T-79.5204 Combinatorial Models and Stochastic Algorithms

Lecture Notes

Pekka Orponen
Helsinki University of Technology
Laboratory for Theoretical Computer Science

Contents

Preface

This set of lecture notes follows the Spring 2007 instalment of the course. The arduous task of typesetting the notes was kindly performed by Lic. Tech. Vesa Hölttä in Spring 2003 when the course — then under its former course code T-79.250 — was first lectured. This script has then been corrected and revised first in Spring 2005 and then in Spring 2007.

The reader should be warned that even though a large number of typos and even factual errors (due to the author, not the typesetter) were corrected in these two revisions, even the present edition is still likely to contain quite a few of both. According to current schedule, the next instalment of the course, and consequently the next major update of the notes, is due in Spring 2009.

Helsinki, 22 April 2007

Pekka Orponen

Part I

Markov Chains and Stochastic Sampling

1 Markov Chains and Random Walks on Graphs

1.1 Structure of Finite Markov Chains

We shall only consider Markov chains with a finite, but usually very large, *state space* $S = \{1, \dots, n\}$.

An *S-valued* (discrete-time) *stochastic process* is a sequence $X_0, X_1, X_2, \dots$ of S-valued random variables over some probability space Ω, i.e. a sequence of (measurable) maps $X_t : \Omega \to S, t = 0, 1, 2, \dots$

Such a process is a *Markov chain* if for all $t \geq 0$ and any $i_0, i_1, \dots, i_{t-1}, i, j \in S$ the following "memoryless" (forgetting) condition holds:

$$\begin{aligned} &\Pr(X_{t+1} = j \mid X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}, X_t = i) \\ &\quad = \Pr(X_{t+1} = j \mid X_t = i). \end{aligned} \tag{1}$$

Consequently, the process can be described completely by giving its *initial distribution (vector)*[1]

$$p^0 = [p_1^0, \dots, p_n^0] = \left[p_i^0\right]_{i=1}^n, \quad \text{where } p_i^0 = \Pr(X_0 = i)$$

[1] By a somewhat confusing convention, distributions in Markov chain theory are represented as row vectors. We shall be denoting the $1 \times n$ *column* vector with components $p_1, \dots, p_n$ as $(p_1, \dots, p_n)$, and the corresponding $n \times 1$ *row* vector as $[p_1, \dots, p_n] = (p_1, \dots, p_n)^T$. All vectors shall be column vectors unless otherwise indicated by text or notation.

and its sequence of *transition (probability) matrices*

$$P^{(t)} = \left(p_{ij}^{(t)}\right)_{i,j=1}^n, \quad \text{where } p_{ij}^{(t)} = \Pr(X_t = j \mid X_{t-1} = i).$$

Clearly, by the rule of total probability, the distribution vector at time $t \geq 1$

$$p^{(t)} = [\Pr(X_t = j)]_{j=1}^n$$

is obtained from $p^{(t-1)}$ simply by computing for each j:

$$p_j^{(t)} = \sum_{i=1}^n p_i^{(t-1)} \cdot p_{ij}^{(t)},$$

or more compactly

$$p^{(t)} = p^{(t-1)} P^{(t)}.$$

Recurring back to the initial distribution, this yields

$$p^{(t)} = p^0 P^{(1)} P^{(2)} \dots P^{(t)}. \tag{2}$$

If the transition matrix is time-independent, i.e. $P^{(t)} = P$ for all $t \geq 1$, the Markov chain is *homogeneous*, otherwise *inhomogeneous*. We shall be mostly concerned with the homogeneous case, in which formula (2) simplifies to

$$p^{(t)} = p^0 P^t.$$

We shall say in general that a vector $q \in \mathbb{R}^n$ is a *stochastic vector* if it satisfies

$$q_i \geq 0 \quad \forall\, i = 1, \dots, n \quad \text{and} \quad \sum_i q_i = 1.$$

A matrix $Q \in \mathbb{R}^{n \times n}$ is a *stochastic matrix* if all its row vectors are stochastic vectors.

Now let us assume momentarily that for a given homogeneous Markov Chain with transition matrix P and initial probability distribution p^0 there exists a limit distribution $\pi \in [0,1]^n$ such that

$$\lim_{t \to \infty} p^{(t)} = \pi \qquad \text{(in any norm, e.g. coordinatewise)}. \tag{3}$$

Then it must be the case that

$$\begin{aligned} \pi &= \lim_{t \to \infty} p^0 P^t = \lim_{t \to \infty} p^0 P^{t+1} \\ &= \left(\lim_{t \to \infty} p^0 P^t\right) P = \pi P. \end{aligned}$$

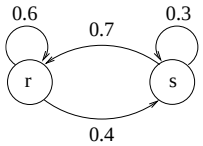


Figure 1: A Markov chain for Helsinki weather.

Thus, any limit distribution satisfying property (3), if such exist, is a left eigenvector of the transition matrix with eigenvalue 1, and can be computed by solving the equation $\pi = \pi P$. Solutions to this equation are called the *equilibrium* or *stationary distributions* of the chain.

Example 1.1 *The weather in Helsinki.* Let us say that tomorrow's weather is conditioned on today's weather as represented in Figure 1 or in the transition matrix:

P	rain	sun
rain	0.6	0.4
sun	0.7	0.3

Then the long-term weather distribution can be determined, in this case uniquely and in fact independent of the initial conditions, by solving

$$\begin{aligned} & \pi P = \pi, \quad \sum_i \pi_i = 1 \\ \Leftrightarrow \quad & \begin{bmatrix} \pi_r & \pi_s \end{bmatrix} \begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix} = \begin{bmatrix} \pi_r & \pi_s \end{bmatrix}, \quad \pi_r + \pi_s = 1 \\ \Leftrightarrow \quad & \begin{cases} \pi_r = 0.6\pi_r + 0.7\pi_s \\ \pi_s = 0.4\pi_r + 0.3\pi_s \end{cases}, \quad \pi_r + \pi_s = 1 \\ \Leftrightarrow \quad & \begin{cases} \pi_r = 0.64 \\ \pi_s = 0.36 \end{cases} \end{aligned}$$

Every finite Markov chain has at least one stationary distribution, but as the following examples show, this need not be unique, and even if it is, then the chain does not need to converge towards it in the sense of equation (3).

Example 1.2 *A reducible Markov chain.* Consider the chain represented in Figure 2. Clearly any distribution $p = [p_1 \quad p_2]$ is stationary for this chain. The

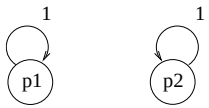


Figure 2: A reducible Markov chain.

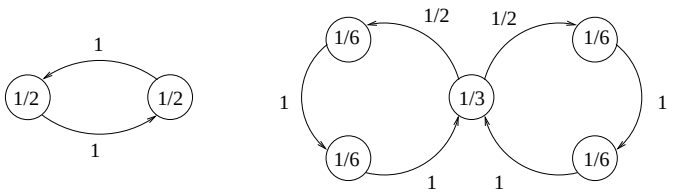


Figure 3: Periodic Markov chains.

underlying cause for the existence of several stationary distributions is that the chain is *reducible*, meaning that it consists of several "noncommunicating" components. (Precise definitions are given below.)

Any irreducible ("fully communicating") chain has a unique stationary distribution, but this does not yet guarantee convergence in the sense of equation (3).

Example 1.3 *Periodic Markov chains.* Consider the chains represented in Figure 3. These chains are *periodic*, with periods 2 and 3. While they do have unique stationary distributions indicated in the figure, they only converge to those distributions from the corresponding initial distributions; otherwise probability mass "cycles" through each chain.

So when is a unique stationary limit distribution guaranteed? The brief answer is as follows.

Consider a finite, homogeneous Markov chain with state set S and transition matrix P. The chain is:

(i) *irreducible*, if any state can be reached from any other state with positive probability, i.e.

$$\forall\, i, j \in S \quad \exists\, t \geq 0 \;:\; P_{ij}^t > 0;$$

(ii) *aperiodic* if for any state $i \in S$ the greatest common divisor of its possible recurrence times is 1, i.e. denoting

$$N_i = \{t \geq 1 \mid P_{ii}^t > 0\}$$

we have $\gcd(N_i) = 1, \quad \forall\, i \in S$.

Theorem (Markov Chain Convergence) *A finite homogeneous Markov chain that is irreducible and aperiodic has a unique stationary distribution π, and the chain will converge towards this distribution from any initial distribution p^0 in the sense of Equation (3).* □

Irreducible and aperiodic chains are also called *regular* or *ergodic*.

We shall prove this important theorem below, establishing first the existence and uniqueness of the stationary distribution, and then convergence. Before going into the proof, let us nevertheless first look into the structure of arbitrary, possibly nonregular, finite Markov chains somewhat more closely.

Let the finite state space be S and the homogeneous transition matrix be P.

A set of states $C \subseteq S, C \neq \varnothing$ is *closed* or *invariant*, if $p_{ij} = 0 \quad \forall\, i \in C, j \notin C$.

A singleton closed state is *absorbing* (i.e. $p_{ii} = 1$).

A chain is *irreducible* if S is the only closed set of states. (This definition can be seen to be equivalent to the one given earlier.)

Lemma 1.1 *Every closed set contains a **minimal** closed set as a subset.* □

State j is *reachable* from state i, denoted $i \to j$, if $P_{ij}^t > 0$ for some $t \geq 0$.

States $i, j \in S$ *communicate*, denoted $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

Lemma 1.2 *The communication relation "$\leftrightarrow$" is an equivalence relation. All the minimal closed sets of the chain are equivalence classes with respect to "$\leftrightarrow$". The chain is irreducible if and only if all its states communicate.* □

States which do not belong to any of the minimal closed subsets are called *transient*.

One may thus partition the chain into equivalence class with respect to "$\leftrightarrow$". Each class is either a minimal closed set or consists of transient states. This is illustrated in Figure 4. By "reducing" the chain in this way one obtains a DAG-like structure, with the minimal closed sets as leaves and the transient components as internal nodes. (Actually a "forest" if the chain is disconnected.) An irreducible chain of course reduces to a single node.

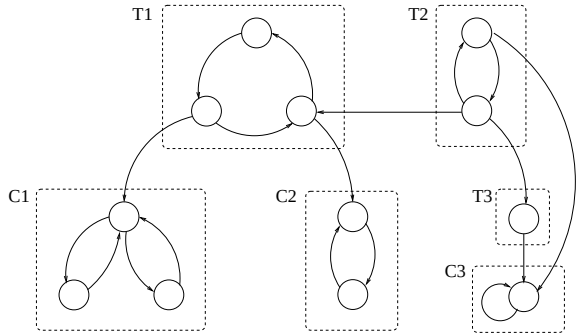


Figure 4: Partitioning of a Markov chain into communicating classes.

The *period* of state $i \in S$ is

$$\gcd\{\underbrace{t \geq 1 \mid P_{ii}^t > 0}_{N_i}\}.$$

A state with period 1 is *aperiodic*.

Lemma 1.3 *Two communicating states have the same period. Hence, every component of the "$\leftrightarrow$" relation has a uniquely determined period.* □

Define the *first hit* (or *first passage*) probabilities for states $i \to j$ and $t \geq 1$ as:

$$f_{ij}^{(t)} = \Pr(X_1 \neq j, X_2 \neq j, \ldots, X_{t-1} \neq j, X_t = j \mid X_0 = i),$$

and the *hitting* (or *passage*) *probability* for $i \to j$ as

$$f_{ij}^* = \sum_{t \geq 1} f_{ij}^{(t)}.$$

Then the *expected hitting* (*passage*) *time* for $i \to j$ is

$$\mu_{ij} = \begin{cases} \sum_{t \geq 1} t\, f_{ij}^{(t)}, & \text{if } f_{ij}^* = 1; \\ \infty & \text{if } f_{ij}^* < 1 \end{cases}$$

For $i = j$, μ_{ii} is called the *expected return time*, and often denoted simply μ_i.

State $i \in S$ is *recurrent* (or *persistent*) if $f_{ii}^* = 1$, otherwise it is *transient*. (In infinite Markov chains the recurrent states are further divided into *positive recurrent*

with $\mu_i < \infty$ and *null recurrent* with $\mu_i = \infty$, but the latter case does not occur in finite Markov chains and thus need not concern us here.)

The following theorem provides an important characterisation of the recurrent states.

Notation: $P^k = \left(p_{ij}^{(k)}\right)_{i,j=1}^{n}$.

Theorem 1.4 *State $i \in S$ is recurrent if and only if $\sum_{k\geq 0} p_{ii}^{(k)} = \infty$. Correspondingly, $i \in S$ is transient if and only if $\sum_{k\geq 0} p_{ii}^{(k)} < \infty$.*

Proof. Recall the relevant definitions:

$$\begin{aligned} p_{ii}^{(k)} &= \Pr(X_k = i \mid X_0 = i), \\ f_{ii}^{(t)} &= \Pr(X_1 \neq i, \ldots, X_{t-1} \neq i, X_t = i \mid X_0 = i). \end{aligned}$$

Then it is fairly clear that

$$p_{ii}^{(k)} = \sum_{t=1}^{k} f_{ii}^{(t)} p_{ii}^{(k-t)} = \sum_{t=0}^{k-1} f_{ii}^{(k-t)} p_{ii}^{(t)}.$$

Consequently, for any K:

$$\begin{aligned} \sum_{k=1}^{K} p_{ii}^{(k)} &= \sum_{k=1}^{K} \sum_{t=0}^{k-1} f_{ii}^{(k-t)} p_{ii}^{(t)} \\ &= \sum_{t=0}^{K-1} p_{ii}^{(t)} \sum_{k=t+1}^{K} f_{ii}^{(k-t)} \\ &\leq \sum_{t=0}^{K} p_{ii}^{(t)} f_{ii}^{*} \\ &= \left(1 + \sum_{t=1}^{K} p_{ii}^{(t)}\right) f_{ii}^{*} \end{aligned}$$

Since K was arbitrary, we obtain:

$$(1 - f_{ii}^{*}) \sum_{k=1}^{\infty} p_{ii}^{(k)} \leq f_{ii}^{*}.$$

Now if $i \in S$ is transient, i.e. $f_{ii}^{*} < 1$, then

$$\sum_{k\geq 1} p_{ii}^{(k)} \leq \frac{f_{ii}^{*}}{1 - f_{ii}^{*}} < \infty.$$

Conversely, assume that $i \in S$ is recurrent, i.e. $f_{ii}^{*} = 1$. Now one can see that

$$\begin{aligned} \Pr(X_t = i \text{ for at least two } t \geq 1 \mid X_0 = i) &= \sum_{t,t'\geq 1} f_{ii}^{(t)} f_{ii}^{(t')} = \left(\sum_{t\geq 1} f_{ii}^{(t)}\right)^2 \\ &= (f_{ii}^{*})^2 = 1, \end{aligned}$$

and by induction that

$$\Pr(X_t = i \text{ for at least } s \text{ times } | X_0 = i) = (f_{ii}^{*})^s = 1.$$

Consequently,

$$P_{kk}^{\infty} \triangleq \Pr(X_k = i \text{ infinitely often} \mid X_0 = i) = \lim_{s\to\infty} (f_{ii}^{*})^s = 1.$$

However, if $\sum_{k\geq 0} p_{ii}^{(k)} < \infty$, then by the Borel-Cantelli lemma (see below) it should be the case that $p_{kk}^{\infty} = 0$.

Thus it follows that if $f_{ii}^{*} = 1$, then also $\sum_{k\geq 0} p_{ii}^{(k)} = \infty$. □

Lemma (Borel-Cantelli, "easy case") *Let $A_0, A_1, \ldots$ be events, and A the event "infinitely many of the A_k occur". Then*

$$\sum_{k\geq 0} \Pr(A_k) < \infty \Rightarrow \Pr(A) = 0.$$

Proof. Clearly

$$A = \bigcap_{m\geq 0} \bigcup_{k\geq m} A_k.$$

Thus for all $m \geq 0$,

$$\Pr(A) \leq \Pr\left(\bigcup_{k\geq m} A_k\right) \leq \sum_{k\geq m} \Pr(A_k) \to 0 \text{ as } m \to \infty,$$

assuming the sum $\sum_{k\geq 0} \Pr(A_k)$ converges. □

Let $C_1, \ldots, C_m \subseteq S$ be the minimal closed sets of a finite Markov chain, and $T \triangleq S \setminus (C_1 \cup \cdots \cup C_m)$.

Theorem 1.5 *(i) Any state $i \in C_r$, for some $r = 1, \ldots, m$, is recurrent.*
(ii) Any state $i \in T$ is transient.

Proof. (i) Assume $i \in C$, C minimal closed subset of S. Then for any $k \geq 1$,

$$\sum_{j \in S} p_{ij}^{(k)} = \sum_{j \in C} p_{ij}^{(k)} = 1,$$

because C is closed and P is a stochastic matrix. Consequently,

$$\sum_{k \geq 0} \sum_{j \in C} p_{ij}^{(k)} = \infty,$$

and because C is finite, there must be some $j_0 \in C$ such that

$$\sum_{k \geq 0} p_{ij_0}^{(k)} = \infty.$$

Since $j_0 \leftrightarrow i$, there is some $k_0 \geq 0$ such that $p_{j_0 i}^{(k_0)} = p_0 > 0$. But then

$$\sum_{k \geq 0} p_{ii}^{(k)} \geq \sum_{k \geq k_0} p_{ij_0}^{(k-k_0)} p_{j_0 i}^{(k_0)} = \left(\sum_{k \geq k_0} p_{ij_0}^{(k-k_0)} \right) \cdot p_0 = \infty.$$

By Theorem 1.4 i is thus recurrent.

(ii) Denote $C = C_1 \cup \cdots \cup C_m$. Since for any $j \in Y$ the set $\{l \in S \mid j \to l\}$ is closed, it must intersect C; thus for any $j \in T$ there is some $k \geq 1$ such that

$$p_{iC}^{(k)} \triangleq \sum_{l \in C} p_{jl}^{(k)} > 0.$$

Since T is finite, we may find a $k_0 \geq 1$ such that for any $j \in T$, $p_{jC}^{(k_0)} = p > 0$. Then one may easily compute that for any $i \in T$,

$$p_{iT}^{(k_0)} \leq 1 - p,\ p_{iT}^{(2k_0)} \leq (1-p)^2,\ p_{iT}^{(3k_0)} \leq (1-p)^3, \text{ etc.}$$

and so

$$\sum_{k \geq 1} p_{ii}^{(k)} \leq \sum_{k \geq 1} p_{iT}^{(k)} \leq \sum_{r \geq 0} k_0 p_{iT}^{(rk_0)} \leq k_0 \sum_{r \geq 0} (1-p)^r < \infty.$$

By Theorem 1.4, i is thus transient. □

1.2 Existence and Uniqueness of Stationary Distribution

A matrix $A \in \mathbb{R}^{n \times n}$ is

(i) *nonnegative*, denoted $A \geq 0$, if $a_{ij} \geq 0 \quad \forall\, i, j$

(ii) *positive*, denoted $A \gtrsim 0$, if $a_{ij} \geq 0 \quad \forall\, i, j$ and $a_{ij} > 0$ for at least one ij

(iii) *strictly positive*, denoted $A > 0$, if $a_{ij} > 0 \quad \forall\, i, j$

We denote also $A \geq B$ if $A - B \geq 0$, etc.

Lemma 1.6 *Let $P \geq 0$ be the transition matrix of some regular finite Markov chain with state set S. Then for some $t_0 \geq 1$ it is the case that $P^t > 0 \quad \forall\, t \geq t_0$.*

Proof. Choose some $i \in S$ and consider the set

$$N_i = \{t \geq 1 \mid p_{ii}^{(t)} > 0\}.$$

Since the chain is (finite and) aperiodic, there is some finite set of numbers $t_1, \ldots, t_m \in N_i$ such that

$$\gcd N_i = \gcd\{t_1, \ldots, t_m\} = 1,$$

i.e. for some set of coefficients $a_1, \ldots, a_m \in \mathbb{Z}$,

$$a_1 t_1 + a_2 t_2 + \cdots + a_m t_m = 1.$$

Let P and N be the absolute values of the positive and negative parts of this sum, respectively. Thus $P - N = 1$. Let $T \geq N(N-1)$ and consider any $s \geq T$. Then $s = aN + r$, where $0 \leq r \leq N-1$ and, consequently, $a \geq N-1$. But then $s = aN + r(P-N) = (a-r)N + P$ where $a - r \geq 0$, i.e. S can be represented in terms of $t_1, \ldots, t_m$ with nonnegative coefficients $b_1, \ldots, b_m$. Thus

$$p_{ii}^{(s)} \geq p_{ii}^{(b_1 t_1)} p_{ii}^{(b_2 t_2)} \cdots p_{ii}^{(b_m t_m)} > 0.$$

Since the chain is irreducible, the claim follows by choosing t_0 sufficiently larger than T to allow all states to communicate with i. □

Let then $A \geq 0$ be an arbitrary nonnegative $n \times n$-matrix. Consider the set

$$\Lambda = \{\lambda \in \mathbb{R} \mid Ax \geq \lambda x \text{ for some } x \geq 0\}.$$

Clearly $0 \in \Lambda$, so $\Lambda \neq \varnothing$. Also, it is easy to see that the values in Λ are upper bounded by the maximal rowsum M of A. Thus $\Lambda \subseteq [0, M]$, and we may define

$$\lambda^* = \sup \Lambda.$$

To see that the supremum of Λ is actually attained by some $\lambda^* \in \Lambda$ and vector $x^* \geq 0$, observe that one may also define λ^* as

$$\lambda^* = \max_{x \in [0,1]^n} \min_{i=1,\ldots,n} \frac{(Ax)_i}{x_i},$$

where in the case of $x_i = 0$, the quotient $\frac{(Ax)_i}{x_i}$ is defined as the appropriate limit to maintain continuity.

Theorem 1.7 (Perron-Frobenius) *For any strictly positive matrix $A > 0$ there exist a positive real number $\lambda^* > 0$ and a strictly positive vector $x^* > 0$ such that:*

(i) $Ax^* = \lambda^* x^*$;

(ii) *if $\lambda \neq \lambda^*$ is any other (in general complex) eigenvalue of A, then $|\lambda| < \lambda^*$;*

(iii) λ^* *has geometric and algebraic multiplicity 1.*

Proof. Define λ^* as above, and let $x^* \geq 0$ be a vector such that $Ax^* \geq \lambda^* x^*$. Since $A > 0$, also $\lambda^* > 0$.

(i) Suppose that it is not the case that $Ax^* = \lambda^* x^*$, i.e. that $Ax^* \geq \lambda^* x^*$, but not $Ax^* = \lambda^* x^*$. Consider the vector $y^* = Ax^*$. Since $A > 0$, $Ax > 0$ for any $x \gtrsim 0$; in particular now $A(y^* - \lambda^* x^*) = Ay^* - \lambda^* Ax^* = Ay^* - \lambda^* y^* > 0$, i.e. $Ay^* > \lambda^* y^*$; but this contradicts the definition of λ^*.

Consequently $Ax^* = \lambda^* x^*$, and furthermore $x^* = \frac{1}{\lambda^*} Ax^* > 0$.

(ii) Let $\lambda \neq \lambda^*$ be an eigenvalue of A and $y \neq 0$ the corresponding eigenvector, $Ay = \lambda y$. Denote $|y| = (|y_1|, \ldots, |y_n|)$. Since $A > 0$, it is the case that

$$A|y| \geq |Ay| = |\lambda y| = |\lambda||y|.$$

By the definition of λ^*, it follows that $|\lambda| \leq \lambda^*$.

To prove strict inequality, let $\delta > 0$ be so small that the matrix $A_\delta = A - \delta I$ is still strictly positive. Then for any eigenvalue λ of A, $\lambda - \delta$ is an eigenvalue of A_δ and vice versa. Since $A_\delta > 0$, its largest eigenvalue is $\lambda^* - \delta$, i.e. for any other eigenvalue λ of A, $|\lambda - \delta| \leq \lambda^* - \delta$.

But this implies that A cannot have any eigenvalues $\lambda \neq \lambda^*$ on the circle $|\lambda| = \lambda^*$, because such would have $|\lambda - \delta| > |\lambda^* - \delta|$. (See Figure 5.)

(iii) We shall consider only the geometric multiplicity. Suppose there was another (real) eigenvector $y > 0$, linearly independent of x^*, associated to λ^*. Then one could form a linear combination $w = x^* + \alpha y$ such that $w \gtrsim 0$, but not $w > 0$. However, since $A > 0$, it must be the case that also $w = \frac{1}{\lambda^*} Aw > 0$. □

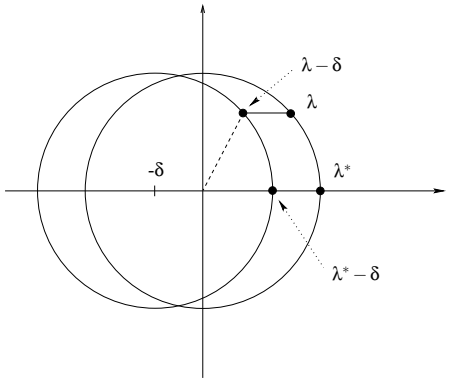


Figure 5: Maximality of the Perron-Frobenius eigenvalue.

Corollary 1.8 *If A is a nonnegative matrix ($A \geq 0$) such that some power of A is strictly positive ($A^n > 0$), then the conclusions of Theorem 1.7 hold also for A.* □

Note: In fact every nonnegative matrix $A \geq 0$ has a real "Perron-Frobenius" eigenvalue $\lambda^* \geq 0$ of maximum modulus, i.e. such that $|\lambda| \leq \lambda^*$ holds for all eigenvalues λ of A. But in this general case there may also be complex eigenvalues of equal modulus, and λ^* itself may be nonsimple, i.e. have multiplicity greater than one.

Proposition 1.9 *Let $A \geq 0$ be a nonnegative $n \times n$ matrix with row and column sums*

$$r_i = \sum_j a_{ij}, \quad c_j = \sum_i a_{ij}, \quad i, j = 1, \ldots, n$$

Then for the Perron-Frobenius eigenvalue λ^* of A the following bounds hold:

$$\min_i r_i \leq \lambda^* \leq \max_i r_i, \quad \min_j c_j \leq \lambda^* \leq \max_j c_j.$$

Proof. Let $x^* = (x_1, x_2, \ldots, x_n)$ be an eigenvector corresponding to λ^*, normalised so that $\sum_i x_i = 1$. Summing up the equations for $Ax^* = \lambda^* x^*$ yields:

$$\begin{array}{ccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = \lambda^* x_1 \\
a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = \lambda^* x_2 \\
 & & & & \vdots & & & \\
a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = \lambda^* x_n \\
\hline
c_1x_1 & + & c_2x_2 & + & \ldots & + & c_nx_n & = \lambda^* \underbrace{(x_1 + \cdots + x_n)}_{1} & = \lambda^*
\end{array}$$

Thus λ^* is a "weighted average" of the column sums, so in particular $\min_j c_j \leq \lambda^* \leq \max_j c_j$.

Applying the same argument to A^T, which has the same λ^* as A, yields the row sum bounds. □

Corollary 1.10 *Let $P \geq 0$ be the transition matrix of a regular Markov chain. Then there exists a unique distribution vector π such that $\pi P = \pi$ ($\Leftrightarrow$ $P^T \pi^T = \pi^T$).*

Proof. By Lemma 1.6 and Corollary 1.8, P has a unique largest eigenvalue $\lambda^* \in \mathbb{R}$. By Proposition 1.9, $\lambda^* = 1$, because as a stochastic matrix all row sums of P (i.e. the column sums of P^T) are 1. Since the geometric multiplicity of λ^* is 1, there is a unique stochastic vector π (i.e. satisfying $\sum_i \pi_i = 1$) such that $\pi P = \pi$. □

1.3 Convergence of Regular Markov Chains

In Corollary 1.10 we established that a regular Markov chain with transition matrix P has a unique stationary distribution vector π such that $\pi P = \pi$.

By elementary arguments (page 3) we know that starting from any initial distribution q, if the iteration $q, qP, qP^2, \ldots$ converges, then it must converge to this unique stationary distribution.

However, it remains to be shown that if the Markov chain determined by P is regular, then the iteration always converges.

The following matrix decomposition is well known:

Lemma 1.11 (Jordan canonical form) *Let $A \in \mathbb{C}^{n\times n}$ be any matrix with eigenvalues $\lambda_1, \ldots, \lambda_l \in \mathbb{C}$, $l \leq n$. Then there exists an invertible matrix $U \in \mathbb{C}^{n\times n}$ such that*

$$UAU^{-1} = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & J_r \end{bmatrix}$$

*where each J_i is a $k_i \times k_i$ **Jordan block** associated to some eigenvalue λ of A:*

$$J_i = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

The total number of blocks associated to a given eigenvalue λ corresponds to λ's geometric multiplicity, and their total dimension $\sum_i k_i$ to λ's algebraic multiplicity. □

Now let us consider the Jordan canonical form of a transition matrix P for a regular Markov chain. Assume for simplicity that all the eigenvalues of P are real and distinct. (The general argument is similar, but needs more complicated notation.) Then the rows of U may be taken to be left eigenvectors of the matrix P, and the Jordan canonical form reduces to the familiar eigenvalue decomposition:

$$UPU^{-1} = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$

In this case one notes that in fact the columns of $U^{-1} = V$ are precisely the *right* eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$. By Lemma 1.6 and Corollary 1.8, P has a unique largest eigenvalue $\lambda_1 = 1$, and the other eigenvalues may be ordered so that $1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$. The unique (up to normalisation) left eigenvector associated to eigenvalue 1 is the stationary distribution π, and the corresponding unique (up to normalisation) right eigenvector is $\mathbf{1} = (1, 1, \ldots, 1)$. If the first row of U is normalised to π, then the first column of V must be normalised to $\mathbf{1}$ because $UV = UU^{-1} = I$, and hence $(UV)_{11} = u_1 v_1 = \pi v_1 = 1$.

Denoting

$$\Lambda = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix},$$

we have then:

$$P^2 = (V\Lambda U)^2 = V\Lambda^2 U = V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^2 \end{bmatrix} U,$$

and in general

$$P^t = V\Lambda^t U = V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^t & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^t \end{bmatrix} U$$

$$\xrightarrow[t\to\infty]{} V \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} U = \begin{bmatrix} v_{11}u_1 \\ v_{12}u_1 \\ \vdots \\ v_{1n}u_1 \end{bmatrix} = \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}.$$

To make the situation even more transparent, represent a given initial distribution $q = q^0$ in the (left) eigenvector basis as

$$\begin{aligned} q &= \tilde{q}_1 u_1 + \tilde{q}_2 u_2 + \cdots + \tilde{q}_n u_n \\ &= \pi + \tilde{q}_2 u_2 + \cdots + \tilde{q}_n u_n, \qquad \text{where } \tilde{q}_i = \langle q^T, v_i \rangle = q v_i. \end{aligned}$$

Then

$$qP = (\pi + \tilde{q}_2 u_2 + \cdots + \tilde{q}_n u_n) P = \pi + \tilde{q}_2 \lambda_2 u_2 + \cdots + \tilde{q}_n \lambda_n u_n,$$

and generally

$$q^{(t)} = qP^t = \pi + \sum_{i=2}^{n} \tilde{q}_i \lambda_i^t u_i,$$

implying that $q^{(t)} \xrightarrow[t\to\infty]{} \pi$, and if the eigenvalues are ordered as assumed, then

$$||q^{(t)} - \pi|| = o(|\lambda_2|^t).$$

1.4 Transient Behaviour of General Chains

So what happens to the transient states in a reducible Markov chain?

A moment's thought shows that the transition matrix of an arbitrary (finite) Markov chain can be put in the following *canonical form*:

$$P = \left[\begin{array}{ccc|c} P_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & P_r & \\ \hline & & & \\ & R & & Q \\ & & & \end{array}\right]$$

where the r square matrices $P_1, \ldots, P_r$ in the upper left corner represent the transitions within the r minimal closed classes, Q represents the transitions among transient states, and R represents the transitions from transient states to one of the closed classes.

In this ordering, stationary distributions (left eigenvectors of P corresponding to eigenvalue 1) must apparently be of the form $\pi = [\pi_1 \ \cdots \ \pi_r \quad 0 \ \cdots \ 0]$. (Note that since Q has at least one row sum less than 1, by the proof argument in Proposition 1.9 also all of its eigenvalues have modulus less than 1. Thus the only solution of the stationarity equation $\mu Q = \mu$ is $\mu = 0$.)

Consider then the *fundamental matrix* $M = (I - Q)^{-1}$ of the chain. Intuitively, if M is well-defined, it corresponds to $M = I + Q + Q^2 + \ldots$, and represents all the possible transition sequences the chain can have without exiting Q.

Theorem 1.12 *For any finite Markov chain with transition matrix as above, the matrix $I - Q$ is invertible, and its inverse can be represented as the convergent series $M = I + Q + Q^2 + \ldots$*

Proof. Since for any $t \geq 1$,

$$(I - Q)(I + Q + \cdots + Q^{t-1}) = I - Q^t,$$

and $Q^t \to 0$ as $t \to \infty$, the result follows. □

A transparent stochastic interpretation of the fundamental matrix may be obtained by considering any two transient states i, j in a Markov chain as above. Then:

$$\Pr(X_t = j \mid X_0 = i) = Q_{ij}^t \triangleq q_{ij}^{(t)}.$$

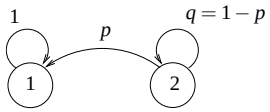


Figure 6: A Markov chain representing the geometric distribution.

Thus,

$$\begin{aligned} E[\text{number of visits to } j \in T \mid X_0 = i \in T] &= q_{ij}^{(0)} + q_{ij}^{(1)} + q_{ij}^{(2)} + \dots \\ &= I_{ij} + Q_{ij} + Q_{ij}^2 + \dots \\ &= M_{ij} \triangleq m_{ij}. \end{aligned}$$

Furthermore,

$$\begin{aligned} & E[\text{number of moves in } T \text{ before exiting to } C \mid X_0 = i \in T] \\ = & \sum_{j \in T} E[\text{number of visits to } j \in T \mid X_0 = i \in T] \\ = & \sum_{j \in T} m_{ij} \\ = & (M\mathbf{1})_i. \end{aligned}$$

As another application, let b_{ij} be the probability that the chain when started in transient state $i \in T$ will enter a minimal closed class via state $j \in C$. Denote $B = (b_{ij})_{i \in T, j \in C}$. Then in fact $B = MR$.

