Lecture Notes on Stochastic Processes

Frank Noé, Bettina Keller and Jan-Hendrik Prinz

July 17, 2013

frank.noe@fu-berlin.de, bettina.keller@fu-berlin.de, jan-hendrik.prinz@fu-berlin.de

DFG Research Center Matheon, FU Berlin, Arnimallee 6, 14195 Berlin, Germany

July 17, 2013

Contents

Chapter 1

Introduction and Overview

1.1 Overview

Markov random processes

	Space Discrete	Space Continuous
Time Discrete	Markov chain	Time-discretized Brownian / Langevin Dynamics
Time Continuous	Markov jump process	Brownian / Langevin Dynamics

Corresponding Transport equations

	Space Discrete	Space Continuous
Time Discrete	Chapman-Kolmogorow	Fokker-Planck
Time Continuous	Master Equation	Fokker-Planck

Examples Space discrete, time discrete: Markov state models of MD, Phylogenetic trees/molecular evolution

time cont: Chemical Reactions

Space cont, time disc: Single Particle Tracking / FRET Experiments, Financial systems

time cont: Particle motion, Molecular Dynamics, Weather system

1.2 Some terms from measure theory

Samples / Outcomes $\omega \in \Omega$ is a sample or outcome from sample space Ω. A sample is the result of a single execution of the model.

Examples:

Coin toss: $\Omega = \{\text{heads}, \text{tails}\}$

Die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Roulette: $\Omega = \{0, ..., 36\}$

Gaussian random variable: $\Omega = \mathbb{R}$.

Events An event E is a set containing zero or more outcomes. Events are defined because in many situations individual outcomes are of little practical use. More complex events are defined in order to characterize groups of outcomes.

Examples:

Die roll: even $= \{2, 4, 6\}$, odd $= \{1, 3, 5\}$

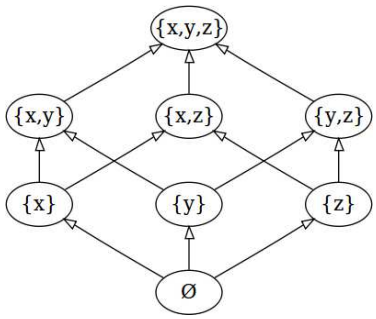
Events may be non-exclusive, e.g.

Roulette: red $= \{1, 3, 5, ...\}$, black $= \{2, 4, 6, ...\}$, $0 = \{0\}$, even, odd, ...

All intervals on the real axis: $\{i_{x_1,x_2} = [x_1, x_2] \mid x_1 \leq x_2; x_1, x_2 \in \mathbb{R}\}$

Algebra In stochastics, the term algebra is used to refer to the event set above, and it is a collection of sets of samples ω with particular properties. It is not related to the term algebra as a field of mathematics.

In order to define algebras, we first introduce the term **Power Set**: Given a set Ω, the power set Ω, written $\mathcal{P}(\Omega)$ or 2^{Ω}, is the set of all subsets of Ω, including the empty set $\emptyset$ and Ω itself. Thus, the elements of $\mathcal{P}(\Omega)$ contain all events that could possibly defined for the sample space Ω. For example the elements of the powerset of $\{x, y, z\}$ are:

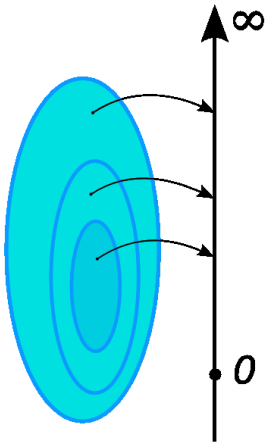


σ-algebra Definition (σ-algebra): *over a set Ω is a nonempty collection Σ of subsets of Ω (including Ω itself) that is closed under complementation and countable unions of its members. It is a Boolean algebra, completed to include countably infinite operations.*

Example: if $\Omega = \{a, b, c, d\}$, one possible σ algebra on Ω is:

$$\Sigma = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}.$$

Measure A measure on a set is a systematic way to assign to each suitable subset a number, intuitively interpreted as the size of the subset. In this sense, a measure is a generalization of the concepts of length, area, volume, *et cetera*. A particularly important example is the **Lebesgue measure** on a Euclidean space, which assigns the conventional length, area and volume of Euclidean geometry to suitable subsets of $\mathbb{R}^n$, $n = 1, 2, 3, ...$ For instance, the Lebesgue measure of [0,1] in the real numbers is its length in the everyday sense of the word, specifically 1. Measures are monotonic in the sense that the measure of a set is nondecreasing when the set is augmented:



Definition (measure): *Let Σ be a σ-algebra over a set Ω. A function $\mu : \Sigma \to \mathbb{R}$ is called a measure if it satisfies the following properties:*

1. *Non-negativity:*
$$\mu(E) \geq 0 \ \forall E \in \Sigma.$$
2. *Null empty set:*
$$\mu(\emptyset) = 0.$$
3. *Countable additivity (or σ-additivity): For all countable collections $\{E_i\}$ of pairwise disjoint sets in Σ:*
$$\mu\Big(\bigcup_{i \in I} E_i\Big) = \sum_{i \in I} \mu(E_i).$$

Measure space Definition (measure space): is a triple (Ω, Σ, μ) containing a nonempty set Ω a σ-Algebra Σ on Ω and a measure $\mu : \Sigma \to \mathbb{R}$

1.3 Probabilities, probability spaces and discrete random variables

Probability measure Definition (probability measure): $\mathbb{P}$ *is a measure with the additional property that* $\mu(\Sigma) = 1$.

This can be ensured by normalizing another measure:

$$\mathbb{P}(E) = \frac{\mu(E)}{\mu(\Sigma)}.$$

We call $\mathbb{P}(E)$ the probability of E. As a consequence, the measure of the complement is:

$$\mathbb{P}(E^C) = 1 - \mathbb{P}(E). \ \forall E \in \Sigma$$

In the case that all samples are equally likely, the probability of a particular event E is given by the ratio of the number of combinations leading to E over the total number of combinations:

$$\mathbb{P}(E) = \frac{n_E}{n_{tot}}$$

For example consider drawing black or white balls from a large urn given that for each draw the probability of black or white is equal. We are interested in the probability of getting k black balls in $n \geq k$ draws, where we are not interested in the sequence of the draw. The number of ways to draw k black balls is given by the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k \cdot (k-1) \cdots 1} \quad \text{if } k \in \{0, 1, \ldots, n\},$$

The total number of combinations is 2^n, and thus the probability is

$$\mathbb{P}(k) = 2^{-n} \binom{n}{k}$$

Conditional Probability Definition (conditional probability): *The conditional probability of event A given B is given by*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, (\text{whence } \mathbb{P}(B) \neq 0)$$

Here $A \cap B$ is the intersection of A and B, that is, it is the event that both events A and B occur.

Independence Definition (independence): *Two events A and B are independent if and only if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

More generally, any collection of events-possibly more than just two of them-are mutually independent if and only if for any finite subset A_1, ..., An of the collection we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i).$$

This is called the multiplication rule for independent events.

Probability space Definition (probability space): *is a measure space $(\Omega, \Sigma, \mathbb{P})$ with a probability measure $\mathbb{P}$. Ω is called sample space and $E \in \Sigma$ are the events. E^C is called counter event. $\mathbb{P}(E)$ is called probability and $\mathbb{P}(E^C)$ counter probability of E.*

Once the probability space is established, it is assumed that "nature" makes its move and selects a single outcome, ω, from the sample space Ω. Then we say that all events from $\mathcal{F}$ containing the selected outcome ω (recall that each event is a subset of Ω) "have occurred". The selection performed by nature is done in such a way that if we were to repeat the experiment an infinite number of times, the relative frequencies of occurrence of each of the events would have coincided with the probabilities prescribed by the function $\mathbb{P}$.

1.4 Random Variables

Given a probabilty space $(\Omega, \Sigma, \mathbb{P})$, a discrete random variable $X(\omega) : \Omega \to \mathbb{R}$ is a map from outcomes to values. We will first take a look at discrete random variables

Example 1: Coin toss. State space $\Omega = \{\text{heads}, \text{tails}\}$. We can introduce a random variable Y as follows:

$$Y(\omega) = \begin{cases} 1, & \text{if } \omega = \text{heads}, \\ 0, & \text{if } \omega = \text{tails}. \end{cases}$$

Example 2: Die roll. State space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The "even" function $e(\omega)$ is given by

$$e(\omega) = \begin{cases} 0, & \text{if one of \{1,3,5\} is rolled}, \\ 1, & \text{if one of \{2,4,6\} is rolled}, \end{cases}$$

Probability density function In probability theory, a probability density function (abbreviated as pdf, or just density) of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space. The probability of a random variable falling within a given set is given by the integral of its density over the set.

A probability density function is most commonly associated with continuous univariate distributions. In the discrete case, the probability density

$$f_X(x) = \mathbb{P}[x]$$

is identical with the probability of an outcome, and is also called probability distribution.

Example 1: coin toss

$$f_Y(y) = \begin{cases} \frac{1}{2}, & \text{if } y = 1, \\ \frac{1}{2}, & \text{if } y = 0. \end{cases}$$

Example 2: die roll

$$f_X(x) = \begin{cases} \frac{1}{6}, & \text{if } x = 1, 2, 3, 4, 5, 6, \\ 0, & \text{otherwise.} \end{cases}$$

We have that probability density can be summed over sets:

$$\mathbb{P}[A] = \sum_{x \in A} \mathbb{P}[x] = \sum_{x \in A} f_X(x)$$

and the cumulative density function is given by:

$$F_X(x_0) = \mathbb{P}[x \leq x_0] = \sum_{x \leq x_0} \mathbb{P}[x] = \sum_{x \leq x_0} f_X[x]$$

As a result of normalization:

$$\begin{aligned} F_X(-\infty) &= 0 \\ F_X(\infty) &= 1 \end{aligned}$$

Moments: expectation, variance, covariance If f is a probability density function, then the value of the following integral above is called the nth moment of the probability distribution.

$$\mu_n = \mathbb{E}(x^n) = \sum_{x \in \Omega} x^n f(x)$$

The first moment is the mean (μ_1 or in short μ). For any two random variables X, Y it holds that

$$\mu_1(X+Y) = \mathbb{E}(X+Y) = \sum_{x \in \Omega} \sum_{y \in \Omega} x\, f(x) + y\, g(y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

The nth central moment of the probability distribution of a random variable X is

$$\mu_n = \mathbb{E}((X-\mu)^n).$$

For example:

$$\begin{aligned} \mu_2 &= Var(x) = \mathbb{E}((x-\mu)^2) \\ &= \mathbb{E}(x^2 - 2x\mu + \mu^2) = \mathbb{E}(x^2) - 2\mu\mathbb{E}(x) + \mu^2 \\ &= \mathbb{E}(x^2) - \mu^2 \end{aligned}$$

For the variance of $X+Y$ we have the integral

$$\begin{aligned} Var(X+Y) &= \mathbb{E}((x-\mu_x) + (y-\mu_y))^2) \\ &= \mathbb{E}((x-\mu_x)^2) + 2\mathbb{E}((x-\mu_x)(y-\mu_y)) + \mathbb{E}((y-\mu_y))^2) \\ &= Var(X) + 2Cov(X,Y) + Var(Y) \end{aligned}$$

Correlation or Pearson's correlation coefficient is defined by:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

For uncorrelated random variables it holds that:

$$Var(X+Y) = Var(X) + Var(Y)$$

as then $Cov(X,Y) = 0$. Note that correlation does not imply independence.

Independence Definition (independent random variables): *Two random variables X and Y are independent if and only if for any numbers a and b the events $\{X \leq a\}$ and $\{Y \leq b\}$ are independent events as defined above. Similarly an arbitrary collection of random variables – possible more than just two of them—is independent precisely if for any finite collection $X_1, ..., X_n$ and any finite set of numbers $a_1, ..., a_n$, the events $X_1 \leq a_1, ..., X_n \leq a_n$ are independent events as defined above.*

For independent random variables it holds for any moments:

$$\mu_n(X+Y) = \mu_n(X) + \mu_n(Y)$$

If two variables are independent, then they are also uncorrelated. The converse of these, i.e. the proposition that if two random variables have a correlation of 0 they must be independent, is not true.

Furthermore, random variables X and Y with probability densities $f_X(x)$ and $f_Y(y)$, are independent if and only if the combined random variable (X,Y) has a joint distribution

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$$

1.5 Central Limit Theorem

The central limit theorem states that when a random effect results from the sum of many independent random variables, each of them having a finite variance both otherwise from arbitrary distributions, then this summed effect will tends

towards a Gaussian distribution. This is the reason why Gaussian distributions are so ubiquitously found: Many real-life processes are complex in the sense that they result from the interaction of many stochastic degrees of freedom, thus the central limit theorem kicks in and produces Gaussian distributions of observed variables.

The proof of the CLT will come later in the continuous random variable section. For now, we'll just give the following statements: Let S_n be the sum of n independent random variables with zero mean and variance σ^2, given by

$$S_n = X_1 + \cdots + X_n.$$

Then, if we define new random variables

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}} = \frac{S_n}{\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}},$$

then

$$\lim_{n\to\infty} \mathbb{P}(Z_n) = \mathcal{N}(0, \sigma^2) :$$

they will converge in distribution to the normal distribution $\mathcal{N}(0, \sigma^2)$ as n approaches infinity. $\mathcal{N}(0, \sigma^2)$ is thus the asymptotic distribution of Z_n. Z_n can also be expressed as

$$Z_n = \sqrt{n}(\overline{X}_n - \mu) = \sqrt{n}\,\overline{X}_n,$$

where

$$\overline{X}_n = \frac{S_n}{n} = \frac{1}{n}(X_1 + \cdots + X_n)$$

is the sample mean. It directly follows that the quantity $Z_n/\sqrt{n} = (\bar{X}_n - \mu) = \bar{X}_n$, i.e. the estimation error of the mean, has a variance of σ^2/n. For any random variable X:

$$Var\{\bar{X}_n - \mathbb{E}(X)\} = \frac{Var(X)}{n}$$

for n independent samples.

Chapter 2

Markov chains

We will now focus our attention to Markov chains and come back to space-continuous processes later. The motivation for this is that whenever studying a space-continuous process in practice, it needs to be in some way discretized, and thus effectively becomes a Markov chain. The numerical properties of such a discretization will be studied later.

Let $X = \{1, ..., n\}$ be a discrete state space and let $x(t)$ be a Markov chain (with the Markov property as defined above) on X, where t may be either discrete or continuous. The system "jumps" between the states of X in time. Such a jump is called a transition.

Time-discrete Markov processes All of the processes we had described here are Markov processes they are defined by the fact that the propagation of the system is entirely determined by knowing its present state x_t, and is thus independent on its past. Formally, for discrete times:

$$\mathbb{P}(x(t) \mid x(t-1),\, x(t-2), ..., x(1)) = \mathbb{P}(x(t) \mid x(t-1))$$

The dynamics then is entirely defined by the transition probabilities $\mathbb{P}(x(t) \mid x(t-1)) : [X \times X] \to \mathbb{R}$. In the state-continuous case this is a transition density that acts between points of the state space. In the state-discrete case this is a transition (probability) matrix, often denoted by $P = [P_{ij}]$ or $T = [T_{ij}]$.

Higher-order Markov chains A Markov chain of order m (or a Markov chain with memory m) where m is finite, is a process satisfying

$$\mathbb{P}(x(t) \mid x(t-1), x(t-2), ..., x(1)) = \mathbb{P}(x(t) \mid x(t-1), x(t-2), ..., x(t-m)) \;\text{ for } t > m$$

In other words, the future state depends on the past m states. It is possible to construct a chain $(y(t))$ from $(x(t))$ which has the 'classical' Markov property as follows:

Let $y(t) = (x(t), x(t-1), ..., y(t-m+1))$, the ordered m-tuple of x values. Then $y(t)$ is a Markov chain with state space X^m and has the classical Markov property. This is a result of Taken's embedding theorem.

Applications (see Wiki pages)

Internet applications The PageRank of a webpage as used by Google is defined by a Markov chain. It is the probability to be at page i in the stationary distribution on the following Markov chain on all (known) webpages. If N is the number of known webpages, and a page i has k_i links then it has transition probability $\frac{\alpha}{k_i} + \frac{1-\alpha}{N}$ for all pages that are linked to and $\frac{1-\alpha}{N}$ for all pages that are not linked to. The parameter α is taken to be about 0.85.

Markov models have also been used to analyze web navigation behavior of users. A user's web link transition on a particular website can be modeled using first- or second-order Markov models and can be used to make predictions regarding future navigation and to personalize the web page for an individual user.

Economics and finance Markov chains are used in Finance and Economics to model a variety of different phenomena, including asset prices and market crashes. The first financial model to use a Markov chain was the regime-switching model of James D. Hamilton (1989), in which a Markov chain is used to model switches between periods of high volatility and low volatility of asset returns. A more recent example is the Markov Switching Multifractal asset pricing model, which builds upon the convenience of earlier regime-switching models It uses an arbitrarily large Markov chain to drive the level of volatility of asset returns.

Dynamic macroeconomics heavily uses Markov chains. An example is using Markov chains to exogenously model prices of equity (stock) in a general equilibrium setting.

Mathematical biology Markov chains also have many applications in biological modelling, particularly population processes, which are useful in modelling processes that are (at least) analogous to biological populations. The Leslie matrix is one such example, though some of its entries are not probabilities (they may be greater than 1). Another important example is the modeling of cell shape in dividing sheets of epithelial cells. The distribution of shapes—predominantly hexagonal—was a long standing mystery until it was explained by a simple Markov Model, where a cell's state is its number of sides.

Empirical evidence from frogs, fruit flies, and hydra further suggests that the stationary distribution of cell shape is exhibited by almost all multicellular animals. Yet another example is the state of Ion channels in cell membranes.

Gambling Markov chains can be used to model many games of chance. Persi Diakonis, a famous mathematician currently in Stanford, has proven several theorems concerning the decorrelation of Markov chains. One such proof shows that the number of times a deck of cards needs to be shuffled in order to be considered to be well shuffled is 7.

2.1 General Properties of space-discrete Markov processes

(also valid for Markov jump processes, see below)

Time-homogeneity Time-homogeneous Markov chains (or stationary Markov chains) are processes where

$$\mathbb{P}(x(t+1) = i | x(t) = j) = \mathbb{P}(x(s+t+1) = i | x(s+t) = j)$$

for all s. The probability of the transition is an invariant property of the system, i.e. independent of the time when we evaluate the Markov chain.

Reducibility A state j is said to be **accessible** from a state i (written $i \to j$) if a system started in state i has a non-zero probability of transitioning into state j at some point. Formally, state j is accessible from state i if there exists a time $t \geq 0$ such that

$$\mathbb{P}(x(t) = j | x(0) = i) > 0.$$

Allowing t to be zero means that every state is defined to be accessible from itself.

A state i is said to **communicate** with state j (written $i \leftrightarrow j$) if both $i \to j$ and $j \to i$. A set of states C is a **communicating class** if every pair of states in C communicates with each other, and no state in C communicates with any state not in C. A communicating class is closed if the probability of leaving the class is zero, namely that if i is in C but j is not, then j is not accessible from i.

Finally, a Markov chain is said to be **irreducible** if its state space is a single communicating class; in other words, if it is possible to get to any state from any state.

Recurrence A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i. Formally, let the random variable T_i be the first return time to state i:

$$T_i = \inf\{t \geq 1 : x(t) = i \mid x(0) = i\}.$$

Then, state i is **transient** if and only if:

$$\mathbb{P}(T_i = \infty) > 0.$$

If a state i is not transient (it has finite hitting time with probability 1), then it is said to be recurrent or persistent. Although the hitting time is finite, it need not have a finite expectation. Let M_i be the expected return time,

$$M_i = \mathbb{E}[T_i].$$

Then, state i is **positive recurrent** if M_i is finite; otherwise, state i is **null recurrent** (the terms non-null persistent and null persistent are also used, respectively).

A state i is called **absorbing** if it is impossible to leave this state.

Steady-state analysis and limiting distributions A Markov process has a unique stationary distribution $\boldsymbol{\pi}$ if and only if it is irreducible and all of its states are positive recurrent. Note that this in particular includes ergodic processes, because ergodicity is a stronger requirement. In that case, π is related to the expected return time:

$$\pi_i = \frac{1}{M_i}.$$

Further, if the chain is both irreducible and aperiodic, then for any i and j,

$$\lim_{t \to \infty} \mathbb{P}(x(t) = j \mid x(0) = i) = \frac{1}{M_j}.$$

Note that there is no assumption on the starting distribution; the chain converges to the stationary distribution regardless of where it begins. Such π is called the equilibrium distribution of the chain.

If a chain has more than one closed communicating class, its stationary distributions will not be unique (consider any closed communicating class in the chain; each one will have its own unique stationary distribution. Any of these will extend to a stationary distribution for the overall chain, where the probability outside the class is set to zero).

2.2 Time-Discrete Markov Chains

A Markov chain on X, named after Andrey Markov, is a discrete random process with the Markov property. A discrete random process means a system which can be in various states, and which changes randomly in discrete steps. It can be helpful to think of the system as evolving once a minute, although strictly speaking the "step" may have nothing to do with time. The Markov property states that the probability distribution for the system at the next step (and in fact at all future steps) only depends on the current state of the system, and not additionally on the state of the system at previous steps. Since the system changes randomly, it is generally impossible to predict the exact state of the system in the future. However, the statistical properties of the system at a great many steps in the future can often be described. In many applications it is these statistical properties that are important.

An example of a Markov chain is a random walk on the number line which starts at zero and transitions +1 or −1 with equal probability at each step. The position reached in the next transitions only depends on the present position and not on the way this present position is reached.

Transition Matrix We define the propagator / transition matrix $P \in \mathbb{R}^{n \times n}$ which describes the evolution of the chain:

$$\begin{aligned} p_{ij} &\geq 0 \; \forall i, j \\ \sum_{j=1...n} p_{ij} &= 1 \; \forall i \end{aligned}$$

These conditions define a *stochastic* matrix. p_{ij} represents the transition probability of state i to state j within one step of the system (which may correspond to a fixed physical time step τ):

$$p_{ij} = \mathbb{P}[x(t+1) = j \mid x(t) = i]$$

Markov chain (definition) We are given a countable set $X = \{1, ..., n\}$ called state space, a stochastic matrix $P \in \mathbb{R}^{n \times n}$, and a distribution $p(0) \in \mathbb{R}^n$. The series of random variables $x(t)$, $t = 0, 1, ..., T$ is called Markov chain of length T with initial distribution $p(0)$ and transition matrix P if:

- $x(0)$ has distribution $p(0)$
- $x(t+1)$ has distribution $p_{x(t)}$, i.e. the $x(t)$-th row of P for all $t \geq 0$.

