn} x_j^{mn} \quad (3.11) \]
where we have made use of the definition of \( P^-(V_a \land H_b) \) and the definition of \( P^-(V_a) \). Eq. (3.11) can now be substituted into Eq. (3.7) to give

\[
\frac{\partial G}{\partial w_{ij}} = -\frac{1}{T} \sum_{a,b} \frac{P^+(V_a)}{P^-(V_a)} P^-(V_a \land H_b) x_i^{ab} x_j^{ab} + \frac{\sum_a P^+(V_a)}{T} \sum_{mn} P^-(V_m \land H_a) x_i^{mn} x_j^{mn} \tag{3.12}
\]

To simplify this equation, we use the followings:

\[
\sum_a P^+(V_a) = 1
\]

\[
P^+(V_a \land H_b) = P^+(H_b | V_a) P^+(V_a)
\]

\[
P^-(V_a \land H_b) = P^-(H_b | V_a) P^-(V_a)
\]

If \( V_a \) is on the visible layer, then the probability that \( H_a \) will occur on the hidden layer should not depend on whether \( V_a \) got there by being clamped to that state or by free-running to that state. Therefore, it must be true that

\[
P^+(H_a | V_a) = P^-(H_a | V_a)
\]

Then

\[
\frac{P^-(V_a \land H_b)}{P^+(V_a \land H_b)} = \frac{P^-(V_a)}{P^+(V_a)}
\]

and

\[
P^-(V_a \land H_b) \frac{P^+(V_a)}{P^+(V_a)} = P^+(V_a \land H_b)
\]

Using the results, we can write

\[
\frac{\partial G}{\partial w_{ij}} = \frac{1}{T} (p_{ij}^* - p_{ij})
\]

where

\[
p_{ij}^* = \sum_{a,b} P^-(V_a \land H_b) x_i^{ab} x_j^{ab} \tag{3.13}
\]

and

\[
p_{ij}^* = \sum_{a,b} P^+(V_a \land H_b) x_i^{ab} x_j^{ab} \tag{3.14}
\]

Weight updates are computed according to

\[
\Delta w_{ij} = \epsilon (p_{ij}^* - p_{ij}) \tag{3.15}
\]

where \( \epsilon \) is a constant learning rate and

\( p_{ij}^* \): co-occurrence probability when the \( V_a \) patterns are being clamped on the visible units.\( p_{ij}^* \): co-occurrence probability when the network is free-running.

**Training a Boltzmann Machine by Simulated Annealing**

1. **Clamp** the outputs of the known visible units to the input vector \( \mathbf{x} \).
2. **Assign** all unknown visible units, and all hidden units, random output values from \( \{0, 1\} \).
3. **Select** a unit, \( x_k \), at random and calculate its net-input value, \( \text{net}_k \), (see below for \( \Delta E_k = \text{net}_k \)).
4. Regardless of the current value of the input, assign the output value, \( x_k = 1 \), with probability

\[
p_k = \frac{1}{1 + e^{-\frac{\text{net}_k}{\epsilon}}}
\]

5. Repeat steps 3 and 4 until all units have had some probability of being selected for update. This
number of unit-updates defines a processing cycle.

6. Repeat step 5 for several processing cycles, until thermal equilibrium has been reached at the given temperature, $T$.

7. Lower the temperature, $T$, and repeat steps 3 through 7.

Derivation of $\Delta E_k$:

$$E = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} x_i x_j$$

The energy difference between the system with $x_k = 0$ and $x_k = 1$ is given by

$$\Delta E_k = (E_{k=0} - E_{k=1}) = \sum_{j=1, j\neq k}^{n} w_{kj} x_j = \text{net}_k$$