Munich University of Technology (TUM)
Centre for Mathematical Sciences

Statistical Methods in Genetics

Volker Hösel

WS 02/03
SS 03

Contents

1 Introduction

The completion of the human genom project gives another major incentive for genetic research. New biotechnology companies emerge and high expectations prevail that genetics could for example help to fight diseases or aid efficient food production.

The mere amount of data collected in gene data bases and the variability of biological phenomena call for mathematical and statistical modelling.

From these models conclusions can be derived and hypotheses generated. These have then to be validated with already collected data or new designed experiments.

A lot of rather advanced statistical techniques are employed in connection with genetic research. The aim of our lectures is to introduce some of the most important ones and show how they work.

1.1 Rough outline of the lectures

- Fundamental Methods:
  Basics, Bayesian paradigm Markov chains, Gibbs sampler
- Hidden Markov Models:
  Architecture, Inference
- Biological Sequences:
  Alignment Methods, Gene Finding

1.2 Literature

Warren J.Ewens, Gregory R. Grant: Statistical Methods in Bioinformatics. Statistics for Biology and Health, Springer, New York 2001.

Kenneth Lang: Mathematical and Statistical Methods for Genetic Analysis (2nd edition). Statistics for Biology and Health, Springer, New York 2001.

Rick Durrett: Probability Models for DNA Sequence Evolution. Probability and Its Applications, Springer, New York 2002.

M.Elizabeth Halloran, Seymour Geisser: Statistics in Genetics. The IMA Volumes in Mathematics and its Applications, Voliume 112, Springer, New York 1999.

Richard Durbin et al.: Biological sequence analysis, Cambridge university press 1998.

Paul Berg, Maxine Singer: Die Sprache der Gene. Grundlagen der Molekulargenetik, Spektrum Akademischer Verlag, Heidelberg, 1993

1.3 Basic facts from biology

The genetic information is contained in the nucleic acid DNA which sequences of the nucleotides **adenine** (A), **cytosine** (C), **guanine** (G) and **thymine** (T). These sequences group together forming double strands of complementary base pairs (A and T or C and G) like:

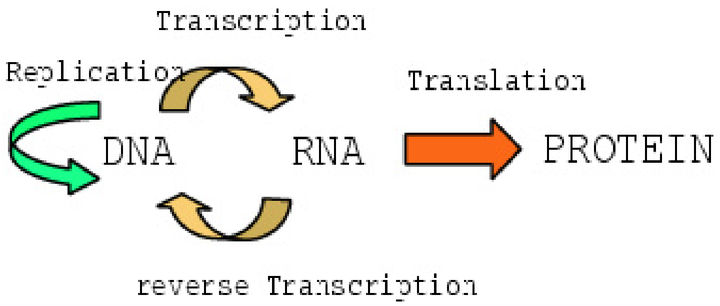
$$\begin{array}{ccccccccc} 5' & C & C & T & G & A & T & T & 3' \\ & | & | & | & | & | & | & | & \\ 3' & G & G & A & C & T & A & A & 5' \end{array}$$

The direction of a sequence is indicated by the ends 5' and 3'. The symbol | stands for a hydrogen bond. The human genom consists of about $3 \cdot 10^9$ letters of the alphabet $\mathbf{A} = \{A,C,G,T\}$. These are grouped in 46 **chromosomes**: 22 pairs of **homologous** chromosomes and two sex chromosomes (male: XY, female XX).

Genes are regions in the genom which induce the production of specific parts of a cell (especially **proteins**). In most cases this, so called, **gene expression** runs in two steps:

1. **Transcription**: The DNA double strand is separated in the gene region. One strand serves as template for the composition of an RNA strand (m-RNA). RNA is a nucleic acid very similar to DNA. It is composed from the alphabet $\mathbf{A} = \{A,C,G,U\}$ with thymine of the DNA substituted by uracil (U).

2. **Translation**: Small strands of t-RNA each connected to one of the 20 amino acids string along the m-RNA, thereby producing proteins. A group of 3 successive nucleic acids (**codons**) on the m-RNA (or t-RNA) represents one amino acid (this correspondence is the **genetic code**).



Currently one estimates that only about 10^{-2} of the genom corresponds to genes.

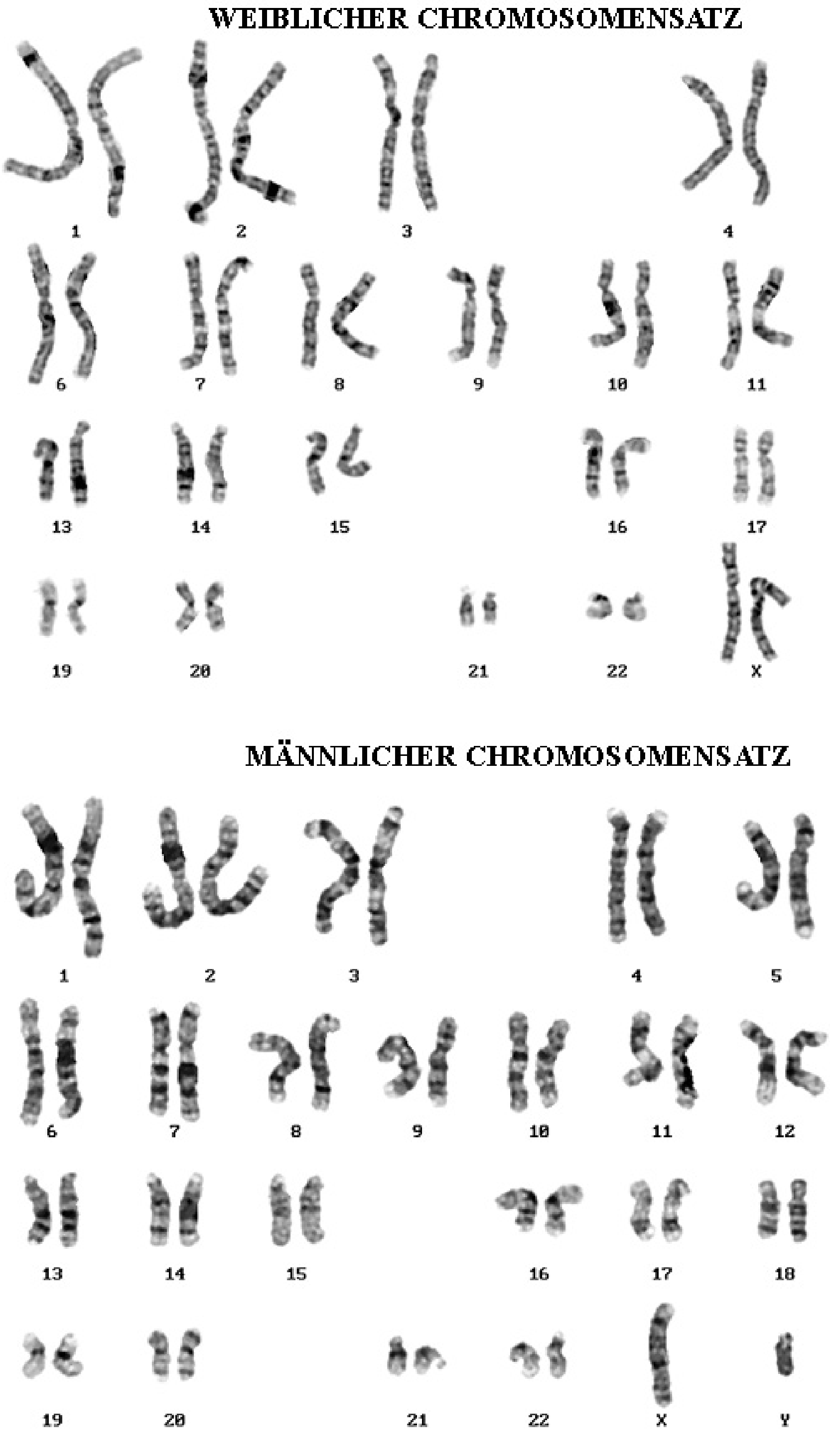


Figure1: Female and male **karyogram**.

Homologous chromosomes are paired. Notice the differences of the genders regarding the sex chromosomes X and Y.

1.4 Mendel's laws

As starting date of the science of genetics usually the year 1865 is seen:

Mendel, J.G. (1865). Verhandlungen des naturforschlichen Vereines in Brünn, Abhandlungen 4, 3-47.

The Bohemian monk Mendel designed experiments with seven discrete traits of garden peas. From his observations he derived the following rules.

Mendel's first law (law of segregation)

It describes the inheritance of a single trait (for example: blood group):

Such a trait is determined by a discrete inherited factor (gene). Genes occur in different variants called **alleles** (blood group alleles: A, B, 0). The **genotype** of an individual is determined by (usually not ordered) pairs of alleles (blood group genotypes: AA, A0, 00, B0, BB). When **gametes** (egg/ovum or sperm/pollen) are formed (**meiosis**) the pairs **segregate** resulting in gametes including only single alleles. Gametes from male and female parents fuse to form a **zygote** and hereby restore the doubling factors. Genes and their alleles persist unchanged in the generations. All alleles from a genotype have **equal chance** of being passed to an offspring.

Mendel's second law (law of independence)

Describes the joint behavior of loci controlling different traits:

Alleles at **different loci** segregate **independently**.

Remarks 1.1 *a) The pairing of alleles might not occur on sex chromosomes. Some genes are located on X chromosomes only and this is the reason for X-linked diseases like red-green blindness or hemophilia.*

b) Mendel's second law is valid for genes on different chromosomes. Genes on the same chromosomes can be regarded as independent only if they are spatially far apart. Dependence and location is explored in the ***linkage analysis****.*

2 Fundamental Methods

In this section we review the more fundamental methods and their applications.

2.1 Basic definitions from probability theory

Probability models are used to model complex and heterogeneous biological interactions. These models allow to derive quantitative estimations or qualitative conclusions.

Definition 2.1

- *A system* F *of subsets of a set* Ω *is a* $\sigma - field$ *if it satisfies:*
  1. $\emptyset, \Omega \in \mathrm{F}$.
  2. For all $A \in \mathrm{F}$ *is* $A^c := \{\omega \in \Omega : \omega \notin \mathrm{A}\}$ *in* F.
  3. If $A_i \in \mathrm{F}$ *for all* $i = 1, 2, 3, ...$ *then so is* $\bigcup_{i=1}^{\infty} A_i$.
- *A* ***probability measure*** P *on a* σ*-field is a function* $P : \mathrm{F} \longrightarrow [0, 1]$ *with:*
  1. $P(\emptyset) = 0, P(\Omega) = 1$.
  2. If $A_1, A_2, ...$ *are pairwise disjoint (*$A_i \cap A_j = \emptyset$ *for* $i \neq j$*) then*
  $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.
- *A* ***probability space*** *is a triple* (Ω, F, P) *with a set* Ω*, a* σ *-field* F *of subsets of* Ω *and a probability measure* P *defined on* F.
- *A* ***real valued random variable*** X *is a function* $X : \Omega \longrightarrow \mathbb{R}$
  with $X^{-1}((-\infty, c)) \in \mathrm{F}$ *for every* $c \in \mathbb{R}$.

Remarks 2.2 F *is called the algebra of events to which* P *assigns a probability.*

$\{A : A \subseteq \Omega\}$ *is always a* σ*-field and can be used for* ***finite*** Ω *to define a probability space.*

For "large" Ω *(for example the real numbers) the class of probability measures defined on all subsets does not provide a satisfying consistent theory. Thus smaller families of subsets (the* σ *-fields* F*) are introduced.*

Remember that two events $A, B \in \mathrm{F}$ *are called* ***(statistically) independent****, if*
$P(A \bigcup B) = P(A)P(B)$.

Mendel's first law implies the independence of the two alleles in a genotype:
$p(AA) = p(A)^2, \quad p(Aa) = p(A)p(a), \quad p(aa) = p(a)^2$.

2.2 Example of probability modelling : Hardy-Weinberg equilibrium I

Consider an autosomal locus with two alleles A and a. The possible genotypes define $\Omega = \{AA, Aa, aa\}$.

How is the frequency (probability) of a genotype pertained in a population?

Assumptions for the simplest model and the consequences for probability:

- *Infinite population size:* use frequencies to determine probabilities.
- *Random mating:* independence of genotypes.
- *No selection:* equal chance of genotypes to produce offsprings (equal **fitness**).
- *discrete generations:* parent generations do not produce offsprings in the grandchild generation.

Given the frequency of the phenotypes at a starting point:

$p_0(AA) = u_0$, $p_0(Aa) = v_0$ and $p_0(aa) = w_0$.

Then the mating outcomes in the following generation are

mating type	offsprings	frequency
AA × AA	AA	u_0^2
AA × Aa	1/2 AA + 1/2 Aa	$2u_0v_0$
AA × aa	Aa	$2u_0v_0$
Aa × Aa	1/4 Aa + 1/2Aa + 1/4 aa	v_0^2
Aa × aa	1/2 Aa+ 1/2 aa	$2u_0v_0$
aa × aa	aa	w_0^2

From the above assumptions one can calculate the next generation:

$$u_1 = u_0^2 + u_0v_0 + \tfrac{1}{4}v_0^2 = (u_0 + \tfrac{1}{2}v_0)^2$$
$$v_1 = u_0v_0 + 2u_0w_0 + \tfrac{1}{2}v_0^2 + v_0w_0 = 2(u_0 + \tfrac{1}{2}v_0)(w_0 + \tfrac{1}{2}v_0)$$
$$w_1 = \tfrac{1}{4}v_0^2 + v_0w_0 + w_0^2 = (w_0 + \tfrac{1}{2}v_0)^2$$

Defining the frequency of the alleles A and a according to Mendel's first law as

$$p_A = u_0 + \tfrac{1}{2}v_0$$
$$p_a = w_0 + \tfrac{1}{2}v_0$$

one finds for the next generation stabilization:

$$u_2 = (p_A^2 + \tfrac{1}{2}2p_Ap_a)^2 = (p_A(p_a + p_A))^2 = p_A^2 = u_1$$
$$v_2 = 2(p_A^2 + \tfrac{1}{2}2p_Ap_a)(p_a^2 + \tfrac{1}{2}2p_Ap_a) = p_Ap_A = v_1$$
$$w_2 = (p_a^2 + \tfrac{1}{2}2p_Ap_a)^2 = (p_a(p_a + p_A))^2 = p_a^2 = u_1$$

Remarks 2.3

- *The Hardy-Weinberg law states that any next generation is already stable.*
- *For alleles on sex chromosomes one can show that equilibrium will be attained only asymptotically.*
- *Allele frequency of recessive genes can be assessed:*
  For recessive a the genotype aa is observable from its phenotype. The independence condition $p_{aa} = (p_a)^2$ then shows that p_a can be estimated by (frequency of the phenotype of aa)$^{1/2}$.
- *Hardy-Weinberg law is also valid for genes with many alleles: $p_{ij} = (2-\delta_{i,j})p_i p_j$*
- *One can show that stability of frequencies (probabilities) is equivalent to*

  $$p_{ij}^2 = 4p_{ii}p_{jj}.$$

- *Weakening the above assumptions may cause differing results for the population dynamics. The adequacy of assumptions has thus to be investigated carefully.*

The case with selection (results only): Now the genotypes have different **fitness** denoted by w_{AA}, w_{Aa}, and w_{aa}.

Definition 2.4 *The **mean fitness** $\bar{w}$ of a population regarding two alleles A and a with frequency p_A and p_a is defined by*

$$\bar{w}(p_A, p_a) = w_{AA}p_A^2 + 2w_{Aa}p_A p_a + w_{aa}p_a^2$$

The frequencies in the following generation are then given by

$$p_{A,1} = \frac{w_{AA}p_A^2 + w_{Aa}p_A p_a}{\bar{w}(p_A, p_a)}, \qquad p_{a,1} = \frac{w_{aa}p_a^2 + w_{Aa}p_A p_a}{\bar{w}(p_A, p_a)}$$

which allows to derive a **main theorem of population genetics**:

$$\bar{w}(p_{A,1}, p_{a,1}) \geq \bar{w}(p_A, p_a).$$

The fitness increases with the generations eventually attaining one of the three stable points: $p_A = 0$ or $p_a = 0$ or $p_A(w_{AA} - w_{Aa}) = p_a(w_{aa} - w_{Aa})$.
The first two alternatives indicate **survival of the fittest** (alleles with low fitness die out). The third alternative is a stable point (convergence from all points in a neighborhood) only if $w_{aA} \geq w_{AA}, w_{aa}$. Coexistence of all genotypes is thus restricted to the case of **heterozygote advantage**. This mechanism is believed to maintain several recessive diseases at high frequencies. For example, a single dose of sickle cell gene protects against malaria.