That is, the next value of the chain depends only on the current value, not any previous values. Thus it is often said that "Markov chains have no memory". Note that this is not exactly correct: Markov chains have memory depth 1. A process without memory would be given by a sequence of independently drawn random variables

Evolution P can be used to generate trajectories according to the definition above:

1. Draw $x(1)$ from the initial distribution p_0
2. Draw $x(t+1)$ from the discrete distribution $[p_{x(t),1}, ..., p_{x(t),n}]$

Alternatively, we can also consider the probability to be in a given state i at time t, $p_i(t)$. This can be viewed as the fraction of particles in each state. The entire distribution of the system is described by the vector $p_t \in \mathbb{R}^n$. Thus the time evolution can be described by the set of equations:

$$p_j(t+1) = \sum_i p_{ij} p_i(t) \; \forall j$$

which is equivalently written by the matrix equation

$$p(t+1)^T = p(t)^T P$$

Chapman-Kolmogorow Equation We use the transition matrix twice and obtain:

$$\begin{aligned} p_j(t+2) &= \sum_i p_{ij} p_i(t+1) \\ &= \sum_i \sum_k p_{ki} p_{ij} p_k(t+1) \end{aligned}$$

or

$$p(t+2)^T = p(t)^T P^2$$

Note that this is possible because of Markovianity. Think about the form of $p_j(t+2)$ if the transition probabilities would not only depend on the current, but also the previous state.

P can be used to transport this probability vector over longer times:

$$p(t)^{\mathrm{T}} = p(0)^T P^t.$$

The probability to go from i to j in t time steps is:

$$\mathbb{P}(x(t) = j \mid x(0) = i) \quad = \quad [P^t]_{ij}$$

Stationary Distribution A Markov process has a unique stationary distribution π if and only if it is irreducible and all of its states are positive recurrent. The stationary distribution, by definition is the distribution that is unchanged by the action of P:

$$\pi^T = \pi^T P$$

we can also write this as

$$\begin{aligned} \pi^T P &= 1\pi^T \\ P^T \pi &= 1\pi \end{aligned}$$

and see that this is an eigenvalue equation. Thus, π (and any multiple $c\pi$, $c \in \mathbb{R}$) is a left eigenvector of P with eigenvalue 1. Since an irreducible and positively recurrent chain has a unique stationary distribution, for a chain the eigenvalue 1 is unique. For all other eigenvalues, it can be shown that they have a norm strictly smaller than 1.

We can also ask for the corresponding right eigenvalue and find that

$$P\mathbf{1} = \mathbf{1}$$

Proof:

$$P \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \sum_j p_{1j} 1 = \sum_j p_{1j} \\ \vdots \\ \sum_j p_{nj} 1 = \sum_j p_{nj} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Convergence towards the stationary distribution For any diagonalizable matrix P we can write:

$$\begin{aligned} Pr_i &= r_i\lambda_i \\ P\begin{bmatrix} r_1 & \cdots & r_n \end{bmatrix} &= \begin{bmatrix} r_1 & \cdots & r_n \end{bmatrix}\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \\ PR &= R\Lambda \\ P &= R\Lambda R^{-1} \end{aligned}$$

and likewise

$$\begin{aligned} l_iP &= \lambda_i l_i \\ \begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} P &= \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}\begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} \\ LP &= \Lambda L \\ P &= L^{-1}\Lambda L \end{aligned}$$

or

$$P = R\Lambda L$$

We assume a chain that is irreducible and positively recurrent. We consider the Chapman-Kolmogorow equation:

$$p(t)^T = p(0)^T P^t$$

and diagonalize P:

$$\begin{aligned} p(t)^T &= p(0)^T (R\Lambda L)^t \\ &= p(0)^T R\Lambda^t L \\ &= p(0)^T \sum_{i=1}^{n} \lambda_i^t r_i l_i^T \end{aligned}$$

Since the chain is irreducible and positively recurrent, it will have a single eigenvalue 1 and otherwise eigenvalues with norm <1. We order them as follows:

$$\lambda_1 = 1, \lambda_2, ..., \lambda_n \quad \text{with } |\lambda_k| \leq |\lambda_{k+1}| \quad \text{for } k > 1.$$

and can thus write:

$$\begin{aligned} p(t)^T &= p(0)^T \lambda_1^t r_1 l_1^T + p(0)^T \sum_{i=2}^{n} \lambda_i^t r_i l_i^T \\ &= p(0)^T \mathbf{1} \pi^T + p(0)^T \sum_{i=2}^{n} \lambda_i^t r_i l_i^T \end{aligned}$$

thus, the infinite time distribution is

$$\begin{aligned} p_\infty^T &= \lim_{t \to \infty} p(t)^T \\ &= p(0)^T \mathbf{1} \pi^T + \lim_{t \to \infty} \left(p(0)^T \sum_{i=2}^{n} \lambda_i^t r_i l_i^T \right) \end{aligned}$$

but since $|\lambda_i| < 1$ for all $i > 1$, the terms on the right cancel and we see that the infinite time distribution is indeed π:

$$\begin{aligned} p_\infty^T &= p(0)^T \mathbf{1} \pi^T \\ &= p(0)^T \begin{bmatrix} \pi_1 & \cdots & \pi_n \\ \vdots & & \vdots \\ \pi_1 & \cdots & \pi_n \end{bmatrix} \\ &= \begin{bmatrix} \pi_1 \sum_j p_j(0) & \cdots & \pi_n \sum_j p_j(0) \end{bmatrix} \\ &= \begin{bmatrix} \pi_1 & \cdots & \pi_n \end{bmatrix} \\ &= \pi^T \end{aligned}$$

Speed of convergence We are interested in how fast the stationary distribution is reached. Consider the error

$$\begin{aligned} E(t) &= \left\| \pi^T - p(t)^T \right\| \\ &= \left\| \pi^T - \pi^T - p(0)^T \sum_{i=2}^{n} \lambda_i^t r_i l_i^T \right\| \\ &= \left\| p(0)^T \sum_{i=2}^{n} \lambda_i^t r_i l_i^T \right\| \\ &= \left\| p(0)^T \lambda_2^t r_2 l_2^T + p(0)^T \sum_{i=3}^{n} \lambda_i^t r_i l_i^T \right\| \end{aligned}$$

Asymptotically, i.e. for times $t \gg -1/\ln|\lambda_3|$, all terms on the right will approximately zero, and we obtain for this regime:

$$E(t) \approx \lambda_2^t \left\| p(0)^T r_2 l_2^T \right\|$$

which means the error decays exponentially with λ_2^t. When comparing to an exponential decay function $\exp(-kt)$, The rate of this decay is $k_2 = -\ln\lambda_2$ and the timescale of this decay is $t_2 = k_2^{-1} = -1/\ln\lambda_2$.

Note that if the initial distribution already happens to be π^T, we see that we get the coefficients: $\|\pi^T r_i l_i^T\| = \|l_1^T r_i l_i^T\|$ for $i > 1$. However, as a result of the diagonalization we have $l_i^T r_j = \langle l_i, r_j \rangle = 0$ for all $i \neq j$, and thus all coefficients become 0, showing that convergence is indeed immediate in this case.

Periodicity A state i has period k if any return to state i must occur in multiples of k time steps. Formally, the period of a state is defined as

$$k = gcd\{n : \mathbb{P}(X_n = i | X_0 = i) > 0\}$$

(where "gcd" is the greatest common divisor). Note that even though a state has period k, it may not be possible to reach the state in k steps. For example, suppose it is possible to return to the state in {6, 8, 10, 12, ...} time steps; then k would be 2, even though 2 does not appear in this list.

If $k = 1$, then the state is said to be **aperiodic** i.e. returns to state i can occur at irregular times. Otherwise ($k > 1$), the state is said to be periodic with period k.

It can be shown that every state in a communicating class must have overlapping periods with all equivalent-or-larger occurring sample(s).

Ergodicity A state i is said to be ergodic if it is aperiodic and positive recurrent. If a Markov process is irreducible and all its states are ergodic, then the process is said to be ergodic.

It can be shown that a finite state irreducible Markov chain is ergodic if it has an aperiodic state.

Reversible Markov chain / Detailed balance A Markov chain is said to be reversible if there is a π such that

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

This condition is also known as the **detailed balance** condition.

Summing over i gives

$$\sum_i \pi_i p_{ij} = \pi_j$$

so for reversible Markov chains, π is always a stationary distribution.

The idea of a reversible Markov chain comes from the ability to "invert" a conditional probability using Bayes' Rule:

$$\begin{aligned}
\mathbb{P}(x(t) = i \mid x(t+1) = j) &= \frac{\mathbb{P}(x(t) = i,\, x(t+1) = j)}{\mathbb{P}(x(t+1) = j)} \\
&= \frac{\mathbb{P}(x(t) = i)\mathbb{P}(x(t+1) = j \mid x(t) = i)}{\mathbb{P}(x(t+1) = j)} \\
&= \frac{\pi_i}{\pi_j}\mathbb{P}(x(t+1) = j \mid x(t) = i) \\
p_{ji} &= \frac{\pi_i}{\pi_j} p_{ij}.
\end{aligned}$$

It now appears as if time has been reversed.

We can consequently define a **backward propagator**, which will be used later, as:

$$\tilde{p}_{ij} = \frac{\pi_j}{\pi_i} p_{ji}$$

which is in the reversible case identical to the forward propagator ($\tilde{p}_{ij} = p_{ij} \forall i, j$).

Further Reading:

1. J. Norris: "Markov Chains". Cambride University Press.
   Parts available from: http://www.statslab.cam.ac.uk/~james/Markov/

2.3 Markov chain Monte Carlo (MCMC)

Monte Carlo idea The idea of Monte Carlo methods is to approximate a deterministic quantity with a probabilistic method. While the idea is quite old, the first use of the term Monte Carlo can be traced back to von Neumann and Ulam in the 1940s, who collaborated on the Manhattan project in Los Alamos.

For example, given that a circle inscribed in a square and the square itself have a ratio of areas that is $\pi/4$, the value of π can be approximated using a Monte Carlo method (see Wiki page on Monte Carlo):

1. Draw a square on the ground, then inscribe a circle within it.
2. Uniformly scatter N objects of uniform size (e.g. grains of rice) over the square.
3. Count the number of objects inside the circle, N_x.
4. The ratio N_x/N is an estimate of $\pi/4$. Multiply the result by 4 to estimate π.

Monte Carlo methods are often useful in cases where integrals need to be computed and deterministic numerical approaches fail, e.g. because the domain cannot be explicitly written or the integration space is too high-dimensional.

Generally, if we draw N samples $(x_1, ..., x_N)$ from a discrete domain X and accept them according to probability distribution $\mathbb{P}(x)$, then we can approximate the probability of a point x as

$$\lim_{N\to\infty} \frac{N_x}{N} = \mathbb{P}(x)$$

which allows expectations to be calculated, e.g.

$$\begin{aligned} \mathbb{E}(x) &= \sum_{x\in\Omega} x\mathbb{P}(x) = \lim_{N\to\infty} \frac{1}{N}\sum_i x_i \\ \mathbb{E}(a(x)) &= \sum_{x\in\Omega} a(x)\mathbb{P}(x) = \lim_{N\to\infty} \frac{1}{N}\sum_i a(x_i) \end{aligned}$$

or in the continuous case, where $f(x)$ is the probability density:

$$\begin{aligned}\mathbb{E}(x) &= \int_{x\in\Omega} dx\, x\, f(x) = \lim_{N\to\infty} \frac{1}{N}\sum_i x_i \\ \mathbb{E}(a(x)) &= \int_{x\in\Omega} dx\, a(x)\, f(x) = \lim_{N\to\infty} \frac{1}{N}\sum_i a(x_i)\end{aligned}$$

Metropolis Monte Carlo Metropolis Monte Carlo[1] was the first MCMC algorithm proposed. MCMC algorithms use the Monte Carlo idea above but construct a Markov Chain of Monte Carlo moves that has an invariant distribution identical to the desired sample distribution $\mathbb{P}$. The MCMC approach is useful for estimating probability density functions for which probability ratios $\mathbb{P}(x_1)/\mathbb{P}(x_2)$ are easy to calculate, but integrals $\sum_{x\in A}\mathbb{P}(x)$ are hard or impossible to calculate analytically or with deterministic numerical methods because the domain A is very complex or the integration space is very high-dimensional. Moreover, if there is only a small fraction of Ω where $\mathbb{P}(x)$ is significantly greater than 0, then application of the direct Monte Carlo method (see circle example above) is not useful, as it leads to mostly low-probability samples.

The Metropolis Monte Carlo algorithm is a special MCMC method designed for physical or chemical systems that have an energy $E(x)$ and where the stationary distribution is given by the Boltzmann distribution

$$\mathbb{P}(x) = Z^{-1}\exp(-\beta E(x))\,, \tag{2.1}$$

x is the the state of the system, *i.e.* the conformation of a molecule, Z is the partition function,

$$Z = \sum_{x\in\Omega}\exp(-\beta E(x))$$

with $\beta = \frac{1}{k_B T}$ where k_B is the Boltzmann constant and T is the temperature. We assume that we have a model that allows us to calculate $E(x)$ and thus also $\exp(-\beta E(x))$ for a given x, but we cannot calculate probabilities of large sets of states, especially Z, because we simply cannot afford to enumerate all states. Metropolis Monte Carlo defines a Markov chain with propagator P with the following properties:

1. The chain is reversible, i.e. we have a stationary distribution π with $\pi_i p_{ij} = \pi_j p_{ji}$
2. The stationary distribution is exactly the distribution we want to sample from $\pi_i = \mathbb{P}(i)$.

[1] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller, "Equation of State Calculations by Fast Computing Machines", Journal of Chemical Physics **21**, 1087 (1953)

If we can construct a Markov chain we can generate realizations $x(t)$ and estimate expectation values from it. For a given distribution $\mathbb{P}(x)$ there are many ways to construct a Markov chain which fulfills the above conditions and is thus able to generate estimates of $\mathbb{P}(x)$. They differ mainly by their computational efficiency for different problems.

In the following we sketch the Metropolis Monte Carlo algorithm:

1. Pick a starting state x_0
2. For $k = 0$ to $N - 1$:
   (a) Pick a trial conformation $x_k \to x'$
   (b) Calculate the probability ratio
   $$\frac{\mathbb{P}(x')}{\mathbb{P}(x_k)} = \exp(-\beta[E(x') - E(x_k)])$$
   (c) Accept the trial move with probability
   $$p_{acc} = \min\{1, \frac{\mathbb{P}(x')}{\mathbb{P}(x_k)}\}$$
   this can be realized by drawing a random number $k \in [0, 1[$ and accepting if $k \leq p_{acc}$.
   (d) If the trial move is accepted, $x_{k+1} = x'$, else $x_{k+1} = x_i$

By looking at the acceptance criterion, we see that:

$$\begin{array}{lcll} E(x') > E(x_i) & \Rightarrow & \exp(-\beta\Delta E) < 1 & \Rightarrow \quad \text{accepted if } k < \exp(-\beta\Delta E) \\ E(x') \leq E(x_i) & \Rightarrow & \exp(-\beta\Delta E) \geq 1 & \Rightarrow \quad \text{accepted if } k < 1, \textit{i.e.} \text{ always} \end{array}$$

The algorithm is correct, i.e. it indeed generates a reversible Markov chain with stationary distribution $\mathbb{P}(x)$, provided that following requirements are fulfilled:

1. The Boltzmann distribution can be evaluated at any point x.
2. The probability of making a trial move $x \to x'$ is equal to the probability of making the reverse trial move, *i.e.* $\mathbb{P}(x \to x') = \mathbb{P}(x' \to x)$
3. Any point x' can be reached from any other point x by making trial moves, *i.e.* the Markov chain is irreducible.

Proof of Metropolis Monte Carlo

- Unique stationary distribution: Assumption 2 means that the constructed Markov chain is positively recurrent, Assumption 3 means that it is irreducible. As a result, the constructed propagator P has a unique stationary distribution.
- Reversibility: We call the probability to propose a move x_i to x_j is $\mathbb{P}(x_i \to x_j)$ and vice versa $\mathbb{P}(x_j \to x_i)$. The probability to move from x_i to x_j, i.e. to propose and accept, are p_{ij} and vice versa p_{ji}. Without restriction of generality we assume $E(x_i) < E(x_j)$. We find that:

$$\begin{aligned} p_{ij} &= \mathbb{P}(x_i \to x_j) \exp(-\beta[E(x_j) - E(x_i)]) = \mathbb{P}(x_i \to x_j) \frac{\exp(-\beta E(x_j))}{\exp(-\beta E(x_i))} = \mathbb{P}(x_i \to x_j) \frac{\mathbb{P}(x_j)}{\mathbb{P}(x_i)} \\ p_{ji} &= \mathbb{P}(x_j \to x_i) \end{aligned}$$

by construction,

$$\mathbb{P}(x_i \to x_j) = \mathbb{P}(x_j \to x_i)$$

$$p_{ij} \frac{\mathbb{P}(x_i)}{\mathbb{P}(x_j)} = p_{ji}$$

resulting in the equations:

$$\mathbb{P}(x_i) p_{ij} = \mathbb{P}(x_j) p_{ji} \quad \forall i, j$$

- It follows that P is a reversible chain with respect to the unique stationary distribution $\pi = \mathbb{P}(x)$.

Metropolis-Hastings algorithm The Metropolis-Hastings algorithm is a generalization of the above idea by Hastings [2]. Let us not assume that the proposal probabilites are symmetric, i.e. in general $\mathbb{P}(x_i \to x_j) = \mathbb{P}(x_j \to x_i)$ is not true for all i, j. Let us consider two points x_i, x_j with $\mathbb{P}(x_i) \geq \mathbb{P}(x_j)$ and accept $x_i \to x_j$ with probability $\frac{\mathbb{P}(x_j)\mathbb{P}(x_j \to x_i)}{\mathbb{P}(x_i)\mathbb{P}(x_i \to x_j)}$ and the reverse move with probability 1. We reconsider the proof of the Metropolis Monte Carlo algorithm:

$$\begin{aligned} p_{ij} &= \mathbb{P}(x_i \to x_j) \frac{\mathbb{P}(x_j)\mathbb{P}(x_j \to x_i)}{\mathbb{P}(x_i)\mathbb{P}(x_i \to x_j)} = \frac{\mathbb{P}(x_j)\mathbb{P}(x_j \to x_i)}{\mathbb{P}(x_i)} \\ p_{ji} &= \mathbb{P}(x_j \to x_i) \end{aligned}$$

[2] Hastings, W.K., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", Biometrika **57**, pp. 97-109 (1970).

such that again:

$$p_{ij}\mathbb{P}(x_i) = p_{ji}\mathbb{P}(x_j).$$

This provides the following algorithm:

1. Pick a starting state x_0
2. For $k = 0$ to $N - 1$:
   (a) Propose a trial conformation $x_k \to x'$ with probability $\mathbb{P}(x_i \to x')$
   (b) Accept the trial move with probability

$$p_{acc} = \min\{1, \frac{\mathbb{P}(x')\mathbb{P}(x' \to x_i)}{\mathbb{P}(x_i)\mathbb{P}(x_i \to x')}\}$$

   (c) If the trial move is accepted, $x_{k+1} = x'$, else $x_{k+1} = x_i$

Additionally to the requirements of the Metropolis Monte Carlo method we need to be able the proposal probabilities $\mathbb{P}(x_i \to x')$.

Remarks on the processes and convergences Consider the Markov process $x(t)$ with stationary probability π and let us assume that we can simulate its dynamics with the propagator P_x. Along the previous analysis, the direct simulation of this process will converge to π asymptotically with rate $k_x = -\ln \lambda_x$, where λ_x is the largest eigenvalue of P_x that is smaller than 1.

We can construct a MCMC method which samples π via a different propagator P_y. This MCMC process will converge to π asymptotically with rate $k_y = -\ln \lambda_y$, where λ_y is the largest eigenvalue of P_y that is smaller than 1.

Thus, it is important to distinguish the the original process and the MCMC process constructed to sample the stationary distribution of the original process. We can construct different processes that sample the same stationary distribution. Clearly, depending on λ_y these processes can have very different efficiencies. Ideally we would chose a strategy with $\lambda_y < \lambda_x$ and thus obtain a faster convergence than the original process itself.

Chapter 3

Analysis of Markov Chains

3.1 Hitting and Splitting/Committor Probabilities

Hitting probabilites for Markov Chains Given a stochastic process on state space $X = \{1, ..., n\}$, the hitting time of set A: $H^A : X \to \{0, 1, 2, ...\} \cup \{\infty\}$ is defined as:

$$H_i^A = \inf\{t \geq 0 : \ x(t) \in A | x(0) = i\}$$

and the hitting probability is the probability that starting from i we ever hit A:

$$h_i^A = \mathbb{P}(H_i^A < \infty).$$

For a Markov chain on X with transition matrix P, the vector of hitting probabilities $h^A = (h_i^A : i \in I)$ is the minimal non-negative solution to the system of linear equations:

$$\begin{aligned} h_i^A &= 1 \text{for } i \in A \\ h_i^A &= \sum_{j \in X} p_{ij} h_j^A \text{for } i \notin A. \end{aligned}$$

(Minimality means that if $x = (x_i : i \in X)$ is another solution with $x_i \geq 0$ for all i, then $x_i \geq h_i$ for all i.)

Proof:

First we show that h^A satisfies the above equation.

1) If $x(0) = i \in A$, then $H_i^A = 0$, so $h_i^A = 1$.

2) If $x(0) = i \notin A$, then $H_i^A \geq 1$, so by the Markov property:

$$\mathbb{P}(H_i^A < \infty \mid x(1) = j) = \mathbb{P}(H_j^A < \infty) = h_j^A$$

and

$$\begin{aligned} h_i^A &= \mathbb{P}(H_i^A < \infty) \\ &= \sum_{j \in X} \mathbb{P}(H_i^A < \infty, x(1) = j) \\ &= \sum_{j \in X} \mathbb{P}(H_i^A < \infty \mid x(1) = j)\mathbb{P}(x(1) = j \mid x(0) = i) \\ &= \sum_{j \in X} h_j^A p_{ij}. \end{aligned}$$

Next, we show that h^A are the minimal solution

1) Suppose that $x = (x_i : i \in I)$ is any solution to the equation. then $h_i^A = x_i = 1$ for $i \in A$.

