Introduction to Bayesian Statistics - 8

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Our next important topic: Bayesian estimates often require complex numerical integrals. How do we confront this problem?

⇨ enter the Monte Carlo methods!

1. acceptance-rejection sampling
2. importance sampling
3. statistical bootstrap
4. Bayesian methods in a sampling-resampling perspective
5. Introduction to Markov chains and to Random Walks (RW)
6. **Detailed balance and Boltzmann's H-theorem**
7. **The Gibbs sampler**
8. **Simulated annealing and the Traveling Salesman Problem (TSP)**
9. **The Metropolis algorithm**
10. Markov Chain Monte Carlo (MCMC)
11. Affine-invariant MCMC algorithms (EMCEE)

6. Detailed balance and Boltzmann's H-theorem

From the definition of conditional probabilities we find

$$\begin{aligned} P[S(n) = S_i \text{ and } S(n+1) = S_j] &= P[S(n) = S_i | S(n+1) = S_j] P[S(n+1) = S_j] \\ &= P[S(n+1) = S_j | S(n) = S_i] P[S(n) = S_i] \end{aligned}$$

therefore, when a Markov chain is time reversed we find

$$\begin{aligned} &P[S(n) = S_i | S(n+1) = S_j] \\ &\qquad = P[S(n+1) = S_j | S(n) = S_i] \frac{P[S(n) = S_i]}{P[S(n+1) = S_j]} \end{aligned}$$

i.e.,

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^{(n)}}{\pi_j^{(n+1)}}$$

which shows that the reversed chain is time-dependent.

However, if states are distributed according to the invariant distribution, we have

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^*}{\pi_j^*}$$

which means that the backward transition probabilities are again time-independent, and in particular they must coincide with the forward transition probabilities, i.e.,

$$p_{ji}\pi_j^* = p_{ij}\pi_i^*$$

a condition which is called *detailed balance*.

So, *if* stationary distribution *then* detailed balance ... however the reverse also holds

$$\pi_j^{(n+1)} = \sum_i \pi_i^{(n)} p_{ij} = \sum_i \pi_j^{(n)} p_{ji} = \pi_j^{(n)} \sum_i p_{ji} = \pi_j^{(n)}$$

i.e., *a distribution is stationary if and only if it satisfies the condition of detailed balance*

Physical aside: continuous-time Markov processes

The time-dependence of the reversed chain is a manifestation of the dissipative character of the chain. Another important related result is the validity of the H-theorem for Markov processes.

In the case of continuous-time processes we can write

$$P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = \\ = P\left(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) P\left(S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right)$$

Memoryless processes

$$P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = P\left(S_{i_k}, t_k\right)$$

Markov processes

$$P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = P\left(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}\right) P\left(S_{i_{k-1}}, t_{k-1}\right)$$

For Markov processes the following equation also holds

$$P(S_n, t+\Delta t) = P(S_n, t) + \\ + \sum_j \left[P\left(S_n, t+\Delta t | S_j, t\right) P(S_j, t) - P\left(S_j, t+\Delta t | S_n, t\right) P(S_n, t) \right]$$

(*master equation*).

When we assume that the transition probabilities are time-invariant, and we define the transition rates T

$$P\left(S_n, t+\Delta t | S_j, t\right) = T_{n,j} \Delta t$$

we find the differential form of the master equation

$$\frac{d}{dt} P(S_n, t) = \sum_j \left[T_{n,j} P(S_j, t) - T_{j,n} P(S_n, t) \right]$$

Using the previous notation for the probability distribution on states, we can rewrite the master equation as follows

$$\frac{d\pi_n}{dt} = \sum_j \left[T_{n,j}\pi_j(t) - T_{j,n}\pi_n(t)\right]$$

Next, we assume that transition probabilities are "reversible"

$$T_{n,j} = T_{j,n}$$

so that

$$\frac{d\pi_n}{dt} = \sum_j T_{n,j}\left[\pi_j(t) - \pi_n(t)\right]$$

and therefore, at equilibrium

$$\sum_j T_{n,j}\left(\pi_j^* - \pi_n^*\right) = 0 \quad \Rightarrow \quad \pi_j^* = \pi_n^*$$

all states are equally likely at equilibrium

Now consider the following sum

$$\boxed{H \sim -S_G}$$

$$H = \sum_n \pi_n \ln \pi_n$$

Using the master equation we find a differential equation for H

$$\frac{dH}{dt} = \sum_n \frac{d}{dt}(\pi_n \ln \pi_n) = \sum_n \frac{d\pi_n}{dt}(\ln \pi_n + 1)$$

$$= \sum_{n,j} T_{n,j} \left(\pi_j - \pi_n\right) \left(\ln \pi_n + 1\right)$$

Exchanging indexes ...

$$\frac{dH}{dt} = \sum_{n,j} T_{n,j} \left(\pi_n - \pi_j\right) \left(\ln \pi_j + 1\right)$$

Adding the two differential equations we find

$$\frac{dH}{dt} = \frac{1}{2}\sum_{n,j} T_{n,j}\left(\pi_n - \pi_j\right)\left(\ln \pi_j - \ln \pi_n\right)$$

