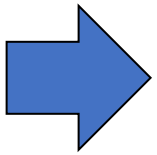
# Introduction to Bayesian Methods - 9

*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. More on Gibbs sampling
9. Simulated annealing and the Traveling Salesman Problem (TSP)
10. The Metropolis algorithm
11. Image restoration and Markov Random Fields (MRF)
12. The Metropolis-Hastings algorithm and Markov Chain Monte Carlo (MCMC)
13. The efficiency of MCMC methods
14. Affine-invariant MCMC algorithms (emcee)

# 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

$$X'_j \sim f(x \mid Y'_j = y'_j)$$

$$Y'_{j+1} \sim f(y \mid X'_j = x'_j). \qquad (2.4)$$

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 $k \to \infty$. 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

$$f(x|y) \sim \text{Binomial}(n,y)$$
$$f(y|x) \sim \text{Beta}(x+\alpha, n-x+\beta)$$

It is also easy to see that the properly normalized distribution is (verify this!)

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

using

$$\text{B}(m,n) = \int_0^1 t^{m-1}(1-t)^{(n-1)}dt = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

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

marginal distribution

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

We generate a "Gibbs sequence" of random variables

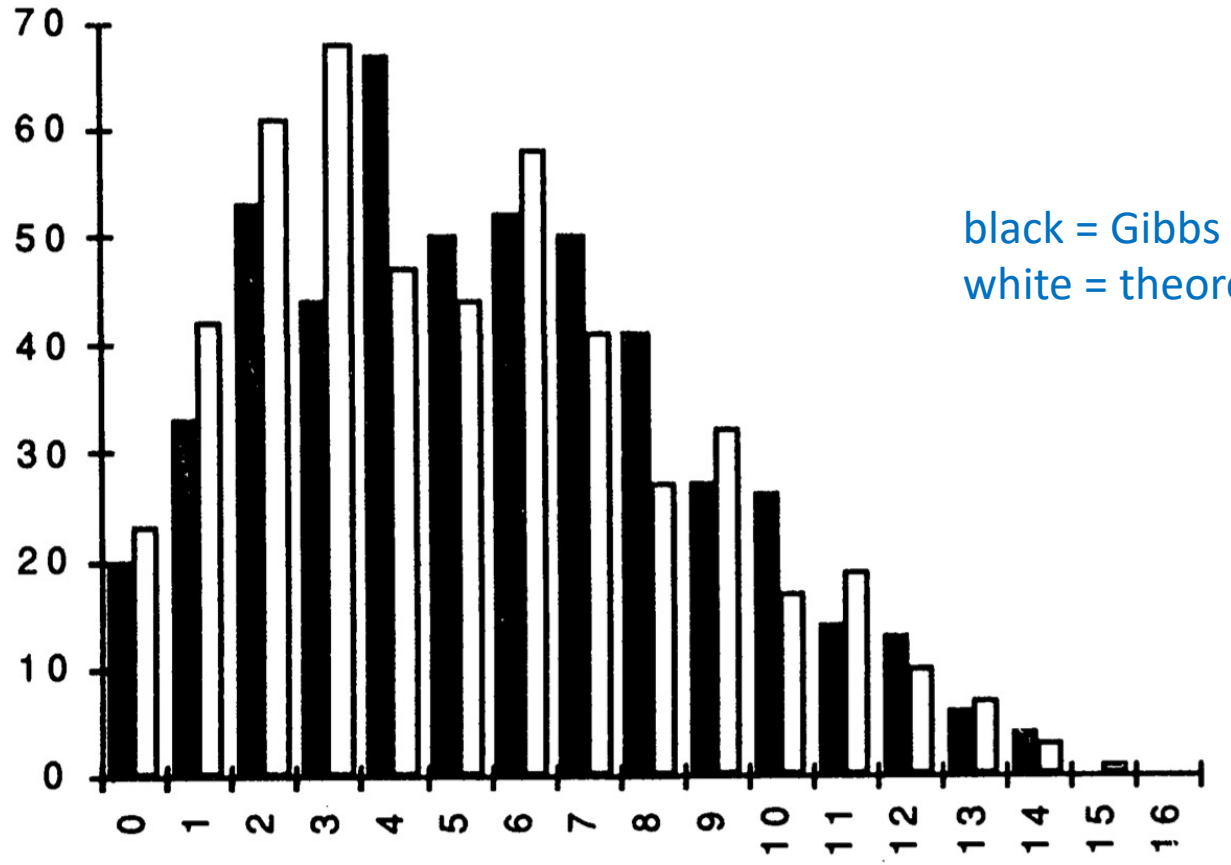
$$Y'_0, X'_0, Y'_1, X'_1, Y'_2, X'_2, \ldots, Y'_k, X'_k$$

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

$$X'_j \sim f(x|Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y|X'_j = x'_j)$$

(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

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

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)

# A simple view of the convergence of Gibbs sampling in the bivariate case.

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

| $Y$ \ $X$ | 0 | 1 |
|---|---|---|
| 0 | $p_1$ | $p_2$ |
| 1 | $p_3$ | $p_4$ |

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

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

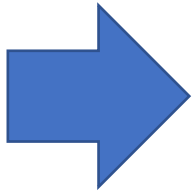
$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}$$

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

marginal distribution

from the usual formula for conditional probabilities

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



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

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

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 single Markov chain in X-space and from the theory of Markov chains we know that iterating this produces a fixed probability distribution, i.e., our marginal distribution for X.

## What about the continuous case? 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

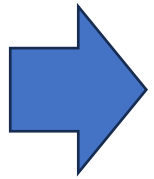
$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t)\, dt\, dy \\ &= \int \left[ \int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy \right] f_X(t)\, dt \\ &= \int h(x, t) f_X(t)\, dt, \end{aligned}$$

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

this is a continuous transition kernel

Let's consider 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

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

**this corresponds to recasting the continuous problem to a discrete problem and we obtain again the eigenvalue problem that must be solved to find the asymptotic distribution in Markov processes.**

# Question: how do we determine the number of steps needed to reach the stationary state?

# Question: how do we determine the number of steps needed to reach the stationary state?

Consider the diagonalized representation of the N-step transition kernel (eigenvalues on the diagonal), such that all N-step transition probabilities are positive in the non-diagonalized version of the kernel (all states can be reached)

$$P^N = \begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & \lambda_2 & 0 & \dots \\ 0 & 0 & \lambda_3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad (1 > \lambda_2 > \lambda_3 > \dots > 0)$$

and the corresponding representation of the initial state vector in terms of eigenvectors

$$\boldsymbol{\pi} = \alpha^* \boldsymbol{\pi}^* + \alpha_2 \boldsymbol{\pi}_2 + \alpha_3 \boldsymbol{\pi}_3 + \dots$$

Then, we see that the repeated action of the N-step transition kernel preserves the stationary state, while it gradually reduces the amplitude of the other states by a factor

$$\lambda_k^{nN}$$

after $nN$ steps.