2) Suppose $i \notin A$, then:

$$x_i = \sum_{j \in X} p_{ij} x_j = \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} x_j$$

Substitute for x_j to obtain:

$$\begin{aligned} x_i &= \sum_{j \in X} p_{ij} x_j = \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} \left(\sum_{k \in A} p_{jk} + \sum_{k \notin A} p_{jk} x_k \right) \\ &= \mathbb{P}(x(1) \in A \mid x(0) = i) + \mathbb{P}(x(1) \notin A, x(2) \in A \mid x(0) = i) + \sum_{j \notin A} \sum_{k \notin A} p_{ij} p_{jk} x(k). \end{aligned}$$

By repeated substitution for x in the final term, we obtain after n steps:

$$\begin{aligned} x_i \quad = \quad & \mathbb{P}(x(1) \in A \mid x(0) = i) + ... + \mathbb{P}(x(1) \notin A, ..., x(n-1) \notin A, x(n) \in A \mid x(0) = i). \\ & + \sum_{j_1 \notin A} ... \sum_{j_n \notin A} p_{ij_1} p_{j_1 j_2} ... p_{j_{n-1} j_n} x_{j_n} \end{aligned}$$

Now if x is non-negative, so is the last term on the right, and the remaining terms sum to $\mathbb{P}(H_i^A < n)$. So $x_i \geq \mathbb{P}(H_i^A < n)$ for all n and then:

$$x_i \geq \lim_{n \to \infty} \mathbb{P}(H_i^A < n) = \mathbb{P}(H_i^A < \infty) = h_i.$$

Committor Probabilities The committor probability, q_i^+ pertaining to two sets A, B is the probability that starting in state i, we will go to B next rather than to A:

$$q_i^+ = \mathbb{P}_i(H^B < H^A).$$

In order to compute this, we define an A-absorbing process as

$$\hat{p}_{ij} = \begin{cases} p_{ij} & i \notin A, j \in S \\ 1 & i \in A, i = j \\ 0 & i \in A, i \notin j \end{cases}$$

and then compute the hitting probability to B. Since the process is absorbing in A only, the hitting probability to B will reflect the probability to go to B next rather than to A.

Using the hitting probability equations:

$$\begin{aligned} q_i^+ &= 1 \text{for } i \in B \\ q_i^+ &= \sum_{j \in X} p_{ij} q_j^+ \text{for } i \notin B. \end{aligned}$$

with the absorbing process yields:

$$\begin{aligned} q_i^+ &= 0 \text{for } i \in A \\ q_i^+ &= 1 \text{for } i \in B \\ q_i^+ &= \sum_{j \in X} p_{ij} q_j^+ \text{for } i \notin \{A, B\}. \end{aligned}$$

The backward committor probability, q_i^- pertaining to two sets A, B is the probability that being in state i, we have been in A last rather than in B. In order to get the backward committor, we use the backwards propagator $\tilde{p}_{ij} = \frac{\pi_j}{\pi_i} p_{ji}$, consider a B-absorbing process for the reverse dynamics and compute the hitting probability for A:

$$\begin{aligned} q_i^- &= 1 \text{for } i \in A \\ q_i^- &= 0 \text{for } i \in B \\ q_i^- &= \sum_{j \in X} \tilde{p}_{ij} q_j^- \text{for } i \notin \{A, B\}. \end{aligned}$$

for reversibility / detailed balance, $p_{ij} = \frac{\pi_j}{\pi_i} p_{ji} = \tilde{p}_{ij}$ and thus:

$$\begin{aligned} q_i^- &= 1 \text{for } i \in A \\ q_i^- &= 0 \text{for } i \in B \\ q_i^- &= \sum_{j \in X} p_{ij} q_j^- \text{for } i \notin \{A, B\}. \end{aligned}$$

it can be easily checked that:

$$q^- = 1 - q^+$$

satisfies this equation as it transforms it to the forward committor equation.

3.2 Transition path theory / Reactive Flux:

Probability weight of reactive trajectories:

$$m_i^R = \pi_i q_i^- q_i^+$$

with $Z_{AB} = \sum_i m_i^R = \sum_i \pi_i q_i^- q_i^+ < 1$ it is clear that we need to normalize:

$$m_i^{AB} = Z_{AB}^{-1} \pi_i q_i^- q_i^+.$$

For detailed balance, we have:

$$m_i^{AB} = Z_{AB}^{-1} \pi_i (1 - q_i^+) q_i^+.$$

to obtain the probability distribution of reaction trajectories, i.e. the probability to be at state i and to be reactive.

Probability current of reactive trajectories:

$$f_{ij}^{AB} = \begin{cases} \pi_i q_i^- p_{ij} q_j^+ & i \neq j \\ 0 & i = j \end{cases}.$$

(for detailed balance, we have $q_i^- = 1 - q_i^+$). *The probability current is the number of jumps $i \to j$ which lie on reactive $A \to B$ trajectories.*

We have a number of nice properties:

1) **Flux conservation (Kirchhoff's 1st law)**

$$\sum_{j\in X}(f_{ij}^{AB} - f_{ji}^{AB}) = 0 \quad \forall i \notin \{A, B\}$$

Proof:

$$\begin{aligned}\sum_{j\in X}(f_{ij}^{AB} - f_{ji}^{AB}) &= \pi_i q_i^- \sum_{j\neq i} p_{ij} q_j^+ - q_i^+ \sum_{j\neq i} \pi_j q_j^- p_{ji} \\ &= \pi_i q_i^- \sum_{j\neq i} p_{ij} q_j^+ - \pi_i q_i^+ \sum_{j\neq i} q_j^- \tilde{p}_{ij}\end{aligned}$$

Due to the committor equations, we have:

$$\sum_{j\in X}(f_{ij}^{AB} - f_{ji}^{AB}) = \pi_i q_i^- q_i^+ - \pi_i q_i^+ q_i^- = 0.$$

From $q_i^+ = 1 \forall i \in A$ and $q_i^- = 0 \forall i \in B$ we see that

$$\begin{aligned}f_{ij}^{AB} &= 0 \forall j \in A \\ f_{ij}^{AB} &= 0 \forall i \in B\end{aligned}$$

thus flux is not conserved at A and B, but throughout the network such that:

$$\sum_{i\in A, j\notin A} f_{ij}^{AB} = \sum_{j\notin B, i\in B} f_{ji}^{AB}.$$

Remarks:

- It is worth noting that by setting q_i^+ as negative potential, f_{ij}^+ as current and $\pi_i p_{ij}$ as conductance provides an electric network theory with Ohm's law and Kirchhoff's laws being valid.

- All TPT is valid when substituting p_{ij} in l_{ij}.

The **Effective current** is defined as

$$f_{ij}^+ = \max\{f_{ij}^{AB} - f_{ji}^{AB}, 0\}$$

and gives the *net average number of reactive trajectories per time unit making a transition from i to j on their way from A to B.*

The **total flux**, i.e. the total number of reactive $A \to B$ trajectories per time unit is simply given by the effective current flowing out of A and into B:

$$\tau_{cyc}^{-1} = K = \sum_{i \in A, j \notin A} p_{ij}^{AB} = \sum_{i \in A, j \notin A} \pi_i p_{ij} q_j^+ = \sum_{j \notin B, i \in B} f_{ji}^{AB} = \sum_{i \in B, j \notin B} \pi_j q_j^- p_{ij}$$

If the system is ergodic, every trajectory must go back from B to A in order to be able to transit to B again. Thus, K is also equal to the number of reactive $B \to A$ trajectories and equal to the number of $A \to B \to A$ cycles. Correspondingly, the inverse of K, τ_{cyc} is the average cycle time.

The $A \to B$ **total transition rate**, i.e. number of reaction events given that we start in A is given by

$$\tau_{AB}^{-1} = k_{AB} = \frac{K}{\sum_{i=X} \pi_i q_i^-}.$$

and this is the inverse $A \to B$ mean first passage time.

Transition Pathways - A reaction pathway $w = (i_0, i_1, ..., i_n)$ from A to B is a simple pathway, such that

$$i_0 \in A,\ i_n \in B,\ i_1 ... i_n \notin \{A, B\}.$$

- the capacity of a pathway w is the minimal effective current:

$$c(w) = \min_{(i,j) \in w} \{f_{ij}^+\}$$

- the bottleneck of a reaction pathway w is the edge with the minimal effective current:

$$(b_1, b_2) = \arg \min_{(i,j) \in w} \{f_{ij}^+\}$$

- The best pathway is one that maximizes the minimal current. This is only necessarily unique at the bottleneck. However, following algorithm is a rational way to find a unique best pathway in graph

$$G = \{X, f^+\}$$

:

BestPath(G,A,B)

1. Determine bottleneck (b_1, b_2) in G

2. Decompose G into L and R, which are the parts of G "left" of b_1 (nodes $\{i : q_i^+ \leq q_{b_1}^+\}$) and "right" of b_2 (nodes $\{i : q_i^+ \geq q_{b_2}^+\}$)

3. $w_L = \begin{cases} b_1 & if\ b_1 \in A \\ BestPath(L, A, \{b_1\}) & else \end{cases}$

4. $w_R = \begin{cases} b_2 & if\ b_2 \in B \\ BestPath(R, \{b_2\}, B) & else \end{cases}$

5. return (w_L,w_R)

In order to decompose the network into individual pathways, let $w = BestPath(G, A, B)$ and subtract that pathway from the network:

$$(f_{ij}^+)' = f_{ij}^+ - c(w) \quad \text{if } (i,j) \in w$$
$$(f_{ij}^+)' = f_{ij}^+ \quad \text{else.}$$

It directly follows from the flux conservation laws that in the new pathways we still have flux conservation and

$$K' = K - c(w).$$

We will have $K' = 0$ when all $A \to B$ pathways have been subtracted and the decomposition is finished. This results in a set of $A \to B$ pathways whose statistical contribution to the $A \to B$ is given by the capacity of each, $c(w)$.

Further Reading:

1. P. Metzner, C. Schütte, and E. Vanden-Eijnden: "Transition Path Theory for Markov Jump Processes". Mult. Mod. Sim. (2007)

2. F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, T. Weikl: "Constructing the Full Ensemble of Folding Pathways from Short Off-Equilibrium Simulations". PNAS (2009)

3.3 Eigenvector decomposition and interpretation

Right and Left eigenvectors can be defined by

$$Pr_i = \lambda_i r_i$$

and

$$l_i P = \lambda_i l_i$$

This can be written using the diagonal matrix of eigenvalues $\Lambda = diag\,(\lambda_1, \ldots, \lambda_n)$ by

$$PR = R\Lambda$$

and

$$LP = \Lambda L$$

We can use $L := R^{-1}$, which is a valid matrix. We diagonalize an transition matrix with the matrix of right eigenvectors R, its inverse denoted by $L := R^{-1}$ and the diagonal matrix of eigenvalues $\Lambda_{ab} = \delta_{ab}\lambda_a$ by

$$\begin{aligned} P^t &= R\Lambda L \\ &= \sum_{i=1}^{n} r_i \lambda_i^t l_i^T \end{aligned}$$

It can be easily seen, that the matrix L is also a matrix of left eigenvectors, which helps with the interpretation. Note, that in principle any matrix of left eigenvectors (in the rows) can be used, but only the one which fulfills $L := R^{-1}$ is suitable for diagonalizing. This definition still allows for some scaling freedom and requires only

$$\begin{aligned} \langle l_i, r_i \rangle &= 1 \;\; \forall i \\ \langle l_i, r_j \rangle &= 0 \;\; \forall i \neq j \end{aligned}$$

We can see by this separation, that the transition matrix is working for each eigenvector-eigenvalue pair working on different timescales, because the time dependence is structurally only found in the potency of the eigenvalue. If we apply this matrix to vector p, which we want to propagate, we get

$$\begin{aligned} p(t)^T &= p(0)^T P^t \\ &= \sum_{i=1}^{n} \lambda_i^t p(0)^T r_i l_i^T \\ &= \sum_{i=1}^{n} \lambda_i^t \langle p(0), r_i \rangle l_i^T \\ &= \sum_{i=1}^{n} \gamma_i \lambda_i^t l_i^T \end{aligned}$$

Which means, that the scalar product with the right eigenvector is the intensity γ_i, how much the ith relaxation process is involved when relaxing from the initial distribution $p(t)$. The left eigenvector denotes the change of probability density that is brought about by the ith relaxation process. If the system is reversible we show that when r_i is a right eigenvector of P, then the vector l_i with

$$l_i = \Pi r_i$$

is a left eigenvector of P. Here, $\Pi = \text{diag}\{\pi_i\}$ with $\pi^T = \pi^T P$ being the stationary distribution.

Proof: reversibility implies detailed balance:

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j$$

In Matrix form:

$$\begin{aligned} \Pi P &= P^T \Pi \\ \Pi R \Lambda R^{-1} &= \left(R \Lambda R^{-1}\right)^T \Pi \\ \Pi R \Lambda R^{-1} &= \left(R^{-1}\right)^T \Lambda R^T \Pi \end{aligned}$$

we note that R^{-1} is a left eigenvector matrix: $L := R^{-1}$

$$\Pi R \Lambda L = L^T \Lambda R^T \Pi$$

which means that ΠR and L^T span the same eigenspace and that when r_i is a right eigenvector,

$$\Pi r_i = l_i$$

is a left eigenvector.

Based on this we can rewrite the decomposition of $P = R\Lambda L$: If P is reversible, there exists a set of left and right eigenvectors, such that

$$\begin{aligned} P &= R \Lambda R^T \Pi \\ &= \Pi^{-1} L^T \Lambda L \end{aligned}$$

provided that the eigenvectors are properly normalized, i.e. $R^T \Pi R = Id$ and $L\Pi^{-1}L^T = Id$. This can be enforced by taking an arbitrary set of left eigenvectors, $\hat{l}_i$, or right eigenvectors, $\hat{r}_i$ and scaling them:

$$\begin{aligned} l_i &= \hat{l}_i / \left(\hat{l}_i^T \Pi^{-1} \hat{l}_i\right)^{\frac{1}{2}} \\ r_i &= \hat{r}_i / \left(\hat{r}_i^T \Pi \hat{r}_i\right)^{\frac{1}{2}} \end{aligned}$$

This can be used for the development of a probability distribution

$$\begin{aligned} p(t)^T &= p(0)^T P^t \\ &= \sum_{i=1}^{n} \lambda_i^t p(0)^T r_i (\Pi r_i)^T \\ &= \sum_{i=1}^{n} \lambda_i^t p(0)^T r_i r_i^T \Pi \\ &= \sum_{i=1}^{n} \lambda_i^t p(0)^T W_i \Pi \end{aligned}$$

with $W_i = r_i r_i^T$ being a projection matrix onto the basis of the right eigenvectors. A probability distribution is projection onto this basis, weighted with the stationary distribution and scaled according to the eigenvalue.

Fig. 3.1 shows an example of the relation between eigenvectors and the underlying dynamics.

3.4 Timescales and timescale test

The eigenvalues λ_i are related to a timescale in the manner, that the process vanishes with increasing t slower or faster. Compare the discrete-time decay λ_i^t and compare it to the exponential decay in continuous time with timescale t_i, $\exp(-t/t_i)$:

$$\begin{aligned} \lambda_i^t &= \exp(-t/t_i) \\ t_i &= \frac{-1}{\log \lambda_i} \end{aligned}$$

and if we associate the transition matrix with a real time step τ then then we have the implied timescales

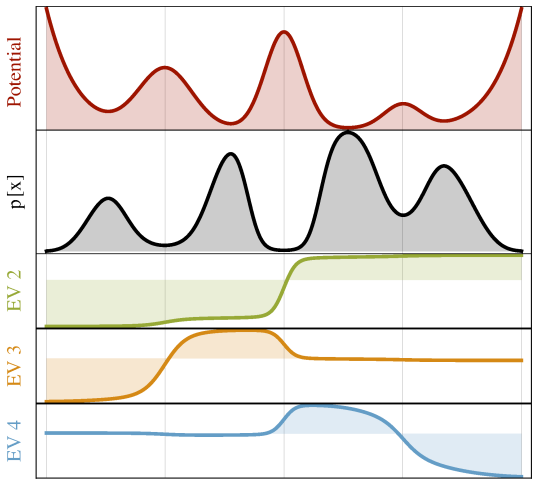


Figure 3.1: Example potential with computed eigenvectors indicatind processes. The underlying dynamics was computed from a 1000 state transition matrix with non-zero transition probabilities between neighboring states given by a metropolis criterion.

$$t_i[P(\tau)] = \frac{-\tau}{\log \lambda_i(\tau)}$$

which can, using the semi-group property of the Markov system

$$P(m\tau) = P^{m\tau}$$

shown to be independent of the multiple m

$$t_i[P(m\tau)] = t_i[P(\tau)^m] = \frac{-m\tau}{\log \lambda[P^m]} = \frac{-m\tau}{\log \lambda[P]^m} = \frac{-m\tau}{m \log \lambda[P]} = t_i[P]$$

3.5 Correlation functions

We would now like to investigate correlation functions of observables $a(t)$ which are observables of Markov chains. Let $x(t)$ be a Markov chain in state space $X = \{1, ..., n\}$, discrete time $t \in \mathbb{N}_0$, with transition matrix $P \in \mathbb{R}^{n \times n}$ and let $f_a : X \to \mathbb{R}$ be a function that maps each state to an observable value. We consider now that the observed time series $a(t)$ has been produced by such an observation of a Markov chain:

$$a(t) = f_a(x(t))$$

As a shorthand notation we also define a vector $\mathbf{a} = (f_a(1), ..., f_a(n))^T$ containing the observable of each state of X in the corresponding entry, i.e. we can write in short:

$$a(t) = a_{x(t)}$$

The autocorrelation function of $a(t)$ can then be written as:

$$\mathbb{E}(a(t)\, a(t+\tau)) = \mathbb{E}(a_{x(t)} a_{x(t+\tau)}) = \sum_{i,j \in X} \mathbb{P}(x(t) = i, x(t+\tau) = j)\, a_i\, a_j$$

where $\tau \in \mathbb{N}_0$ is a discrete time lag. In the stationary and ergodic case this is equivalent to:

$$\mathbb{E}(a(t)\, a(t+\tau)) = \lim_{T \to \infty} \frac{1}{T-\tau} \sum_{t=0}^{T-\tau} a(t)\, a(t+\tau)$$

We notice that we can rewrite:

$$\begin{aligned} \mathbb{E}(a(t)\, a(t+\tau)) &= \sum_{i,j \in X} \mathbb{P}(x(t) = i, x(t+\tau) = j)\, a_i\, a_j \\ &= \sum_{i,j \in X} \pi_i\, p_{ij}(\tau)\, a_i\, a_j \\ &= a^T \Pi P^\tau a \end{aligned}$$

Using spectral decomposition we can rewrite this as:

$$\begin{aligned} \mathbb{E}(a(t)\, a(t+\tau)) &= \sum_{i=1}^{n} \lambda_i^\tau a^T \Pi r_i r_i^T \Pi a \\ &= \sum_{i=1}^{n} \lambda_i^\tau a^T l_i l_i^T a \\ &= \sum_{i=1}^{n} \lambda_i^\tau \langle a, l_i \rangle^2 \\ &= \sum_{i=1}^{n} \lambda_i^\tau \gamma_i \end{aligned}$$

with the choice $\gamma_i = \langle a, l_i \rangle^2$. Thus, we can explain correlation functions in terms of multi-exponential decays with timescales/rates given by the eigenvalues and intensities depending on the overlap of the observable a with the eigenvectors. On the other hand, we can estimate $\mathbb{E}(a(t)\, a(t+\tau))$ from given trajectory and explore its spectral properties by observing this multiexponential decay over τ.

Further Reading:

- Timescales [4]

Chapter 4

Continuous Random variables

We now consider continuous random variables. The basic object is as usual a probability space $(\Omega, \Sigma, \mathbb{P})$, which describes the randomness in our experiment. Ω is the set of basic samples that can occur, the algebra Σ is the set of all possible events that we are interested in characterizing. $\mathbb{P}$ is a probability measure that assigns a probability to each event $S \in \Sigma$. Since we are aiming at continuous random variables, a useful sample space is $\Omega = \mathbb{R}$, and a useful algebra is the Borel algebra:

Borel algebra is the smallest σ-algebra on the real numbers $\mathbb{R}$. It contains all intervals on the real axis, i.e.:

$$\{i_{x_1,x_2} = [x_1, x_2] \mid x_1 \leq x_2; x_1, x_2 \in \mathbb{R}\} \cup \mathbb{R} \cup \emptyset$$

The probability measure $\mathbb{P}$ is a measure with the normalization condition $\mathbb{P}(\Omega) = 1$. A common measure for continuous spaces is the Lebesque measure μ, which can be turned into a probability measure by normalizing with $\mu(\Omega)$:

Lebesque measure is the ordinary notion of length, area, volume of subsets of Euclidean spaces.

Example: It is useful to think of the probability space as a model for a computer that generates high-quality random variables (high quality here means almost uncorrelated). Computational random number generators usually have a sample space $\Omega = [0, 1] \subset \mathbb{R}$, the corresponding algebra is a Borel algebra on the subset $[0, 1]$ and the probability is given by the normalized Lebesque measure: $\mathbb{P}([x_1, x_2]) = \int_{x=x_1}^{x_2} dx / \int_{x=0}^{1} dx$.

4.1 Continuous random variables

Random variable (continuous) A measurable function $X : \Omega \to \mathbb{R}$ between a probability space $(\Omega, \Sigma, \mathbb{P})$ and a measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel-Algebra of $\mathbb{R}$, is a continuous random variable.

Measurable function Definition (measurable function): When (Ω_1, Σ_1) and (Ω_2, Σ_2) are measurable spaces, a function $f : \Omega_1 \to \Omega_2$ is (Σ_1, Σ_2)-measureable if

$$f^{-1}(E_2) := \{\omega \in \Omega_1 : f(\omega) \in E_2\} \in \Sigma_1 \, \forall E_2 \in \Sigma_2$$

Lebesgue Integration A measure space (Ω, Σ, μ) is associated with the theory of Lebesgue integration. Let $g : \Omega \to \mathbb{R}$ be a measurable function, then the integral is defined as:

$$G := \int_\Omega g \, d\mu = \int_\Omega g(\omega) \, d\mu(\omega)$$

In the "nice" case that the measure is absolutely continuous we can rewrite this in terms of the ordinary Riemann integration

$$G = \int_\Omega g(\omega) \, \mu(\omega) \, d\omega$$

Probability density function In probability theory, a probability density function (abbreviated as pdf, or just density) of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space. The probability of a random variable falling within a given set is given by the integral of its density over the set.

A probability density function is most commonly associated with continuous univariate distributions. A random variable X has density f, where f is a non-negative Lebesgue-integrable function, if:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b \mathrm{d}\mu(x) = \int_a^b f(x) \, \mathrm{d}x.$$

Cumulative distribution function Hence, if F is the cumulative distribution function of X, then:

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u,$$

and

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x).$$

Intuitively, one can think of $f(x)dx$ as being the probability of X falling within the infinitesimal interval $[x, x + dx]$.

Example: As in the example above, consider the probability space of a computer that generates $[0, 1]$ random variables with normalized Lebesque measure, $([0, 1], \mathcal{B}([0, 1]), \mu/\mu([0, 1]))$. We define a random variable $x \in \mathbb{R}$ which we want to be distributed according to $f(x)$. This can be realized by defining $x = F^{-1}(\omega)$.

