Markov Chains and MCMC

Markov chains

- Let $S = \{1, 2, \ldots, N\}$ be a finite set consisting of N states.
- A **Markov chain** $Y_0, Y_1, Y_2, \ldots$ is a sequence of random variables, with $Y_t \in S$ for all points in time $t \in \mathbb{N}$, that satisfies the **Markov property**, namely, given the present state, the future and past states are independent:

$$\Pr(Y_{n+1} = x | Y_0 = y_0, \ldots, Y_n = y_n) = \Pr(Y_{n+1} = x | Y_n = y_n)$$

- A Markov chain is **homogeneous** if for all $n \geq 1$:

$$\Pr(Y_{n+1} = j | Y_n = i) = \Pr(Y_n = j | Y_{n-1} = i)$$

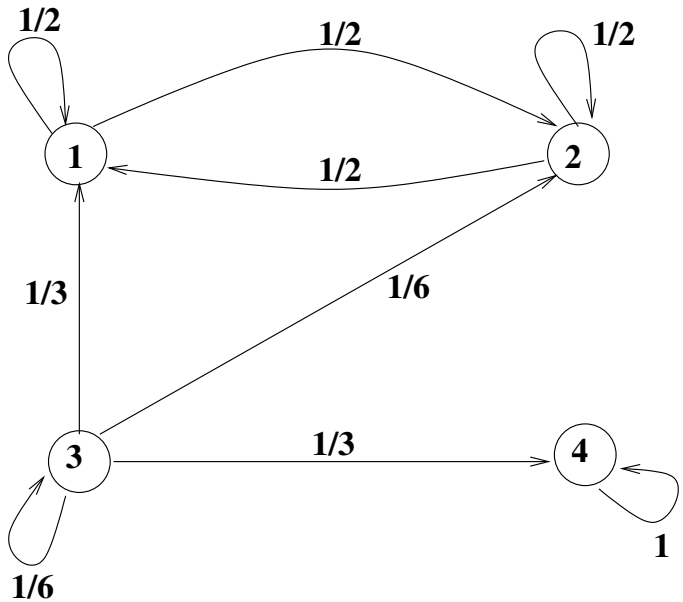
  i.e., transitional probabilities are time independent.
- There is an $N \times N$ **transition** matrix P such that for all i, j, n we have: $P_{ij} := \Pr(Y_{n+1} = j | Y_n = i) \geq 0 \ \bigwedge \ \sum_{j=1}^{N} P_{ij} = 1$.
- Any matrix $P \in \mathbb{R}^{n \times n}$ with $P_{ij} \geq 0$ and $\sum_{j=1}^{N} P_{ij} = 1$, for $1 \leq i, j \leq N$, is called a **stochastic** matrix.

Example

- Let $S = \{1, 2, 3, 4\}$ and

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/3 & 1/6 & 1/6 & 1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

- Here is the **graphical representation** of P, i.e., only the edges with $P_{ij} > 0$ are drawn:



Markov chains as Dynamical Systems

- A homogeneous Markov chain defines a dynamical system:
- Let $X = \{x \in \mathbb{R}^N : \ x_i \geq 0, \ \sum_{i=1}^N x_i = 1\}$: the space of probability vectors over S.
- $d(x, y) := \frac{1}{2} \sum_{k=1}^N |x_k - y_k| = \frac{1}{2} \|x - y\|_1$ is a metric on X.
- $P : X \to X$ defines a linear map by right multiplication.
- $P : x \mapsto xP$, i.e., $(xP)_n = \sum_{k=1}^N x_k P_{kn}$, is well-defined:
- $\sum_{n=1}^N (xP)_n = \sum_{n=1}^N \sum_{k=1}^N x_k P_{kn} = \sum_{k=1}^N \sum_{n=1}^N x_k P_{kn} = \sum_{k=1}^N x_k \sum_{n=1}^N P_{kn} = \sum_{k=1}^N x_k = 1$
- The n-step transition matrix is simply $P^{(n)} = P^n$.
- If $x \in X$ then the probability vector over S evolves as

$$\text{orbit of } x : \quad x, xP, xP^2, \ldots, xP^n, \ldots$$

- Interested in the long term behaviour of orbits.