Proof. For given $i \in T$, $j \in C$,

$$b_{ij} = p_{ij} + \sum_{k \in T} p_{ik} b_{kj}.$$

Thus,

$$B = R + QB \quad \Rightarrow \quad B = (I - Q)^{-1} R = MR.$$

Example 1.4 *The geometric distribution.* Consider the chain of Figure 6, arising e.g. from biased coin-flipping The transition matrix in this case is

$$P = \begin{bmatrix} 1 & 0 \\ p & q \end{bmatrix}.$$

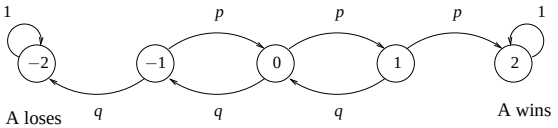


Figure 7: A Markov chain representing a coin-flipping game.

Now $Q = (q)$, $M = (1-q)^{-1} = 1/p$. Thus, e.g.

$$E[\text{number of visits to 2 before exiting to 1} \mid X_0 = 2] = M\mathbf{1} = \frac{1}{p}.$$

An elementary way to obtain the same result would be:

$$\begin{aligned} E[\text{number of visits}] &= \sum_{k \geq 0} \Pr[\text{number of visits} = k] \cdot k \\ &= \sum_{k \geq 0} \Pr[\text{number of visits} \geq k] \\ &= 1 + q + q^2 + \dots = \frac{1}{1-q} = \frac{1}{p}. \end{aligned}$$

Example 1.5 *Gambling tournament.* Players A and B toss a biased coin with A's success probability equal to p and B's success probability equal to $1 - p = q$. The person to first obtain n successes over the other wins. What are A's chances of winning, given that he initially has k successes over B, $-n \leq k \leq n$? (A more technical term for this process is "one-dimensional random walk with two absorbing barriers.")

For simplicity, let us consider only the case $n = 2$. Then the chain is as represented in Figure 7, with transition matrix:

	−2	−1	0	1	2
−2	1	0	0	0	0
−1	q	0	p	0	0
0	0	q	0	p	0
1	0	0	q	0	p
2	0	0	0	0	1

i.e. in canonical form:

	-2	2	-1	0	1
-2	1	0	0	0	0
2	0	1	0	0	0
-1	q	0	0	p	0
0	0	0	q	0	p
1	0	p	0	q	0

Thus, $M = (I-Q)^{-1}$

$$= \begin{bmatrix} 1 & -p & 0 \\ -q & 1 & -p \\ 0 & -q & 1 \end{bmatrix}^{-1} = \frac{1}{p^2+q^2} \begin{bmatrix} p+q^2 & p & p^2 \\ q & 1 & p \\ q^2 & q & q+p^2 \end{bmatrix}$$

and so $B = MR$

$$= \frac{1}{p^2+q^2} \begin{bmatrix} p+q^2 & p & p^2 \\ q & 1 & p \\ q^2 & q & q+p^2 \end{bmatrix} \begin{bmatrix} q & 0 \\ 0 & 0 \\ 0 & p \end{bmatrix} = \frac{1}{p^2+q^2} \begin{bmatrix} qp+q^3 & p^3 \\ q^2 & p^2 \\ \underbrace{q^3}_{\text{A loses}} & \underbrace{pq+p^3}_{\text{A wins}} \end{bmatrix}.$$

1.5 Reversible Markov Chains

We now introduce an important special class of Markov chains often encountered in algorithmic applications. Many examples of these types of chains will be encountered later.

Intuitively, a "reversible" chain has no preferred time direction at equilibrium, i.e. any given sequence of states is equally likely to occur in forward as in backward order.

A Markov chain determined by the transition matrix $P = (p_{ij})_{i,j\in S}$ is *reversible* if there is a distribution π that satisfies the *detailed balance* conditions:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall\, i,j \in S.$$

Theorem 1.13 *A distribution satisfying the detailed balance conditions is stationary.*

Proof. It suffices to show that, assuming the detailed balance conditions, the following stationarity condition holds for all $i \in S$:

$$\pi_i = \sum_{j\in S} \pi_j p_{ji}.$$

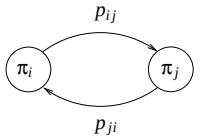


Figure 8: Detailed balance condition $\pi_i p_{ij} = \pi_j p_{ji}$.

But this is straightforward:

$$\sum_{j\in S} \pi_j p_{ji} = \sum_{j\in S} \pi_i p_{ij} = \pi_i \sum_{j\in S} p_{ji} = \pi_i.$$

□

Observe the intuition underlying the detailed balance condition: At stationarity, an equal amount of probability mass flows in each step from i to j as from j to i.(The "ergodic flows"' between states are in pairwise balance; cf. Figure 8.)

Example 1.6 *Random walks on graphs.*

Let $G = (V,E)$ be a (finite) graph, $V = \{1,\ldots,n\}$. Define a Markov chain on the nodes of G so that at each step, one of the current node's neigbours is selected as the next state, uniformly at random. That is,

$$p_{ij} = \begin{cases} \frac{1}{d_i}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \qquad (d_i = \deg(i))$$

Let us check that this chain is reversible, with stationary distribution

$$\pi = \begin{bmatrix} \frac{d_1}{d} & \frac{d_2}{d} & \cdots & \frac{d_n}{d} \end{bmatrix},$$

where $d = \sum_{i=1}^n d_i = 2|E|$. The detailed balance condition is easy to verify:

$$\pi_i p_{ij} = \begin{cases} \frac{d_i}{d} \cdot \frac{1}{d_i} = \frac{1}{d} = \frac{d_j}{d} \cdot \frac{1}{d_j} = \pi_j p_{ji}, & \text{if } (i,j) \in E \\ 0 = \pi_j p_{ji}, & \text{if } (i,j) \notin E \end{cases}$$

Example 1.7 *A nonreversible chain.*

Consider the three-state Markov chain shown in Figure 9. It is easy to verify that this chain has the unique stationary distribution $\pi = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$. However, for any $i = 1,2,3$:

$$\pi_i p_{i,(i+1)} = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9} > \pi_{i+1} p_{(i+1),i} = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}.$$

Thus, even in a stationary situation, the chain has a "preference" of moving in the counter-clockwise direction, i.e. it is not time-symmetric.

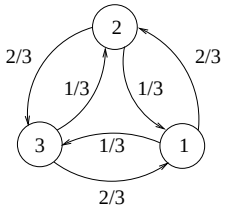


Figure 9: A nonreversible Markov chain.

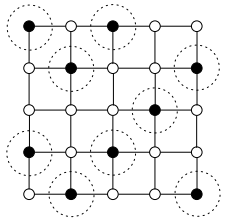


Figure 10: Hard-core colouring of a lattice.

2 Markov Chain Monte Carlo Sampling

We now introduce *Markov chain Monte Carlo (MCMC) sampling*, which is an extremely important method for dealing with "hard-to-access" distributions.

Assume that one needs to generate samples according to a probability distribution π, but π is too complicated to describe explicitly. A clever solution is then to construct a Markov chain that converges to stationary distribution π, let it run for a while and then sample states of the chain. (However, one obvious problem that this approach raises is determining how long is "for a while"? This leads to interesting considerations of the convergence rates and "rapid mixing" of Markov chains.)

Example 2.1 *The hard-core model.*

A *hard-core colouring* of a graph $G = (V,E)$ is a mapping

$$\xi : V \to \{0,1\} \qquad \text{("empty" vs. "occupied" sites)}$$

such that

$$(i,j) \in E \;\Rightarrow\; \xi(i) = 0 \vee \xi(j) = 0 \qquad \text{(no two occupied sites are adjacent)}$$

E.g. on a lattice graph, the hard-core colouring condition models an exclusion principle, whereby a "particle" at one site excludes the presence of "particles" at neighbouring sites, cf. Figure 10. In computer science terms, a hard-core colouring of a graph G corresponds to an independent set of nodes from G.

Denote by μ_G the uniform distribution over all the Z_G valid hard-core colourings of G. We would like to sample colourings according to μ_G, e.g. in order to compute the expected number of ones in a valid colouring:

$$E(n(X)) = \sum_{\xi \in \{0,1\}^V} n(\xi)\mu_G(\xi) = \frac{1}{Z_G} \sum_{\xi \in \{0,1\}^V} n(\xi) I_{[\xi \text{ is valid}]},$$

where $n(\xi)$ denotes the number of ones in colouring ξ.

However, the combinatorial structure of distribution μ_G is quite complicated; it is far from clear how one could pick a random valid hard-core colouring of graph G. (Even computing their total number Z_G is likely to be a so called #P-complete problem, and thus not solvable in polynomial time unless P = NP.)

Given a graph $G = (V,E)$, $V = \{1,\dots,n\}$, let us consider the following Markov chain $(X_0, X_1, \dots)$ on the space of valid hard-core colourings of G:

- Initially choose X_0 to be any valid hard-core colouring of G.
- Then, given colouring X_t, generate colouring X_{t+1} as follows:
  1. Choose some node $i \in V$ uniformly at random.
  2. If all the neighbours of i have colour 0 in X_t, then let $X_{t+1}(i) = 1$ with probability $1/2$ and $X_{t+1}(i) = 0$ with probability $1/2$.
  3. At all other nodes j, let $X_{t+1}(j) = X_t(j)$.

It can be seen that the chain thus determined is irreducible (since all colourings communicate via the all-zeros colouring) and aperiodic (since for any colouring ξ, $P_{\xi\xi} > 0$).

To see that the chain has μ_G as its unique stationary distribution, it suffices to check the detailed balance conditions with respect to μ_G. Let ξ, ξ' be two different colourings. If they differ at more than one node, then $P_{\xi\xi'} = P_{\xi'\xi} = 0$, so it suffices to check the case where $\xi(i) \neq \xi'(i)$ at a single node i. But then

$$\mu_G(\xi) P_{\xi\xi'} = \frac{1}{Z_G} \cdot \frac{1}{n} \cdot \frac{1}{2} = \mu_G(\xi') P_{\xi'\xi}.$$

The above hard-core sampling algorithm is a special case of a *Gibbs sampler* for a target distribution π on a state space of the form $S = C^V$.

The general principle is to choose in step 2 of the state update rule the new value for $X_{t+1}(i)$ according to the *conditional* π*-distribution*:

$$\Pr_{MC}(X_{t+1}(i) = c) = \Pr_{\pi}(\xi(i) = c \mid \xi(j) = X_t(j), j \neq i).$$

(In addition, the chain needs to be initialised in a state X_0 that has nonzero π-probability.) It can be seen that the chain so obtained is aperiodic and has π as a stationary distribution. Whether the chain is also irreducible depends on which states ξ have nonzero π-probability.

Example 2.2 *Sampling graph k-colourings.* Let $G = (V, E)$ be a graph. The following is a Gibbs sampler for the uniform distribution in the space $S = \{1, \ldots, k\}^V$ of k-colourings of G:

- Initially choose X_0 to be any valid k-colouring of G. (Of course, finding a valid k-colouring is an NP-complete problem for $k \geq 3$, but let us not worry about that).
- Then, given colouring X_t, generate colouring X_{t+1} as follows:
  1. Choose some node $i \in V$ uniformly at random.
  2. Let C' be the set of colours assigned by X_t to the neighbours of i:

     $$C' = \{X_t(j) \mid (i, j) \in E\}.$$

     (Note that $|C'| < k$.) Choose a colour for $X_{t+1}(i)$ uniformly at random from the set $\{1, \ldots, k\} \setminus C'$.
  3. At all other nodes j, let $X_{t+1}(j) = X_t(j)$.

Note that it is a nontrivial question whether this chain is irreducible or not.

Another general family of MCMC samplers are the *Metropolis chains*.

Let the state space S have some neighbourhood structure, so that it may be viewed as a (large) connected graph (S, N). Denote by $N(i)$ the set of neighbours of state i, and let $d_i = |N(i)|$. We assume that the neighbourhood structure is symmetric, so that $i \in N(j)$ if and only if $j \in N(i)$.

Then the (basic) *Metropolis sampler* for distribution π on S operates as follows:

- Initially choose X_0 to be some state $i \in S$.
- Then, given state $X_t = i$, state X_{t+1} is obtained as follows:

  1. Choose some $j \in N(i)$ uniformly at random.
  2. With probability $\min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}$, accept $X_{t+1} = j$. Otherwise let $X_{t+1} = i$.

Thus, fully written out the transition probabilities are:

$$p_{ij} = \begin{cases} \frac{1}{d_i} \min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}, & \text{if } j \in N(i) \\ 0, & \text{if } j \notin N(i), j \neq i \\ 1 - \sum_{j \in N(i)} p_{ij}, & \text{if } j = i \end{cases}$$

To show that this chain has π as its stationary distribution, it suffices to check the detailed balance conditions:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall\, i, j \in S.$$

The conditions are trivial if $i = j$ or $j \notin N(i)$, so let us consider the case $j \in N(i)$. There are two subcases:

(i) Case $\frac{\pi_j d_i}{\pi_i d_j} \geq 1$: Then:

$$\begin{cases} \pi_i p_{ij} = \pi_i \cdot \frac{1}{d_i} \cdot 1 \\ \pi_j p_{ji} = \pi_j \cdot \frac{1}{d_j} \cdot \frac{\pi_i d_j}{\pi_j d_i} = \frac{\pi_i}{d_i} \end{cases}$$

(ii) Case $\frac{\pi_j d_i}{\pi_i d_j} < 1$: Then:

$$\begin{cases} \pi_i p_{ij} = \pi_i \cdot \frac{1}{d_i} \cdot \frac{\pi_j d_i}{\pi_i d_j} = \frac{\pi_j}{d_j} \\ \pi_j p_{ji} = \pi_j \cdot \frac{1}{d_j} \cdot 1 \end{cases}$$

(Note that in both cases $\pi_i p_{ij} = \pi_j p_{ji} = \min\{\frac{\pi_i}{d_i}, \frac{\pi_j}{d_j}\}$.) Hence π is a stationary distribution of the chain.

Furthermore, the chain is guaranteed to be aperiodic if there is at least one $i \in S$ such that $\frac{\pi_j d_i}{\pi_i d_j} < 1$ ($\Rightarrow\ p_{ii} > 0$) i.e. it is not the case that for all $i, j \in S$:

$$\frac{\pi_i}{d_i} = \frac{\pi_j}{d_j} = \text{ const.}$$

In the latter case the chain reduces to a simple random walk on the graph (S,N) with stationary distribution

$$\pi = \left[\frac{d_1}{d} \quad \frac{d_2}{d} \quad \cdots \quad \frac{d_n}{d}\right]$$

as seen earlier. Such a random walk is aperiodic, if and only if the graph (S,N) contains at least one odd cycle, i.e. if (S,N) is not bipartite.

3 Estimating the Convergence Rate of a Markov Chain

3.1 Second Eigenvalue, Conductance, Canonical Paths

Consider a regular Markov Chain on state set $S = \{1,\ldots,n\}$, with transition probability matrix $P = (p_{ij})$ and stationary distribution π.

We would like to measure the rate of convergence of the chain to π, e.g. in terms of the *total variation distance*:

$$\Delta_V^{(i)}(t) = d_V(\pi^{(i,t)}, \pi),$$

where $\pi_j^{(i,t)} = p_{ij}^{(t)}$, and

$$d_V(\rho,\pi) = \max_{A\subseteq S} |\rho(A) - \pi(A)| = \frac{1}{2}\sum_{j\in S} |\rho_j - \pi_j|.$$

However, we get somewhat tighter results by using instead of d_V the *relative pointwise distance*

$$d_{rp}^U(\rho,\pi) = \max_{j\in U} \frac{|\rho_j - \pi_j|}{\pi_j}.$$

Hence, we define our convergence rate function as:

$$\Delta^U(t) = \max_{i\in U} d_{rp}^U(\pi^{(i,t)},\pi) = \max_{i,j\in U} \frac{|p_{ij}^{(t)} - \pi_j|}{\pi_j}.$$

When we consider convergence over the whole state space, i.e. $U = S$, we denote simply:

$$\Delta(t) = \Delta^S(t).$$

Proposition 3.1 *For any two distributions* ρ, π, *where* $\pi_j > 0$ *for all* j*:*

$$d_V(\rho,\pi) \le \frac{1}{2} d_{rp}^S(\rho,\pi) \le \frac{1}{\min_j \pi_j} d_V(\rho,\pi).$$

Consequently, $\Delta_V^{(i)}(t) \le \frac{1}{2}\Delta(t)$ *for all* i,t. □

Define the *mixing time* of a given regular chain as

$$\tau(\varepsilon) = \min\{t \mid \Delta(t') \le \varepsilon \quad \forall\, t' \ge t\}.$$

In algorithmic applications, the details of the chain are often determined by some input x, in which case we write $\Delta_x(t)$, $\tau_x(\varepsilon)$ correspondingly.

A chain (more precisely, a family of chains determined by inputs x) is *rapidly mixing* if

$$\tau_x(\varepsilon) = \text{poly}\left(|x|, \ln\frac{1}{\varepsilon}\right).$$

Our goal is now to establish some techniques for analysing the convergence rates of Markov chains and proving them to be rapidly mixing.

Lemma 3.2 *A regular Markov chain with transition matrix* P *and stationary distribution* π *is reversible, if and only if the matrix* $D^{1/2}PD^{-1/2}$ *is symmetric, where* $D^{1/2} = diag(\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_n})$.

Proof. $D^{1/2}PD^{-1/2} = \left(D^{1/2}PD^{-1/2}\right)^T \Leftrightarrow DP = P^T D.$

Inspecting this condition coordinatewise shows that it is exactly the same as the reversibility condition $\pi_i p_{ij} = p_{ji}\pi_j \quad \forall\, i,j$. □

Now it is easy to see that the matrix $A = D^{1/2}PD^{-1/2}$ has the same eigenvalues as P: if λ is an eigenvalue of P with left eigenvector u, then for the vector $v = uD^{-1/2}$ we obtain

$$vA = uD^{-1/2}\left(D^{1/2}PD^{-1/2}\right) = uPD^{-1/2} = \lambda uD^{-1/2} = \lambda v.$$

Since matrix A is symmetric for reversible P, this shows that reversible P have real eigenvalues. By the Perron-Frobenius theorem they can thus be ordered as

$$\lambda_1 = 1 > \lambda_2 \ge \lambda_3 \ge \cdots \ge \lambda_n > -1.$$

Denote $\lambda_{max} = \max\{|\lambda_i| : 2 \le i \le n\} = \max\{\lambda_2, -\lambda_n\}$.

Theorem 3.3 *Let P be the transition matrix of a regular, reversible Markov chain, and other notations as above. Then for any $U \subseteq S$,*

$$\Delta^U(t) \leq \frac{\lambda_{max}^t}{\min_{i\in U} \pi_i}.$$

Proof. Let $e^1, \ldots, e^n$ be an orthonormal basis for $\mathbb{R}^n$ consisting of left eigenvectors of A, where vector e^i is associated to eigenvalue λ_i. Especially, $e^1 = \pi D^{-1/2} = [\sqrt{\pi_1}, \sqrt{\pi_2}, \ldots, \sqrt{\pi_n}]$.

Then A has a spectral representation

$$A = \sum_{i=1}^n \lambda_i (e^i)^T e^i = \sum_{i=1}^n \lambda_i E_i,$$

where $E_i = (e^i)^T e^i$. Clearly $E_i^2 = E_i$, and $E_i E_j = 0$ if $i \neq j$.

Thus, for any $t \geq 0$, $A^t = \sum_{i=1}^n \lambda_i^t E_i$, and hence

$$\begin{aligned} P^t &= D^{-1/2} A^t D^{1/2} = \sum_{i=1}^n \lambda_i^t \left(D^{-1/2}(e^i)^T\right)\left(e^i D^{1/2}\right) \\ &= 1\pi + \sum_{i=2}^n \lambda_i^t \left(D^{-1/2}(e^i)^T\right)\left(e^i D^{1/2}\right). \end{aligned}$$

In component form, this means:

$$p_{jk}^{(t)} = \pi_k + \sqrt{\frac{\pi_k}{\pi_j}} \sum_{i=2}^n \lambda_i^t e_j^i e_k^i.$$

Computing the relative pointwise distance convergence rate, we thus get for any $U \subseteq S$:

$$\begin{aligned} \Delta^U(t) &= \max_{j,k\in U} \frac{\left|\sum_{i=2}^n \lambda_i^t e_j^i e_k^i\right|}{\sqrt{\pi_j \pi_k}} \qquad (4) \\ &\leq \lambda_{max}^t \frac{\max_{j,k\in U}\left|\sum_{i=2}^n e_j^i e_k^i\right|}{\min_{j\in U}\pi_j} \\ &\leq \frac{\lambda_{max}^t}{\min_{j\in U}\pi_j} \qquad \text{(by the Cauchy-Schwarz inequality and normality).} \ \square \end{aligned}$$

Theorem 3.4 *With notation and assumptions as above,*

$$\Delta(t) \geq \lambda_{max}^t$$

for all even t. Moreover, if all eigenvalues of P are nonnegative, then the bound holds for all t.

Proof. Continuing from equation (4) above, when t is even or all eigenvalues are nonnegative, the following holds:

$$\Delta(t) = \Delta^S(t) \geq \max_{j\in S} \frac{\left|\sum_{i=2}^n \lambda_i^t (e_j^i)^2\right|}{\pi_j} \geq \lambda_{max}^t \max_{j\in S} \frac{(e_j^{i_0})^2}{\pi_j},$$

where e^{i_0} is a normalised eigenvector corresponding to eigenvalue with absolute value λ_{max}. Necessarily $(e_j^{i_0})^2 \geq \pi_j$ for some j for otherwise

$$||e^{i_0}|| = \sum_{j=1}^n (e_j^{i_0})^2 < \sum_{j=1}^n \pi_j = 1,$$

contradicting the normality of e^{i_0}. $\square$

Negative eigenvalues are often a nuisance, but they can always be removed, without affecting the convergence properties of the chain much, by adding appropriate self-loops to the states. E.g.:

Proposition 3.5 *With notation and assumptions as above, consider the chain determined by transition matrix $P' = \frac{1}{2}(I+P)$. This chain is then also regular and reversible, has same stationary distribution π, and its eigenvalues satisfy $\lambda_n' > 0$ and $\lambda_{max}' = \lambda_2' = \frac{1}{2}(1+\lambda_2)$.* $\square$

With Theorem 3.3 and Proposition 3.5 in mind, it is clear that the key to analysing convergence rates of reversible Markov chains is to find good techniques for bounding the second eigenvalue λ_2 away from 1.

An interesting and intuitive approach to this task is via the notion of "conductance" of a chain.

Given a finite, regular, reversible Markov chain $\mathcal{M}$ on the state space $S = \{1, \ldots, n\}$, transition probability matrix $P = (p_{ij})$ and stationary distribution $\pi = (\pi_i)$, we associate to $\mathcal{M}$ a weighted graph $G = (S, E, W)$, where $E = \{(i,j) \mid p_{ij} > 0\}$, and

the weights, or "capacities" on the edges correspond to the *ergodic flows* between states:

$$w_{ij} = \pi_i p_{ij} = \pi_j p_{ji}.$$

Given a state set $A \subseteq S$, the *volume* of A is defined as

$$V_A = \pi(A) = \sum_{i \in A} \pi_i,$$

and the *ergodic flow* out of A as

$$F_A = \sum_{\substack{i \in A \\ j \notin A}} \pi_i p_{ij} = \sum_{\substack{i \in A \\ j \notin A}} w_{ij} = w(A, \bar{A}).$$

(Note that $0 < F_A \leq V_A < 1$.)

Then the *conductance* of the cut $(A, \bar{A})$, or the *(weighted) expansion* of A is

$$\Phi_A = \frac{F_A}{V_A} = \frac{w(A, \bar{A})}{\pi(A)},$$

and finally the *conductance* of $\mathcal{M}$, or G, is obtained as

$$\Phi_M = \Phi(G) = \min_{0 < \pi(A) \leq 1/2} \Phi_A.$$

Since clearly $F_A = F_{\bar{A}}$ for any $\varnothing \neq A \subsetneq S$, this may equally well be defined as:

$$\Phi = \min_{\varnothing \neq A \subsetneq S} \max(\Phi_A, \Phi_{\bar{A}}).$$

Theorem 3.6 *For a regular reversible Markov chain with underlying graph G, the second eigenvalue of the transition matrix satisfies:*

(i)

$$\lambda_2 \leq 1 - \frac{\Phi(G)^2}{2};$$

(ii)

$$\lambda_2 \geq 1 - 2\Phi(G).$$

Proof. Later. □

Corollary 3.7 *With notation and assumptions as above, the convergence rates for the chain under consideration satisfy, for any $\varnothing \neq A \subsetneq S$ and $t \geq 0$:*

(i)

$$\Delta^A(t) \leq \frac{\left(1 - \Phi^2/2\right)^t}{\min\limits_{i \in A} \pi_i};$$

(ii)

$$\Delta(t) \geq (1 - 2\Phi)^t.$$

Corollary 3.8 *Consider a family of regular reversible chains where all eigenvalues are nonnegative, parameterised by some input string x, and with underlying graphs G_x. Then the chains are rapidly mixing, if and only if*

$$\Phi(G_x) \geq \frac{1}{p(|x|)},$$

for some polynomial p and all x.

Proof. According to Corollary 3.7 (i):

$$\begin{array}{lrcll}
 & \Delta(t) & \leq & \varepsilon & \\
\text{if} & \frac{\left(1-\Phi^2/2\right)^t}{\min_{i \in A} \pi_i} & \leq & \varepsilon & \\
\text{if} & t \cdot \underbrace{\ln\left(1 - \frac{\Phi^2}{2}\right)}_{\leq -\Phi^2/2} & \leq & \ln \varepsilon + \ln \pi_{min} & \\
\text{if} & -t\Phi^2/2 & \leq & \ln \varepsilon + \ln \pi_{min} & \\
\text{if} & t & \geq & \frac{2}{\Phi^2}\left(\ln \frac{1}{\varepsilon} + \ln \frac{1}{\pi_{min}}\right). &
\end{array}$$

Conversely, by Theorem 3.4 and Corollary 3.7 (ii):

$$\begin{array}{lrcll}
 & \Delta(t) & > & \varepsilon & \\
\text{if} & \lambda_2^t & > & \varepsilon & \\
\text{if} & t \ln \lambda_2 & > & \ln \varepsilon & \\
\text{if} & t \ln \frac{1}{\lambda_2} & < & \ln \frac{1}{\varepsilon} & \\
\text{if} & t \cdot \frac{1-\lambda_2}{\lambda_2} & < & \ln \frac{1}{\varepsilon} & \ln \frac{1}{\lambda} = \ln\left(1 + \frac{1-\lambda}{\lambda}\right) \leq \frac{1-\lambda}{\lambda},\ 0 < \lambda \leq 1 \\
\text{if} & t & < & \frac{\lambda_2}{1-\lambda_2} \cdot \ln \frac{1}{\varepsilon} & \\
\text{if} & t & < & \frac{1-2\Phi}{2\Phi} \ln \frac{1}{\varepsilon} & \frac{\lambda}{1-\lambda} \text{ increasing in } \lambda,\ 1 - 2\Phi \leq \lambda_2.
\end{array}$$

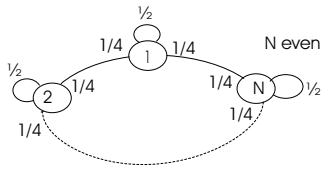


Figure 11: Random walk on a ring.

Consequently,

$$\frac{1-2\Phi(G_x)}{2\Phi(G_x)}\ln\frac{1}{\varepsilon} \le \tau_x(\varepsilon) \le \frac{2}{\Phi(G_x)^2}\left(\ln\frac{1}{\varepsilon}+\ln\frac{1}{\pi^x_{min}}\right).\Box$$

Example 3.1 *Random walk on a ring.* Consider the regular, reversible Markov chain described by the graph in Figure 11.

Clearly the stationary distribution is $\pi = [\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}]$.

The conductance $\Phi_A = F_A/V_A$ of a cut $(A,\bar{A})$ is minimised by choosing A to consist of any $n/2$ consecutive nodes on the cycle, e.g. $A = \{1,2,\ldots,n/2\}$. Then

$$\Phi = \Phi_A = \frac{F_A}{V_A} = \frac{\sum_{\substack{i\in A\\ j\notin A}}\pi_i p_{ij}}{\sum_{i\in A}\pi_i} = \frac{2\cdot\frac{1}{n}\cdot\frac{1}{4}}{\frac{n}{2}\cdot\frac{1}{n}} = \frac{1/2n}{1/2} = \frac{1}{n}.$$

Thus, by Theorem 3.6, the second eigenvalue satisfies:

$$1-\frac{2}{n} \le \lambda_2 \le 1-\frac{1}{2n^2},$$

by Corollary 3.7, the convergence rate satisfies

$$\left(1-\frac{2}{n}\right)^t \le \Delta(t) \le n\cdot\left(1-\frac{1}{2n^2}\right)^t,$$

and by Corollary 3.8, the mixing time satisfies:

$$\begin{aligned} &\frac{1-2/n}{2/n}\ln\frac{1}{\varepsilon} \le \tau(\varepsilon) \le 2n^2\left(\ln\frac{1}{\varepsilon}+\ln n\right)\\ \Leftrightarrow\quad &\left(\frac{n}{2}-1\right)\cdot\ln\frac{1}{\varepsilon} \le \tau(\varepsilon) \le 2n^2\left(\ln n+\ln\frac{1}{\varepsilon}\right).\end{aligned}$$

It is an intriguing, and nontrivial, exercise to work out the value of λ_2 exactly in this case, in order to determine whether the mixing times $\tau(\varepsilon)$ are closer to the given lower or upper bounds as a function of n.

Let us now return to the proof of Theorem 3.6, establishing the connection between the second-largest eigenvalue and the conductance of a Markov chain. Recall the statement of the Theorem:

Theorem 3.6 *Let $\mathcal{M}$ be a finite, regular, reversible Markov chain and λ_2 the second-largest eigenvalue of its transition matrix. Then:*

(i) $\lambda_2 \le 1-\frac{\Phi^2}{2}$,

(ii) $\lambda_2 \ge 1-2\Phi$.

Proof. (i) The approach here is to bound Φ in terms of the eigenvalue gap of $\mathcal{M}$, i.e. to show that $\Phi^2/2 \le 1-\lambda_2$, from which the claimed result follows.

Thus, consider the eigenvalue $\lambda = \lambda_2$. (The following proof does not in fact depend on this particular choice of eigenvalue $\lambda \ne 1$, but since we are proving an upper bound of the form $\Phi^2/2 \le 1-\lambda$, all other eigenvalues yield weaker bounds than λ_2.)

Let e be a left eigenvector $e \ne 0$ such that $eP = \lambda e$. Now e must contain both positive and negative components, since $\sum_i e_i = 0$ as can be seen:

$$\begin{aligned} eP = \lambda e \;&\Leftrightarrow\; \sum_i e_i p_{ij} = \lambda e_j \quad \forall\, j\\ &\Rightarrow\; \sum_j\sum_i e_i p_{ij} = \sum_i e_i \underbrace{\sum_j p_{ij}}_{=1} = \lambda\sum_j e_j\\ &\overset{\lambda\ne 1}{\Rightarrow}\; \sum_i e_i = 0.\end{aligned}$$

Define $A = \{i \mid e_i > 0\}$. Assume, without loss of generality, that $\pi(A) \le 1/2$. (Otherwise we may replace e by $-e$ in the following proof.)

Define further a "π-normalised" version of $e \restriction A$:

$$u_i = \begin{cases} e_i/\pi_i, & \text{if } i \in A\\ 0, & \text{if } i \notin A\end{cases}$$

Without loss of generality we may again assume that the states are indexed so that $u_1 \ge u_2 \ge \ldots \ge u_r > u_{r+1} = \ldots = u_n = 0$, where $r = |A|$.

In the remainder of the proof, the following quantity will be important:

$$D = \frac{\sum_{i<j} w_{ij}(u_i^2 - u_j^2)}{\sum_i \pi_i u_i^2}.$$

We shall prove the following claims:

(a) $\Phi \le D$,

(b) $D^2/2 \le 1-\lambda$,

which suffice to establish our result.

Proof of (a): Denote $A_k = \{1,\dots,k\}$, for $k = 1,\dots,r$. The numerator in the definition of D may be expressed in terms of the ergodic flows out of the A_k as follows:

$$\begin{aligned}\sum_{i<j} w_{ij}(u_i^2 - u_j^2) &= \sum_{i<j} w_{ij} \sum_{i\le k<j} (u_k^2 - u_{k+1}^2) \\ &= \sum_{k=1}^{r} (u_k^2 - u_{k+1}^2) \sum_{\substack{i\in A_k \\ j\notin A_k}} w_{ij} \\ &= \sum_{k=1}^{r} (u_k^2 - u_{k+1}^2) F_{A_k}.\end{aligned}$$

Now the capacities of the A_k satisfy $\pi(A_k) \le \pi(A) \le 1/2$, so by definition $\Phi_{A_k} \ge \Phi \;\Rightarrow\; F_{A_k} \ge \Phi \cdot \pi(A_k)$. Thus,

$$\begin{aligned}\sum_{i<j} w_{ij}(u_i^2 - u_j^2) &= \sum_{k=1}^{r} (u_k^2 - u_{k+1}^2) F_{A_k} \\ &\ge \Phi \cdot \sum_{k=1}^{r} (u_k^2 - u_{k+1}^2)\pi(A_k) \\ &= \Phi \cdot \sum_{k=1}^{r} (u_k^2 - u_{k+1}^2) \sum_{i=1}^{k} \pi_i \\ &= \Phi \cdot \sum_{i=1}^{r} \pi_i \sum_{k=i}^{r} (u_k^2 - u_{k+1}^2) \\ &= \Phi \cdot \sum_{i\in A} \pi_i u_i^2.\end{aligned}$$

Hence,

$$\Phi \le \frac{\sum_{i<j} w_{ij}(u_i^2 - u_j^2)}{\sum_i \pi_i u_i^2} = D.$$

Proof of (b): We introduce one more auxiliary expression:

$$E = \frac{\sum_{i<j} w_{ij}(u_i - u_j)^2}{\sum_i \pi_i u_i^2},$$

and establish that: (b') $D^2 \le 2E$, (b") $E \le 1-\lambda$. This will conclude the proof of Theorem 3.6 (i).