4.2 Properties of Random Variables $X \in \mathbb{R}$

Moments: expectation, variance, covariance If f is a probability density function, then the value of the following integral above is called the nth moment of the probability distribution.

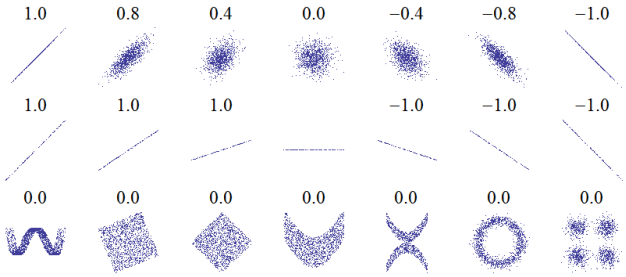
$$\mu_n = \mathbb{E}(x^n) = \int_{-\infty}^{\infty} x^n\, f(x)\, dx$$

For any two random variables X, Y it holds that

$$\mu_1(X + Y) = \mathbb{E}(X + Y) = \int\int x\, f(x)\, y\, g(y)\, dx\, dy = \mathbb{E}(X) + \mathbb{E}(Y)$$

The properties discussed for discrete random variables can be transferred to the continuous case likewise.

Correlations Consider the following distributions of two random variables and their correlation coefficients:



4.3 Transformation between random variables

If the probability density function of an independent random variable x is given as $f_X(x)$, it is possible to calculate the probability density function of some variable $y = g(x)$. This is also called a "change of variable" and is in practice used to generate a random variable of arbitrary shape $f_{g(X)} = f_Y$ using a known (for instance uniform) random number generator.

In order to perform this variable change we will need to rescale the new variable. In order to get an expression for this we make use of the fact that the probability contained in a differential area must be invariant under change of variables. That is,

$$|f_Y(y)\,dy| = |f_X(x)\,dx|\,,$$

If the function g is monotonic, i.e. invertible, we can compute $x = g^{-1}(y)$, and the resulting density function is then given by:

$$f_Y(y) = \left|\frac{dx}{dy}\right| f_X(x) = \left|\frac{1}{g'(g^{-1}(y))}\right| f_X(g^{-1}(y)).$$

Multidimensional case:

Let $\mathbf{x}$ be a n-dimensional random variable with joint density $f_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{y} = h(\mathbf{x})$ a bijective and diffentiable function, then $\mathbf{y}$ has density

$$f_{\mathbf{Y}}(\mathbf{y}) = \left|\det\left(\frac{d\mathbf{x}}{d\mathbf{y}}\right)\right| f_{\mathbf{X}}(\mathbf{x})$$

where

$$\left(\frac{d\mathbf{x}}{d\mathbf{y}}\right) = J_{\mathbf{x}} = \begin{pmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ & & \\ \frac{dx_n}{dy_1} & & \frac{dx_n}{dy_n} \end{pmatrix}.$$

Example 1:

$$\begin{aligned} f_X(x) &= a\exp(-\frac{x}{2\sigma^2}) \\ y &= \sqrt{x} \\ x &= y^2 \\ \frac{dx}{dy} &= 2y \\ f_Y(y) &= 2y\ \exp(-\frac{y^2}{2\sigma^2}) \end{aligned}$$

Example 2:

Let us consider a gas with identical particles of mass m at sufficiently high thermal motion. The kinetic energy of a particle is given by:

$$E = \frac{1}{2}mv^2$$

where $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$ is the speed of the particle. Each state of the system, characterized by v_x, v_y, v_z has a probability weight given by the Boltzmann distribution:

$$p(v_x, v_y, v_z) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left(-\frac{m}{2kT}(v_x^2 + v_y^2 + v_z^2)\right)$$

We use the transformation $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$ and obtain

$$\begin{aligned} p(v) &= \int_{v^2 = v_x^2 + v_y^2 + v_z^2} p(v_x, v_y, v_z) dv_x dv_y dv_z \\ &= 4\pi v^2 p(v_x, v_y, v_z) \\ &= v^2 \sqrt{16\pi^2 \left(\frac{m}{2\pi kT}\right)^3} \exp\left(-\frac{mv^2}{2kT}\right) \\ &= v^2 \sqrt{\frac{2}{\pi}\left(\frac{m}{kT}\right)^3} \exp\left(-\frac{mv^2}{2kT}\right). \end{aligned}$$

This is the probability distribution of speeds v, which is known as the Maxwell-Boltzmann distribution.

We next consider the distribution of kinetic energies which can for each speed v be calculated as $E = \frac{1}{2}mv^2$. Here we perform a variable transformation and obtain via the function $v = \sqrt{2E/m}$:

$$\begin{aligned} p(E) &= \frac{dv}{dE} p(v(E)) \\ &= \frac{d}{dE}\sqrt{\frac{2E}{m}} p\left(\frac{2E}{m}\right) \\ &= \sqrt{\frac{1}{2Em}}\frac{2E}{m}\sqrt{\frac{2}{\pi}\left(\frac{m}{kT}\right)^3}\exp\left(-\frac{E}{kT}\right) \\ &= 2\sqrt{\frac{E}{\pi(kT)^3}}\exp\left(-\frac{E}{kT}\right) \end{aligned}$$

4.4 Linear Error Propagation

We consider n random variables $x_1, ..., x_n$ that we aggregate in the vector $\mathbf{x} = (x_1, ..., x_m)^T$. x_j are distributed according to probability distribution $p(x_1, ..., x_n) = p(\mathbf{x})$. We assume that we have the probability distribution p available and can compute its maximum over $\mathbf{x}$ and the variances and covariances of the x-variables. However, we are interested in another set of random variables $\mathbf{y} = (y_1, ..., y_m)^T$ that are given by a generally nonlinear function of $x_1, ..., x_n$:

$$\begin{aligned} y_1 &= f_1(x_1, ..., x_n) \\ &\vdots \\ y_m &= f_n(x_1, ..., x_n) \end{aligned}$$

Let us assume that the direct evaluation of the distribution $p(\mathbf{y})$ is difficult, such that we cannot easily calculate maxima and variances of the y-variables directly. In particular we are interested in "error propagation", i.e. how the variances in $\mathbf{x}$ propagate through the f onto $\mathbf{y}$. We are now seeking an approximate way to do this.

One approach is to estimate the uncertainties on y_i by Monte Carlo sampling. Here, we investigate linear error propagation, which provides an analytical result, but at the sacrifice of making two approximations: (1) The distribution function of x_i is approximated by a multivariate Gaussian, and (2) the functions f_j are linearized. Clearly, these approximations are only meaningful in certain situations, especially if the distribution is monomodal and near-Gaussian around its maximum, and if the f_j are close-to linear within the high-probability range of x_i.

Gaussian approximation of the density We first seek the $\mathbf{x}$ that maximizes the probability density $p(\mathbf{x})$:

$$\hat{\mathbf{x}} = \arg\max p(\mathbf{x})$$

By definition, the gradient at $\hat{\mathbf{x}}$ is zero. Next, we calculate second derivative matrix H, the Hessian matrix:

$$\Sigma^{-1} = H = \begin{bmatrix} \frac{\partial p(\mathbf{x})}{\partial x_1 x_1} & \cdots & \frac{\partial p(\mathbf{x})}{\partial x_n x_1} \\ \vdots & & \vdots \\ \frac{\partial p(\mathbf{x})}{\partial x_1 x_n} & \cdots & \frac{\partial p(\mathbf{x})}{\partial x_n x_n} \end{bmatrix}.$$

It can be shown that if $p(\mathbf{x})$ is a multivariate Gaussian distribution with covariance matrix Σ containing the covariances:

$$\sigma_{ij} = \mathbb{E}(x_i x_j) - \mathbb{E}(x_i)\mathbb{E}(x_j),$$

it is $\Sigma^{-1} = H(\hat{\mathbf{x}})$, i.e. the inverse of the Hessian at the maximum yields the covariance matrix. Based on this fact, we approximate $p(\mathbf{x})$ around $\hat{\mathbf{x}}$ via:

$$p(\mathbf{x}) \approx \mathcal{N}(\hat{\mathbf{x}}, \Sigma_{\hat{\mathbf{x}}}).$$

Note that if $p(\mathbf{x})$ is not actually a Gaussian, the σ_{ij} are generally smaller than the true covariances $\mathbb{E}(x_i x_j) - \mathbb{E}(x_i)\mathbb{E}(x_j)$. The Gaussian approximation is only valid within the vicinity of $\hat{\mathbf{x}}$, and therefore especially useful if $p(\hat{\mathbf{x}})$ is very peaked around $\hat{\mathbf{x}}$. For distributions coming from estimation procedures this is usually the case when a relatively large amount of data has been collected.

Linear approximation of the transfer function Next we will approximate our functions f_j which can generally be very nonlinear. We take the first-order (linear) approximation of the Taylor series around $\hat{\mathbf{x}}$:

$$\begin{aligned} f_j(\mathbf{x}) &= f_j(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})^T \nabla_{f_j}(\hat{\mathbf{x}}) + \mathcal{O}(\|\mathbf{x} - \hat{\mathbf{x}}\|^2) \\ &\approx f_j(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}})^T \nabla_{f_j}(\hat{\mathbf{x}}) \\ &\approx f_j(\hat{\mathbf{x}}) - \hat{\mathbf{x}}^T \nabla_{f_j}(\hat{\mathbf{x}}) + \mathbf{x}^T \nabla_{f_j}(\hat{\mathbf{x}}) \\ &\approx a_j + \mathbf{x}^T \mathbf{b}_j \end{aligned}$$

where we have defined $a_j = f_j(\hat{\mathbf{x}}) - \hat{\mathbf{x}}^T \nabla_{f_j}(\hat{\mathbf{x}})$ and $\mathbf{b}_j := \nabla_{f_j}(\hat{\mathbf{x}})$. If we aggregate as follows: $\mathbf{y} = (y_1, ..., y_m)^T$, $\mathbf{a} = (a_1, ..., a_m)^T$ and $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_m]^T$, we can write the entire system as:

$$\mathbf{y} \approx \mathbf{a} + \mathbf{B}\mathbf{x}$$

This is a second approximation. We assume that the functions f_j, although generally nonlinear, are almost linear within the range of $\mathbf{x}$ values around $\hat{\mathbf{x}}$ that are accessible with high probability.

Propagation Generally, if $\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}$ is an affine transformation of the Gaussian-distributed variables $\mathbf{x} \sim \mathcal{N}(\hat{\mathbf{x}}, \Sigma)$, then it can be shown that:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\hat{\mathbf{x}}, \mathbf{B}\Sigma\mathbf{B}^T).$$

Thus, the covariance matrix of the y-Variables is given by $\mathbf{B}\Sigma\mathbf{B}^T$. Of special interest are the variances of the y variables which are found on the diagonal:

$$(\mathbf{B}\Sigma\mathbf{B}^T)_{ii} = \mathbf{b}_i\Sigma\mathbf{b}_i^T$$

We can thus estimate the maximum probability $\hat{\mathbf{y}}$ and the covariances $\Sigma_{\mathbf{y}}$ of the target variables through Gaussian error propagation without further approximations.

Example We consider again the Boltzmann-distributed particles. Let us assume that we have a Device that allows us to measure the kinetic energy E and we are interested in calculating the velocity the particle had. We can make use of the equation $v = \sqrt{\frac{2E}{m}}$. However, given that the measuremed value $\tilde{E}$ has a measurement error of magnitude s_E with respect to the true value E, such that the distribution of $\tilde{E}$ is given by:

$$p(\tilde{E}|E) = \mathcal{N}(E, s_E^2).$$

What can we now say about the error in v, s_v?

Let us assume we measure a value $\tilde{E}$. We first used Bayes inversion in order to write down the probability distribution of E in terms of $\tilde{E}$:

$$
\begin{aligned}
p(E|\tilde{E}) &\propto p(E)p(\tilde{E}|E) \\
&\propto \frac{1}{2\pi s_E^2}\exp\left(-\frac{(E-\tilde{E})^2}{2s_E^2}\right)2\sqrt{\frac{E}{\pi(kT)^3}}\exp\left(-\frac{E}{kT}\right) \\
&\propto \sqrt{E}\exp\left(-\frac{E}{kT}\right)\exp\left(-\frac{(E-\tilde{E})^2}{2s_E^2}\right) \\
&\propto \sqrt{E}\exp\left(-\frac{(E-\tilde{E})^2kT+2s_E^2E}{2s_E^2kT}\right) \\
&\propto \sqrt{E}\exp\left(-\frac{E^2kT-2(\tilde{E}kT-s_E^2)E+\tilde{E}^2kT}{2s_E^2kT}\right) \\
&\propto \sqrt{E}\exp\left(-\frac{(\sqrt{kT}E-\frac{(\tilde{E}kT-s_E^2)}{\sqrt{kT}})^2-\frac{(\tilde{E}kT-s_E^2)^2}{kT}+\tilde{E}^2kT}{2s_E^2kT}\right) \\
&\propto \sqrt{E}\exp\left(-\frac{(E-\frac{(\tilde{E}kT-s_E^2)}{kT})^2}{2s_E^2}\right)
\end{aligned}
$$

We abbreviate $m_E = \frac{(\tilde{E}kT-s_E^2)}{kT}$ and obtain:

$$
\frac{dp(E|\tilde{E})}{dE} = \left(\frac{1}{2\sqrt{E}}-\frac{\sqrt{E}}{s_E^2}(E-m_E)\right)\exp\left(-\frac{(E-m_E)^2}{2s_E^2}\right)
$$

which becomes zero for:

$$
\hat{E} = \arg\max p(E|\tilde{E}) = \frac{m_E \pm \sqrt{m_E^2+2s_E^2}}{2}
$$

of which we choose the positive solution:

$$
\hat{E} = \frac{m_E + \sqrt{m_E^2+2s_E^2}}{2}.
$$

Note that for vanishing error $\hat{E} = m_E = \tilde{E}$.

The variance of E in a Gaussian approximation around $\hat{E}$ is given by the inverse second derivative at $\hat{E}$. Using the abbreviation $\alpha = \sqrt{m_E^2+2s_E^2}$, we obtain:

$$\sigma_E^2 = \left(\frac{d^2 p(E|\tilde{E})}{dE^2}\bigg|_{\hat{E}} \right)^{-1} = \frac{s_E^2 \sqrt{2m_E + 2\alpha}}{2\exp(-\frac{(\alpha - m_E)^2}{8s_E^2})\alpha}$$

Thus, we have approximated:

$$p(E|\tilde{E}) \approx \mathcal{N}(\hat{E}, \sigma_E^2).$$

Now we use the equation $v = \sqrt{\frac{2E}{m}}$ which is linearized by:

$$\begin{aligned} v(E) &\approx \sqrt{\frac{2\hat{E}}{m}} + \sqrt{\frac{1}{2m\hat{E}}}(E - \hat{E}) \\ &\approx \sqrt{\frac{2\hat{E}}{m}} - \sqrt{\frac{\hat{E}}{2m}} + \sqrt{\frac{1}{2m\hat{E}}}E \\ &\approx \sqrt{\hat{E}}\left(\sqrt{\frac{2}{m}} - \sqrt{\frac{1}{2m}}\right) + \sqrt{\frac{1}{2m\hat{E}}}E \end{aligned}$$

And hence

$$\sigma_v^2 = \left(\frac{dv}{dE}\right)^2 \sigma_E^2 = \frac{s_E^2 \sqrt{2m_E + 2\alpha}}{2m(m_E + \alpha)\exp(-\frac{(\alpha - m_E)^2}{8s_E^2})\alpha}$$

if we take $m = 1$ unitless then

$$\sigma_v^2 = \frac{s_E^2}{\sqrt{2m_E + 2\alpha}\exp(-\frac{(\alpha - m_E)^2}{8s_E^2})\alpha}$$

For small errors s_E, this is approximately:

$$\sigma_v^2 \approx \frac{s_E^2}{2\tilde{E}^{3/2}}$$

4.5 Characteristic Functions

An alternative way to represent a probability density $f_X(x)$ is by its Fourier transform. This is called characteristic Function $G(\omega)$, defined for all $\omega \in \mathbb{R}$:

$$G(\omega) = G(\omega)\{x\} = \mathbb{E}[e^{i\omega x}] = \int e^{i\omega x} f_X(x)\, dx$$

This is also the moment generating function of the distribution because its Taylor expansion turns out to be

$$G(\omega) = \sum_{m=0}^{\infty} \frac{(i\omega)^m}{m!}\mu_m = 1 + i\omega\mu - \frac{1}{2}\omega^2\mu_2 + ...$$

with the moments μ_m.

The characteristic function completely determines the behavior and properties of the probability distribution of the random variable X. The two approaches are equivalent in the sense that knowledge of one of the functions can always be used in order to find the other one, yet they both provide different insight for understanding the features of our random variable. However, in particular cases, there can be differences in whether these functions can be represented as expressions involving simple standard functions.

If a random variable admits a density function, then the characteristic function is its dual, in the sense that each of them is a Fourier transform of the other.

Proof of the central limit theorem For a theorem of such fundamental importance to statistics and applied probability, the central limit theorem has a remarkably simple proof using characteristic functions. It is similar to the proof of a (weak) law of large numbers. For any random variable, X_i, with zero mean, the characteristic function of X_i is, by Taylor's theorem,

$$G(\omega) = 1 - \frac{\sigma^2\omega^2}{2} + o(\omega^2)$$

where $o(\omega^2)$ is "little o notation" for some function of ω that goes to zero more rapidly than ω^2. Thus, the characteristic function of Z_n is

$$\begin{aligned} G(\omega)\{Z_n\} &= G(\omega)\left\{\sum_{i=1}^{n} X_i\right\} \\ &= \prod_{i=1}^{n} G\left(\frac{\omega}{\sqrt{n}}\right)\{X_i\} \\ &= \left[1 - \frac{\sigma^2\omega^2}{2n} + o\left(\frac{\omega^2}{n}\right)\right]^n \to e^{-\omega^2\sigma^2/2}, \quad n \to \infty. \end{aligned}$$

Chapter 5

Markov chain estimation

5.1 Bayesian Approach

Bayes Theorem: Consider two events A and B. Based on the definition of the conditional probability we can follow:

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \cdot \mathbb{P}(A)}{P(B)} \\ &= \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}.\end{aligned}$$

Let us consider a model M and observation O. Bayes' rule states that:

$$\mathbb{P}(M \mid O) = \mathbb{P}(M)\frac{\mathbb{P}(O \mid M)}{\mathbb{P}(O)}$$

in particular, when we consider only one given observation O, we can state:

$$\mathbb{P}(M \mid O) \propto \mathbb{P}(M)\ \mathbb{P}(O \mid M).$$

where we call

$\mathbb{P}(M \mid O)$	posterior probability
$\underline{\mathbb{P}(M)}$	prior probability of the model M
$\underline{\mathbb{P}(O)}$	prior probability of the data O
$\underline{\mathbb{P}(O \mid M)}$	likelihood

since usually we only work with a given dataset $\mathbb{P}(O)$ is constant, and it is sufficient to know that:

$$\mathbb{P}(M \mid O) \propto \mathbb{P}(M)\,\mathbb{P}(O \mid M)$$

The likelihood $\mathbb{P}(O \mid M)$ is usually easy to compute if we have specified the type of model and we have the data at hand. The prior $\mathbb{P}(M)$ is a probability that has the function to bias the estimator towards models that are reasonable. If no prior is used that is equivalent to using a uniform prior, which may not always be a good choice. Choosing $\mathbb{P}(M)$ is a modeling problem, so there is "right" or "wrong" here and it must be chosen with expertise for the process observed. The significance of the prior is to be able to have a well-defined probability distribution even in the case that no or almost no data is at hand. When more and more data is collected, the likelihood $\mathbb{P}(O \mid M)$ will get sharper and thus eventually dominate the prior. Thus, it is generally a good idea to use a rather weak prior that rules out unphysical results but can be easily overcome when some data is collected.

When seeking optimial models we will maximize the posterior:

$$\mathbb{P}(M \mid O) \to \max$$

or alternatively we can draw samples x_k from it, e.g. with a MCMC approach:

$$x_k \sim \mathbb{P}(M \mid O)$$

5.2 Transition Matrix Estimation from Observations; Likelihood

We consider an observed sequence $x(t) \in X$, $t = \{0, ..., T\}$ in a state space $X = \{1, ..., n\}$.

We assuming that $x(t)$ has been generated from a Markov chain with transition matrix P, which we would like to infer from the data. This is a very common problem in many applications, such as finance, game theory, molecular physics and biology.

Let the frequency matrix $Z = (z_{ij}) \in \mathbb{R}^{n \times n}$ count the number of observed transitions between states, *i.e.* z_{ij} is the number of observed transitions from state i at time t to state j at time $t + 1$, summed over all times t:

$$z_{ij} = |\{x(t) = i,\, x(t+1) = j \mid t = 0...T-1\}|.$$

It is easy to treat the case of observing multiple sequences $x^{(1)}(t), ..., x^{(N)}(t)$ by noticing that their count matrices add up: $Z = Z^{(1)} + ... + Z^{(N)}$. As a shorthand notation we define:

$$z_i := \sum_{k=1}^{n} z_{ik}.$$

is the total number of observed transitions leaving state i. It is intuitively clear that in the limit of an infinitely long trajectory, the elements of the true transition matrix are given by the trivial estimator:

$$\hat{p}_{ij} = \frac{z_{ij}}{\sum_k z_{ik}} = \frac{z_{ij}}{z_i}, \tag{5.1}$$

For a trajectory of limited length, the underlying transition matrix P cannot be unambiguously computed. The probability that a particular P would generate the observed trajectory $x(t)$ is given by:

$$\mathbb{P}(x(0), ..., x(T) \mid P) = \prod_{t=0}^{T-1} p_{x(t),x(t+1)} = \mathbb{P}(Z|P) = \prod_{i,j=1}^{n} p_{ij}^{z_{ij}}$$

Vice versa, the probability that the observed data was generated by a particular transition matrix P is

$$\mathbb{P}(P|Z) \propto \mathbb{P}(P)\mathbb{P}(Z|P) = \mathbb{P}(P) \prod_{i,j=1}^{n} p_{ij}^{z_{ij}}, \tag{5.2}$$

where $\mathbb{P}(P)$ is the prior probability of transition matrices before observing any data. $\mathbb{P}(Z|P)$ is called likelihood. Transition matrix estimation is often approached by identifying the maximum of $\mathbb{P}(Z|P)$, i.e., the maximum likelihood estimator. For the case of a uniform prior, this is identical to the transition matrix with maximum posterior probability. Otherwise, we now restrict ourselves to prior distributions which are conjugate to the likelihood ("conjugate prior"), i.e. have the same functional form. This leads to:

$$\mathbb{P}(P|Z) \propto \prod_{i,j} p_{ij}^{b_{ij}+z_{ij}} = \prod_{i,j} p_{ij}^{c_{ij}}, \tag{5.3}$$

with the prior count matrix $B = [b_{ij}]$ and we have defined the total (posterior) number of counts $C = B + Z$. In the following we will always work with C. The likelihood estimation $C = Z$ is a special case.