Communicating Classes

- A state j is **accessible** from i, denoted by $i \to j$, if there exists $n \geq 0$ with $P_{ij}^{(n)} > 0$, i.e., j can be reached from i after a finite number of steps.
- Two states i and j **communicate**, denoted by $i \leftrightarrow j$, if $i \to j$ and $j \to i$.
- Note that for any state i, we have $i \leftrightarrow i$. Why?
- Communication is an equivalence relation, which induces **communicating classes**.
- By analysing the transition matrix, we determine the communication classes.
- Find the communicating classes of P in Equation (1).
- A Markov chain is **irreducible** if it has a single communicating class, i.e., if any state can be reached from any state.

Aperiodic Markov Chains

- The **period** of a state i is defined as $d(i) := \gcd\{n \geq 1 : P_{ii}^{(n)} > 0\}$.
- Here, gcd denotes the greatest common divisor.
- **Example:** Consider a Markov Chain with transition matrix,

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{2}$$

  All diagonal elements of P^{2n} are positive and those of P^{2n+1} are all zero. Thus, every state has period 2.
- If $d(i) = 1$ then i is said to be aperiodic.
- If all states are aperiodic then the chain is called **aperiodic**.

Irreducibility and Periodicity: basic properties

- If a Markov chain is irreducible then all its states have the same period.
- An irreducible Markov chain is aperiodic if there is a state i with $P_{ii} > 0$.
- An irreducible Markov chain is aperiodic iff there exists $n \geq 1$ such that $\forall i, j . P_{ij}^{(n)} > 0$.
- If P is the transition matrix of an irreducible Markov chain and $0 < a < 1$, then $aI + (1 - a)P$ is the transition matrix of an irreducible Markov chain, where I is the $N \times N$ identity matrix.
- Thus, by choosing a small $a > 0$, we can allow a small self transition to make the Markov chain aperiodic.
- Moreover, we will see later that P and $aI + (1 - a)P$ have the same stationary distribution.

Stationary distribution

- A probability vector $\pi \in X$ is a **stationary distribution** of a Markov chain if $\pi P = \pi$.
- Any transition matrix has an eigenvalue equal to 1. Why?
- Thus, find the stationary distributions by solving the eigenvalue problem $\pi P = \pi$ or $P^T \pi^T = \pi^T$.
- **Fundamental Theorem of Markov chains.** An irreducible and aperiodic Markov chain has a unique stationary distribution π which satisfies:
  - $\lim_{n\to\infty} xP^n = \pi$ for all $x \in X$.
  - $\lim_{n\to\infty} P^n$ exists and is the matrix with all rows equal to π.
- This means that whatever our starting probability vector, the dynamics of the chain takes it to the unique stationary distribution or the steady-state of the system.
- Thus, we have an attractor π with basin X.
- **Check:** $\pi P = \pi \iff \pi(aI + (1-a)P) = \pi$, for $0 < a < 1$.

Reversible Chains and Detailed Balanced Condition

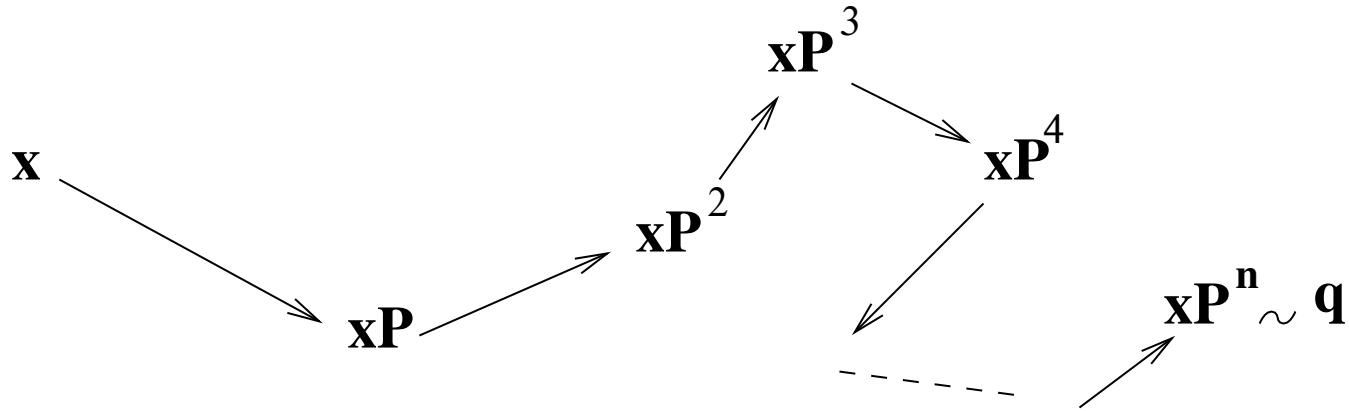
- A Markov chain is **reversible** if there is $\pi \in X$ that satisfies the **detailed balanced condition**:

