Introduction to Bayesian Statistics - 7

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. Simulated annealing
7. The Metropolis algorithm
8. Markov Chain Monte Carlo (MCMC)
9. The Gibbs sampler
10. The efficiency of MCMC algorithms
11. Affine-invariant MCMC algorithms (EMCEE)

4. *Bayesian methods in a sampling-resampling perspective (Smith & Gelfand, 1992)*

Bayesian Statistics Without Tears: A Sampling–Resampling Perspective

A. F. M. SMITH and A. E. GELFAND*

Even to the initiated, statistical calculations based on Bayes's Theorem can be daunting because of the numerical integrations required in all but the simplest applications. Moreover, from a teaching perspective, introductions to Bayesian statistics—if they are given at all—are circumscribed by these apparent calculational difficulties. Here we offer a straightforward sampling–resampling perspective on Bayesian inference, which has both pedagogic appeal and suggests easily implemented calculation strategies.

In Bayesian methods we have to evaluate many integrals, like, e.g.,

$$p(\theta|x) = \frac{l(\theta; x)p(\theta)}{\int l(\theta; x)p(\theta)\, d\theta}$$

normalization (evidence)

$$p(\phi|x) = \int p(\phi, \psi|x)\, d\psi.$$

marginalization

$$E[m(\theta)|x] = \int m(\theta)p(\theta|x)\, d\theta$$

averages (statistical estimators)

except in simple cases, explicit evaluation of such integrals will rarely be possible, and realistic choices of likelihood and prior will necessitate the use of sophisticated numerical integration or analytic approximation techniques (see, for example, Smith et al. 1985, 1987; Tierney and Kadane, 1986). This can pose problems for the applied practitioner seeking routine, easily implemented procedures. For the student, who may already be puzzled and discomforted by the intrusion of too much calculus into what ought surely to be a simple, intuitive, statistical learning process, this can be totally off-putting.

Bayesian learning as a resampling procedure (importance sampling-like scheme)

$$p(\theta_k|D) \propto p(D|\theta_k)p(\theta_k) = \ell(D|\theta_k)p(\theta_k)$$

1. prior distribution defined by the empirical distribution of the initial samples

(sampling)

2. the Likelihood distorts the distribution of initial samples (corresponds to a sample acceptance probability)

(resampling))

3. the posterior distribution is represented by the resampled empirical distribution

Example (McCullagh & Nelder): take two sets of binomially distributed independent random variables X_{i1} and X_{i2} (i=1,2,3)