2.3 Basic statistical notations

To set our notations we recall statistical definitions which are required later.

Definition 2.5

- *A* ***probability distribution*** *(on the real line) is a function $F : \mathbb{R} \to [0,1]$ satisfying:*
  i) $F(x)$ is non decreasing;
  ii) $F(-\infty) = \lim_{x\to-\infty} F(x) = 0, \quad F(+\infty) = \lim_{x\to+\infty} F(x) = 1$;
  iii) F(x) is continuous on the right and has a limit on the left at each $x \in \mathbb{R}$.

- *F is called* ***discrete*** *if it is piecewise constant. With $\triangle F(x_k)$ the height of a jump at point x_k one obtains by $P(x_k) = \triangle F(x_k)$ probability assignments for the jump points.*

- *A non negative function f satisfying $F(x) = \int_{-\infty}^{x} f(t)dt$ for all $x \in \mathbb{R}$ is called* ***density*** *of the distribution function F.*

- *A real valued random variable X defined on a probability space (Ω, F, P) is* ***distributed according to*** *F if*

  $$P(X \le x) := P(\omega : X(\omega) \le x) = F(x).$$

  A random variable is called discrete or continuous if its distribution has this property.

- *The* ***mean*** *$\mu := E(X)$ and the* ***variance*** *$\sigma^2 := E(X-\mu)^2$ is defined for a*

  - *discrete random variable X as*

    $$E(X) = \sum_k x_k P(X = x_k);$$

    $$E(X-\mu)^2 = \sum_k (x_k - \mu)^2 P(X = x_k)$$

    and for a

  - *real valued continuous variable X with density f as*

    $$E(X) = \int_{-\infty}^{\infty} x f(x) dx;$$

    $$E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx.$$

Comments:

a) A density function f may not exist for a distribution function F.
b) Neither the mean nor the variance may exist for a given random variable.

We give examples which are most important in biological applications.

- Given a sequence of n identical, independent (Bernoulli) trials with two possible outcomes (”success”, ”failure”) and the probability p of ”success” in any such trial. The numbers

  $$B(n, p\ )(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

  define the discrete **binomial distribution** assessing the probability of k ”successes” in the sequence of trials. The mean is np and the variance is $np(1-p)$.

- The random variable X counting the ”successes” in a sequence of independent Bernoulli trials before the first ”failure” occurs is distributed according to the **geometric distribution**:

  $$P(X = n) = (1-p)p^n \tag{2}$$

  The corresponding distribution function can be calculated therefrom (exercise!) as

  $$F(x) = P(X \leq n) = 1 - p^{n+1}$$

  The mean is $p/(1-p)$ and the variance is $p/(1-p)^2$.

- BLAST (Basic Local Alignment Search Tool, Altschul & al., 1990) algorithms employ random variables which behave asymptotically like geometric distributed ones: A variable X defined on non negative integers $0, 1, 2, \ldots$ is **geometric-like** if

  $$\lim_{k\to\infty} \frac{P(X \geq k)}{Cp^k} = 1 \tag{3}$$

  for some fixed constant C with $0 < C < 1$.

- An important discrete distribution is the **Poisson distribution** given by

  $$P(\lambda)(k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{4}$$

  It gives the probability of k events in a time interval, if an average of λ observations can be expected in such an interval. Formally such distributions may be derived within the theory of Poisson processes and queuing theory. Here the mean and the variance are both λ.

- If one records the waiting time in the above Poisson setting one gets a continuous random variable which is **exponentially distributed**. It has

$$f(x) = \lambda e^{-\lambda x} \tag{5}$$

with $x \geq 0$ as density function. This gives by integration $F(x) = 1 - e^{-\lambda x}$. Further one gets $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.

- The exponential distribution is a special case of the **gamma distribution** which has a density function

$$f_{(\lambda,k)}(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}, \quad x > 0. \tag{6}$$

The gamma distribution has two arbitrary positive parameters λ and k and covers other important distributions. For $k = 1$ one gets the exponential distribution and for $\lambda = 1/2$ and $k = 1/2\ \nu$ with a positive integer ν the chi-square distribution with ν degrees of freedom. The mean μ and the variance σ^2 are given by

$$\mu = \frac{k}{\lambda}; \quad \sigma^2 = \frac{k}{\lambda^2}$$

.

- Finally we mention the familiar **Gaussian distribution**. The density is

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{7}$$

where the mean μ and the variance σ^2 are parameters for the density. The usual notation for stating that a random variable X is distributed according to a Gaussian distribution is: $X \sim N(\mu, \sigma^2)$

Definition 2.6

- *The **joint distribution** F of n random variables $X_1, \ldots, X_n$ defined on the same probability space is given by*

$$F(x_1, \ldots, x_n) = P(\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n)$$

where $x_1, \ldots, x_n$ are possible values of the random variables.

- *$X_1, \ldots, X_n$ are **independent** if*

$$F(x_1, \ldots, x_n) = \prod_{k=1}^{n} F_{X_k}(x_k)$$

where F_{X_k} is the distribution of X_k.

Often assumed: $X_1, \ldots, X_n$ independent with the same distribution (**i.i.d**).

Note that for **discrete** random variables independence is given by

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} P(X_k = x_k). \tag{8}$$

If densities f_{X_k} exist for all F_{X_k} the joint density f associated to F is given in the independent case as

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k). \tag{9}$$

Definition 2.7

- *The* ***marginal distribution*** *of a subset $X_1, \ldots, X_s$ of n* ***discrete*** *random variables $X_1, \ldots, X_n$ with joint distribution $P(X_1 = x_1, \ldots, X_n = x_n)$ is given by*
$$P(X_1 = x_1, \ldots, X_s = x_s) = \sum_{x_{s+1},\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n)$$
where the summation runs over all possible values of $X_{s+1} \ldots, X_n$.
- *The* ***conditional probability*** *that $X_{s+1} = x_{s+1}, \ldots, X_n = x_n$ is valid, given $X_1 = x_1, \ldots, X_s = x_s$ is*
$$P(X_{s+1} = x_{s+1}, \ldots, X_n = x_n | X_1 = x_1, \ldots, X_s = x_s) = \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(X_1 = x_1, \ldots, X_s = x_s)},$$
assumed that the denominator is positive.

The right hand side of the definition can be memorized as "**joint distribution divided by marginal distribution**".

For independent random variables one has with equation (8)

$$P(X_{s+1} = x_{s+1}, ..., X_s = x_n | X_1 = x_1, ..., X_s = x_s) = P(X_{s+1} = x_{s+1}, ..., X_s = x_n). \tag{10}$$

The concept of marginal distribution and conditionality applies likewise to continuous random variables involving integration.

Remember the similarity of the elementary definitions of conditional probability for two events A and B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

The expectation of a function $g(X)$ of a random variable X is given in the discrete case by

$$E(g(X)) = \sum_i g(x_i)P(x_i)$$

and as

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

for continuous variables with density f, provided all entities make sense.

Joint densities (continuous case) and joint probabilities (discrete case) likewise allow to define expectations for functions $g(X_1, \dots, X_n)$ of dependent or independent random variables $X_1, \dots, X_n$.

A most important example is:

Definition 2.8

- *The **covariance** $\sigma_{X,Y}$ of continuous random variables X, Y with joint density $f(x, y)$ is defined as*

$$\sigma_{X,Y} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))f(x, y)dxdy;$$

- *for discrete variables one has*

$$\sigma_{X,Y} = \sum_{x,y} (x - E(X))(y - E(Y))P(X = x, Y = y).$$

Remember that the **correlation** $\rho_{X,Y}$ of X and Y is calculated by

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{11}$$

with the standard deviation $\sigma_X = \sqrt{\sigma_X^2}$ of X and that of Y respectively.

Another important concept required is conditional expectation.

Definition 2.9 *The **conditional expectation** of X given Y is for discrete random variables defined as*

$$E(X|Y = y) = \sum_x xP(X|Y = y).$$

The definition holds for all values y with $P(Y = y) > 0$. Note that the conditional expectation $E(X|Y = y)$ considered as a function of y is a discrete random variable.

For the more involved continuous case we refer to the textbooks.

2.4 Bayes's formula and odds ratios

The definition of conditional probabilities for events A, B with positive probability shows

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

and thus **Bayes's formula**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{12}$$

Example 2.10 (Color-blindness)
Suppose 2% of a population are color-blind ($P(C) = 0.02$) and let the probability that a color-blind person is a woman be $P(w|C) = 0.01$ What is the probability $P(C|m)$ for a man to be color-blind?

With $P(m) = P(w) = 1/2$ and $P(m|C) = 1 - P(w|C)$ we get

$$P(C|m) = \frac{P(m|C)P(C)}{P(m)} = \frac{0.99 \cdot 0.02}{0.5} \approx 4\% \qquad \diamond$$

More general notations of conditional probability involve parameters and hypotheses.

For example, assume one observes n trials, each with two possible outcomes. Under the hypothesis of i.i.d. Bernoulli trials with probability p for "success" the (conditional) probability of k observed "successes" can be calculated as

$$P(k|B(n,p)) = B(n,p)(k).$$

If $P(k|B(n,p))$ is considered as function of p alone (fixed observation k) one gets a **likelihood function**.

Statistical **point estimation** derives parameters like p from observations like k. One strategy is to take as estimators for the parameters those values which maximize the likelihood function.

These **maximum likelihood estimators** can sometimes be found with elementary techniques. In the Bernoulli example:

$$\partial_p B(n,p)(k) = \binom{n}{k} p^{k-1}(1-p)^{n-k-1}(k - np)$$

For $0 < p < 1$ this derivative is zero if and only if $k = np$ and moreover this defines a maximum of $P(k|B(n,p))$. The maximum likelihood estimator $\hat{p}$ for binomial experiments with given n and observed k is therefore:

$$\hat{p} = \frac{k}{n} \tag{13}$$

Example 2.11 (Comparing genetic sequences) *Given two (observed) sequences $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_m$ with letters from an alphabet* ***A*** *(e.g.* ***A****=\{A, C, G, T \}) . We would like to know wether x and y are related (homologous) or not. Thereto the likelihoods of the models I (independent) and R (related) are compared:*

Model I*: Assume that a letter a occurs independently in the sequences with some probability p_a. The conditional probability for the observations x and y is then*

$$P(x, y|I) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{m} q_{x_i}.$$

Now assume the x and y are ***aligned:*** *$n = m$. This could be achieved by inserting gaps _. A lot of alignment strategies exist to find "optimal" alignments.*

Model R*: Aligned pairs ab occur with a joint probability p_{ab}. In genetics this is usually defined as the probability that a and b result from common ancestors.* ***These probabilities have to be derived from a mathematical model describing the mechanism of heredity****. The conditional probability for the whole alignment is then given as*

$$P(x, y|R) = \prod_{i=1}^{n} p_{x_i y_i}.$$

The ratio of the two likelihoods $P(x, y|R)$ and $P(x, y|I)$ is the ***odds ratio****:*

$$\frac{P(x, y|R)}{P(x, y|I)} = \frac{\prod_{i=1}^{n} p_{x_i y_i}}{\prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}} = \prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

$P(x, y|R) > P(x, y|I)$ corresponds to an odds ratio > 1. In this case hypothesis R is favored over I. Analogously an odds ratio < 1 supports I.

Usually one prefers to apply log *and work with* ***log-likelihood functions*** *and* ***log-odds ratios****:*

$$S = \log(\prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}) = \sum_{i=1}^{n} \log(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}) = \sum_{i=1}^{n} s(x_i, y_i)$$

The last equality uses the log-likelihood ratio for aligned pairs

$$s(a, b) = \log(\frac{p_{ab}}{q_a q_b}).$$

Now, $S > 0$ recommends R as the more probable model and $S < 0$ recommends model I.

2.5 Entropy, Kullback-Leibler distance, mutual information

Entropy is well established in information theory and could be seen as a measure of the average uncertainty of an outcome.

Definition 2.12 *The (Shannon)* ***entropy*** *of a discrete random variable X having only a finite number N of possible outcomes x_i with $P(x_i) > 0$ is defined as*

$$H = -\sum_{i=1}^{N} P(x_i) \log P(x_i). \tag{14}$$

High entropy corresponds to high uncertainty about the outcome of X. The **uniform distribution**, that is $P(x_k) = 1/N$ for all $k = 1, \ldots, N$ has the maximal entropy

$$H = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N.$$

The distribution $P(x_0) = 1$ for one value x_0 has the minimal entropy $H = 0$, the outcome of X is certain.

If the entropy H_{before} before a measurement is high, the reduction to certainty ($H_{after} = 0$) or at least to smaller values of H_{after} by the measurement gains high information. The **information content** is

$$I = H_{before} - H_{after}. \tag{15}$$

If the logarithm is chosen with base 2 one gets **bit** as unit of the entropy. The number of bits can be interpreted as number of dichotomous questions necessary to achieve certainty. This is why entropy is seen as information, producing the apparently startling fact that high information and high uncertainty go together.

Example 2.13 (Entropy of DNA) *The distribution of $\{A, C, D, T\}$ at a specific position can be estimated from the frequencies in a number of related sequences. The entropy H_i at a conserved position i (one prevailing nucleotide) is smaller than the maximal entropy $H_{max} = \ln_2(4) = 2bit$. $H_{max} - H_i$ is positive (larger than some threshold) for conserved positions and almost zero (smaller than some threshold) for not conserved positions . The information content of a sequence*

$$I = \sum_i (H_{max} - H_i)$$

is thus a measure for the amount of conserved positions.

Definition 2.14 *For two probability distributions P, Q defined on the same set $x_1, \dots, x_N$ the **relative entropy** or **Kullback-Leibler distance** is defined by*

$$H(P||Q) = \sum_{i=1}^{N} P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

For uniform Q the relative entropy is an information content.

Proposition 2.15 *$H(P||Q) \geq 0$ holds with $H(P||Q) = 0$ if and only if $P = Q$.*

Proof. Consider $f(x) = \log(x) - x + 1$ for $x > 0$ with derivatives $f'(x) = 1/x - 1$ and $f''(x) = -(x)^{-2} < 0$. The unique maximum of f is $x = 1$ with $f(1) = 0$. This gives for $x > 0$ the elementary inequality

$$\log(x) \leq x - 1,$$

with equality holding for $x = 1$ only. Therefrom one gets:

$$-H(P||Q) = \sum_{i=1}^{N} P(x_i) \log \frac{Q(x_i)}{P(x_i)} \leq \sum_{i=1}^{N} P(x_i)(\frac{Q(x_i)}{P(x_i)} - 1) = \sum_{i=1}^{N} Q(x_i) - \sum_{i=1}^{N} P(x_i) = 0,$$

showing $H(P||Q) \geq 0$. Equality is attained if and only if $Q(x_i)/P(x_i) = 1$ is valid for all $i = 1, \dots, N$. ◇

Note that in general $H(P||Q) \neq H(Q||P)$; the relative entropy is not a mathematical distance.

If P and Q represent distributions according to different hypotheses, the relative entropy is the expectation of the log-odds ratio.

If Q represents the model with independence of letters in two sequences x and y (compare second example in 2.4) the relative entropy is a **mutual information**:

$$M(x, y) = \sum_{i=1}^{n} p_{x_i y_i} \log \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \tag{16}$$

The mutual information quantifies the information one gets about x by observing y.

2.6 Pattern in sequences

Entropy and mutual information can be used to find unusual patterns in biological sequences. For example, the last two bases of introns are often coded with AG. Introns are parts of the DNA spliced out before the transcription and thus not copied to m-RNA.

