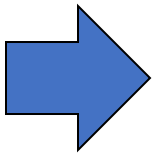
Introduction to Bayesian Statistics - 7

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. 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)

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.

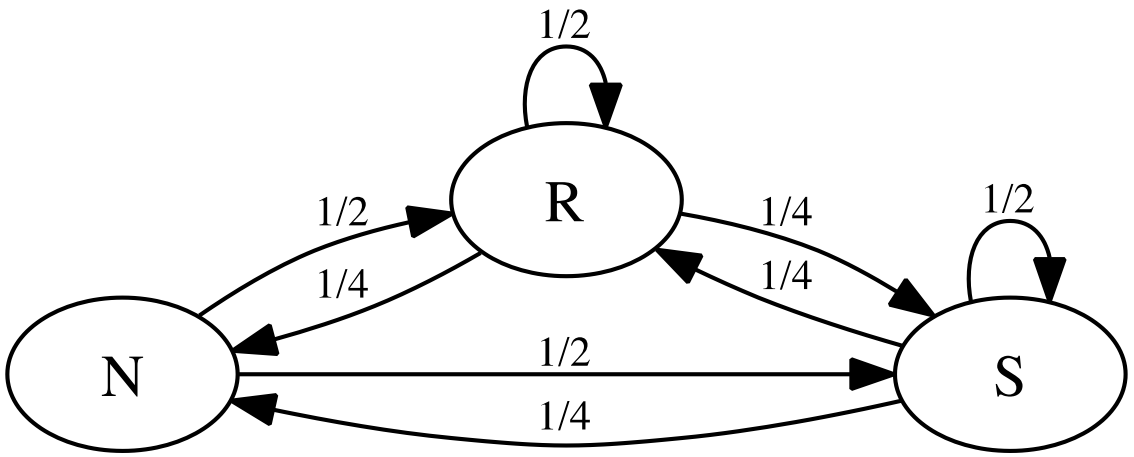
Such a system is a special type of discrete time stochastic process, which is called *Markov chain*.