# Question: how do we determine the number of steps needed to reach the stationary state?

In the case of the Land of Oz wheather model, we find the following diagonalized 2-step transition kernel

$$\mathbf{P}^2 = \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.25 & 0.5 & 0.5 \\ 0.125 & 0.3125 & 0.3125 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} 1. & 0. & 0. \\ 0. & 0.0625 & 0. \\ 0. & 0. & -4.44 \times 10^{-18} \end{pmatrix}$$

## *8. Another view of Gibbs sampling*

Consider again the sequence that generates a "Gibbs sequence" of random variables

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

where one initial value is specified and the others are computed with the rule

$$X_j' \sim f(x|Y_j' = y_j')$$
$$Y_{j+1}' \sim f(y|X_j' = x_j')$$

# Example: bivariate Gaussian distribution

- Bivariate Gaussian likelihood

$$p(y_1, y_2|\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{|\det V|}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})^T V^{-1}(\mathbf{y}-\boldsymbol{\theta})\right]$$

  where the known covariance matrix is

$$V = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- Posterior pdf from Bayes theorem with improper priors, with a single datapoint $(y_1,y_2)$

$$p(\theta_1, \theta_2|y_1, y_2) \propto p(y_1, y_2|\theta_1, \theta_2)$$

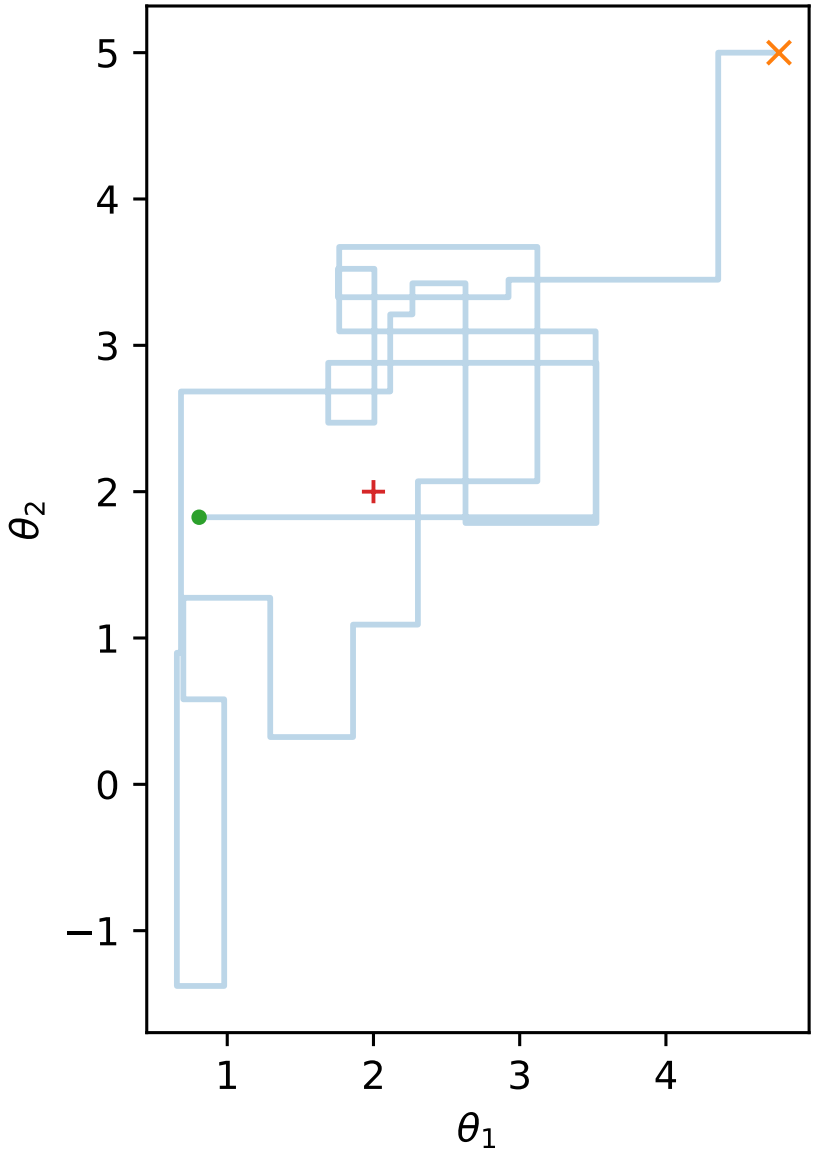
- Expanding, we find

$$p(\theta_1, \theta_2 | y_1, y_2) \sim \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[ (\theta_1 - y_1)^2 - 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right] \right\}$$

- Then, from the Gaussian structure of the posterior, we find the conditional PDFs

$$p(\theta_1 | \theta_2, y_1, y_2) \sim \exp\left[ -\frac{1}{2(1-\rho^2)} \left( \theta_1 - (y_1 + \rho(\theta_2 - y_2)) \right)^2 \right]$$

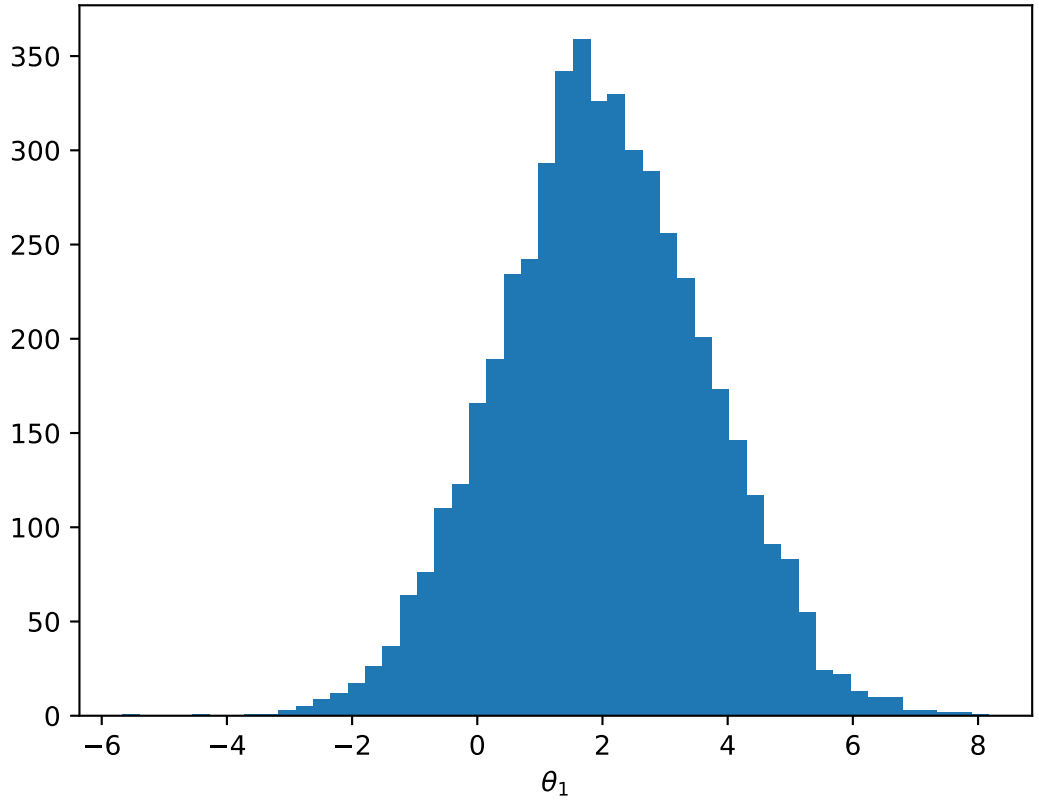
$$p(\theta_2 | \theta_1, y_1, y_2) \sim \exp\left[ -\frac{1}{2(1-\rho^2)} \left( \theta_2 - (y_2 + \rho(\theta_1 - y_1)) \right)^2 \right]$$