Proof of (b'): Observe first that

$$\sum_{i<j} w_{ij}(u_i + u_j)^2 \le 2\sum_{i<j} w_{ij}(u_i^2 + u_j^2) \le 2\sum_{i\in A} \pi_i u_i^2.$$

Then, by the Cauchy-Schwartz inequality:

$$\begin{aligned}D^2 &= \left(\frac{\sum_{i<j} w_{ij}(u_i^2 - u_j^2)}{\sum_i \pi_i u_i^2}\right)^2 \\ &\le \left(\frac{\sum_{i<j} w_{ij}(u_i + u_j)^2}{\sum_i \pi_i u_i^2}\right)\left(\frac{\sum_{i<j} w_{ij}(u_i - u_j)^2}{\sum_i \pi_i u_i^2}\right) \le 2E.\end{aligned}$$

Proof of (b"): Denote $Q = I - P$. Then $eQ = (1-\lambda)e$ and thus

$$eQu^T = (1-\lambda)eu^T = (1-\lambda)\sum_{i=1}^{r} \pi_i u_i^2.$$

On the other hand, writing eQu^T out explicitly:

$$\begin{array}{rcl|l} eQu^T & = & \sum_{i=1}^{n}\sum_{j=1}^{r} q_{ij}e_iu_j & q_{ij} = -p_{ij} = -\frac{w_{ij}}{\pi_i}, \quad i \neq j \\ & \geq & \sum_{i=1}^{r}\sum_{j=1}^{r} q_{ij}e_iu_j & q_{ii} = 1 - p_{ii} = \sum_{i\neq j} p_{ij} \\ & = & -\sum_{i\in A}\sum_{\substack{j\in A\\ j\neq i}} w_{ij}u_iu_j + \sum_{i\in A}\sum_{\substack{j\in \bar{A}\\ j\neq i}} w_{ij}u_i^2 & e_i = \pi_iu_i, \quad i \in A \\ & = & -2\sum_{i<j} w_{ij}u_iu_j + \sum_{i<j} w_{ij}(u_i^2+u_j^2) & \\ & = & \sum_{i<j} w_{ij}(u_i-u_j)^2. & \end{array}$$

Thus,

$$E\cdot\sum_i \pi_iu_i^2 = \sum_{i<j} w_{ij}(u_i-u_j)^2 \leq eQu^T = (1-\lambda)\cdot\sum_i \pi_iu_i^2 \;\Rightarrow\; E \leq 1-\lambda.$$

(ii) Given the stationary distribution vector $\pi \in \mathbb{R}^n$, define an inner product $\langle\cdot,\cdot\rangle_\pi$ in $\mathbb{R}^n$ as:

$$\langle u, v\rangle_\pi = \sum_{i=1}^{n} \pi_iu_iv_i.$$

By (a slight modification of) a standard result (the Courant-Fischer minimax theorem) in matrix theory, and the fact that P is reversible with respect to π, implying $\langle u, Pv\rangle_\pi = \langle Pu, v\rangle_\pi$, one can characterise the eigenvalues of P as:

$$\lambda_1 = \max\left\{\frac{\langle u,Pu\rangle_\pi}{\langle u,u\rangle_\pi} \mid u \neq 0\right\},$$
$$\lambda_2 = \max\left\{\frac{\langle u,Pu\rangle_\pi}{\langle u,u\rangle_\pi} \mid u \perp \pi, u \neq 0\right\}, \text{ etc.}$$

In particular,

$$\lambda_2 \geq \frac{\langle u,Pu\rangle_\pi}{\langle u,u\rangle_\pi} \text{ for any } u \neq 0 \text{ such that } \sum_i \pi_iu_i = 0. \tag{5}$$

Given a set of states $A \subseteq S$, $0 < \pi(A) \leq 1/2$, we shall apply the bound (5) to the vector u defined as:

$$u_i = \begin{cases} \dfrac{1}{\pi(A)}, & \text{if } i \in A \\ -\dfrac{1}{\pi(A)}, & \text{if } i \in \bar{A} \end{cases}$$

Clearly

$$\sum_i \pi_iu_i = \sum_{i\in A}\frac{\pi_i}{\pi(A)} - \sum_{i\in\bar{A}}\frac{\pi_i}{\pi(A)} = 1-1 = 0, \text{ and}$$

$$\langle u,u\rangle_\pi = \sum_i \pi_iu_i^2 = \sum_{i\in A}\frac{\pi_i}{\pi(A)^2} + \sum_{i\in\bar{A}}\frac{\pi_i}{\pi(\bar{A})^2} = \frac{1}{\pi(A)} + \frac{1}{\pi(\bar{A})},$$

so let us compute the value of $\langle u, Pu\rangle_\pi$.

The task can be simplified by representing P as $P = I_n - (I_n - P)$, and first computing $\langle u, (I-P)u\rangle_\pi$:

$$\begin{aligned} \langle u,(I-P)u\rangle_\pi &= \sum_i \pi_iu_i\sum_j (I-P)_{ij}u_j \\ &= -\sum_i\sum_{j\neq i}\pi_iu_ip_{ij}u_j + \sum_i\sum_{j\neq i}\pi_iu_ip_{ij}u_i \\ &= \sum_i\sum_{j\neq i}\pi_ip_{ij}(u_i^2 - u_iu_j) \\ &= \sum_{i<j}\pi_ip_{ij}(u_i-u_j)^2 \\ &= \sum_{\substack{i\in A\\ j\neq i}}\pi_ip_{ij}\left(\frac{1}{\pi(A)}+\frac{1}{\pi(\bar{A})}\right)^2 \\ &= \left(\frac{1}{\pi(A)}+\frac{1}{\pi(\bar{A})}\right)^2 F_A. \end{aligned}$$

Thus,

$$\begin{aligned} \lambda_2 &\geq \frac{\langle u,Pu\rangle_\pi}{\langle u,u\rangle_\pi} = \frac{1}{\langle u,u\rangle_\pi}\Big(\langle u,u\rangle_\pi - \langle u,(I-P)u\rangle_\pi\Big) \\ &= 1 - \frac{1}{\langle u,u\rangle_\pi}\cdot\langle u,(I-P)u\rangle_\pi \\ &= 1 - \left(\frac{1}{\pi(A)}+\frac{1}{\pi(\bar{A})}\right)^{-1}\left(\frac{1}{\pi(A)}+\frac{1}{\pi(\bar{A})}\right)^2\cdot F_A \\ &= 1 - \left(\frac{1}{\pi(A)}+\frac{1}{\pi(\bar{A})}\right)\cdot F_A \\ &\geq 1 - \frac{2}{\pi(A)}\cdot F_A = 1 - 2\Phi_A. \end{aligned}$$

Since the bound (6) holds for any $A \subseteq S$ such that $0 < \pi(A) \leq 1/2$, it follows that it holds also for the conductance

$$\Phi = \min_{0<\pi(A)\leq 1/2} \Phi_A.$$

Thus, we have shown that $\lambda_2 \geq 1 - 2\Phi$, which completes the proof. □

Despite the elegance of the conductance approch, it can be sometimes (often?) difficult to apply in practice – computing graph conductance can be quite difficult. Also the bounds obtained are not necessary the best possible; in particular the square in the upper bound $\lambda_2 \leq 1 - \Phi^2/2$ is unfortunate.

An alternative approch, which is sometimes easier to apply, and can even yield better bounds, is based on the construction of so called "canonical paths" between states of a Markov chain.

Consider again a regular, reversible Markov chain with stationary distribution π, represented as a weighted graph with node set S and edge set $E = \{(i,j) \mid p_{ij} > 0\}$. The weight, or capacity, w_e associated to edge $e = (i,j)$ corresponds to the ergodic flow $\pi_i p_{ij}$ between states i and j.

Specify for each pair of states $k, l \in S$ a *canonical path* γ_{kl} connecting them. The paths should intuitively be chosen as short and as nonoverlapping as possible. (For precise statements, see Theorems 3.9 and 3.11 below.)

Denote $\Gamma = \{\gamma_{kl} \mid k, l \in S\}$ and define the unweighted and weighted *edge loading* induced by Γ on an edge $e \in E$ as:

$$\begin{aligned} \rho_e &= \frac{1}{w_e} \sum_{\gamma_{kl} \ni e} \pi_k \pi_l \\ \bar{\rho}_e &= \frac{1}{w_e} \sum_{\gamma_{kl} \ni e} \pi_k \pi_l |\gamma_{kl}|, \end{aligned}$$

where $|\gamma_{kl}|$ is the length (number of edges) of path γ_{kl}. (Note that here the edges are considered to be *oriented*, so that only paths crossing an edge $e = (i,j)$ in the direction from i to j are counted in determining the loading of e.) The *maximum edge loading* induced by Γ is then:

$$\begin{aligned} \rho &= \rho(\Gamma) = \max_{e \in E} \rho_e \\ \bar{\rho} &= \bar{\rho}(\Gamma) = \max_{e \in E} \bar{\rho}_e. \end{aligned}$$

Theorem 3.9 *For any regular, reversible Markov chain and any choice of canonical paths,*

$$\Phi \geq \frac{1}{2\rho}.$$

Proof. Represent the chain as a weighted graph G, where the weight (capacity) on edge $e = (i,j)$ is defined as:

$$w_{ij} = \pi_i p_{ij} = \pi_j p_{ji}.$$

Every set of states $A \subseteq S$ determines a cut $(A, \bar{A})$ in G, and the conductance of the cut corresponds to its *relative capacity*:

$$\Phi_A = \frac{w(A,\bar{A})}{V_A} = \frac{1}{\pi(A)} \sum_{i \in A, j \in \bar{A}} w_{ij}.$$

Let then A be a set with $0 < \pi(A) \leq \frac{1}{2}$ that minimises Φ_A, and thus has $\Phi_A = \Phi$. Assume some choice of canonical paths $\Gamma = \{\gamma_{ij}\}$, and assign to each path γ_{ij} a "flow" of value $\pi_i \pi_j$. Then the total amount of flow crossing the cut $(A, \bar{A})$ is

$$\sum_{i \in A, j \in \bar{A}} \pi_i \pi_j = \pi(A)\pi(\bar{A}),$$

but the cut edges, i.e. edges crossing the cut, have only total capacity $w(A,\bar{A})$. Thus, some cut edge e must have loading

$$\rho_e = \frac{1}{w_e} \sum_{\gamma_{ij} \ni e} \pi_i \pi_j \geq \frac{\pi(A)\pi(\bar{A})}{w(A,A)} \geq \frac{\pi(A)}{2w(A,A)} = \frac{1}{2\Phi}.$$

The result follows. □

Corollary 3.10 *With notations and assumptions as above,*

$$\lambda_2 \leq 1 - \frac{1}{8\rho^2}.$$

Proof. From Theorems 3.6 and 3.9. □

A more advanced proof yields a tighter result:

Theorem 3.11 *With notations and assumptions as above:*

(i) $\lambda_2 \leq 1 - \dfrac{1}{\bar{\rho}}$

(ii) $\Delta(t) \leq \dfrac{(1 - 1/\bar{\rho})^t}{\min\limits_{i \in A} \pi_i}$

(iii) $\tau(\varepsilon) \le \bar{\rho}\left(\ln\frac{1}{\varepsilon} + \ln\frac{1}{\pi_{min}}\right)$.□

Example 3.2 *Random walk on a ring.* Let us consider again the cyclic random walk of Figure 11. Clearly the stationary distribution is $\pi = [\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}]$, and the ergodic flow on each edge $e = (i, i\pm 1)$ is

$$w_e = \pi_i p_{i,i\pm 1} = \frac{1}{n}\cdot\frac{1}{4} = \frac{1}{4n}.$$

An obvious choice for a canonical path connecting nodes k, l is the shortest one, with length

$$|\gamma_{kl}| = \min\{|l-k|, n-|l-k|\}.$$

It is fairly easy to see that each (oriented) edge is now travelled by 1 canonical path of length 1, 2 of length 2, 3 of length 3, ..., $\frac{n}{2}$ of length $\frac{n}{2}$ (actually the last one is just an upper bound). Thus:

$$\begin{array}{rcl} \bar{\rho} & = & \max_e \frac{1}{w_e}\sum_{\gamma_{kl}\ni e}\pi_k\pi_l|\gamma_{ij}| \le 4n\sum_{r=1}^{n/2}\frac{1}{n^2}\cdot r^2 \\ & = & \frac{4}{n}\cdot\frac{1}{6}\cdot\frac{n}{2}\cdot\left(\frac{n}{2}+1\right)\cdot(n+1) = \frac{1}{6}(n+1)(n+2) \\ \Rightarrow & & \\ \tau(\varepsilon) & \le & \frac{1}{6}(n+1)(n+2)\left(\ln n + \ln\frac{1}{\varepsilon}\right) \\ & = & \frac{1}{6}n^2\left(\ln n + \frac{1}{\varepsilon}\right) + O\left(n\left(\ln n + \ln\frac{1}{\varepsilon}\right)\right). \end{array}$$

Example 3.3 *Sampling permutations.* Let us consider the Markov chain whose states are all possible permutations of $[n] = \{1, 2, \ldots, n\}$, and for any permutations $s, t \in S_n$:

$$p_{st} = \begin{cases} \frac{1}{2}, & \text{if } s = t, \\ \frac{1}{2}\cdot\binom{n}{2}^{-1}, & \text{if } s \text{ can be changed to } t \text{ by transposing two elements,} \\ 0, & \text{otherwise} \end{cases}$$

Thus, e.g. for $n = 3$ we obtain the transition graph in Figure 12.

Clearly, the stationary distribution for this chain is $\pi = \left[\frac{1}{n!}, \frac{1}{n!}, \ldots, \frac{1}{n!}\right]$, and the ergodic flow on each edge $\tau = (s,t)$, with $s \ne t$, $p_{st} > 0$, is:

$$w_\tau = \pi_s p_{st} = \frac{1}{n!}\cdot\frac{1}{2}\cdot\binom{n}{2}^{-1}.$$

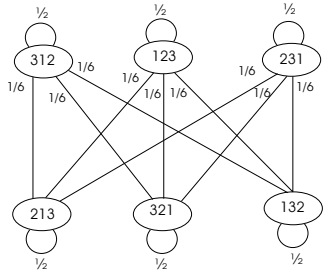


Figure 12: Transition graph for three-element permutations.

A natural canonical path connecting permutation s to permutation t is now obtained as follows:

$$s = s_0 \to s_1 \to s_2 \to \cdots \to s_{n-1} = t.$$

where at each $s_k, s_k(k) = t(k)$. (Thus, each s_k matches t up to element k, $s_k(1\ldots k) = t(1\ldots k)$.)

Thus, e.g. the canonical path connecting $s = (1234)$ to $t = (3142)$ is as follows:

$$(1234) \to \overbrace{(3|214)}^{\omega} \xrightarrow{\tau} \overbrace{(31|24)}^{\omega'} \to (314|2).$$

Now let us denote the set of canonical paths containing a given transition $\tau : \omega \to \omega'$ by $\Gamma(\tau)$. We shall upper bound the size of $\Gamma(t)$ by constructing an injective mapping $\sigma_\tau : \Gamma(\tau) \to S_n$. Obviously, the existence of such a mapping implies that $|\Gamma(\tau)| \le n!$.

Suppose τ transposes locations $k+1$ and l, $k+1 < l$, of permutation ω. Then for any $\langle s,t\rangle \in \Gamma(\tau)$, define the permutation $z = \sigma_\tau(s,t)$ as follows:

1. Place the elements in $\omega(1\ldots k)$ in the locations they appear in s. (Note that permutation ω is given and fixed as part of τ.)
2. Place the remaining elements in the remaining locations in the order they appear in t.

Thus, for example in the above example case:

$$\sigma_\tau(\langle 1234\rangle, \langle 3142\rangle) \to (\_\ \_\ 3\ \_) \to \underbrace{(1432)}_{z}$$

$$\omega = (3|214), \qquad k = 1$$

Why is this mapping an injection, i.e. how do we recover s and t from a knowledge of τ and $z = \sigma_\tau(s,t)$? The reasoning goes as follows:

1. $t = \omega(1\ldots k)+$ "other elements in same order as in z"
2. $s =$ "elements in $\omega(1\ldots k)$ at locations indicated in z" + "other elements in locations deducible from the transposition path $s = s_0 \to s_1 \to \cdots \to s_k = \omega$"

This is somewhat tricky, so let us consider an example. Say $\omega = (3\quad 1|2\quad 4)$, $k = 2$, $z = (1\quad 4\quad 3\quad 2)$. Then:

1. $t = (3\quad 1|\_\quad \_) + (\_\quad \_|4\quad 2) = (3\quad 1|4\quad 2)$
2.

$$
\begin{array}{rrcllcl}
 & s & = & s_0 = (1\ \_\ 3\ \_) & & s_0 = (1\ \_\ 3\ \_) \\
 & & & s_1 = (3|\ \_\ \_\ \_) & \Rightarrow & s_1 = (3|\ 2\ 1\ \_) \\
 & \omega & = & s_2 = (3\ 1|\ 2\ 4) & & s_2 = (3\ 1|\ 2\ 4) \\
\hline
\therefore & s & = & s_0 = (1\ 2\ 3\ 4) & & s_0 = (1\ 2\ 3\ 4) \\
 & & & s_1 = (3|\ 2\ 1\ 4) & \Rightarrow & s_1 = (3|\ 2\ 1\ 4) \\
 & \omega & = & s_2 = (3\ 1|\ 2\ 4) & & s_2 = (3\ 1|\ 2\ 4)
\end{array}
$$

Thus, we know that for each transition τ,

$$|\Gamma(\tau)| \le n!$$

We can now obtain a bound on the unweighted maximum edge loading induced by our collection of canonical paths:

$$
\begin{aligned}
\rho &= \max_{\tau\in E} \frac{1}{q_\tau} \sum_{\langle s,t\rangle \in \Gamma(\tau)} \pi_s \pi_t \le \left(\frac{1}{n!}\cdot\frac{1}{2}\cdot\binom{n}{2}^{-1}\right)^{-1} \cdot n! \cdot \left(\frac{1}{n!}\right)^2 \\
&= 2n!\binom{n}{2}\cdot n!\cdot(\frac{1}{n!})^2 = 2\cdot\binom{n}{2} = n(n-1).
\end{aligned}
$$

By Theorem 3.9, the conductance of this chain is thus $\Phi \ge \frac{1}{2n(n-1)}$, and by Corollary 3.8, its mixing time is thus bounded by

$$
\begin{aligned}
\tau_n(\varepsilon) &\le \frac{2}{\Phi^2}\left(\ln\frac{1}{\varepsilon} + \ln\frac{1}{\pi_{\min}}\right) \le 2(2n(n-1))^2\left(\ln\frac{1}{\varepsilon} + \ln n!\right) \\
&= O\left(n^4\left(n\ln n + \ln\frac{1}{\varepsilon}\right)\right).
\end{aligned}
$$

3.2 Coupling

An important "classical" approach to obtaining convergence results for Markov chains is the *coupling method*. As a simple case, let $\mathcal{M} = (X_0, X_1, \ldots)$ and $\mathcal{N} = (Y_0, Y_1, \ldots)$ be two independent Markov chains with the same state space $S = \{1,\ldots,n\}$ and the same regular transition matrix $P = (p_{ij})$, and consequently the same stationary distribution π.

Thus, if one considers the Markov chain $\mathcal{M} \times \mathcal{N}$ with random variables $Z_t = (X_t, Y_t)$, one obtains transition probabilities

$$
\begin{aligned}
p^Z_{ij,kl} &= \Pr(Z_t = (k,l) \mid Z_{t-1} = (i,j)) \\
&= \Pr(X_t = k \mid X_{t-1} = i)\cdot\Pr(Y_t = l \mid Y_{t-1} = j) \\
&= p_{ik}p_{jl}.
\end{aligned}
$$

Moreover, since $\mathcal{M}$ and $\mathcal{N}$ are regular with stationary distribution π, then so is $\mathcal{M} \times \mathcal{N}$ with stationary distribution $\pi^Z = \pi^T\pi$ (i.e. $\pi^Z_{ij} = \pi_i\pi_j$).

Note once more that "projected" (marginalised) to its first or second component, $\mathcal{M} \times \mathcal{N}$ yields realisations of the same process, i.e.

$$
\begin{aligned}
&\Pr(Z_t = (k,*) \mid Z_0 = (k_0,l_0)) = \Pr(X_t = k \mid X_0 = k_0) \\
&= p^{(t)}_{k_0k}, \text{ independent of } l_0; \\
&\Pr(Z_t = (*,l) \mid Z_0 = (k_0,l_0)) = \Pr(Y_t = l \mid Y_0 = l_0) \\
&= p^{(t)}_{l_0l}, \text{ independent of } k_0.
\end{aligned}
\tag{6}
$$

Now define a random variable T that for any realisation of $\mathcal{M} \times \mathcal{N}$ indicates the first time at which X_t and Y_t have the same value, i.e. their *coupling time*:

$$T = \inf\{t \ge 0 | X_t = Y_t\}.$$

One can in fact modify the chain $\mathcal{M} \times \mathcal{N}$ so that after coupling the X- and Y-components not just have the same distributions, but in fact strictly the same values (i.e. $X_t = Y_t\ \forall\ t \ge T$), yet the marginal properties (6) stay the same. Simply define $X'_t = (X'_t, Y_t)$, where

$$X'_t = \begin{cases} X_t, & t < T, \\ Y_t, & t \ge T. \end{cases}$$

Let us denote the resulting nonhomogeneous chain by $\mathcal{M}\,|\mathcal{N}$. Now the projections of $\mathcal{M}\,|\mathcal{N}$ to its X- and Y-components are surely not independent, but viewed in isolation, as marginals of $\mathcal{M}\,|\mathcal{N}$, they have exactly the same stochastic properties.

In particular, in a coupled chain $\mathcal{M}\ |\mathcal{N}$, let us fix an arbitrary initial state $X_0 = k_0$ for $\mathcal{M}$, and similarly $Y_0 = l_0$ for $\mathcal{N}$, and denote the respective time t distributions as $p^{(t)} = (p^{(t)}_{k_0 k})_k$ and $q^{(t)} = (p^{(t)}_{l_0 l})_l$. Then for any $A \subseteq S$:

$$\begin{aligned} p^{(t)}(A) &= \Pr(X_t \in A) \\ &\geq \Pr(Y_t \in A \wedge X_t = Y_t) \\ &= 1 - \Pr(Y_t \notin A \vee X_t \neq Y_t) \\ &\geq 1 - \Pr(Y_t \notin A) - \Pr(X_t \neq Y_t) \\ &= \Pr(Y_t \in A) - \Pr(t < T) \\ &= q^{(t)}(A) - \Pr(t < T), \end{aligned}$$

i.e. $q^{(t)}(A) - p^{(t)}(A) \leq \Pr(t < T)$. A similar argument shows that also $p^{(t)}(A) - q^{(t)}(A) \leq \Pr(t < T)$, and so for any $A \subseteq S$, $|p^{(t)}(A) - q^{(t)}(A)| \leq \Pr(T > t)$, implying that

$$d_V(p^{(t)}, q^{(t)}) = \sup_{A \subseteq S} |p^{(t)}(A) - q^{(t)}(A)| \leq \Pr(T > t). \tag{7}$$

If one establishes the coupling bound (7) so that it holds for arbitrary pairs of initial states, then it also holds for arbitrary initial distributions.

In particular, if the initial state of the chain Y is chosen according to the stationary distribution π, then $q^{(t)} = \pi$ for all $t \geq 0$, and one obtains the convergence bound:

$$d_V(p^{(t)}, \pi) = \frac{1}{2} \sum_i |p^{(t)}_i - \pi_i| \leq \Pr(T > t). \tag{8}$$

Example 3.4 *Random walk on a ring.* Consider again the cyclic random walk of Figure 11 with n states, n even. To obtain an upper bound on the coupling probability $\Pr(T > t)$, consider two independent copies (X_t), (Y_t) of the walk, initiated at $X_0 = 1$ and $Y_0 = \frac{n}{2} + 1$.

Denote $D_t = \min\{|Y_t - X_t|, n - |Y_t - X_t|\}$. Then $D_0 = \frac{n}{2}$, $0 \leq D_t \leq \frac{n}{2}$ for all t, $\Pr(D_{t+1} < D_t \mid D_t > 0) \geq \frac{1}{4}$, and $T = \inf\{t \mid D_t = 0\}$ (cf. Figure 13). Thus for any $k \geq 0$,

$$\Pr(T \leq k + \frac{n}{2} \mid T > k) \geq (\frac{1}{4})^{n/2} = (\frac{1}{2})^n,$$

and consequently

$$\Pr(T > t) \leq (1 - 2^{-n})^{\lfloor t/(n/2) \rfloor}.$$

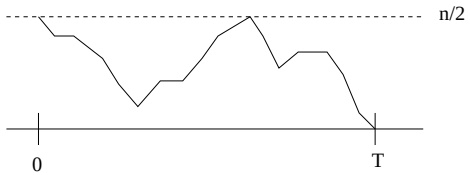


Figure 13: A realisation of the (D_t) chain.

Hence we obtain a geometric bound on the convergence rate of this walk:

$$d_V(p^{(t)}, \pi) \leq (1 - 2^{-n})^{\lfloor 2t/n \rfloor}.$$

The bound is not very tight, mainly because there is no systematic "drift" effect that would bring the chains (X_t) and (Y_t) closer to each other: they just eventually coalesce by random "fluctuation". A much more interesting application of the coupling technique will be presented below.

Generally speaking, a *coupling* of two Markov chains (X_t) and (Y_t) (or any stochastic processes) is a process $Z_t = (X'_t, Y'_t)$ that has (X_t) and (Y_t) as its marginal distributions.

In the case of finite Markov chains this means that:

$$\begin{aligned} \Pr(X'_{t+1} = k | X'_t = i, Y'_t = j) &= \Pr(X_{t+1} = k | X_t = i) = p^X_{ik}, \\ \Pr(Y'_{t+1} = l | X'_t = i, Y'_t = j) &= \Pr(X_{t+1} = l | Y_t = j) = p^Y_{jl}. \end{aligned} \tag{9}$$

The coupling conditions (9) are trivially satisfied by the independent coupling, where $p^Z_{ij,kl} = p^X_{ik} p^Y_{jl}$, but the more interesting couplings are the non-independent ones.

In the following Lemma, and also later in this section, mixing times are considered with respect to the total variation distance, i.e. for now

$$\tau(\varepsilon) = \tau^V(\varepsilon) = \min\left\{ t \mid d_V(p^{(i,s)}, \pi) \leq \varepsilon \quad \forall\, s \geq t \text{ and } \forall \text{ initial states } i \right\}.$$

Lemma 3.12 ("Coupling lemma") *Let $\mathcal{M}$ be a finite, regular Markov chain and $Z_t = (X_t, Y_t)$, $t \geq 0$, a coupling of two copies of $\mathcal{M}$ (i.e. (Z_t) is a Markov chain whose X- and Y-marginals satisfy the coupling conditions* (9) *with respect to the transition probabilities of $\mathcal{M}$). Suppose further that $t : (0,1] \to \mathbb{N}$ is a function such that given any $\varepsilon \in (0,1]$, $\Pr(X_t \neq Y_t) \leq \varepsilon$ holds for all $t \geq t(\varepsilon)$, uniformly over the choice of the initial state (X_0, Y_0). Then the mixing time $\tau(\varepsilon)$ of $\mathcal{M}$ is bounded above by $t(\varepsilon)$.*

Proof. Let $X_0 = i$ be arbitrary, and choose Y_0 according to the stationary distribution π of $\mathcal{M}$. Fix $\varepsilon \in (0,1]$ and let $t \geq t(\varepsilon)$. Then for any set of states A:

$$\begin{aligned} p^{(i,t)}(A) &= \Pr(X_t \in A) \\ &\geq \Pr(Y_t \in A \wedge X_t = Y_t) \\ &\geq 1 - \Pr(Y_t \notin A) - \Pr(X_t \neq Y_t) \\ &\geq \Pr(Y_t \in A) - \varepsilon \\ &= \pi(A) - \varepsilon, \end{aligned}$$

and similarly for the set $\bar{A} = S \setminus A$. Thus

$$|p^{(i,t)}(A) - \pi(A)| \leq \varepsilon \quad \forall\, t \geq t(\varepsilon),$$

and because A was chosen arbitrarily, also

$$d_V(p^{(i,t)}, \pi) = \max_{A \subseteq S} |p^{(i,t)}(A) - \pi(A)| \leq \varepsilon \quad \forall\, t \geq t(\varepsilon).$$

Thus $\tau(\varepsilon) \leq t(\varepsilon)$. □

Example 3.5 *Gibbs sampler for graph colourings.* Let $G = (V,E)$ be an undirected graph with maximum node degree Δ. Without loss of generality assume that $V = \{1,\ldots,n\}$. A *q-colouring* of G is a map $\sigma : V \to \{1,\ldots,q\} = Q$ such that

$$(i,j) \in E \;\Rightarrow\; \sigma(i) \neq \sigma(j).$$

According to so called Brooks' Theorem, G has a q-colouring for any $q \geq \Delta + 1$. (In fact, already for $q \geq \Delta$ unless G contains a $(\Delta+1)$-clique $K_{\Delta+1}$ as a component.)

For $q \geq \Delta + 2$, one can set up the following Gibbs sampler Markov chain $\mathcal{M}$ to sample q-colourings of G asymptotically uniformly at random (cf. Example 2.2, p. 24):

Given a colouring $\sigma \in Q^V$:

(i) select a node $i \in V$ uniformly at random;

(ii) select a legal colour c for i uniformly at random (c is legal for i if $c \neq \sigma(j)\ \forall\ j \in \Gamma(i)$);

(iii) recolour i with colour c (i.e. move from σ to σ', where $\sigma'(i) = c$ and $\sigma'(j) = \sigma(j)$ for $j \neq i$).

Let us verify that $\mathcal{M}$ is regular for $q \geq \Delta + 2$:

1. Irreducibility: Any colouring can be reached from any other by recolouring the nodes in increasing order; because $q \geq \Delta + 2$ one can avoid conflicts by if necessary first adjusting the colours at higher-numbered neighbours of the present node.

2. Aperiodicity: Each colouring has a nonzero self-loop probability, so aperiodicity follows from regularity.

It is easy to verify that by reversibility $\mathcal{M}$ has as its stationary distribution π the uniform distribution over the set of legal colourings $S \subseteq Q^V$.

Let us then consider how quickly the chain $\mathcal{M}$ converges to π, in terms of the d_V distance. To introduce the ideas, consider first the trivial case $E = \varnothing$ ($\Rightarrow\ S = Q^V$).

In this case one can effect a coupling between two copies of $\mathcal{M}$ as follows: in a transition $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$:

(i) select a node $i \in V$ uniformly at random;

(ii) select a colour $c \in Q$ uniformly at random and recolour i with colour c in both X_t and Y_t; let the resulting colourings be X_{t+1} and Y_{t+1}.

Now clearly (X_t) and (Y_t) are both faithful copies of $\mathcal{M}$, i.e. the marginal transition probabilities work out OK:

$$\begin{aligned} \Pr(X_{t+1} = \sigma' \mid X_t = \sigma, Y_t = \eta) &= \Pr(\sigma, \sigma'), \\ \Pr(Y_{t+1} = \eta' \mid X_t = \sigma, Y_t = \eta) &= \Pr(\eta, \eta'). \end{aligned}$$

On the other hand, it is clear that the time required for the chains (X_t) and (Y_t) to coalesce is not very much larger than n, because at each step of the coupled chain, a randomly chosen node is coloured similarly in both (X_t) and (Y_t).

More precisely, introduce the random variable

$$D_t = \#\{i \in V \mid X_t(i) \neq Y_t(i)\}.$$

Thus $D_t = 0 \;\Leftrightarrow\; X_t = Y_t \Leftrightarrow\; T \leq t$.

Furthermore,

$$E(D_{t+1} \mid D_t) = \frac{D_t}{n} \cdot (D_t - 1) + \frac{n - D_t}{n} \cdot D_t = \left(1 - \frac{1}{n}\right) \cdot D_t$$

$$\Rightarrow \quad E(D_t \mid D_0) = \left(1 - \frac{1}{n}\right)^t \cdot D_0$$

$$\overset{\text{(Markov)}}{\Rightarrow} \Pr(D_t > 0 \mid D_0) \leq E(D_t \mid D_0) \leq \left(1 - \frac{1}{n}\right)^t \cdot n \leq n e^{-t/n}.$$

Thus, choosing $t \geq n \ln \frac{n}{\varepsilon}$ suffices to guarantee that $\Pr(X_t \neq Y_t) \leq \varepsilon$, which by Lemma 3.12 implies that the mixing time satisfies $\tau(\varepsilon) \leq n \ln \frac{n}{\varepsilon}$.

For the general case we need a more complicated coupling, in order to take into account the constraints on colour choice caused by the edges in E.