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

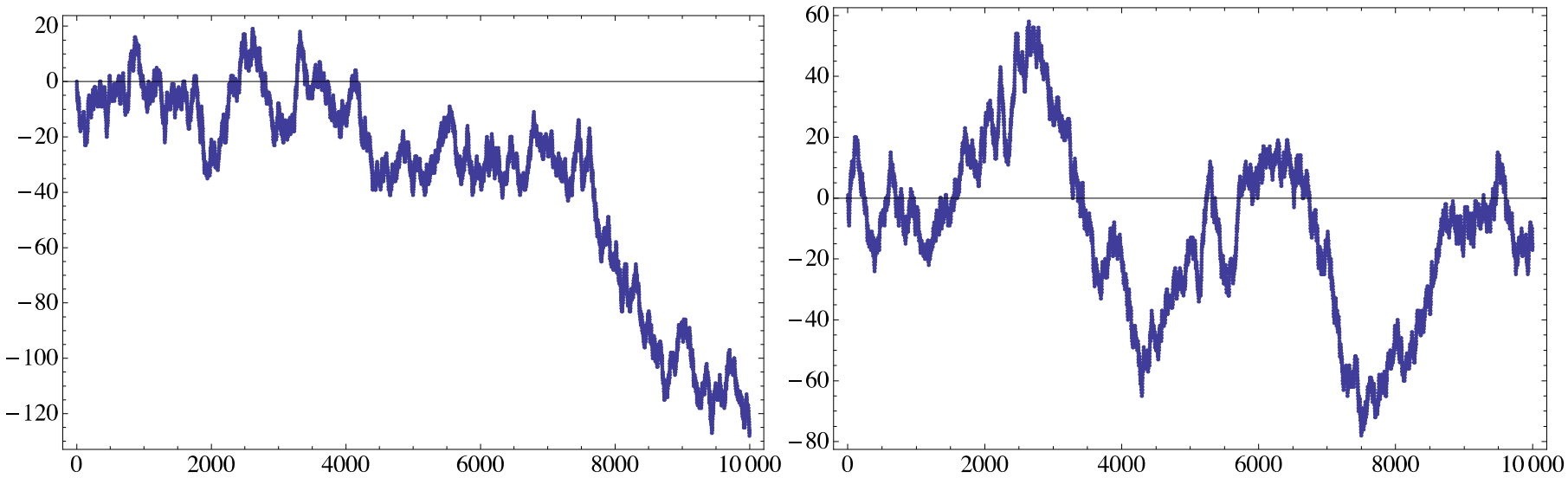
$$\sum_j p_{ij} = 1$$

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

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

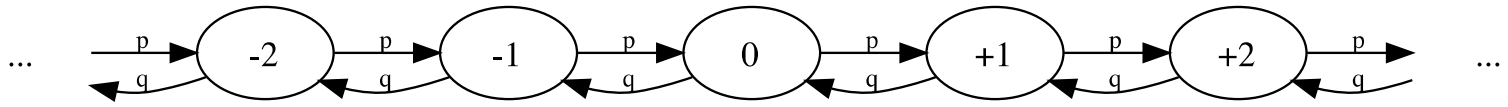
$$\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{m} 1 = m$$

$$m = n$$



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

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

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

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:

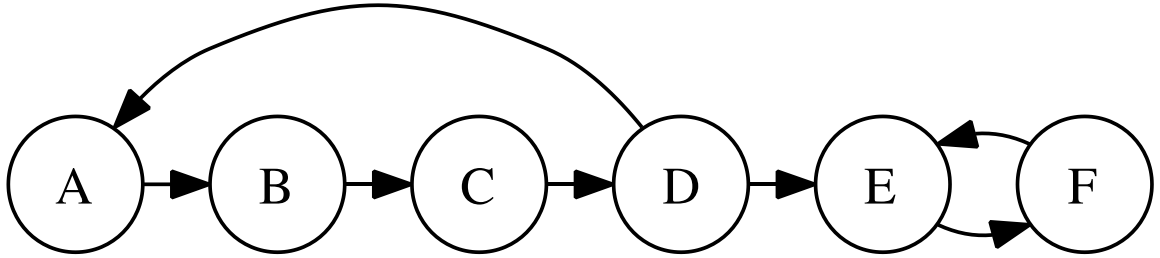
$$\mathbf{P}^n \xrightarrow[n\to\infty]{} \mathbf{P}_\infty \qquad (\mathbf{P}_\infty)_{i,j} = f_j$$

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*.

6. Detailed balance and Boltzmann's H-theorem

From the definition of conditional probabilities we find

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

therefore, when a Markov chain is time reversed we find

$$P[S(n) = S_i | S(n+1) = S_j]$$
$$= P[S(n+1) = S_j | S(n) = S_i] \frac{P[S(n) = S_i]}{P[S(n+1) = S_j]}$$

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

$$H \sim -H_B$$

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

Using the master equation we find a differential equation for H

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

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

Exchanging indexes ...

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

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)

Explaining the Gibbs Sampler

RGE*

applications of the Gibbs sampler have been in Bayesian models, it is also extremely useful in classical (likelihood) calculations [see Tanner (1991) for many examples]. Furthermore, 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. This point is wonderfully illustrated by Smith and Gelfand (1992).

In the next section we describe and illustrate the application of the Gibbs sampler in bivariate situations.

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

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

$$Y_{j+1}' \sim f(y \mid X_j' = x_j'). \tag{2.4}$$

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 $f(x)$.

The convergence (in distribution) of the Gibbs sequence (2.3) can be exploited in a variety of ways to obtain an approximate sample from $f(x)$. For example, Gelfand and Smith (1990) suggest generating m independent Gibbs sequences of length k, and then using the final value of X_k' from each sequence. If k is chosen large enough, this yields an approximate iid sample from $f(x)$. Methods for choosing such k, as well as alternative approaches to extracting information from the Gibbs sequence, are discussed in Section 5. For the

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

of the marginal distribution $f(x)$ of X. T

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

For the following joir

(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.

can be obtained analytically from (2.5) as

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

$$x = 0, 1, \ldots, n, \quad (2.7)$$

black = Gibbs sampling
white = theoretical expectation

the beta-binomial distribution. Here, characteristics of $f(x)$ can be directly obtained from (2.7), either analytically or by generating a sample from the marginal and not fussing with the conditional distributions. However, this simple situation is useful for illustrative purposes. Figure 1 displays histograms of two samples $x_1, \ldots, x_m$ of size $m = 500$ from the beta-binomial distribution of (2.7) with $n = 16$, $\alpha = 2$, and $\beta = 4$.

The two histograms are very similar, giving credence to the claim that the Gibbs scheme for random variable generation is indeed generating variables from the marginal 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

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)