5.3 Maximum Probability Estimator

We will now derive the Maximum Probability Estimator by finding the transition matrix that maximizes $\mathbb{P}(P|Z)$. The probability is difficult to work with due to the product. For optimization purposes it is therefore a common "trick" to instead work with the logarithm of the likelihood (log-likelihood):

$$Q = \log \mathbb{P}(P|C) = \sum_{i,j} c_{ij} \log p_{ij}.$$

This is useful since the logarithm is a monotonic function: as a result, the maximum of $\log f$ is also the maximum of f. However, this function is not bounded from above, since for $p_{ij} \to \infty$, $Q \to \infty$. Of course, we somehow need to restrict ourselves to sets of variables which actually form transition matrices, i.e., they satisfy the constraint:

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

When optimizing with equality constraints, one uses Langrangian multipliers. The Lagrangian for Q is given by:

$$F = Q + \lambda_1(\sum_j p_{1j} - 1) + ... + +\lambda_m(\sum_j p_{mj} - 1).$$

This function is maximized by the maximum likelihood transition matrix. It turns out that F only has a single stationary point, which can be easily found by setting the partial derivatives to zero. Those are given by

$$\frac{\partial \log F}{\partial p_{ij}} = \frac{c_{ij}}{p_{ij}} + \lambda_i.$$

Set to 0:

$$\frac{c_{ij}}{\hat{p}_{ij}} + \lambda_i = 0$$
$$\lambda_i \hat{p}_{ij} = -c_{ij}.$$

We now make use of the transition matrix property:

$$\lambda_i \sum_{j=1}^{m} \hat{p}_{ij} = \lambda_i = -\sum_{j=1}^{m} c_{ij} = -c_i$$

and thus:

$$\frac{c_{ij}}{\hat{p}_{ij}} - c_i = 0$$
$$\hat{p}_{ij} = \frac{c_{ij}}{c_i}.$$

It turns out that $\hat{P}(\tau)$, as provided by Eq. (5.1), is the maximum of $\mathbb{P}(C|P)$ and thus also of $\mathbb{P}(P|C)$ when transition matrices are assumed to be uniformly distributed *a priori*. In the limit of infinite sampling, $\mathbb{P}(P|C)$ converges towards a delta distribution with its peak at $\hat{P}(\tau)$.

5.4 Maximum Likelihood Estimator of Reversible Matrices

It will turn out that in particular when estimating conformation dynamics from short trajectories, it is essential to additionally require that the transition matrix is reversible. The reason is that the system itself is in equilibrium, and a statistical model for the equilibrium case is desired, but the individual trajectories are far off equilibrium, as they are too short to be considered "relaxed" on the timescale of the molecule. George Boxer (Princeton) has suggested the following way to compute the maximum likelihood reversible transition matrix:

Let $x_{ij} = \pi_i p_{ij}$ be the unconditional transition probabilities with the additional constraint that $\sum_{i,j} x_{ij} = 1$, the detailed balance condition is given by

$$x_{ij} = x_{ji}.$$

The transition probabilities are given in terms of x_{ij} as:

$$p_{ij} = \frac{x_{ij}}{\sum_{k=1}^{n} x_{ik}},$$

such that the log-likelihood is given by:

and then our goal is to find $X = (x_{ji})$ in order to maximize

$$Q = \sum_{i,j=1}^{m} c_{ij} \left(\log x_{ij} - \log \left(\sum_{k=1}^{m} x_{ik} \right) \right).$$

The partial derivatives are given by:

$$\frac{\partial Q}{\partial x_{ij}} = \frac{c_{ij}}{x_{ij}} + \frac{c_{ji}}{x_{ji}} - \sum_{j'=1}^{m} \frac{c_{ij'}}{\sum_{k=1}^{m} x_{ik}} - \sum_{j'=1}^{m} \frac{c_{jj'}}{\sum_{k=1}^{m} x_{jk}}$$

and writing $c_i = \sum_{k=1}^{n} c_{ki}$ and $x_i = \sum_{k=1}^{n} x_{ki}$ we have

$$\frac{\partial Q}{\partial x_{ji}} = \frac{c_{ji} + c_{ij}}{x_{ji}} - \frac{c_i}{x_i} - \frac{c_j}{x_j}$$

When Q is maximized $\frac{\partial Q}{\partial x_{ji}} = 0$ and so this gives the condition that

$$x_{ji} = \frac{c_{ij} + c_{ji}}{\frac{c_i}{x_i} + \frac{c_j}{x_j}}$$

and we can iterate this condition to convergence.

5.5 Error Propagation

We start again with the posterior distribution of transition matrix $P \in \mathbb{R}^{n \times n}$:

$$\mathbb{P}(P|Z) \propto \prod_{i,j} p_{ij}^{c_{ij}}, \tag{5.4}$$

setting $U = (u_{ij}) = (c_{ij} + 1)$, we can rewrite this as:

$$\mathbb{P}(P|Z) \propto \prod_{i} \prod_{j} p_{ij}^{u_{ij}-1} = \prod_{i} Dir\,(p_i, u_i)$$

where *Dir* is a Dirichlet distribution, a well-known distribution. Based on known properties of this distribution we can state using the abbreviation $u_i = \sum_j u_{ij}$:

$$\begin{aligned}
\bar{p}_{ij} = [\mathbb{E}(P)]_{ij} &= \frac{u_{ij}}{u_i} = \frac{c_{ij}+1}{c_i+n} \\
\hat{p}_{ij} = [\hat{P}]_{ij} = (\arg\max \mathbb{P}(P|Z))_{ij} &= \frac{u_{ij}-1}{u_{ij}-n} = \frac{c_{ij}}{c_i} \\
Var(p_{ij}) &= \frac{u_{ij}(u_i-u_{ij})}{u_i^2(u_i+1)} = \frac{\bar{p}_{ij}(1-\bar{p}_{ij})}{(u_i+1)} = \frac{\bar{p}_{ij}(1-\bar{p}_{ij})}{c_i+n+1} \\
Cov(p_{ij}, p_{ik}) &= \frac{-u_{ij}u_{ik}}{u_i^2(u_i+1)} \quad \forall i \neq j
\end{aligned}$$

We now are interested in the question how the uncertainties given by $Var(p_{ij})$ propagate onto uncertainties of functions derived from transition matrices, such as eigenvalues. If we do not use constraints between different rows such as detailed balance, the rows can be treated as independent sets of random variables and thus:

$$Cov\left(p_{ij}, p_{lk}\right) = 0,\ i \neq l$$

We can thus define a Covariance matrix separately for each row as:

$$\begin{aligned}\Sigma_{jk}^{(i)} := Cov\left(p_{ij}, p_{ik}\right) &= \frac{1}{u_i^2(u_i+1)}\left[u_i\delta_{jk}u_{ij} - u_{ij}u_{ik}\right] \\ &= \frac{1}{(u_i+1)}\left[\delta_{jk}\bar{p}_{ij} - \bar{p}_{ij}\bar{p}_{ik}\right]\end{aligned}$$

where δ is the Kronecker delta. Alternatively, we can write the covariance matrix in vector notation:

$$\Sigma^{(i)} = \frac{1}{(u_i+1)}\left[diag\left(\bar{p}_i\right) - \bar{p}_i\bar{p}_i^T\right]$$

This means, that the covariance for the Dirichlet processes scales with the total number of count in a row.

Using this first two moments of the distribution we can approximate each row in the transition matrix P_i by a multivariate Gaussian distribution of the form

$$P_i \sim \mathcal{N}\left(\hat{p}_i, \Sigma^{(i)}\right)$$

This we can use in Gaussian error propagation for linear functions of the transition matrix (see above). Let us assume we have a scalar function $f(P) : \mathbb{R}^{n\times n} \to \mathbb{R}$. The first order Taylor approximation is given by:

$$f(P) = f(\hat{P}) + \sum_{i,j}\frac{\partial f}{\partial p_{ij}}\left(\hat{P}\right)\left(p_{ij} - \hat{p}_{ij}\right).$$

Since we know that the rows are independent we define a sensitivity vector for each row separately

$$s_j^{(i)} = \frac{\partial f}{\partial p_{ij}}(\hat{P})$$

and with the function for the error propagation we get

$$\hat{f} = f\left(\hat{P}\right)$$

and

$$Var\left(f\right) = Cov(f,f) = \sum_i \left(s^{(i)}\right)^T \Sigma^{(i)} s^{(i)}$$

Example: Eigenvalues For properly normalized eigenvectors we have:

$$\begin{aligned} \Lambda &= LPR \\ \lambda^{(k)} &= l^{(k)} P r^{(k)} \\ &= \sum_{i,j} l_i^{(k)} p_{ij} r_j^{(k)} \\ \frac{\partial \lambda^{(k)}}{\partial p_{ij}} &= l_i^{(k)} r_j^{(k)} \end{aligned}$$

and for unnormlized eigenvectors this is corrected by:

$$\frac{\partial \lambda^{(k)}}{\partial p_{ij}} = \frac{l_i^{(k)} r_j^{(k)}}{\langle l^{(k)}, r^{(k)} \rangle}$$

and thus, using the linear perturbation approach:

$$\begin{aligned} Var\left(\lambda^{(k)}\right) &= \sum_{i=1}^{n} \sum_{a,b} \frac{\partial \lambda^{(k)}}{\partial p_{ia}} Cov(p_{ab}) \frac{\partial \lambda^{(k)}}{\partial p_{ib}} \\ &= \frac{1}{\langle l^{(k)}, r^{(k)} \rangle^2} \sum_{i=1}^{n} \sum_{a,b} l_i^{(k)} r_a^{(k)} \left(\sum_a \frac{u_{ia}(u_i - u_{ia})}{u_i^2(u_i+1)} + \sum_{a,b \neq a} \frac{-u_{ia} u_{ib}}{u_i^2(u_i+1)} \right) l_i^{(k)} r_b^{(k)} \end{aligned}$$

In the special case of $p = 1$ we have $l_i = \pi_i$, $r_i = 1$ with $l \cdot r = 1$ and we get

$$
\begin{aligned}
Var\left(\lambda^{(1)}\right) &= \sum_{i=1}^{n}\sum_{a\neq b}\frac{\partial\lambda^{(k)}}{\partial p_{ia}}Var(p_{aa}^{(i)})\frac{\partial\lambda^{(k)}}{\partial p_{ia}} + \sum_{a\neq b}\frac{\partial\lambda^{(k)}}{\partial p_{ia}}Cov(p_{ab}^{(i)})\frac{\partial\lambda^{(k)}}{\partial p_{ib}} \\
&= \sum_{i=1}^{n}\pi_i^2\left(\sum_a\frac{u_{ia}(u_i-u_{ia})}{u_i^2(u_i+1)} + \sum_{a,b\neq a}\frac{-u_{ia}u_{ib}}{u_i^2(u_i+1)}\right) \\
&= \sum_{i=1}^{n}\frac{\pi_i^2}{u_i^2(u_i+1)}\left(\sum_{a\neq b}u_{ia}u_i - u_{ia}u_{ia} - \sum_{a,b\neq a}u_{ia}u_{ib}\right) \\
&= \sum_{i=1}^{n}\frac{\pi_i^2}{u_i^2(u_i+1)}\left(u_i\sum_a(u_{ia}-u_{ia})\right) \\
&= 0
\end{aligned}
$$

which is the expected result, since the first eigenvalue is constant.

The limitation of this approach is that it does not work well in situations where the Transition matrix distribution is far from Gaussian (especially in the situation of little data). Furthermore, the more nonlinear a given function of interest is in terms of p_{ij}, the more the estimated uncertainty on this function might be wrong.

Further Reading:

- Error Estimation [3, 1]

5.6 Full Bayesian Estimation

Here, a general method to sample transition matrices according to the posterior probability (5.3) based on Markov Chain Monte Carlo (MCMC) is proposed. While it is computationally more expensive than the linear error analysis and the Dirichlet sampling, it is more general than these methods. In particular, it allows (i) the complete distribution of arbitrary observables to be approximated to the desired degree of accuracy, (ii) the sampling to be restricted to transition matrices fulfilling additional constraints, such as detailed balance and predefined π, and (iii) arbitrary prior distributions, $p(T)$, to be employed. The method is illustrated on μs MD simulations of a hexapeptide for which the distributions and uncertainties of the free energy differences between conformations, the transition matrix elements and the transition matrix eigenvalues are estimated.

To sample the distribution 5.3, a sampling procedure is proposed based on taking Monte-Carlo steps in T-space: Given a current matrix T and a proposed new matrix T', the acceptance probability is computed by:

$$p_{accept} = \frac{p(T' \to T)}{p(T \to T')} \frac{p(T'|C)}{p(T|C)} \tag{5.5}$$

where $p(T \to T')$ and $p(T' \to T)$ denote the probability to propose T' given T and vice versa. Any proposal step can be used to sample the probability $p(T|C)$, provided that two conditions are satisfied:

- $p(T \to T')$ and $p(T' \to T)$ can be evaluated for every possible proposal step, such that (5.5) can be evaluated.
- The proposal steps generate an ergodic chain, *i.e.* if $\mathcal{T}$ denotes the set of matrices to be sampled from, then from any matrix $T \in \mathcal{T}$ any other matrix $T' \in \mathcal{T}$ must be accessible with a finite number of steps.

MCMC sampling of transition matrices: See paper F. Noé: "Probability Distributions of Molecular Observables computed from Markov Models", J. Chem. Phys **128**, 244103 (2008).

5.6.1 Metropolis-Hastings sampling

We want to sample

$$p(P|Z) \propto p(P)p(Z|P) = p(P) \prod_{i,j} p_{ij}^{c_{ij}} \tag{5.6}$$

where p_{ij} is the transition probability from state i to state j and c_{ij} are the number of observed transitions in the data set. For the sake of simplicity we assume a uniform prior

$$p(P|Z) \propto p(Z|P) = \prod_{i,j} p_{ij}^{c_{ij}}, \tag{5.7}$$

however, the use of non-uniform priors is straight forward. Using the Metropolis-Hastings algorithm, we have

$$p_{MH} = \frac{p(P' \to P)\, p(P'|Z)}{p(P \to P')\, p(P|Z)} = \frac{p(P' \to P)\, p(Z|P')}{p(P \to P')\, p(Z|P)} = \frac{p(P' \to P) \prod_{i,j} p'^{c_{ij}}_{ij}}{p(P \to P') \prod_{i,j} p^{c_{ij}}_{ij}} \tag{5.8}$$

The goal is to find proposal steps $P \to P'$ and the corresponding proposal probability $p(P \to P')$, such that the proposed transition matrices P' fulfill certain constraints, *e.g.* stochasticitiy, detailed balance or fixed stationary distribution.

5.6.2 Non-reversible shift element

Shifts a single off-diagonal element p_{ij} of the original transition matrix P by Δ. The reverse element p_{ji} is not altered and, therefore, detailed balance is not guaranteed. Stochasticity is preserved by an appropriate modification of the diagonal element p_{ii}:

$$p'_{ij} = p_{ij} - \Delta \tag{5.9}$$
$$p'_{ii} = p_{ii} + \Delta\,, \tag{5.10}$$

where Δ is a uniform random number in $\left[-p_{ii}, p_{ij}\right]$. The forward and backward proposal probabilities are symmetric:

$$p(P \to P') = \frac{1}{p_{ij} + p_{ii}} \tag{5.11}$$
$$p(P' \to P) = \frac{1}{p_{ij} - \Delta + p_{ii} + \Delta} = p(P \to P')\,. \tag{5.12}$$

and, thus,

$$p_{MH} = \frac{\prod_{i,j} {p'}_{ij}^{c_{ij}}}{\prod_{i,j} p_{ij}^{c_{ij}}} = \left(\frac{p_{ij} - \Delta}{p_{ij}}\right)^{c_{ij}} \left(\frac{p_{ii} + \Delta}{p_{ii}}\right)^{c_{ii}} \tag{5.13}$$

5.6.3 Reversible shift element

Shifts a off-diagonal element p_{ij} of the original transition matrix P by Δ. The reverse element p_{ji} is altered such that detailed balance

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{5.14}$$

is maintained. Stochasticity is preserved by an appropriate modification of the diagonal elements p_{ii} and p_{jj}:

$$p'_{ij} = p_{ij} - \Delta \tag{5.15}$$
$$p'_{ji} = p_{ji} - \frac{\pi_i}{\pi_j}\Delta \tag{5.16}$$
$$p'_{ii} = p_{ii} + \Delta \tag{5.17}$$
$$p'_{jj} = p_{jj} + \frac{\pi_i}{\pi_j}\Delta \tag{5.18}$$

Allowed range of Δ:

upper bound:

(a) $p'_{ij} = p_{ij} - \Delta \geq 0 \quad \rightarrow \quad \Delta \leq p_{ij}$

(b) $p'_{ji} = p_{ji} - \frac{\pi_i}{\pi_j}\Delta \geq 0 \quad \rightarrow \quad \Delta \leq \frac{\pi_j}{\pi_i} p_{ji} = p_{ij}$

lower bound:

(c) $p'_{ii} = p_{ii} + \Delta \geq 0 \quad \rightarrow \quad \Delta \geq -p_{ii}$

(d) $p'_{jj} = p_{jj} + \frac{\pi_i}{\pi_j}\Delta \geq 0 \quad \rightarrow \quad \Delta \geq -\frac{\pi_j}{\pi_i} p_{jj}$

and thus

$$\Delta \in \left[\max\left(-p_{ii}, -\frac{\pi_j}{\pi_i} p_{jj} \right), p_{ij} \right] . \tag{5.19}$$

The proposal probability in the forward direction is

$$p(P \rightarrow P') = \frac{1}{p_{ij} - \max\left(-p_{ii}, -\frac{\pi_j}{\pi_i} p_{jj} \right)} \tag{5.20}$$

and in the backward direction

$$\begin{aligned} p(P' \rightarrow P) &= \frac{1}{p'_{ij} - \max\left(-p'_{ii}, -\frac{\pi_j}{\pi_i} p'_{jj} \right), p'_{ij}} && (5.21) \\ &= \frac{1}{p_{ij} - \Delta - \max\left(-p_{ii} - \Delta, -\frac{\pi_j}{\pi_i}(p_{jj} + \frac{\pi_i}{\pi_j}\Delta) \right)} && (5.22) \\ &= \frac{1}{p_{ij} - \Delta - \max\left(-p_{ii} - \Delta, -\frac{\pi_j}{\pi_i} p_{jj} - \Delta \right)} && (5.23) \\ &= \mathbb{P}(P \rightarrow P') . && (5.24) \end{aligned}$$

The accepting probability is

$$\begin{aligned} p_{acc} &= \min\{1, p_{MH}\} && (5.25) \\ & && (5.26) \\ p_{MH} &= \frac{\prod_{i,j} p'^{c_{ij}}_{ij}}{\prod_{i,j} p^{c_{ij}}_{ij}} = \left(\frac{p_{ij} - \Delta}{p_{ij}} \right)^{c_{ij}} \left(\frac{p_{ii} + \Delta}{p_{ii}} \right)^{c_{ii}} \left(\frac{p_{ji} - \frac{\pi_i}{\pi_j}\Delta}{p_{ji}} \right)^{c_{ji}} \left(\frac{p_{jj} + \frac{\pi_i}{\pi_j}\Delta}{p_{jj}} \right)^{c_{jj}} && (5.27) \end{aligned}$$

If P fulfills detailed balance, the stationary distribution remains unchanged ($\pi' = \pi$):

$$\begin{aligned} \pi P' = [&\pi_1, ..., \pi_{i-1}, \\ &\pi_1 p_{1i} + ... + \pi_i(p_{ii} + \Delta) + ... + \pi_j(p_{ji} - \frac{\pi_i}{\pi_j}\Delta) + ... + \pi_m p_{mi}, \pi_{i+1}, ..., \pi_{j-1}, \\ &\pi_1 p_{1j} + ... + \pi_i(p_{ij} - \Delta) + ... + \pi_j(p_{jj} + \frac{\pi_i}{\pi_j}\Delta) + ... + \pi_m p_{mj}, \pi_{j+1}, ..., \pi_m] \\ = [&\pi_1, ..., \pi_m] = \pi. \end{aligned}$$

□

Furthermore, if P fulfills detailed balance, P' fulfills detailed balance as well:

$$\pi'_i p'_{ij} = \pi_i(p_{ij} - \Delta) = \pi_i p_{ij} - \pi_i \Delta$$
$$\pi'_j p'_{ji} = \pi_j(p_{ji} - \frac{\pi_i}{\pi_j}\Delta) = \pi_j p_{ji} - \pi_i \Delta$$

Thus

$$\pi_i p_{ij} = \pi_j p_{ji} \Rightarrow \pi'_i p'_{ij} = \pi'_j p'_{ji}.$$

□

5.6.4 Row Shift

Finally a step is considered which scales the self-transition probability, p_{ii}, and all outgoing transition probabilities, p_{ij} as follows:

$$\begin{aligned} p'_{ij} &= \alpha p_{ij} \\ p'_{ii} &= 1 - \sum_{k \neq i} p'_{ik} \\ &= 1 - \alpha \sum_{k \neq i} p_{ik} \\ &= 1 - \alpha(1 - p_{ii}) \\ &= \alpha p_{ii} - \alpha + 1 \end{aligned}$$

The step thus changes the ith row of P. The parameter α is subject to the following constraints:

$$\begin{array}{rcrl} \alpha p_{ij} \geq 0 & \rightarrow & \alpha \geq 0 & (a) \\ \alpha p_{ii} - \alpha + 1 \leq 1 & \rightarrow & \alpha \geq 0 & (b) \\ \alpha p_{ii} - \alpha + 1 \geq 0 & \rightarrow & \alpha \leq \frac{1}{1 - T_{ii}} & (c) \\ \alpha p_{ij} \leq 1 \ \forall j \neq i & \rightarrow & \alpha \leq \frac{1}{\max(T_{ij})} \ \forall j \neq i & (d) \end{array}$$

Note that $(1 - p_{ii}) \geq p_{ij}$ for all $j \neq i$, and thus $(1 - p_{ii})^{-1} \leq (\max(p_{ij}))^{-1}$, making (d) redundant with (c). Consequently, α is drawn uniformly from following range:

$$\alpha \in \left[0, \frac{1}{1 - p_{ii}}\right]$$

The ratio of proposal probabilities is given by:

$$\frac{p(P' \to P)}{p(P \to P')} = \frac{p_\alpha(P' \to P)}{p_\alpha(P \to P')} \frac{dA'}{dA} = \frac{dr' dA'}{dr\, dA},$$

where p_α is the proposal probability along the line parametrized by α, while dA is an area element orthogonal to that line and intersecting with P, and dA' is the scaled area element. With $p(\alpha) = 1 - p_{ii}$ and $\alpha = p'_{ij}/p_{ij} = (1 - p'_{ii})/(1 - p_{ii})$ we obtain (see paper) $p_\alpha(P' \to P)/p_\alpha(P \to P') = 1$. The area element is proportial to the $(m-2)$nd power of the distance of the ith row from $(0, ..., 0, p_{ii} = 1, 0, ..., 0)$, denoted by r:

$$dA \propto (r \cdot dc)^{(m-2)}$$
$$dA' \propto (r' \cdot dc)^{(m-2)}.$$

With $r = \sqrt{(1 - p_{ii})^2 + \sum_{j \neq i} p_{ij}^2}$ and $r' = \sqrt{\alpha^2(1 - p_{ii})^2 + \sum_{j \neq i} \alpha^2 p_{ij}^2} = \alpha r$ one obtains:

$$\frac{p(P' \to P)}{p(P \to P')} = \alpha^{(m-2)}.$$

Acceptance probability:

$$\begin{aligned} p_{acc} &= \frac{p(P' \to P)}{p(P \to P')} \frac{p(P'|C)}{p(P|C)} \\ &= \alpha^{(m-2)} \left(\frac{p'_{ii}}{p_{ii}}\right)^{C_{ii}} \prod_{j \neq i} \left(\frac{p'_{ij}}{p_{ij}}\right)^{C_{ij}} \\ &= \alpha^{(m-2)} \left(\frac{1 - \alpha(1 - p_{ii})}{p_{ii}}\right)^{C_{ii}} \prod_{j \neq i} \alpha^{C_{ij}} \\ &= \alpha^{(m-2+C_i-C_{ii})} \left(\frac{1 - \alpha(1 - p_{ii})}{p_{ii}}\right)^{C_{ii}} \end{aligned}$$

with $C_i = \sum_{j=1}^{m} C_{ij}$.