$$\pi_i P_{ij} = \pi_j P_{ji}, \text{ for } 1 \leq i, j \leq N$$

- This means that
  $\Pr(Y_n = j, Y_{n-1} = i) = \Pr(Y_n = i, Y_{n-1} = j)$ when
  $\Pr(Y_{n-1} = i) = \pi_i$ for all i, i.e., the chain is **time reversible**.
- **Exercise:** If π satisfies the detailed balanced condition, then it is a stationary distribution.
- This shows that satisfying the detailed balanced condition is sufficient for π to be the stationary distribution.
- However, the detailed balanced condition is not a necessary condition for π to be a stationary distribution.

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) methods are based on the convergence of the orbit of any initial probability vector to the unique stationary distribution of an irreducible and aperiodic Markov chain.
- If we need to sample from a distribution q say on a finite state space, we construct an irreducible and aperiodic Markov chain P with unique stationary distribution $\pi = q$.
- Then, since $\lim_{n\to\infty} xP^n = \pi$, i.e. $\lim_{n\to\infty} |xP^n - \pi| = 0$ for any $x \in X$, it follows that for large n, we have $xP^n \approx \pi = q$.



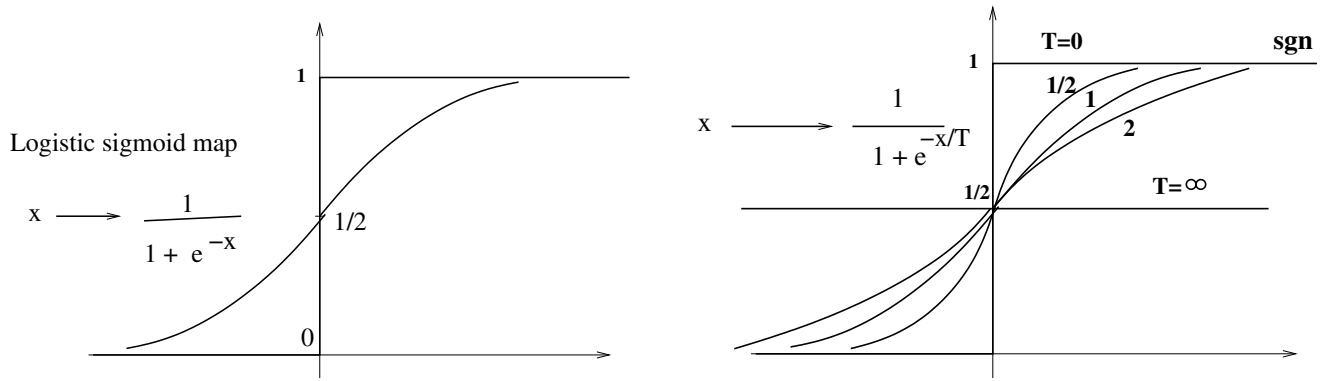
- Metropolis-Hastings algorithms such as **Gibbs sampling** used in stochastic Hopfield networks and RBMs are examples of MCMC, as we will see.

Stochastic Hopfield Networks

- Replace the deterministic updating $x_i \leftarrow \mathrm{sgn}(h_i)$, where $h_i = \sum_{j=1}^{N} w_{ij} x_j$, with an asynchronous probabilistic rule:

$$\Pr(x_i = \pm 1) = \frac{1}{1 + \exp(\mp 2h_i/T)} = \frac{1}{1 + \exp(-2h_i x_i/T)} \tag{3}$$

  where $T > 0$ is the **pseudo temperature**.
- As $T \to 0$, this is reduced to the deterministic rule.