Consider the following 20 sequences each with length 30. They are produced with MATHEMATICA by a random mechanism and some additional manipulation.

```
T A G T A A G T G A C T G A G A A G G T G C T A T C C A G A
G C A T A C C T C T A A T T A G A T T G T G C C A C C G A T
T T G A G G G A A G G C C A A G T T G C G A G C T C A C A T
C A T A T T G A A A G C C G A G T A A G T T A G C G A C G T
G C G T T T A A T T C T G C A G G T A C C G G G G C T T T C
T G T C C T A T T G G A T A C T C C G A C G T T G T A G A A
T T C C C T C C T C A T A A G A C A A C A C T C C G G T T G
C G A C A T T C C G C A A A A G A T T A C T A C A C T A A G
A T G T G G C T C T A A C G A G C A G T C A C C C C T C G G
C C C T A C T A C C G A T C A G G A T G G G A T C A T A A T
A T G G G T G A C T A A G A C T C A A A G A T C C A C T T C
G G C C C G G A T A C A T T A G T G C G C T C C T A A T C G
C G G G A G T G C T T T T A G A G G G T G G A A A T G G A A
T A T C T A T C C A C T G A A G G A G C T G C A A G G G C C
G A A C C C C A G A T T A G A G G G T T A T C G C G C A A C
T A C C T A C G T G G A A T C T G A A G C A A A G T G G G G
C A G A T T C A C G G A C A A G C A T A G C A C C C G C C C
A A A G C A G G A A C C A A A G A G C G C A T A C A T C T T
A T T G T G G G A G G A T T G A T G T T T A A T G A A C T G
G G G A A G G A T T T G T T A G T A A A T T C C A A T C A G
```

The overall frequency of bases is $P_{total} = (p_A, p_C, p_G, p_T) = (0.285, 0.225, 0.255, 0.235)$.

For each column s we now calculate the frequency of the occurring bases $P_{column}(s) = (p_A(s), p_C(s), p_G(s), p_T(s))$ and therefrom the relative entropy

$$H(P_{column}(s)||P_{total}) = \sum_{X=A,C,G,T} p_X(s) \log_2 \left(\frac{p_X(s)}{p_X} \right)$$

Plotting the relative entropy for every column reveals peaks for $s = 15$ and $s = 16$. These are just the columns where manipulation took place.

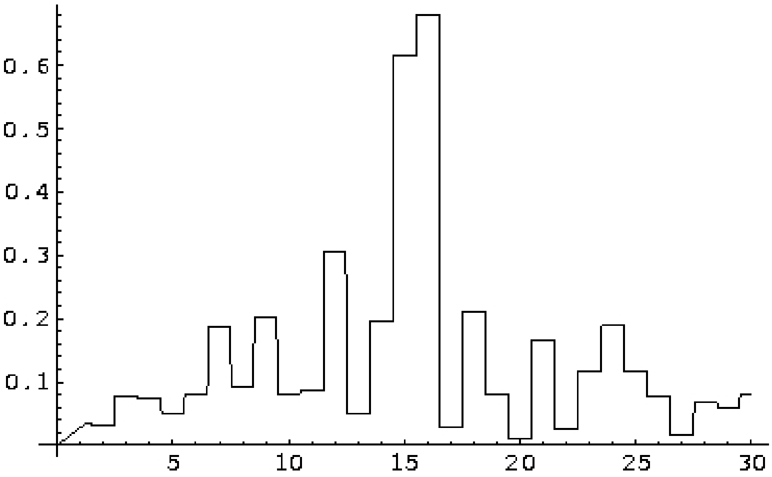


Figure 2: Relative entropy $H(P_{column}(s)||P_{total})$ as function of column number.

To investigate whether columns are correlated we calculate the mutual information between site s and site $s+1$ ($s = 1 \dots 39$):

$$M(s) = \sum_{X,Y=A,C,G,T} p_{X,Y}(s) \log_2 \left(\frac{p_{X,Y}(s)}{p_X(s) p_Y(s+1)} \right).$$

Here $p_{X,Y}(s)$ represents one of the 16 frequencies of base pairs with base X at position s and Y at position $s+1$. The mutual information is maximal at position $s = 15$ indicating that the bases in column $s = 15$ and $s = 16$ are not independent.

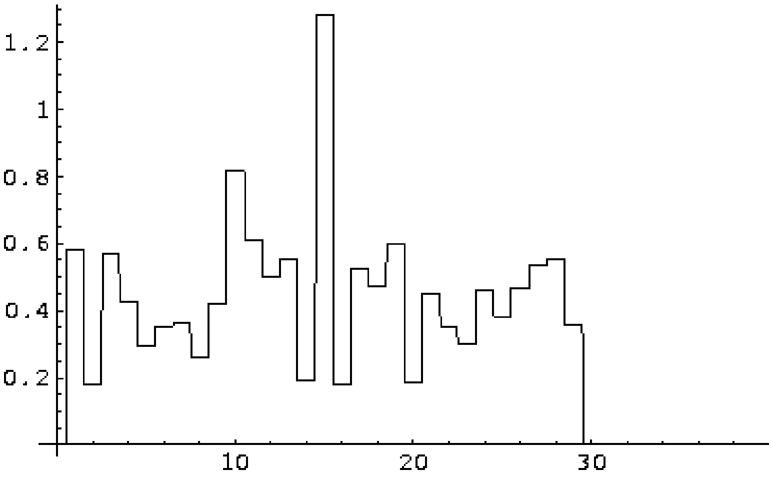


Figure 3: Mutual information of adjacent columns as function of the first of them.

3 Estimation Techniques

In case of incomplete information there exist different strategies to estimate interesting parameters. In this chapter we introduce the EM (Expectation Maximization) algorithm (Dempster & al. 1977) and the Bayes paradigm to update information.

3.1 Gene counting: Hardy-Weinberg equilibrium II

A) Consider a locus with s **codominant** alleles. The frequency of alleles can be assessed by counting the frequency of phenotypes.

In a population of size n the counts $(X_1, ..., X_s)$ of the different alleles have a **multinomial distribution** (joint distribution)

$$P(X_1 = n_1, ..., X_s = n_s) = \frac{n!}{\prod_{k=1}^{s}(n_k)!} \prod_{k=1}^{s} p_k^{n_k}. \tag{17}$$

The probabilities p_k of the different alleles satisfy $p_1 + \ldots + p_s = 1$. For the counts $n_1 + \ldots + n_s = n$ is valid. Note that the random variables $X_1, \ldots, X_s$ are **not** independent! The (generally unknown) p_k are usually estimated by the relative frequencies:

$$\hat{p_k} = \frac{n_k}{n} \tag{18}$$

This is a maximum likelihood estimator of p_k.

B) Estimating allele frequencies in the presence of **recessive alleles**. For example, the blood group genotypes AA and A0 yield the same phenotype. So, how to estimate the frequencies of allele A?
Assume Hardy Weinberg equilibrium!

Blood group example:

Let the true probabilities of the alleles be p_A, p_B and p_0 and the counts of phenotypes in a population with size n be n_A, n_B and n_0.
The probability of genotype AA then is p_A^2 and that of $A0$ is $2p_Ap_0$. This gives the numbers of respective genotypes in the population

$$n_{AA} = \frac{n_A p_A^2}{p_A^2 + 2p_Ap_0} \tag{19}$$

and

$$n_{A0} = \frac{n_A 2p_Ap_0}{p_A^2 + 2p_Ap_0} \tag{20}$$

The estimator for p_A is now:

$$p_A = \frac{2n_{AA} + n_{A0} + n_{AB}}{2n} \tag{21}$$

A and B are codominant and thus n_{AB} can be observed from phenotypes.

Now an iteration process is applied:

Start: guess $p_{A,1}$, $p_{B,1}$ and $p_{0,1}$.

Iteration: For $k = 1, 2, 3, ...$

$$n_{AA,k} = n_A \frac{p_{A,k}^2}{p_{A,k}^2 + 2p_{A,k}p_{0,k}}, \tag{22}$$

$$n_{A0,k} = n_A \frac{2p_{A,k}p_{0,k}}{p_{A,k}^2 + 2p_{A,k}p_{0,k}}, \tag{23}$$

and therefrom calculate the update

$$p_{A,k+1} = \frac{2n_{AA,k} + n_{A0,k} + n_{AB}}{2n}. \tag{24}$$

The calculations for $p_{B,k}$ are identical after interchanging in the above equations A, and B. The value for $p_{0,k}$ may be found by the equation

$$p_{A,k} + p_{B,k} + p_{0,k} = 1$$

.

Now continue until some stabilization of the values occurs after say M iterations. The values of $p_{A,M}$, $p_{B,M} p_{0,M}$ are then (maximum likelihood) estimates for p_A, p_B and p_0 respectively.

Example: $n = 100$, $n_{AB} = 10$, $n_A = 60$, $n_B = 30$ gives with the initial values $p_{A,0} = 3/5$, $p_{B,0} = 3/10$ and $p_{0,0} = 1/10$ after 5 iterations $(0.35, 0.153, 0.496)$. The shown digits are already stable.

Why does this work?

This is a special case of the EM algorithm ! (Exercise!)

3.2 EM algorithm

Assume we want to estimate a vector of parameters Θ (like (p_A, p_B, p_0) in the blood group example) with maximum likelihood techniques. The log-likelihood function $\log P(Y|\Theta)$ depends on complete information Y (e.g. n_{AA}, n_{A0}, n_{AB}, n_{BB}, n_{B0}, n_{00}). If the necessary information is only partially available (e.g. only n_{AB} and n_0 from phenotype) one can employ the **expectation maximization (EM) algorithm**. We consider the discrete case, an extension to continuous variables is straightforward (compare the textbooks).

First observe how the log-likelihood function for observable but incomplete X can be composed from log-likelihoods with complete information:

$$\log P(X|\Theta) = \log \sum_Y P(X, Y|\Theta). \tag{25}$$

The sum is to be taken over all possible values for the missing data Y.

The definition of conditional probabilities yields

$$P(X, Y|\Theta) = P(Y|X, \Theta)P(X|\Theta)$$

and thus

$$\log P(X|\Theta) = \log P(X, Y|\Theta) - \log P(Y|X, \Theta). \tag{26}$$

Now multiply (26) with the probability $P(Y|X, \Theta_n)$ for the unobservable data Y given the observations X and "true" parameters Θ_n. Summation over all possible Y gives:

$$\log P(X|\Theta) = \sum_Y P(Y|X, \Theta_n) \log P(X, Y|\Theta) - \sum_Y P(Y|X, \Theta_n) \log P(Y|X, \Theta). \tag{27}$$

The first term on the right side of (27) is the central function

$$Q(\Theta|\Theta_n) := \sum_Y P(Y|X, \Theta_n) \log P(X, Y|\Theta) \tag{28}$$

The **E-step** of the EM algorithm is now to calculate this conditional expectation:

$$Q(\Theta|\Theta_n) = E(\log(P(X, Y|\Theta)|X, \Theta_n) \tag{29}$$

The **M-step** maximizes $Q(\Theta|\Theta_n)$ as function of Θ to obtain new parameters Θ_{n+1}.

The 2-step procedure is iterated until equilibrium is attained.

Central to the theory of EM iteration is the following inequality:

Theorem 3.1 *Successive EM parameters* Θ_n *and* Θ_{n+1} *satisfy*

$$\log P(X|\Theta_n) \leq \log P(X|\Theta_{n+1}).$$

Strict inequality is valid if the conditional distributions $P(Y|X, \Theta_n)$ *and* $P(Y|X, \Theta_{n+1})$ *differ.*

Proof. From (27) one gets

$$\log P(X|\Theta) - \log P(X|\Theta_n) = Q(\Theta|\Theta_n) - Q(\Theta_n|\Theta_n) + \sum_Y P(Y|X, \Theta_n) \log \frac{P(Y|X, \Theta_n)}{P(Y|X, \Theta)}.$$

The last term of this equation is a relative entropy and therefore always non negative (proposition 2.1).

Θ_{n+1} is chosen as to maximize $Q(\Theta|\Theta_n)$. Therefore

$$Q(\Theta_{n+1}|\Theta_n) - Q(\Theta_n|\Theta_n)$$

is non negative too and thus

$$\log P(X|\Theta_{n+1}) - \log P(X|\Theta_n) \geq 0.$$

For equality it is necessary that the relative entropy satisfies

$$H(P(Y|X, \Theta_n)|P(Y|X, \Theta_n)) = 0.$$

This is only possible if the involved distributions are equal (again proposition 2,1). ◇

The theorem states that the log-likelihood is never decreased by an EM iteration and gives a condition when every EM iteration achieves an improvement (increases the log-likelihood).

Note that even a convergent EM algorithm might only find **local maxima** of the likelihood function.

3.3 Bayesian paradigm

Maximum likelihood estimation may especially in case of few data produce undesirable results. If, for example, some outcome has zero counts in a multivariate scenario the estimated probability for this outcome is also zero. This may be contrary to former observations. Thus, a more "robust" estimation, down-weighting the actual experiment, may be adequate.

A consistent framework for information update is given by the **Bayesian paradigm**. This is a technique which integrates prior knowledge (eventually from different sources) with Bayes's formula:

$$P(\theta|M, Y) = \frac{P(\theta|M)P(Y|\theta, M)}{P(Y|M)}. \tag{30}$$

Here, $P(\theta|M)$ is the a priori distribution (**prior**) of the parameter of interest θ. It is a conditional distribution given some initial information (measurement, model, ...) M. The **posterior probability** $P(\theta|M, Y)$ for θ is calculated from the prior and the likelihood $P(Y|\theta, M)$ with new information Y.

Different strategies exist to estimate θ from the posterior probability.

The maximal a posteriori probability (**MAP**) takes as estimate θ^{MAP} which maximizes the posterior or equivalently

$$\theta^{MAP} = \text{argmax}_\theta P(\theta|M)P(Y|\theta, M).$$

The posterior mean estimator (**PME**) takes the mean of the posterior with respect to the domain Θ of possible parameters θ

$$\theta^{PME} = \int_\Theta \theta P(\theta|M, Y) \mathrm{d}\theta.$$

Here the integral is applied to every component of the parameter vector θ.

Remarks 3.2 *a) The determination of the prior allows some subjectivity making the Bayes method controversial.*

b) In contrast to ML estimates, MAP and PME are not invariant under nonlinear transformations of the parameters. This unfavorable feature makes scaling a relevant issue.

An important Bayesian prior is the **Dirichlet distribution**:

$$\mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \prod_{k=1}^{N} \theta_k^{\alpha_k - 1}. \tag{31}$$

Here, the parameters of interest $\theta = (\theta_1, \ldots, \theta_N)$ define probability distributions. The domain Θ is thus the $(n-1)$-dimensional simplex

$$\Theta = \{\theta \ | \ \theta_k \geq 0 \text{ for all } k = 1, \ldots N \text{ and } \sum_{k=1}^{N} \theta_k = 1\}.$$

The constants $\alpha = (\alpha_1, \ldots, \alpha_N)$ are positive parameters representing a priori information and determine the prior.

$Z(\alpha)$ is a normalization factor to make $\mathcal{D}(\theta|\alpha)$ a density with respect to θ:

$$Z(\alpha) = \int\limits_{\Theta} \prod_{k=1}^{N} \theta_k^{\alpha_k - 1} \mathrm{d}\theta = \frac{\prod\limits_{k=1}^{N} \Gamma(\alpha_k)}{\Gamma(\sum\limits_{k=1}^{N} \alpha_k)}. \tag{32}$$

The last equation (stated without proof) involves the Gamma function which satisfies $\Gamma(x+1) = x\Gamma(x)$.

Remarks 3.3 *a) It can be shown that the expectation of θ_k for a Dirichlet distribution is*

$$E(\theta_k | \mathcal{D}(\theta|\alpha)) = \alpha_k \cdot (\sum_{k=1}^{N} \alpha_k)^{-1}.$$

b) The Dirichlet distribution is connected to unit scale gamma distributions

$$f_{(1,\alpha_k)}(x_k) = \frac{1}{\Gamma(\alpha_k)} \theta_k^{\alpha_k - 1} e^{-x_k} \ :$$

For independent variables $X_1, \ldots, X_N$ with X_k distributed according to $f_{(1,\alpha_k)}(x_k)$ the ratios

$$\theta_k = \frac{X_k}{\sum_{k=1}^{N} X_k}$$

have as joint density over Θ the distribution $\mathcal{D}(\theta|\alpha)$. This fact allows to sample Dirichlet distributions from the more accessible gamma distributions.