Since

$$\left(\pi_n - \pi_j\right)\left(\ln \pi_j - \ln \pi_n\right) \leq 0$$

we find

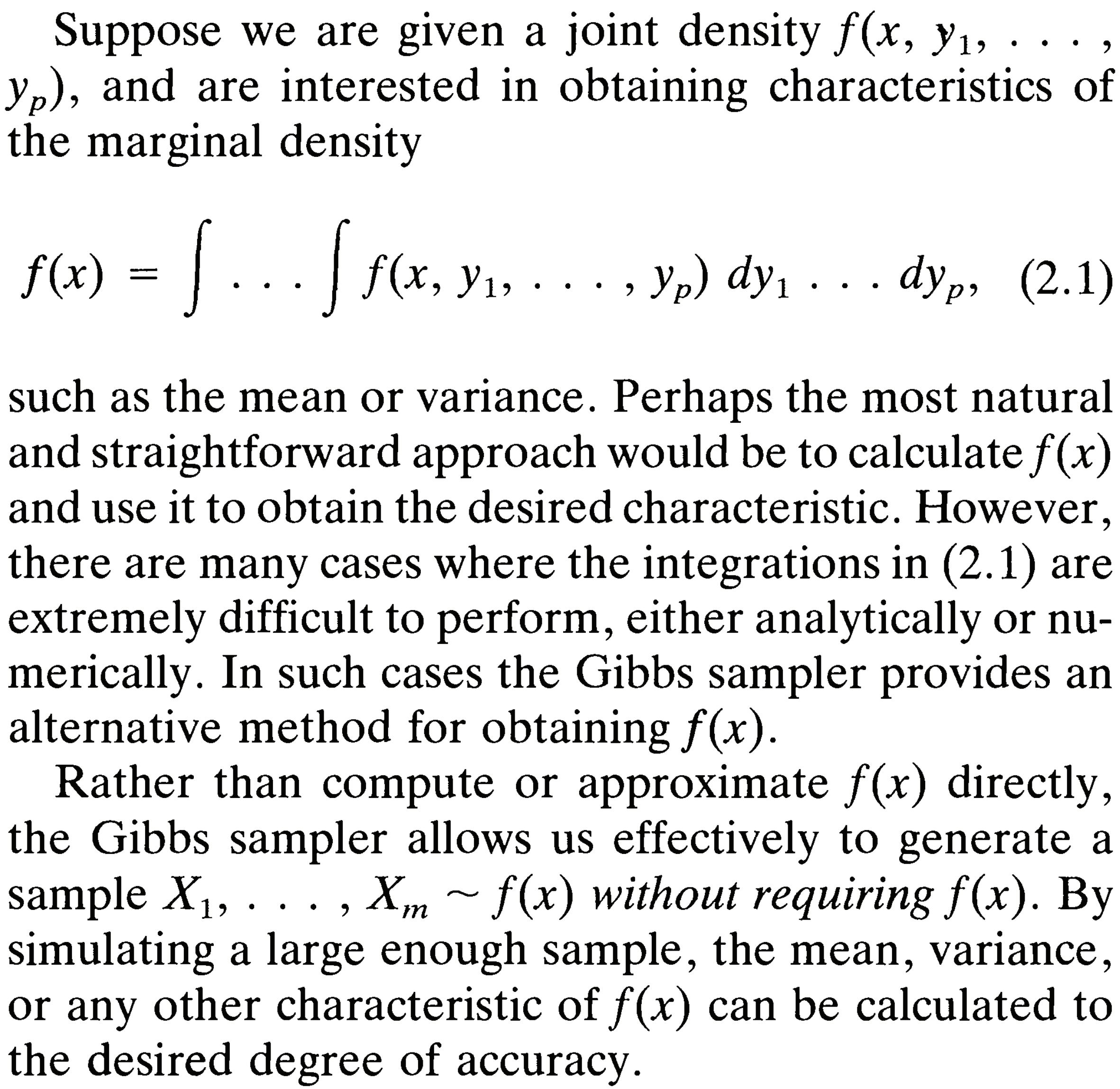
$$\boxed{\frac{dH}{dt} \leq 0}$$

Boltzmann's H-theorem

The derivative vanishes at equilibrium, and we find that it is a stable point for *H*. Since *H* is essentially the negative of Gibbs' entropy, the theorem states that the entropy of a Markov chain increases up to a maximum which is reached at equilibrium.

7. The Gibbs sampler

(adapted from Casella and George,
Explaining the Gibbs sampler Am.Stat. 46 (1992) 167)