The row shift operation will change the stationary distribution π. π is, for example, required to conduct the reversible element shifts. Instead of recomputing the stationary distribution expensively by solving an eigenvalue problem, it may be efficiently updated as follows:

$$\pi'_i = \frac{\pi_i}{\pi_i + \alpha(1-\pi_i)}$$
$$\pi'_j = \frac{\alpha\pi_j}{\pi_i + \alpha(1-\pi_i)}.$$

Proof:

$$\pi' = \pi' P'$$

has the elements π_i and π_j, $j \neq i$, given by:

$$\begin{aligned}
\pi'_i &= \pi'_1 p'_{1i} + ... + \pi'_{i-1} p'_{i-1,i} + \pi'_i p'_{ii} + \pi'_{i+1} p'_{i+1,i} + ... + \pi'_m p'_{mi} \\
\frac{\pi_i}{\pi_i + \alpha(1-\pi_i)} &= \frac{\alpha\pi_1 p_{1i} + ... + \alpha\pi_{i-1} p_{i-1,i} + \pi_i[1-\alpha(1-p_{ii})] + \alpha\pi_{i+1} p_{i+1,i} + ... + \alpha\pi_m p_{mi}}{\pi_i + \alpha(1-\pi_i)} \\
\pi_i &= \alpha[\pi_1 p_{1i} + ... + \pi_m p_{mi}] + \pi_i - \alpha\pi_i \\
\pi_i &= \pi_1 p_{1i} + ... + \pi_m p_{mi}
\end{aligned}$$

and

$$\begin{aligned}
\pi'_j &= \pi'_1 p'_{1j} + ... + \pi'_{i-1} p'_{i-1,j} + \pi'_i p'_{ij} + \pi'_{i+1} p'_{i+1,j} + ... + \pi'_m p'_{mj} \\
\frac{\alpha\pi_j}{\pi_i + \alpha(1-\pi_i)} &= \frac{\alpha\pi_1 p_{1j} + ... + \alpha\pi_{i-1} p_{i-1,j} + \pi_i \alpha p_{ij} + \alpha\pi_{i+1} p_{i+1,j} + ... + \alpha\pi_m p_{mj}}{\pi_i + \alpha(1-\pi_i)} \\
\pi_j &= \pi_1 p_{1j} + ... + \pi_m p_{mj}.
\end{aligned}$$

Thus:

$$\pi = \pi P \Leftrightarrow \pi' = \pi' P'$$

□

Chapter 6

Markov Jump Processes

Time-continuous Markov processes When time steps are taken to be spaced differentially small, the above definition is also useful to define time-continuous Markov processes. For example Brownian motion, Langevin dynamics, and Master equation dynamics are all time-continuous Markov processes.

Intuitively, one can define a time-homogeneous Markov process as follows. Let $x(t) \in X = \{1, ..., n\}$ be the random variable describing the state of the process at time t. Now prescribe that, given that the process starts in a state i at time t, it has made the transition to some other state $j \neq i$ at time $t + h$ with probability given by

$$p(x(t+h) = j \mid x(t) = i) = k_{ij}h + o(h),$$

where $o(h)$ represents a quantity that goes to zero faster than h goes to zero. In other words, the limit

$$\lim_{h \to 0^+} \frac{dp(x(t+h) = j \mid x(t) = i)}{dh} = \lim_{h \to 0^+} \frac{d}{dh}(k_{ij}h + o(h)) = k_{ij}$$

exists. The transition rates k_{ij}, $i, j \in X$ defined the transition rate matrix $K \in \mathbb{R}^{n \times n}$. For the probabilities to be conserved, i.e., the probabilities associated with starting in a given state must add up to one, the off-diagonal elements of K must be non-negative and the diagonal elements must satisfy

$$k_{ii} = -\sum_{j \neq i} k_{ij}.$$

With this notation, and letting $\mathbf{p}_t \in \mathbb{R}^n$ the probability to be at any state at time t, the evolution of a continuous-time Markov process is given by the first-order differential equation

$$\frac{d}{dt}\mathbf{p}_t = \mathbf{p}_t\mathbf{K}$$

To see the reason of the definition of k_{ii}, we expand this equation out by components:

$$\begin{aligned}\frac{d}{dt}p_i(t) &= \sum_{j=1}^{n} p_j(t)k_{ji} \\ &= \sum_{j\neq i}^{n} p_j(t)k_{ji} + p_i(t)k_{ii} \\ &= \underbrace{\sum_{j\neq i}^{n} p_j(t)k_{ji}}_{\text{gain}} - \underbrace{\sum_{j\neq i} p_i(t)k_{ij}}_{\text{loss}}.\end{aligned}$$

which is known as the Master equation or gain-loss equation.

The probability that no transition happens in some time r is

$$\mathbb{P}(x(s) = i \,\forall\, s \in (t, t+h] \,|\, x(t) = i) = e^{k_{ii}h}.$$

That is, the probability distribution of the waiting time until the first transition is an exponential distribution with rate parameter $-k_{ii}$, and continuous-time Markov processes are thus memoryless processes.

A time dependent (time heterogeneous) Markov process is a Markov process as above, but with the k-rate a function of time, denoted $k_{ij}(t)$.

6.1 Poisson process

This is an important stochastic process in physics modelling countable, singular events in continuous time, such as:

- The arrival of "customers" in a queue.
- The number of raindrops falling over an area.
- The number of photons hitting a photodetector.
- The number of particles emitted via radioactive decay by an unstable substance, where the rate decays as the substance stabilizes.

It is parametrized by a "rate" or "intensity" constant k, or equivalently the characteristic timescale $\tau = k^{-1}$. A time-homogeneous Poisson process is defined as a counting process $N(t)$ where the number of counts $(N(t+\Delta t) - N(t))$ in the time interval $(t, t+\Delta t]$ follows a Poisson distribution with parameter $k\Delta t$. This relation is given as:

$$\mathbb{P}[(N(t+\Delta t) - N(t)) = n] = \frac{e^{-k\Delta t}(k\Delta t)^n}{n!} \qquad n \in \mathbb{N}_0 \tag{6.1}$$

k is the expected number of events per unit time and $\tau = k^{-1}$ is the expected duration between two events. It can be shown that the moments of this distribution are:

$$\mathbb{E}[n] = k\Delta t,$$

i.e. the mean number of counts is simply the waiting time times the rate, and also

$$Var[n] = k\Delta t.$$

For large enough statistics ($\Delta t \gg k^{-1}$), the central limit theorem makes the Poisson distribution converge towards a Gaussian distribution with corresponding moments:

$$\lim_{k\Delta t \to \infty} \mathbb{P}[(N(t+\Delta t) - N(t)) = n] = \mathcal{N}(k\Delta t, k\Delta t).$$

For short enough times, the probability to get more than one count in Δt vanishes:

$$\begin{aligned}
\lim_{k\Delta t \to 0} \frac{\mathbb{P}[(N(t+\Delta t) - N(t)) > 1]}{\mathbb{P}[(N(t+\Delta t) - N(t)) = 1]} &= \lim_{k\Delta t \to 0} \frac{1}{e^{-k\Delta t} k\Delta t} \sum_{n=2}^{\infty} \frac{e^{-k\Delta t}(k\Delta t)^n}{n!} \\
&= \lim_{k\Delta t \to 0} \sum_{n=2}^{\infty} \frac{(k\Delta t)^{n-1}}{n!} \\
&= 0
\end{aligned}$$

Thus in the short-time limit, we can ignore the possibility to get more than one count:

$$\begin{aligned}
\mathbb{P}[(N(t+\Delta t) - N(t)) = 0] &= e^{-k\Delta t} \\
\mathbb{P}[(N(t+\Delta t) - N(t)) = 1] &= k\Delta t\, e^{-k\Delta t} \\
\mathbb{P}[(N(t+\Delta t) - N(t)) > 1] &\approx 0.
\end{aligned}$$

for $k\Delta t \to 0$, we also have $e^{-k\Delta t} \approx 1 - k\Delta t$

$$\begin{aligned}\mathbb{P}[(N(t+\Delta t) - N(t)) = 0] &\approx 1 - k\Delta t \\ \mathbb{P}[(N(t+\Delta t) - N(t)) = 1] &\approx k\Delta t(1 - k\Delta t) \approx k\Delta t.\end{aligned}$$

which is identical to a binomial distribution $\mathbb{P}[(N(t+\Delta t) - N(t)) = n] = (k\Delta t)^n(1-k\Delta t)^{1-n}$ for $n \in \{0,1\}$.

Let us go back to the general case (6.1). Without loss of generality we start the process at time $t = 0$ and ask for the probability of t_1, the time at which the first count occurs:

$$\begin{aligned}dt\, f_{t_1}(t) &= \mathbb{P}[(N(t_1) - N(0) = 0]\, \mathbb{P}[(N(t_1 + dt) - N(t_1) = 1] \\ &= e^{-kt_1} e^{-k\,dt} k\, dt \\ &= dt\, k\, e^{-kt_1}\end{aligned}$$

Due to the independence of events in subsequent intervals, the probability density of any intermittance time t between two counts is:

$$\begin{aligned}f_t(t) &= k\, \exp(-kt) \\ &= \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right).\end{aligned}$$

This also shows immediately that the mean waiting time is equal to τ:

$$\mathbb{E}[t] = \int_{t=0}^{\infty} f_t(t)\; t\; dt = \tau = k^{-1}$$

while the variance is

$$Var[t] = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \tau^2 = k^{-2}$$

6.2 Markov Jump Process

Consider a system that can exist in two states such as the chemical reaction:

$$1 \xrightarrow{k} 2$$

and let the time-dependent populations of agents "1" and "2" be denoted by $p_1(t)$ and $p_2(t)$. We can model their evolution by the following differential equations

$$-\frac{dp_1(t)}{dt} = kp_1(t) = \frac{dp_2(t)}{dt}$$

Solving for p_1:

$$\begin{aligned} \frac{dp_1}{p_1} &= -kdt \\ \int_{p_1(0)}^{p_1(t)} \frac{dp_1}{p_1} &= -k\int_0^t dt \\ \ln\frac{p_1(t)}{p_1(0)} &= -kt \\ p_1(t) &= p_1(0)\exp(-kt). \end{aligned}$$

and for p_2:

$$\begin{aligned} \int_{p_2(0)}^{p_2(t)} dp_2(t) &= -\int_{p_1(0)}^{p_1(t)} dp_1(t) \\ p_2(t) - p_2(0) &= p_1(0) - p_1(t) \\ p_2(t) &= p_1(0) + p_2(0) - p_1(0)\exp(-kt) \end{aligned}$$

Generally we can say that both time evolutions have the functional form $p_i(t) \sim \exp(-kt)$, while multiplicative and additive constants can be used to realize different solutions.

Here, p_1 and p_2 are concentrations or amounts of the corresponding chemicals. Now we change the viewpoint and look at a single copy of molecule 1. We ask at which time, t, this copy will decay to molecule 2. Since this decay occurs with a constant rate and independent of the past, it has to be $f_t(t) \propto \exp(-kt)$, which normalized becomes:

$$f_t(t) = k\exp(-kt),$$

thus each exit process of a Markov jump process is a Poisson process. In other words, a Markov jump process (which can generally consist of many states that interconvert), is a superposition of Poisson processes, each occurring with a rate that corresponds to the exit rate out of the current state.

As an example consider $C_2H_6 \rightarrow 2CH_3$. This reaction is irreversible and has a rate of $5.46 \cdot 10^{-4}s^{-1}$ under normal temperature and pressure conditions.

Let us now consider the general case: We have m states (e.g. chemical species), for each of them we consider a concentration, population or probability

$$p_i \in \mathbb{R},\ i \in \{1, ..., m\}.$$

And we have a number of simple reactions between species of the sort

$$p_i \overset{k_{ij}}{\to} p_j$$

between some pairs $i, j \in \{1, ..., m\}$. We can formally define all rates

$$k_{ij} \in \mathbb{R},\ i, j \in \{1, ..., m\},$$

and set those to 0 where no reaction exists.

6.3 Master equation

Now the time evolution of the p_i is given by a differential equation with a sum of gains and losses, called *Master equation*:

$$\begin{aligned}
\frac{dp_i}{dt} &= \sum_{j \neq i} \left[k_{ji} p_j - k_{ij} p_i \right] \\
&= \sum_{j \neq i} k_{ji} p_j + p_i \sum_{j \neq i} -k_{ij} \\
&= \sum_{j \neq i} k_{ji} p_j + k_{ii} p_i \\
&= \sum_{j=1}^{m} k_{ji} p_j
\end{aligned}$$

where we have formally defined

$$k_{ii} = \sum_{j \neq i} -k_{ij},$$

allowing us to write the Master equation in matrix form simply as:

$$\frac{d\mathbf{p}^T(t)}{dt} = \mathbf{p}^T(t)\mathbf{K}.$$

with generator / rate matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ containing nonnegative off-diagonal and nonpositive diagonal entries:

$$\begin{aligned}
k_{ij} &\geq 0 \ \forall i \neq j \\
k_{ii} &= -\sum_{i \neq j} k_{ij} \ \forall i
\end{aligned}$$

Note that we have conservation of mass, i.e.

$$\|\mathbf{p}\|_1 = \sum_{i=1}^{m} p_i = p_{tot} = const.$$

Using the normalization convention $p_{tot} = 1$, we can look at the same equation in terms of moving probability densities:

$$\frac{d\mathbf{p}(t)^T}{dt} = \mathbf{p}(t)^T\mathbf{K}.$$

The Master equation has the following formal solution:

$$\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t)\exp(\mathbf{K}\tau) \tag{6.2}$$

(of course all constant shifts are also solutions).

This suggests to compare to Markov chains, where we have:

$$\mathbf{p}^T(t+\tau) = \mathbf{p}^T(t)\mathbf{T}(\tau), \tag{6.3}$$

showing that we have the equivalence:

$$\mathbf{T}(\tau) = \exp(\mathbf{K}\tau). \tag{6.4}$$

where $\exp(\cdot)$ is the Matrix exponential, which is defined as:

$$\begin{aligned}\exp(\mathbf{K}\tau) &= \exp(\tau\mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1}) \\ &= \exp(\tau\mathbf{R}diag(\kappa_1,...,\kappa_m)\mathbf{R}^{-1}) \\ &= \mathbf{R}diag(\exp(\tau\kappa_1),...,\exp(\tau\kappa_m))\mathbf{R}^{-1}\end{aligned}$$

where

$$\mathbf{R} = [\mathbf{r}_1,...\mathbf{r}_m]$$

is the right eigenvector matrix with eigenvectors $\mathbf{r}_i$ in the columns, and

$$diag(\kappa_1,...,\kappa_m)$$

is a diagonal matrix with the eigenvalues κ_i on the diagonal.

Thus, the transition and the rate matrix are closely related. $\mathbf{T}(\tau)$ is the propagator for the *solution* of the Master equation for time interval τ. In other

words, Eq. (6.3) is an exact solution of Eq. (6.2) at time points τ, i.e. an exact time-discretization. We can compare this to the *approximative* solution that is obtained when we make the time step τ very small, such that $\tau\kappa_i \to 0 \, \forall i \in \{1, ..., m\}$, and $\exp(\tau\kappa_i) \approx 1 + \tau\kappa_i \, \forall i \in \{1, ..., m\}$:

$$\begin{aligned} \exp(\mathbf{K}\tau) &= \mathbf{R}diag(\exp(\tau\kappa_1), ..., \exp(\tau\kappa_m))\mathbf{R}^{-1} \\ &\approx \mathbf{R}diag(1 + \tau\kappa_1, ..., 1 + \tau\kappa_m)\mathbf{R}^{-1} \\ &\approx \mathbf{Id} + \tau\mathbf{K}, \end{aligned}$$

which is a first-order Taylor approximation. The resulting equation is

$$\mathbf{T}(\tau) \approx \mathbf{Id} + \tau\mathbf{K}$$

and the associated solution approximation to the Master equation is

$$\mathbf{p}^T(t + \tau) \approx \mathbf{p}^T(t) + \tau\mathbf{p}^T(t)\mathbf{K},$$

which is just the Euler discretization of the Master equation.

Note that the formal inverse operation $\mathbf{K} = \tau^{-1}\log(\mathbf{T}(\tau))$ should be avoided. This operation is numerically very unstable (fluctuations of eigenvalues close to 0 become strongly amplified by the log), and also $\mathbf{T}(\tau)$ has a greater real-valued support than $\mathbf{K}$: $\mathbf{T}(\tau)$ may have negative eigenvalues, in which case no real-valued solution of the log exists. However the operation

$$\mathbf{K}_p(\tau) = \frac{\mathbf{T}(\tau) - \mathbf{I}}{\tau}$$

is always permitted. $\mathbf{K}_p(\tau) \approx \mathbf{K}$ is called *pseudogenerator*, and it is an approximation to the true rate matrix in a Markov jump process. The approximation becomes exact in the limit:

$$\mathbf{K} = \lim_{\tau \to 0^+} \frac{\mathbf{T}(\tau) - \mathbf{I}}{\tau},$$

and if that limit exists, then the process is said to have the generator $\mathbf{K}$. When in fact the underlying process is already a time-discrete Markov chain at some finite time τ, then $\mathbf{K}_p(\tau)$ is exact and we can treat the Markov chain as a Markov jump process with $\mathbf{K}_p(\tau)$ as rate matrix.

Due to the close relationship to Markov chains, many properties of Markov chains have direct equivalents for Markov jump processes. If $\mathbf{K}$ is irreducible and positively recurrent it has a stationary distribution given by

$$0 = \boldsymbol{\pi}^T \mathbf{K}.$$

The backward generator is defined as:

$$\tilde{k}_{ij} = \frac{\pi_j}{\pi_i} k_{ji}$$

If $\mathbf{K}$ is furthermore reversible it fulfills the detailed balance equation:

$$\tilde{k}_{ij} = k_{ij} \, \forall i, j.$$
$$\pi_i k_{ij} = \pi_j k_{ji} \, \forall i, j.$$

6.4 Solving very large systems

Consider a situation where state space is huge. For example, in the limit of few particles, we would no longer like to use particle concentrations, but would like to resolve particle number. For example, a virus A can replicate with some rate k_1, be killed by a drug B with some other rate k_2, but the drug also degenerates with some rate k_3:

$$\begin{aligned} A &\xrightarrow{k_1} 2A \\ A + B &\xrightarrow{k_2} B \\ B &\xrightarrow{k_3} 0 \end{aligned}$$

Now we are interested in the question: Given a set of rate constant and the fact that a patient has initially $[A]$ virus particles and receives $[B]$ drug molecules, will the virus be killed, or will it survive? We need to resolve few particle numbers, because a single surviving virus could spawn a new population, or a single surviving drug can still kill a small population. So the difference between 0 and 1 copies is important and cannot be resolved by a concentration. Now the state space of our system is huge, because each combination of copy numbers of A and B is a state, e.g. [(0,0), (0,1), (1,0), ...]. In principle the state space is infinite, but even when we limit ourselves to 1000 copies per species at most (which is only reasonable for certain settings of initial concentrations and rate constants), then we already have 10^6 states, and we don't want to make linear algebra operations with a $10^6 \times 10^6$ rate matrix. Therefore we need a way to simulate this system without ever evaluating the entire matrix.

We notice two facts:

1. At a given time t, consider we are in state i. Exit out of this state occurs via a Poisson process with rate $k_i = \sum_j k_{ij} = -k_{ii}$. This means the lifetime, or waiting time, τ_i, is distributed as:

$$p(\tau_i) = k_i \exp(-k_i \tau_i).$$

with the cdf

$$\begin{aligned} F(\tau_i) = \mathbb{P}(\tau_i' \leq \tau_i) &= k_i \int_0^{\tau_i} dt \ \exp(-k_i \tau_i') \\ &= k_i[-\frac{1}{k_i}\exp(-k_i\tau_i) + \frac{1}{k_i}] \\ &= 1 - \exp(-k_i \tau_i) \end{aligned}$$

whose inverse is:

$$\tau_i = -\frac{\ln[1 - F(t_0)]}{k_i}$$

thus, if v is a uniform random variable in $[0, 1]$, then $-\ln u / k_i$ is distributed as $p(t)$.

2. Given that we in a state i and have two options to leave, using rates k_1 and k_2, what is the probability to choose 1 versus 2? Call the exit time to state 1 t_2 and to state 2 t_2:

$$\begin{aligned} \mathbb{P}(t_1 < t_2) &= \int_{t=0}^{\infty} p(t_1 = t)\, \mathbb{P}(t_2 > t)\, dt \\ &= \int_{t=0}^{\infty} k_1 \exp(-k_1 t)\, [1 - (1 - \exp(-k_2 t)]\, dt \\ &= \int_{t=0}^{\infty} k_1 \exp(-k_1 t)\ \exp(-k_2 t)\, dt \\ &= \int_{t=0}^{\infty} k_1 \exp(-t(k_1 + k_2))\, dt \\ &= \frac{k_1}{k_1 + k_2} \end{aligned}$$

and

$$\mathbb{P}(t_2 < t_1) = \frac{k_2}{k_1 + k_2}.$$

This argument can be easily extended to more than 2 states.