The following theorem states that the Dirichlet distribution is the **conjugate** prior for the multinomial distribution.

Theorem 3.4 *For multinomial observations*

$$P(n|\theta) = \frac{N!}{\prod_{k=1}^{s} n_k!} \prod_{k=1}^{s} \theta_k^{n_k},$$

with Dirichlet prior $\mathcal{D}(\theta|\alpha)$ and a posteriori distribution

$$P(\theta|n,\alpha) = \frac{P(n|\theta)\mathcal{D}(\theta|\alpha)}{P(n|\alpha)}, \quad n = (n_1, \ldots, n_s)$$

the following is valid ($A = \sum_{k=1}^{s} \alpha_k$):

$$\begin{array}{lrcl} a) & P(\theta|n,\alpha) & = & \mathcal{D}(\theta|n+\alpha) \quad , \\ b) & \theta_i^{PME} & = & \frac{n_i+\alpha_i}{N+A}, \quad i = 1, \ldots, s. \end{array}$$

Proof. First one gets

$$P(\theta|n,\alpha) = \frac{\frac{N!}{\prod_{k=1}^{s} n_k!} \prod_{k=1}^{s} \theta_k^{n_k} \frac{1}{Z(\alpha)} \prod_{k=1}^{s} \theta_k^{\alpha_k - 1}}{P(n|\alpha)} \propto \prod_{k=1}^{s} \theta_k^{n_k+\alpha_k-1}.$$

Thus $P(\theta|n,\alpha)$ is proportional to $\mathcal{D}(\theta|n+\alpha)$. But both distributions are normalized with respect to θ and this forces the equality a).

From this result and (32) we calculate for the posterior mean estimator θ^{PME}:

$$\theta_i^{PME} = \int_\Theta \theta_i \mathcal{D}(\theta|n+\alpha) \mathrm{d}\theta = \frac{1}{Z(n+\alpha)} \int_\Theta \prod_{k=1}^{s} \theta_k^{n_k+\alpha_k-1+\delta_{ki}} = \frac{Z(n+\alpha+e_i)}{Z(n+\alpha)}.$$

e_i is a vector having 1 at position i and 0 else. Now

$$\frac{Z(n+\alpha+e_i)}{Z(n+\alpha)} = \frac{\prod\limits_{k=1}^{s} \Gamma(n_k+\alpha_k+e_i)\Gamma(\sum\limits_{k=1}^{s} n_k+\alpha_k)}{\prod\limits_{k=1}^{s} \Gamma(n_k+\alpha_k)\Gamma(\sum\limits_{k=1}^{s} n_k+\alpha_k+e_i)} = \frac{\Gamma(n_i+\alpha_i+1)\Gamma(N+A)}{\Gamma(n_i+\alpha_i)\Gamma(N+A+1)} = \frac{n_i+\alpha_i}{N+A}$$

and this is statement b). ◇

Remarks 3.5 *a) Invoking a Dirichlet prior to multinomial observations and taking the PME mimics adding extra observations α to the the actual measurement and taking the ML estimator. This is why the α_i are called* ***pseudocounts****.*

b) For fixed pseudocounts and growing number of actual observations θ^{PME} approaches the ML and is thus a consistent estimator.

c) For large values of α, the estimated parameters do not change "too much" after adding few observations. This makes results more "robust".

Example 3.6 *The 15^{th} column of the above sequence example has the following base counts:*

$$n = (n_A, n_C, n_G, n_T) = (13, 3, 4, 0)$$

The ML estimator for the probabilities in this column is

$$p^{ML} = (13/20, 3/20, 4/20, 0/20) = (0.65, 0.15, 0.20, 0.00).$$

Assume we take a Dirichlet prior with

$$\alpha = (20, 20, 20, 20)$$

The corresponding posterior mean estimator is then

$$p^{PME} = ((13{+}20)/100, (3{+}20)/100, (4{+}20)/100, (0{+}20)/100) = (0.33, 0.23, 0.24, 0.2).$$

This result is the same as if we had 80 further sequences where the bases in the respective column are univariate distributed. As a thumb rule one could thus state:

Choose the size of pseudocounts α similar to the number of already investigated sequences (if the results are comparable).

4 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods have a variety of statistical applications. They allow to simulate complex stochastic processes and to calculate (suboptimal) estimators in parameter spaces with high dimensionality. Especially in genetics these techniques are nowadays frequently applied.

4.1 Markov chains

The theory of Markov processes is well developed and deserves lectures of its own. We restrict our attention to **discrete time finite Markov chains**. These are more simple than general processes, but are highly relevant for applications like the construction of substitution matrices (PAM) and the modelling of evolutionary processes.

Definition 4.1 *Let $(X_n)_{n\in\mathbb{N}_0}$ be random variables with values in a finite set S (**state space**). $(X_n)_{n\in\mathbb{N}_0}$ is a **Markov chain** if for all $n\in \mathbb{N}_0$ and $(s_i)_{i=0}^n \subseteq S$ the conditions*

$$P(X_{n+1} = s_{n+1}|X_n = s_n, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) = P(X_{n+1} = s_{n+1}|X_n = s_n) \tag{33}$$

are satisfied.

*The probabilities $q(k) = P(X_0 = k)$ define the **initial distribution** and the matrix*

$$\boldsymbol{P}_{n+1} := (p_{n+1}(i,j))_{1\leq i,j\leq |S|}$$

*with $p_{n+1}(i,j) = P(X_{n+1} = j|X_n = i)$ is the **transition matrix** composed of the transition probabilities from state i to state j at time $n+1$.*

Note that we have, as usual, identified S with the finite set of integers $\{1,\ldots,|S|\}$ and interpreted the index of X_{n+1} is as time.

The condition (33) is a **property of memoryless**:

The transition probabilities

$$p_{n+1}(i,j)$$

between two **states** i and j depend on the last state i only and not on the further history.

If the transition matrices do not change with time:

$$\mathbf{P}_{n+1} = \mathbf{P}_1 =: \mathbf{P}$$

for all $n \in \mathbb{N}$ the sequence $(X_n)_{n\in\mathbb{N}_0}$ is called **homogeneous** Markov chain.

Note that the matrices $\mathbf{P}_{n+1}$ are **stochastic**: Matrix elements are non-negative and rows sum up to one.

Remarks 4.2 *Every sequence $(\boldsymbol{M}_n)_{n\in\mathbb{N}}$ of stochastic $|S|\times|S|$-matrices together with a probability distribution q on S determines a Markov chain. To establish this, one defines for all $t\in\mathbb{N}_0$ the joint probabilities for $(X_n)_{n\in\mathbb{N}_0}$ as*

$$P(X_0=s_0,\dots,X_t=s_t):=q(s_0)\boldsymbol{M}_1(s_0,s_1)\cdots\boldsymbol{M}_t(s_{t-1},s_t).$$

The consistency theorem of Kolmogorov allows to extend these finite dimensional distributions to a joint distribution of the entire Markov chain.

Definition 4.3 *A Markov chain is **stationary** if $(X_n)_{n\in\mathbb{N}_0}$ and $(X_{n+1})_{n\in\mathbb{N}_0}$ have the same (joint) distribution.*

Remarks 4.4 *For a homogeneous Markov chain $(X_n)_{n\in\mathbb{N}_0}$ the following statements are equivalent (Exercise):*

1. *$(X_n)_{n\in\mathbb{N}_0}$is stationary*

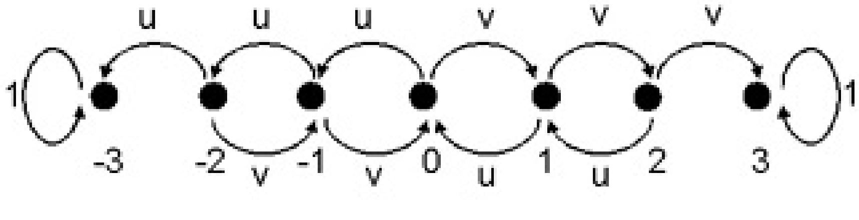
2. *X_0 and X_1 have the same distribution*

3. *The initial distribution q is **invariant** under $\boldsymbol{P}$:* $\quad q\boldsymbol{P}=q.$

Homogeneous Markov chains can be visualized by **(directed) graphs**: vertices are the states and an arrow from i to j, with a number $p(i,j)$ attached, indicates the possibility to pass from i to j with probability $p(i,j)$. If $p(i,j)=0$ no arrow is shown. If $p(i,i)=1$, the state i is **absorbing**. Once the chain gets to this state it remains there.

Example 4.5 *Let $S=\{0,\pm1,\dots,\pm N\}$, $q(0)=1$, $p(N,N)=p(-N,-N)=1$, and, for $|i|<N$:*

*$p(i,i+1)=u$, $p(i,i-1)=v$ with $u+v=1$ and $p(i,j)=0$ in all other cases. Such a Markov chain is called **simple random walk** with **absorbing barrier**.*



It could model a two-player game with each player having a bankroll N. At each turn the first player wins $+1$ from the second with probability u and loses 1 with probability v. The vertices $-N$ and N are absorbing. They represent the ruin of the players.

In the sequel we assume that all Markov chains are homogeneous (if not otherwise stated).

Definition 4.6 *The probability of a transition from state i to state j in k steps is given by*

$$p^{(k)}(i,j) := P(X_k = j | X_0 = i),$$

and the probability of finding the state j at time k is

$$q^{(k)}(j) := P(X_k = j).$$

*Further denote as $q^{(k)}$ and $\boldsymbol{P}^{(k)}$ the respective distribution and **k-step transition** matrices.*

Theorem 4.7 *The k-step transition probabilities $p^{(k)}(i,j)$ satisfy the* ***Kolmogorov-Chapman equation***

$$p^{(k+l)}(i,j) = \sum_{s \in S} p^{(k)}(i,s) p^{(l)}(s,j) \tag{34}$$

or in matrix form

$$\boldsymbol{P}^{(k+l)} = \boldsymbol{P}^{(k)} \boldsymbol{P}^{(l)}.$$

Proof. With the formula for total probability and the Markov property one obtains:

$$p^{(k+l)}(i,j) = P(X_{k+l} = j | X_0 = i) = \sum_{s \in S} P(X_{k+l} = j, X_k = s | X_0 = i)$$

$$= \sum_{s \in S} P(X_{k+l} = j | X_k = s) P(X_k = s | X_0 = i) = \sum_{s \in S} p^{(k)}(i,s) p^{(l)}(s,j).$$

◇

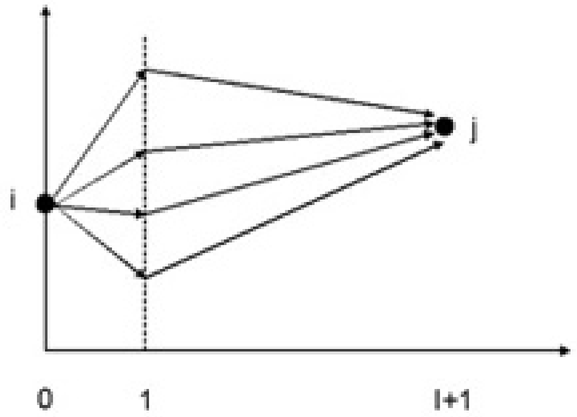
The following special cases are especially important:

The **backward equation**

$$p^{(1+l)}(i,j) = \sum_{s \in S} p(i,s) p^{(l)}(s,j) \tag{35}$$

or

$$\mathbf{P}^{(1+l)} = \mathbf{P}\mathbf{P}^{(l)}$$

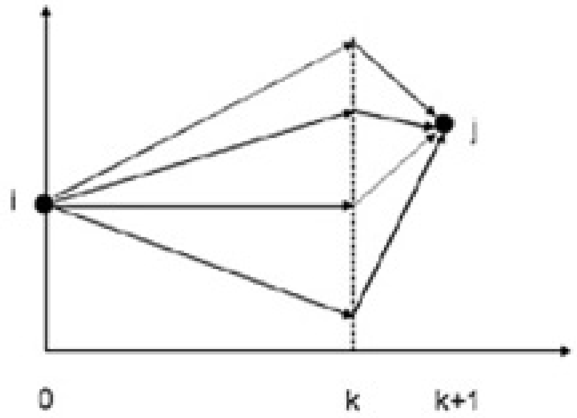


and the **forward equation**

$$p^{(k+1)}(i,j) = \sum_{s \in S} p^{(k)}(i,s) p(s,j) \tag{36}$$

or

$$\mathbf{P}^{(k+1)} = \mathbf{P}^{(k)}\mathbf{P}.$$



Analogously, one has for the unconditional probabilities $q^{(k)}(j)$

$$q^{(k+l)}(j) = \sum_{s \in S} p^{(k)}(s) p^{(l)}(s, j) \tag{37}$$

or

$$q^{(k+l)} = q^{(k)} \mathbf{P}^{(l)}.$$

In particular one has again a forward equation

$$q^{(k+1)} = q^{(k)} \mathbf{P}$$

and a backward equation

$$q^{(1+l)} = q \mathbf{P}^{(l)}.$$

These equations establish a main result:

Theorem 4.8 *For homogeneous Markov chains the k-step transition probabilities $p^{(k)}(i, j)$ are the elements of k^{th} power of the matrix $\boldsymbol{P}$:*

$$\boldsymbol{P}^{(k)} = \boldsymbol{P}^{k}$$

The asymptotic behavior of a homogeneous Markov chain can thus be studied by examination of the stochastic matrix $\mathbf{P}$.

To discuss convergence of $\mathbf{P}^k$ or $q^{(k)}$ for growing k we need some properties of stochastic matrices. We do not give proofs here, instead refer to textbooks (look up for Perron and Frobenius theorems).

For a matrix, $\mathbf{P} > 0$ denotes that all entries are strictly positive.

Definition 4.9 *A stochastic matrix $\boldsymbol{P}$ is called **primitive** if there exists a $k_0 \in \mathbb{N}$ with $\boldsymbol{P}^{k_0} > 0$.*

Remarks 4.10 *For a transition matrix $\boldsymbol{P}$ of a homogeneous Markov chain primitivity implies that from any initial state i any state j is reached in k_0-steps with a positive probability $p^{(k_0)}(i, j)$ (j is **accessible** from i).*

$\boldsymbol{P}^{k_0} > 0$ implies $\boldsymbol{P}^{k} > 0$ for all $k > k_0$ (Exercise).

A primitive $\boldsymbol{P}$ has $r = 1$ as simple eigenvalue with eigenvector $\lambda(1, \ldots, 1)^T$ (T means transposed and λ is any real number). The other eigenvalues have absolute values smaller than 1. These facts are essential for the following theorem.

Theorem 4.11 *Let $\boldsymbol{P}$ be a primitive stochastic $n \times n$-matrix. Then*

$$\lim_{k \longrightarrow \infty} \boldsymbol{P}^k = \begin{pmatrix} y_1 & \dots & y_n \\ \vdots & & \vdots \\ y_1 & \dots & y_n \end{pmatrix} \tag{38}$$

with the left-eigenvector $y = (y_1, \dots, y_n)$ uniquely determined by $y\boldsymbol{P} = y$ and $y_1 + \dots + y_n = 1$.

Proof. For example: chapter IV in B. Huppert, Angewandte Lineare Algebra, deGruyter, Berlin, 1990. ◇

Remarks 4.12 *The left-eigenvector y defines an invariant measure on S. Moreover the Markov chain starting with any initial distribution q ends up with this* ***limit distribution*** *(k growing to infinity):*

$$q^{(k)} = q\boldsymbol{P}^k \longrightarrow q \begin{pmatrix} y_1 & \dots & y_n \\ \vdots & & \vdots \\ y_1 & \dots & y_n \end{pmatrix} = y$$

To handle more general cases than primitive Markov chains one defines an equivalence relation on the states of a Markov chain.