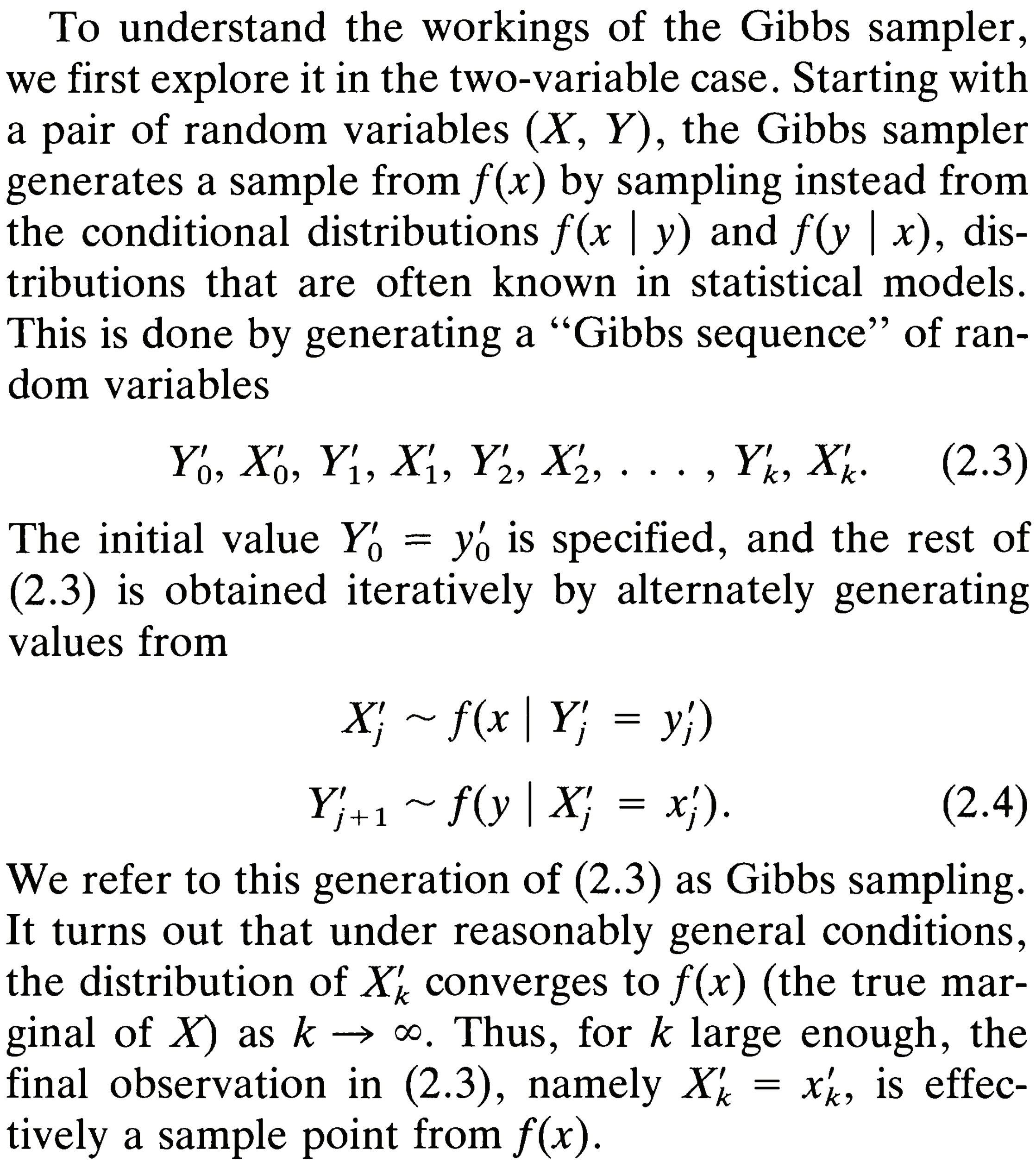


Suppose we are given a joint density $f(x, y_1, \ldots, y_p)$, and are interested in obtaining characteristics of the marginal density

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p)\, dy_1 \ldots dy_p, \quad (2.1)$$

such as the mean or variance. Perhaps the most natural and straightforward approach would be to calculate $f(x)$ and use it to obtain the desired characteristic. However, there are many cases where the integrations in (2.1) are extremely difficult to perform, either analytically or numerically. In such cases the Gibbs sampler provides an alternative method for obtaining $f(x)$.

Rather than compute or approximate $f(x)$ directly, the Gibbs sampler allows us effectively to generate a sample $X_1, \ldots, X_m \sim f(x)$ *without requiring* $f(x)$. By simulating a large enough sample, the mean, variance, or any other characteristic of $f(x)$ can be calculated to the desired degree of accuracy.



To understand the workings of the Gibbs sampler, we first explore it in the two-variable case. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$, distributions that are often known in statistical models. This is done by generating a "Gibbs sequence" of random variables

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'. \qquad (2.3)$$

The initial value $Y_0' = y_0'$ is specified, and the rest of (2.3) is obtained iteratively by alternately generating values from

$$\begin{aligned} X_j' &\sim f(x \mid Y_j' = y_j') \\ Y_{j+1}' &\sim f(y \mid X_j' = x_j'). \end{aligned} \qquad (2.4)$$

We refer to this generation of (2.3) as Gibbs sampling. It turns out that under reasonably general conditions, the distribution of X_k' converges to $f(x)$ (the true marginal of X) as $k \rightarrow \infty$. Thus, for k large enough, the final observation in (2.3), namely $X_k' = x_k'$, is effectively a sample point from $f(x)$.

Let's start with an example, and consider the following joint distribution:

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \quad x = 0,\dots,n \quad 0 \le y \le 1$$

We see that

$$\begin{aligned} f(x|y) &\sim \mathrm{Binomial}(n,y) \\ f(y|x) &\sim \mathrm{Beta}(x+\alpha, n-x+\beta) \end{aligned}$$

It is also easy to see that the properly normalized distribution is

$$p(x,y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}$$

using

$$\begin{aligned} \mathrm{B}(m,n) &= \int_0^1 t^{m-1}(1-t)^{(}n-1)dt \\ &= \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \end{aligned}$$

$$p(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}$$

marginal distribution

How do we recover a marginal pdf when we cannot carry out explicit calculations???

We generate a "Gibbs sequence" of random variables

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'$$

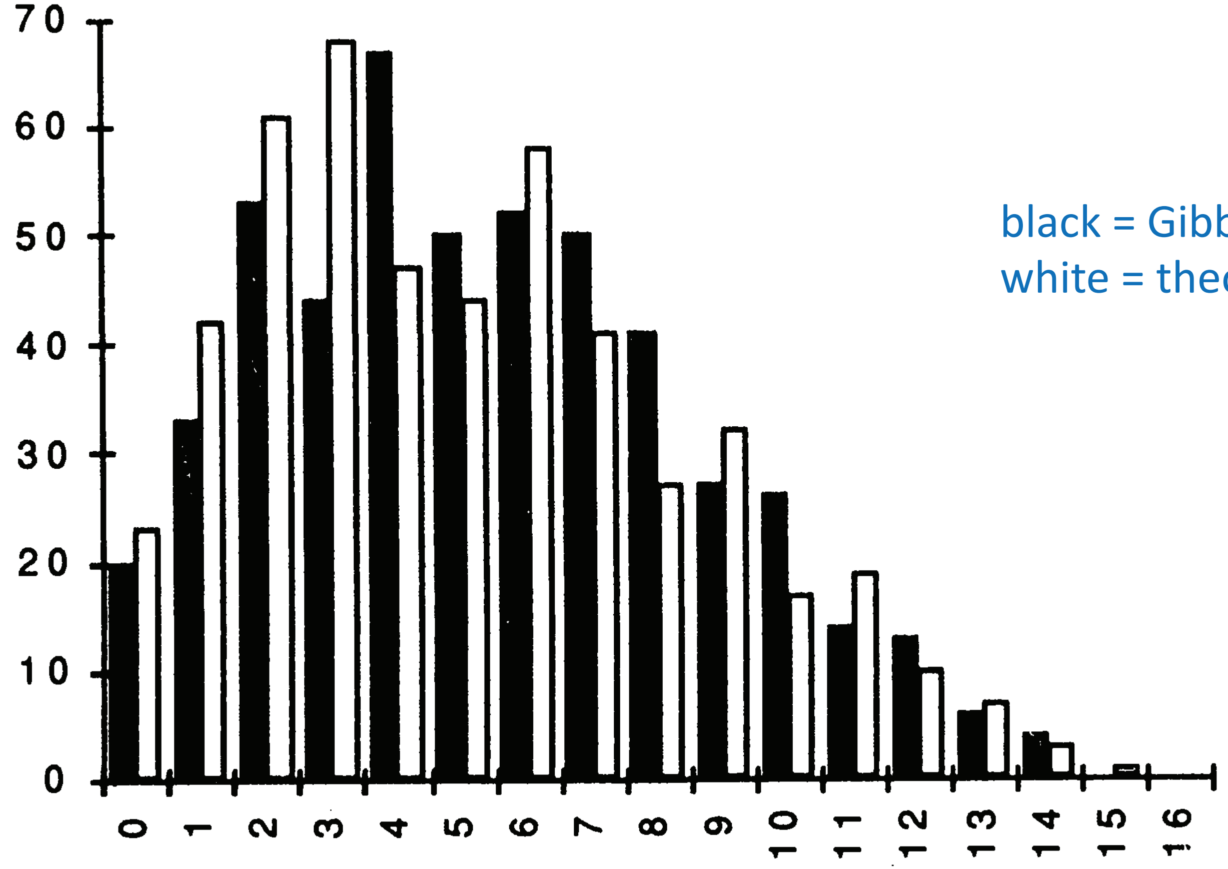
where the initial values are specified and the others are computed with the rule

$$X_j' \sim f(x \mid Y_j' = y_j')$$

$$Y_{j+1}' \sim f(y \mid X_j' = x_j')$$

(Gibbs sampling).