- **Exercise.** From (3), obtain the probability of flipping:

$$\Pr(x_i \to -x_i) = \frac{1}{1 + \exp(\Delta E/T)}, \tag{4}$$

  where $\Delta E = E' - E$ is the change in energy.
- There is now some probability that the energy increases.

Stationary Distribution

- A stochastic Hopfield network can be viewed as a Markov process with 2^N states: $x_i = \pm 1$ for $1 \leq i \leq N$.
- The flipping probability (4) defines the transition matrix P.
- Irreducible: The flipping rule is applied asynchronously at all nodes, one at a time, and therefore, there is a non-zero probability of going from any state to any state.
- Aperiodic: There is a non-zero probability that at any time a state does not change.
- Thus, the network has a unique stationary distribution.
- **Exercise:** Show that the distribution

$$\pi(x) := \Pr(x) = \frac{\exp(-E(x)/T)}{Z}, \tag{5}$$

  where Z is the normalisation factor, satisfies the detailed balanced condition, i.e., it is the stationary distribution.
- Start with any configuration of the network; successively apply the flipping rule. After a large number of iterations, we have a sample from the stationary distribution.

Computing average values

- Suppose now we are interested to find not a sample from the stationary distribution π but the average value of a function $f : X \to \mathbb{R}$ on the configurations of a stochastic Hopfield network with respect to π.
- For example, we may take $f(x) = \sum_{i=1}^{N} x_i$ and want to compute $\mathbb{E}_\pi(\sum_{i=1}^{N} x_i)$, i.e., the average value of the sum of all binary node values with respect to the stationary distribution π of the stochastic network.
- As we have seen, we can find a sample of π by starting with any distribution p_0 and applying the transitional probability P a large number of times, n say, to obtain $p_0 P^n$ as a sample of π.
- We thus reasonably expect that we can find a good estimate for $\mathbb{E}_\pi(\sum_{i=1}^{N} x_i)$ by computing the expected value of $\sum_{i=1}^{N} x_i$ with respect to $p_0 P^n$, i.e., find $\mathbb{E}_{p_0 P^n}(\sum_{i=1}^{N} x_i)$.

Convergence of average value

- First recall that given a probability vector $p \in X$ and a function $f : \{1, \ldots, N\} \to \mathbb{R}$ the **average value** or **expected value** of f with respect to p is given by

$$\mathbb{E}_p(f) = \sum_{i=1}^{N} p_i f(i)$$

- If a sequence $p^{(n)} \in X$, $(n \in \mathbb{N})$ of probability vectors converges to p, then we have:

$$\mathbb{E}_{p^{(n)}}(f) \to \mathbb{E}_p(f), \qquad \text{as } n \to \infty.$$

- In fact, $\lim_{n\to\infty} p^{(n)} = p$ implies

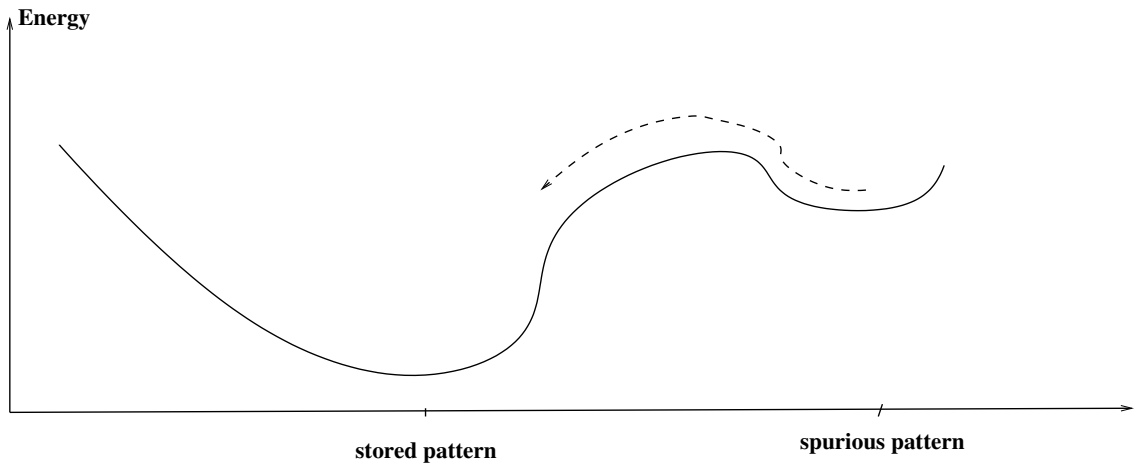
$$\lim_{n\to\infty} \|p^{(n)} - p\|_1 = \lim_{n\to\infty} \sum_{i=1}^{N} |p_i^{(n)} - p_i| = 0$$