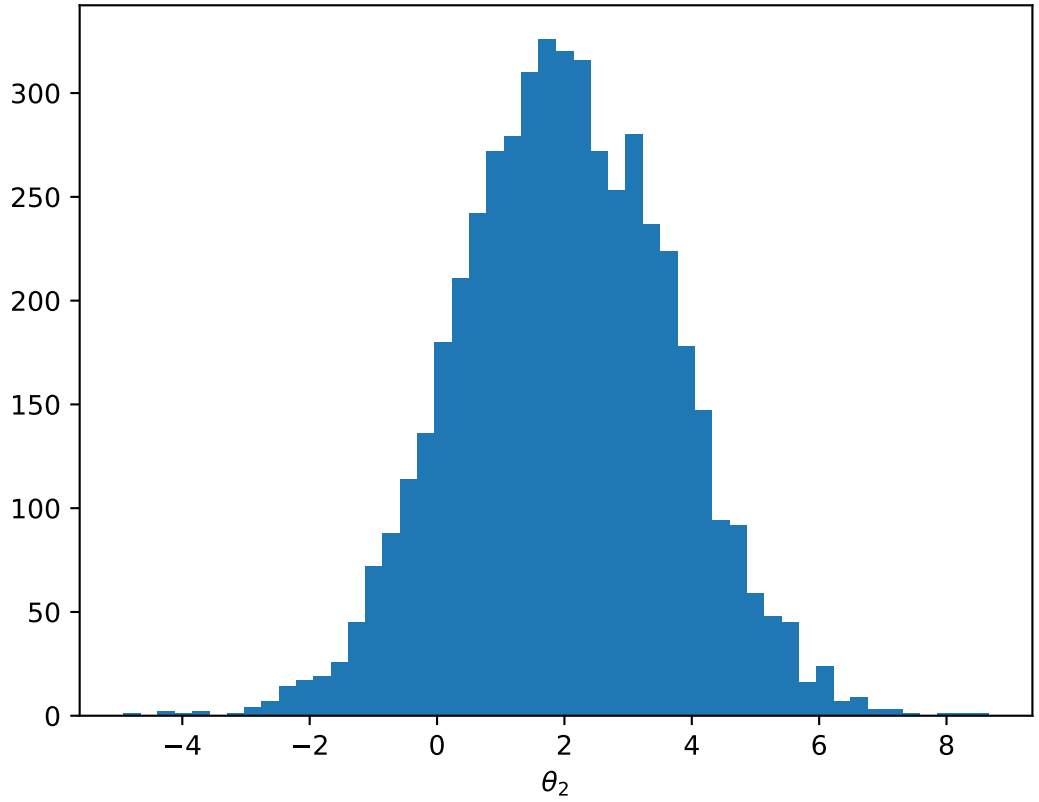



20 initial steps in a Gibbs sampler run:

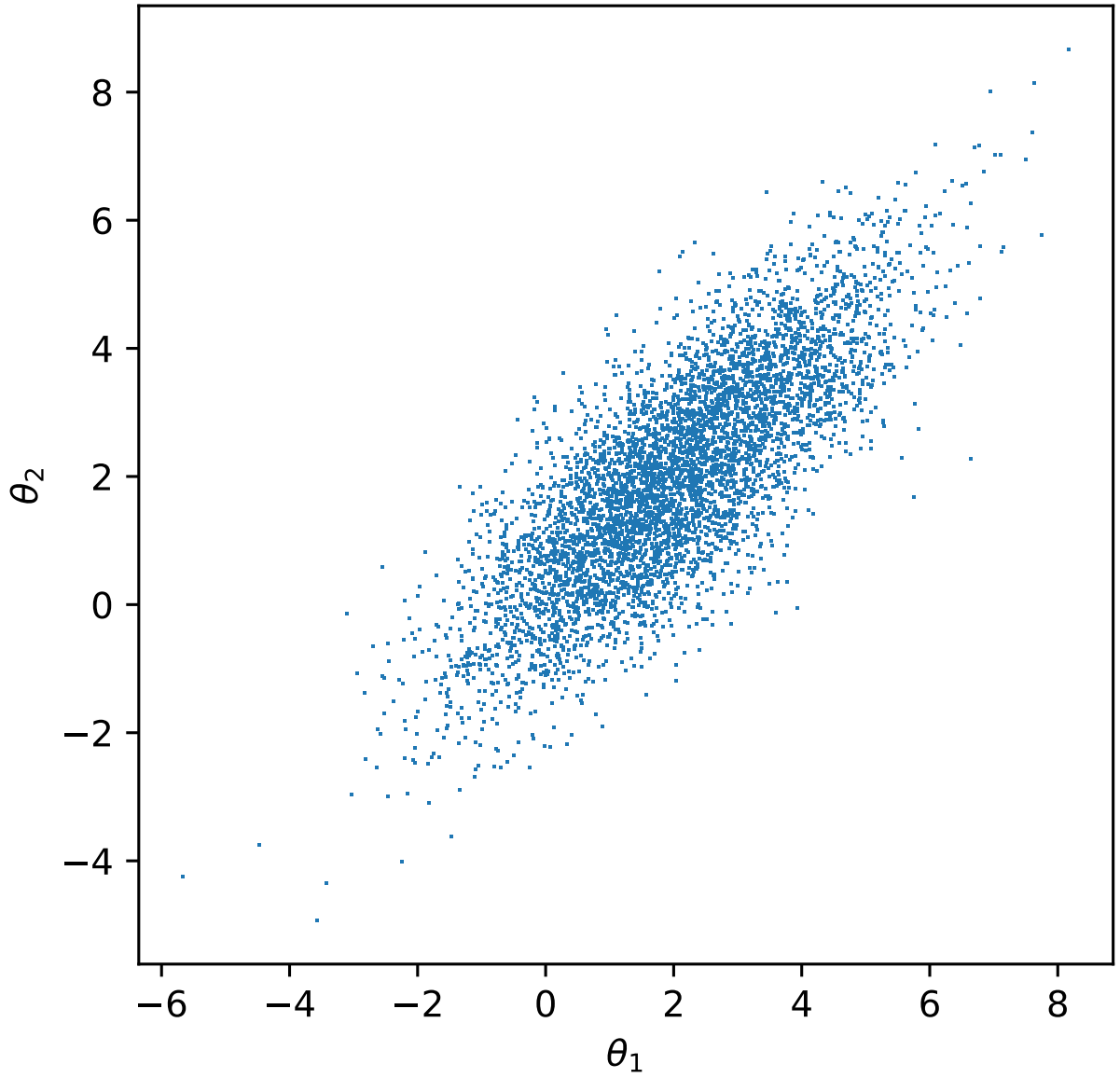
- orange cross: starting pair
- green dot: position after 20 steps
- red cross: bivariate mean