We observe that by a simple construction, it is possible to combine two finite state sets A and B to a single state set C so that there are random variables X_A and X_B such that

$$\begin{aligned} &\text{(i)} \quad \Pr(X_A = x) = \begin{cases} 1/|A|, & x \in A, \\ 0, & x \notin A; \end{cases} \\ &\qquad \Pr(X_B = x) = \begin{cases} 1/|B|, & x \in B, \\ 0, & x \notin B; \end{cases} \\ &\text{(ii)} \quad \Pr(X_A = X_B) = \frac{|A \cap B|}{\max\{|A|, |B|\}}. \end{aligned} \tag{10}$$

Denote $\Gamma(i) = \{j \in V \mid (i,j) \in E\}$, $X_t(i)$ = colour of node i in colouring X_t, and $X_t(U) = \{X_t(i) \mid i \in U\}$.

Consider the following coupling $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$:

(i) select a node $i \in V$ uniformly at random;

(ii) select colours $c_X \in Q \setminus X_t(\Gamma(i))$, $c_Y \in Q \setminus Y_t(\Gamma(i))$ uniformly (but not independently) at random, using the joint sample space indicated in (10);

(iii) recolour node i with colour c_X in X_t to yield X_{t+1}; similarly with colour c_Y in Y_t to yield Y_{t+1}.

Denote $A = A_t = \{i \in V \mid X_t(i) = Y_t(i)\}$. Thus $D_t = |\bar{A}| = |V \setminus A|$.

Now clearly $D_{t+1} \in \{D_t + 1, D_t, D_t - 1\}$. Let us compute the probabilities $P(D_{t+1} \mid D_t)$ for each of these cases:

(i) $D_{t+1} = D_t + 1$. In this event the chosen $i \in A$, and $c_X \neq c_Y$.

Denote by $\xi = |Q \setminus X_t(\Gamma(i))|$, $\eta = |Q \setminus Y_t(\Gamma(i))|$, $\zeta = |Q \setminus (X_t(\Gamma(i)) \cup Y_t(\Gamma(i)))|$ the number of legal values for c_X, c_Y, and their overlap, respectively. Thus, the probability that the same colour is chosen for i in both X_{t+1} and Y_{t+1} is $\zeta / \max\{\xi, \eta\}$. Denote $d'(i) = |\Gamma(i) \setminus A|$ (recall that $i \in A$). Then

$$q - \Delta \leq \xi, \eta \leq \zeta + d'(i).$$

Hence:

$$\begin{aligned} \Pr(c_X = c_Y) &= \frac{\zeta}{\max\{\xi, \eta\}} \geq \frac{\max\{\xi, \eta\} - d'(i)}{\max\{\xi, \eta\}} \\ &\geq 1 - \frac{d'(i)}{q - \Delta} \end{aligned}$$

and consequently:

$$\Pr(D_{t+1} = D_t + 1) \leq \frac{1}{n} \sum_{i \in A} \frac{d'(i)}{q - \Delta} = \frac{m'}{(q - \Delta)n},$$

where $m' = \sum_{i \in A} d'(i)$.

(ii) $D_{t+1} = D_t - 1$. In this event the chosen $i \in \bar{A}$, and $c_X = c_Y$.

Denote ξ, η, ζ as in case (i), and $d''(i) = |\Gamma(i) \cap A|$. Now

$$q - \Delta \leq \xi, \eta \leq \zeta + (\Delta - d''(i)).$$

As in case (i), we obtain:

$$\begin{aligned} \Pr(c_X = c_Y) &= \frac{\zeta}{\max\{\xi, \eta\}} \geq \frac{\max\{\xi, \eta\} - (\Delta - d''(i))}{\max\{\xi, \eta\}} \\ &\geq 1 - \frac{\Delta - d''(i)}{q - \Delta} = \frac{q - 2\Delta + d''(i)}{q - \Delta} \end{aligned}$$

and consequently:

$$\begin{aligned} \Pr(D_{t+1} = D_t - 1) &\geq \frac{1}{n} \sum_{i \in \bar{A}} \left(\frac{q - 2\Delta}{q - \Delta} + \frac{d''(i)}{q - \Delta} \right) \\ &= \frac{q - 2\Delta}{(q - \Delta)n} D_t + \frac{m'}{(q - \Delta)n}, \end{aligned}$$

where $m' = \sum_{i \in \bar{A}} d''(i) = \sum_{i \in A} d'(i)$.

Denoting for brevity

$$a = \frac{q-2\Delta}{(q-\Delta)n}, \quad b = b(m') = \frac{m'}{(q-\Delta)n},$$

we see that

$$\Pr(D_{t+1} = D_t + 1) \le b, \quad \Pr(D_{t+1} = D_t - 1) \ge aD_t + b.$$

Assume that $a > 0$, i.e. that $q > 2\Delta$. Then

$$\begin{aligned} E(D_{t+1}|D_t) &\le b(D_t+1) + (aD_t+b)(D_t-1) + (1-aD_t-2b)D_t \\ &= (1-a)D_t. \end{aligned}$$

Thus, $E(D_t) \le (1-a)^t D_0 \le (1-a)^t n$, and hence by Markov's inequality

$$\Pr(D_t > 0) \le (1-a)^t n \le ne^{-at}.$$

Thus $\Pr(X_t \ne Y_t) \le \varepsilon$ for $t \ge \frac{1}{a}\ln\frac{n}{\varepsilon}$, and so by Lemma 3.12, the mixing time of the chain satisfies

$$\tau(\varepsilon) \le \frac{q-\Delta}{q-2\Delta} \cdot n \ln\frac{n}{\varepsilon} \le (\Delta+1)n\ln\frac{n}{\varepsilon}$$

for $q > 2\Delta$.

4 Exact Sampling with Coupled Markov Chains

In 1996 J. Propp and D. Wilson introduced an intriguing method for producing samples from a Markov chain *exactly* according to its stationary distribution. This *exact sampling* (or "coupling from the past") technique eliminates the need to compute Markov chain convergence rates for quality control: when the algorithm stops, it is guaranteed to produce a perfect sample. However for slowly converging chains stopping will take a long time, so convergence rates are still of importance from the point of view of algorithm efficiency. (There are also some other efficiency caveats in the method besides slow convergence of the simulated chain. These are discussed below.)

Let $\mathcal{M}$ be a regular reversible Markov chain with state set $S = \{1,\dots,n\}$, transition probability matrix $P = (p_{ij})$, and stationary distribution π.

Consider an explicit simulation of $\mathcal{M}$ by the following method: at each step t, a uniformly distributed random number $R_t \in [0,1)$ is chosen, and the state transition of $\mathcal{M}$ is determined as $X_{t+1} = s(X_t, R_t)$, where

$$s(i,r) = \begin{cases} 1, & \text{if } r \in [0, p_{i1}), \\ 2, & \text{if } r \in [p_{i1}, p_{i1}+p_{i2}), \\ \vdots & \\ n, & \text{if } r \in [p_{i1}+\ldots+p_{i(n-1)}, 1). \end{cases}$$

It is clear that transition probabilities according to the chain $\mathcal{M}$ can equivalently be computed with respect to sequences (R_t) and the above deterministic transition rule, e.g.

$$P_{ij}^{(t)} = \Pr(X_t = j | X_0 = i) = \Pr_{\vec{R}}(s^{(t)}(i,\vec{R}) = j),$$

where

$$s^{(t)}(i, \langle r_0, r_1, \dots, r_{t-1}\rangle) = \underbrace{s(s(\cdots s(}_{t}(i,r_0),r_1)\cdots),r_{t-1}).$$

Now let us consider the following curious simulation method for the chain $\mathcal{M}$, from further and further away in the *past* ($t = -T$, $T = 1,2,4,8,\dots$) to the present ($t = 0$):

Algorithm PW (Propp-Wilson):

set $T \leftarrow 1$
generate random numbers $r_{-T},\dots,r_{-1} \in [0,1)$ uniformly at random;
(1) simulate the chain $\mathcal{M}$ as above, using the random numbers $r_{-T},\dots,r_{-1}$, from every possible initial state $X_{-T} \in S$;
if all the simulations lead to the same state $X_0 = i_0$, then output i_0 and stop;
otherwise generate T more random numbers $r_{-2T},\dots,r_{-T-1} \in [0,1)$ uniformly at random;
set $T \leftarrow 2T$; go to (1).

For a three-state chain, a run of the PW algorithm might look as illustrated in Figure 14. Here the algorithm has required two restarts, but the third run from $T = -4$ has resulted in all the simulated realisations of the chain coalescing, with common result $i_0 = 2$.

In the following, we shall assume that the PW algorithm always converges with probability 1. Ensuring this may require some care in verifying that the deterministic update rule $s(i,r)$, and the chosen numbering of the Markov chain states do not interact in a bad way.

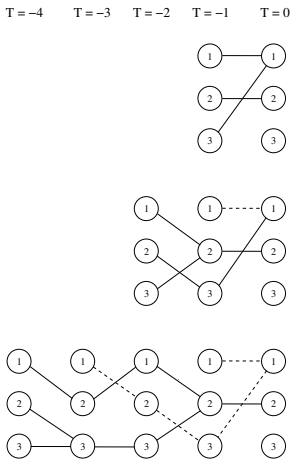


Figure 14: A Propp-Wilson simulation of a 3-state Markov chain.

Theorem 4.1 *Let Y be a random variable indicating the eventual output state of the PW algorithm, under the above assumptions and notations. Then*

$$\Pr_R(Y = i) = \pi_i, \quad \forall\, i \in S.$$

Proof. Fix some value $i \in S$. To prove the Theorem, it suffices to show that for any $\varepsilon > 0$

$$|\Pr_R(Y = i) - \pi_i| \leq \varepsilon.$$

So fix an arbitrary $\varepsilon > 0$. Since we assume that the PW algorithm terminates with probability 1, there is some value of T such that

$$\Pr_R(\text{PW simulation converges for chains of length } T) \geq 1 - \varepsilon. \tag{11}$$

Now consider running the actual chain from time $-T$ to time 0, starting with the stationary distribution:

$$\Pr(X_{-T} = i) = \pi_i.$$

In this case, of course also the variable X_0 is distributed according to the stationary distribution:

$$\Pr_R(X_0 = i) = \pi_i.$$

However, if the coalescence event (11) occurs for a given sequence R of random numbers, then $X_0 = Y$, and so $\Pr_R(X_0 \neq Y) \leq \varepsilon$. Thus,

$$\begin{aligned} \Pr(Y = i) - \pi_i &= \Pr(Y = i) - \Pr(X_0 = i) \\ &\leq \Pr(Y = i, X_0 \neq i) \\ &\leq \varepsilon, \end{aligned}$$

and by a similar argument

$$\pi_i - \Pr(Y = i) \leq \varepsilon.$$

Thus, $|\Pr(Y = i) - \pi_i| \leq \varepsilon$, and the claim is proved. □

Note that the PW algorithm cannot be "simplified" by simulating the chains forwards from time $T = 0$ until they coalesce. This yields biased samples.

The PW algorithm as described above still has two shortcomings:

1. The need to store long sequences of random numbers for reuse (can be a serious problem in long simulations); and
2. The need to simulate the chains starting from all possible initial states (infeasible in many applications where the number of system states is exponential in the size of the system itself).

Problem (1) has been addressed in a recent (2000) modification to the algorithm ("CFTP with read once randomness") by D. Wilson.

For problem (2), Propp & Wilson (1996) proposed a solution that can be applied when the states of the system have a suitable partial order $\sqsubseteq$ respected by the update rule.

Specifically, assume that the states of the system to be simulated form a partial order $(S = \{\sigma_1, \ldots, \sigma_n\}, \sqsubseteq)$ with a unique largest element $\top$ ("top") and unique smallest element $\bot$ ("bottom"), and satisfying the condition

$$\sigma \sqsubseteq \sigma' \;\Rightarrow\; s(\sigma, r) \sqsubseteq s(\sigma', r), \quad \forall\, \sigma, \sigma' \in S \text{ and } r \in [0, 1). \tag{12}$$

Then it suffices to simulate the "top" and "bottom" chains until they couple, since their coupling implies the coalescence of all the other chains as well (cf. Figure 15).

This is of course a huge improvement: reducing the simulation of, say, 2^n parallel chains to just 2.

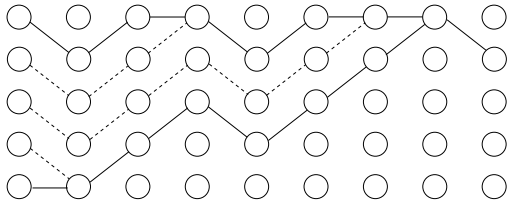


Figure 15: Coalescence of an ordered Propp-Wilson simulation.

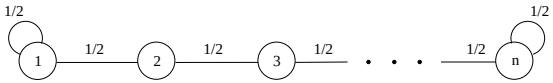


Figure 16: A one-dimensional random walk with semi-reflecting barriers.

So what systems admit this simplification?

A simple example would be a one-dimensional random walk on the state set $S = \{1,\dots,n\}$ with, say, semi-reflecting barriers to ensure regularity of the chain (Figure 16). Assume the state transition rule is:

$$s(i,r) = \begin{cases} \max\{i-1,1\}, & \text{if } 0 \le r < \frac{1}{2}, \\ \min\{i+1,n\}, & \text{if } \frac{1}{2} \le r < 1. \end{cases}$$

The the natural ordering of states fulfills the condition (12):

$$i \le j \;\Rightarrow\; s(i,r) \le s(j,r) \qquad \forall\, i,j = 1,\dots,n,\; r \in [0,1).$$

Interestingly, also complicated systems such as the Ising spin glass model admit such orderings. In the case of the Ising model, the order between states $\sigma, \sigma' \in \{-1,+1\}^n$ is determined simply by

$$\sigma \sqsubseteq \sigma' \quad \text{if} \quad \sigma_i \le \sigma'_i \;\forall\, i = 1,\dots,n.$$

Clearly $\bot = (-1,\dots,-1)$ and $\top = (1,\dots,1)$ with respect to $\sqsubseteq$, and also condition (12) can be verified.

Part II

Combinatorial Models

5 A Sketch of Basic Statistical Physics

Statistical physics = Thermodynamics (macroscopic)
+ Statistical mechanics (microscopic)

5.1 Thermodynamics

A *thermodynamic system* is characterised by (macroscopic, observable) variables T ("temperature") and $X_1,\dots,X_n$. These variables determine "all interesting" properties of the system.

E.g. in the classical ideal gas model a sufficient set of variables is T, p, V and N. ($N \sim$ the number of molecules is here for simplicity thought of as a continuous quantity. This might be easier if N was replaced by $n = N/N_0$, the amount in moles of gas, where $N_0 = 6.02 \cdot 10^{23}$ is Avogadro's number.)

The system is in (*thermal*) *equilibrium* if it satisfies a characteristic *state equation*

$$g(T,X_1,\dots,X_n) = 0$$

E.g. ideal gas: $pV - NkT = 0$, where $k = 1.38 \cdot 10^{-23} J/K$ is *Boltzmann's constant*, or $pV - nRT = 0$, where $R = 8.32 J/K\text{mol}$ is the *gas constant*.

A *potential* or *energy function* for the system is some sufficiently smooth function

$$F = F(T,X_1,\dots,X_n).$$

In classical thermodynamics, a key role is given to the *total energy* function determined by the *First Law of Thermodynamics*:

$$dU = dQ + dW, \tag{1}$$

where dQ is the amount of "heat" added to a system and dW is the amount of "work" performed on it.

Integrating the potential given e.g. the state equation of the ideal gas yields

$$U(T,p,N) = U_0 + \left(\frac{1}{2}fN + N - S_0\right)(T - T_0) - NT\ln\left(\left(\frac{T}{T_0}\right)^{1+f/2}\frac{p_0}{p}\right),$$

where U_0, S_0, T_0 and p_0 are reference values and $f/2$ a constant ("specific heat").[1]

In classical thermodynamics, the system variables are divided into *extensive* and *intensive*, depending on whether their values depend on the "size" of the system or not. E.g. T and p are intensive, V and N extensive.

Two systems at the same temperature may be "combined", and if F is otherwise a function of extensive variables only, then it is linear, i.e.

$$F(T, X_1 + X_1', \ldots, X_n + X_n') = F(T, X_1, \ldots, X_n) + F(T, X_1', \ldots, X_n').$$

By the total derivative formula:

$$dF = \left(\frac{\partial F}{\partial T}\right)dT + \sum_{i=1}^{n}\left(\frac{\partial F}{\partial X_i}\right)dX_i. \tag{2}$$

State variables are *conjugate* (with respect to F), if

$$X = \frac{\partial F}{\partial Y} \qquad \text{or} \qquad Y = \frac{\partial F}{\partial X}.$$

In classical thermodynamics conjugates of extensive variables are intensive, and vice versa. The conjugate of T w.r.t. U,

$$S = \frac{\partial U}{\partial T}$$

is called the *entropy* of the system.

[1] To be precise, since T and p are not "natural" variables of the energy function U arising from its differential definition (1), this equation refers to a variant of U expressed in terms of T, p and N, so called "Gibbs free energy".

Conjugate variables may be interchanged via the *Legendre transform*, yielding new forms of a given potential function. E.g. in the case of the ideal gas with fixed N, $U = U(S,V)$ and

$$dU = TdS - pdV.$$

Here we may interchange S for T by considering instead of U the *Helmholz free energy* $F = U - ST$. This satisfies:[2]

$$dF = dU - SdT - TdS = TdS - pdV - SdT - TdS = -SdT - pdV.$$

For this potential function the "natural" variables are T and V, i.e. $F = F(T,V)$.

In the classical setting, it is a law of nature (the *Second Law of Thermodynamics*) that in equilibrium processes (evolutions) entropy never decreases:

$$dS \geq 0.$$

Processes for which entropy stays constant ($dS = 0$) are called *adiabatic*.

5.2 Statistical Mechanics

Let us consider a thermodynamic energy function framed in terms of extensive variables:

$$U = U(S, X_1, \ldots, X_n),$$

and assume that the value of U expresses in fact only the *average* of a large number of microscopic potentials:

$$U = \langle H \rangle = \sum_{\omega} p_\omega H(\omega).$$

The micropotential function $H(\omega)$ is also called the *Hamiltonian* of the system. We shall furthermore assume, motivated by the additivity of U, that the Hamiltonian of a system consisting of two independent subsystems at thermal equilibrium can be decomposed as:

$$H(\langle \omega_1, \omega_2 \rangle) = H(\omega_1) + H(\omega_2).$$

What is now the distribution of the microstates p_ω, given the constraint that $\langle H \rangle = U$? We assume that all microstates with the same value of the Hamiltonian are equally probable, so that p_ω has the form $p_\omega = g(H(\omega))$.

[2] There is an unfortunate sign difference here as compared to formula (2). We could have fixed this by defining $F = ST - U$, but this would have been against convention.

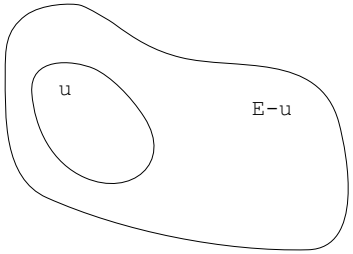


Figure 1: A heat bath.

To further specify the functional form of the distribution, think of our system $\mathcal{S}$ as being in thermal equilibrium with, but otherwise independent of, a much larger system or "reservoir" $\mathcal{R}$. Denote the total system consisting of $\mathcal{S}$ and $\mathcal{R}$ by $\mathcal{T}$ (This is called a "heat bath" arrangement; cf. Figure 1.)

For any given system, denote by $\Omega(u) = |H^{-1}(u)|$ the number of its microstates at potential u. (Whether we are referring to $\mathcal{S}$, $\mathcal{R}$ or $\mathcal{T}$ should always be clear from the context.) Fix some reference potential level $E \gg U$ for the total system $\mathcal{T}$, and observe that by our assumption, all microstates of $\mathcal{T}$ with potential E have the same probability.

Now for every microstate ω of $\mathcal{S}$, there are exactly $\Omega(E - H(\omega))$ microstates ω^r of $\mathcal{R}$ such that the combined state $\langle \omega, \omega^r \rangle$ of $\mathcal{T}$ has potential E. Since all of these are equally probable, it follows that $p_\omega \propto \Omega(E - H(\omega))$. Taking logarithms and applying Taylor's formula yields:

$$\begin{aligned} \ln p_\omega &= \ln \Omega(E - H(\omega)) + \text{const.} \\ &= \ln \Omega(E) - \left(\frac{\partial \ln \Omega(E')}{\partial E'} \right)_{E'=E} H(\omega) + \cdots \\ &= \ln \Omega(E) - \beta H(\omega) + \cdots, \end{aligned}$$

where $\beta = \partial \ln \Omega / \partial E$ is a parameter whose value is to be determined later.

Taking exponentials again, we obtain the so called *Gibbs* (or *Boltzmann*) *distribution*

$$p_\omega \propto e^{-\beta H(\omega)} \tag{3}$$

with normalisation constant (actually, function)

$$Z = Z_\beta = \sum_\omega e^{-\beta H(\omega)}, \tag{4}$$

known as the *partition function*. [3] Now the value of β is in principle determined implicitly by the condition

$$\langle H \rangle = \frac{1}{Z} \sum_\omega e^{-\beta H(\omega)} H(\omega) = U,$$

but we shall obtain a more transparent representation for it below.

The (logarithm of the) partition function (4) can be used to compute several macroscopic quantities:

First:

$$\begin{aligned} \frac{\partial \ln Z}{\partial \beta} &= \frac{1}{Z} \frac{\partial Z}{\partial \beta} \\ &= \frac{1}{Z} \frac{\partial}{\partial \beta} \sum_\omega e^{-\beta H(\omega)} \\ &= \frac{1}{Z} \sum_\omega e^{-\beta H(\omega)} \left(-H(\omega) \right) \\ &= -\sum_\omega p_\omega H(\omega) \\ &= -U. \end{aligned}$$

Second: Consider an extensive variable X_i and its conjugate $\mu_i = \partial U / \partial X_i$.

$$\begin{aligned} \frac{\partial \ln Z}{\partial X_i} &= \frac{1}{Z} \sum_\omega \frac{\partial}{\partial X_i} e^{-\beta H(\omega; X_i)} \\ &= \frac{1}{Z} \sum_\omega e^{-\beta H(\omega; X_i)} \left(-\beta \frac{\partial H(\omega; X_i)}{\partial X_i} \right) \\ &= -\beta \sum_\omega p_\omega \frac{\partial H(\omega; X_i)}{\partial X_i} \\ &= -\beta \left\langle \frac{\partial H(\omega; X_i)}{\partial X_i} \right\rangle \\ &= -\beta \mu_i. \end{aligned}$$

[3] In fact, $Z = Z(\beta, X_1, \ldots, X_n)$. Note also that Z is a kind of a *generating function* for the sequence of values $\Omega(u)$, since $Z(\beta) = \sum_u \Omega(u) \cdot (e^{-\beta})^u$.

Third:

$$\begin{aligned} d\ln Z &= \frac{\partial \ln Z}{\partial \beta} d\beta + \sum_{i=1}^{n} \frac{\partial \ln Z}{\partial X_i} dX_i \\ &= -U d\beta - \beta \sum_{i=1}^{n} \mu_i dX_i \\ &= -d(\beta U) + \underbrace{\beta dU - \beta \sum_{i=1}^{n} \mu_i dX_i}_{\beta T dS}. \end{aligned}$$

$$\therefore T dS = \frac{1}{\beta} d(\ln Z + \beta U)$$

$$\therefore \frac{1}{\beta} = kT, \quad dS = k d(\ln Z + \beta U), \quad k = \frac{1}{\beta T} = \text{ constant}$$

$$\therefore \frac{1}{\beta} = kT, \quad S = k \ln Z + \frac{U}{T} + \text{const.} \sim k \ln Z + \frac{U}{T}$$

$$\therefore \beta = \frac{1}{kT}, \quad -kT \ln Z \sim U - TS = F \quad \text{(Helmholz free energy)}$$

Conversely, let us expand the entropy variable as a microscopic average:

$$\begin{aligned} S &= k \ln Z + k\beta U \\ &= k \ln Z + k \sum_{\omega} p_\omega \beta H(\omega) \\ &= k \left(\ln Z - \sum_{\omega} p_\omega (\ln Z + \ln p_\omega) \right) \\ &= -k \sum_{\omega} p_\omega \ln p_\omega. \end{aligned} \quad \left| \begin{array}{l} p_\omega = \frac{1}{Z} e^{-\beta H(\omega)} \\ \Rightarrow \beta H(\omega) = -\ln(Z p_\omega) \\ \\ \sum_{\omega} p_\omega = 1 \end{array} \right.$$

One more, simplified expression for entropy: partition the range of possible potential values into narrow bands (of width ΔU, say), and denote the number of microstates falling in band r as

$$\Omega(U_r) = \left| \{ \omega : U_r \le H(\omega) < U_r + \Delta U \} \right|$$

Then the partition function is approximately

$$Z \approx \sum_r \Omega(U_r) e^{-\beta U_r}$$

In fact, since the number of microstates in a typical system is huge, the microstate potentials are highly concentrated around the average $U = \langle H \rangle$, and so in fact

$$Z \approx \Omega(U) e^{-\beta U},$$

whence

$$S = \frac{1}{T}(-F + U) = k \ln Z + \frac{U}{T} \approx k \ln \Omega(U) \underbrace{- \beta k U + \frac{U}{T}}_{=0} \approx k \ln \Omega(U).$$

6 The Ising Model, Spin Glasses and Neural Networks

6.1 The Ising Model

The following model was introduced by Ernst Ising in 1925 to explain magnetism in materials.

At a microscopic level, Ising's model system consists of N *sites* arranged in a lattice, either 1-D, 2-D ($N = L^2$), or maybe even 3-D. At each site $i = 1, \ldots, N$ is located a magnetic ion or *spin* pointing either *up* or *down* ($S_i = \pm 1$). Neighbouring sites $\langle ij \rangle$ are related by an *interaction coefficient* J_{ij}, which in Ising's model is uniformly either a positive $J > 0$ ("ferromagnetic case") or a nonpositive $J \le 0$ ("antiferromagnetic case"). A system whose internal interactions are all weak ($J_{ij} \approx 0$) is "paramagnetic". In addition, there may be an *external field* h influencing the orientation of each of the spins. (More generally, one could have separate fields h_i for each spin S_i.)

The Hamiltonian of spin state $\sigma = \langle S_1, \ldots, S_N \rangle$ is

$$H(\sigma) = -J \sum_{\langle ij \rangle} S_i S_j - h \sum S_i,$$

where the sum is taken over *nearest neighbour pairs* $\langle ij \rangle$ and periodic boundary conditions are assumed for simplicity.

States σ yielding the global minimum value of $H(\sigma)$ are called *ground states* of the system. For a ferromagnetic system, the ground state has either all $S_i = +1$ if $h > 0$, or all $S_i = -1$ if $h < 0$. If $h = 0$, these two states are both equally good.

As a very simple example, let us compute the partition function for a trivial Ising paramagnet with N spins and $J = 0$. Denote $\Omega = \{+1, -1\}^N$. Then:

$$\begin{aligned} Z_\beta &= \sum_{\sigma\in\Omega} e^{-\beta H(\sigma)} \\ &= \sum_{\sigma\in\Omega} \exp(\beta h \sum_i S_i) \\ &= \sum_{S_1=\pm1} \sum_{S_2=\pm1} \cdots \sum_{S_N=\pm1} e^{\beta h S_1} e^{\beta h S_2} \ldots e^{\beta h S_N} \\ &= \left(\sum_{S=\pm1} e^{\beta h S} \right)^N \\ &= \big(2\cosh(\beta h)\big)^N \end{aligned} \qquad \cosh x = \frac{e^x + e^{-x}}{2}$$

Define the *(total) magnetisation* of state σ as

$$M(\sigma) = \sum_{i=1}^{N} S_i.$$

The corresponding thermodynamic average at given β is

$$\begin{aligned} \langle M \rangle &= \frac{1}{Z} \sum_{\sigma\in\Omega} M(\sigma) \exp(-\beta H(\sigma)) \\ &= \frac{1}{Z} \underbrace{\sum_{\sigma\in\Omega} \big(\sum_i S_i\big) \exp(-\beta H(\sigma))}_{(\bigstar)}. \end{aligned}$$

However now in fact $(\bigstar) = \frac{\partial Z}{\partial(\beta h)}$, so fortuitously:

$$\begin{aligned} \langle M \rangle &= \frac{1}{Z} \frac{\partial Z}{\partial(\beta h)} = \frac{\partial \ln Z}{\partial(\beta h)} \\ &= N \frac{\partial \ln(2\cosh(\beta h))}{\partial(\beta h)} \\ &= N \frac{2(\partial \cosh(\beta h)/\partial(\beta h))}{2\cosh(\beta h)} \\ &= N \frac{2\sinh(\beta h)}{2\cosh(\beta h)} \\ &= N \tanh(\beta h). \end{aligned}$$

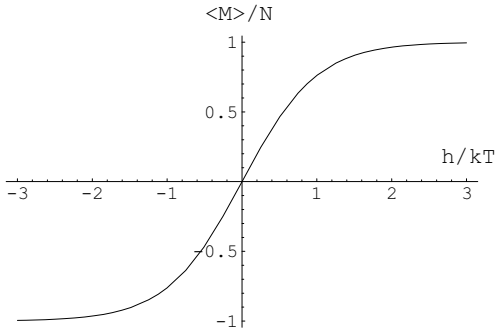


Figure 2: Magnetisation of an Ising paramagnet.

Thus the average magnetisation per site or "magnetisation density" of a totally decoupled Ising paramagnet at external field h and temperature $T = 1/k\beta$ equals

$$\langle M \rangle = \tanh\left(\frac{h}{kT}\right).$$

A plot of this function is presented in Figure 2.

The ferromagnetic 1-D Ising model is also explicitly solvable with somewhat more work. The 2-D ferromagnetic case with $h = 0$ was solved by L. Onsager in 1944, and in a simpler way by Kasteleyn & Fisher in 1961. The 2-D case with $h \neq 0$ and higher dimensions are still open.

6.2 Spin Glasses

Spin glasses generalise the Ising model with more general interactions. Also the spins may be nonbinary, in which case such models are called *Potts glasses*.

The general form of the (binary-state) spin glass Hamiltonian is

$$H(\sigma) = -\sum_{\langle ij \rangle} J_{ij} S_i S_j - \sum_i h_i S_i,$$

where $J_{ij}, h_i \in \mathbb{R}$. Also the neighbourhood relation may correspond to an arbitrary *graph*, not necessary a lattice.

Several varieties of spin glass models have been introduced, e.g.:

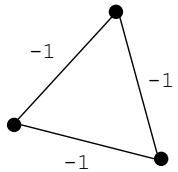


Figure 3: Frustrated spin glass configuration.

- The Sherrington-Kirkpatrick model: Hamiltonian as above, complete interconnection graph, coefficients J_{ij} according to a specific probability distribution.

- The Edwards-Anderson model: Hamiltonian

$$H(\sigma) = -\sum_{\langle ij \rangle} J_{ij} S_i S_j,$$

regular lattice topology (e.g. cubic), J_{ij} independent Gaussian variables.

A phenomenon that makes spin glass models even less tractable than the Ising model is *frustration*. E.g. in the spin glass neighbourhood in Figure 3 there is no completely "consistent" choice of spin values.

Frustration means that the "landscape" determined by the Hamiltonian can have a very complicated structure, with large numbers of local minima, and no obvious location for the globally minimal ground state.

In fact, the problem of determining the ground state of a given SK-spin glass instance $\langle \bar{J}, h \rangle$ is *NP-complete*, even with $h = 0$.

This can be seen by reduction from the well-known NP-complete MAX CUT problem: Given a graph $G = (V, E)$, determine the partition $V = V_1 \cup V_2$ that maximises $w(V_1, V_2) = \left| \{(i, j) \in E : i \in V_1 \wedge j \in V_2\} \right|$.

The reduction is as follows:

Given a graph $G = (V, E)$, let $\bar{J}$ be an SK system with sites corresponding to V, and J_{ij} determined by

$$J_{ij} = \begin{cases} -1, & \text{if } \langle i, j \rangle \in E, \\ 0, & \text{otherwise.} \end{cases}$$

Let then $C = (V_1, V_2)$ be a cut in G, and divide the edges in G corresponding as

$$\begin{aligned} E_1 &= \{\langle i, j \rangle \in E : i, j \in V_1\}, \\ E_2 &= \{\langle i, j \rangle \in E : i, j \in V_2\}, \\ E_C &= \{\langle i, j \rangle \in E : i \in V_1 \wedge j \in V_2\}. \end{aligned}$$

Consider the spin glass state σ determined as

$$S_i = \begin{cases} +1, & \text{if } i \in V_1, \\ -1, & \text{if } i \in V_2. \end{cases}$$

For this,

$$\begin{aligned} H(\sigma) &= -\sum_{\langle ij \rangle} J_{ij} S_i S_j = \sum_{\langle ij \rangle \in E} S_i S_j \\ &= \sum_{\langle ij \rangle \in E_1} S_i S_j + \sum_{\langle ij \rangle \in E_2} S_i S_j + \sum_{\langle ij \rangle \in E_C} S_i S_j \\ &= |E_1| + |E_2| - |E_C| \\ &= |E| - 2|E_C| \\ &= |E| - 2w(C). \end{aligned}$$

Conversely, given any spin glass state σ, one obtains a cut C satisfying $w(C) = \frac{1}{2}|E| - \frac{1}{2}H(\sigma)$.

Thus, graph cuts and spin glass states correspond one-to-one, with $w(C) \propto -H(\sigma)$, and minimising one is equivalent to maximising the other.

The result means that the SK spin glass ground state problem is in a sense "universal" difficult problem, i.e. it contains as special cases all the ∼2000 other known NP-complete problems.

For $J_{ij} > 0$ and arbitrary $\bar{h}$ the problem reduces to network flow, and can be solved in polynomial time. For planar G and $\bar{h} = 0$ the problem also has a polynomial time algorithm (Fisher 1966 (2-D lattices), Barahona 1982). However, for planar G with $h \neq 0$, and for 3-D lattices the problem is NP-complete (Barahona 1982). It is also NP-complete for every other nonplanar crystal lattice graph (Istrail 2000). Thus, the dimensionality of the system is not crucial to the complexity of the ground state problem; the key is rather the planarity of the interconnection graph.