We observe that for large enough *k*, the final *X* values have a fixed distribution that corresponds to the marginal pdf of the *x* variate.



black = Gibbs sampling
white = theoretical expectation

Figure 1. Comparison of Two Histograms of Samples of Size $m = 500$ From the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram sample was obtained using Gibbs sampling with $k = 10$. The white histogram sample was generated directly from the beta-binomial distribution.

Should we expect this result?

Consider the following expectation value

$$E_y[f(x|y)] = \int_Y f(x|y)f(y)dy = \int_Y f(x,y)dy = f(x)$$

therefore we can estimate *f*(*x*) with the sum

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} f(x \mid y_i)$$

where the y's are generated according to their marginal distribution; finally the Gibbs sampling provides representative samples that correspond to the marginal distribution of the x's. (for a mathematically accurate proof, check the paper by Casella&George)

Does Gibbs sampling converge?

We consider the following case: two discrete random variables with marginally Bernoulli distributions and with a joint probability distribution described by this matrix

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{X} \\ & & 0 & 1 \\ \cline{3-4} & 0 & p_1 & p_2 \\ \cline{3-4} Y & 1 & p_3 & p_4 \\ \cline{3-4} \end{array}$$

$$p_i \geq 0,\ p_1 + p_2 + p_3 + p_4 = 1$$

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$

$$f_x = [f_x(0) \quad f_x(1)] = [p_1 + p_3 \quad p_2 + p_4]$$

marginal distribution

from the usual formula for conditional probabilities

$$f_{y|x}(y|x) = \frac{f(x,y)}{f_x(x)}$$

$$A_{y|x} = \begin{bmatrix} \dfrac{p_1}{p_1 + p_3} & \dfrac{p_3}{p_1 + p_3} \\ \dfrac{p_2}{p_2 + p_4} & \dfrac{p_4}{p_2 + p_4} \end{bmatrix}$$

$$A_{x|y} = \begin{bmatrix} \dfrac{p_1}{p_1 + p_2} & \dfrac{p_2}{p_1 + p_2} \\ \dfrac{p_3}{p_3 + p_4} & \dfrac{p_4}{p_3 + p_4} \end{bmatrix}$$

transition probabilities

Since we are only interested in the X sequence

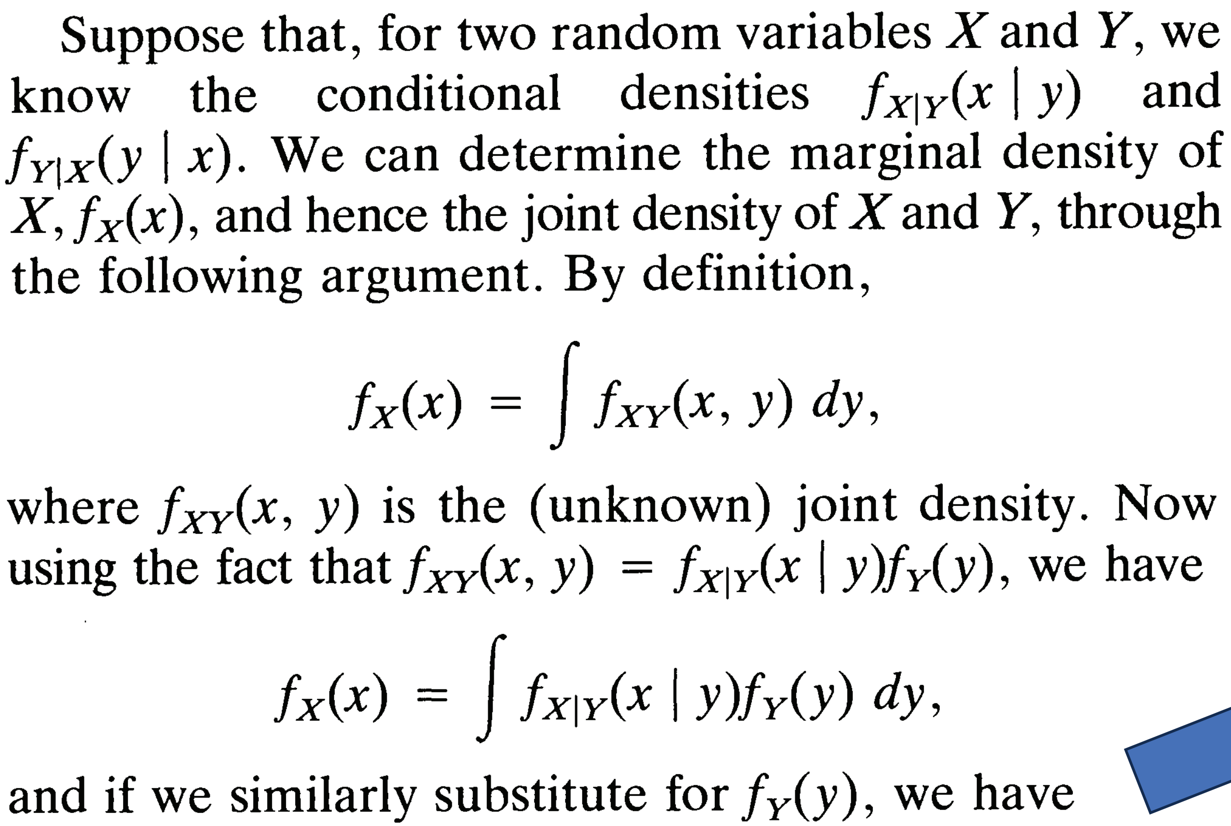
$$P(X_1' = x_1 \mid X_0' = x_0) = \sum_y P(X_1' = x_1 \mid Y_1' = y) \\ \times P(Y_1' = y \mid X_0' = x_0).$$

the transition matrix for the X sequence is

$$A_{x|x} = A_{y|x} A_{x|y}$$

This defines the transition probabilities for a Markov chain and from the theory of Markov chains we know that iterating this produces a fixed probability distribution, i.e., our marginal distribution for X.