marginal distributions (second half of the simulation values in a run with 10000 generated pairs)



posterior distribution (second half of the simulation values in a run with 10000 generated pairs)




**Exercise**:

extend this treatment to more than one pair of measured $y$ values and write a computer program to implement it

# So, what's the use of all this?

Consider the case where we want to compute the marginal pdf

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

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

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

by means of the conditional PDFs

# Final exercise:

how would you implement Gibbs sampling with a multivariate empirical distribution?

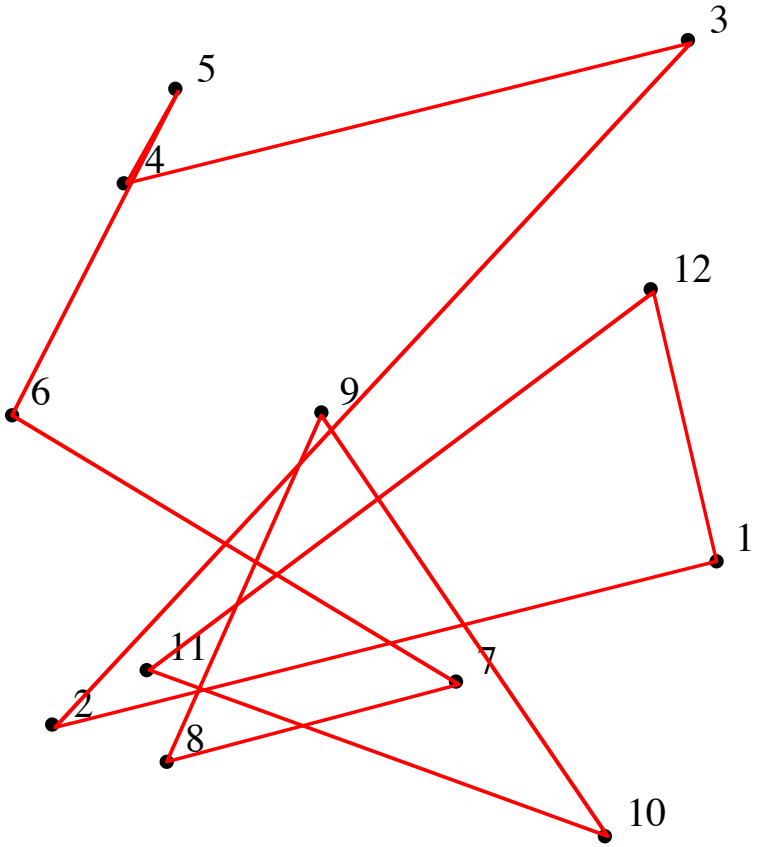
# *9. The Traveling Salesman Problem and Simulated Annealing*

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.