6.3 Neural Networks

John Hopfield proposed, in an influential paper in 1982, to use the SK model as a basis for "neural associative memories". The idea is to create an N-site SK

system whose local potential minima correspond to a set of N-bit vectors to be stored. These local minima are also stable states of the system's deterministic (0-temperature) "Glauber dynamics". When such a system is initialised at a state which is "close" to one of the stored stable states, the dynamics (presumably) tends to return it to the nearby local minimum. Thus small perturbations in the stable states tend to get corrected, and the system has "error-correcting" or "associative" capabilities.

More precisely, the deterministic dynamics of such a system is as follows: at a given discrete time instant, a randomly (or in a round-robin manner) chosen site k is updated according to the local rule:

$$\begin{aligned} S'_k &= \operatorname{sgn} \underbrace{\left(\sum_{\langle kj \rangle} J_{kj} S_j + h_k \right)}_{(\bigstar)} \\ &= \begin{cases} +1, & \text{if } (\bigstar) > 0, \\ -1, & \text{if } (\bigstar) < 0, \\ S_k, & \text{if } (\bigstar) = 0, \end{cases} \end{aligned}$$

It can be seen that each time a site changes state, the value of $H(\sigma)$ decreases: Assume $S'_k \neq S_k$. Consider

$$\begin{aligned} H(\sigma') - H(\sigma) &= -\sum_{\langle ij \rangle} J_{ij} S'_i S'_j - \sum_i h_i S'_i \\ &\quad + \sum_{\langle ij \rangle} J_{ij} S_i S_j + \sum_i h_i S_i \\ &= -\sum_{\langle kj \rangle} J_{kj} S'_k S_j + \sum_{\langle kj \rangle} J_{kj} S_k S_j - h_k (S'_k - S_k) \\ &= -\underbrace{(S'_k - S_k)}_{\blacktriangle} \underbrace{\left(\sum_{\langle kj \rangle} J_{kj} S_j + h_k \right)}_{\blacktriangledown} \\ &< 0, \end{aligned}$$

where ▼ and ▲ have the same sign.

Thus, since the value of $H(\sigma)$ is lower bounded by

$$H(\sigma) \geq -\sum_{\langle ij \rangle} |J_{ij}| - \sum_i |h_i|,$$

the system converges eventually to a local minimum of its Hamiltonian.

How should one then craft the interaction coefficients so that a given set of patterns become stable states of the system's dynamics? This can in principle be done in various ways, of which Hopfield proposed the following adaptation of "Hebb's rule":[4]

Consider first a single pattern $\sigma = (S_1, \ldots, S_N) \in \{+1, -1\}^N$ and choose $J = \sigma\sigma^T - I = [S_i S_j]_{ij} - I, h = 0$. Then the dynamics operates as follows:

$$\operatorname{sgn}(J\sigma) = \operatorname{sgn}\left((\sigma\sigma^T - I)\sigma\right) = \operatorname{sgn}\left((||\sigma||^2 - 1)\sigma\right) = \sigma,$$

i.e. σ is a stable state of the dynamics.

Given then a (smallish) set of patterns $\sigma_1, \ldots, \sigma_m$, choose

$$J = \sum_{p=1}^{m} \sigma_p \sigma_p^T - mI \qquad \left(\text{or normalised } J = \frac{1}{m} \sum_p \sigma_p \sigma_p^T - I \right).$$

If the patterns are random, independent identically distributed bit vectors, and there are only $m \ll N$ of them, they are "almost orthogonal", and we may approximate:

$$\begin{aligned} \operatorname{sgn}(J\sigma_k) &= \operatorname{sgn}\left(\left(\sum_{p=1}^{m} \sigma_p \sigma_p^T - mI \right) \sigma_k \right) \\ &= \operatorname{sgn}\left(\underbrace{(||\sigma_k||^2 - m)\sigma_k}_{\text{"signal"}} + \underbrace{\sum_{p \neq k} \overbrace{(\sigma_p^T \sigma_k)}^{\approx 0} \sigma_p}_{\text{"noise"}} \right) \\ &= \sigma_k, \end{aligned}$$

"with high probability".

This analysis has been performed rigorously many times under different assumptions, and the number of patterns m that can be reliably stored has been estimated under different criteria. Typically, the "reliable" storage capacity comes out as $m \approx 0.14N \ldots 0.18N$.

The deterministic Glauber dynamics of SK spin glasses has also other computationally interesting features. One can e.g. show that convergence to a stable state

[4] In a 1949 book, D. O. Hebb suggested as a basic mechanism of neuronal memory that simultaneous activity reinforces the interconnections between neurons. Physiologically this suggestion is still controversial, but mathematically the idea has been used as a basis of several learning mechanisms in artificial neural networks.

can require a number of spin flips that is exponential in N (A. Haken et al. ca. 1989), and that one can in fact embed arbitrary computations in the dynamics (Orponen 1995). (More precisely, determining whether a given "output spin" is $+1$ or -1 in the local minimum reached from a given initial state is a "PSPACE-complete" problem.)

6.4 The NK Model

Introduced by Stuart Kauffman (ca. 1986) as a "tunable family of fitness landscapes".

A *fitness landscape* is a triple $\langle X, R, f \rangle$, where X is the *configuration* (or *state*) *space*, $R \subseteq X \times X$ is a *neighbourhood relation* on X, and $f : X \to \mathbb{R}$ is a *fitness* (or *objective*) *function* .

A point $x \in X$ is a *local optimum* (of f on X) if

$$f(y) \leq f(x) \quad \forall\, yRx$$

and a *global optimum* (*maximum*) if

$$f(y) \leq f(x) \quad \forall\, y \in X$$

Questions of the "ruggedness" of landscapes (correlation structure), number and height of local optima, sizes of "attraction basins" of local optima with respect to "hill-climbing" algorithms etc. are of great interest for natural landscapes.

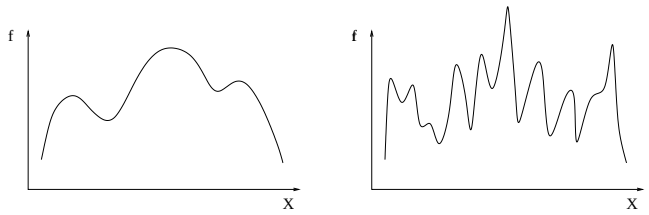
In Kauffman's NK models, $X = A^N$ (usually just $X = \{0,1\}^N$) and K is a tunable neighbourhood size parameter that influences the landscape characteristics, especially its ruggedness (cf. Figure 4).

The model can be seen as a toy model of "epigenetic interactions in chromosomes" — or also a generalisation of the spin glass model.

In Kauffman's model, a *chromosome* is an N-vector of *loci* (*genes*, "positions"), each of which has a value from a set of *alleles* A (usually just $A = \{0,1\}$). A "filled-in" chromosome $\alpha \in A^N$ is called a *genotype*.

The fitness of each gene $i \in \{1,\ldots,N\}$ in a genotype $\alpha = (a_1,\ldots,a_N) \in A^N$ depends on the allele a_i and K other alleles $a^i_1,\ldots,a^i_K$ via some local fitness function $f^i(\alpha) = f^i(a_i; a^i_1,\ldots,a^i_K)$, usually normalised so that $f^i(\alpha) \in [0,1]$. The total fitness of a genotype $\alpha \in A^N$ is the normalised sum of its genes' local fitnesses:

$$f(\alpha) = \frac{1}{N}\sum_{i=1}^{N} f^i(a_i; a^i_1,\ldots,a^i_k) \quad \in [0,1].$$



(a) $K \ll N$ (b) $K \approx N$

Figure 4: A smooth (a) and a rugged (b) NK fitness landscape.

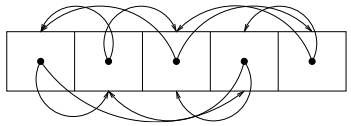


Figure 5: An NK interaction network with $N = 5$, $K = 2$.

Figure 5 illustrates an NK network with five loci and two "epigenetic interactions" per locus.

In Kauffman's versions of the model, the K loci affecting locus i can either be systematically selected as e.g. $i+1,\ldots,i+K \pmod N$, or the chromosome can be simply "randomly wired". The f^i are usually determined as randomly generated 2^{K+1}-element "interaction tables".

From the spin glass perspective, e.g. a 1-D Ising model with N spins can be seen as an $N2$ network where $f^i(S_i; S_{i-1}, S_{i+1}) = \frac{J}{2}(S_{i-1}S_i + S_iS_{i+1})$, and an SK spin glass with coefficients J_{ij} and local fields h_i as an $N(N-1)$ network where

$$f^i(S_i; \sigma \setminus \{S_i\}) = \frac{1}{2}\sum_{\langle ij \rangle} J_{ij}S_iS_j + h_iS_i.$$

Basic properties of the NK model, for binary alleles $A = \{0,1\}$ and varying values of K, include the following:

$K = 0$**:**

If $f^i(0) \neq f^i(1)\ \forall i = 1,\ldots,N$, then there is a unique global optimum, which is easily found by e.g. the obvious 1-locus mutation "hill-climbing" algorithm.

Expected length of the hill-climbing path is $N/2$. (Half of the alleles are "right" in the beginning, after that one allele gets fixed at each step.)

Neighbouring genotypes α, α' are always highly correlated, as necessarily $|f(\alpha) - f(\alpha')| \leq 1/N$.

$1 \leq K < N-1$:

For $K = 1$, a global optimum can still be found in polynomial time. For $K \geq 2$, global optimisation is NP-complete. However, for adjacent affecting loci $(i \curvearrowleft i+1, \ldots, i+K)$, the problem can be solved in time $o(2^K N)$ (Weinberger).

$K = N-1$:

Neighbouring genotypes are totally uncorrelated.

$\Rightarrow$ Probability that a given genotype α is a local optimum is equal to the probability that α has the highest rank within its 1-mutant neighbourhood. This probability is equal to $1/(N+1)$.

$\Rightarrow$ The expected number of local optima is $2^N/(N+1)$.

The expected number of improvement steps for 1-mutant hill-climbing to hit a local optimum is proportional to $\log_2 N$ (each improvement step typically halves the rank of the genotype within the neighbourhood).

The expected waiting time for finding an improvement step is proportional to N.

7 Random Graphs

7.1 The Erdős-Rényi Model(s)

Two closely related "uniform" random graph models introduced in 1959 by P. Erdős & A. Rényi and E. N. Gilbert.

Consider the family $\mathcal{G}_n$ of all (labelled, undirected) graphs on n nodes. Denote $N = \binom{n}{2}$; then $|\mathcal{G}_n| = 2^N$.

Define the following two probability spaces

[Erdős & Rényi:] $\mathcal{G}(n,M)$ = all $G \in \mathcal{G}_n$ with exactly $M \leq N$ edges, taken with uniform probability, i.e.

$$\Pr(G_M = H) = \begin{cases} \binom{N}{M}^{-1}, & \text{if } H \text{ has } M \text{ edges} \\ 0; & \text{otherwise.} \end{cases}$$

[Gilbert:] $\mathcal{G}(n,p)$ = all $G \in \mathcal{G}_n$, taken so that each edge has occurrence probability p, $0 \leq p \leq 1$, independently of the other edges, i.e.

$$\Pr(G_p = H) = p^M \underbrace{(1-p)}_{q}^{N-M}, \text{ if } H \text{ has } M \text{ edges.}$$

These spaces are in a precise sense "close" if $M \sim pN$, and are often both referred to (unfairly to Gilbert) as the "Erdős-Rényi random graph model", or alternatively as the $\mathcal{G}(n,M)$ and $\mathcal{G}(n,p)$ random graph models.

Let $\Omega_n, n = 0,1,2,\ldots$ be a sequence of probability spaces of n-node graphs. Say that *almost every* (*a.e.*) graph in Ω_n has property Q if

$$\Pr(G \in \Omega_n \text{ has } Q) \to 1, \text{ as } n \to \infty.$$

Conversely, *almost no* graph in Ω_n has property Q if a.e. graph in Ω_n has property $\neg Q$, i.e.

$$\Pr(G \in \Omega_n \text{ has } Q) \to 0, \text{ as } n \to \infty.$$

Theorem 7.1 *Let H be a fixed graph and p a constant, $0 < p < 1$. Then a.e. $G \in \mathcal{G}(n,p)$ contains an induced copy of H.*

Remark: an "induced copy" means here a subset of nodes whose induced subgraph is isomorphic to H.

Proof. Let $k = |H| =$ number of nodes in H. Then a graph G with $n = |G| \geq k$ nodes can be partitioned into $\lfloor n/k \rfloor$ disjoint sets of k nodes (with some left over). For each of these sets, the probability that it forms an induced copy of H is $r > 0$. (Precisely, $r = \frac{k!}{|\mathrm{Aut}(H)|} p^{e(H)} q^{\binom{k}{2} - e(H)}$.)

Thus, the probability that none of these sets forms an induced copy of H is

$$(1-r)^{\lfloor n/k \rfloor} \to 0, \text{ as } n \to \infty. \square$$

Let $k, l \in \mathbb{N}$. Say that a graph $G = (V,E)$ has property Q_{kl} if $\forall\, U, W, |U| \leq k, |W| \leq l, U \cap W = \varnothing$, G contains a node $v \in V \setminus (U \cup W)$ such that v is adjacent to all $u \in U$ and no $w \in W$ (cf. Figure 6).

Lemma 7.2 *For every constant p, $0 < p < 1$, and all $k, l \in \mathbb{N}$, a.e. $G \in \mathcal{G}(n,p)$ has property Q_{kl}.*

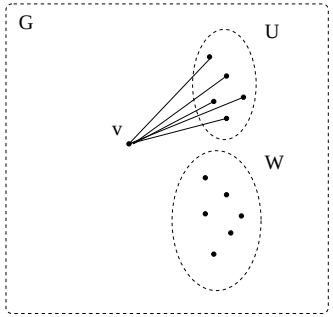


Figure 6: Property Q_{kl}.

Proof. For a fixed $U,W,v \in V \setminus (U \cup W)$, the probability that the condition is satisfied is

$$p^{|U|}q^{|W|} \geq p^k q^l$$

The events are independent for different v, so the probability that no appropriate v exists is

$$\left(1-p^{|U|}q^{|W|}\right)^{n-|U|-|W|} \leq \left(1-p^k q^l\right)^{n-k-l}.$$

There are at most n^{k+l} (U,W)-pairs to be considered, so the probability that some pair has no good v is bounded by

$$n^{k+l}\underbrace{(1-p^k q^l)}_{<1}{}^{n-k-l} \to 0, \text{ as } n \to \infty.$$

Thus in a.e. $G \in \mathcal{G}(n,p)$ all (U,W)-pairs have some appropriate v. □

Corollary 7.3 *Let p, $0 < p < 1$, be a constant. Then (i) a.e. $G \in \mathcal{G}(n,p)$ has minimum degree $\geq k$, for given constant k (ii) a.e. $G \in \mathcal{G}(n,p)$ has diameter 2 (iii) a.e. $G \in \mathcal{G}(n,p)$ is k-connected for given constant k.*

Proof. (i) and (ii) are immediate.

(iii) In a.e. $G \in \mathcal{G}(n,p)$, no two nodes u_1, u_2 can be separated by a cutset of size $k-1$, because we may choose in Lemma 7.2 $U = u_1,u_2$, $W = w_1,\ldots,w_{k-1}$ for arbitrary $w_1,\ldots,w_{k-1}$, and obtain a path u_1—v—u_2 connecting u_1, u_2 and avoiding $w_1,\ldots,w_{k-1}$. □

Corollary 7.4 *Let ϕ be any first-order sentence about graphs (i.e. quantification over nodes, relations $E(u,v)$ + identity). Then either $G \models \phi$ or $G \models \neg\phi$ for a.e. $G \in \mathcal{G}(n,p)$.*

Proof. Skipped. □

Thus, all the first-order properties of $\mathcal{G}(n,p)$ for fixed p are easily captured. Things are more interesting when the number of nodes discussed and/or the probability p depends on n.

Given graph G, denote:

independence number $\alpha(G)$	=	size of the largest independent set in G,
clique number $\omega(G)$	=	size of the largest clique in G,
chromatic number $\chi(G)$	=	smallest number of colours needeed for colouring nodes in G so that no two adjacent nodes get the same colour.

Lemma 7.5 *Given $n \geq k \geq 2$, random $G \in \mathcal{G}(n,p)$:*

$$\Pr(\alpha(G) \geq k) \leq \binom{n}{k} q^{\binom{k}{2}}.$$

Proof. Probability that given k-set of nodes in G is independent is $q^{\binom{k}{2}}$. Total number of k-sets is $\binom{n}{k}$. □

Theorem 7.6 *Let $p, 0 < p < 1$ and $\varepsilon > 0$ be constant. Then for a.e. $G \in \mathcal{G}(n,p)$:*

$$\chi(G) \geq \frac{\ln 1/q}{2+\varepsilon} \cdot \frac{n}{\ln n} = \Omega\left(\frac{n}{\ln n}\right) = \textit{large!}$$

Proof. By Lemma 7.5, for any fixed $n \geq k \geq 2$:

$$\begin{aligned}\Pr(\alpha(G) \geq k) &\leq \binom{n}{k} q^{\binom{k}{2}} \leq n^k q^{\binom{k}{2}} \\ &= q^{k\frac{\ln n}{\ln q} + \frac{1}{2}k(k-1)} \\ &= q^{\frac{k}{2}\left[-\frac{2\ln n}{\ln 1/q} + k - 1\right]} \\ &\to 0 \text{ for } k \text{ large,}\end{aligned}$$

i.e. when

$$\frac{k}{2}\left[-\frac{2\ln n}{\ln 1/q} + k - 1\right] \to \infty.$$

A sufficient condition for this to hold is that $k \geq k(n,\varepsilon) = (2+\varepsilon)\frac{\ln n}{\ln 1/q}$. Thus for large n, almost no graph $G \in \mathcal{G}(n,p)$ can have a colouring that would assign the same colour to $k(n,\varepsilon)$ or more nodes. Hence, a proper colouring of almost any $G \in \mathcal{G}(n,p)$ requires at least $\frac{n}{k(n,\varepsilon)} = \frac{\ln 1/q}{2+\varepsilon} \cdot \frac{n}{\ln n}$ colours. □

Theorem 7.7 *Let p, $0 < p < 1$ be constant. Then for a.e. $G \in \mathcal{G}(n,p)$:*

$$\omega(G) \in \{d, d+1\},$$

where $d = d(n,p)$ is the largest integer such that

$$\binom{n}{d} p^{\binom{d}{2}} \geq \ln n.$$

(This implies $d = 2\log_{1/p}(n) + O(\log\log n.)$.) □

A *graph property* Q is an isomorphism-closed family of graphs, i.e. if $G \in Q$ (or "G has Q") and $G \approx G'$, then also $G' \in Q$.

A *threshold function* for a graph property Q is a function $t : \mathbb{N} \to \mathbb{R}$ such that

$$\Pr(G \in \mathcal{G}(n,p(n)) \text{ has } Q) \xrightarrow[n\to\infty]{} \begin{cases} 1, \text{ if } p \succ t, \\ 0, \text{ if } p \prec t, \end{cases}$$

where:

$$p \succ t \Leftrightarrow \lim_{n\to\infty} \frac{p(n)}{t(n)} = \infty,$$
$$p \prec t \Leftrightarrow \lim_{n\to\infty} \frac{p(n)}{t(n)} = 0.$$

Further notation:

$$p \sim t \Leftrightarrow \lim_{n\to\infty} \frac{p(n)}{t(n)} = 1,$$
$$p \approx t \Leftrightarrow p(n) = \Theta(t(n)).$$

Denote: $P_n^Q(p) = \Pr(G \in \mathcal{G}(n,p) \text{ has } Q)$.

For technical reasons, we will actually use the following slightly stronger definition for a threshold function: $t(n)$ is a threshold function for graph property Q if

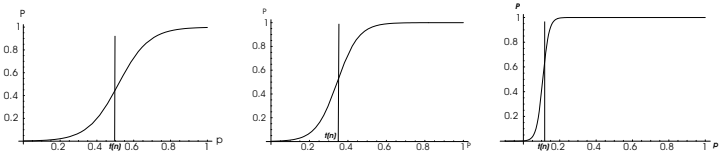


Figure 7: $P_n^Q(p)$ for (a) small, (b) intermediate and (c) large n.

for any sequence $n_1 < n_2 < \ldots$ of graph sizes and $p(n_1), p(n_2), \ldots$ of associated edge probabilities,

$$\lim_{k\to\infty} \frac{p(n_k)}{t(n_k)} = \infty \Rightarrow P_{n_k}^Q(p(n_k)) = 1, \qquad (*)$$
$$\lim_{k\to\infty} \frac{p(n_k)}{t(n_k)} = 0 \Rightarrow P_{n_k}^Q(p(n_k)) = 0. \qquad (**)$$

A graph property is *monotone* if it is preserved under addition of edges, i.e. if $G = (V,E)$ and $G' = (V,E')$ are graphs such that $E \subseteq E'$ and G has Q, then also G' has Q. For monotone Q it is the case that $p_1 \leq p_2 \Rightarrow P_n^Q(p_1) \leq P_n^Q(p_2)$, so the inverse of $P_n^Q(p)$ is well-defined:

$$p_n^Q(\alpha) = \text{the smallest } p \text{ such that } P_n^Q(p) \geq \alpha.$$

In fact for monotone Q one can show that $P_n^Q(p)$ is a continuous, strictly increasing function of p, so actually $p_n^Q(\alpha) = \textit{unique } p$ such that $P_n^Q(p) = \alpha$.

Figure 7 illustrates the evolution of the function P_n^Q, and a corresponding threshold function $t(n)$, for a monotone graph property Q from small to large values of n.

Lemma 7.8 *A function $t(n)$ is a threshold for monotone graph property Q if and only if $t(n) \approx p_n^Q(\alpha)$ for all $0 < \alpha < 1$.*

Proof. Suppose that $t(n)$ is threshold function for Q, but $t(n) \not\approx p_n^Q(\alpha)$ for some $0 < \alpha < 1$. Denoting for brevity $p(n) = p_n^Q(\alpha)$, this means that either there is a sequence $n_1, n_2, \ldots$ such that

$$p(n_k)/t(n_k) \to \infty,$$

or there is a sequence $n_1, n_2, \ldots$ such that

$$p(n_k)/t(n_k) \to 0.$$

However, since for all n it holds that $P_n^Q(p(n)) = P_n^Q(p_n^Q(\alpha)) = \alpha$, $0 < \alpha < 1$, the former case violates condition (*) and the latter case condition (**) in the definition of a threshold function.

"$\Leftarrow$" Assume then that $t(n)$ is *not* a threshold function for Q. Then there are either a sequence $n_1, n_2, \ldots$ and a constant $\alpha < 1$ such that

$$p(n_k)/t(n_k) \to \infty \quad \text{but} \quad P_{n_k}^Q(p(n_k)) \leq \alpha,$$

or a sequence $n_1, n_2, \ldots$ and a constant $\alpha > 0$ such that

$$p(n_k)/t(n_k) \to 0 \quad \text{but} \quad P_{n_k}^Q(p(n_k)) \geq \alpha.$$

In the former case,

$$t(n_k) \prec p(n_k) \leq p_{n_k}^Q(\alpha),$$

and in the latter case

$$t(n_k) \succ p(n_k) \geq p_{n_k}^Q(\alpha).$$

Thus in either case, $t(n) \not\approx p_n^q(\alpha)$ for some $0 < \alpha < 1$. □

Theorem 7.9 *Every monotone graph property Q has a threshold function.*

Proof. For brevity, denote $p_n^Q(\alpha) = p(\alpha)$. Choose some arbitrary $0 < \alpha < \frac{1}{2}$. The goal is to prove that $p(\alpha) \approx p(1-\alpha)$, thus establishing e.g.

$$t(n) = p\left(\frac{1}{2}\right) = p_n^Q\left(\frac{1}{2}\right)$$

as a threshold function for Q. (Since $p(\alpha) \leq p(\frac{1}{2}) \leq p(1-\alpha)$.)

Let $m \in \mathbb{N}$ be such that $(1-\alpha)^m \leq \alpha$. Let $p = p_n(\alpha)$ and consider a sample of m independent graphs $G_1, \ldots, G_m$ from $\mathcal{G}(n,p)$. Then the graph $G_1 \cup \cdots \cup G_m \in \mathcal{G}(n,q)$, where $q = 1-(1-p)^m \leq mp$, and so

$$\Pr(G_1 \cup \cdots \cup G_m \text{ has } Q) \leq \Pr(G \in \mathcal{G}(n, mp_n(\alpha)) \text{ has } Q).$$

On the other hand, since Q is monotone, if any G_i has Q, then so does $G_1 \cup \cdots \cup G_m$. Thus,

$$\begin{aligned} \Pr(G_1 \cup \cdots \cup G_m \text{ does not have } Q) &\leq (1 - \Pr(G_i \text{ has } Q))^m \\ &= (1-\alpha)^m \leq \alpha. \end{aligned}$$

Hence,

$$\Pr_n^Q(mp_n(\alpha)) \geq \Pr(G_1 \cup \cdots \cup G_m \text{ has } Q) \geq 1-\alpha,$$

and so

$$p_n(\alpha) \leq p_n(1-\alpha) \leq mp_n(\alpha),$$

i.e. $p(\alpha) \approx p(1-\alpha)$. (Since m depends only on α, not on n.) □

Consider a graph property Q defined as "G has Q" if $X(G) > 0$, where $X \geq 0$ is a random variable on $\mathcal{G}(n,p)$.

E.g. if $X(G)$ denotes the number of spanning trees of G, then property Q corresponds to connectedness.

Recall the two properties characterising a threshold function $t(n)$:

(i) $p(n) \prec t(n) \Rightarrow$ almost no $G \in \mathcal{G}(n, p(n))$ has Q.
(ii) $p(n) \succ t(n) \Rightarrow$ almost all $G \in \mathcal{G}(n, p(n))$ have Q.

If X is integral, then one can aim to verify conditions (i) and (ii) by the so called "first-moment method" and "second-moment method", respectively.

The first-moment method consists simply of upper-bounding the expectation $E[X]$ and applying Markov's inequality:

$$\Pr(X \geq 1) \leq E[X] \qquad \begin{array}{l}(\text{ more generally, for } a > 0 \\ \quad p(X \geq a) \leq E[X]/a\ ).\end{array}$$

More specifically, one aims to show that if the choice of edge probabilities satisfies $p(n) \prec t(n)$, then $E[X_n] \to 0$. By Markov's inequality it then follows that also $P_n^Q(p(n)) = \Pr(X_n \geq 1) \to 0$.

The second-moment method is based on lower-bounding $E[X]$ *and* upper-bounding $\mathrm{Var}[X]$.

Denote $\mu_n = E[X_n]$, $\sigma_n^2 = \mathrm{Var}[X_n] = E[(X_n - \mu_n)^2] = E[X_n^2] - \mu_n^2$. Recall Chebyshev's inequality (a simple consequence of Markov's inequality): for any $\lambda > 0$,

$$\Pr(|X - \mu| \geq \lambda) \leq \frac{\sigma^2}{\lambda^2}.$$

Lemma 7.10 *If $\mu_n > 0$ for n large, and $\frac{\sigma_n^2}{\mu_n^2} \to 0$ as $n \to \infty$, then $\Pr(X_n > 0) \to 1$ as $n \to \infty$.*

Proof. If $X_n = 0$, then $|X_n - \mu_n| = \mu_n$. Hence

$$\Pr(X_n = 0) \leq \Pr(|X_n - \mu_n| \geq \mu_n) \leq \frac{\sigma_n^2}{\mu_n^2} \to 0 \text{ as } n \to \infty. \quad \square$$

For the next result, denote the number of nodes in a graph G by $|G|$, the number of edges by $e(G)$, and define its *density* as $\delta(G) = \frac{e(G)}{|G|}$. Aay that a graph G is *balanced* if $\delta(G') \leq \delta(G)$ for all subgraphs G' of G.

Theorem 7.11 *Let H be a balanced graph. Then the graph property "G has a subgraph isomorphic to H" has threshold function $n^{-1/\delta(H)}$.*

Proof. Denote $X(G)$ =number of H-subgraphs of a given graph G. Let $k = |H|$, $l = e(H)$, so $\delta(H) = l/k$, and let $G \in \mathcal{G}(n,p)$, where $p = \gamma n^{-1/\delta(H)} = \gamma n^{-k/l}$ for some $\gamma = \gamma_n$. Let us first apply the first-moment method to show that if $\gamma \to 0$, then almost no G contains a subgraph isomorphic to H. Denote

$$\mathcal{H} = \{\text{all copies of } H \text{ on vertex-set of } G\}.$$

Then $|\mathcal{H}| = \binom{n}{k} h \leq \binom{n}{k} k! \leq n^k$, where h is the number of different arrangements of H on a set of k vertices, $h = k!/|\mathrm{Aut}(H)|$. Thus

$$\begin{aligned} E[X] &= \sum_{H' \in \mathcal{H}} \Pr(H' \subseteq G) = |\mathcal{H}| \cdot p^l \\ &\leq n^k p^l = n^k (\gamma n^{-k/l})^l = \gamma^l \xrightarrow[\gamma \to 0]{} 0, \end{aligned}$$

and by Markov's inequality the desired result follows.

For the other part, we wish apply the second-moment method to show that if $\gamma \to \infty$, then almost every graph G contains a subgraph isomorphic to H. For this, we need to verify that $\mu = E[X] > 0$ for all sufficiently large n, and then show that

$$\frac{\sigma^2}{\mu^2} = \frac{1}{\mu^2}(E[X^2] - \mu^2) \to 0 \quad \text{as } n \to \infty.$$

The first condition is easy to check: without loss of generality, assume that $\gamma = \gamma_n \geq 1$ for all n. Then:

$$\begin{aligned} \mu = E[X] &= |\mathcal{H}| \cdot p^l \\ &= \binom{n}{k} h \cdot \gamma_n^l \cdot n^{-k} \\ &\geq \text{const} \cdot n^k \cdot h \cdot \gamma_n^l \cdot n^{-k} \\ &> 0. \end{aligned}$$

For the other requirement, let us try to compute:

$$\begin{aligned} E[X^2] &= \sum_{H',H'' \in \mathcal{H}} \Pr(H' \cup H'' \subseteq G) \\ &= \sum_{H',H'' \in \mathcal{H}} p^{e(H')+e(H'')-e(H' \cap H'')} \\ &\leq \sum_{H',H'' \in \mathcal{H}} p^{2l - i\delta(H)}, \end{aligned}$$

where $i = |H' \cap H''|$. (Note that $\delta(H' \cap H'') \leq \delta(H)$.)
Denote then $\mathcal{H}_i^2 = \{(H',H'') \in \mathcal{H}^2 : |H' \cap H''| = i\}$ and compute separately for each i the sum

$$A_i = \sum_{\mathcal{H}_i^2} \Pr(H' \cup H'' \subseteq G)$$

Case $i = 0$:

$$\begin{aligned} A_0 &= \sum_{\mathcal{H}_0^2} \Pr(H' \cup H'' \subseteq G) \\ &= \sum_{\mathcal{H}_0^2} \Pr(H' \subseteq G) \cdot \Pr(H'' \subseteq G) \qquad H', H'' \text{ independent} \\ &\leq \sum_{\mathcal{H}^2} \Pr(H' \subseteq G) \cdot \Pr(H'' \subseteq G) \\ &= \left(\sum_{\mathcal{H}} \Pr(H' \subseteq G) \right)^2 \\ &= \mu^2. \end{aligned}$$

Case $i \geq 1$:

$$\begin{aligned}
A_i &= \sum_{\mathcal{H}_i^2} \Pr(H' \cup H'' \subseteq G) \\
&= \sum_{H' \in \mathcal{H}} \sum_{\substack{H'': \\ |H' \cap H''| = i}} \Pr(H' \cup H'' \subseteq G) \\
&\leq |\mathcal{H}| \cdot \binom{k}{i}\binom{n-k}{k-i} h p^{2l} p^{-il/k} \qquad\qquad h = \frac{k!}{|\mathrm{Aut}(H)|} \\
&\leq |\mathcal{H}| \cdot c_1 n^{k-i} h p^{2l} (\gamma n^{-k/l})^{-il/k} \\
&= \mu \cdot c_1 n^{k-i} h p^l \gamma^{-il/k} n^i \\
&= \mu \cdot c_1 n^k h p^l \gamma^{-il/k} \\
&= \mu c_2 \underbrace{\binom{n}{k}}_{|\mathcal{H}|} h\, p^l \gamma^{-il/k} \\
&= \mu^2 \cdot c_2 \gamma^{-il/k} \\
&\leq \mu^2 \cdot c_2 \gamma^{-l/k}.
\end{aligned}$$

Thus, denoting $c_3 = kc_2$, we get the estimate

$$\frac{E[X^2]}{\mu^2} = \left(\frac{A_0}{\mu^2} + \frac{\sum_i A_i}{\mu^2}\right) \leq 1 + c_3 \gamma^{-l/k}$$

and hence

$$\frac{\sigma^2}{\mu^2} = \frac{E[X^2] - \mu^2}{\mu^2} \leq c_3 \gamma^{-l/k} \xrightarrow[\gamma \to \infty]{} 0.$$

The desired result then follows by Lemma 7.10. □

Corollary 7.12 *For $k \geq 3$, the property of containing a k-cycle has threshold $t(n) = n^{-1}$. (Note that the threshold is independent of k.)* □

Corollary 7.13 *For $k \geq 2$, the property of containing a specific tree structure T on k nodes has threshold function $t(n) = n^{-k/(k-1)}$.* □

Corollary 7.14 *For $k \geq 2$, the property of containing a k-clique ($\approx K_k$) has threshold function $t(n) = n^{-2/(k-1)}$.* □

Denote $\delta^*(H) = \max\{\delta(H') | H' \text{ is subgraph of } H\}$.

Theorem 7.11' *The graph property "G has a subgraph isomorphic to H" has threshold function $n^{-1/\delta^*(H)}$.* □

Threshold functions for global graph properties

Also known as the "phase transition".