How do we prove that the conditionals determine the marginals? Consider the bivariate case



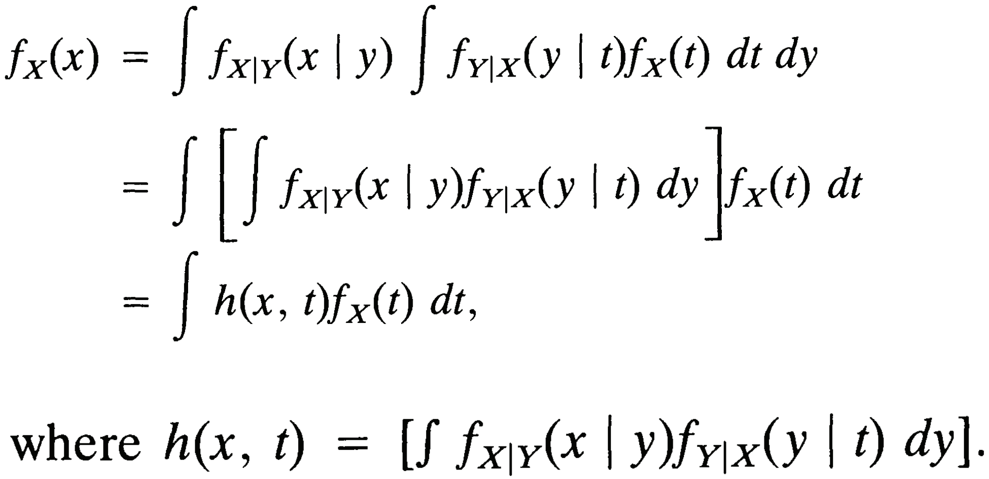
Suppose that, for two random variables X and Y, we know the conditional densities $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. We can determine the marginal density of X, $f_X(x)$, and hence the joint density of X and Y, through the following argument. By definition,

$$f_X(x) = \int f_{XY}(x, y)\, dy,$$

where $f_{XY}(x, y)$ is the (unknown) joint density. Now using the fact that $f_{XY}(x, y) = f_{X|Y}(x \mid y) f_Y(y)$, we have

$$f_X(x) = \int f_{X|Y}(x \mid y) f_Y(y)\, dy,$$

and if we similarly substitute for $f_Y(y)$, we have



$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t)\, dt\, dy \\ &= \int \left[\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy \right] f_X(t)\, dt \\ &= \int h(x, t) f_X(t)\, dt, \end{aligned}$$

where $h(x, t) = [\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy]$.

Let's look at the integral equation, how would it look like in a discrete setting? (in a computer, for instance)

$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t)\, dt\, dy \\ &= \int \left[\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy \right] f_X(t)\, dt \\ &= \int h(x, t) f_X(t)\, dt, \end{aligned}$$

⇒

$$[f_X]_i = \sum_j [h]_{ij} [f_X]_j = \sum_j [f_X]_j [h^T]_{ji}$$

or also in vector-matrix notation

$$\mathbf{f}_X = \mathbf{f}_X \mathbf{h}^T$$

which is exactly the eigenvalue problem that must be solved to find the asymptotic distribution in Markov processes.

So, what's the use of all this?

Consider the case where we want to compute the marginal pdf

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_p)\, dy_1 \ldots dy_p$$

in a situation where the multidimensional integral can be hard to compute.

The Gibbs sampler completely bypasses the calculation of the multidimensional integral and affords an easy path to marginalization.

Indeed, the procedure can be easily extended to multidimensional distributions, for example with two nuisance variables we produce the sequence

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2, X'_2, \ldots$$

by means of the conditional PDFs

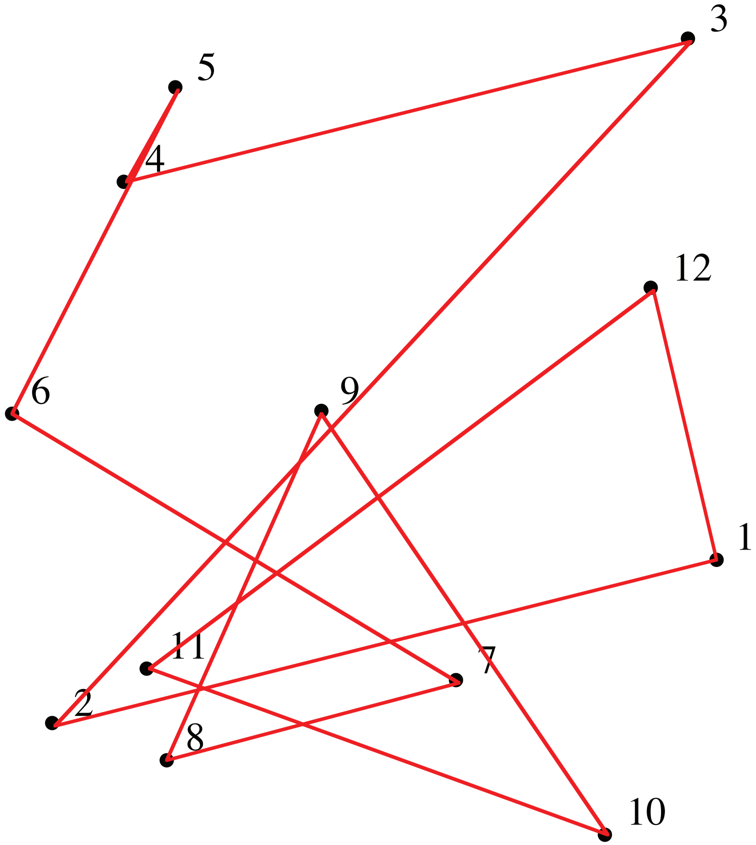
8. The Traveling Salesman Problem and Simulated Annealing

To introduce the method, we consider the *Traveling Salesman Problem* (TSP), where we want to find the shortest closed path that connects N cities.