# *10. The Metropolis algorithm and its application to the TSP*

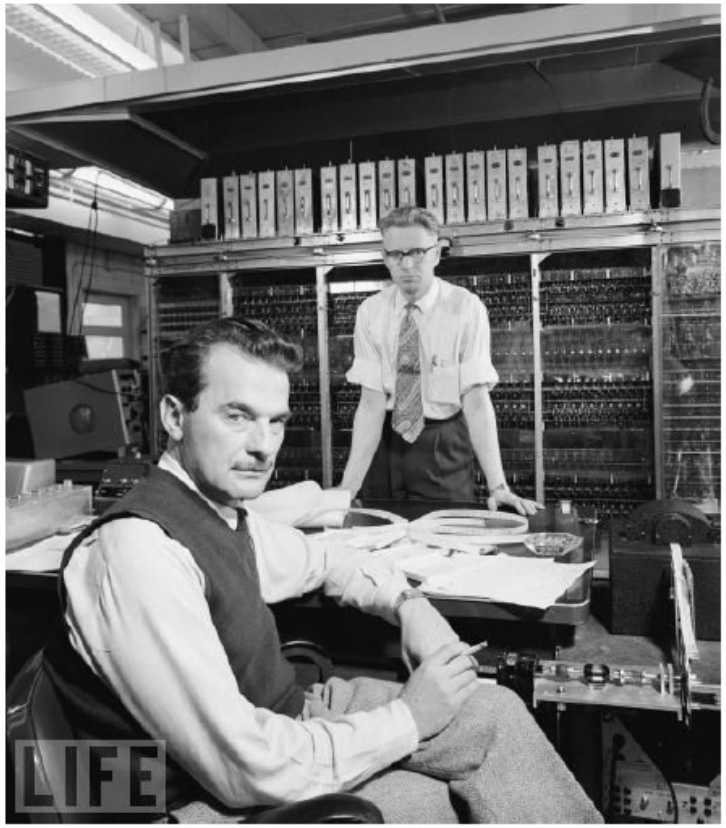


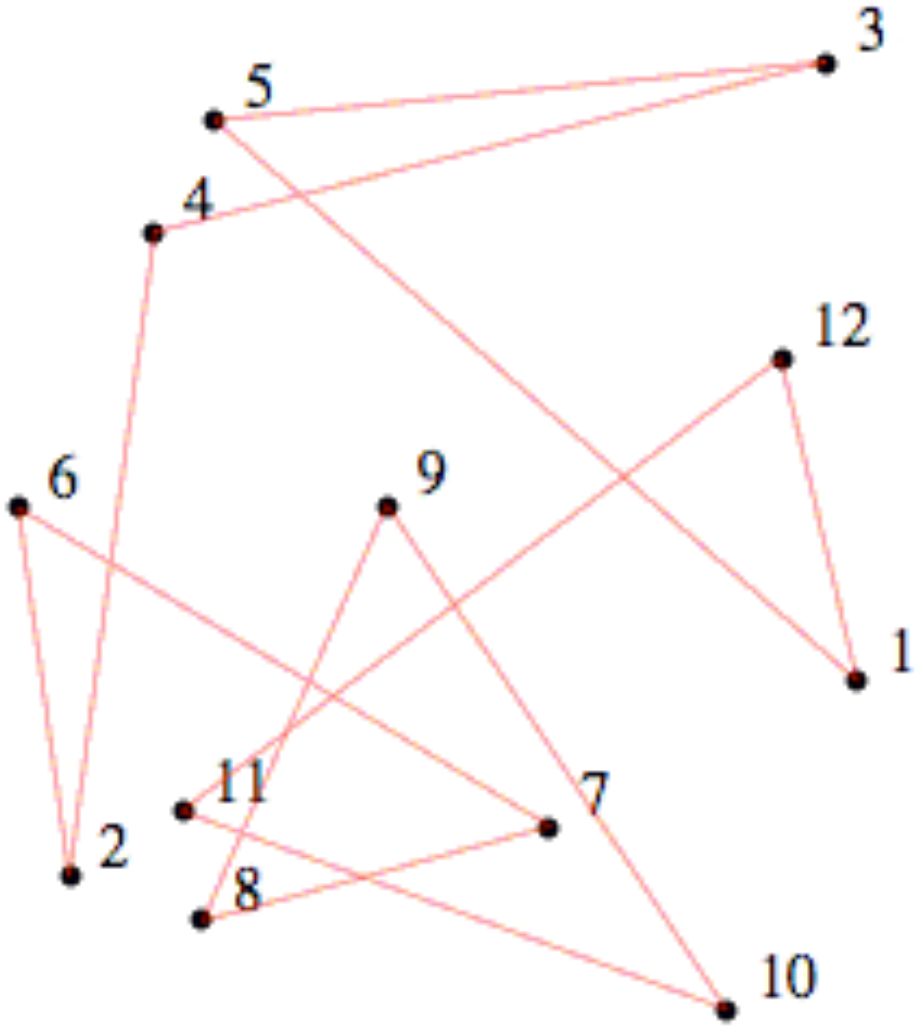

Figure 8.14: Portrait of American computer scientists Nicholas Metropolis (1915 - 1999) (seated) and James Henry Richardson (1918 - 1996) at Los Alamos National Laboratory, Los Alamos, New Mexico, November 1953 (from `http://www.life.com`).

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 $p$

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

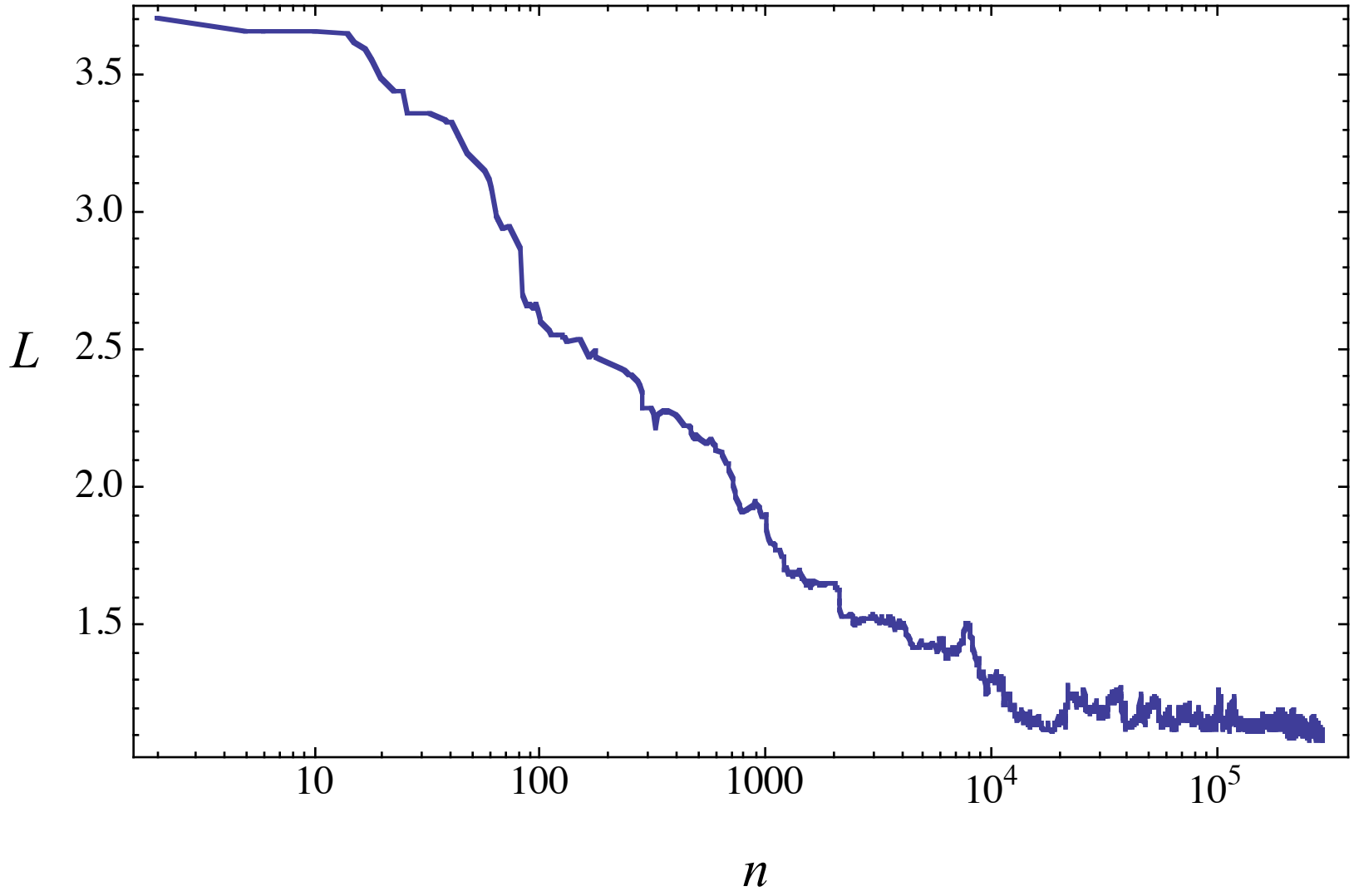
Additional details

- 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

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

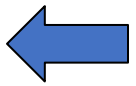
Moreover, it is easy to show that the algorithm preserves detailed balance

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

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

# *11. Image restoration and Markov Random Fields (MRF)*

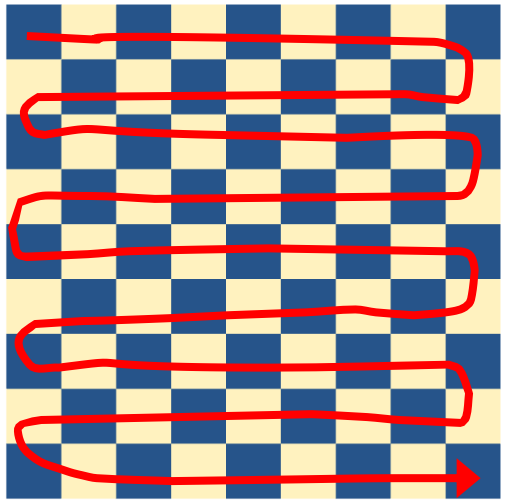
The optimization problem.