i.e., $\lim_{n\to\infty} |p_i^{(n)} - p_i| = 0$ for each $i = 1, \ldots, N$. Thus:

$$\mathbb{E}_{p^{(n)}}(f) = \sum_{i=1}^{N} p_i^{(n)} f(i) \to \sum_{i=1}^{N} p_i f(i) = \mathbb{E}_p(f)$$

Simulated Annealing

- In optimisation problems, **Simulated Annealing** is performed in order to avert getting stuck in local minima.
- In Hopfield networks, local minima correspond to spurious patterns with higher energy than the stored patterns.
- Start from a reasonably high value of T so that the states have a good probability of jumping over from the basins of local minima to basins of stored patterns.
- Then steadily lower T so that the states gradually follow a downhill road in the energy landscape to a stable state.



Markov Random Fields I

- For applications in Machine Learning, Computer Vision, Image Processing etc., we need a generalisation of Markov chains to Markov Random Fields (MRF).
- In a Markov chain we have a sequence of random variables satisfying the Markov property that the future is independent of the past given the present.
- In a MRF we have a vector of random variables presented in an undirected graph that describes the conditional dependence and independence of any pair of the random variables given a third one.
- Given a random variable Y, two random variables Y_1 and Y_2 are **independent** if

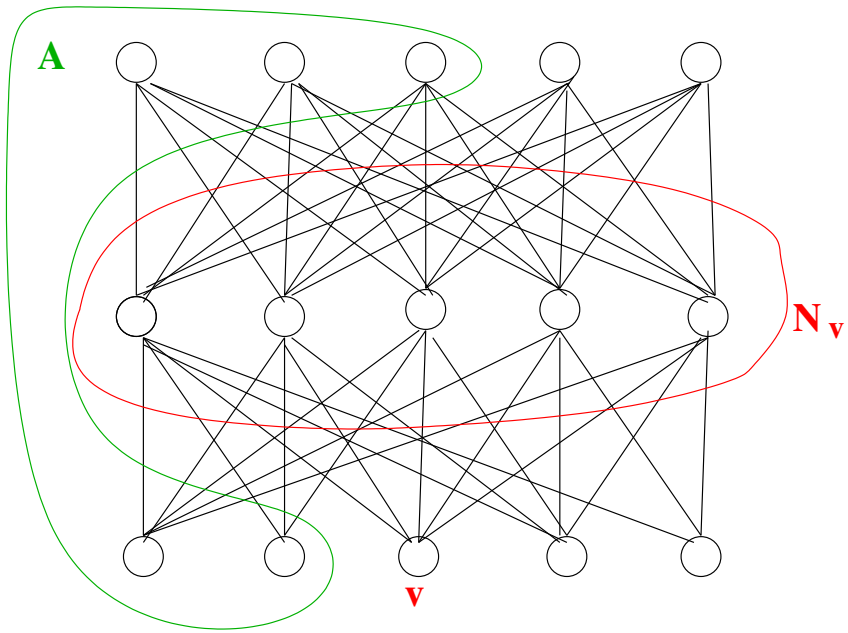
$$\Pr(Y_1, Y_2 | Y) = \Pr(Y_1 | Y) \Pr(Y_2 | Y)$$

Markov Random Fields II

- Assume $G = (V, E)$ is an undirected graph such that each node $v \in V$ is associated with a random variable Y_v taking values in a finite state space S. Thus, $S^{|V|}$ is the set of configurations.
- Assume **local Markov property**: For all $A \subset V$, $v \in V \setminus A$:

$$\Pr(v|\mathrm{N}_v, A) = \Pr(v|\mathrm{N}_v),$$

where N_v denotes the set of neighbours of v in G.



- $(Y_v)_{v \in V}$ is called a **Markov Random Field**.