The problem was first stated by the Viennese mathematician Karl Menger in 1930 and is one of the most studied problems in combinatorial optimization.

For many up-to-date links, see
http://www.math.uwaterloo.ca/tsp/index.html

See also the history page
http://www.math.uwaterloo.ca/tsp/history/index.html



12 “cities” randomly distributed in the (0,1) square: the path corresponds to a random permutation of the sequence of cities.

(path length L=1.93834)

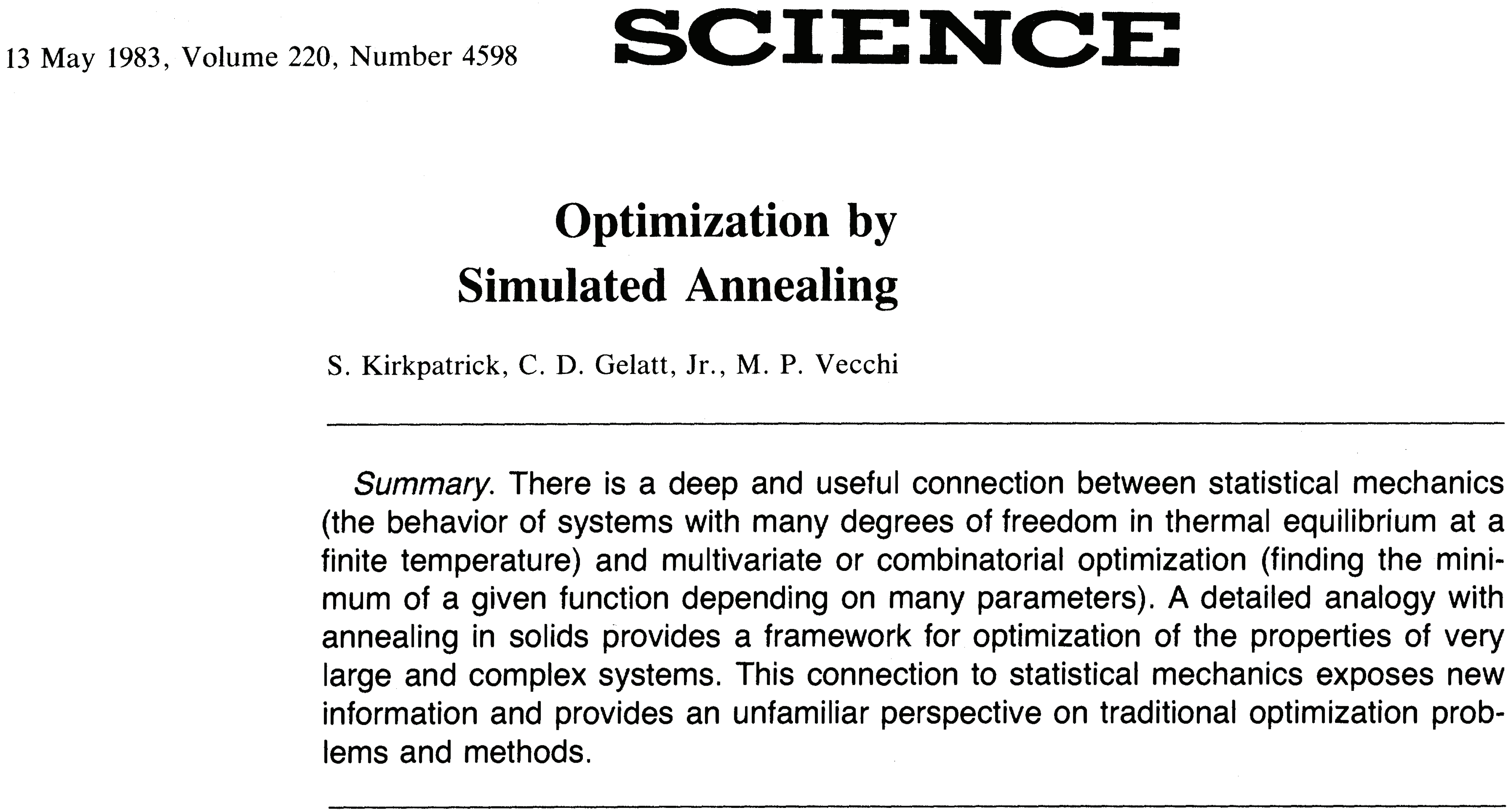
Paths are enumerated by permutations of “city names”, e.g., {9, 2, 7, 8, 1, 12, 4, 5, 3, 10, 11, 6} (start at 9, step to 2, and so on until you reach 6 and then return to 9).

The total number of configurations (undirected paths) is

$$\frac{1}{2}(n-1)!$$

The problem belongs to the class of NP-complete problems (Non-Polynomial complexity, a class of particularly hard problems)

In such cases there is only one known exact solution: the full enumeration of all paths.