When dealing with signals, we usually assume that data are corrupted by noise

$$d = x + w$$

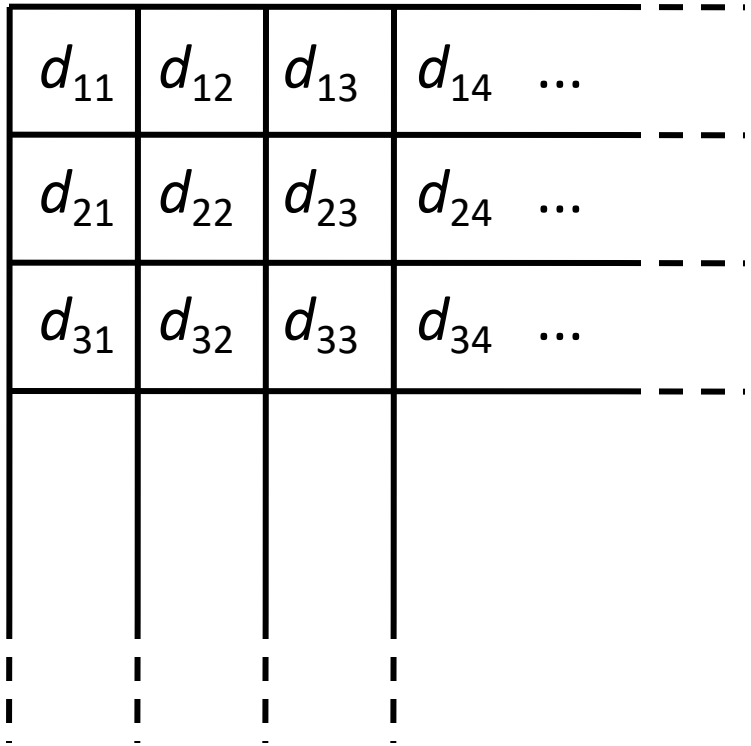
data vector — signal vector — (vector) noise process

An image can be viewed as a vector, for instance unfolding the sequence of pixels as shown on the right, we obtain the equivalent of a long signal vector.



If there are $n$ pixels on each side, there are in all $n^2$ pixels, and if there are L gray levels, then the number of possible configurations that define an image is

$$N = L^{n^2}$$




pixel map

true pixel vector

$\mathbf{x}$

*posterior* pixel distribution

*likelihood*

*prior* pixel distribution

$$P(\mathbf{x}|\mathbf{d}) = \frac{P(\mathbf{d}|\mathbf{x})}{P(\mathbf{d})} P(\mathbf{x}) \propto P(\mathbf{d}|\mathbf{x}) P(\mathbf{d})$$

Bayesian estimate of true pixel vector from observed pixel vector

Given the number of possible configurations

$$N = L^{n^2}$$

we see that even for small image sizes, say $n = 100$, with binary levels, $L = 2$, we find

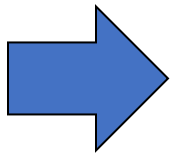
$$N = 2^{10^4}$$

therefore, the problem of finding the MAP estimate in a Bayesian context is a hard computational task.

The MAP estimate depends on the prior distribution

Possible priors:

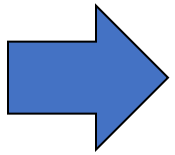
$P(\mathbf{x})$ flat prior



Maximum Likelihood Estimate (MLE)

$$P(\mathbf{x}|\mathbf{d}) \propto P(\mathbf{d}|\mathbf{x})P(\mathbf{d}) \propto P(\mathbf{d}|\mathbf{x})$$

$P(\mathbf{x})$ entropic prior



Maximum Entropy Method (MEM)

Notice also that

$$\ln P(\mathbf{x}|\mathbf{d}) \approx \ln P(\mathbf{d}|\mathbf{x}) - [-\ln P(\mathbf{d})]$$

therefore, the MAP estimate is equivalent to maximizing the likelihood with a *penalty function*

$$-\ln P(\mathbf{d})$$

Experiments have been tried with many different penalties, many of them barely justified on probabilistic grounds (or not at all!)

Now, let $\boldsymbol{x}$ be the vector of "true values" (uncorrupted intensities of an image, a spectrum, etc. ...), and translate these values into counts

$$n_i = \lfloor \alpha d_i \rfloor$$

($i$ = 1, ..., $M$). The least informative prior corresponds to a structureless image, and pixelwise it is once again the uniform prior. Then, the probability of one count at the $i$-th position is just $1/M$.

Likewise, the probability of a given vector of values where the total number of counts is $N$, is given by the multinomial probability

$$P(\mathbf{n}) = \frac{N!}{n_1! n_2! \ldots n_M!} \left(\frac{1}{M}\right)^N ; \quad \sum_k n_k = N$$