$$X_{i1} = \text{Binomial}\left(n_{i1}, \theta_1\right)$$

$$X_{i2} = \text{Binomial}\left(n_{i2}, \theta_2\right)$$

The observed random variables are the sums

$$Y_i = X_{i1} + X_{i2}$$

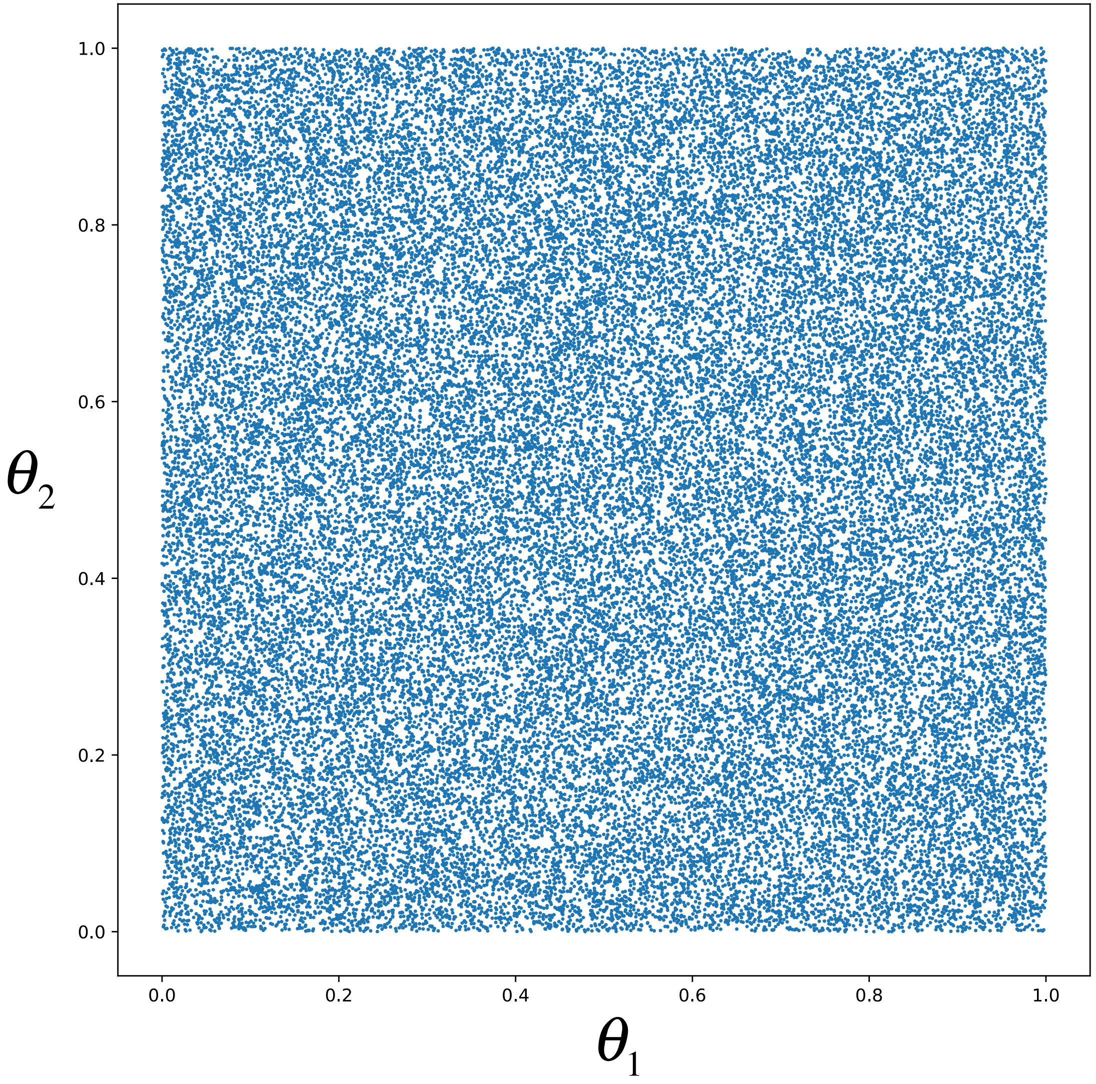
$$\text{likelihood} = \prod_{i=1}^{3} \sum_{j_i} \binom{n_{i1}}{j_i} \binom{n_{i2}}{y_i - j_i} \theta_1^{j_1} (1-\theta_1)^{n_{i1}-j_i} \theta_2^{y_i - j_i} (1-\theta_2)^{n_{i2}-y_i+j_i}$$

$$\max(0, y_i - n_{i2}) \le j_i \le \min(n_{i1}, y_i)$$

Sample data

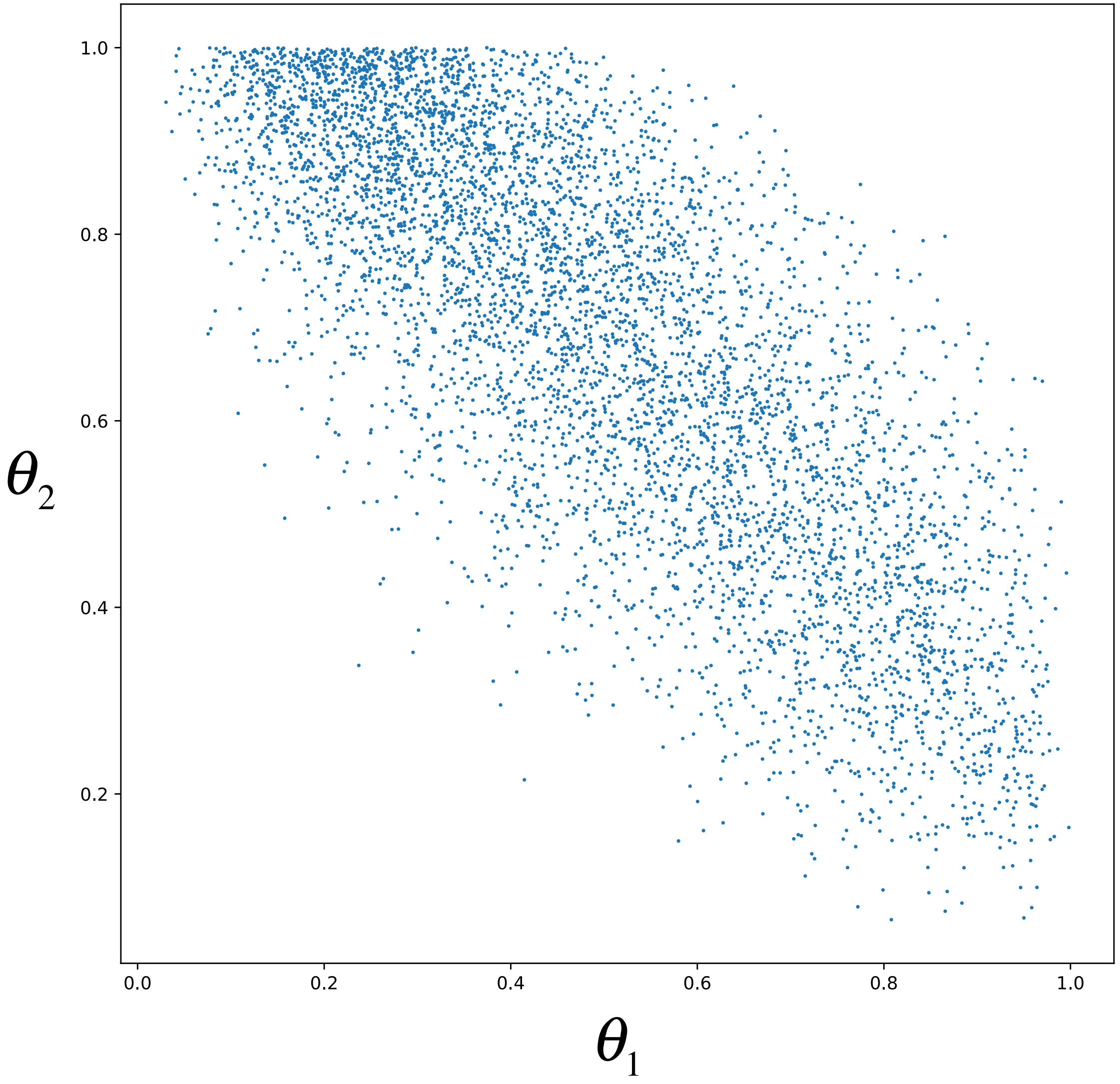
	1	**2**	**3**
n_{i1}	5	6	4
n_{i2}	5	4	6
y_i	7	5	6