* Gibbs Distribution (Hammersley-Clifford Theorem)

- Recall that a **clique** of G is a fully connected subgraph.
- Let $\mathrm{cl}(G)$ denote the set of cliques of G and $|A|$ the number of elements in any finite set A.
- Any strictly positive probability distribution $q : S^{|V|} \to [0, 1]$ of the configurations $S^{|V|}$ of a MRF **factorises** over $\mathrm{cl}(G)$.
- This means that for each $C \in \mathrm{cl}(G)$ there exists a function $\phi_C : S^{|C|} \to \mathbb{R}$ such that for $x \in S^{|V|}$

$$q(x) = \frac{1}{Z} \prod_{C \in \mathrm{cl}(G)} \phi_C(x_C),$$

where $x_C \in S^{|C|}$ denotes the components of x in C and

$$Z = \sum_{x \in S^{|V|}} \prod_{C \in \mathrm{cl}(G)} \phi_C(x_C),$$

is the normalisation constant, or the **partition function**.

* Logistic model

- Since $\phi_C(x_C) > 0$ for each clique C, we can define the **energy** of a clique for state $x \in X$ as

$$E(x_C) := -\log \phi_C(x_C), \qquad \text{where}$$

$$q(x) = \frac{1}{Z} \prod_{C \in \mathrm{cl}(G)} \exp -E(x_C) = \frac{1}{Z} \exp \left(\sum_{C \in \mathrm{cl}(G)} -E(x_C) \right),$$

$$\text{with} \quad Z = \sum_{x \in S^{|V|}} \exp \left(\sum_{C \in \mathrm{cl}(G)} -E(x_C) \right).$$

- In a **Logistic model**, i.e., **log linear** model, the energy is assumed to have the form:

$$E(x_C) = -w_C^T f_C(x_C) = -w_C \cdot f_C(x_C),$$

  where the vector w_C represents a model parameter at clique C and f_C is a clique dependent vector function.
- Stochastic Hopfield networks, Boltzmann Machines and RBM's all have this Logistic model.

Gibbs Sampling

- The stochastic Hopfield network can be viewed as a Markov random field, with a fully connected graph.
- The asynchronous probabilistic rule for flipping states one at a time is an example of a general method:
- **Gibbs sampling** in a Markov random field with a graph $G = (V, E)$ updates each variable based on its conditional distribution given the state of the other variables.
- Assume $X = (X_1, \ldots, X_N)$, where $V = \{1, \ldots, N\}$, with X_i taking values in a finite set.
- Suppose $\pi(x) = e^{-E(x)}/Z$ is the joint distribution of X.
- Assume q is a strictly positive distribution on V.
- At each step, pick $i \in V$ with probability $q(i)$; sample a new value for X_i based on its conditional probability distribution given the state $(x_v)_{v \in V \setminus i}$ of all other variables $(X_v)_{v \in V \setminus i}$.
- This defines a transition matrix for which π has the detailed balanced condition, i.e., is the stationary distribution.
- Instead of q, a pre-defined order is usually used for selecting nodes and the result still holds.

Gibbs Sampling in Hopfield networks

- Suppose a probability distribution q is given on the N nodes of a Hopfield network with $q(i) > 0$ for $1 \leq i \leq N$.
- If at each point of time we select node i with probability $q(i)$ and then flip its value according to the flipping probability as in Equation (4), then we are performing Gibbs sampling.
- What is the transition matrix P_{xy} for $x, y \in \{-1, 1\}^N$?
- If $H(x, y) \geq 2$ then no transition takes place, i.e., $P_{xy} = 0$.
- If $H(x, y) = 1$, then $x_i \neq y_i$ for some i with $1 \leq i \leq N$. Then node i is selected with probability $q(i)$ and thus $P_{xy} = q(i)/(1 + \exp \Delta E_i)$ where $\Delta E_i = E(-x_i) - E(x_i)$ is the change in the energy.
- And the probability that node i is selected and x_i is not flipped is $q(i)(1 - 1/(1 + \exp \Delta E_i))$.
- Note that the sum of all all these probabilities will add to 1:

$$\sum_{i=1}^{N} \frac{q(i)}{1 + \exp \Delta E_i} + q(i) \left(1 - \frac{1}{1 + \exp \Delta E_i}\right) = \sum_{i=1}^{N} q(i) = 1.$$