Using Stirling's approximation

$$n! \approx n^n e^{-n}; \quad \ln n! \approx n \ln n - n$$

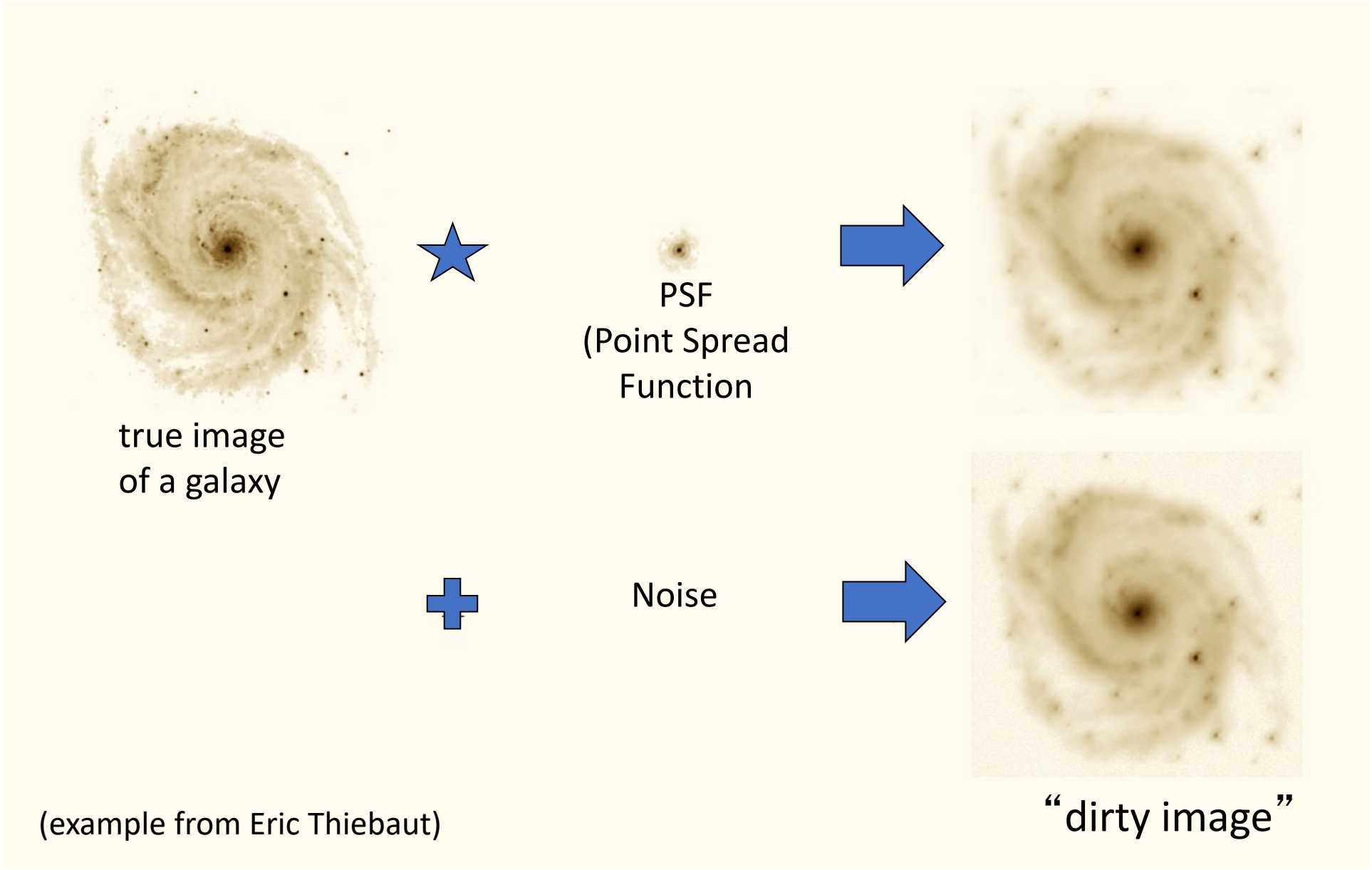
we find, with the definition $p_i = \dfrac{d_i}{\sum_{k=1}^{M} d_k}$

$$\begin{aligned}
\ln P(\mathbf{n}) &\approx (N \ln N - N) - \sum_k (n_k \ln n_k - n_k) \\
&= N \ln N - \sum_k n_k \ln n_k \\
&\approx -\alpha \sum_k d_k \ln d_k + \text{const.}
\end{aligned}$$

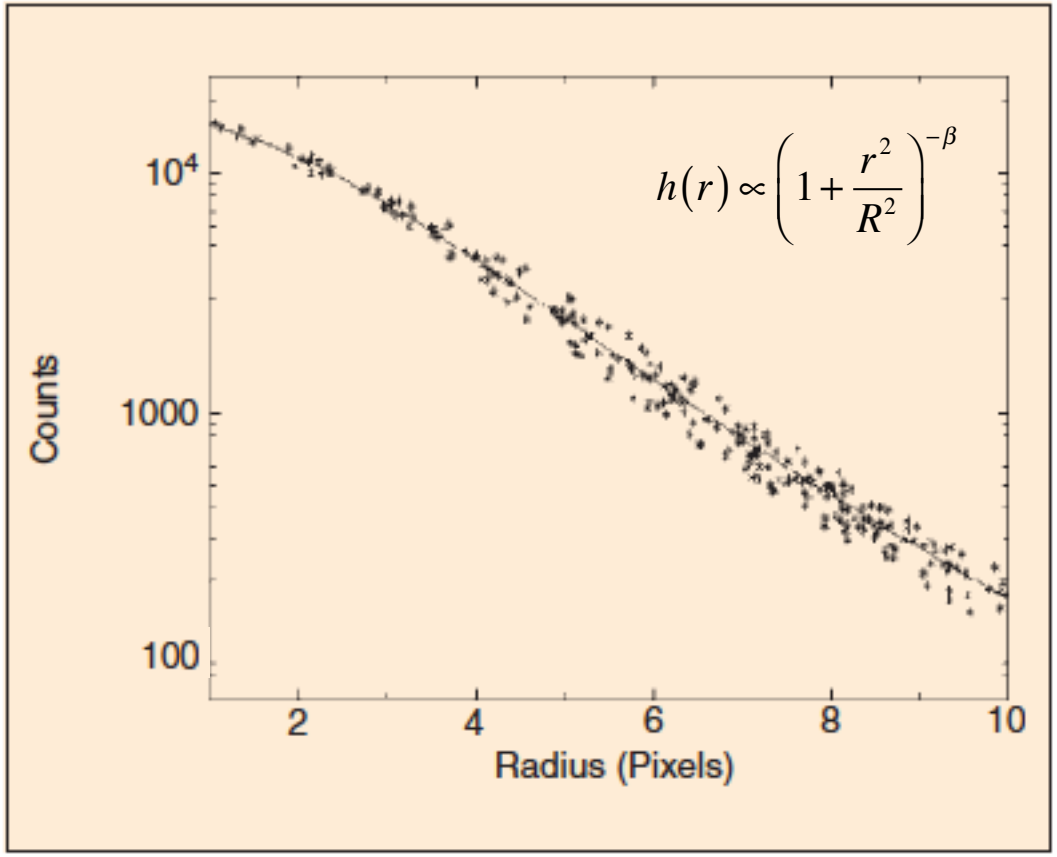
*entropic prior*

$$P(\mathbf{n}) \sim \exp\left(-\alpha \sum_k d_k \ln d_k\right) \sim \exp\left(-\alpha' \sum_k p_k \ln p_k\right) = \exp\left[\alpha' S(\mathbf{d})\right]$$

# Image likelihood: 1. the observation model




(example from Eric Thiebaut)