Example of implementation in Python (see Jupyter notebook)

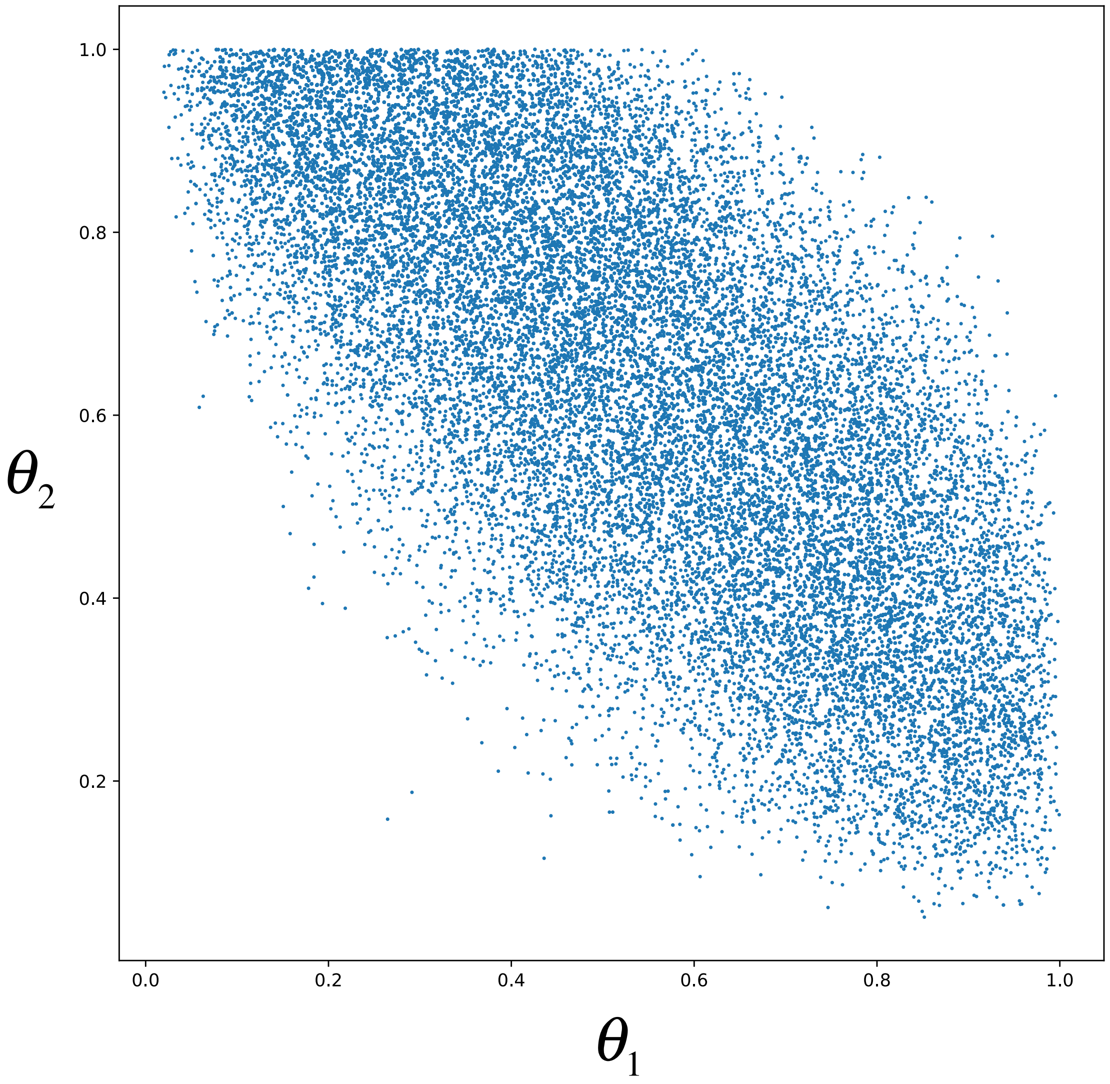


prior distribution
(50000 samples, uniform in 2D parameter space)

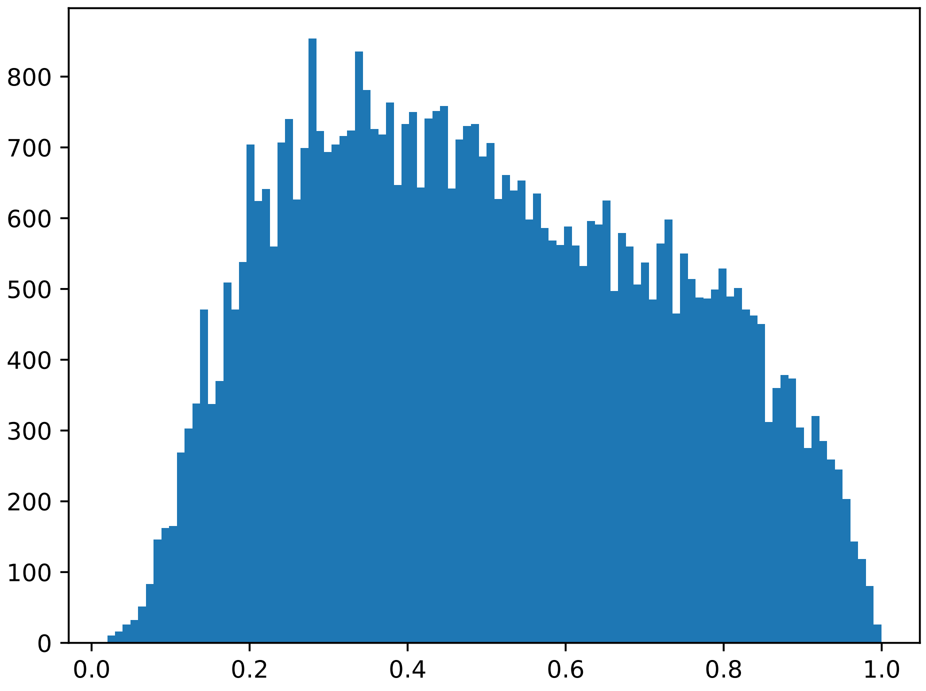
Posterior as a resampled prior using acceptance-rejection