Definition 4.13 *States i and j of S* ***communicate*** *$(i \sim j)$ if i is accessible from j and j is accessible from i.*

Remarks 4.14 *It is easily seen that $\sim$ is symmetric, reflexive and transitive, and thus defines a decomposition of S in disjoint equivalence classes C (Exercise).*

Definition 4.15 *A class C of communicating states is* ***recurrent*** *(or* ***essential****) if all states accessible from states within C belong to C; otherwise C is called* ***transient*** *(or* ***inessential****).*

Remarks 4.16 *In the above example (two player game) one has three classes:*
The recurrent classes ("game over") $C_1 = \{-N\}$ and $C_2 = \{N\}$ and the transient class ("game ongoing") $C_3 = \{-N+1, \dots, N-1\}$.

Theorem 4.17 *Transient classes C are left in the long run:*

$$\lim_{k \longrightarrow \infty} P(X_k \in C) = 0$$

Recurrent classes C give rise to stationary distributions y^q with $y^q_k = 0$ for k not in C and $y^q_k > 0$ for k in C.

A general stationary distribution y has the form

$$y = \sum_{C \ recurrent} \lambda_C \ y^C \quad with \quad \lambda_C \geq 0 \quad and \quad \sum_{C \ recurrent} \lambda_C = 1. \tag{39}$$

Proof. See for example VIII 3 Theorem 1 and VIII 4 Theorem 2 in A.N.Shiryayev, Probability, Springer 1984. ◇

The theorem implies that eventually it might be enough to include only recurrent states in the modelling of long running processes. Transient states (if any) could in such cases be considered as died out and a reduced model might do the job.

Definition 4.18 *A Markov chain is called **indecomposable** if all its states are recurrent and communicate (one equivalence class of $\sim$ only).*

Remarks 4.19 *The transition matrix of a Markov chain with essential states only can be brought to block form (by changing the enumeration of states if necessary):*

$$\boldsymbol{P} = \begin{pmatrix} A_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \dots & \boldsymbol{0} & A_r \end{pmatrix}. \tag{40}$$

*The matrices $A_1, \dots, A_r$ correspond to r equivalence classes of communicating states. Every matrix together with the corresponding **indecomposable class** of recurrent communicating states defines an indecomposable Markov chain. These r chains can be be studied independently.*

Definition 4.20 *The period $d(j)$ of a state j is the largest positive integer l satisfying:*

$p^{(n)}(j, j) > 0$ is valid only if n has the form $n = l \cdot m$ ($m \in \mathbb{N}$).

Theorem 4.21 *All states of an indecomposable class the same period.*

Proof. Any two states i and j communicate. Thus, there are numbers k and l with $p^{(k)}(i,j) > 0$ and $p^{(l)}(j,i) > 0$. By the Kolmogorov-Chapman equations one gets

$$p^{(k+l)}(i,i) \geq p^{(k)}(i,j)\, p^{(l)}(j,i) > 0.$$

Therefore $k+l$ is divisible by the period $d(i)$ of state i.

Suppose there is an integer $n > 0$ not divisible by $d(i)$. Then the number $k+l+n$ is also not divisible by $d(i)$ and thus $p^{(k+l+n)}(i,i) = 0$. But again from the Kolmogorov-Chapman equations one has

$$0 = p^{(k+l+n)}(i,i) \geq p^{(k)}(i,j)\, p^{(n)}(j,j)\, p^{(l)}(j,i).$$

This forces $p^{(n)}(j,j) = 0$.

Thus, $p^{(n)}(j,j) > 0$ requires that n is divisible by $d(i)$ and thus $d(i) \leq d(j)$.

By changing roles of i and j one gets $d(j) \leq d(i)$ and finally $d(i) = d(j)$ $\diamond$

The theorem shows that the following definition makes sense.

Definition 4.22 *Denote the (common) period of states of an indecomposable class C as $d(C)$. If $d(C) = 1$ the class (and the corresponding Markov chain) is called* ***aperiodic****.*

Classes C which are not aperiodic can be divided in **cyclic subclasses** as follows ($d := d(C) > 1$):

Chose a state $s \in C$ and define the subsets (dependent of s) by:

$$C_0 = \{j \in C : p^{(n)}(s,j) > 0 \Rightarrow n \equiv 0 (\text{mod } d)\}$$

$$C_1 = \{j \in C : p^{(n)}(s,j) > 0 \Rightarrow n \equiv 1 (\text{mod } d)\}$$

$$\dots$$

$$C_{d-1} = \{j \in C : p^{(n)}(s,j) > 0 \Rightarrow n \equiv (d-1) (\text{mod } d)\}$$

Then, obviously, one gets a disjoint partition

$$C = C_0 \cup C_1 \cup \ldots \cup C_{d-1}.$$

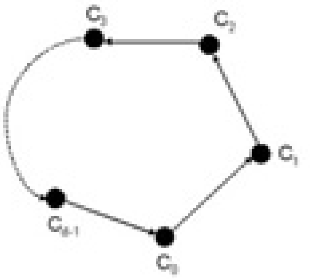
Theorem 4.23 *In one step a state of class C_k goes exclusively to a state of class $C_{(k+1) mod\ d}$.*

Proof. Let i be in C_k and $p^{(1)}(i,j) > 0$ for a state j. $p^{(n)}(s,i) > 0$ requires $n \equiv k(\text{mod } d)$ or $n+1 \equiv (k+1)(\text{mod } d)$. The last equation and the inequality

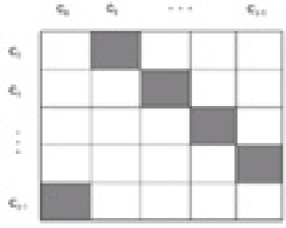
$$p^{(n+1)}(s,j) \geq p^{(n)}(s,i)\ p^{(1)}(i,j) > 0$$

show that j is in $C_{(k+1) mod\ d}$. ◇

Remarks 4.24 *The theorem states the cyclic character of transitions for states of period d :*



The corresponding transition matrix has block structure:



A state initially in C_0 will be in a specific class C_k at "times" t with $t = k + sd$ $(s = 0, 1, \ldots)$. Thus d-step transitions restricted to the classes C_k define transition matrices

$$\boldsymbol{P}(k) := (p(i,j)^{(d)}_{i,j \in C_k})$$

for indecomposable aperiodic Markov chains.

The matrices $\boldsymbol{P}(k)$ are primitive (Exercise!).

The discussed partition of homogeneous (recurrent) Markov chains into communicating aperiodic classes is why Markov models are often assumed to be aperiodic and irreducible.

4.2 Example: Wright Fisher Model

Recall the the situation of section 2.2. where the Hardy Weinberg equilibrium for two alleles is obtained. We assumed infinite population. What changes if the population is (more realistically) considered as finite?

Now assume one has in each discrete generation $2N$ gametes of allele type A_1 or A_2. The random variables X_n count the number and gametes A_1 in generation n. In every daughter generation the gametes are independently and with equal probability chosen from the parent gametes. Then, the probability to encounter l gametes of type A_1 in a daughter generation if k are in the parent generation is given by a binomial distribution:

$$p(k,l) := P(X_{n+1} = l | X_n = k) = \binom{2N}{l} \left(\frac{k}{2N}\right)^l \left(1 - \frac{k}{2N}\right)^{2N-l} = B(2N, \frac{k}{2N}\ )(l) \tag{41}$$

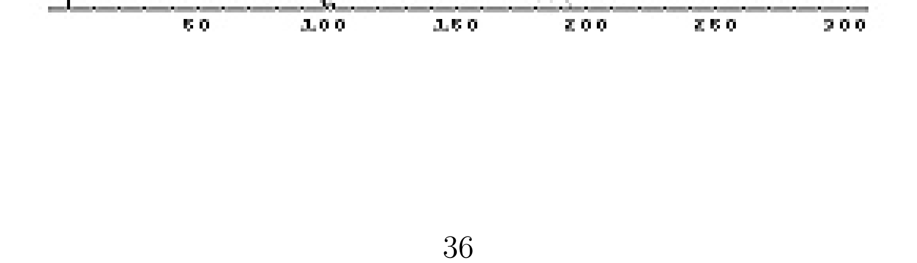
Therefore $(X_n)_{n \in \mathbb{N}_0}$ is a homogeneous Markov chain.
The $2N+1$ states fall into 3 communicating classes:
$k = 0$ implies $p(0,0) = 1$ and $p(0,l) = 0$ for all $l \neq 0$.
$k = 2N$ implies $p(2N,2N) = 1$ and $p(2N,l) = 0$ for all $l \neq 2N$.
For $1 \leq l, k \leq 2N-1$ one has $p(k,l) > 0$. All those states communicate and define one transient class with respect to $\sim$.

All stationary distributions are of the form (compare (39)) $p\delta_0 + (1-p)\delta_{2N}$. Furthermore one can show that every initial distribution (of X_0) tends to such a stationary distribution with

$$p = \frac{1}{2N} E X_0 = \frac{1}{2N} \sum_{k=0}^{2N} k P(X_0 = k).$$

This is the average fraction of chains terminating in $k = 0$.

The following figure shows several realizations of a Wright-Fisher process with
$2N = 100$ and $p = 1/2$



4.3 Gibbs Sampling

Sampling and annealing techniques are applied for simulation and parameter estimation. Such methods help when direct calculations are not feasible due to the computational size of the problem. As example one may regard the distribution of specific patterns in DNA sequences.

If all elements x of a finite probability space Ω have positive probability $P(x) > 0$ one can write P in **Gibbs form**:

$$P(x) = \frac{1}{Z} \exp(-H(x)), \tag{42}$$

with **energy** H and the normalization constant

$$Z = \sum_{z \in \Omega} \exp(-H(z)), \tag{43}$$

called **partition function**. P is also called a **Gibbs distribution**.

Thereto, one only has to define

$$H(x) := -\ln(P(x)) - \ln(Z)$$

with any positive constant Z to gain

$$\exp(-H(x)) = P(x) \cdot Z.$$

Choosing Z as in (43) provides the representation (42).

In many applications Ω can be regarded as product

$$\Omega = \prod_{t \in A} S_t$$

of finite state spaces $S_t = \{s_1, \ldots, s_n\}$.

S_t could, for example, be the alphabet of amino acids at site t and A be the collection of considered sites.

More generally, A may describe a multidimensional domain with a neighborhood structure defined on the sites. In image analysis A is often an area of pixels and S_t comprises possible grey values for pixel t. In the following x_t denotes the t-th component of $x \in \Omega$.

Definition 4.25 *A* ***neighborhood system*** *on a set A is a family $N = \{A_t\}_{t \in A}$ of subsets of A such that for all $t \in A$:*

$$a)\ t \notin A_t \qquad and \qquad b)\ t \in A_s \Rightarrow s \in A_t.$$

The subset A_t is called ***neighborhood*** *of site s.*

The couple (A, N) defines a **graph** with **vertices** A and **edges** N: Sites s and t are linked by an edge if and only if they are neighbors ($s \in A_t$).

Definition 4.26 *Any element $t \in A$ is a **clique**. A subset $C \subset A$ with more than one element is called clique of the graph (A, N) if and only if any two distinct sites of C are mutual neighbors. A clique is called **maximal** if for any site $t \notin C$, $C \cup \{t\}$ is not a clique.*

With neighborhoods and cliques we can define the important term Gibbs potential.

Definition 4.27 *A **Gibbs potential** on $\Omega = \prod_{t \in A} S_t$ relative to a neighborhood system N is a collection $\{V_C\}_{C \subset A}$ of real functions V_C defined on Ω with*
i) $V_C \equiv 0$ if C is not a clique,
ii) for all $x, y \in \Omega$ and all $C \subset A$ the equality $x_t = y_t$ for all $t \in C$ implies

$$V_C(x) = V_C(y).$$

*An energy H **derives from the potential** $\{V_C\}_{C \subset A}$ if*

$$H(x) = \sum_{C:Clique} V_C(x)$$

Remarks 4.28 *Potentials V_C are determined by the values on C alone. These cliques are often small, including only the nearest neighbors. As a consequence, the possible values of the energy function H are also restricted. Therefore, the calculation workload may considerably be reduced if Gibbs potentials are involved.*

The Gibbs sampler uses conditional distributions like the following:

$$P(x_t | x_{A \setminus t}) = \frac{\exp(-H(x_t x_{A \setminus t}))}{\sum_{z_t \in S_t} \exp(-H(z_t x_{A \setminus t}))}. \tag{44}$$

Here we denote as $x_t x_{A \setminus t}$ the (product) state with value x_t at site t and value $x_{A \setminus t}$ for the product of all other sites.

As application of the developed terminology we calculate conditional distributions in the case of the famous Ising model. This model was introduced 1925 by Ising to understand qualitatively the phenomenon of phase transition in ferromagnetic materials.

Example 4.29 *The Ising model is defined on a space $\Omega = \prod_{t \in A} S_t$ with state spaces $S_t = \{-1, 1\}$. There are only 2-element cliques $C = \{s, t\}$ of neighbored elements s and t (denoted here $s \sim t$). The Gibbs potentials are defined by $V_{s,t}(x) = -x_s x_t$ for $s \sim t$ and $V_C \equiv 0$ else. Thus, the energy is given by*

$$H(x) = -\sum_{s \sim t} x_s x_t.$$

For the conditional distributions one gets

$$P(x_s | x_{S \backslash s}) = \frac{\exp(\sum_{s \sim t} x_s x_t)}{\exp(-\sum_{s \sim t} x_t) + \exp(\sum_{s \sim t} x_t)},$$

and especially

$$P(x_s = 1 | x_{S \backslash s}) = \frac{1}{1 + \exp(-2 \sum_{s \sim t} x_t))},$$
$$P(x_s = -1 | x_{S \backslash s}) = \frac{1}{1 + \exp(+2 \sum_{s \sim t} x_t))}.$$

Conditional probabilities, like the above, are encountered in the more general context of Markov random fields.

Definition 4.30 *Given a finite product space $\Omega = \prod_{t \in A} S_t$ with a neighborhood system $N = \{A_t\}_{t \in A}$ as above. The set of random variables $(X_t)_{t \in A}$, taking as values the coordinates x_t, is a **Markov random field** with respect to N, if for all sites t one has:*

$$P(X_t = x_t | X(A \backslash t) = x(A \backslash t)) = P(X_t = x_t | X(A_t) = x(A_t)).$$

*The conditional probabilities $P(X_t = x_t | X(A_t) = x(A_t))$ are called **local characteristics** of the Markov random field.*

Remarks 4.31 *A Gibbs distribution with respect to a neighborhood system is the distribution of a Markov random field with respect to the same neighborhood system. The Gibbs-Markov equivalence theorem moreover states that Markov fields satisfying certain positivity conditions are associated to Gibbs distributions.*

In the field of Gibbs sampling local characteristics may be invoked for the definition of the relevant (homogeneous) Markov chains:

Given a Gibbs distribution P on a finite space $\Omega = \prod_{t \in A} S_t$. With the above notations we define for elements $x, y \in \Omega$ and subsets $I \subset A$

$$P_I(x,y) := \begin{cases} Z_I^{-1} \exp(-H(y_I x_{A \setminus I})) & \text{for } x_{A\setminus I} = y_{A \setminus I} \ ; \\ 0 & \text{else.} \end{cases} \tag{45}$$

The $P_I(x,y)$ describe a transition from x to y which changes x at most on the set I. If I is chosen to consist of one element $t \in A$ only, the values of $P_I(x,y)$ which are not zero are just the local characteristics of the underlying Markov field.

Definition 4.32 *A probability distribution μ and a transition matrix $\boldsymbol{P}$ of a homogeneous Markov chain satisfy the* ***detailed balance*** *equation if one has for all $x, y \in \Omega$*

$$\mu(x)p(x,y) = \mu(y)p(y,x). \tag{46}$$

Remarks 4.33 *Detailed balance means that the homogeneous Markov chain associated to μ and $\boldsymbol{P}$ is* ***reversible in time****.*

Further, μ is invariant for P. This can be seen by summing up (46) with respect to x:

$$\sum_x \mu(x)p(x,y) = \sum_x \mu(y)p(y,x) = \mu(y). \tag{47}$$

These equations constitute the ***global balance***

$$\mu\boldsymbol{P} = \mu.$$

Theorem 4.34 *The Gibbs distribution P and its local characteristics P_I satisfy the detailed balance equations*

$$P(x)P_I(x,y) = P(y)P_I(y,x). \tag{48}$$

In particular, P is invariant for P_I.