The "epochs of evolution": Consider the structure of random graphs $G \in \mathcal{G}(n,p)$, as $p = p(n)$ increases. The following results can be shown (note that np = average node degree):

0. If $p \prec n^{-2}$, then a.e. G is empty.
1. If $n^{-2} \prec p \prec n^{-1}$, then a.e. G is a forest (a collection of trees).
   - The threshold for the apperarance of any k-node tree structure is $p = n^{-k/(k-1)}$.
   - The threshold for the appearance of cycles (of all constant sizes) is $p = n^{-1}$.
2. If $p \sim cn^{-1}$ for any $c < 1$ (i.e. $np \to c < 1$ as $n \to \infty$), then a.e. G consists of components with at most one cycle and $\Theta(\log n)$ nodes.
3. "Phase transition" or "emergence of the giant component" at $p \sim n^{-1}$ (i.e. $np \to 1$).
4. If $p \sim cn^{-1}$ for any $c > 1$ (i.e. $np \to c > 1$), then a.e. G consists of a unique "giant" component with $\Theta(n)$ nodes and small components with at most one cycle.
5. If $n^{-1} \prec p \prec \frac{\ln n}{n}$, then a.e. G is disconnected, consisting of one giant component and trees.
6. If $p \succ \frac{\ln n}{n}$, then a.e. G is connected (in fact Hamiltonian).

Theorem 7.15 *Let $p_l(n) = \frac{\ln n - \omega(n)}{n}$, $p_u(n) = \frac{\ln n + \omega(n)}{n}$ where $\omega(n) \to \infty$. Then*

(i) *a.e. $G \in \mathcal{G}(n, p_l)$ is disconnected;*

(ii) *a.e. $G \in \mathcal{G}(n, p_u)$ is connected.*

Proof. We shall use the second moment method on random variables $X_k = X_k(G)$ = number of components on G with exactly k nodes.

Assume without loss of generality that $\omega(n) \leq \ln\ln n$ and $\omega(n) \geq 10$.

(i) Set $p = p_l$ and compute $\mu = E(X_1), \sigma^2 = \mathrm{Var}(X_1)$. By linearity of expectation,

$$\begin{aligned}\mu = E(X_1) &= n(1-p)^{n-1} = ne^{(n-1)\ln(1-p)} \\ &\le ne^{-np} = ne^{-\ln n+\omega(n)} = e^{\omega(n)} \xrightarrow[n\to\infty]{} \infty.\end{aligned}$$

Furthermore, the expected number of ordered pairs of isolated nodes is

$$E(X_1(X_1-1)) = n(n-1)(1-p)^{2n-3}.$$

Hence,

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X_1) = E(X_1^2) - \mu^2 \\ &= E(X_1(X_1-1)) + \mu - \mu^2 \\ &= n(n-1)(1-p)^{2n-3} + n(1-p)^{n-1} - n^2(1-p)^{2n-2} \\ &\le n(1-p)^{n-1} + pn^2(1-p)^{2n-3} \\ &\le \mu + (\ln n - \omega(n))ne^{-2\ln n+2\omega(n)} \underbrace{(1-p)^{-3}}_{\le 2} \\ &\le \mu + \frac{2\ln n}{n}e^{2\omega(n)} \le \mu + 1 \qquad \text{for large } n.\end{aligned}$$

Thus, $\frac{\sigma^2}{\mu^2} \le \frac{\mu+1}{\mu^2} \to 0$ as $n \to \infty$, and by lemma 7.10,

$$\Pr(G \text{ is disconnected}) \ge \Pr(X_1(G) > 0) \to 1 \text{ as } n \to \infty.$$

(ii) (Here basic expectation estimation, or "1st moment method" suffices.)
Set $p = p_u = \frac{\ln n+\omega(n)}{n}$ and compute

$$\begin{aligned}\Pr(G \text{ is disconnected}) &= \Pr\left(\sum_{k=1}^{\lfloor n/2\rfloor} X_k \ge 1\right) \\ &\le E\left(\sum_{k=1}^{\lfloor n/2\rfloor} X_k\right) = \sum_{k=1}^{\lfloor n/2\rfloor} E(X_k) \\ &\le \sum_{k=1}^{\lfloor n/2\rfloor} \binom{n}{k}(1-p)^{k(n-k)} \qquad (5)\end{aligned}$$

Split the sum (5) in two parts:

$$\begin{aligned}\text{(a)} &\sum_{1\le k\le n^{3/4}} \binom{n}{k}(1-p)^{k(n-k)} \\ &\le \sum_{1\le k\le n^{3/4}} \left(\frac{en}{k}\right)^k e^{k(n-k)(-p)} \\ &= \sum_{1\le k\le n^{3/4}} \left(\frac{en}{k}\right)^k e^{-knp}e^{k^2p} \\ &\le \sum_{1\le k\le n^{3/4}} k^{-k}n^k e^k e^{-k(\ln n+\omega(n))}e^{k^2\cdot 2\ln n/n} \\ &= \sum_{1\le k\le n^{3/4}} k^{-k}e^{(1-\omega(n))k}e^{2k^2\ln n/n} \\ &\le e^{-\omega(n)} \cdot \underbrace{\sum_{1\le k\le n^{3/4}} \exp\left(-k\ln k + k + 2k^2\frac{\ln n}{n}\right)}_{\le 3} \\ &\le 3e^{-\omega(n)}.\end{aligned}$$

$$\begin{aligned}\text{(b)} &\sum_{n^{3/4}\le k\le n/2} \binom{n}{k}(1-p)^{k(n-k)} \\ &\le \sum_{n^{3/4}\le k\le n/2} \left(\frac{en}{k}\right)^k e^{k(n-k)(-p)} \\ &\le \sum_{n^{3/4}\le k\le n/2} \left(en^{1/4}\right)^k n^{-n/4} \\ &\le \frac{n}{2}e^{n/2}n^{-\frac{1}{4}n^{3/4}} \\ &\le n^{-n^{3/4}/5} \\ &= \exp\left(-\frac{n^{3/4}}{5}\ln n\right) \\ &\le e^{-\omega(n)} \text{ for large } n.\end{aligned}$$

Thus, altogether

$$\Pr(G \text{ is disconnected}) \le 4e^{-\omega(n)} \xrightarrow[n\to\infty]{} 0. \ \square$$

What happens at the "phase transition" $p \sim n^{-1}$? For fixed values of n and $N = \binom{n}{2}$, consider the space of "graph processes" $\widetilde{G} = (G_t)_{t=0}^{N}$, where at each "time instant" t a new edge is selected uniformly at random for insertion into an n-node graph. (Thus, picking graph G_t from a randomly chosen process $\widetilde{G} \in \mathcal{G}(n,M)$, where $M = t$.)

Theorem 7.16 *Let $c > 0$ be a constant and $\omega(n) \to \infty$. Denote $\beta = (c - 1 - \ln c)^{-1}$ and $t = t(n) = \lfloor cn/2 \rfloor$. Then*

(i) *At $c < 1$, every component C of a.e. G_t satisfies*

$$\left| |C| - \beta\left(\ln n - \frac{5}{2}\ln\ln n\right)\right| \leq \omega(n).$$

(ii) *At $c = 1$, for any fixed $h \geq 1$ the h largest components C of a.e. G_t satisfy*

$$|C| = \Theta(n^{2/3}).$$

(iii) *At $c > 1$, the largest component C_0 of a.e. G_t satisfies*

$$||C_0| - \gamma n| \leq \omega(n) \cdot n^{1/2},$$

where $0 < \gamma = \gamma(c) < 1$ is the unique root of

$$e^{-c\gamma} = 1 - \gamma.$$

The other components C of a.e. G_t satisfy also in this case

$$\left| |C| - \beta\left(\ln n - \frac{5}{2}\ln\ln n\right)\right| \leq \omega(n).$$

Thus, the fraction of nodes in the "giant" component of a.e. G_t for $t = cn/2$ behaves as illustrated in Figure 8.

Let us prove one part of this result, the emergence of a gap in the component sizes of $G \in \mathcal{G}(n,p)$ at $p \sim n^{-1}$. (This corresponds to $t \sim N_p \sim n/2$.)

Theorem 7.17 *Let $a \geq 2$ be fixed. Then for large n, $\varepsilon = \varepsilon(n) < 1/3$ and $p = p(n) = (1+\varepsilon)n^{-1}$, with probability at least $1 - n^{-a}$, a random $G \in \mathcal{G}(n,p)$ has no component C that satisfies*

$$\frac{8a}{\varepsilon^2}\ln n \leq |C| \leq \frac{\varepsilon^2}{12}n.$$

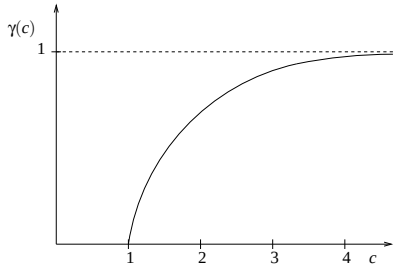


Figure 8: Fraction of nodes in the giant component.

Proof. Let us consider "growing" the component $C(u)$ of an arbitrary node u in G incrementally as follows:

1. (Stage 0:) Set $A_0 = \varnothing, B_0 = \{u\}$.
2. (Stage $i+1$:) If $B_i = A_i$, then stop with $C(u) = B_i$. Otherwise pick an arbitrary $v \in B_i \setminus A_i$; set $A_i = A_i \cup \{v\}$, $B_{i+1} = B_i \cup \{\text{neighbours of } v \text{ in } G\}$.

Now what is the probability distribution of $|B_i|$ (=size of set B_i)?

Consider any node $v \in G \setminus \{u\}$. It participates in i independent Bernoulli trials for being included in B_i, each with success probability equal to p. Thus the inclusion probability for any fixed $v \neq u$ is $1 - (1-p)^i$, independently of each other.

Consequently, the size of each B_i obeys a simple binomial distribution

$$\Pr(|B_i| = k) = \binom{n-1}{k}(1-(1-p)^i)^k(1-p)^{i(n-k-1)}.$$

This gives also for each k an upper bound on the probability

$$\Pr(|C(u)| = k) = \Pr(|B_i| = k \wedge \text{ process stops at stage } i).$$

Denoting $p_k = \Pr(|C(u)| = k)$ for any fixed $u \in G$, it is clear that

$$\Pr(G \text{ contains a component of size } k) \leq np_k,$$

and to prove the theorem it suffices to show that

$$\sum_{k=k_0}^{k_1} p_k \leq n^{-a-1},$$

where $k_0 = \lceil 8a\varepsilon^{-2}\ln n\rceil$, $k_1 = \lceil \varepsilon^2 n/12\rceil$.

Since presumably $k_0 \le k_1$, we may assume $\varepsilon^4 \ge \frac{96a\ln n}{n} \ge \frac{1}{n}$.

We may now estimate

$$p_k \le \Pr(|B_i| = k) \le \frac{n^k}{k!} e^{-\frac{k^2}{2n}} (kp)^k (1-p)^{k(n-k-1)}, \tag{6}$$

because

$$\binom{n-1}{k} = \frac{n^k}{k!}\prod_{j=1}^{k}\left(1-\frac{j}{n}\right) \le \frac{n^k}{k!}e^{-\frac{k^2}{2n}}, \text{ and}$$

$$(1-p)^k \ge 1-kp.$$

Applying Stirling's formula

$$\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \le k! \le e^{\frac{1}{12k}}\sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

and the bounds $k_0 \le k \le k_1$ to (6) we obtain

$$\begin{aligned} p_k &\le \exp\left(\frac{-k^2}{2n} - \frac{\varepsilon^3 k}{3} + \frac{k^2(1+\varepsilon)}{n}\right) \\ &\le \exp\left(\frac{-\varepsilon^2 k}{3} + \frac{k^2}{n}\right) \\ &\le \exp\left(\frac{-\varepsilon^2 k}{4}\right), \end{aligned}$$

and consequently

$$\begin{aligned} \sum_{k=k_0}^{k_1} p_k &\le \sum_{k=k_0}^{k_1} e^{-\varepsilon^2 k/4} \le e^{-\varepsilon^2 k_0/4}\cdot(1-e^{-\varepsilon^2/4})^{-1} \\ &\le \frac{5}{\varepsilon^2}\cdot e^{-\varepsilon^2 k_0/4} \le 5\sqrt{n}\cdot n^{-2a} \\ &= 5n^{-2a+1/2} < n^{-a-1}. \end{aligned}$$

for large n. □

7.2 Nonuniform Models

Introduction

Obviously (in hindsight), most large "real-world" networks do not conform to the Erdös-Rényi random graph model. Consider e.g. the Internet, the WWW, traffic networks (airline connections, roads), collaboration networks (scientists, artistic, business), etc. All these exhibit strong nonuniformities: clustering, nodes with exceptionally high degree, ("hubs") etc.

This was noted (vaguely) in the social sciences at least in the 1960's (Milgram, "six degrees of separation") and also in popular culture ("small worlds", "the Kevin Bacon game").

Curiously, the first serious mathematical (physical) investigation of the phenomenon seems to have been Duncan Watts' Ph.D. thesis (under Steven Strogatz) in 1998 (?), and the "letter" to Nature by Watts and Strogatz in June 1998.

The Watts & Strogatz paper set off a veritable avalanche of work in the area – fueled in no small part by the current interest in modeling the Internet and the WWW.

"Small World" Networks

Watts & Strogatz 1998 etc.

Empirical measurements of real networks vs. predictions of the ER random graph model showed that the ER model is not an adequate model of practical networks.

Statistical measures on a graph $G = (V,E)$, $|V| = n$:

- **Characteristic path length** = average distance between nodes:

  $$\mathcal{L}(G) = \binom{n}{2}^{-1}\sum_{u\neq v}\operatorname{dist}(u,v),$$

  where $\operatorname{dist}(u,v)$ is the length of the shortest path between u and v.

- **Clustering coefficient**

  $$\mathcal{C}(G) = n^{-1}\sum_{v}\rho(\Gamma_v),$$

  where Γ_v is the subgraph of G induced by the neighbours of node v in G,

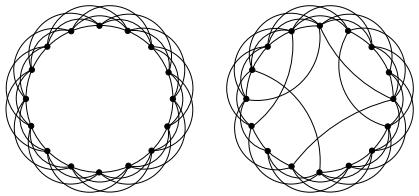


Figure 9: The SW random graph model: circulant graph and rewired graph.

and for a graph Γ with k nodes and l edges, the *density* of Γ is[5]

$$\rho(\Gamma) = l/\binom{k}{2}.$$

Watts and Strogatz considered the following three empirical graphs (n = number of nodes, δ = average node degree; only the largest component of each graph was chosen):

- Hollywood film actors collaboration network: $n = 225226$, $\delta = 61$
- Power grid of the western US: $n = 4941$, $\delta = 2.67$
- Neural network of nematode *Caenorhabditis elegans*: $n = 282$, $\delta = 14$

Watts and Strogatz obtained the following comparisons ($\mathcal{L}_{ER}$ and $\mathcal{C}_{ER}$ denote the corresponding values for ER random graphs of comparable size and density):

	$\mathcal{L}$	$\mathcal{L}_{ER}$	$\mathcal{C}$	$\mathcal{C}_{ER}$
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.08	0.0005
C. elegans	2.65	2.25	0.28	0.05

The empirical conclusion is thus that "real networks" have path length comparable to ER random graphs (= short) but considerably higher clustering. To model such observations, Watts and Strogatz introduced a specific "small world" (SW) random graph model, whereby one starts with a "circulant graph" $C_{n,k}$, and then randomly "rewires" some small fraction p of the edges. (Cf. Figure 9.)

[5]To be precise, the definition requires that $k \geq 2$. For nodes v with 0 or 1 neighbours, it is most convenient to stipulate that the neighbourhood density corresponds to the global density, i.e. that $\rho(\Gamma_v) = |E|/|V|$.

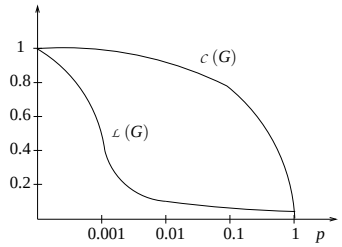


Figure 10: Path length and clustering coefficient in SW random graphs.

Watts & Strogatz experimented on the effect of the rewiring probability p on $\mathcal{L}(G)$ and $\mathcal{C}(G)$ in this model and obtained results as indicated in Figure 10 (curves normalised by $\mathcal{C}(C_{n,k})$ and $\mathcal{L}(C_{n,k})$; $n = 1000$, $k = 5$). Thus, the "small world" phenomenon of small $\mathcal{L}$ and large $\mathcal{C}$ seems to occur for p in the range $0.0005 \ldots 0.05$.

Watts and Strogatz call all graph families with this qualitative property "small world graphs". The notion has also been quantified by Walsh (1999) in terms of the *proximity ratio*

$$\mu = \frac{\mathcal{C}/\mathcal{L}}{\mathcal{C}_{ER}/\mathcal{L}_{ER}}.$$

Thus, presumably $\mu \gg 1$ for small world graphs. However, this quantity does not seem to be very invariant over various SW graph families. E.g. for *C. elegans*, $\mu \approx 4.8$ and for the power grid graph $\mu \approx 106$, but for the actors' network $\mu \approx 2400$.

For analytical simplicity, Newman et al. (1999, 2000) modified the Watts-Strogatz SW model to simply adding a fraction p of random cross edges, rather than rewiring. This variant of the model is called the "solvable SW", or SSW model.

Other Small World Models

- **Kleinberg's (2000) lattice model:** Basis is an $s \times s$ square lattice, with Manhattan (L_1) metric:

  $$\mathrm{d}(u,v) = \mathrm{d}((i,j),(k,l)) = |k-i| + |l-j|.$$

  Each node u has local connections to all nodes within distance $d \leq p$, and in addition $q \geq 0$ directed "long distance" connections. The probability of creating a long distance connection between u and v is proportional to thei distance, $\Pr((u,v)) \propto \mathrm{d}(u,v)^{-r}, r \geq 0$.

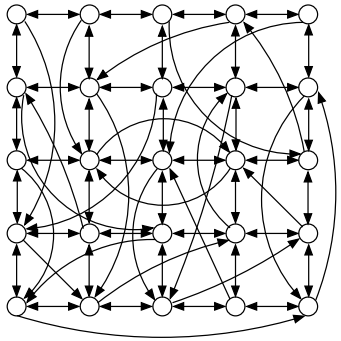


Figure 11: A Kleinberg lattice.

- **"Caveman graphs":** (Watts 1999; old idea?) Deterministic SW graph model. Connect a collection of r "k-man caves" (k-cliques) together in a systematic manner.

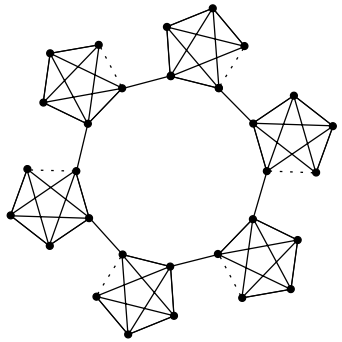


Figure 12: A collection of six 5-caves connected together in a 6-cycle.

Scale Free Networks

So are small world graphs a good model of real world networks? Not always. (Usually not?)

One aspect of real networks that SW graphs often do not model well is the degree

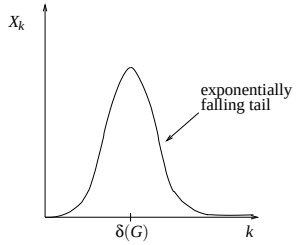


Figure 13: Degree distribution of an ER random graph.

distribution. In an ER random graph $G \in \mathcal{G}(n,p)$, the degree distribution is almost binomial with parameters $n-1, p$. For large n and small p, the distribution approaches Poisson(λ), where $\lambda = np$.

More precisely, if $X_k = X_k(G)$ = number of nodes in G with deg = k, then

$$P(k) = \frac{E(X_k)}{n} = \binom{n-1}{k} p^k (1-p)^{n-1-k} \approx e^{-np} \frac{(np)^k}{k!} \approx e^{-\delta} \frac{\delta^k}{k!},$$

where δ = average degree of graph G. Thus, the degree distribution of a typical ER graph G looks as illustrated in Figure 13.

The degree distributions of SW graphs are typically even more peaked around $\delta(G)$. E.g. in WS graphs based on the circulant $C_{n,t}$, approximately fraction $1-2tp$ of the nodes has degree equal to $2t$ (recall that $p \ll 1$ is the rewiring probability).

However, many real world networks seem to have very heavy tailed degree distributions, well matched by "power laws"

$$P(k) \propto k^{-\gamma},$$

where $\gamma = 2 \ldots 4$. This indicates that there are some nodes with unreasonably large (in the ER or SW models) degrees. Also, such networks are called "scale free", because there is no characteristic "scale" or node degree value at which large networks would concentrate.

On a log-log plot, the degree distributions of such networks look somewhat as in Figure 14

For instance, the following values for γ have been estimated for real world networks (Barabási & Albert 1999)

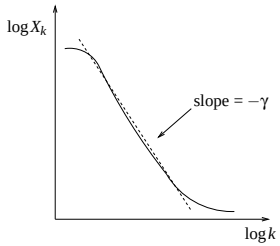


Figure 14: Degree distribution of a "scale-free" random graph.

	n	δ	γ
Film actors	212250	28.8	2.3 ± 0.1
WWW (local)	325729	5.46	2.1 ± 0.1
Power grid	4941	2.67	4

Barabási & Albert (1999) proposed the following attractive "growth and preferential attachment" model (BA model) to explain the emergence of such power law degree distributions in networks:

- The network is initialised at time $t = 0$ with some small set of nodes and edges, $G_0 = (V_0, E_0)$
- At time $t+1$, a new node u is introduced to the network, with d_0 edges that are preferentially attached to the existing nodes $v \in V_t$ so that

$$\Pr((u,v) \in E_{t+1}) \propto \deg_t(v).$$

Barabási and Albert argue heuristically and experimentally that this growth process yields networks with power law degree distributions

$$P(k) \propto k^{-3}.$$

They also claim that with nonlinear preferences the exponent γ can be adjusted also to values different than 3.

These arguments have been made rigorous by Eriksen & Hörnquist (2002) and by Krapivsky (2000). (However some problems still remain with nonlinear preferences?)

Finally, note that the popular experimental graphs (Internet, actors, power grid, etc.) have both small world and scale free properties, so neither the SW nor the BA model (which are mutually contradictory) provides a fully satisfactory explanation for them.

Part III

Stochastic Algorithms

8 Simulated Annealing

Global optimisation (say, minimisation) of an objective function $H(\sigma)$, framed as a Hamiltonian of a statistical mechanics system, via a sequence of Metropolis samplers for the Gibbs distributions determined by $H(\sigma)$ at decreasing values of the temperature parameter $T \to 0$.

Let $H : S \to \mathbb{R}$ be a function to be minimised over a finite (but typically very large) state space S. Assume that S has some neighbourhood structure $S = (S, N)$ (cf. page 24).

In any specific application of the method, the algorithm designer typically has a lot of freedom in the choice of the most appropriate N. This choice can have a significant effect on the efficiency of the algorithm: one would like to have N such that $N(\sigma)$ is small for each $\sigma \in S$, yet the resulting Metropolis chains converge rapidly.

The Gibbs distribution determined by H at temperature T is (recall page 58):

$$\pi_\sigma^{(T)} = \Pr_T(\sigma) = \frac{1}{Z_T} e^{-H(\sigma)/kT} = \frac{1}{Z_T} e^{-\beta H(\sigma)},$$

where $\beta = 1/kT$.

A relevant observation is that as $T \to 0$ (or $\beta \to \infty$), the distribution $\Pr_T(\sigma)$ gets more peaked according to H. Denoting by $S^* = \{\sigma^* \in S \mid H(\sigma^*) = \min\}$ the set of global optima of H, one observes that:

$$\frac{\Pr_T(\sigma)}{\Pr_T(\sigma^*)} = e^{-\beta(H(\sigma)-H(\sigma^*))} \xrightarrow[(\beta\to\infty)]{T\to 0} \begin{cases} 0, & \sigma \notin S^* \\ 1, & \sigma \in S^* \end{cases}$$

Thus, in the limit one obtains:

$$\pi^*_\sigma = \lim_{T\to 0} \Pr_T(\sigma) = \begin{cases} 0, & \sigma \notin S^* \\ 1/|S^*|, & \sigma \in S^* \end{cases}$$

Of course, one cannot directly sample according to π^*, but the idea is that by starting at a high value of T and then slowly (but how slowly?) decreasing it, one obtains a nonhomogenous Metropolis chain that converges reasonably fast (?) to π^*.

As regards the convergence of the chains at each fixed $T > 0$, we can appeal to the general results concerning Metropolis samplers from page 24 onwards.

Let us just check the form of the acceptance probabilities: a proposed move $\sigma \to \tau$, where $\tau \in N(\sigma)$, is accepted with probability:

$$\begin{aligned} p_{\sigma\tau} &= \min\left\{\frac{\pi_\tau d_\sigma}{\pi_\sigma d_\tau}, 1\right\} \\ &= \min\left\{\frac{e^{-\beta H(\tau)}}{e^{-\beta H(\sigma)}} \cdot \frac{d_\sigma}{d_\tau}, 1\right\} \\ &= \min\left\{e^{-\beta(H(\tau)-H(\sigma))} \cdot \frac{d_\sigma}{d_\tau}, 1\right\} \\ &= \min\left\{e^{-\beta(H(\tau)-H(\sigma))}, 1\right\}, \end{aligned}$$

if (S,N) is regular i.e. $|N(\sigma)| = |N(\tau)|$ for all σ, τ.

Thus, for a regular neighbourhood structure, and denoting $\Delta H = H(\tau) - H(\sigma)$, a proposed transition $\sigma \to \tau$ is accepted always if $\Delta H \le 0$, and with probability $e^{-\beta \Delta H}$, if $\Delta H > 0$.[1]

In summary, one obtains the following general method for minimising a function H over a state space S with regular neighbourhood structure N:

Algorithm SA(H, S, N):

$T \leftarrow T_{init}$;
$\sigma \leftarrow \sigma_{init}$;
while $T > T_{final}$ do
    $L \leftarrow$ sweep(T);
    for L times do
        choose $\tau \in N(\sigma)$ uniformly at random;
        $\Delta H \leftarrow H(\tau) - H(\sigma)$;
        if $\Delta H \le 0$ then $\sigma \leftarrow \tau$;
        else choose $r \in [0,1)$ uniformly at random;
            if $r \le \exp(-\Delta H/T)$
            then $\sigma \leftarrow \tau$;
    end for;
    $T \leftarrow$ lower(T);
end while;
result $\leftarrow \sigma$;

[1] In the general case of nonregular neighbourhoods, potential-increasing transitions should be accepted with probability $e^{-\beta\Delta H} \cdot d_\sigma/d_\tau$.

The obvious question is now how to choose appropriate functions lower(T) and sweep(T), i.e. what is a good "cooling schedule" $\langle T_0, L_0\rangle, \langle T_1, L_1\rangle, \ldots$

In practice, it is customary to just start from some "high" temperature T_0, and after each "sufficiently long" sweep L decrease the temperature by some "cooling factor" $\alpha \approx 0.8 \ldots 0.99$:

$$T_{k+1} = \alpha T_k.$$

Theoretically this is much too fast, as we shall see, but often seems to work well enough.

Consider an inhomogenous Markov chain with transition matrices $P^{(0)}, P^{(1)}, P^{(2)}, \ldots$ Denote

$$P(m,k) = P^{(m)} P^{(m+1)} \cdots P^{(m+k-1)}$$

i.e. $P_{ij}(m,k) = \Pr(X_{m+k} = j \mid X_m = i)$.

The chain $\mathcal{M}$ is *weakly ergodic* if for all $m \ge 0$:

$$\lim_{k\to\infty} \sup_{\mu,\nu} d_V\left(\mu^T P(m,k), \nu^T P(m,k)\right) = 0$$

and *strongly ergodic* if there is some distribution π such that for all $m \ge 0$:

$$\lim_{k\to\infty} \sup_{\mu} d_V(\mu^T P(m,k), \pi) = 0$$

Let Q be an $n \times m$ stochastic matrix. The *(Dobrushin) ergodic coefficient* of Q is defined as:

$$\begin{aligned} \rho = \rho(Q) &= \max_{i,j} d_V(q_i, q_j) \qquad \begin{matrix} q_i = (q_{i1}, \ldots, q_{im}) \\ q_j = (q_{j1}, \ldots, q_{jm}) \end{matrix} \\ &= \frac{1}{2} \max_{i,j} \sum_{k=1}^{m} |q_{ik} - q_{jk}| \end{aligned}$$

The following key technical lemmas will possibly be proved later. The proofs are not exceedingly difficult.

Lemma 8.1 ("Dobrushin's inequality")
Given the stochastic matrices $Q_1 \in [0,1]^{n\times m}, Q_2 \in [0,1]^{n\times l}$:

$$\rho(Q_1Q_2) \le \rho(Q_1)\rho(Q_2).$$

Lemma 8.2 ("Dobrushin convergence rate bound")
Given the stochastic matrix P and the distributions μ,ν:

$$d_V(\mu^T P^n, \nu^T P^n) \le d_V(\mu,\nu)\rho(P)^n.$$

Lemma 8.3
An inhomogeneous Markov chain $\mathcal{M}$ with transition probability matrices $P^{(0)}$, $P^{(1)},\ldots$ is weakly ergodic if and only if either (and hence both) of the following conditions hold:

(i) *for any $m \ge 0$: $\lim_{k\to\infty}\rho(P(m,k)) = 0$;*

(ii) *for some increasing sequence $0 \le m_0 < m_1 < \cdots$*

$$\sum_{i=0}^{\infty}(1-\rho(P(m_i,m_{i+1}))) = \infty.$$

Lemma 8.4
Let $\mathcal{M}$ be a weakly ergodic Markov chain with transition probability matrices $P^{(0)},P^{(1)},\ldots$ Suppose that there exists a sequence of distributions $\pi^{(0)},\pi^{(1)},\ldots$ such that

(i) *$\pi^{(m)}P^{(m)} = \pi^{(m)}$, for each $m \ge 0$;*

(ii) $\sum_{m=0}^{\infty} ||\pi^{(m)} - \pi^{(m+1)}||_1 < \infty.$

Then $\mathcal{M}$ is also strongly ergodic, with limit distribution

$$\pi_i^* = \lim_{m\to\infty}\pi_i^{(m)}.$$

Theorem 8.5
Consider a simulated annealing computation on input (H,S,N). Assume the neighbourhood graph (S,N) is connected and regular of degree r. Denote:

$$\Delta = \max\{H(\tau) - H(\sigma) \mid \sigma \in S, \tau \in N(\sigma)\}.$$

Suppose the cooling schedule used is of the form $\langle T_0,L\rangle,\langle T_1,L\rangle,\langle T_2,L\rangle,\ldots$, where

$$L \ge \min_{\sigma^*\in S^*}\max_{\sigma\notin S^*}\mathrm{dist}(\sigma,\sigma^*), \tag{1}$$

where $\mathrm{dist}(\sigma,\sigma^*)$ is the distance in graph (S,N) from σ to σ^*, and for each cooling stage $l \ge 2$:

$$T_l \ge \frac{L\Delta}{\ln l} \quad (\textit{but } T_l \xrightarrow[l\to\infty]{} 0). \tag{2}$$

Then the distribution of states visited by the computation converges in the limit to π^*, where

$$\pi_\sigma^* = \begin{cases} 0, & \text{if } \sigma \notin S^* \\ 1/|S^*|, & \text{if } \sigma \in S^* \end{cases}$$

Proof: Denote by $P^{(0)},P^{(1)},\ldots$ the sequence of transition matrices for the Markov chain on S determined by the SA algorithm with the given parameters. We shall show, based on Lemma 8.4, that this chain is strongly ergodic with the given limit distribution.

Let us first verify weak ergodicity using Lemma 8.3 (ii). Let $\sigma^* \in S^*$ be some ground state achieving the lower bound in condition (1). We shall show that for any $\sigma \in S$ and $k \ge k_0$, where k_0 is sufficiently large:

$$P_{\sigma\sigma^*}(k,k+L) \ge \left(\frac{1}{r}e^{-\Delta/t_k}\right)^L, \tag{3}$$

where $t_k = T_{\lfloor k/L\rfloor}$ = cooling temperature at step k.