Based on these two arguments, we can propose an algorithm to sample trajectories from rate systems without evaluating the full matrix. The Gillespie algorithm for simulating the time evolution of a system where some processes can occur with known rates K can be written as follows:

1. Set the time $t = 0$ and initial state $i = i_0$.

2. Form a list of all possible states that can be reached from the current state i and corresponding rates $k_{i1}, ..., k_{in}$. Let the total reaction rate be $k_i = \sum_{j=1}^{n} k_{ij}$

3. Calculate the cumulative function $p_j = \frac{\sum_{s=1}^{j} k_{is}}{k_i}$ for $j = 0, \ldots, n$.

4. Generate a uniform random number $u \in (0, 1]$ and find the event to carry out j by finding the j for which $p_{j-1} < u \leq p_j$ (this can be achieved efficiently using binary search). Update state i

5. Generate a uniform random number $v \in (0, 1]$ and update time by $t = t - \frac{\log v}{k_i}$

6. Return to step 2.

This algorithm is known in different sources variously as the residence-time algorithm or the n-fold way or the Bortz-Kalos-Liebowitz (BKL) algorithm, Gillespie algorithm, or just the kinetic Monte Carlo (KMC) algorithm. It is important to note that the timestep involved is a function of the probability that all events i, did not occur.

6.5 Hitting Probabilities, Committors and TPT fluxes

Similar as for transition matrices, we can derive simply expressions for transition path theory using rate matrices.

Hitting probabilites for Markov Jump Processes Using

$$\begin{aligned} h_i^A &= 1 \text{for } i \in A \\ h_i^A &= \sum_{j \in I} p_{ij} h_j^A \text{for } i \notin A. \end{aligned}$$

with $K = P - I$ (The linear expansion is sufficient for solving a linear system of equations) yields

$$\begin{aligned} h_i^A &= 1 \text{for } i \in A \\ h_i^A &= \sum_{j \in I, j \neq i} k_{ij} h_j^A + (k_{ii} + 1) h_i^A \text{for } i \notin A. \end{aligned}$$

and thus

$$\begin{aligned} h_i^A &= 1 \text{for } i \in A \\ \sum_{j \in I} k_{ij} h_j^A &= 0 \text{for } i \notin A. \end{aligned}$$

Committor Probabilities

$$\hat{k}_{ij} = \begin{cases} k_{ij} & i \notin A \\ 0 & i \in A \end{cases}$$

substituted into the Dirichlet problem:

$$\begin{aligned} q_i^+ &= 1 \text{for } i \in B \\ \sum_{j \in I} k_{ij} q_j^+ &= 0 \text{for } i \notin B. \end{aligned}$$

yields

$$\begin{aligned} q_i^+ &= 0 \text{for } i \in A \\ q_i^+ &= 1 \text{for } i \in B \\ \sum_{j \in I} k_{ij} q_j^+ &= 0 \text{for } i \notin \{A, B\}. \end{aligned}$$

In order to get the backward committor, we define the backwards propagator $\tilde{k}_{ij} = \frac{\pi_j}{\pi_i} k_{ji}$ and obtain:

$$\begin{aligned} q_i^- &= 1 \text{for } i \in A \\ q_i^- &= 0 \text{for } i \in B \\ \sum_{j \in I} \tilde{k}_{ij} q_j^- &= 0 \text{for } i \notin \{A, B\}. \end{aligned}$$

for reversibility / detailed balance, $k_{ij} = \frac{\pi_j}{\pi_i} k_{ji}$ and it can be easily checked that:

$$q^- = 1 - q^+$$

The TPT fluxes are exactly equivalent:

$$\begin{aligned} f_{ij} &= \pi_i q_i^- k_{ij} q_i^+ \\ f_{ij}^+ &= \max\{0, f_{ij} - f_{ji}\}. \end{aligned}$$

Chapter 7

Continuous Markov Processes

7.1 Random Walk

Define a random process on the set of whole numbers that in each turn makes a step, randomly and equally probable by +1 or -1, starting with position 0. The number of different walks of n steps where each step is +1 or −1 is clearly 2^n. For the simple random walk, each of these walks are equally likely. Thus, to have m "+1" steps and $n - m$ "-1" steps we have $\binom{n}{m}$ combinations. Such a walk will proceed to the coordinate $s_n = 2m - n$, such that we can also write the number of combinations as $\binom{n}{(n+s_n)/2}$, which is 0 if $n + s_n$ is odd. Therefore, the probability density of s_n is equal to

$$\mathbb{P}(s_n) = 2^{-n}\binom{n}{(n+s_n)/2}.$$

As an alternative view, we can draw n random numbers x_i uniformly from $\{-1, 1\}$. We define the random variable s_n as

$$S_n = \sum_{i=1}^{n} x_i$$

7.1.1 Long-time approximation and transition to continuous variables

Since the x_i are independent, we have

$$Var(s_n) = \sum_{i=1}^{n} Var(x_i) = n$$

Following the central limit theorem (see above), we obtain for large n the approximation

$$s_n \sim \mathcal{N}(0, \frac{n^2\sigma^2}{n}) = \mathcal{N}(0, n),$$

i.e. the position of the walker is distributed as a Gaussian with zero mean and variance n. This is identical to saying that after an initial burnin phase the mean square displacement of the walker grows linearly with time n, or its mean displacement (standard deviation) with square root of time, $\sqrt{n}$.

Thus,

$$\mathbb{P}(s_n) \approx \frac{1}{\sqrt{2\pi n}} \exp(-\frac{s_n^2}{2n})$$

this can also be shown using the de Moivre-Laplace theorem, which is a central limit theorem for Binomial distributions.

Next, we go to a continuous space. We make the subsitutions:

$$\begin{aligned} n &= t/\Delta t \\ s_n &= x/\Delta x \end{aligned}$$

Since variable $s_n = x/\Delta x$ is substituted, we must make a change of variables (see above). We have $ds_n/dx = 1/\Delta x$:

$$\begin{aligned} p(x) &\approx \left|\frac{dS_n}{dx}\right| \mathbb{P}(S_n) \\ &= \frac{1}{|\Delta x|\sqrt{2\pi\frac{t}{\Delta t}}} \exp(-\frac{x^2\Delta t}{2t(\Delta x)^2}) \\ &= \frac{1}{\sqrt{4\pi D}} \exp(-\frac{x^2}{4Dt}) \\ &= \mathcal{N}_x(0, 2Dt) \end{aligned}$$

where we have defined the diffusion constant

$$D := \frac{\Delta x^2}{2\Delta t}.$$

which determines the proportionality constant of the increase of the mean square displacement:

$$2D\Delta t = \Delta x^2.$$

Hence we can also write down a transition density, i.e. the conditional probability density to jump from point x to point y in time step τ:

$$p(x, y; \tau) = \frac{1}{\sqrt{4\pi D}} \exp(-\frac{(x-y)^2}{4D\tau})$$

7.1.2 Wiener Process and Brownian dynamics

In mathematics, the Wiener process is a continuous-time stochastic process named in honor of Norbert Wiener. It is often called Brownian motion, after Robert Brown. It is one of the best known Lévy processes (càdlàg stochastic processes with stationary independent increments) and occurs frequently in pure and applied mathematics, economics and physics.

Definition:

1. $W_0 = 0$
2. W_t is almost surely continuous
3. W_t has independent increments with distribution $W_t - W_s \sim \mathcal{N}(0, t-s)$, (for $0 \leq s \leq t$).

The condition that it has independent increments means that if $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$ then $W_{t1} - W_{s1}$ and $W_{t2} - W_{s2}$ are independent random variables, and the similar condition holds for n increments.

Consider the unit Brownian motion (normalized white noise) process:

$$dx = dW_t$$

Numeric realization at time step Δt:

1. $x_0 = 0$
2. $x_{t+1} = x_t + \sqrt{t}\Delta W$ with $\Delta W \sim \mathcal{N}(0, 1)$

Brownian dynamics

$$dx = \sigma dW_t = \sqrt{2D} dW_t$$

This scaled Wiener process is simply the continuous version of the 1-dimensional random walk, also known as one-dimensional free diffusion with diffusion, because:

1. $x_0 = 0$

2. $x_t \sim \mathcal{N}(0, 4Dt)$

Since we have an explicit description of the time-dependent probability distribution, we can also imagine an ensemble of trajectories to evolve in time, whose distribution has the time-dependent moments:

$$\begin{aligned} \mu(t) &= 0 \\ \sigma(t) &= \sqrt{2Dt} \end{aligned}$$

At $t \to 0_+$ the distribution starts out as a Dirac function at $x = 0$ and then broadens into a Gaussian that gets flatter and flatter. This is a Fokker-Planck type description and will be studied in more detail later.

7.2 Langevin and Brownian Dynamics

Consider the one-dimensional Langevin equation in $x \in \mathbb{R}$

$$m\frac{d^2x}{dt^2} = -\nabla V(x) - \gamma m\left[\frac{dx}{dt} + \sqrt{2D}dW\right],$$

with mass m, friction γ and the Wiener process dW. In the overdamped case, the drag force due to friction is much larger than the inertial force ($|\gamma\dot{x}| \gg |m\ddot{x}|$), and it is thus assumed $|m\ddot{x}| \approx 0$, obtaining Brownian dynamics:

$$0 = -\nabla V(x) - \gamma m\left[\frac{dx}{dt} + \sqrt{2D}dW\right],$$

which is more conveniently written as:

$$\frac{dx}{dt} = -\frac{\nabla V(x)}{\gamma m} + \sqrt{2D}\dot{W},$$

The fluctuation-dissipation theorem (Einstein-Smoluchowski) relates friction and temperature T *via* the diffusion constant D:

$$D = \frac{k_B T}{\gamma m}$$

with k_B being the Boltzmann constant. This allows us to write:

$$\frac{dx}{dt} = -D\frac{\nabla V(x)}{k_B T} + \sqrt{2D}dW.$$

The stationary density of Brownian dynamics is:

$$\pi(x) \propto \exp\left(-\frac{V(x)}{k_B T}\right)$$

Using an Euler discretization, we can integrate the Brownian dynamics equation as:

$$\begin{aligned} x_{t+\tau} - x_t &= -D\tau\frac{\nabla V(x_t)}{k_B T} + \sqrt{2D\tau}\eta_t. \\ \eta_t &= \mathcal{N}(0,1). \end{aligned}$$

Calling the present position x and the subsequent position y, we can write down the conditional transition probability density:

$$\begin{aligned} p(x,y;\tau) &= \mathcal{N}_y\left(x - D\tau\frac{\nabla V(x)}{k_B T}, 2D\tau\right). \\ &= \frac{1}{\sqrt{4\pi D}}\exp\left[-\frac{\left(y - x + D\tau\frac{\nabla V(x)}{k_B T}\right)^2}{4D\tau}\right] \end{aligned}$$

Example: The local potential is assumed to be harmonic:

$$V(x) = \frac{\alpha}{2}(x-\mu)^2$$

yielding

$$\frac{dx}{dt} = -\alpha D\frac{x-\mu}{k_B T} + \sqrt{2D}dW.$$

and the stationary density:

$$\pi(x) \propto \exp\left(-\frac{\alpha(x-\mu)^2}{2k_B T}\right) = \mathcal{N}\left(\mu, \frac{k_B T}{\alpha}\right).$$

We will later see that when starting with a delta- or Gaussian distribution in a Brownian dynamics in a harmonic potential, then the solution is still Gaussian at any time later.

7.2.1 Applications

Molecular Dynamics Consider the motion of an ion in a channel

Cellular Dynamics Consider the motion of proteins in or on a membrane

7.2.2 Further reading

- Stochastic Processes: [5]

Chapter 8

Markov model discretization error

8.1 Basics

Let Ω be a real-valued vector space $\mathbb{R}^d$ (state space) and μ, p, l, r functions on a Hilbert space $\mathcal{H} = \{p : \Omega \to \mathbb{R} : \int_\Omega d\mathbf{x}\, p^2(\mathbf{x}) < \infty\}$ with scalar product $\langle u, v\rangle_w = \int_\Omega d\mathbf{x}\, u(\mathbf{x})\, v(\mathbf{x})\, w(\mathbf{x})$. We consider a Markov process $\mathbf{z}_t$ on Ω which is stationary and ergodic with respect to its unique stationary (invariant) distribution $\mu(\mathbf{x}) \equiv p(\mathbf{z}_t = \mathbf{x})\ \forall t$. We use variables $\mathbf{x}, \mathbf{y} \in \Omega$ to denote points in state space. The dynamics of the process $\mathbf{z}_t$ are characterized by the transition density

$$p(\mathbf{x}, \mathbf{y}; \tau) = p(\mathbf{z}_{t+\tau} = \mathbf{y} \mid \mathbf{z}_t = \mathbf{x}),$$

and the correlation density, i.e., the probability density of finding the process at points $\mathbf{x}$ and $\mathbf{y}$ at a time spacing of τ, is defined by

$$C(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{x})\, p(\mathbf{x}, \mathbf{y}; \tau) = p(\mathbf{z}_{t+\tau} = \mathbf{y},\ \mathbf{z}_t = \mathbf{x}).$$

We further assume $\mathbf{z}_t$ to be reversible with respect to its stationary distribution, i.e.:

$$\mu(\mathbf{x})\, p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y})\, p(\mathbf{y}, \mathbf{x}; \tau) \tag{8.1}$$

$$C(\mathbf{x}, \mathbf{y}; \tau) = C(\mathbf{y}, \mathbf{x}; \tau). \tag{8.2}$$

Reversibility is not strictly necessary but tremendously simplifies the forthcoming expressions and their interpretation [**?**]. In physical simulations, reversibility is the consequence of the simulation system being in thermal equilibrium with its environment, i.e. the dynamics in the system is purely a consequence of thermal fluctuations and there are no external driving forces.

Consider a probability density $p_t(\mathbf{x}) \equiv p(\mathbf{z}_t = \mathbf{x})$. We can write the propagation of this density *via* propagator $\mathcal{P}(\tau)$ as:

$$p_\tau(\mathbf{y}) = \mathcal{P}(\tau) \circ p_0(\mathbf{x}) = \int_{\mathbf{x}} d\mathbf{x}\, p_0(\mathbf{x})\, p(\mathbf{x}, \mathbf{y};\, \tau)$$

8.2 Spectral decomposition

The transition density can be decomposed into spectral components of the propagator [**?**]:

$$p(\mathbf{x}, \mathbf{y};\, \tau) = \sum_{i=1}^{\infty} \lambda_i(\tau)\, \mu^{-1}(\mathbf{x})\, l_i(\mathbf{x})\, l_i(\mathbf{y})$$

Suppose we are interested in the dominant m eigenvalues. We then only need to consider a finite sum, and a "fast" part that quickly decays in τ:

$$p(\mathbf{x}, \mathbf{y};\, \tau) = \sum_{i=1}^{m} \lambda_i(\tau)\, \mu^{-1}(\mathbf{x})\, l_i(\mathbf{x})\, l_i(\mathbf{y}) + p_{\text{fast}}(\mathbf{x}, \mathbf{y};\, \tau)$$

where l are eigenfunctions and λ_i are eigenvalues of the propagator, i.e.

$$\lambda_i(\tau)\, l_i = \mathcal{P}(\tau)\, l_i.$$

and eigenvalues shall be sorted as $\lambda_1 = 1 > \lambda_2 \geq \lambda_3$.... The eigenfunctions are normalized by:

$$\langle l_i, l_j \rangle_{\mu^{-1}} = \int_{\mathbf{x}} d\mathbf{x}\, \mu^{-1}(\mathbf{x})\, l_i(\mathbf{x})\, l_j(\mathbf{x}) \quad = \quad \delta_{ij}. \tag{8.3}$$

where δ_{ij} is the Kronecker delta. For the first term we have

$$l_1 = \mathcal{P}(\tau)\, l_1,$$

i.e. l_1 is proportional to the stationary (invariant) density. As a result of the normalization conditions we have $\|l_1\|_1 = \int_{\mathbf{x}} d\mathbf{x}\, l_1(\mathbf{x}) = 1$ and

$$l_1(\mathbf{x}) = \mu(\mathbf{x})$$

is identical to the stationary density. As a result of reversibility, the fast part p_{fast} vanishes for long times τ and we are left with:

$$p(\mathbf{x}, \mathbf{y};\, \tau) \quad = \quad \sum_{i=1}^{m} \lambda_i(\tau)\, \mu^{-1}(\mathbf{x})\, l_i(\mathbf{x})\, l_i(\mathbf{y}) \tag{8.4}$$

$$\quad = \quad \sum_{i=1}^{m} \exp(-\kappa_i \tau)\, \mu^{-1}(\mathbf{x})\, l_i(\mathbf{x})\, l_i(\mathbf{y}) \tag{8.5}$$

where κ_i is an implied rate. Correspondingly, the correlation density can be written as:

$$C(\mathbf{x}, \mathbf{y}; \tau) = \sum_{i=1}^{m} \lambda_i(\tau)\, l_i(\mathbf{x})\, l_i(\mathbf{y})$$

Example

Consider a two-well potential with a spectral gap, $\kappa_2 \ll \kappa_3$. Then, there exists a family of $\tau > p\kappa_3^{-1}$ where p is a sufficiently large number (usually 3 or greater), such that

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}; \tau) &\approx \mu(\mathbf{x}) + \lambda_2(\tau)\, \mu^{-1}(\mathbf{x})\, l_2(\mathbf{x})\, l_2(\mathbf{y}) \\ C(\mathbf{x}, \mathbf{y}; \tau) &\approx \mu(\mathbf{x})\mu(\mathbf{y}) + \lambda_2(\tau)\, l_2(\mathbf{x})\, l_2(\mathbf{y}). \end{aligned}$$

I.e., for describing the dynamics on long timescales, we only need to approximate μ, l_2 and λ_2.

8.3 Raleigh variational principle

In nontrivial dynamical systems neither the correlation densities $p(\mathbf{x}, \mathbf{y}; \tau)$ and $C(\mathbf{x}, \mathbf{y}; \tau)$ nor the eigenvalues λ_i and eigenfunctions l_i are analytically available. This section provides a variational principle based on which these quantities can be estimated from simulation data generated by the dynamical process $\mathbf{z}_t$. For this, the formalism introduced above is used to formulate the Raleigh variational principle used in quantum mechanics **[?]** for Markov processes.

The density $C(\mathbf{x}, \mathbf{y}; \tau)$ is considered to act as a the kernel of a correlation operator. Let f be a real-valued function of state, $f = f(\mathbf{x}) : \Omega \to \mathbb{R}$, its autocorrelation with respect to the stochastic process $\mathbf{z}_t$ is given by:

$$\mathrm{acf}(f; \tau) \quad = \quad \mathbb{E}[f(\mathbf{z}_0)\, f(\mathbf{z}_\tau)] = \int_{\mathbf{x}} \int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, f(\mathbf{x})\, C(\mathbf{x}, \mathbf{y}; \tau)\, f(\mathbf{y}) \qquad (8.6)$$

Remark 1. In the Dirac notation often used in physical literature, integrals such as the one above may be abbreviated by $\mathbb{E}[f(\mathbf{x}_0)\, f(\mathbf{x}_\tau)] = \langle f \mid \mathcal{C} \mid f \rangle$ with the correlation operator defined as $\mid \mathcal{C} \mid f \rangle = \int_{\mathbf{x}} d\mathbf{x}\, f(\mathbf{x})\, C(\mathbf{x}, \mathbf{y}; \tau)$.

Theorem 2. The autocorrelation function of a weighted eigenfunction $r_k = \mu^{-1} l_k$ is its eigenvalue λ_k:

$$\mathrm{acf}(r_k; \tau) = \mathbb{E}\left[r_k(\mathbf{z}_0)\, r_k(\mathbf{z}_\tau)\right] = \lambda_k$$

Proof. Using (8.6) with $f = \mu^{-1} l_k$, it directly follows that:

$$\begin{aligned}
\mathrm{acf}(r_k;\, \tau) &= \int_{\mathbf{x}} \int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, \mu^{-1}(\mathbf{x})\, l_k(\mathbf{x})\, C(\mathbf{x}, \mathbf{y};\, \tau)\, \mu^{-1}(\mathbf{y})\, l_k(\mathbf{y}) \qquad (8.7)\\
&= \sum_{i=1}^{m} \lambda_i(\tau) \int_{\mathbf{x}} \int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, \mu^{-1}(\mathbf{x})\, l_k(\mathbf{x})\, l_i(\mathbf{x})\, l_i(\mathbf{y})\, \mu^{-1}(\mathbf{y})\, l_k(\mathbf{y})\\
&= \sum_{i=1}^{m} \lambda_i(\tau)\, \langle l_k, l_i \rangle^2_{\mu^{-1}}\\
&= \sum_{i=1}^{m} \lambda_i(\tau)\, \delta_{ik}\\
&= \lambda_k(\tau).
\end{aligned}$$

□

Theorem 3. Let $\hat{r}_2 = \mu^{-1} \hat{l}_2$ be an approximate model for the second eigenfunction, which is normalized and orthogonal to the true first eigenfunction:

$$\langle \hat{l}_2, \mu \rangle_{\mu^{-1}} = 0 \qquad (8.8)$$

$$\langle \hat{l}_2, \hat{l}_2 \rangle_{\mu^{-1}} = 1, \qquad (8.9)$$

then

$$\mathrm{acf}(\hat{r}_2;\, \tau) = \mathbb{E}\left[\hat{r}_2(\mathbf{z}_0)\, \hat{r}_2(\mathbf{z}_\tau)\right] \leq \lambda_2$$

Proof. $\hat{l}_2$ is written in terms of the basis of eigenfunctions l_i:

$$\hat{l}_2 = \sum_i a_i l_i.$$

Due to $\langle \hat{l}_2, \mu \rangle_{\mu^{-1}} = \langle \hat{l}_2, 1 \rangle = 0$, $\mu(x) \geq 0$, and $\mu^{-1} \hat{l}_2 = \sum_i a_i \mu^{-1} l_i = a_1 1 + \sum_{i \geq 2} a_i \mu^{-1} l_i$ it follows $a_1 = 0$. Hence:

$$\hat{l}_2 = \sum_{i \geq 2} a_i l_i = a_2 l_2 + \sum_{i > 2} a_i l_i = a_2 l_2 + \epsilon.$$

Using the normalization condition (8.9), the following equality can be derived for the amplitudes:

$$\begin{aligned}
1 &= \langle \hat{l}_2, \hat{l}_2 \rangle_{\mu^{-1}}\\
&= \langle \sum_{i \geq 2} a_i l_i, \sum_{j \geq 2} a_j l_j \rangle_{\mu^{-1}}\\
&= \sum_{i \geq 2} a_i \sum_{j \geq 2} a_j \langle l_i, l_j \rangle_{\mu^{-1}}\\
&= \sum_{i \geq 2} a_i^2 \qquad (8.10)
\end{aligned}$$

then

$$\begin{aligned}
\mathrm{acf}(\mu^{-1}\hat{l}_2;\,\tau) &= \sum_i \lambda_i(\tau) \int_{\mathbf{x}}\int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, \mu^{-1}(\mathbf{x}) \left(\sum_{j\geq 2} a_j l_j(\mathbf{x})\right) l_i(\mathbf{x})\, l_i(\mathbf{y})\, \mu^{-1}(\mathbf{y}) \left(\sum_{j\geq 2} a_j l_j(\mathbf{y})\right) \\
&= \sum_i \lambda_i(\tau) \left[\int_{\mathbf{x}} d\mathbf{x}\, \mu^{-1}(\mathbf{x}) \left(\sum_{j\geq 2} a_j l_j(\mathbf{x})\right) l_i(\mathbf{x})\right]^2 \\
&= \sum_i \lambda_i(\tau) \left[\sum_{j\geq 2} a_j \langle l_j, l_i\rangle_{\mu^{-1}}\right]^2 \\
&= \sum_i a_i^2 \lambda_i(\tau) \\
&= a_2^2\lambda_2(\tau) + \sum_i a_i^2 \lambda_i(\tau) \\
&\leq a_2^2\lambda_2(\tau) + \sum_i a_i^2 \lambda_2(\tau) = \lambda_2(\tau) \sum_i a_i^2 = \lambda_2(\tau)
\end{aligned} \tag{8.11}$$

□

Corollary 4. Let $\hat{r}_k = \mu^{-1}\hat{l}_k$ be an approximate model for the k'th eigenfunction, with the normalization and orthogonality constraints:

$$\begin{aligned}
\langle \hat{l}_k, l_i\rangle_{\mu^{-1}} &= 0, \quad \forall i < k \\
\langle \hat{l}_k, \hat{l}_k\rangle_{\mu^{-1}} &= 1,
\end{aligned} \tag{8.12}$$

then

$$\mathrm{acf}(\hat{r}_k;\,\tau) = \mathbb{E}\left[\hat{r}_k(\mathbf{z}_0)\, \hat{r}_k(\mathbf{z}_\tau)\right] \leq \lambda_k$$

The proof is analogous to Theorem 3, with Eq. (8.10) being $\sum_{i\geq k} a_i^2 = 1$.