13 May 1983, Volume 220, Number 4598 SCIENCE

Optimization by Simulated Annealing

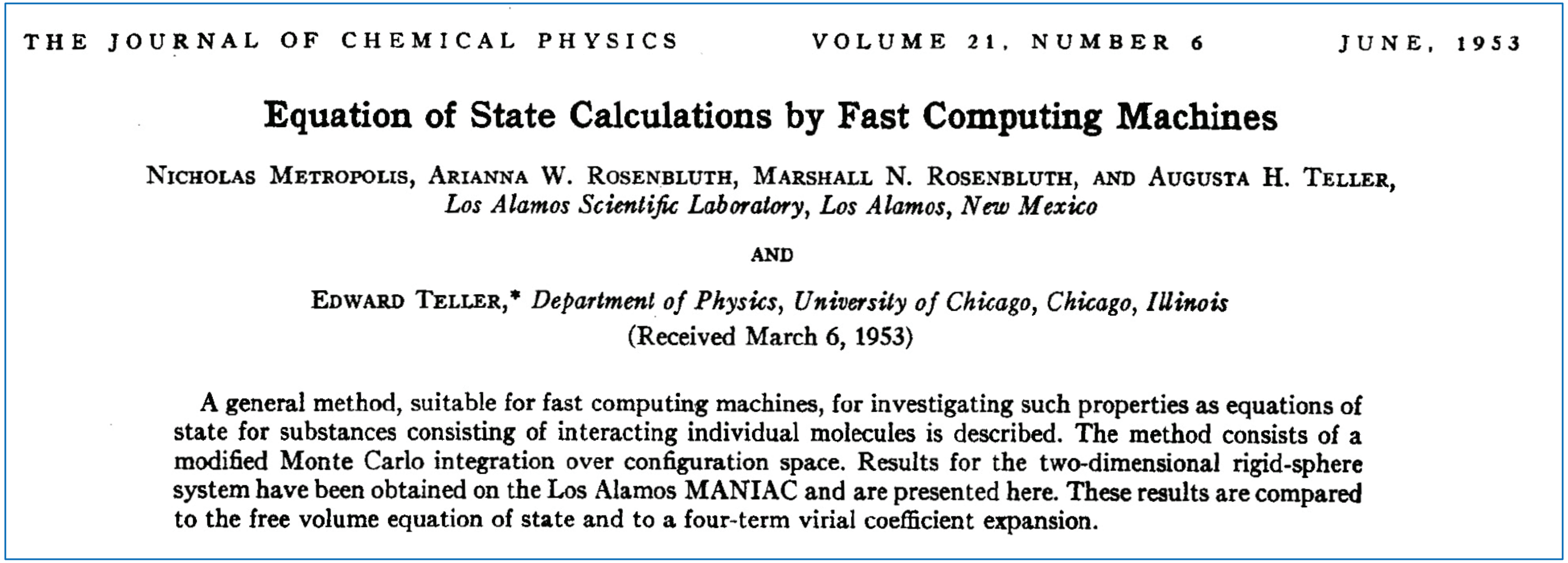
S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

Summary. There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters). A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods.

Approximate solution of the TSP with the Simulated Annealing algorithm

path length → **energy of the system**

exploration of the configuration space with the *Metropolis algorithm* (51909 citations to date, April 10, 2024)
(Metropolis, Rosenbluth Rosenbluth ,Teller and Teller, 1953)



THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

9. The Metropolis algorithm and its application to the TSP

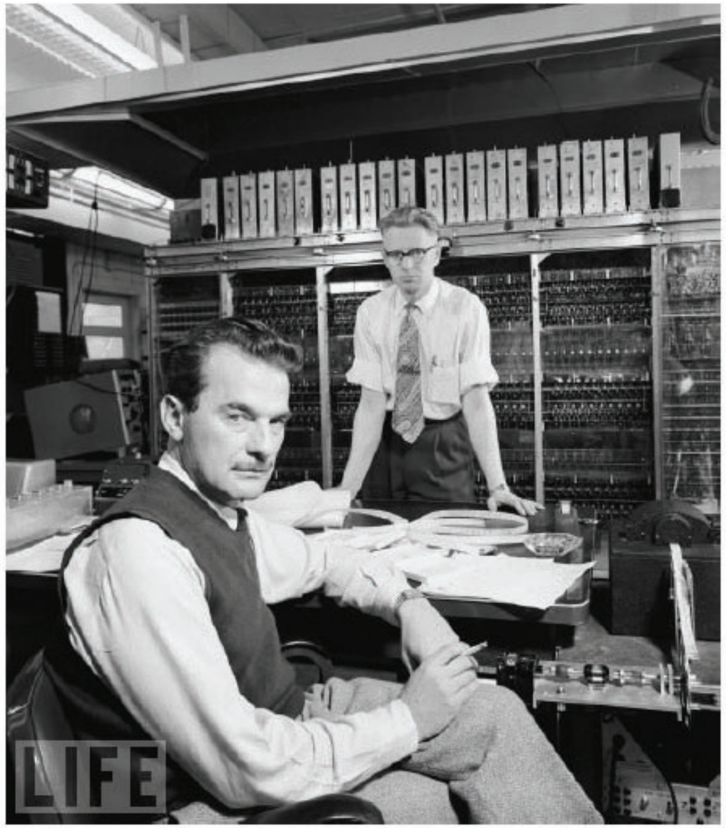


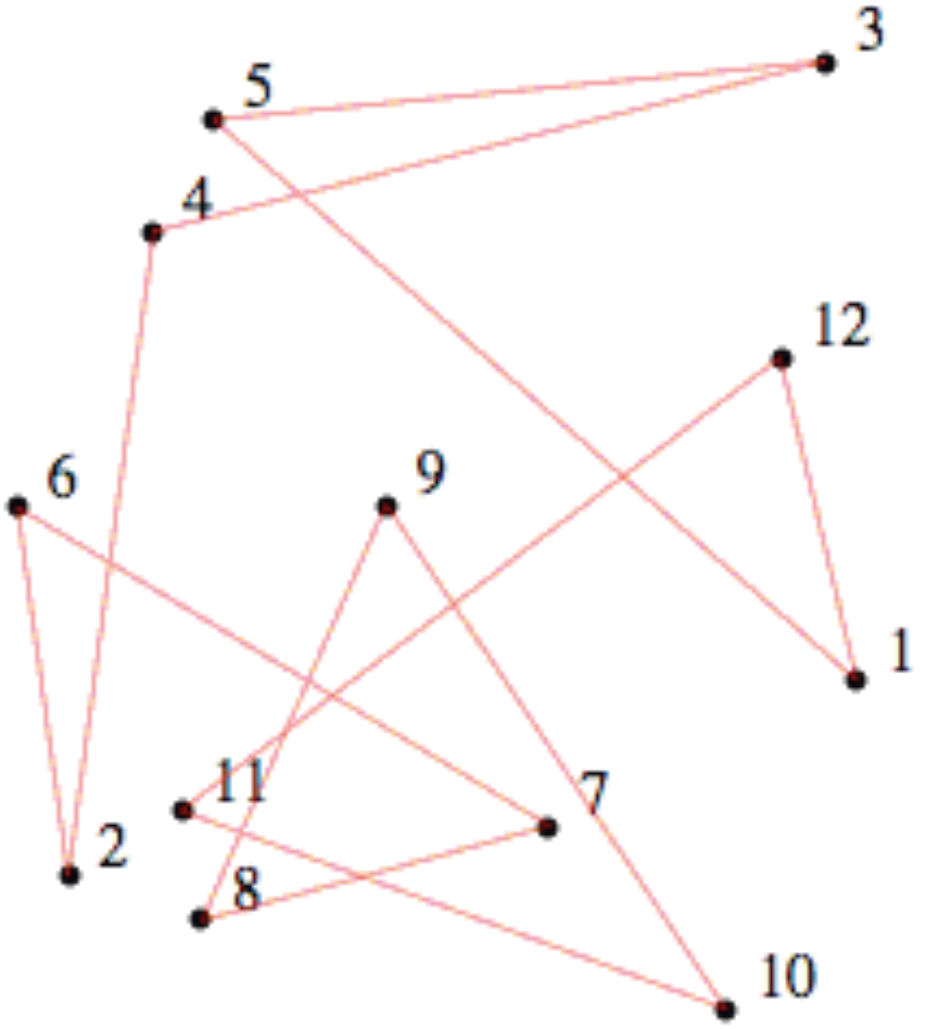
Figure 8.14: Portrait of American computer scientists Nicholas Metropolis (1915 - 1999) (seated) and James Henry Richardson (1918 - 1996) at Los Alamos National Laboratory, Los Alamos, New Mexico, November 1953 (from `http://www.life.com`).