Proof. If $x_{A\setminus I} \neq y_{A \setminus I}$ both sides of (48) are zero.

For $x_{A\setminus I} = y_{A\setminus I}$ one has $x = x_I y_{A\setminus I}$ and $y = y_I x_{A\setminus I}$. Therefrom it follows that

$$\exp(-H(x)) \frac{\exp(-H(y_I x_{A\setminus I}))}{\sum_{z_I} \exp(-H(z_I x_{A\setminus I}))} = \exp(-H(y)) \frac{\exp(-H(x_I y_{A\setminus I}))}{\sum_{z_I} \exp(-H(z_I y_{A\setminus I}))}.$$

This equation is just

$$Z \cdot P(x) \cdot P_I(x,y) = Z \cdot P(y) \cdot P_I(y,x),$$

establishing the detailed balance. The above remark shows that P is invariant. ◇

An enumeration $E = \{t_1, \ldots, t_N\}$ of the sites $t \in A$ ($N = |A|$) fixes a **visiting scheme**. We identify t_k and k and define with the associated local characteristics the transition probabilities

$$P_E(x, y) = P_{\{1\}} \cdot \ldots \cdot P_{\{N\}}(x, y). \tag{49}$$

The corresponding Markov chain is realized by the following algorithm:

1) Draw a initial configuration x according to a start distribution (e.g. $\mu = \delta_x$).

2) Update x in position one by y_1 randomly drawn from $P_{\{1\}}(x, y)$. The new configuration $y = y_1 x_{\{2,\ldots,N\}}$ has to be updated at position two.

Continue this way until a **sweep** is finished, that is position N has been reached.

3) Carry out many sweeps.

The justification for the above algorithm is:

Theorem 4.35 *For every initial distributions μ*

$$\lim_{n\to\infty} \mu \, (\boldsymbol{P}_E)^n(x) = P(x)$$

is valid for all configurations $x \in \Omega$.

Proof. The Gibbs distribution P is invariant for P_I and thus also for compositions of local characteristics. Since the probability to get y_t at position t is positive, the transition probability $P_E(x, y)$ is strictly positive. Therefore the Markov chain is primitive and the theorem follows from Theorem 4.11 and Remark 4.12. ◇

4.4 Application to multiple sequence alignment

Different individuals or species may have descended from a common ancestor. Their DNA or protein sequences are then expected to contain domains of great similarity. It is the issue of **multiple sequence alignment** to find such domains and to align the sequences there along. The 'best' multiple alignment is usually found by maximizing an alignment score. If the number of sequences is large, brute force methods are, due to exploding running times, not feasible. Different alternative techniques exist. We discuss an application of Gibbs sampling for ungapped local alignments published in:

Lawrence & al. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262, 208-214.

Assume we have N protein sequences with a background (overall) frequency of the 20 amino acids $p_1, \dots, p_{20}$. The task is to find in every sequence a segment of length W, such that the resulting 20 segments are most 'similar'.

Every possible alignment can be seen as an array with N rows and W columns.

We construct a Markov chain with the possible arrays as states. Transition from one state to another occurs with positive probability only, if states differ at most in one row (the alignment is changed for one sequence only).

Example for an update in row 3:

$$\begin{array}{cccccc} \mathrm{V} & \mathrm{Q} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{N} \\ \mathrm{V} & \mathrm{R} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{R} \\ \mathrm{R} & \mathrm{Q} & \mathrm{B} & \mathrm{A} & \dots & \mathrm{C} \\ \mathrm{V} & \mathrm{Q} & \mathrm{R} & \mathrm{A} & \dots & \mathrm{N} \\ & & & \vdots & & \\ \mathrm{V} & \mathrm{Q} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{N} \end{array} \qquad \longrightarrow \qquad \begin{array}{cccccc} \mathrm{V} & \mathrm{Q} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{N} \\ \mathrm{V} & \mathrm{R} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{R} \\ \mathrm{V} & \mathrm{Q} & \mathrm{B} & \mathrm{A} & \dots & \mathrm{N} \\ \mathrm{V} & \mathrm{Q} & \mathrm{R} & \mathrm{A} & \dots & \mathrm{N} \\ & & & \vdots & & \\ \mathrm{V} & \mathrm{Q} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{N} \end{array}$$

One defines for an amino acid j in column i of an array a the probability estimate

$$q_{i,j}(a) = \frac{c_{i,j}(a) + b_j}{N(a) + B}.$$

Here, $c_{i,j}(a)$ is the count of j in column i of a and $N(a)$ the number of columns in a. The b_j are pseudocounts ($B = \sum_j b_j$) which are introduced to make estimations more 'robust' and to assure that $q_{i,j}(a) > 0$ (compare Theorem 3.5).

The transitions from one state s of the Markov chain to another state is defined as follows:

To update s, a row r is chosen randomly (equal chance).

Then, the probabilities $q_{i,j}(s^{(r)})$ are calculated. The reduced array $s^{(r)}$ results by removing row r from s.

The probability of a segment $x = x_1 x_2 \dots x_W$ with amino acid x_i at position i is under the background probability

$$P_x = p_{x_1} \dots p_{x_W}, \tag{50}$$

and under the estimates for the reduced array

$$Q_x(s^{(r)}) = q_{1,x_1}(s^{(r)}) \dots q_{W,x_W}(s^{(r)}). \tag{51}$$

The likelihood ratios $l_x(s^{(r)}) = Q_x(s^{(r)})/P_x$ define the probabilities

$$P(s,t) = \frac{l_x(s^{(r)})}{\sum_y l_y(s^{(r)})} \tag{52}$$

for a transition to a neighbored alignment t with segment x at row r. The sum in the denominator runs over the $(L_r - W + 1)$ segments of length W (in sequence r with length L_r).

Denote as z the segment originally at row r in state s. Now, the definition of $P(s,t)$ immediately yields:

$$\frac{P(s,t)}{P(t,s)} = \frac{q_{1,x_1}(s^{(r)}) \dots q_{W,x_W}(s^{(r)})}{q_{1,z_1}(s^{(r)}) \dots q_{W,z_W}(s^{(r)})} \cdot \frac{p_{z_1} \dots p_{z_W}}{p_{x_1} \dots p_{x_W}}. \tag{53}$$

Note, that $q_{i,y_i}(s^{(r)}) = q_{i,y_i}(t^{(r)})$, since the probabilities come from the same reduced array.

Further, the relative entropy of the considered probabilities p and q is given by

$$H(q(s)||p) = \sum_{i=1}^{W} \sum_{j=1}^{20} q_{i,j}(s) \log\Big(\frac{q_{i,j}(s)}{p_j}\Big).$$

States s with high relative entropy correspond to good alignments. These strongly differ from alignments randomly selected according to p (compare Section 2.5).

Now, associate to the states s the probabilities

$$\lambda_s = C \cdot \prod_{i=1}^{W} \prod_{j=1}^{20} \Big(\frac{q_{i,j}(s)}{p_j} \Big)^{c_{i,j}(s)}, \tag{54}$$

with a suitable constant C.

Neighbored states s and t differ only in one row r. The corresponding counts $c_{i,j}(s)$ and $c_{i,j}(t)$ for column i are either identical for all 20 values of j or are identical for 18 values of j and differ by $+1$ and -1 for the remaining two values of j.

Therefore it is reasonable to **approximate** $q_{i,j}(s)$ and $q_{i,j}(t)$ by $q_{i,j}(s^{(r)}) = q_{i,j}(t^{(r)})$.

From this approximation and the definition of λ one gets

$$\frac{\lambda_t}{\lambda_s} \approx \frac{q_{1,x_1}(s^{(r)}) \ldots q_{W,x_W}(s^{(r)})}{q_{1,z_1}(s^{(r)}) \ldots q_{W,z_W}(s^{(r)})} \cdot \frac{p_{z_1} \ldots p_{z_W}}{p_{x_1} \ldots p_{x_W}}, \tag{55}$$

where again $x_1, \ldots, x_W$ correspond to row r in state t, and respectively $z_1, \ldots, z_W$ to state s.

Equation (53) yields

$$\frac{\lambda_t}{\lambda_s} \approx \frac{P(s,t)}{P(t,s)}.$$

Thus, the detailed balance equation

$$\lambda_t P(t,s) \approx \lambda_s P(s,t)$$

is (with good approximation) satisfied for neighbored states.

For states not neighbored, the equation is also valid (both sides are zero). Therefore λ is stationary for P.

Now, simulate a Markov chain with the described transition probabilities for neighbored state and any initial distribution.

This is just a Gibbs sampling with **random visiting scheme** (remember the non-deterministic choice of a row r in state s). States with high probability according to the stationary distribution are visited comparatively frequently and may thus be recognized. Furthermore, $\log \lambda_s$ is a linear function of the relative entropy.

Thus frequently visited states and good alignments go together.

4.5 Metropolis-Hastings algorithms

Monte Carlo methods are used to get random samples from a complicated probability distribution μ. The purpose is to calculate from those "observations" quantities (e.g. the mean value of μ) which may not be accessible analytically.

Markov Chain Monte Carlo (MCMC) techniques provide the samples by running a Markov chain with limit distribution μ for a "long" time. An important example is the Gibbs sampling introduced in the preceding sections.

More general is the **Metropolis-Hastings algorithm**:

Assume (again) that the distribution of interest μ is given on a finite space Ω with $\mu(x) > 0$ for all $x \in \Omega$.

Choose a transition matrix Q describing transition probabilities between elements of Ω with entries $q(x, y)$ **strictly positive**.

In this framework the probability distributions (x fixed)

$$P(y|x) = q(x, y)$$

are called **proposal distributions**. From the current state x a new state y is proposed with a given probability.

Now, define the probability $a(x, y)$ that the state y proposed from current state x is **accepted**:

$$a(x, y) = \min(1, \frac{\mu(y)q(y, x)}{\mu(x)q(x, y)}) \tag{56}$$

Theorem 4.36 *A homogeneous Markov chain with transition matrix* $\boldsymbol{P}$ *given by*

$$p(x, y) = q(x, y)a(x, y) \quad for \quad x \neq y$$

and

$$p(x, x) = 1 - \sum_{x \neq y} p(x, y)$$

has limit distribution μ.

Proof. All transition probabilities $p(x, y)$ including $p(x, x)$ are positive (Exercise!).

The Markov chain is thus trivially primitive (irreducible and aperiodic). Theorem 4.11 and Remarks 4.12 state that there is just one stationary distribution which is also the limit distribution for any start distribution.

Next, we show that μ and $\mathbf{P}$ satisfy the detailed balance equation

$$\mu(x)p(x,y) = \mu(y)p(y,x).$$

μ is then invariant for $\mathbf{P}$ (compare Remarks 4.33) and thus the unique limit distribution.

First, assume without loss of generality $\frac{\mu(y)q(y,x)}{\mu(x)q(x,y)} < 1$.

Then, we have

$$a(x,y) = \frac{\mu(y)q(y,x)}{\mu(x)q(x,y)}, \qquad p(x,y) = \frac{\mu(y)q(y,x)}{\mu(x)}$$

and, because of $\frac{\mu(x)q(x,y)}{\mu(y)q(y,x)} > 1$, also

$$a(y,x) = 1, \qquad\qquad p(y,x) = q(y,x).$$

In the case $\frac{\mu(y)q(y,x)}{\mu(x)q(x,y)} = 1$ both equations for $p(x,y)$ and $p(y,x)$ are also valid.

Therefrom, one gets in any case the detailed balance $\mu(x)p(x,y) = \mu(y)p(y,x)$. ◇

Remarks 4.37 *A Metropolis-Hastings algorithm for the simulation of a Markov chain with limit distribution μ is very simple:*

```
Chose a proposal matrix $Q$.

Initialize $X_0 = x$; set $t = 0$.

Repeat {

        Sample a state $y$ from proposal distribution $q(x,.)$

        Sample a Uniform(0,1) random variable $U$

        If $U \leq a(x,y)$ set $X_{t+1} = y$

        otherwise set $X_{t+1} = X_t$

        Increment $t$

    }.
```

Remarks 4.38 *Main problems concerning the application of Metropolis-Hastings algorithms (and MCMC in general) are:*

- *How many iterations are needed to be close to the interesting distribution μ (**burn-in**)?*
- *After burn in, how many further steps are necessary to get good estimations for the quantities of interest (stopping time)?*

*There is a lot of research ongoing to address such questions. Formal tools, like **convergence diagnostics**, have been proposed. But often, when it comes to real applications, visual inspection and thumb rules prevail. Thus, obtained results often need cautious interpretation.*

Check for equilibrium: Gelman-Rubin statisics

One strategy to check the convergence of Markov chains is Gelman-Rubin statisics. Assume we have several runs (say r) of same length N and we monitor some scalar function $f_{it} = f(X_{it})$, where X_{it} is the value of the i-th chain at time t. Convergence of the chains is judged by comparing the **between-sequence variances**

$$B = \frac{N}{r-1}\sum_{i=1}^{r}(\bar{f}_{i.} - \bar{f}_{..})^2, \qquad \text{where} \quad \bar{f}_{i.} = \frac{1}{N}\sum_{t=1}^{N} f_{it}, \quad \bar{f}_{..} = \frac{1}{r}\sum_{i=1}^{r}\bar{f}_{i.}$$

and the **within-sequence variances**

$$W = \frac{1}{r}\sum_{i=1}^{r} s_i^2, \qquad \text{where} \qquad s_i^2 = \frac{1}{N-1}\sum_{t=1}^{N}(f_{it} - \bar{f}_{i.})^2.$$

From B and W one constructs two estimates of the variance of f:

$$\widehat{var}(f) = \frac{N-1}{N}W + \frac{1}{N}B \qquad \text{and} \qquad \sqrt{\hat{R}} = \sqrt{\frac{\widehat{var}(f)}{W}}.$$

$\widehat{var}(f)$ is an **unbiased** estimator ($E(\widehat{var}(f) = var(f)$) of the variance under stationarity. In praxis, chains are only asymptotically stationary. Therefore, $\widehat{var}(f)$ usually overestimates $var(f)$ (called ”conservative” estimator).

On the other hand, the within-sequence variance W usually underestimates $var(f)$ for finite realizations. Both estimators tend to $var(f)$, but from different directions. The ratio of these lower and upper bounds for $var(f)$ is the Gelman-Rubin statistics $\sqrt{\hat{R}}$, also called **estimated potential scale reduction**. Its closeness to 1 is regarded as indictor for equilibrium.

Some prominent strategies for choosing proposal matrices are:

The **metropolis algorithm** (Metropolis et al., 1953) considers only symmetric proposals of the form

$$q(x, y) = q(y, x)$$

for all x, y. The acceptance probability reduces to

$$a(x, y) = \min(1, \frac{\mu(y)}{\mu(x)}).$$

Here, a new state y is always taken, if it is at least as probable as x. If y is less probable than x, it has a chance $\mu(y)/\mu(x)$ to be taken.

A special case is **random-walk Metropolis** with

$$q(x, y) = q(||x - y||),$$

where the dependence is a function of some defined distance $||x - y||$ of x and y.

The **independence sampler** has proposals which do not depend on the current state:

$$q(x, y) = q(y).$$

Here the acceptance is denoted as

$$a(x, y) = \min(1, \frac{w(y)}{w(x)}),$$

with $w(x) = \mu(x)/q(x)$. For the independence sampler to work well the distribution q should be chosen "close" to μ. Recommendation is, to choose q heavier-tailed (more spread out) than μ. This may prevent the Markov chain to get stuck:

Assume, q is chosen lighter-tailed than μ and x is accidentally in the tail of μ. A new proposal y is probably not in the tail of μ. In this case the acceptance $w(y)/w(x)$ is very low and this freezes the state x.