Posterior as a resampled prior using weighted bootstrap

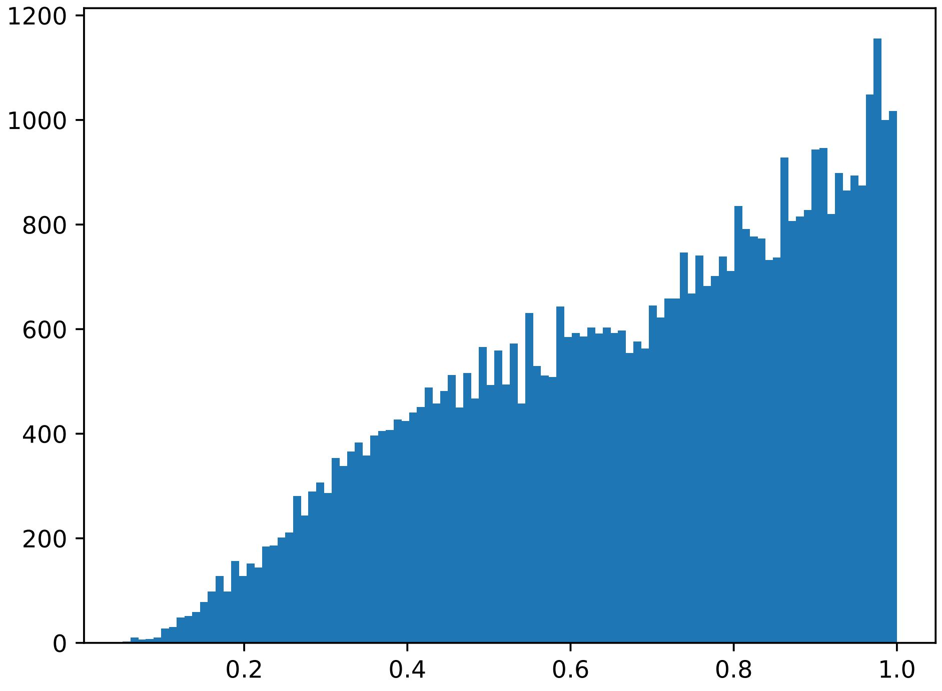


The resampled points are representative of the posterior distribution and can be used to evaluate any sample estimate



Marginalized histogram of θ_1

Sample mean: 0.500±0.001



Marginalized histogram of θ_2

Sample mean: 0.677±0.001

... these calculational methodologies have also had an impact on theory. By freeing statisticians from dealing with complicated calculations, the statistical aspects of a problem can become the main focus.

Casella & George, in their description of the Gibbs sampler. *Am. Stat.* **46** (1992) 167

5. Very short introduction to Markov chains

Consider a system such that

- the system can occupy a finite or countably infinite set of states S_n;
- the system changes state randomly at discrete times $t = 1, 2, \ldots$;
- if the system is in state S_i, then the probability that the system goes into state S_j is

$$p_{ij} = P[S(n+1) = S_j | S(n) = S_i] \qquad i, j = 1, 2, \ldots$$

  i.e., this probability depends only on the previous state, and is independent o all previous states (this is the *Markov property*);

- the *transition probabilities* p_{ij} do not depend on time n.

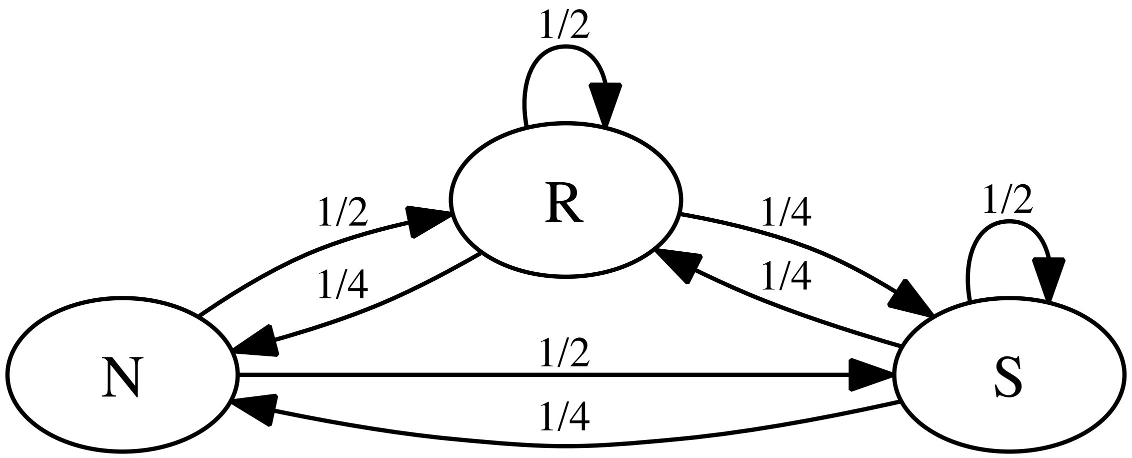
Such a system is a special type of discrete time stochastic process, which is called *Markov chain*.

Example:

in the Land of Oz they never have two nice days in a row, rather, after a sunny day it either rains or snows.

If they have a nice day, they are just as likely to have snow as rain the next day. If they have snow or rain, they have an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day. When we denote the three states with the symbols N (Nice), R (Rain), or S (Snow), the transition probabilities are:

$$\begin{array}{lll} p_{NN} = 0; & p_{NR} = 1/2; & p_{NS} = 1/2 \\ p_{RN} = 1/4; & p_{RR} = 1/2; & p_{RS} = 1/4 \\ p_{SN} = 1/4; & p_{SR} = 1/4; & p_{SS} = 1/2 \end{array}$$



(representation as a directed graph)

Matrix of transition probabilities (also called *transition kernel*)

$$\mathbf{P} = \begin{pmatrix} p_{NN} & p_{NR} & p_{NS} \\ p_{RN} & p_{RR} & p_{RS} \\ p_{SN} & p_{SR} & p_{SS} \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