PSF from atmospheric turbulence

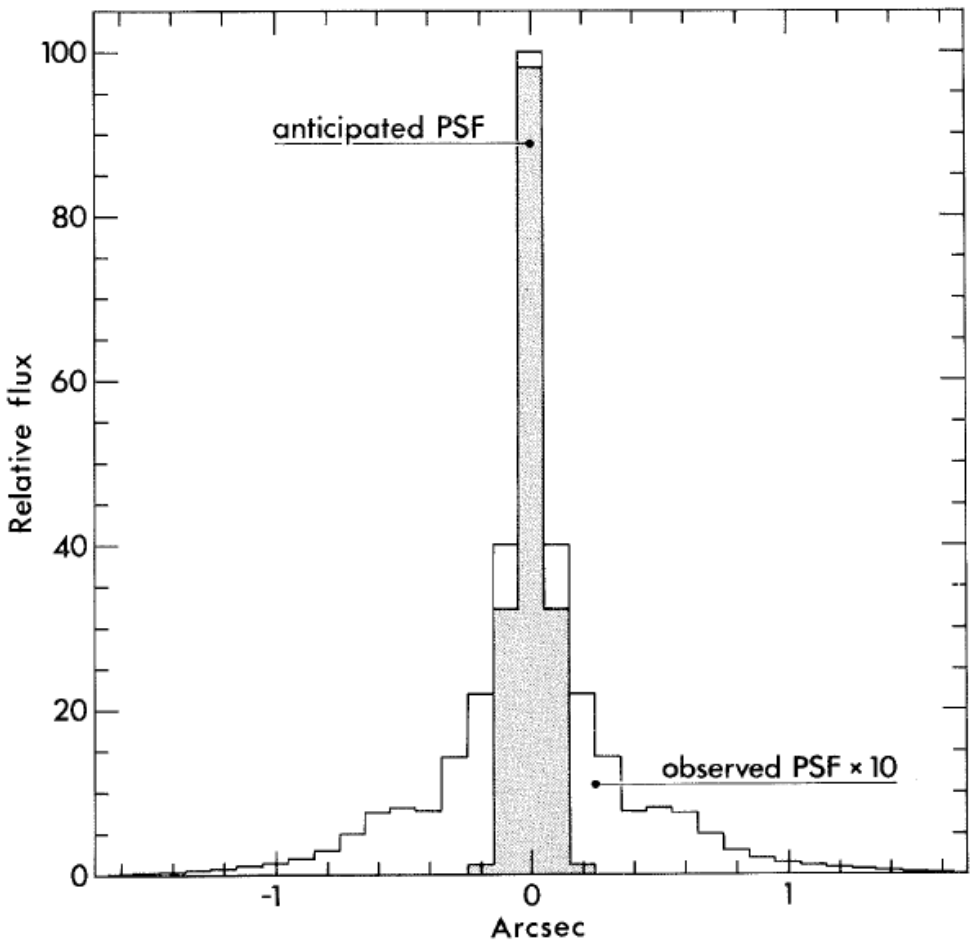
In general, the effect of the PSF is modeled by a linear operator

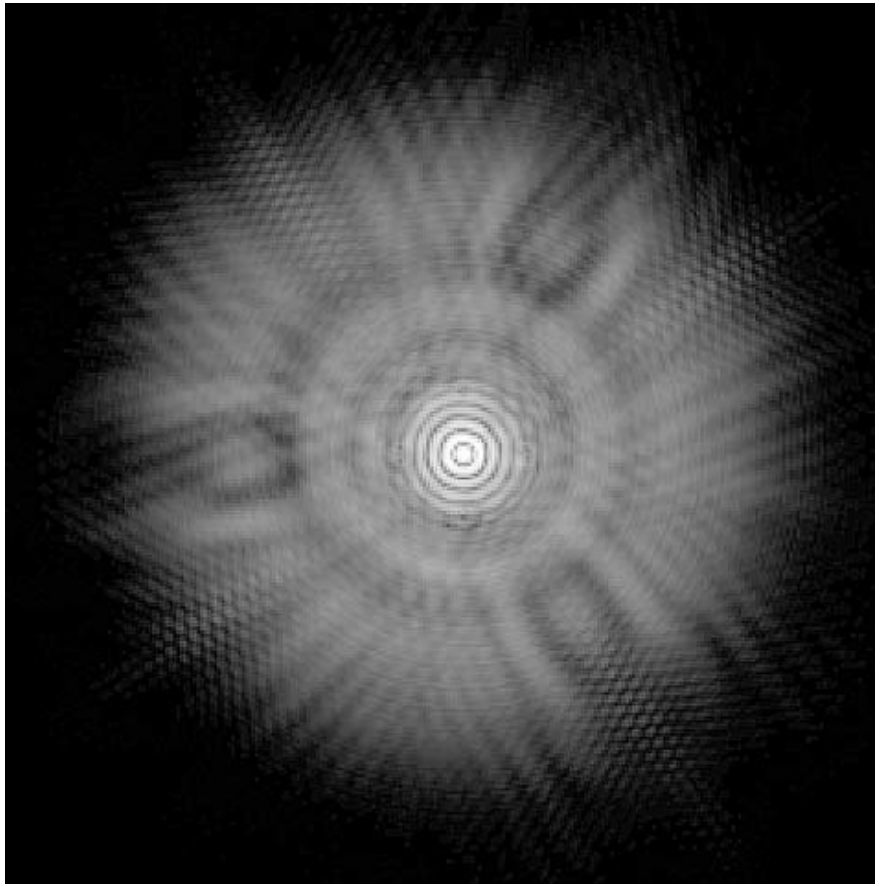
$$\mathbf{x} \rightarrow \mathbf{Hx}$$

"true" pixel vector

the action of the optical system on the "true image" is modeled by the matrix **H**

# The Hubble PSF before the first servicing mission



# Image likelihood: 2. the noise model (degradation model)

Gaussian noise model

$$P(\mathbf{x}|\mathbf{d}) \propto \exp\left[-\frac{(\mathbf{x}-\mathbf{H}\mathbf{d})^2}{2\sigma^2}\right]$$

Poisson noise model

$$P(\mathbf{x}|\mathbf{d}) \propto \prod_n \frac{(\mathbf{H}\mathbf{d})_n^{d_n}}{d_n!} \exp\left[-(\mathbf{H}\mathbf{d})_n\right]$$

Poisson noise mostly from detection process, Gaussian noise mostly from electronics or from approximation of Poisson noise. Sometimes we can use the Gaussian approximation of Poisson noise

$$P(\mathbf{d}|\mathbf{x}) \propto \prod_n \frac{(\mathbf{H}\mathbf{x})_n^{d_n}}{d_n!} \exp\left[-(\mathbf{H}\mathbf{x})_n\right] \approx \prod_n \exp\left[-\frac{(d_n-(\mathbf{H}\mathbf{x})_n)^2}{2(\mathbf{H}\mathbf{x})_n}\right] = \exp\left[-\sum_n \frac{(d_n-(\mathbf{H}\mathbf{x})_n)^2}{2(\mathbf{H}\mathbf{x})_n}\right]$$