It then follows from condition (3) and from the fact $|p-q| = p+q-2\min\{p.q\}$ that

$$\begin{aligned}
&1-\rho(P(k,k+L)) \\
&= 1-\frac{1}{2}\max_{\sigma,\tau}\sum_{\nu\in S}|P_{\sigma\nu}(k,k+L) - P_{\tau\nu}(k,k+L)| \\
&= \min_{\sigma,\tau}\sum_{\nu\in S}\min\{P_{\sigma\nu}(k,k+L), P_{\tau\nu}(k,k+L)\} \\
&\ge \min_{\sigma\in S}P_{\sigma\sigma^*}(k,k+L) \\
&\ge r^{-L}e^{-L\Delta/t_k},
\end{aligned}$$

and so (choosing $m_l = l \cdot L$):

$$\sum_{l=0}^{\infty}(1-\rho(P(m_l,m_{l+1}))) \geq \sum_{l=k_0}^{\infty}(1-\rho(P(lL,lL+L)))$$
$$\geq \sum_{l=k_0}^{\infty} r^{-L}e^{-L\Delta/t_k} \geq r^{-L}\sum_{l=k_0}^{\infty}\frac{1}{l} = \infty.$$

Thus, let us check that condition (3) holds for some sufficiently large k_0. Observe first that for any $\sigma \in S$ and $\tau \in N(\sigma)$:

$$P_{\sigma\tau}(k) = \frac{1}{r}\min\{e^{-(H(\tau)-H(\sigma))/t_k},1\} \geq \frac{1}{r}e^{-\Delta/t_k}.$$

Similarly, for any $\sigma^* \in S^*$ there is some k_0 such that for all $k \geq k_0$:

$$P_{\sigma^*\sigma^*}(k) \geq \frac{1}{r}e^{-\Delta/t_k}.$$

Namely, let $\delta = \min\{H(\tau)-H(\sigma^*) \mid \sigma^* \in S^*, \tau \in N(\sigma^*)\setminus S^*\}$. Now $\delta > 0$, unless H is a constant function. Thus for all $k \geq k_0$, where k_0 is sufficiently large:

$$1-e^{-\delta/t_k} \geq e^{-\Delta/t_k},$$

and so

$$\begin{aligned}
P_{\sigma^*\sigma^*} &= 1-\sum_{\tau\in N(\sigma^*)}P_{\sigma^*\tau}(k)\\
&= 1-\sum_{\tau\in N(\sigma^*)}\frac{1}{r}e^{-(H(\tau)-H(\sigma))/t_k}\\
&\geq 1-\frac{1}{r}(r-1+e^{-\delta/t_k})\\
&= \frac{1}{r}(1-e^{-\delta/t_k})\\
&\geq \frac{1}{r}e^{-\Delta/t_k}.
\end{aligned}$$

Thus, for any $\sigma \in S$ and $k \geq k_0$:

$$\begin{aligned}
&P_{\sigma\sigma^*}(k,k+L)\\
&= \sum_{\tau_1}\sum_{\tau_2}\cdots\sum_{\tau_{L-1}}P_{\sigma\tau_1(k)}P_{\tau_1\tau_2}(k+1)\cdots P_{\tau_{L-1}\sigma^*}(k+L-1)\\
&\geq P_{\sigma\sigma_1}(k)P_{\sigma_1\sigma_2}(k+1)\cdots P_{\sigma_{L-1}\sigma^*}(k+L)\\
&\geq \left(\frac{1}{r}e^{-\Delta/t_k}\right)^L,
\end{aligned}$$

where $\sigma,\sigma_1,\sigma_2,\ldots,\sigma_{L-1},\sigma^*$ is a shortest path from σ to σ^* in (S,N), with possibly state σ^* repeated several times if the length of the actual path is less than L.

Having now established the weak ergodicity of our chain, let us check conditions (i) and (ii) of Lemma 8.4 to complete the proof.

For condition (i) it suffices to observe that the stationary distribution at stage l of the algorithm:

$$\pi_\sigma^{(l)} = \frac{1}{Z_l}e^{-H(\sigma)/T_l}, \quad Z_l = \sum_{\sigma\in S}e^{-H(\sigma)/T_l},$$

satisfies the condition $\pi^{(l)}P^{(m)} = \pi^{(l)}$, for values of m from lL to $(l+1)L-1$.

For condition (ii), one can show by a somewhat tedious calculation (cf. Aarts & Korst, "Simulated Annealing … ", p. 22) that for each of the intermediate stationary distributions $\pi^{(l)}$:

$$\text{if } \sigma^* \in S^*, \text{ then } \frac{\partial}{\partial T}\pi_{\sigma^*}^{(l)} < 0;$$
$$\text{if } \sigma \notin S^*, \text{ then } \frac{\partial}{\partial T}\pi_\sigma^{(l)} > 0 \text{ for } l \geq l_1 \text{ sufficiently large.}$$

As $T_{l+1} \leq T_l$ at each stage l, it thus follows that:

$$\pi_{\sigma^*}^{(l+1)} \geq \pi_{\sigma^*}^{(l)} \text{ for } \sigma^* \in S^*$$
$$\pi_\sigma^{(l+1)} \leq \pi_\sigma^{(l)} \text{ for } \sigma \notin S^* \text{ and } l \geq l_1$$

Thus, for $l \geq l_1$:

$$\begin{aligned}
\left|\left|\pi^{(l)}-\pi^{(l+1)}\right|\right|_1 &= \sum_{\sigma\in S}\left|\pi_\sigma^{(l)}-\pi_\sigma^{(l+1)}\right|\\
&= \sum_{\sigma^*\in S^*}\left|\pi_{\sigma^*}^{(l)}-\pi_{\sigma^*}^{(l+1)}\right| + \sum_{\sigma\notin S^*}\left|\pi_\sigma^{(l)}-\pi_\sigma^{(l+1)}\right|\\
&= 2\left(\sum_{\sigma^*\in S^*}\pi_{\sigma^*}^{(l+1)} - \sum_{\sigma^*\in S^*}\pi_{\sigma^*}^{(l)}\right).
\end{aligned}$$

Hence, denoting $\hat{\pi}^{(m)} = \pi^{(\lfloor m/L \rfloor)}$:

$$\begin{aligned}\sum_{m=0}^{\infty} \left|\left|\hat{\pi}^{(m)} - \hat{\pi}^{(m+1)}\right|\right|_1 &= \sum_{l=0}^{\infty} \left|\left|\hat{\pi}^{(l)} - \hat{\pi}^{(l+1)}\right|\right|_1 \\ &= \sum_{l=0}^{l_1} \left|\left|\hat{\pi}^{(l)} - \hat{\pi}^{(l+1)}\right|\right|_1 + \sum_{l=l_1+1}^{\infty} \left|\left|\hat{\pi}^{(l)} - \hat{\pi}^{(l+1)}\right|\right|_1 \\ &\leq 2l_1 + 2\left(\sum_{\sigma^* \in S^*} \pi^*_{\sigma^*} - \sum_{\sigma^* \in S^*} \pi^{(l_1+1)}_{\sigma^*}\right) \\ &\leq 2l_1 + 2 < \infty.\end{aligned}$$

This completes the proof, because according to Lemma 8.4 the chain has the limit distribution π^*, where

$$\pi^*_\sigma = \lim_{l\to\infty} \pi^{(l)}_\sigma = \lim_{l\to\infty} \frac{1}{Z_l} e^{-H(\sigma)/T_l} = \begin{cases} 0, & \text{if } \sigma \notin S^* \\ 1/|S^*|, & \text{if } \sigma \in S^* \end{cases} \quad \square$$

9 Approximate counting

Let Σ be an alphabet (without loss of generality $\Sigma = \{0,1\}$) and $R \subseteq \Sigma^* \times \Sigma^*$ an NP relation over Σ^*, i.e.

- for some polynomial $p(n)$, $R(x,w) \Rightarrow |w| \leq p(|x|)$, where $|z|$ denotes the length of string z
- the condition $R(x,w)$ can be tested in polynomial time, for any given $\langle x,w\rangle$

Well-known examples of NP relations:

- SAT(ϕ,t), where ϕ is (an encoding of) a Boolean formula and $t : \mathrm{Var}_\phi \to \{T,F\}$ is a truth assignment to its variables; relation holds if ϕ evaluates to T under t.
- $\mathrm{COL}_q(G,\sigma)$, where $G = (V,E)$ is a graph and $\sigma : V \to \{1,\ldots,q\}$ is a candidate q-colouring of its nodes; relation holds if σ is valid for G, i.e. if $(u,v) \in E \Rightarrow \sigma(u) \neq \sigma(v)\ \forall\, u,v \in V$.

Denote $R(x) = \{w \in \Sigma^* | R(x,w) \text{ holds}\}$.

One may consider different computational problems related to R:

- *existence problem*: given x, determine if $R(x) \neq \varnothing$
- *counting problem*: given x, determine $N_R(x) = |R(x)|$
- *sampling problem*: given x, provide $w \in R(x)$ uniformly at random

A *randomised approximation scheme (ras)* for the counting problem associated to R is a randomised algorithm $A(x,\varepsilon)$ such that for any $x \in \Sigma^*, \varepsilon > 0$:

$$\Pr((1-\varepsilon)N_R(x) \leq A(x,\varepsilon) \leq (1+\varepsilon)N_R(x)) \geq \frac{3}{4},$$

where the probability is with respect to the random choices made by the algorithm. The ras is *fully polynomial (fpras)* if its running time is polynomial in $|x|$ and $1/\varepsilon$.

An *almost uniform sampler (aus)* for R is a randomised algorithm $S(x,\delta)$ such that for any $x \in \Sigma^*, S(x,\delta) \in R(x)$ and $d_V(S(x,\delta), U_R(x)) \leq \delta$, where $S(x,\delta)$ denotes (by slight abuse of notation) the distribution of the output of $S(x,\delta)$, and $U_R(x)$ denotes the uniform distribution over $R(x)$. An aus is *fully polynomial (fpaus)* if its running time is polynomial in $|x|$ and $\ln 1/\delta$.

It can be shown (Jerrum et al. 1986, Sinclair 1993) that if R is "self-reducible", then R has an fpras if and only if it has an fpaus.

Self-reducibility of R means roughly (the exact definition is somewhat more general) that there is a small collection of polynomial time functions $f_i, g_i, i = 1,\ldots,k$, such that for any $x \in \Sigma^*$, $|f_i(x)| < |x|$ and

$$R(x) = \bigcup_{i=1}^{k} g_i(x, R(f_i(x))).$$

E.g. for the SAT relation SAT(ϕ) = SAT(ϕ_T) $\cup$ SAT(ϕ_F), where ϕ_T (ϕ_F) is the formula obtained from ϕ by substituting T (F) for the first variable and simplifying. Almost all "natural" NP-complete relations are self-reducible.

Let us see concretely, in the case of low-degree graph colouring, how an efficient fpaus (pages 46-50) can be converted into an efficient fpras.

Given a graph $G = (V,E)$ with maximum node degree $\Delta < q$, denote for brevity $\Omega(G) = \mathrm{COL}_q(G)$, and assume the existence of a fpaus $S(G,\delta)$ for q-colourings. (Actually, the fpaus-construction on pages 46-50 requires more strongly that $\Delta < q/2$.)

One possible self-reduction for graph colouring is

$$\Omega(G) = g(G, \Omega(G')),$$

where $G' \sim G$ with one edge (e.g. highest-numbered one) removed, and

$$g(G,\sigma) = \begin{cases} \sigma & \text{if } \sigma \text{ is valid for } G \\ \perp & \text{otherwise} \end{cases}$$

where $\perp$ is a "null-value" ($S \cup \{\perp\} = S$ for any S).

Assuming $|E| = m$, denote $G = G_m$, $G' = G_{m-1}, \ldots, G^{(m)} = G_0 = (V, \varnothing)$. Now clearly $|\Omega(G_0)| = q^n$, where $n = |V|$. Then the quantity we are interested in can be expressed as:

$$\begin{aligned} N(G) = |\Omega(G)| &= \frac{|\Omega(G)_m|}{|\Omega(G)_{m-1}|} \cdot \frac{|\Omega(G)_{m-1}|}{|\Omega(G)_{m-2}|} \cdots \frac{|\Omega(G)_1|}{|\Omega(G)_0|} \cdot |\Omega(G)_0| \\ &= \rho_m \cdot \rho_{m-1} \cdots \rho_1 \cdot q^n, \end{aligned} \tag{4}$$

where

$$\rho_k = \frac{|\Omega(G)_k|}{|\Omega(G)_{k-1}|}.$$

Now each of the ratios in ρ_k and hence the product (4) can be estimated using our presumed fpaus to generate a "sufficiently large" number of samples form each $\Omega(G_{k-1})$ and seeing how many of those fall also in $\Omega(G_k)$. Some analysis is needed to determine the appropriate numbers.

Before going into the analysis, let us note that the same approach, combined with more complicated samplers, has been used to provide fpras for such important problems as:

- approximating the volume of a convex body (Dyer, Frieze, Kannan 1991)
- approximating the partition function of a ferromagnetic Ising model (Jerrum & Sinclair 1993)
- approximating the permanent of a positive matrix (Jerrum, Sinclair & Vigoda 2001)

Let us then complete the analysis of the graph colouring fpras. Recall that

$$|\Omega(G)| = \rho_m \cdot \rho_{m-1} \cdots \rho_1 \cdot q^n,$$

where each

$$\rho_k = \frac{|\Omega(G)_k|}{|\Omega(G)_{k-1}|}.$$

Now clearly each $\Omega(G_k) \subseteq \Omega(G_{k-1})$, so that $\rho_k \leq 1$. On the other hand, each colouring $\sigma \in \Omega(G_{k-1}) \setminus \Omega(G_k)$ must be such that it assigns the same colour to both endpoints u, v of the edge e removed from G_k to obtain G_{k-1}. Let u be the lower-numbered of the nodes. Then σ can be transformed to a valid colouring of G_k by recolouring u with one of the $\geq q - \Delta \geq 1$ colours free for it. On the other hand, each colouring in $\Omega(G_k)$ is generated by this process in at most one way. Thus

$$|\Omega(G_{k-1}) \setminus \Omega(G_k)| \leq |\Omega(G_k)|,$$

and so $\rho_k \geq \frac{1}{2}$.

Assume then without loss of generality that $m \geq 1$ and $0 < \varepsilon \leq 1$. (Recall $\varepsilon \sim$ error tolerance for the fpras to be constructed).

Let $Z_k \in \{0,1\}$ be a random variable obtained by running the presumed fpaus for G_{k-1} and testing whether the resulting colouring is also valid for G_k ($\rightarrow Z_k = 1$) or not ($\rightarrow Z_k = 0$). Denote $\mu_k = E[Z_k]$.

By setting $\delta = \frac{\varepsilon}{6m}$ in the fpaus one may ensure that

$$\rho_k - \frac{\varepsilon}{6m} \leq \mu_k \leq \rho_k + \frac{\varepsilon}{6m}, \tag{5}$$

and noting the bounds on ρ_k, that

$$\left(1 - \frac{\varepsilon}{3m}\right)\rho_k \leq \mu_k \leq \left(1 + \frac{\varepsilon}{3m}\right)\rho_k. \tag{6}$$

Note also that by (5), $\mu_k \geq \frac{1}{3}$.

To decrease the variance of our ρ_k-estimate, let $Z_k^{(1)}, \ldots, Z_k^{(s)}$ be $s = \lceil 74\varepsilon^{-2}m \rceil \leq 75\varepsilon^{-2}m$ independent copies of variable Z_k, and let

$$\bar{Z}_k = \frac{1}{s}\sum_{i=1}^{s} Z_k^{(i)}$$

be their mean. Then $E[\bar{Z}_k] = E[Z_k] = \mu_k$ and

$$\frac{\mathrm{Var}(\bar{Z}_k)}{\mu_k^2} = \frac{s^{-2} \cdot s \cdot \mathrm{Var}(Z_k)}{\mu_k^2} = \frac{s^{-1}(\mu_k - \mu_k^2)}{\mu_k^2} = s^{-1}(\mu_k^{-1} - 1) \leq 2s^{-1}$$

We shall take as our estimator for $|\Omega(G)|$ the random variable $Y = q^n \mu_1 \cdots \mu_m$.

The variance of Y can be bounded as:

$$\begin{aligned}
\frac{\mathrm{Var}(Y)}{E(Y)^2} &= \frac{\mathrm{Var}(\bar{Z}_1 \cdots \bar{Z}_m)}{(\mu_1 \cdots \mu_m)^2} \\
&= \prod_{k=1}^{m} \left(1 + \frac{\mathrm{Var}(\bar{Z}_k)}{\mu_k^2}\right) - 1 \\
&\le \left(1 + \frac{2}{s}\right)^m - 1 \qquad\qquad s = \lceil 74\tfrac{m}{\varepsilon^2} \rceil \;\Rightarrow\; \tfrac{2}{s} \le \tfrac{2\varepsilon^2}{74m} = \tfrac{\varepsilon^2}{37m} \\
&\le \left(1 + \frac{\varepsilon^2}{37m}\right)^m - 1 \\
&\le e^{\varepsilon^2/37} - 1 \qquad\qquad e^x - 1 = x + \underbrace{\frac{x^2}{2!} + \frac{x^3}{3!} + \cdots}_{\text{small!}} \\
&\le \frac{\varepsilon^2}{36}.
\end{aligned}$$

Since by Chebyshev's inequality:

$$\Pr(|Y - E(Y)| \ge \lambda E(Y)) \le \frac{1}{\lambda^2} \frac{\mathrm{Var}(Y)}{E(Y)^2}$$

i.e.

$$\Pr\left(\left|\frac{Y}{q^n} - \mu_1 \cdots \mu_m\right| \ge \lambda \mu_1 \cdots \mu_m\right) \le \frac{1}{\lambda^2} \frac{\varepsilon^2}{36}$$

we obtain, by choosing $\lambda = \varepsilon/3$, the bound

$$\Pr\left(\left(1 - \frac{\varepsilon}{3}\right) \mu_1 \cdots \mu_m \le q^{-n} Y \le \left(1 + \frac{\varepsilon}{3}\right) \mu_1 \cdots \mu_m\right) \ge \frac{3}{4}.$$

But from inequality (6) we obtain the bound

$$\begin{aligned}
&\left(1 - \frac{\varepsilon}{3m}\right)^m \rho_1 \cdots \rho_m \le \mu_1 \cdots \mu_m \le \left(1 + \frac{\varepsilon}{3m}\right)^m \rho_1 \cdots \rho_m \\
\Rightarrow\; &\left(1 - \frac{\varepsilon}{2}\right) \rho_1 \cdots \rho_m \le \mu_1 \cdots \mu_m \le \left(1 + \frac{\varepsilon}{2}\right) \rho_1 \cdots \rho_m
\end{aligned}$$

Putting these two bounds together yields the desired fpras condition:

$$\Pr\left((1-\varepsilon) \underbrace{q^n \rho_1 \cdots \rho_m}_{|\Omega(G)|} \le Y \le (1+\varepsilon) \underbrace{q^n \rho_1 \cdots \rho_m}_{|\Omega(G)|}\right) \ge \frac{3}{4}.$$

10 Markov Chain Monte Carlo Simulations

This is a very broad area and would actually merit a full main section of its own. Maybe later.

In many practical applications of Markov chains, one is interested not just in sampling according to the stationary distribution π, but also in computing expected values of various quantities with respect to it:

$$E_\pi[f] = \sum_{\sigma \in S} f(\sigma) \pi_\sigma \quad \text{(also denoted } \langle f \rangle_\pi)$$

E.g. one might want to compute the average magnetisation of a spin glass model at a given inverse temperature β (cf. page 62):

$$\langle M \rangle = \sum_{\sigma \in S} M(\sigma) \cdot \underbrace{e^{-\beta H(\sigma)}/Z_\beta}_{\text{Gibbs density}}$$

The task could be approached by producing many independent sample states σ according to π, computing $f(\sigma)$ for each and controlling the estimation error.

However, it is customary to compute the estimates from a single (or a few) long runs of the chain:

$$E_\pi[f] \approx \frac{1}{N} \sum_{k=1}^{N} f(X_k(\omega)), \quad N \text{ large}$$

(More precisely, maybe

$$E_\pi[f] \approx \frac{1}{N - N_0} \sum_{k=N_0+1}^{N} f(X_k(\omega)),$$

where N_0 is an initial "burn-in" time to eliminate systematic effects of choice of the initial state.)

For this approach to work properly, the Markov chains must be "path-ergodic" in the sense that the stationary distribution is sampled properly along almost every individual path of the chain.

In fact, if the word was not already so overused, we could define a Markov chain $\mathcal{M} = (X_1, X_2, \ldots)$ to be *ergodic with stationary distribution* π if for any initial distribution μ and for all states $\sigma \in S$:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} I_\sigma(X_k) = \pi_\sigma \quad \mu\text{-almost surely},$$

i.e.

$$\Pr_\mu\left(\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} I_\sigma(X_k(\omega)) \neq \pi_\sigma\right) = 0,$$

where I_σ is an indicator function for state σ:

$$I_\sigma(\xi) = \begin{cases} 1, & \text{if } \xi = \sigma \\ 0, & \text{if } \xi \neq \sigma \end{cases}$$

Luckily, all regular (finite) Markov chains are ergodic also in this strong sense. In fact, even more is true:

Theorem 10.1 (Ergodic Theorem for Regular Markov Chains)
Let $\mathcal{M} = (X_1, X_2, \dots)$ be a regular Markov chain with state space S, and $f : S \to \mathbb{R}$ any function. Then for any initial distribution μ:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} f(X_k) = E_\pi[f] \quad \mu\text{-almost surely.}$$

We do not have all the tools (or the time) to give a complete proof of Theorem 10.1, but here are the key components:

Theorem 10.2 (Kolmogorov's Strong Law of Large Numbers)
Let $X_1, X_2, \dots$ be a sequence of independent identically distributed random variables defined on probability space $(\Omega, \mathcal{F}, P)$, and such that $E[|X_k|] = E[|X_1|] < \infty$ for all k. Then

$$\lim_{N\to\infty}\frac{1}{N}(X_1 + \dots + X_N) = E[X_1] \quad P\text{-almost surely.}$$

Lemma 10.3 (Regenerative Cycle Lemma / Strong Markov Property)
Let $\mathcal{M} = (X_0, X_1, \dots)$ be a regular finite Markov chain with state space S. Fix any state $0 \in S$. Then 0 is visited on any given sample path of $\mathcal{M}$ infinitely often (almost surely), and denoting $\tau_0, \tau_1, \tau_2, \dots$ the successive times of visit to 0, the sample path segments

$$\{X_{\tau_k}, X_{\tau_k+1}, \dots, X_{\tau_{k+1}-1}\}, \quad k \geq 0,$$

are independent and identically distributed.

Proof of Theorem 10.1: Recall that for any $\sigma \in S$:

$$\pi_\sigma = \frac{\rho_\sigma}{\mu_0} = \frac{1}{\mu_0}\cdot E_0\left[\sum_{n\geq 1} I_{[X_n=\sigma]} I_{[\tau_1 > n]}\right] = \frac{1}{\mu_0} E_0\left[\sum_{n=1}^{\tau_1} I_{[X_n=\sigma]}\right],$$

where $E_0[\cdot] = E[\cdot | X_0 = 0]$, τ_1 is the time of first return to 0, and $\mu_0 = E[\tau_1]$.

Given a sample path starting at state 0, let $\tau_1, \tau_2, \dots$ be the successive return times to 0, and define

$$U_p = \sum_{n=\tau_p+1}^{\tau_{p+1}} f(X_n).$$

By Lemma 10.3, the U_p's are independent and identically distributed random variables. Assuming $f \geq 0$ we obtain:

$$\begin{aligned} E[U_0] &= E_0\left[\sum_{n=1}^{\tau_1} f(X_n)\right] \\ &= E_0\left[\sum_{n=1}^{\tau_1}\sum_{\sigma\in S} f(\sigma) I_{[X_n=\sigma]}\right] \\ &= \sum_{\sigma\in S} f(\sigma) E_0\left[\sum_{n=1}^{\tau_1} I_{[X_n=\sigma]}\right] \\ &= \mu_0 \sum_{\sigma\in S} f(\sigma)\pi_\sigma = \mu_0 E_\pi[f] \end{aligned}$$

By Theorem 10.2 (Strong Law of Large Numbers), then:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{p=1}^{n} U_p = E[U_0] = \mu_0 E_\pi[f] \quad \eta\text{-almost surely,}$$

i.e.

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=\tau_1+1}^{\tau_{n+1}} f(X_k) = \mu_0 E_\pi[f] \quad \eta\text{-almost surely.} \tag{7}$$

Define then random variables $\nu(n)$ as:

$$\nu(n) = \sum_{k=1}^{n} I_{[X_k=0]}$$

($\sim$ number of returns to 0 by time n). Clearly $\tau_{\nu(n)} \le n < \tau_{\nu(n)+1}$ for all n, so that

$$\frac{1}{\nu(n)}\sum_{k=1}^{\tau_{\nu(n)}} f(X_k) \le \frac{1}{\nu(n)}\sum_{k=1}^{n} f(X_k) < \frac{1}{\nu(n)}\sum_{k=1}^{\tau_{\nu(n)+1}} f(X_k) \quad \text{almost surely.}$$

Since by Lemma 10.3, $\nu(n) \to \infty$ as $n \to \infty$, we obtain from equation (7):

$$\lim_{n\to\infty} \frac{1}{\nu(n)}\sum_{k=1}^{n} f(X_k) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{\tau_{n+1}} f(X_k) = \mu_0 E_\pi[f] \quad \text{almost surely.} \qquad (8)$$

However, asymptotically also

$$n \sim \tau_{\nu(n)} = \sum_{i=0}^{\nu(n)-1} (\tau_{i+1} - \tau_i) \quad \text{almost surely,}$$

so by Lemma 10.3 and Theorem 10.2:

$$\frac{n}{\nu(n)} \sim \frac{1}{\nu(n)}\sum_{i=0}^{\nu(n)-1} (\tau_{i+1} - \tau_i) = E[\tau_1] = \mu_0 \quad \text{almost surely.}$$

Thus $\mu_0 \nu(n) \sim n$, and by combining equations (7) and (8):

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_k) &= \lim_{n\to\infty} \frac{1}{\mu_0 \nu(n)}\sum_{k=1}^{n} f(X_k) \quad \text{almost surely} \\ &= E_\pi[f].\end{aligned}$$

The case of general $f : S \to \mathbb{R}$ can be handled by treating separately the nonnegative functions

$$f^+ = \max\{f, 0\} \text{ and } f^- = \max\{-f, 0\}$$

and summing up the resulting equalities. □

Convergence Rates of MCMC Simulation Algorithms

Let $\mathcal{M} = (X_0, X_1, \dots)$ be a regular finite Markov chain with state space $S = \{1, \dots, r\}$, transition probability matrix P, and stationary distribution π. Denote:

$$\Pi = \begin{bmatrix} \pi_1 & \cdots & \pi_r \\ \pi_1 & \cdots & \pi_r \\ & \vdots & \\ \pi_1 & \cdots & \pi_r \end{bmatrix} \quad \text{(i.e. for any distribution } \mu,\ \mu^T \Pi = \pi^T\text{)}.$$

The *fundamental matrix* of chain $\mathcal{M}$ is defined as

$$Z = (I - (P - \Pi))^{-1}.$$

Proposition 10.4 *For a regular chain $\mathcal{M}$, the fundamental matrix Z is well-defined, and*

$$Z = I + \sum_{n\ge 1} (P^n - \Pi).$$

Proof: It is easy to verify that for all $k \ge 1$:

$$P\Pi^k = \Pi^k P = \Pi.$$

Thus,

$$\begin{aligned}(P-\Pi)^n &= \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} P^k \Pi^{n-k} \\ &= P^n + \sum_{k=0}^{n-1} \binom{n}{k} (-1)^{n-k} \Pi \\ &= P^n - \Pi.\end{aligned}$$

Therefore, with $A = P - \Pi$,

$$(I-A)(I + A + A^2 + \ldots + A^{n-1}) = I - A^n = I + P^n - \Pi,$$

and consequently

$$(I-A)(I + \sum_{n\ge 1} A^n) = \lim_{n\to\infty} (I + P^n - \Pi) = I.$$

Hence the matrix $I - A = I - (P - \Pi)$ is invertible, and

$$(I - (P-\Pi))^{-1} = I + \sum_{n\ge 1} (P-\Pi)^n = I + \sum_{n\ge 1} (P^n - \Pi).$$ □

The fundamental matrix has many uses (analogous to the fundamental matrix of transient states) in computing expected recurrence times etc.

We, however, quote only the one of main interest to us (and even that without its somewhat technical proof). Given a Markov chain $\mathcal{M}$ with finite state space S, and any functions $f, g : S \to \mathbb{R}$, denote:

$$\langle f, g \rangle_\pi = E_\pi[f(X)g(X)] = \sum_{i\in S} \pi(i) f(i) g(i)$$

$$\mathrm{Var}_\mu(f) = E_\mu[(f(X) - \overline{f})^2] = E_\mu[f(X)^2] - (\underbrace{E_\mu[f(X)]}_{\overline{f}})^2$$

Theorem 10.5 (Asymptotic variance of Ergodic Estimates)
For a regular chain $\mathcal{M}$, and any function $f : S \to \mathbb{R}$,

$$\lim_{N\to\infty} \frac{1}{N} Var_\mu \left(\sum_{k=1}^{N} f(X_k) \right) = \underbrace{2\langle f, Zf\rangle_\pi - \langle f, (I+\Pi) f\rangle_\pi}_{\text{Denote } v(f,P,\pi)}$$

for any initial distribution μ.

Proof: E.g. Brémaud 1999, pages 232-234. □

Since by Theorem 10.1,

$$\tilde{f}_N = \frac{1}{N} \sum_{k=1}^{N} f(X_k) \underset{\text{a.s.}}{\longrightarrow} \bar{f} = E_\pi[f],$$

by Chebyshev's inequality we see that for any $\delta > 0$ and for "large N":

$$\Pr(|\tilde{f}_N - \bar{f}| \geq \delta) \leq \frac{1}{\delta^2} \mathrm{Var}(\tilde{f}_N) = \frac{1}{\delta^2 N^2} \mathrm{Var}\left(\sum_{k=1}^{N} f(X_k) \right) \approx \frac{v(f,P,\pi)}{\delta^2 N}$$

independent of the initial distribution μ.

Suppose then that the transition probability matrix P has r distinct eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r > -1$, with associated left and right eigenvectors $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$, respectively (normalized so that $u_i^T v_i = 1 \quad \forall\, i$). Then:[2]

$$P^n = \sum_{i=1}^{r} \lambda_i^n v_i u_i^T = \Pi + \sum_{i=2}^{r} \lambda_i^n v_i u_i^T,$$

and so

$$Z = I + \sum_{n\geq 1} (P^n - \Pi) = I + \sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} v_i u_i^T.$$

Thus

$$\begin{aligned} v(f,P,\pi) &= 2\langle f, Zf\rangle_\pi - \langle f, (I+\Pi) f\rangle_\pi \\ &= 2\langle f, f\rangle_\pi + 2\sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} \langle f, v_i\rangle_\pi (u_i^T f) - \langle f, f\rangle_\pi - \langle f, \Pi f\rangle_\pi \\ &= \underbrace{\langle f, (I-\Pi) f\rangle_\pi}_{\mathrm{Var}_\pi(f(X_0))} + 2\sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} \langle f, v_i\rangle_\pi (f^T u_i). \end{aligned}$$

[2] Cf. page 16. Also left eigenvectors are here represented as column vectors, however.

For a reversible chain ($D^{1/2} P D^{-1/2}$ symmetric), $u_i = D v_i$ and therefore $f^T u_i = \langle f, v_i\rangle_\pi$. Applying the decomposition $f = \sum_i \langle f, v_i\rangle_\pi v_i$ we obtain in this case

$$v(f,P,\pi) = \sum_{i=2}^{r} \frac{1+\lambda_i}{1-\lambda_i} |\langle f, v_i\rangle_\pi|^2.$$

Let us then consider the task of designing good "Metropolis-like" reversible Markov chains with given stationary distribution π and as good convergence rate as possible.

To achieve a given stationary distribution π, the detailed balance conditions require only that

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{for all states } i, j \in S \tag{9}$$

There are potentially an infinite number of transition matrices P satisfying conditions (9). Let us focus on solutions of the form

$$p_{ij} = q_{ij} \alpha_{ij},$$

where $Q = (q_{ij})$ is an irreducible *candidate-generation matrix*, and $\alpha_{ij} \in (0,1]$ are the *acceptance probabilities* for given tentative state transitions.

W. Hastings (1970) proposed the following general class of acceptance probability matrices guaranteeing the validity of the detailed balance conditions (9):

$$\alpha_{ij} = \frac{s_{ij}}{1+t_{ij}},$$

where

$$t_{ij} = \frac{\pi_i q_{ij}}{\pi_j q_{ji}}.$$

and $s_{ij} = s_{ji}$ are numbers chosen so that $\alpha_{ij} \in (0,1]$, i.e.

$$0 < s_{ij} \leq 1 + \min\{t_{ij}, t_{ji}\} \quad \forall\, i, j. \tag{10}$$

Enforcing equality in condition (10) results in the Metropolis-Hastings algorithm

$$\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

(check this!), whereas always choosing $s_{ij} = 1$ defines the so called *Barker's algorithm*:

$$\alpha_{ij} = \frac{\pi_j q_{ij}}{\pi_j q_{ji} + \pi_i q_{ij}}.$$

Let us then compare the various Hastings-type MCMC algorithms with respect to their asymptotic variance (Theorem 10.5). We quote the following result without proving it:

Theorem 10.6
Let $P = (p_{ij})$ and $P' = (p'_{ij})$ be regular transition matrices over finite state space S, with the same stationary distribution π. If $p_{ij} \geq p'_{ij}$ for all $i \neq j$, then

$$v(f, P, \pi) \leq v(f, P', \pi)$$

holds for all functions $f : S \to \mathbb{R}$.

Proof: E.g. Brémaud page 300. □

Corollary 10.7
For a given candidate-generation matrix Q, the Metropolis-Hastings algorithm has optimal asymptotic variance in the class of Hastings algorithms.

Proof: Since the α_{ij} are probabilities, the upper bound on s_{ij} given in condition (10) cannot be exceeded. The Metropolis-Hastings algorithm matches the upper bound. □

11 Genetic Algorithms

Genetic algorithms (GA) are a general-purpose "black-box" optimisation method proposed by J. Holland (1975) and K. DeJong (1975).

The method has attracted lots of interest, but its theory is still incomplete and the empirical results somewhat inconclusive. Advantages of the technique are that it is general-purpose, parallelisable, and adapts incrementally to changing cost functions ("on-line optimisation"). Disadvantages, on the other hand include that GA's are typically very slow – thus the technique should be used with moderation for simple serial optimisation of a stable, easily evaluated cost function.

Some claim that GA's typically require fewer function evaluations to reach comparable results as e.g. simulated annealing. Thus the method may be good when function evaluations are expensive (e.g. require some acutal physical measurement).

11.1 The Basic Algorithm

We consider the so called "simple genetic algorithm"; also many other variations exist.

Assume we wish to maximise a cost function c defined on n-bit binary strings:

$$c : \{0,1\}^n \to \mathbb{R}.$$

Other types of domains must be encoded into binary strings, which is a nontrivial problem. View each of the candidate solutions $s \in \{0,1\}^n$ as an *individual* or *chromosome*. At each stage (*generation*) t the algorithm maintains a *population* of individuals $p_t = (s_1, \ldots, s_m)$.