This is a *row stochastic matrix*, where all rows are such that

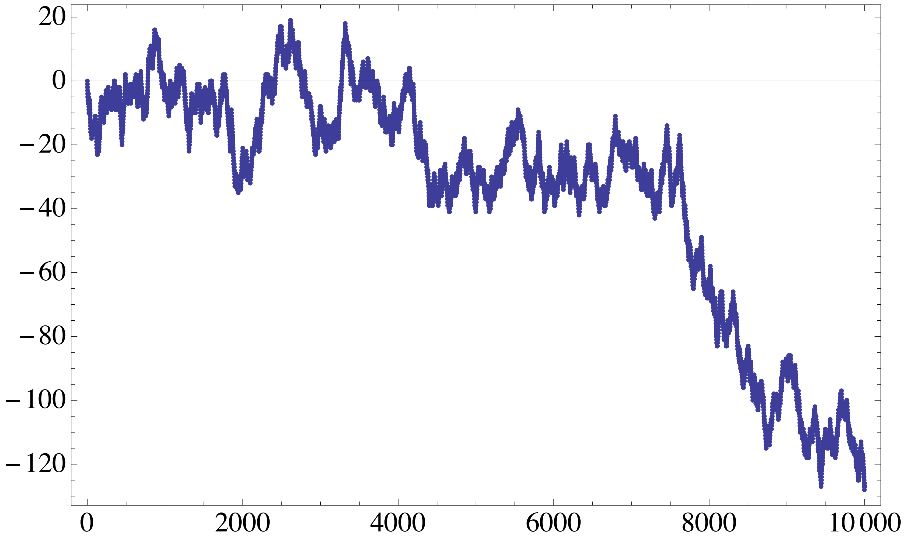
$$\sum_j p_{ij} = 1$$

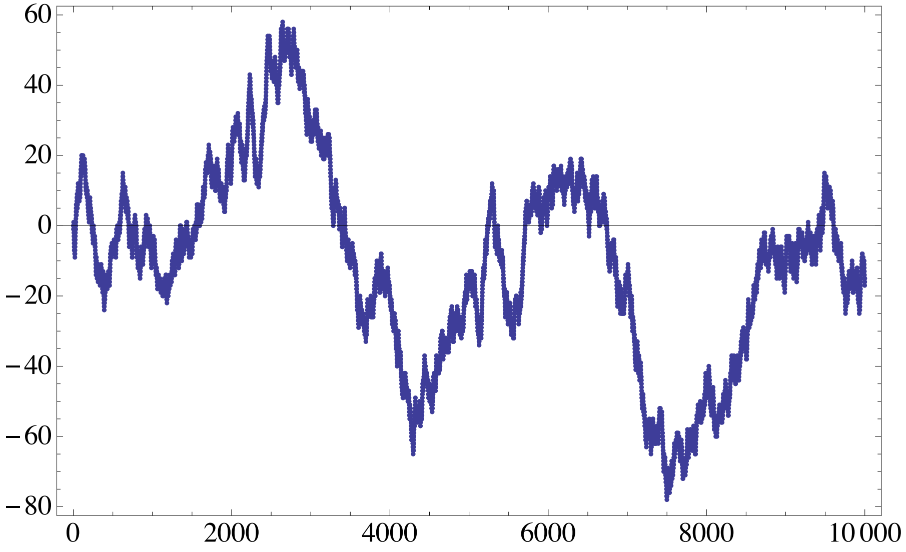
There are also *column stochastic matrices*, and *doubly stochastic matrices* that are necessarily square:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = \sum_{i=1}^{n} 1 = n$$

$$\sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{m} 1 = m$$

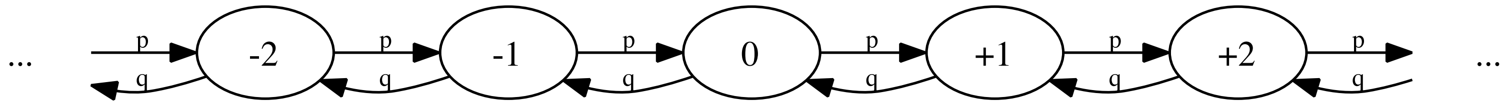
$$\Rightarrow \quad m = n$$





Discrete-time discrete-space random walks are an example of Markov chains with infinite states.

$$p_{i,i+1} = p; \quad p_{i,i-1} = q$$



Now let

$$\pi_i^{(n)} = P[S(n) = S_i]$$

be the probability that at time *n* the system is in state S_i , then:

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

When we define the vector $\boldsymbol{\pi}^{(n)} = \{\pi_j^{(n)}\}$ and the matrix $\mathbf{P} = \{p_{ij}\}$ we see that the equation becomes

$$\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)} \mathbf{P}$$

$$\Rightarrow \quad \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \mathbf{P}^n$$

n-step transition kernel

For example, the transition kernels for the weather in the Land of Oz are

$$\mathbf{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.375 & 0.4375 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} 0.199219 & 0.400391 & 0.400391 \\ 0.200195 & 0.400391 & 0.399414 \\ 0.200195 & 0.399414 & 0.400391 \end{pmatrix}$$

$$\mathbf{P}^{10} = \begin{pmatrix} 0.200001 & 0.4 & 0.4 \\ 0.2 & 0.400001 & 0.4 \\ 0.2 & 0.4 & 0.400001 \end{pmatrix}$$

the transition kernels seem to converge to a fixed matrix ...

$$\mathbf{P}^{20} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

$$\mathbf{P}^{100} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

Notice that if the transition kernel converges to a fixed matrix where all rows are equal, then the distribution of states also converges to a fixed distribution which does not depend on the initial distribution:

$$\mathbf{P}^n \xrightarrow[n\to\infty]{} \mathbf{P}_\infty \qquad (\mathbf{P}_\infty)_{i,j} = f_j$$

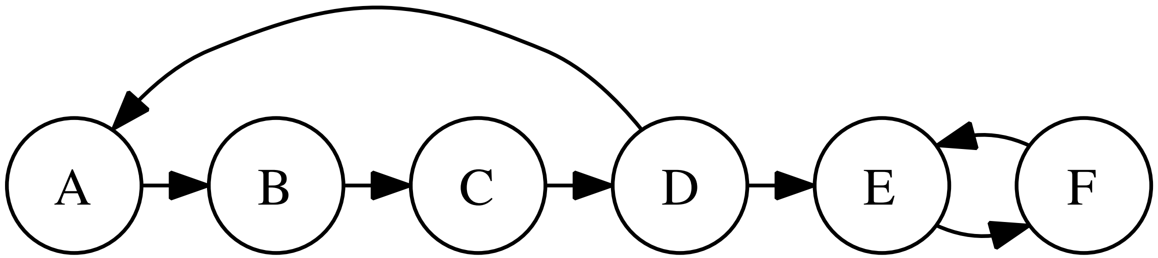
all rows equal

⇓

$$\pi_j^{(\infty)} = \sum_i \pi_i^{(0)} (\mathbf{P}_\infty)_{i,j} = \sum_i \pi_i^{(0)} f_j = f_j$$

Persistent and transient states ...

Type of state	Definition of state (assuming, where applicable, that the state is initially occupied)
Periodic	Return to state possible only at times t, $2t$, $3t$, ..., where $t > 1$
Aperiodic	Not periodic
Recurrent/Persistent	Eventual return to state certain
Transient	Eventual return to state uncertain
Ephemeral	Is a state j such that $p_{ij} = 0$ for every i
Positive-recurrent	Recurrent/persistent, finite mean recurrence time
Null-recurrent	Recurrent, infinite mean recurrence time
Ergodic	Aperiodic, positive-recurrent



This graph represents the states and the transition probabilities of a finite Markov chain with 6 states.

The arrows correspond to nonzero transition probabilities. If the chain starts with any one of states A, B, C or D, it can loop around these four states until a transition D to E occurs, then the system is locked in the E-F loop.

States A, B, C, and D are transient, while states E and F are persistent (and periodic, with period 2). A Markov chain with just one class, such that all states communicate, is said to be irreducible. This Markov chain is not irreducible.

VERY INTERESTING MATH ON PERSISTENT STATES, HOWEVER WE DO NOT PURSUE IT FURTHER, WE DO NOT NEED IT NOW.

Limiting probabilities and stationary distributions

Here we prove that the convergence that we saw in the Land of Oz example is a general feature of Markov chains, under the assumption that the chain is irreducible, and that for some N we have

$$\min_{i,j} p_{ij}^{(N)} = \delta > 0$$

Now let

$$r_j^{(n)} = \min_i p_{ij}^{(n)}; \quad R_j^{(n)} = \max_i p_{ij}^{(n)}$$

be the min and max of the j-the column vector in the n-step transition matrix.