The **single component Metropolis-Hastings** algorithm applies Metropolis-Hastings to the different components of states $x = (x_1, \dots, x_n)$ separately. A candidate for y_k for an update of component k is generated from the proposal distribution

$$q_k(x_k; x_{-k}, y_k).$$

The probability for y_k depends on x_k and all the other components x_{-k} of k. A usual visiting scheme updates the components one after another. In this case component $1, \dots, k-1$ have already been updated when it is k's turn.

The acceptance probability for y_k

$$a(x, y_k) = \min(1, \frac{\mu(y_k|x_{-k}) q_k(y_k; y_{-k}, x_k)}{\mu(x_k|x_{-k}) q_k(x_k; x_{-k}, y_k)}) \tag{57}$$

$\mu(x_k|x_{-k})$ is called the **full conditional distribution** of the k-th coordinate $X., k$.

An important example of a single component technique is **Gibbs sampling**, introduced before. Here, the proposal is the full conditional distribution and acceptance $a(x, y) \equiv 1$. With the Gibbs form of μ one gets exactly (45):

$$p(x, y) = \begin{cases} Z_I^{-1} \exp(-H(y_I x_{A \setminus I})) & \text{for } x_{A \setminus I} = y_{A \setminus I} \ ; \\ 0 & \text{else.} \end{cases} \tag{58}$$

Note, that contrary to Metropolis algorithms the proposal matrix depends on μ.

Remarks 4.39 *Often it is not feasible to calculate the partition function Z for a Gibbs measure*

$$\mu(x) = \frac{1}{Z} \exp(-H(x)).$$

Fortunately, Metropolis-Hastings algorithms usually do not require Z. With a proposal matrix G not depending on μ, one has $a(x, y)$ as function of $H(x) - H(y)$ alone. For independence sampler and Gibbs sampling this is not valid, Z has to be calculated.

Remarks 4.40 *We state some features of ergodic Markov chains which are important for an application of MCMC techniques.*

Irreducible aperiodic (ergodic) Markov chains $(X_n)_{n\in\mathbb{N}_0}$ ***with finite state space and limit distribution*** μ ***satisfy:***

Geometric convergence:

The convergence rate of the chain (convergence towards μ*)is given by the eigenvalue* λ *of* $\boldsymbol{P}$ *with second largest absolute value* $(1 = \lambda_1 > |\lambda_2| \geq \ldots)$*:*

$$\sum_y |p^N(x,y) - \mu(y)| \leq C|\lambda_2|^N. \tag{59}$$

Here C *is a suitable positive constant.*

Ergodicity:

For any real valued function f *which satisfies* $\sum_x f(x)\mu(x) < \infty$ *and every initial distribution* ν *the averages*

$$\overline{f}_N = \frac{1}{N}\sum_{t=1}^{N} f(X_t)$$

converge P_ν*- almost sure to the mean of* f*:*

$$\lim_{N->\infty} \overline{f}_N = \sum_x f(x)\mu(x) = E_\mu(f(X)). \tag{60}$$

The last result is the main incentive to invoke Metropolis-Hastings algorithms.

Proofs and further results may be found in the literature, which is abundant in this area. Good monographs are for example:

Pierre Brémaud: Markov Chains: Gibbs fields, MonteCarlo simulation, and queues. Texts in applied mathematics, 31, Springer, New York 1999.

W.R. Gilks, S. Richardson and D.J. Spiegelhalter: Markov Chain Monte Carlo in Practice. Interdisciplinary statistics, Chapman & Hall/CRC, New York, 1998.

4.6 Simulated Annealing

The algorithms of the last sections simulate from limit distributions of homogeneous Markov chains. In equilibrium, states occur (probably!) frequently if their limit probability is high. In many applications one is interested only in states with highest probability. A lot of numerical strategies exist to find (at least approximately) minima (or maxima) of complicated numeric functions f . Techniques like **steepest descent** start from some initial x_0 and try to find x_1 with $f(x_0) > f(x_1)$. From x_1, x_2 is derived, likewise. Continuing in this way, one gets a sequence $x_0, x_1, x_2, \ldots$, which hopefully converges to a minimum of f. But, in general, such sequences only approach local minima. **Simulated annealing** is invoked to avoid this problem and to find with large probability global minima. We discuss simulated annealing in the context of Gibbs distributions.

As before, let Ω be a finite state space and let the probability of states given by

$$\mu(x) = \frac{\exp(-H(x))}{\sum_y \exp(-H(y))},$$

with some real energy function H.

Definition 4.41 *The* ***Gibbs distribution with temperature*** T ***to energy*** H *is given by* $(T > 0)$

$$\mu_T(x) = \frac{\exp(-\frac{H(x)}{T})}{\sum_y \exp(-\frac{H(y)}{T})}. \tag{61}$$

Example 4.42 *Metropolis sampler with symmetric proposal matrix Q not dependent on T and acceptance probability*

$$a_T(x,y) = \min\{1, \exp(\frac{H(x) - H(y)}{T})\}$$

have limit distributions of form (61).

Now, let M be the set of global minima of energy H and $|M|$ their number.

Theorem 4.43 *As T approaches 0, μ_T is monotonically increasing for $x \in M$ and (finally) monotonically decreasing for $x \notin M$. Further, the following is valid:*

$$\lim_{T \to 0} \mu_T(x) = \begin{cases} \frac{1}{|M|}, & \textit{if } x \in M; \\ \\ 0, & \textit{else.} \end{cases}$$

Proof. Let m be the minimal value of H. Then one has

$$\mu_T(x) \quad = \quad \frac{\exp(-\frac{H(x)}{T})}{\sum_y \exp(-\frac{H(y)}{T})}$$

$$= \frac{\exp(-\frac{(H(x)-m)}{T})}{\sum_{y:H(y)=m} \exp(-\frac{(H(y)-m)}{T}) + \sum_{y:H(y)\neq m} \exp(-\frac{(H(y)-m)}{T})}$$

If x or z are minima, the exponent disappears and the respective summand is 1. The other exponents are strictly negative. The respective summands converge to 0 for $T \to 0$. Therefore, $\mu_T(x)$ increases monotonically to $1/|M|$ if x is a minimum and tends to 0 else.

Now, assume $x \notin M$ and set $a(y) = H(y) - H(x)$ to receive

$$\mu_T(x) = \frac{1}{|H(x) = H(y)| + \sum_{y:a(y)<0} \exp(-\frac{a(y)}{T}) + \sum_{y:a(y)>0} \exp(-\frac{a(y)}{T})}.$$

We have to show, that the denominator is finally growing. Differentiating it with respect to T yields:

$$\sum_{y:a(y)<0} \frac{a(y)}{T^2} \exp(-\frac{a(y)}{T}) + \sum_{y:a(y)>0} \frac{a(y)}{T^2} \exp(-\frac{a(y)}{T}).$$

As $T \to 0$, the second term tends to zero and the first one to ∞. Thus, the derivative finally becomes positive and μ_T decreasing.

◇

Remarks 4.44 *According to the theorem, sampling for "small" values of T provides, almost exclusively, states witch achieve maximal values of the Gibbs distribution.*

Remarks 4.45 *For increasing T, every term in*

$$\frac{\exp(-\frac{H(x)}{T})}{\sum_y \exp(-\frac{H(y)}{T})}$$

tends to 1, and therefore, μ_T tends to the uniform distribution on Ω.

Simulated annealing strategy:

The idea of simulated annealing is borrowed from annealing phenomenons in physics: Compounds may crystallize only if they are slowly annealed from high to low temperatures. Fast cooling may result in a disordered structure corresponding to a local minimum of the free energy.

Simulated annealing is performed by running some Markov chain with limit distribution μ_T. During the run of the chain, the temperature T is "slowly" decreasing.

Remarks 4.46 *According to theorem (4.43), cooling causes μ_T to approach the uniform distribution on global maxima of μ. But, cooling during the run of the Markov chain, causes the chain to be non homogeneous and convergence to a limit distribution is not at all clear. The **cooling schedule**, controlling the decrease of T plays an essential role.*

Several convergence theorems are known for simulated annealing. Without proof, we state one involving the Metropolis sampler.

Theorem 4.47 *There exists a constant $\gamma > 0$ with the property:*

For convergence of Metropolis based simulated annealing, starting from any initial state, it is necessary and sufficient that

$$\sum_{k=1}^{\infty} e^{-\frac{\gamma}{T_k}} = \infty.$$

T_k is the temperature at the step k of the Markov chain.

In particular, a logarithmic cooling schedule

$$T_k = \frac{a}{\ln(k+1)}$$

provides convergence if and only if $a \geq \gamma$.

Proof. Hajek, B.: Cooling schedule for optimal annealing.
Mathematics for Operations Research 13, 311-329, 1988. ◇

Remarks 4.48 *Logarithmic cooling is extremely slow. Especially, if the involved constants are large, low temperatures are not reached in acceptable times. In practice, faster cooling schedules are applied. Convergence becomes then a delicate point, again.*

Remarks 4.49 *To make the proposal matrix in large state spaces more feasible a neighborhood structure is introduced in Ω: Every state x is given a neighborhood $N(x)$ of states. Only these states can be reached from x in one step. This means, $q(x, y) > 0$ for y from the neighborhood of x and $q(x, y) = 0$ else. To get an irreducible Markov chain one has to assure that all states communicate (**communicating neighborhoods**).*

4.7 One example: double digest problem

A **restriction endonuclease** is an enzyme which cuts DNA at specific nucleotide sequences. These molecular scissors have a variety of applications in genetics. In the double digest problem (DDP) two different endonucleases are applied to 3 identical DNA strings of length N. One string is cut by both endonucleases together, the other strings suffer separate applications. As result, one gets three sets of sequence segments (of total length N), which represent different cuts of "the" considered string. The task is to construct from those sets the location of the cuts in the original sequence.

The problem is **NP-complete**, which means, that it has the same complexity as, for example, the travelling salesman problem. For such tasks no polynomial-time algorithms are known and might even not exist.

A thorough discussion of the digest and related problems can be found in Waterman, M.S.: Introduction to Computational Biology. Chapman and Hall, New York, 1995. We only glimpse on the basic ideas of DDP:

Sort the s segments produced by both endonucleases together according to their size $c(i)$:

$$c(1) \leq c(2) \leq \ldots \leq c(s) \tag{62}$$

Let $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, A_m$ be the pieces produced by the respective separate applications. These sets can be represented (jointly) in $m!n!$ different orderings. Each joint ordering l represents $s-1$ cuts which define again s segments. Denote their size as $d_l(i)$ and sort them:

$$d_l(1) \leq d_l(2) \leq \ldots \leq d_l(s) \tag{63}$$

The d-sequences of some l should be equal to (62)

Example: Let the sets of segment length resulting from separate cuts be:

$$\{1, 3, 3, 12\},$$

$$\{1, 2, 3, 3, 4, 6\}$$

and the segment lengths after application of both endonucleases

$$\{1, 1, 1, 1, 2, 2, 2, 3, 6\}.$$

Then the sortings $(1, 3, 12, 3), (2, 4, 6, 3, 3, 1)$ and $(1, 3, 12, 3), (3, 3, 6, 1, 2, 4)$ both produce solutions of the DDP.

Remarks 4.50 *The non uniqueness in the example is typical. Using powerful results from probability theory (Kingman's subadditive ergodic theorem) one can prove that the number of solutions to the DDP increases exponentially with the total string length N (Watermen, section 3.1). This bad performance makes the mapping of long stretches of DNA a difficult task.*

If measurement errors are involved, one looks for d-sequences most similar to the c-sequence.

A common measure for similarity is the χ^2-statistics:

$$f(l) = \sum_{i=1}^{s} \frac{(d_l(i) - c(i))^2}{c(i)} \tag{64}$$

Solving the DDP now means to find the minima of this function.

Waterman discusses the application of simulated annealing. The above f is used as energy function. A neighborhood structure is introduced to the orderings: Two orderings are neighbored if they differ at most in one switch between adjoining segments.

The technique is tested for the bacteriophage λ ($N = 48502$). Here, the the complete sequence and thus the restriction sites are known.

The restrictions enzymes BamHI and EcoRI are applied to λ, producing cuts at the following sites:

BamHI: 5509, 22350, 27976, 34503, 41736

EcoRI: 21230, 26108, 31751, 39172, 44976

The corresponding segment lengths are

$$\{5509, 5626, 6527, 6766, 7233, 16841\},$$
$$\{3526, 4878, 5643, 5804, 7421, 21230\},$$
$$\{1120, 1868, 2564, 2752, 3240, 3526, 3758, 3775, 4669, 5509, 15721\}.$$

Several runs invoking simulated annealing are produced. The initial configurations are randomly chosen and the cooling schedules are proportional to $1/t$. The number of iterations t, required to solve the DDP and thus to recover the restriction map, is in the order of several thousand.

5 Hidden Markov Models

A hidden Markov model (HMM) is an extensions of a discrete-time Markov model. Every state of a HMM emits a letter from some finite output alphabet A. Thus, a running HMM produces a sequence of states $X_1, X_2, \ldots$ and a sequence of output symbols $O_1, O_2, \ldots$. The probability distribution on A, controlling the output, is state dependent, but usually not time dependent. The states of the Markov model themselves are often regarded as not observable (hidden).

As before, we assume a finite state space S, a initial distribution μ on S and a transition matrix $\mathbf{P}$.

The number of states and the transition matrix define the **architecture** of a HMM.

Often it is convenient to make the chain **transient** by adding a **begin state** B and an **end state** E, both producing no output. The chain starts in B, never coming back to B and stays in E (stops there) if it happened to be there.

Essentially, two architectures are distinguished:

Recurrent architecture:

All states (save B and E) are communicating. These states may be visited at any time by the Markov chain.

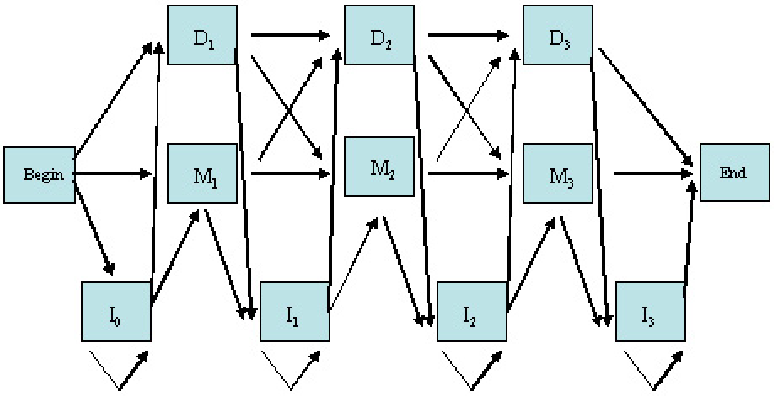
Left-to-right architecture:

Here, a states which may be visited at a certain time are fixed beforehand. The chain propagates, from left to right, through a path of possible states until E is reached.

Example 5.1 ***Mutation-deletion-insertion (MDI) architecture***

MDI is a tool for the aligning multiple DNA or protein sequences. The state space consists of mutation states $(M_1, \ldots, M_L)$*, insertion states* $(I_1, \ldots, I_L)$ *and deletion states* $D_1, \ldots, D_L$*. The output of* M *states are (for example) bases, randomly chosen from a state dependent distribution. Analogously,* I *states produce site-specific insertion of letters.* D *states are considered to produce no output.*

MDI architecture with 3 M states



The next feature of a HMM is the output it produces.

Definition 5.2 *Let $A = \{a_1, \dots a_N\}$ be the set of observation symbols of a HMM. The time independent **emission probability** is for each state x and observation symbol a given by*

$$b_x(a) = P(O_t = a | X_t = x),$$

*defining the $n \times m$ matrix $\boldsymbol{B} = (b_x(a))_{x,a}$. The full **set of parameters** of a HMM is denoted as*

$$\lambda = (\boldsymbol{P}, \boldsymbol{B}, \mu).$$

Given a sequence of observed HMM outputs $O = o_1, o_2, \dots o_T$. Then, in HMM applications one usually wants to answer the following questions:

A) What is the probability of O for a given parameter set λ?

B) What is the hidden sequence $Q = x_1, x_2, \dots, x_T$ of states with highest probability $P(Q|O)$?