There are three operations defined on populations:

- *selection* $\sigma(p)$ ("survival of the fittest")
- *recombination* $\rho(p)$ ("mating", "crossover")
- *mutation* $\mu(p)$

The *Simple Genetic Algorithm* is then as follows:

function SGA(σ, ρ, μ):
    $p \leftarrow$ random initial population;
    while p "not converged" **do**
        $p' \leftarrow \sigma(p)$;
        $p'' \leftarrow \rho(p')$;
        $p \leftarrow \mu(p'')$
    end while;
    return p (or "fittest individual" in p).
end.

Selection

Denote $\Omega = \{0,1\}^n$. The selection operator $\sigma : \Omega^m \to \Omega^m$ maps populations probabilistically: given an individual $s \in p$, the expected number of copies of s in $\sigma(p)$ is proportional to the *fitness* of s *in* p. This is a function of the cost of s compared to the costs of other $s' \in p$.

Some possible fitness functions are:

- Relative *cost* ($\Rightarrow$ "canonical GA"):

$$f(s) = \frac{c(s)}{\frac{1}{m}\sum_{s' \in p} c(s')} \triangleq \frac{c(s)}{\bar{c}}.$$

- Relative *rank* :

$$f(s) = \frac{r(s)}{\frac{1}{m}\sum_{s' \in p} r(s')} = \frac{2}{m+1} \cdot r(s),$$

  where $r(s)$ is the rank of individual s in a worst-to-best ordering of all $s' \in p$.

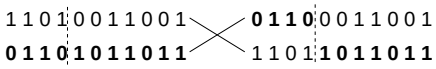
Once the fitness of individuals has been evaluated, selection can be performed in different ways:

- *Roulette-wheel selection* ("stochastic sampling with replacement"):
  - Assign to each individual $s \in p$ a probability to be selected in proportion to its fitness value $f(s)$. Select m individuals according to this distribution.
  - Pictorially: Divide a roulette wheel into m sectors of width proportional to $f(s_1), \ldots, f(s_m)$. Spin the wheel m times.
- *Remainder stochastic sampling*:
  - For each $s \in p$, select deterministically as many copies of s as indicated by the integer part of $f(s)$. After this, perform stochastic sampling on the fractional parts of the $f(s)$.
  - Pictorially: Divide a fixed disk into m sectors of width proportional to $f(s_1), \ldots, f(s_m)$. Place an outer wheel around the disk, with m equally-spaced pointers. Spin the outer wheel once.

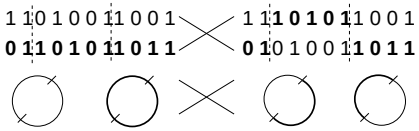
Recombination

Given a population p, choose two random individuals $s, s' \in p$. With probability p_ρ, apply a *crossover operator* $\rho(s, s')$ to produce two new offspring individuals t, t' that replace s, s' in the population. Repeat the operation $m/2$ times, so that on average each individual participates once. Denote the total effect on the population as $p' = \rho(p)$. (A practical implementation: choose $\frac{p_\rho}{2} \cdot m$ random pairs from p and apply crossover deterministically.) Typically $p_\rho \approx 0.7 \ldots 0.9$.

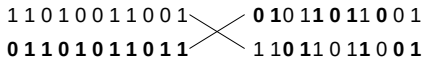
Some possible crossover operators are illustrated in Figure 1.



(a) 1-point



(b) 2-point



(c) uniform

Figure 1: Typical crossover operators.

Mutation

Given population p, consider each bit of each individual and flip it with some small probability p_μ. Denote the total effect on the population as $p' = \mu(p)$. Typically, $p_\mu \approx 0.001 \ldots 0.01$. It appears that for n-bit strings a good choice is $p_\mu = 1/n$.

Theoretically mutation is disruptive. Recombination and selection should take care of optimisation; mutation is needed only to (re)introduce "lost alleles", alternative values for bits that have the bits that have the same value in all current individuals.

In practice mutation plus selection equals local search. Mutation, even with quite high values of p_μ, can be efficient and is often more important than recombination.

Analysis of GA's: Hyperplane sampling

The notion of hyperplane sampling presents a heuristic view of how a genetic algorithm works.

A *hyperplane* (actually subcube) is a subset of $\Omega = \{0,1\}^n$, where the values of

some bits are fixed and other are free to vary. A hyperplane may be represented by a *schema* $H \in \{0,1,*\}^n$. E.g. the schema '$0*1**$' represents the 3-dimensional hyperplane (subcube) of $\{0,1\}^5$ where bit 1 is fixed to 0, bit 3 is fixed to 1, and bits 2, 4, and 5 vary.

An individual $s \in \{0,1\}^n$ *samples* hyperplane H, or *matches* the corresponding schema if the fixed bits of H match the corresponding bits in s. BY some abuse of notation, this situation is denoted as "$s \in H$". Note that a given individual generally samples many hyperplanes simultaneously, e.g. individual '101' samples '$10*$', '$1*1$', etc.

Define the *order* of a hyperplane H as:

$$\begin{aligned} o(H) &= \text{number of fixed bits in } H \\ &= n - \dim H. \end{aligned}$$

The *average cost* of hyperplane H is then:

$$c(H) = \frac{1}{2^{n-o(H)}} \sum_{s \in H} c(s).$$

Denoting by $m(H,p)$ the number of individuals in population p that sample hyperplane H, the *average fitness* of hyperplane H in population p is defined as:

$$f(H,p) = \frac{1}{m(H,p)} \sum_{s \in H \cap p} f(s,p)$$

A heuristic claim is then that selection drives the search towards hyperplanes of higher average cost (quality).

Consider e.g. the cost function and partition of Ω into hyperplanes (in this case, intervals) of order 3 presented in Figure 2. Here the current population of 21 individuals samples the hyperplanes so that e.g. '$000**$' and '$010**$' are sampled by three individuals each, and '$100**$' and '$101**$' by two individuals each. Hyperplane '$010**$' has a rather low average fitness in this population, whereas '$111**$' has a rather high average fitness.

The result of e.g. roulette wheel selection on this population might lead to elimination of some individuals and duplication of others, as presented in Figure 3.

Then, in terms of expected values, one can show that

$$E[m(H,\sigma(p))] = m(H,p) \cdot f(H,p).$$

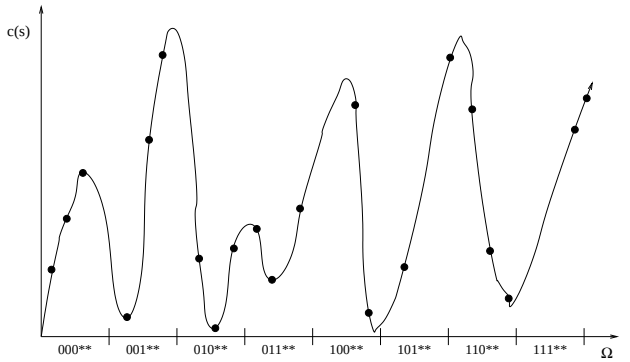


Figure 2: A population sampling hyperplanes.

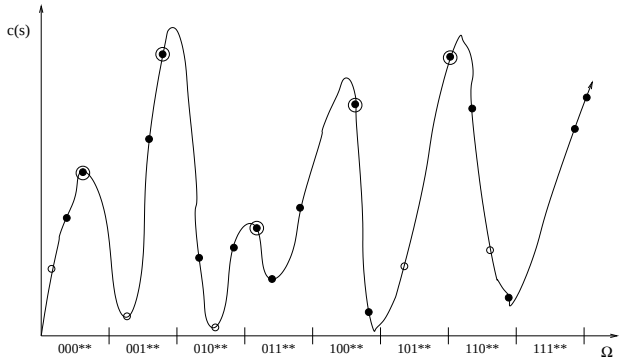


Figure 3: A sampling population after selection.

The effect of crossover on schemata

Consider a schema such as

$$H = **\underbrace{11**01*1}_{\Delta(H)=7}**$$

and assume that it is represented in the current population by some $s \in H$.

If s participates in a crossover operation and the crossover point is located between bit positions 3 and 10, then with large probability the offspring are no longer in H. In this case schema H is said to be *disrupted*. On the other hand, if the crossover point is elsewhere, then one of the offspring stays in H, and H is *retained*.

Generally, the probability that in 1-point crossover a schema $H = \{0,1,*\}^n$ is retained, is (ignoring the possibility of "lucky combinations")

$$\Pr(\text{retain } H) \approx 1 - \frac{\Delta(H)}{n-1},$$

where $\Delta(H)$ is the *defining length* of H, i.e. the distance between the first and last fixed bit in H.

More precisely, if H has $m(H,p)$ representatives in population p of total size m:

$$\Pr(\text{retain } H) \geq 1 - \frac{\Delta(H)}{n-1}\left(1 - \frac{m(H,p)}{m}\right).$$

The Schema "Theorem"

The Schema Theorem, proposed by J. Holland (1975), provides a heuristic estimate of the changes in representation of a given schema H from one generation to the next.

Denote:

$$m(H,t) = \text{number of individuals in population at generation } t \text{ that sample } H.$$

Then:

(i) Effect of selection:

$$m(H,t') \approx m(H,t) \cdot f(H)$$

(ii) Effect of recombination:

$$\begin{aligned} m(H,t'') &\approx (1-p_\rho)m(H,t') + p_\rho\left(m(H,t')\Pr(\text{retain } H) + \underbrace{m\cdot\Pr(\text{luck})}_{\geq 0}\right) \\ &\geq (1-p_\rho)m(H,t') + p_\rho m(H,t')\left(1 - \frac{\Delta(H)}{n-1}\left(1 - \frac{m(H,t')}{m}\right)\right) \\ &= m(H,t')\left(1 - p_\rho\frac{\Delta(H)}{n-1}\left(1 - \frac{m(H,t')}{m}\right)\right) \end{aligned}$$

(iii) Effect of mutation:

$$m(H,t+1) \approx m(H,t'')\cdot(1-p_\mu)^{o(H)}$$

In summary, then:

$$m(H,t+1) \gtrsim m(H,t)\cdot f(H)\cdot\left(1 - p_\rho\frac{\Delta(H)}{n-1}\left(1 - \frac{m(H,t')}{m}\right)\right)\cdot(1-p_\mu)^{o(H)}.$$

The formula leads to so called "*Building Block Hypothesis*": In a genetic search, short, above-average fitness schemata of low order ("building blocks") receive an exponentially increasing representation in the population.

The following criticisms have been expressed as regards the "Schema Theorem" and the Building Block Hypothesis, however:

- Many of the approximations used in deriving the "Schema Theorem" implicitly assume that the population is very large. In particular, it is assumed that all the relevant schemata are well sampled. This is clearly not possible in practice, because there are 3^n possible schemata of length n.
- The result cannot be used to predict the development of the population for much more than one generation, because:
  - firstly, the long-term development depends on the coevolution of the schemata, and the "theorem" considers only one schema in isolation;
  - secondly, an "exponential growth" cannot in any case continue for long in a finite population.

11.2 Genetic Algorithms as Stochastic Processes

A proper mathematical treatment of GA's would view them as stochastic processes. It is unfortunately very difficult to obtain any nontrivial analytical results in this direction. Here we outline a simple Markov chain model presented by Vose & Liepins (1991) and Rudolph (1994).

Consider the "canonical GA", i.e. the Simple Genetic Algorithm using the relative cost fitness function and standard proportional ("roulette-wheel") selection, in the form:

$p \leftarrow$ random initial population;
$p \leftarrow \sigma(p)$; (selection)
while p "not converged" **do**
  $p' \leftarrow \rho(p)$; (recombination)
  $p'' \leftarrow \mu(p')$ (mutation)
  $p \leftarrow \sigma(p'')$; (selection)
end while.

Encode a population of m individuals, each an n-bit string, as an integer (in binary representation)

$$p \in \{0,1\}^{mn} \equiv \underbrace{\{0,1,\ldots,2^{mn}-1\}}_{\mathbb{Z}_{2^{mn}}}.$$

Then the CGA can be modeled as a Markov chain on state space $\mathbb{Z}_{2^{mn}}$, with the transitions probability matrix $P = CMS$, where

C is the recombination ("crossover") transition probability matrix
M is the mutation transition probability matrix
S is the selection transition probability matrix

A stochastic matrix $P = (p_{ij})$ is:

(i) *positive*, if $p_{ij} > 0$ for all i, j;

(ii) *primitive*, if P^k is positive for some $k \geq 0$;

(iii) *reducible*, if it can be converted to the form

$$\tilde{P} = \begin{bmatrix} C & 0 \\ R & T \end{bmatrix},$$

where C and T are square matrices, by applying the same permutation to the rows and the columns;

(iv) *irreducible*, if it is not reducible.

The interpretation of these definitions is that primitive matrices correspond to the irreducible and aperiodic Markov chains defined before. In a reducible matrix, the upper rows correspond to a "closed" or "absorbing" class of states, the lower rows to "transient" states. Note that a positive matrix is trivially primitive.

Theorem 11.1
Let P be a primitive stochastic matrix. Then the sequence P^k converges as $k \to \infty$ to a stochastic matrix P^∞ which has the form

$$P^\infty = \begin{bmatrix} p^\infty \\ \vdots \\ p^\infty \end{bmatrix},$$

where p^∞ is a stochastic vector with all components positive. (The vector p^∞ represents the stationary distribution of the chain.)

Theorem 11.2
Let P be a reducible stochastic matrix of the form

$$P = \begin{bmatrix} C & 0 \\ R & T \end{bmatrix},$$

where C is primitive, and T does not contain an irreducible submatrix. Then the sequence P^k converges as $k \to \infty$ to a stochastic matrix P^∞ of the form

$$P^\infty = \begin{bmatrix} p^\infty & 0 \\ \vdots & \vdots \\ p^\infty & 0 \end{bmatrix},$$

where p^∞ is a stochastic vector with all components positive.

Lemma 11.3
The transition probability matrix $P = CMS$ of the "canonical genetic algorithm", with mutation probability $0 < p_\mu < 1$ is positive and hence primitive.

Proof: Denote $C = (c_{ik}), M = (m_{kl}), S = (s_{lj})$. Then $P = (p_{ij})$, where

$$p_{ij} = \sum_{kl} c_{ik} m_{kl} s_{lj}.$$

Observe:

(i) $\forall\, i \exists\, k_i : c_{ik_i} > 0$ (Because C is stochastic $\Rightarrow \forall\, i : \sum_k c_{ik} = 1$)

(ii) M is positive: denote $N = mn$, $d(k,l)$ = Hamming distance between populations k, l. Then:

$$m_{kl} = p_\mu^{d(k,l)} \cdot (1-p_\mu)^{N-d(k,l)} > 0.$$

(iii) $\forall\, j : s_{jj} > 0$ (Because with nonzero probability, selection does not change the population.)

Thus:

$$p_{ij} = \sum_{kl} c_{ik} m_{kl} s_{lj} \geq c_{ik_i} m_{k_i j} s_{jj} > 0. \;\square$$

Theorem 11.4
The CGA with mutation probability $0 < p_\mu < 1$ converges to a stationary distribution of populations where the probability of every population is nonzero.

Proof: Follows from Theorem 11.1 and Lemma 11.3. □

Assume the CGA is defined so as to maximize the function $c : \{0,1\} \to \mathbb{R}$. Denote

$$c^* = \max\{c(i) \mid i \in \{0,1\}^n\},$$

and for a population $\hat{i} = (i_1, \ldots, i_m)$:

$$c^*(\hat{i}) = \max\{c(i_k) \mid k = 1, \ldots, m\}.$$

Denote by $\hat{i}^{(t)}$ the population of the CGA at time t. The algorithm *converges to global optimum* if

$$\lim_{t\to\infty} \Pr(c^*(\hat{i}^{(t)}) = c^*) = 1.$$

Note that the simulated annealing algorithm converges to global optimum in exactly this sense.

Corollary 11.5
If nonoptimal solutions with respect to the cost function c exist (i.e. if $c(j) < c^*$ for some $j \in \{0,1\}^n$), then the CGA does not converge to the global optimum.

Proof: Let $\hat{j} = (j, j, \ldots, j)$ be a population such that $c^*(\hat{j}) < c^*$ By Theorem 11.2,

$$\lim_{t\to\infty} \Pr(\hat{i}^{(t)} = \hat{j}) = \varepsilon > 0,$$

and thus

$$\lim_{t\to\infty} \Pr(c^*(\hat{i}^{(t)}) = c^*) \leq 1 - \varepsilon < 1. \;\square$$

Theorem 11.6
On the other hand, if the best solution found is always kept in the population ("elitist" selection) and not mutated, then the CGA does converge to the global optimum.

Proof: Simple corollary to Theorem 11.2: the transition probability matrix P reduces in this case to the form

$$P = \begin{bmatrix} C & 0 \\ R & T \end{bmatrix},$$

where the upper rows correspond to the unique closed class of populations containing a globally optimal solution. □

Note that for practical purposes, such (non)convergence results are of course largely irrelevant. The important (but difficult) questions are:

- How fast does the CGA with elitist selection converge towards an optimal solution?
- Does the CGA without elitist selection converge to a population with mostly optimal solutions, and how fast?

12 Combinatorial Phase Transitions

12.1 Phenomena and Models

"Where the Really Hard Problems Are" (Cheeseman et al. 1991)

Many NP-complete problems can be solved in polynomial time "on average" or "with high probability" for reasonable-looking distributions of problem instances. E.g. Satisfiability in time $o(n^2)$ (Goldberg et al. 1982), Graph Colouring in time $o(n^2)$ (Grimmett & McDiarmid 1975, Turner 1984).

Where, then, are the (presumably) exponentially hard instances of these problems located? Could one tell ahead of time whether a given instance is likely to be hard?

Early studies of this issue done by: Yu & Anderson (1985), Hubermann & Hogg (1987), Cheeseman, Kanefsky & Taylor (1991), Mitchell, Selman & Levesque (1992), Kirkpatrick & Selman (1994), etc.

Hard Instances for 3-SAT

Mitchell, Selman & Levesque (AAAI 1992).

Experiments on the behaviour of the Davis-Putnam[-Logemann-Loveland] (DP[LL]) procedure on randomly generated 3-cnf Boolean formulas.

E.g. satisfiable 3-cnf formula

$$(x_1 \vee \overline{x_2} \vee x_3) \wedge (\overline{x_1} \vee x_2 \vee \overline{x_4})$$

The expressions in parenthesis are *clauses* and the x's are *literals*.

Distribution of test formulas:

- number of variables
- $m = \alpha n$ randomly generated clauses of 3 literals, $2 \leq \alpha \leq 8$

TThe Davis-Putnam[-Logemann-Loveland] (DP[LL]) method for testing the satisfiability of a set of clauses Σ on the variable set V:

1. If Σ is empty, return "satisfiable".
2. If Σ contains an empty clause, return "unsatisfiable".
3. If Σ contains a unit clause $c = x^{\pm}$, assign to x a value which satisfies c, simplify the remaining clauses correspondingly, and call DPLL recursively.
4. Otherwise select an unassigned $x \in V$, assign $x \leftarrow 1$, simplify Σ, and call DPLL recursively. If this call returns "satisfiable", then return "satisfiable"; else assign $x \leftarrow 0$, simplify Σ, and call DPLL recursively again.

For each set of 500 formulas, Mitcell et al. plotted the median number of DPLL calls required for solution.

The results of this experiment are illustrated in Figures 4 and 5. Discussion:

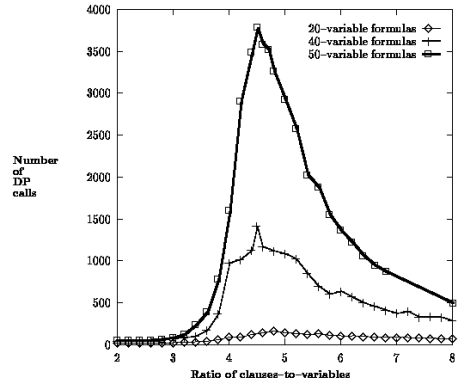


Figure 4: Number of DPLL calls required to determine satisfiability (Mitchell et al. 1992).

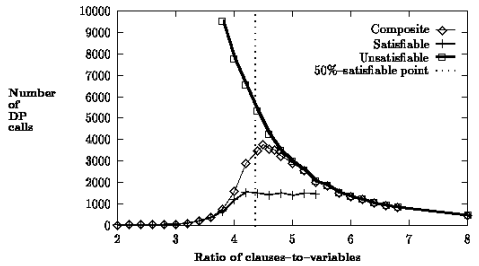


Figure 5: Number of required DPLL calls according to type of formula (Mitchell et al. 1992).

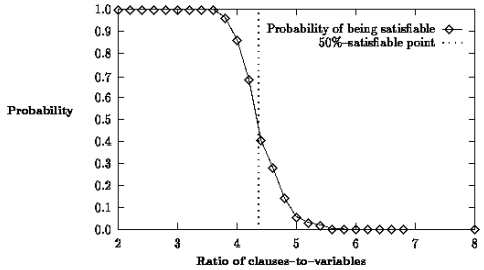


Figure 6: Probability of satisfiability for random 3-cnf formulas (Mitchell et al. 1992).

- A clear peak in running times (number of DPLL calls) near the point where 50% of formulas are satisfiable.
- The "50% satisfiable" point or "satisfiability threshold" seems to be located at roughly $\alpha \approx 4.25$ for large n.
- The peak seems to be caused by relatively short unsatisfiable formulas.

A fundamental question is whether the connection of the running time peak and the satisfiability threshold a characteristic of the DPLL algorithm, or a (more or less) algorithm independent "universal" feature?

The "50% satisfiable" point or "satisfiability threshold" for 3-SAT seems to be located at $\alpha \approx 4.25$ for large n.

12.2 Statistical Mechanics of k-SAT ("1st-Order Analysis")

Kirkpatrick & Selman (Science 1994)

Similar experiments as above for k-SAT, $k = 2, \ldots, 6$, 10000 formulas per data point. Results illustrated in Figure 7. Further observations:

- The "satisfiability threshold" α_c shifts quickly to larger values of α for increasing k.
- For fixed k, the value of α_c drifths slowly to smaller values for increasing n.

A statistical mechanics model of a k-cnf formula:

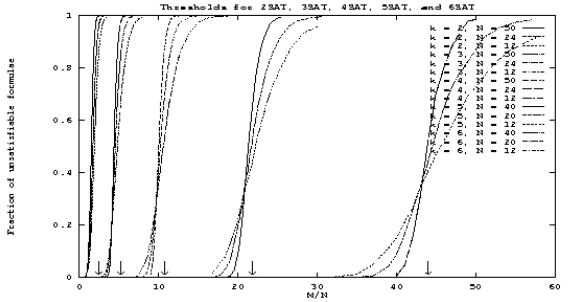


Figure 7: Probability of satisfiability for random k-cnf formulas (Kirkpatrick & Selman 1994).

- variables x_i $\sim$ spins with states ± 1
- clauses c $\sim$ k-wise interactions between spins
- truth assignment σ $\sim$ state of spin system
- Hamiltonian $H(\sigma)$ $\sim$ number of clauses unsatisfied by σ
- α_c $\sim$ critical "interaction density" point for "phase transition" from "satisfiable phase" to "unsatisfiable phase"

Estimates of α_c for various values of k via "annealing approximation", "replica theory", and observation:

k	α_{ann}	α_{rep}	α_{obs}
2	2.41	1.38	1.0
3	5.19	4.25	4.17 ± 0.03
4	10.74	9.58	9.75 ± 0.05
5	21.83	20.6	20.9 ± 0.1
6	44.01	42.8	43.2 ± 0.2

The "annealing approximation" means simply assuming that the different clauses are satisfied independently. This leads to the following estimate:

- The probability that a given clause c is satisfied by a random σ: $p_k = 1 - 2^{-k}$.

- The probability that a random σ satisfies all $m = \alpha n$ clauses assuming independence: $p_k^{\alpha n}$.
- Total number of satisfying assignments $= 2^n p_k^{\alpha n} \triangleq S_k^n(\alpha)$.
- For large n, $S_k^n(\alpha)$ falls rapidly from 2^n tp 0 near a critical value $\alpha = \alpha_c$. Where is α_c?
- One approach: solve for $S_k^n(\alpha) = 1$.

$$S_k^n(\alpha) = 1 \Leftrightarrow 2p_k^\alpha = 1$$
$$\Leftrightarrow \alpha = -\frac{1}{\log_2 p_k} = -\frac{\ln 2}{\ln(1-2^{-k})} \approx -\frac{\ln 2}{2^{-k}} = (\ln 2)\cdot 2^k.$$

It is in fact known that:

- A sharp satisfiability threshold α_c exists for all $k \geq 2$ (Friedgut 1999).
- For $k = 2$, $\alpha_c = 1$ (Goerdt 1982, Chvátal & Reed 1982). Note that 2-SAT $\in$ P.
- For $k = 3$, $3.14 < \alpha_c < 4.51$ (lower bound due to Achlioptas 2000, upper bound to Dubois et al. 1999).
- Current best empirical estimate for $k = 3$: $\alpha_c \approx 4.27$ (Braunstein et al. 2002).

12.3 Local Search Methods for 3-SAT

Local search methods (e.g. simulated annealing, genetic algorithms) can be used for finding (with high probability) satisfying truth assignments to randomly generated 3-cnf formulas in the satisfiable phase ($m/n = \alpha < \alpha_c$).

Consider first a general objective function $E = E(x)$ to be minimised. Then the basic local search scheme is:

- Start with some randomly chosen feasible solution $x = x_0$.
- If value of $E(x)$ is not "good enough", search for some "neighbour" x' of x that satisfies $E(x') \lesssim E(x)$. If such an x' is found, set $x \leftarrow x'$ and repeat.
- If no improving neighbour is found, then either restart at new random $x = x_0$ or relax the neighbourhood condition [algorithm-dependent].

In the setting of the 3-SAT problem, the objective function to be minimised is $E = E_F(s) =$ the number of unsatisfied clauses in formula F under truth assignment s. When $\alpha < \alpha_c$, an assignment s satisfying $E(s) = 0$ exists with high probability, and local search techniques are surprisingly powerful in finding such assignments.

The first systematically tested algorithm of this type was the following procedure GSAT by (Selman et al. 1992):

```
GSAT(F):
  s = initial truth assignment;
  while flips < max_flips do
    if s satisfies F then output s & halt, else:
    - find a variable x whose flipping causes
      largest decrease in E (if no decrease is
      possible, then smallest increase);
    - flip x.
```

An improvement to GSAT is to augment it with a fraction p of random walk moves, leading to algorithm NoisyGSAT (Selman et al. 1996):

```
NoisyGSAT(F,p):
  s = initial truth assignment;
  while flips < max_flips do
    if s satisfies F then output s & halt, else:
    - with probability p, pick a variable x
      uniformly at random and flip it;
    - with probability (1-p), do basic GSAT move:
      - find a variable x whose flipping causes
        largest decrease in E (if no decrease is
        possible, then smallest increase);
      - flip x.
```

A subtle but important change to NoisyGSAT is to *focus* the search on the presently unsatisfied clauses. This leads to the current "industry standard" WalkSAT algorithm (Selman et al. 1996):

```
WalkSAT(F,p):
  s = initial truth assignment;
  while flips < max_flips do
    if s satisfies F then output s & halt, else:
    - pick a random unsatisfied clause C in F;
    - if some variables in C can be flipped without
```

```
    breaking any presently satisfied clauses,
    then pick one such variable x at random; else:
  - with probability p, pick a variable x
    in C at random;
  - with probability (1-p), pick an x in C
    that breaks a minimal number of presently
    satisfied clauses;
  - flip x.
```

The focusing seems to be important: in the (somewhat unsystematic) experiments performed by Selman et al. (1996), WalkSAT outperforms NoisyGSAT by several orders of magnitude.

Also other local search techniques can be applied to the satisfiability problem. Good results have been obtained e.g. with the following Record-to-Record Travel (RRT) method first introduced in the context of the TSP problem (Dueck 1993):

```
RRT(E,d):
  s = initial feasible solution;
  s* = s; E* = E(s);
  while moves < max_moves do
    if s is a global min. of E then output s & halt,
    else:
      pick a random neighbour s' of s;
      if E(s') <= E* + d then let s = s';
      if E(s') < E* then:
        s* = s'; E* = E(s').
```

In applying RRT to SAT, one chooses again $E(s) =$ number of clauses unsatisfied by truth assignment s, together with single-variable flip neighbourhoods. Imposing the *focusing* heuristic of always selecting the flipped variables from unsatisfied clauses (precisely: one unsatisfied clause is chosen at random, and from there a variable at random) leads to the "focused RRT" (FRRT) algorithm for 3-SAT, which is quite competitive with WalkSAT (Seitz & Orponen 2003).

12.4 Statistical Mechanics of K-SAT ("Replica Analysis")

The analyses in this area are rather technical, so we present just some basic ideas.

Consider again the statistical mechanics model of k-SAT formulas discusses on p. 126. I.e. we consider the ensemble of random k-cnf formulas with n variables and

$m = \alpha n$ clauses. The Boolean-valued variables x_i are mapped to binary-state spins as $x_i \in \{\text{true}, \text{false}\} \mapsto \text{spin } S_i \in \{+1, -1\}$.

A formula consists of a set of clauses C_l represented in terms of an "interaction matrix" $C = (C_{li})$:

$$C_{li} = \begin{cases} +1, & \text{if } C_l \text{ includes } x_i \\ -1, & \text{if } C_l \text{ includes } \overline{x_i} \\ 0, & \text{otherwise} \end{cases}$$

Thus,

$$\sum_{i=1}^{n} C_{li} S_i = -K$$

if and only if all the literals in clause C_l are "wrong", i.e. the clause is unsatisfied by truth assignment (spin state) $S = (S_1, \ldots, S_n)$.

We consider the Hamiltonian function

$$E[S,C] = \sum_{l=1}^{m} \delta\left(\sum_{i=1}^{n} C_{li} S_i + K\right) \quad = \text{number of clauses in } C \text{ unsatisfied by } S,$$

$$\delta(u) = \begin{cases} 1, & \text{if } u = 0 \\ 0, & \text{otherwise} \end{cases}$$

The ground state potential (minimum number of unsatisfied clauses) of a given system C is $E^*[C] = \min_S E[S,C]$. For randomly generated C, $\Pr(E^*[C] = 0)$with high probability when α is small, and we would like to approximate the value $\alpha = \alpha_c(K)$ where this property ceases to hold.

This is however a very difficult problem, so we approach it indirectly by considering rather the average of $E^*[C]$ with respect to C, denoted $E_{GS} = \overline{E^*[C]}$. (Such averages with respect to system parameters are called "quenched averages", as opposed to the more usual "thermal averages" computed with respect to system states.)

For large n, the distribution of $E^*[C]$ is highly concentrated around $E_{GS} = E_{GS}(\alpha, K)$. ($E^*$ is said to be "self-averaging".) In particular:

$$E_{GS} \approx 0 \text{ in the sat. phase } (\alpha < \alpha_c(K)),$$
$$E_{GS} > 0 \text{ in the unsat. phase } (\alpha > \alpha_c(K)).$$

Thus, we use the behaviour of E_{GS} as a guide to determining the value of α_c.

It is known that

$$E_{GS} = -T\overline{\ln Z_T[C]} + o(T^2)$$

as $T \to 0$, where

$$Z_T[C] = \sum_S \exp(-E[S,C]/T).$$

(This follows by averaging from the fundamental thermodynamic formula $F = E - TS = -kT \ln Z$ (p. 60).)

The important, but complicated quantity $\overline{\ln Z}$ can be estimated using the so called "replica method".

Consider the Taylor expansion of Z^ν as a function of ν for small ν:

$$Z^\nu = e^{\nu \ln Z} = 1 + \nu \ln Z + o(\nu^2)$$

Thus, for a fixed $Z > 0$:

$$\ln Z = \lim_{\nu \to 0} \frac{Z^\nu - 1}{\nu}.$$

Applying this to $\ln Z_T[C]$ and averaging over C yields:

$$E_{GS} = -T \lim_{\nu \to 0} \frac{1}{\nu} \left(\overline{Z_T[C]^\nu} - 1 \right) + o(T^2) \tag{11}$$

as $T \to 0$.

Now assume that the "small ν" is in fact an integer. Then:

$$\begin{aligned} \overline{Z_T[C]^\nu} &= \overline{\left(\sum_S \exp(-E[S,C]/T) \right)^\nu} \\ &= \sum_{S_1} \dots, \sum_{S_\nu} \overline{\exp\left(-\sum_{r=1}^{\nu} E[S_r,C]/T \right)} \end{aligned}$$

Thus we have transformed the problem of computing $\overline{Z_T^\nu}$ to the consideration of ν interconnected "replicas" of the original system.

This modified structure can further be viewed as a single system consisting of n vector-valued spins $\vec{\sigma}_i \in \{+1,-1\}^\nu, i = 1, \dots, n$, with (non-random) potential function

$$E_{eff}[\vec{\sigma}_1, \dots, \vec{\sigma}_n] = -T \ln \left[\overline{\exp\left(-\sum_{r=1}^{\nu} E[S_r,C]/T \right)} \right].$$

One can easily check that with this choice:

$$\overline{Z_T^\nu} = Z_T^{eff} = \sum_{\{\vec{\sigma}_i\}} \exp(-E_{eff}[\{\vec{\sigma}\}]/T).$$

This partition function may in some cases be so concentrated that for large n:

$$\overline{Z_T^\nu} = Z_T^{eff} \approx e^{-n\tilde{f}_T(\nu)} \approx 1 - n\tilde{f}_T(\nu),$$

where $\tilde{f}_T(\nu)$ is some nonlinear function with $\tilde{f}_T(0) = 0$.

Plugging this estimate in formula (11) yields

$$E_{GS} \approx -T \lim_{\nu \to 0} \frac{-n\tilde{f}_T(\nu)}{\nu} = Tn\tilde{f}_T'(0).$$

The replica method has been partially mathematically vindicated, i.e. the requisite "analytic continuation" from integer to real ν is justified under some conditions, although not generally.

From an application point of view, approximating the function $\tilde{f}_T(\nu)$ is the difficult part of the technique.

Index