Recall the example:

$$\mathbf{P}^2 = \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.375 & 0.4375 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} 0.199219 & 0.400391 & 0.400391 \\ 0.200195 & 0.400391 & 0.399414 \\ 0.200195 & 0.399414 & 0.400391 \end{pmatrix}$$

$$\mathbf{P}^{10} = \begin{pmatrix} 0.200001 & 0.4 & 0.4 \\ 0.2 & 0.400001 & 0.4 \\ 0.2 & 0.4 & 0.400001 \end{pmatrix}$$

$$\mathbf{P}^{20} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

$$\mathbf{P}^{100} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

we shall show that, in each column, the min and the max become closer and closer as n grows and bracket a value that is the asymptotic matrix element (the same for all rows in a given column)

Then we find

$$r_j^{(n+1)} = \min_i p_{ij}^{(n+1)} = \min_i \mathbf{P}_{ij}^{n+1} = \min_i (\mathbf{P}\mathbf{P}^n)_{ij} = \min_i \sum_k p_{ik} p_{kj}^{(n)}$$
$$\geq \min_i \sum_k p_{ik} r_j^{(n)} = r_j^{(n)}$$

and

$$R_j^{(n+1)} = \max_i p_{ij}^{(n+1)} = \max_i \mathbf{P}_{ij}^{n+1} = \max_i (\mathbf{P}\mathbf{P}^n)_{ij} = \max_i \sum_k p_{ik} p_{kj}^{(n)}$$
$$\leq \max_i \sum_k p_{ik} R_j^{(n)} = R_j^{(n)}$$

This means that, as *n* grows, the minimum and the maximum values in a column vector get closer and closer (the components of the column vector get closer and closer). *But do they converge to the same value ???*

We must consider the difference

$$R_j^{(n)} - r_j^{(n)} = \max_i p_{ij}^{(n)} - \min_k p_{kj}^{(n)} = \max_{i,k} \left[p_{ij}^{(n)} - p_{kj}^{(n)} \right]$$

Then, shifting the difference by N, we find

$$R_j^{(n+N)} - r_j^{(n+N)} = \max_{i,k} \left[p_{ij}^{(n+N)} - p_{kj}^{(n+N)} \right] = \max_{i,k} \left\{ \sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)} \right] p_{lj}^{(n)} \right\}$$

Next we split the difference enclosed in braces into sums of negative and positive contributions

$$\begin{aligned} \sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)} \right] p_{lj}^{(n)} &= \sum_l^{+} [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} + \sum_l^{-} [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} \\ &\leq \sum_l^{+} [p_{il}^{(N)} - p_{kl}^{(N)}] R_j^{(n)} + \sum_l^{-} [p_{il}^{(N)} - p_{kl}^{(N)}] r_j^{(n)} \end{aligned}$$

Now consider the structure of the positive sum, it must contain at least one term where one subtracts the smallest element in the column, so that

$$\sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] = \sum_l^+ p_{il}^{(N)} - \sum_l^+ p_{kl}^{(N)} \leq \sum_l p_{il}^{(N)} - \delta = 1 - \delta$$

Similarly, for the negative sum we find

$$\sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] = \sum_l^- p_{il}^{(N)} - \sum_l^- p_{kl}^{(N)} \geq \delta - \sum_l p_{kl}^{(N)} = -(1 - \delta)$$

and therefore

$$\begin{aligned} \sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)} \right] p_{lj}^{(n)} &\leq \sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] R_j^{(n)} + \sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] r_j^{(n)} \\ &\leq (1 - \delta) R_j^{(n)} - (1 - \delta) r_j^{(n)} = (1 - \delta)(R_j^{(n)} - r_j^{(n)}) \end{aligned}$$

so that taking strides of N steps at a time, and recalling that $0 < 1 - \delta < 1$

$$R_j^{(kN)} - r_j^{(kN)} < (1 - \delta)^k \left[R_j^{(N)} - r_j^{(N)} \right] \xrightarrow[k \to \infty]{} 0$$

Since

$$R_j^{(kN)} - r_j^{(kN)} < (1-\delta)^k \left[R_j^{(N)} - r_j^{(N)} \right] \xrightarrow[k \to \infty]{} 0$$

the matrix elements in the *j*-th column converge to a single value p_j^*, i.e.,

$$p_{ij}^* = \lim_{n \to \infty} [\mathbf{P}^n]_{ij} = p_j^*$$

and

$$\pi_j^* = \sum_k \pi_k^{(0)} p_{kj}^* = \sum_k \pi_k^{(0)} p_j^* = p_j^*$$

This asymptotic distribution is stable, indeed from

$$\pi_j^{(n)} = \sum_k \pi_k^{(n-1)} p_{kj}$$

we find

$$[\pi^* \mathbf{P}]_j = \sum_k \pi_k^* p_{kj} = \sum_k p_k^* p_{kj} = \sum_k p_{ik}^* p_{kj} = p_{ij}^* = p_j^* = \pi_j^*$$

or, in matrix form

$$\pi^* = \pi^* \mathbf{P}$$

i.e., the asymptotic probability vector is the left eigenvector with eigenvalue 1 of the transition probability matrix. The distribution expressed by the probability vector π^* is called *invariant distribution* or *stationary distribution*.