1. We generate a new configuration C′ from the present configuration C
2. We compute the energy of the new configuration, E'
3. We compute the energy difference $\Delta E = E' - E$
4. The new configuration is accepted with probability p

$$\begin{cases} p = 1 & \Delta E < 0 \\ p = \exp\left(-\dfrac{\Delta E}{kT}\right) & \Delta E \geq 0 \end{cases}$$

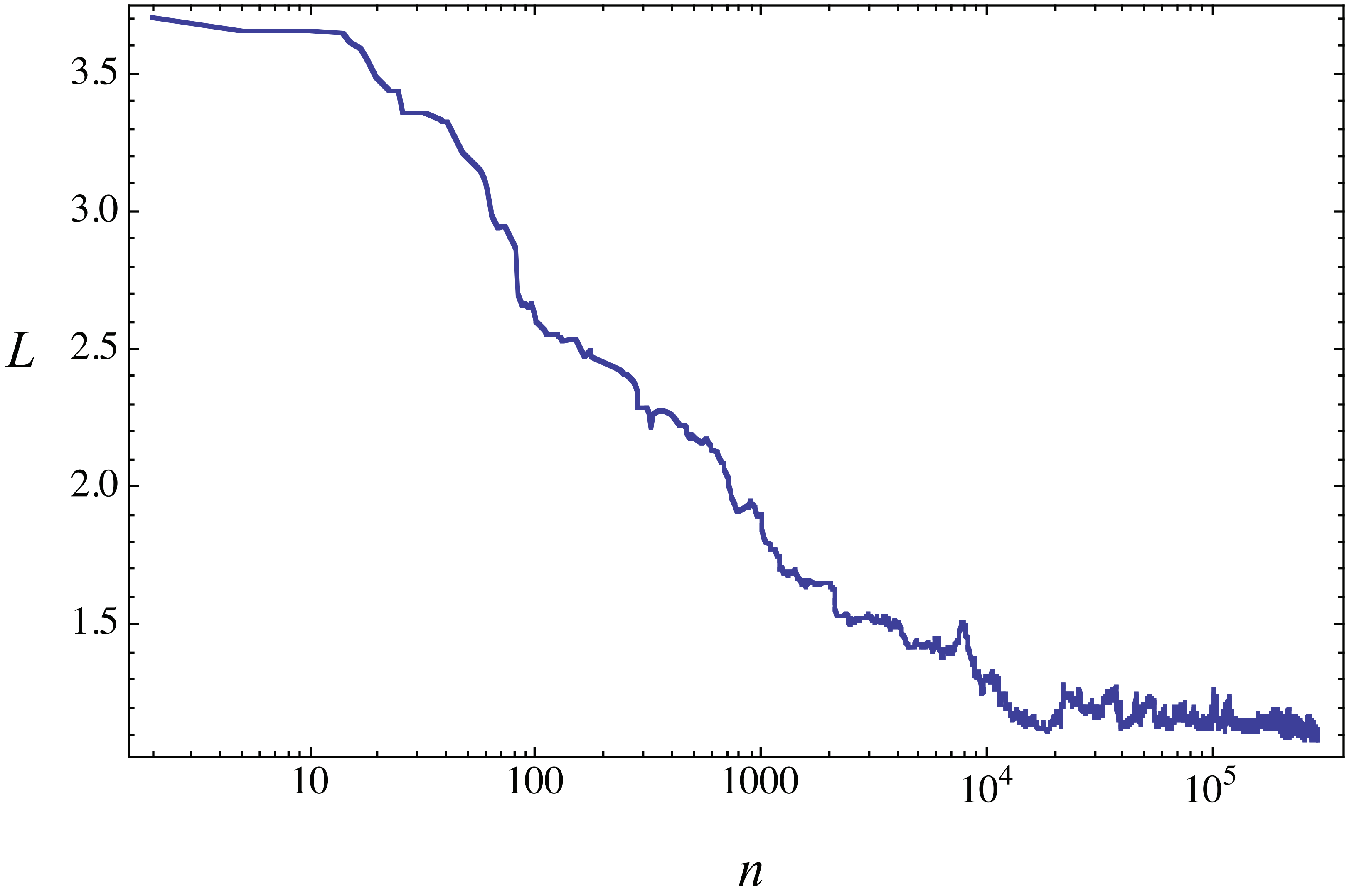
Additional details

- the algorithm needs a slow cooling (it is common to choose an exponential cooling schedule)
- if cooling is not gradual, the system can get stuck into a local minimum
- simple exchanges of pairs of cities are the individual moves in the SA solution of the TSP
- the individual steps from one configuration to the next can be described by a Markov chain



k = 1
T = 0.05
L = 1.84655

Decrease of total path length in a realization of the SA solution of a 50-cities problem



Here we note that the transition probability can be written as follows

$$T(C \to C') = \min\left[1, \exp\left(-\frac{(E' - E)}{kT}\right)\right]$$

Moreover, it is easy to show that the algorithm preserves detailed balance

$$P(C)T(C \to C') = P(C')T(C' \to C)$$

where P(C) is the stationary probability of configuration C. Indeed, at equilibrium we find that, if E' > E,

$$P(C)\exp\left(-\frac{(E' - E)}{kT}\right) = P(C')$$

$$\frac{P(C')}{P(C)} = \exp\left(-\frac{(E' - E)}{kT}\right)$$

Boltzmann's distribution is the equilibrium distribution