C) Which parameter set λ maximizes $P(O|\lambda)$ for fixed graph structure of the underlying Markov chain?

We will describe in the following algorithms which address these questions. Before doing this, we illustrate and motivate the concept with a standard example in genetics: *CpG* islands.

5.1 Some motivation: *CpG* islands

In the human genome wherever the nucleotide CG occurs, the C nucleotide is typically modified and mutates into a T (**methylation**). Write CpG for a dinucleotide to distinguish it from the base pair $C - G$.

Because of the frequent mutation $CpG \longrightarrow TpG$, the CpG dinucleotides are rarer in the genome than would be expected from the independent probabilities of C and G.

Methylation is suppressed in short stretches of the genome (around 'start' regions of genes). here we see more CpG islands than elsewhere. The CpG islands are typically 100-5000 bp long.

Typical questions concerning CpG islands are:

- Given a short stretch of genomic sequence, how would we decide if it comes from a CpG island?
- Given a long piece of sequence, how would we find the CpG islands in it, if there are any?

- The first question can be answered with ordinary Markov chains (Exercise!):

  Two different areas (CpG islands and non-CpG area) have different transition probabilities for the states $\{A, C, G, T\}$. For each kind of area one has a Markov chain model. One is called the + model, the other the − model.

  The transition probabilities $p^+(x, y)$ and $p^-(x, y)$ are estimated from known + and − areas: Define

  $$c^+(x, y) = \text{number of times base y follows x in + areas}$$

  and take the maximum likelihood estimators

  $$p^+(x, y) = \frac{c^+(x, y)}{\sum_z c^+(x, z)}.$$

  Analogously, $p^-(x, y)$ is derived for − areas.

For a given 'short stretch' X we calculate the **sequence probability** with both models and compare the two of them with the log odds ratio

$$S(X) = \log \frac{P(X|model{+})}{P(X|model{-})}.$$

- For the second question a HMM is adequate:

  To simulate in one model the 'islands in a sea of non-island genomic sequence', one combines the Markov chains from the $+$ and the $-$ model in one HMM model.

  Now, the space of hidden states is denoted as

  $$S = \{A^+, C^+, G^+, T^+, A^-, C^-, G^-, T^-\}$$

  and the observation symbols are

  $$\{A, C, G, T\}.$$

  The states A^+, C^+, G^+, T^+ represent $+$ areas and and states A^-, C^-, G^-, T^- representing $-$ areas. The symbols X^+ and X^- exclusively emit the observable symbol X.

  Identifying CpG islands in a genomic sequence corresponds to the search for the most probable hidden sequence in this HMM.

  This is just problem B), stated before.

In the following we present the efficient algorithms which are used to address the problems A) to C).

5.2 The forward and backward algorithm

The **forward algorithm** allows an efficient calculation of the probability $P(O|\lambda)$ mentioned in question A).

Essential for the algorithm is the calculation of the

forward variables

$$\alpha(t,x) = P(o_1,\dots,o_t,X_t = x). \tag{65}$$

This is the joint probability for the sequence of observations $o_1,\dots,o_t$ and that the state at time t is x.

Given the full parameter set λ, $\alpha(t,x)$ can be calculated inductively on t.

initialization :

$$\alpha(1,x) = \mu(x)b_x(o_1) \tag{66}$$

iteration : $\alpha(t+1,x) = \sum_y P(o_1,\dots,o_{t+1},X_t = y, X_{t+1} = x)$ gives

$$\alpha(t+1,x) = \sum_y \alpha(t,y)p(y,x)b_x(o_{t+1}). \tag{67}$$

This algorithm provides $\alpha(T,x)$ for all x.

Therefrom, $P(O|\lambda)$ follows:

$$P(O|\lambda) = \sum_x \alpha(T,x) \tag{68}$$

The algorithm requires on the order of TN^2 computations. This is highly efficient compared to "naive" strategies which calculate the required probability from

$$P(O|\lambda) = \sum_Q P(O|Q,\lambda)P(Q|\lambda) \qquad \text{(exercise!)}.$$

A different approach to solve problem A) is the **backward algorithm**. It runs back in time t. The associated **backward variables** are the conditional probabilities

$$\beta(t,x) = P(o_{t+1},\dots,o_T|X_t = x). \tag{69}$$

The corresponding recursion is:

initialization :

$$\beta(T-1,x) = \sum_y p(x,y) b_y(o_T) \tag{70}$$

and for $t \leq T-1$

iteration :

$$\beta(t-1,x) = \sum_y p(x,y) b_y(o_t) \beta(t,y). \tag{71}$$

From these equations one calculates first $\beta(T-1,x)$ for all x and steps then down to $\beta(1,x)$ to get

$$P(O|\lambda) = \sum_x \mu(x) \beta(1,x) b_x(o_1). \tag{72}$$

Remarks 5.2 *The (posterior) probability $P(X_t = x|O,\lambda)$ of a certain state x at time t, conditioned on the observed sequence $o_1, \ldots, o_T$ can easily be calculated from $P(O|\lambda), \alpha(t,x)$ and $\beta(t,x)$ (exercise!):*

$$P(X_t = x|O,\lambda) = \frac{\alpha(t,x)\beta(t,x)}{P(O|\lambda)}. \tag{73}$$

In some applications one is not interested in the hidden states themselves, but, in derived quantities therefrom. Assume for example, that one has a function $g(x)$ defined on the states. From $P(X_t = x|O,\lambda)$ one could then calculate the derived expression

$$G(t|O) = \sum_x P(X_t = x|O,\lambda) g(x). \tag{74}$$

$G(t|O)$ models some feature of the Markov chain at "time" t. Important special cases have g zero on a subset of states and one on the complement:

Example 5.3 ***Posterior decoding of** CpG **islands***

In the CpG model, introduced above, one may ask if a certain site t belongs to an island or not. One defines $g(x) = 1$ for $x \in \{A^+, C^+, G^+, T^+\}$ and $g(x) = 0$ for $x \in \{A^-, C^-, G^-, T^-\}$. The quantity $G(t|O)$ then defines the (posterior) probability that a base observed at position t is in a CpG island.

5.3 The Viterbi algorithm

The Viterbi algorithm yields a solution to our stated problem B):

A sequence of states which have highest probability $P(Q|O)$ is assigned to the observed sequence O.

First one defines $\delta_t(x)$ as the maximal probability for a path (sequence of states) to end in x at time t and produce output $o_1, \dots, o_t$:

$$\delta_t(x) = \max_{y_1, \cdots y_{t-1}} P(y_1, \dots, y_{t-1}, X_t = x, o_1, \dots, o_t) \tag{75}$$

For the joint probabilities $P(Q, O)$ then

$$\max_Q P(Q, O) = \max_x \delta_T(x)$$

is valid. Since $P(O)$ does not depend on Q, it follows from

$$P(Q|O) = \frac{P(Q, O)}{P(O)}$$

that the maximum of $P(Q, O)$ and the maximum of $P(Q|O)$ are attained at the same Q.

The entities $\delta_t(x)$ can again be calculated by induction:

initialization : For all states x one has

$$\delta_1(x) = \pi(x) b_x(o_1). \tag{76}$$

iteration : For $2 \leq t \leq T$ and all states x set

$$\delta_t(x) = \max_y \delta_{t-1}(y) p(y, x) b_x(o_t). \tag{77}$$

Therefrom, a most probable hidden sequence $\hat{x}_1, \dots, \hat{x}_T$ can be found:

$$\hat{x}_T = \mathrm{argmax}_x \delta_T(x) \tag{78}$$

and for $t \leq T - 1$

$$\hat{x}_t = \mathrm{argmax}_x \delta_t(x) p(x, \hat{x}_{t+1}). \tag{79}$$

Note, that the involved maxima need not be unique. In this case one chooses any solution.

5.4 The Baum-Welch method

Now we are concerned with the problem of identifying the parameters of the HMM, given a predefined topology (question C)).

A) If we have training sequences where all paths of hidden states are known, the parameters can be estimated with maximum likelihood techniques.

This is for example the case in our CpG islands HMM, when all the islands are known beforehand.

In the training set, let $N(x)$ be the **count of first states** being x, $N(x,y)$ the **count of transitions** from state x to state y and $N(x,o)$ the **count of emissions** of observable o from state x.

The maximum likelihood estimators π^{ML} for the initial distribution, $p^{ML}(x,y)$ for the transition probability and $b_x^{ML}(o)$ for the emission probability are then given by

$$\pi^{ML}(x) = \frac{N(x)}{\sum_z N(z)}, \tag{80}$$

$$p^{ML}(x,y) = \frac{N(x,y)}{\sum_z N(x,z)} \tag{81}$$

and

$$b_x^{ML}(o) = \frac{N(x,o)}{\sum_{o'} N(x,o')}. \tag{82}$$

Recall, that if the training set is small, overfitting may occur. In this case it might be recommended to add pseudocounts (compare (3.6) and (3.7)).

B) If the parameters have to be estimated from training sequences with unknown paths (hidden states) the Baum-Welch algorithm is applied.

The Baum-Welch algorithm is an EM algorithm (compare section (3.2)).

Essentially, the unknown numbers $N(x)$, $N(x,y)$ and $N(x,o)$ are substituted by expectations $EN(x)$, $EN(x,y)$ and $EN(x,o)$, calculated from the observations of the training sequences and conditioning on current values of λ.

(**Note of care:** To simplify notations, we often leave out the dependency on λ.)

From $EN(x)$, $EN(x,y)$ and $EN(x,o)$ the estimates for initial, transition and emission probabilities are updated, by again using (80), (81) and (82) (change N to EN).

The mentioned strategy will now be described in more detail.

Let the index g denote training sequence number g.

From equation (73) one gets

$$P(X_1 = x|O, \lambda) = \frac{\alpha(1, x)\beta(1, x)}{P(O)}. \tag{83}$$

Further one has (exercise!)

$$P(X_t = x, X_{t+1} = y|O, \lambda) = \frac{\alpha(t, x)p(x, y)b_y(o_{t+1})\beta(t + 1, y)}{P(O)}. \tag{84}$$

The expected values are obtained by summing up for t and all g:

$$EN(x) = \sum_g \frac{1}{P(O^g)} \alpha^g(1, x)\beta^g(1, x), \tag{85}$$

$$EN(x, y) = \sum_g \frac{1}{P(O^g)} \sum_t \alpha^g(t, x)p(x, y)b_y(o_{t+1})\beta^g(t + 1, y), \tag{86}$$

$$EN(x, o) = \sum_g \frac{1}{P(O^g)} \sum_{t:O_t=o} \alpha^g(t, x)\beta^g(t, x). \tag{87}$$

Remarks 5.4 *The described strategy is rather involving: For each sequence in the training set a forward and a backward algorithm has to be run to get the required variables α and β.*

Values for the EN depend on the size of the training set. This dependence is cancelled in the ratios required for the ML estimators.

Now, we can state the Baum-Welch algorithm.

Baum-Welch algorithm:

Initialization:

Start with some arbitrarily chosen λ.

Choose some initial $EN(x)$, $EN(x, y)$ and $EN(x, o)$.

Recurrence:

1. For all sequences g calculate the forward and backward variables $\alpha^g(t, x)$ and $\beta^g(t, x)$.
2. Calculate new values of $EN(x)$, $EN(x, y)$ and $EN(x, o)$ using (85) (86) and (87).
3. Calculate the new model parameters using (80), (81) and (82) (with N replaced by EN)
4. Calculate the new log likelihood of the model

Termination:

Stop if the change in the log likelihood is less than some threshold.

Else go to 1.

The following two theorems show that the Baum-Welch algorithm is a special case of an EM algorithm.

Recall, that we are looking for maximal values of the log likelihood

$$\log P(O|\lambda) = \log \sum_Q P(O, Q|\lambda).$$

The hidden sequences Q are the missing data, not to be confused with the **Q-function from the EM equations**:

$$Q(\lambda|\lambda_n) = \sum_Q P(Q|O, \lambda_n) \log P(O, Q|\lambda) \tag{88}$$

Theorem 5.5 *The Q-function of the Baum-Welch algorithm is*

$$Q(\lambda|\lambda_n) = \sum_x EN(x)\log(\pi(x)) + \sum_x \sum_o EN(x,o)\log(b_x(o)) + \sum_x \sum_y EN(x,y)\log(p(x,y)).$$

Proof.

Denote for a fixed sequence Q the number of transitions from x to y as $N(x,y|Q)$ and likewise, the number of emissions o from state x as $N(x,o|Q)$. Further, define $N(x|Q)$ as equal to one, if x is the first state in sequence Q and zero else.

The joint probability of an observation O and a hidden sequence Q can be calculated with conditioning on some λ as

$$P(O,Q|\lambda) = \prod_x \pi(x)^{N(x|Q)} \prod_x \prod_o b_x(o)^{N(x,o|Q)} \prod_x \prod_y p(x,y)^{N(x,y|Q)}. \tag{89}$$

Therefrom, one gets with the definition of the Q-function and a current value of λ_n

$$Q(\lambda|\lambda_n) = \sum_Q P(Q|O,\lambda_n)\Big(\sum_x N(x|Q)\log(\pi(x)) + \sum_x \sum_o N(x,o|Q)\log(b_x(o)) \tag{90}$$

$$+ \sum_x \sum_y N(x,y|Q)\log(p(x,y))\Big)$$

Now, the above defined expectations $EN(x)$, $EN(x,y)$ and $EN(x,o)$ calculated under the current model λ_n can be written:

$$EN(x) = \sum_Q P(Q|O,\lambda_n)N(x|Q), \tag{91}$$

$$EN(x,y) = \sum_Q P(Q|O,\lambda_n)N(x,y|Q), \tag{92}$$

$$EN(x,o) = \sum_Q P(Q|O,\lambda_n)N(x,o|Q). \tag{93}$$

Inserting these equations in (90) gives the required representation of the Q function.
◇

Theorem 5.6 *The Q-function $Q(\lambda|\lambda_n)$ of theorem (5.5) is maximized for λ_{n+1} comprising the parameters*

$$p^{(n+1)}(x,y) = \frac{EN(x,y)}{\sum_z EN(x,z)},$$

$$b_x^{(n+1)}(o) = \frac{EN(x,o)}{\sum_{o'} EN(x,o')},$$

$$\pi^{(n+1)}(x) = \frac{EN(x)}{\sum_z EN(z)}),$$

with the EN corresponding to λ_n.

Proof. The difference between $Q(\lambda_{n+1}|\lambda_n)$ with λ_{n+1} defined as above and $Q(\lambda|\lambda_n)$ with any λ is

$$\begin{aligned} Q(\lambda_{n+1}|\lambda_n) - Q(\lambda|\lambda_n) &= \sum_x \sum_y EN(x,y) \log\Big(\frac{p^{(n+1)}(x,y)}{p(x,y)}\Big) \\ & +\sum_x \sum_o EN(x,o) \log\Big(\frac{b_x^{(n+1)}(o)}{b_x(o)}\Big) + \sum_x EN(x) \log\Big(\frac{\pi^{(n+1)}(x)}{\pi(x)}\Big) \\ &= \sum_x \sum_z EN(x,z) \sum_y p^{(n+1)}(x,y) \log\Big(\frac{p^{(n+1)}(x,y)}{p(x,y)}\Big) \\ & +\sum_x \sum_{o'} EN(x,o') \sum_o b_x^{(n+1)}(o) \log\Big(\frac{b_x^{(n+1)}(o)}{b_x(o)}\Big) \\ & +\sum_z EN(z) \sum_x \pi^{(n+1)}(x) \log\Big(\frac{\pi^{(n+1)}(x)}{\pi(x)}\Big). \end{aligned}$$

The last three terms are positive linear combinations of relative entropies. Thus $Q(\lambda_{n+1}|\lambda_n) - Q(\lambda|\lambda_n)$ is nonnegative and zero if one chooses for all x, y and o

$$\begin{aligned} p(x,y) &:= p^{(n+1)}(x,y), \\ b_x(o) &:= b_x^{(n+1)}(o), \\ \pi(x) &:= \pi^{(n+1)}(x). \end{aligned}$$

◇