Remark 5. A crucial assumption of the variational principle given by Theorems (2) to (4) is that for estimating the k-th eigenfunction, the $k-1$ dominant eigenfunctions are already known. In particular, the first eigenfunction, i.e. the stationary density must be known. In practice, these eigenfunctions are approximated via solving a variational principle. Nonetheless, some basic statements can be made even if no eigenfunction is known exactly. For example, it is trivial that when the estimated stationary density $\hat{\mu}$ is used in Theorem 2, then the estimate of the first eigenvalue is still always correctly 1:

$$\mathrm{acf}(\hat{\mu}^{-1}\hat{\mu};\,\tau) = \mathrm{acf}(1;\,\tau) = 1$$

and from theorems 2 and 3 it follows that any function $\hat{r}_k \neq \hat{\mu}$

$$\mathrm{acf}(\hat{r}_k;\,\tau) < 1$$

hence the eigenvalue 1 is simple and dominant also when estimating eigenvalues from data.

Remark 6. An important insight at this point is that a variational principle of conformation dynamics can be formulated in terms of correlation functions. In contrast to quantum mechanics or other fields where the variational principle has been successfully employed, no closed-form expression of the operator $\mathcal{C}$ is needed. The ability to express the variational principle in terms of correlation functions with respect to $\mathcal{C}$, means that the eigenvalues to be maximized can be directly estimated from simulation data. When statistically sufficiently realizations of $\mathbf{z}_t$ are available, then the autocorrelation function can be estimated via:

$$\mathrm{acf}(\hat{r}_k;\, \tau) = \mathbb{E}(\hat{r}_k(\mathbf{z}_0)\hat{r}_k(\mathbf{z}_\tau)) \approx \frac{1}{N}\sum \hat{r}_k(\mathbf{z}_0)\hat{r}_k(\mathbf{z}_\tau)$$

where N are the number of simulated time windows of length τ.

8.4 Ritz method

The Ritz method is a systematic approach to find the best possible approximation to the m first eigenfunctions of an operator $\mathcal{C}$ simultaneously in terms of a linear combination of orthonormal functions [**?**]. Here the Ritz method is simply restated in terms of the present notation. Let $\chi_i : \Omega \to \mathbb{R}$, $i \in \{1, ..., m\}$ be a set of m orthonormal basis functions:

$$\langle \chi_i, \chi_j \rangle_\mu = \delta_{ij}$$

and let $\boldsymbol{\chi}$ denote the vector of these functions:

$$\boldsymbol{\chi}(\mathbf{x}) = [\chi_1(\mathbf{x}), ..., \chi_m(\mathbf{x})]^T.$$

We seek a coefficient matrix $\mathbf{B} \in \mathbb{R}^{m\times m}$ with

$$\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_m] \tag{8.13}$$

with the column vectors $\mathbf{b}_i = [b_{i1}, ..., b_{im}]^T$ that approximate eigenfunctions of operator $\mathcal{C}$ as:

$$\hat{r}_i(\mathbf{x}) = \mu^{-1}(\mathbf{x})\hat{l}_i(\mathbf{x}) \quad = \quad \mathbf{b}_i^T\boldsymbol{\chi}(\mathbf{x}) = \sum_j b_{ij}\,\chi_i(\mathbf{x}) \tag{8.14}$$

The matrix $\mathbf{B}$ that optimally approximates the eigenfunctions in terms of maximizing the Raleigh coefficients $\hat{\lambda}_i(\tau) = \mathrm{acf}(\hat{\mu}^{-1}\hat{l}_i;\, \tau)$ is found by the eigenvectors of the eigenvalue problem:

$$\mathbf{HB} = \mathbf{B}\boldsymbol{\Lambda}$$

with the individual eigenvalue/eigenvector pairs

$$\mathbf{H}\mathbf{b}_i = \mathbf{b}_i\hat{\lambda}_i$$

and the density matrix $\mathbf{H} = [h_{ij}]$ with:

$$h_{ij} = \int_{\mathbf{x}} \int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, \chi_i(\mathbf{x})\, C(\mathbf{x}, \mathbf{y}; \tau)\, \chi_j(\mathbf{y}) \tag{8.15}$$

$$= \mathbb{E}[\chi_i(\mathbf{z}_0)\, \chi_j(\mathbf{z}_\tau)], \tag{8.16}$$

or $h_{ij} = \langle \chi_i \mid \mathcal{C} \mid \chi_j \rangle$ in the Dirac notation.

Remark 7. Due to the equality between Eq. (8.15) and (8.16) the elements of the $\mathbf{H}$ matrix can be estimated as correlation functions of a simulation of the process $\mathbf{z}_t$

8.5 Roothaan-Hall method

The Roothaan-Hall method is a generalization of the Ritz method used for solving the linear parameter optimization problem for the case when the basis set is not orthogonal [**?**, **?**]. Let the matrix $\mathbf{S} \in \mathbb{R}^{m \times m}$ with elements

$$S_{ij} = \langle \chi_i, \chi_j \rangle$$

be the matrix of overlap integrals with the normalization conditions $S_{ii} = 1$. Note that $\mathbf{S}$ has full rank if and only if all χ_i are pairwise linearly independent. The optimal solution $\mathbf{B}$ in the sense of Eqs (8.13)-(8.14) is found by the eigenvectors of the generalized eigenvalue problem:

$$\mathbf{HB} = \mathbf{SB\Lambda}$$

with the individual eigenvalue/eigenvector pairs:

$$\mathbf{Hb}_i = \mathbf{Sb}_i \hat{\lambda}_i$$

The direct approach to solve this problem is given by

$$\mathbf{S}^{-1}\mathbf{HB} = \mathbf{B\Lambda}.$$

Remark 8. The Ritz and Roothaan-Hall methods are useful for eigenfunction models that are expressed in terms linear combinations of basis functions. Nonlinear parameter models can also be handled with nonlinear optimization methods. In such nonlinear cases it needs to be tested whether there is a unique optimum or not.

Markov state model

Let $\{S_1, ..., S_n\}$ be pairwise disjoint sets partitioning Ω and let $\pi_i = \int_{S_i} d\mathbf{x}\mu(\mathbf{x})$ be the stationary probability of set $S_i \subset \Omega$. Consider the piecewise constant functions

$$\chi_i = \mathbf{1}_{S_i}$$

then the S-matrix is given by:

$$\begin{aligned} S_{ij} &= \langle \chi_i, \chi_j \rangle \\ &= \delta_{ij} \int_{S_i} d\mathbf{x}\mu(\mathbf{x}) \\ &= \mathrm{diag}\boldsymbol{\pi} \\ &= \mathbf{\Pi}. \end{aligned}$$

The corresponding $\mathbf{H}$ matrix evaluates to

$$\begin{aligned} h_{ij} &= \int_{\mathbf{x}} \int_{\mathbf{y}} d\mathbf{x}\, d\mathbf{y}\, \mathbf{1}_{S_i}\, C(\mathbf{x}, \mathbf{y};\, \tau)\, \mathbf{1}_{S_j} && (8.17) \\ &= \int_{S_i} \int_{S_j} d\mathbf{x}\, d\mathbf{y}\, C(\mathbf{x}, \mathbf{y};\, \tau) \\ &= \int_{S_i} \int_{S_j} d\mathbf{x}\, d\mathbf{y}\, \mu(\mathbf{x}) p(\mathbf{x}, \mathbf{y};\, \tau) && (8.18) \\ &= \frac{\pi_i}{\pi_i} \int_{S_i} d\mathbf{x}\, d\mathbf{y}\, \mu(\mathbf{x}) p(\mathbf{x}, S_j;\, \tau) && (8.19) \\ &= \pi_i T_{ij} && (8.20) \\ &= c_{ij} \\ &= \mathbf{\Pi T} \end{aligned}$$

and using the Roothan-Hall method thus results in:

$$\begin{aligned} \mathbf{\Pi}^{-1}\mathbf{\Pi T B} &= \mathbf{B\Lambda}. \\ \mathbf{TB} &= \mathbf{B\Lambda}. \\ \mathbf{Tb}_i &= \mathbf{b}_i \lambda_i. \end{aligned}$$

and is therefore associated with the right eigenvector problem of a Markov state model:

$$\mathbf{Tr}_i = \mathbf{r}_i \lambda_i.$$

Therefore, using a Markov model is identical with using the Roothan-Hall method with the choice $\chi_i = \mathbf{1}_{S_i}$, i.e. using step functions as a Basis set. In other words, the Markov model eigenvectors provide an optimal linear combination of step functions to the true eigenfunctions.

8.6 Results

Approximation of κ_2

The autocorrelation function of $\tilde{r}_2$, which we denote as $\langle \tilde{r}_2(0)\tilde{r}_2(\tau)\rangle = \langle \tilde{r}_2(\mathbf{x}_t)\tilde{r}_2(\mathbf{x}_{t+\tau})\rangle$, evaluates to

$$\begin{aligned} \tilde{\lambda}_2(\tau) &= \langle \tilde{r}_2(0)\tilde{r}_2(\tau)\rangle \\ &= \alpha e^{-\kappa_2\tau} + \sum_{i>2} \langle r_i, \tilde{r}_2\rangle_\mu^2 e^{-\kappa_i\tau} \end{aligned} \tag{8.21}$$

where

$$\alpha = \langle r_2, \tilde{r}_2\rangle_\mu^2.$$

This autocorrelation function does not yield the exact eigenvalue $\lambda_2(\tau)$, but some approximation $\tilde{\lambda}_2(\tau)$. For $\tau \gg \kappa_3^{-1}$, which can readily be achieved for clear two-state processes where $\kappa_2 \gg \kappa_3$, the sum on the right hand side disappears:

$$\tilde{\lambda}_2(\tau) \approx \alpha e^{-\kappa_2\tau}. \tag{8.22}$$

This suggests that the true rate, as well as a measure of reaction coordinate quality, could be recovered from large tau decay of an appropriately good trial function even from the projected process.

Single-τ rate estimators: A simple rate estimator is to directly take value of the autocorrelation function of some function $\tilde{\psi}_2$ at a single value of τ, and transform it into a rate estimate by virtue of Eq. (8.22). We call these estimators *single-τ estimators*. Ignoring statistical uncertainties, they yield a rate estimate of the form

$$\tilde{\kappa}_{2,\text{single}} = -\frac{\ln \tilde{\lambda}_2(\tau)}{\tau} \tag{8.23}$$

Quantitatively, the error can be bounded by the expression:

$$\tilde{\kappa}_{2,\text{single}} - \kappa_2 \leq -\frac{\ln \alpha}{\tau}. \tag{8.24}$$

Proof:

$$\begin{aligned}
\hat{\kappa}_2 &= -\tau^{-1} \ln \tilde{\lambda}_2(\tau) \\
&= -\tau^{-1} \ln \left(\alpha \lambda_2(\tau) + \sum_{i>2} a_i^2 \lambda_i(\tau) \right) \\
&= -\tau^{-1} \ln \left(\alpha \mathrm{e}^{-\tau\kappa_2} + \sum_{i>2} a_i^2 \mathrm{e}^{-\tau\kappa_i} \right) \\
&= -\tau^{-1} \ln \left(\mathrm{e}^{-\tau\kappa_2} \left[\alpha + \sum_{i>2} a_i^2 \mathrm{e}^{-\tau(\kappa_i - \kappa_2)} \right] \right) \\
&= -\tau^{-1} \left(\ln \mathrm{e}^{-\tau\kappa_2} + \ln \left[\alpha + \sum_{i>2} a_i^2 \mathrm{e}^{-\tau(\kappa_i - \kappa_2)} \right] \right) \qquad (8.25)
\end{aligned}$$

which leads to the systematic error in the rate $\hat{\kappa}_2$:

$$\Delta\kappa_{2,\tau} = \tilde{\kappa}_{2,\tau} - \kappa_2 = -\tau^{-1} \ln \left[\alpha + \sum_{i>2} a_i^2 \mathrm{e}^{-\tau(\kappa_i - \kappa_2)} \right] \qquad (8.26)$$

Please note that the expression in the logarithm is smaller than unity, such that the rate $\hat{\kappa}_2$ is always overestimated. We can continue to simplify to

$$\begin{aligned}
\Delta\kappa_{2,\tau} &= -\tau^{-1} \ln \left(\alpha (1 + \sum_{i>2} \frac{a_i^2}{\alpha} \mathrm{e}^{-\tau(\kappa_i - \kappa_2)}) \right) \\
&= \tau^{-1} \ln \frac{1}{\alpha} - \tau^{-1} \ln \left(1 + \sum_{i>2} \frac{a_i^2}{\alpha} \mathrm{e}^{-\tau(\kappa_i - \kappa_2)} \right) \qquad (8.27)
\end{aligned}$$

as an expression for the estimation error. This error can then be bounded using $0 \leq \ln(1+x)$ for $x \geq 0$ by

$$0 \quad \leq \Delta\kappa_{2,\tau} \leq \quad \tau^{-1} \ln \frac{1}{\alpha} \qquad (8.28)$$

and since $\kappa_i > \kappa_2$ is true for $i > 2$ we can also find a lower bound on the error that only depends on the spectral gap $\kappa_3 - \kappa_2$ and the RCQ α

$$0 \leq \tau^{-1} \ln \frac{1}{\alpha} - \tau^{-1} \ln \left(1 + \frac{1-\alpha}{\alpha} \mathrm{e}^{-\tau(\kappa_3 - \kappa_2)} \right) \quad \leq \Delta\kappa_{2,\tau} \leq \quad \tau^{-1} \ln \frac{1}{\alpha} \qquad (8.29)$$

where we used that $\ln(1+x) < x$ always holds. We conclude that the estimation error $\Delta\kappa_2$ is dominated by a $1/\tau$ dependence whereas the width of this

error bound decreases exponentially in the spectral gap. For a two-state system with with a large gap $\kappa_2 \gg \kappa_3$ this uncertainty vanishes and the rate error is indeed approximated by:

$$\Delta\kappa_{2,\tau} \approx \tau^{-1} \ln \frac{1}{\alpha}. \tag{8.30}$$

The systematic error of single-τ estimators results from the fact that Eq. (8.23) effectively attempts to fit the tail of a multiexponential decay $\tilde{\lambda}_2(\tau)$ by a single-exponential with the constraint $\tilde{\lambda}_2(0) = 1$. Unfortunately, the ability to improve these estimators by simply increasing τ is limited because the statistical uncertainty of estimating Eq. 8.22 quickly grows in τ [**?**].

Chapter 9

Stochastic vs. Transport Equations

9.1 Stochastic Differential Equations (SDE)

We want to integrate a stochastic differential equation of the following form

$$\frac{dX_t}{dt} = b\left(t, X_t\right) + \sigma\left(t, X_t\right) R_t$$

with b and σ being continuous functions and R_t a random variable representing noise. The time integrated process is again a random variable.

9.1.1 Terminology

A *probability measure space* $(\Omega, \mathcal{A}, P)$ is a triple of

- Ω the set of results
- $\mathcal{A} \in \mathcal{P}(\Omega)$ a σ-Algebra the set of measurable events, which we can associate with a probability
- $P : \mathcal{A} \to [0, 1] \subset \mathbb{R}_0^+$ a measure on Ω, which is normalized $P(\Omega) = 1$ and $P(A)$ the probability that event $A \in \mathcal{A}$ occurs.

A *random variable* is a function $X : \Omega \to \mathbb{R}^n$, which is measurable with respect to the measure spaces $(\Omega, \mathcal{A}, P)$ and $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

We can use an abbreviated form and write for a random variable $f : \Omega \to \mathbb{R}$

$$\langle f \rangle := \int f \, dP = \int_\Omega f(\omega) \, dP(\omega)$$

and in analogy

$$\langle f \, g \rangle := \int f \, g \, dP = \int_\Omega f(\omega) \, g(\omega) dP(\omega)$$

The expectation is then defined as

$$\mathbb{E}(f) := \langle f \rangle$$

All Moments $\langle f^n \rangle$ can be defined in a similar manner.

Each random variable $X : \Omega \to \mathbb{R}^n$ has an induced measure μ_X defined by

$$\mu_X(B) := P(X^{-1}(B)) = P(\{\omega \in \Omega \mid X(\omega) \in B\}), \qquad \forall B \in \mathcal{B}(\mathbb{R}^n)$$

and is thus equals the probability of the inverse image set $X^{-1}(B)$.

Using the induced measure we can use the random variable in the *usual sense* with the probability measure space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu_X)$ and consider $\mu_X(B)$ to be the probability to generate a random number $x \in B \subset \mathcal{B}(\mathbb{R}^n)$.

If we can represent μ_X by a a density function $g : \mathbb{R}^n \to \mathbb{R}$ with

$$\mu_X(B) = P(X^{-1}(B)) = \int_B g(x) dx$$

and we can rewrite the integration with measure μ_X by an integral in $\mathbb{R}^n$.

In analogy a function $f : \mathbb{R}^n \to \mathbb{R}^m$, which is measurable with respect to $(\mathcal{B}(\mathbb{R}^n), \mathcal{B}(\mathbb{R}^m))$, can then be integrated with

$$\langle f(X) \rangle = \int_\Omega f(X(\omega)) dP(\omega) = \int_{\mathbb{R}^n} f(x) g(x) dx$$

9.1.2 Stochastic Processes

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ the measure space with the Borel-Algebra of $\mathbb{R}^n$. If $I \subseteq [0, \infty)$ an index set, then the family $(X_t)_{t \in I}$ consisting of measurable functions $X_t : \Omega \to \mathbb{R}^n, \ t \in I$ is a stochastic process (with state space $\mathbb{R}^n$).

For fixed t the function X_t is a n-dimensional random variable.

For fixed ω the function $(X_t)_{t \in I}(\omega) : I \to \mathbb{R}^n$ is one path of $(X_t)_{t \in I}$ and one can consider ω as a particle and the path its trajectory.

Finally one can consider the function $(X_t)_{t \in I} : (t, \omega) \to \mathbb{R}^n$ and ask for measurability.

9.1.3 Further Reading

- Introduction to SDE[2] (in German)

9.2 Master-Equation to Fokker-Planck

$$\frac{\partial p(x,t)}{\partial t} = \int \left(K(y,x)p(y,t) - K(x,y)p(x,t)\right) dy$$

We express $K(x,y)$ by

$$\hat{K}\left(y, r(x,y)\right) = K(x,y), \qquad r(x,y) = x - y$$

so that we have

$$\hat{K}\left(y,r\right) = K\left(y+r,y\right)$$

This leads to

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= \int \left(\hat{K}(y,y-x)p(y,t) - \hat{K}(x,x-y)p(x,t)\right) dr \\ &= \int \left(\hat{K}(x-r,r)p(x-r,t) - \hat{K}(x,-r)p(x,t)\right) dr\end{aligned}$$

Now we assume that

1. $\hat{K}\left(x,y\right)$ is of short range $\hat{K}\left(y,r\right) \approx 0, \quad |y| > \delta$
2. $\hat{K}\left(x,y\right)$ varies slowly $\hat{K}\left(y+\Delta y,r\right) \approx \hat{K}\left(y,r\right), \quad |\Delta y| < \delta$
3. $p(x,t)$ varies slowly in x

and do a Taylor expansion of the first integral in $x + (-r)$ around x

$$\begin{aligned}\hat{K}(x-r,r)p(x-r,t) &\approx \hat{K}(x,r)p(x,t) + \frac{(-r)^1}{1!}\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right) + \frac{(-r)^2}{2!}\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right) \\ &\approx \hat{K}(x,r)p(x,t) - r\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right) + \frac{1}{2}r^2\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right)\end{aligned}$$

Place this into the equation above and get

$$\frac{\partial p(x,t)}{\partial t} = \int \left(\hat{K}(x,r)p(x,t) - r\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right) + \frac{1}{2}r^2\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right)\right) - \hat{K}(x,-r)p(x,t)dr$$

$$= \int \hat{K}(x,r)p(x,t)dr - \int r\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right)dr + \frac{1}{2}\int r^2\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right)dr - \int \hat{K}(x,-r)p(x,t)dr$$

$$= p(x,t)\int \hat{K}(x,r)dr - \int r\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right)dr + \frac{1}{2}\int r^2\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right)dr - p(x,t)\int \hat{K}(x,-r)dr$$

The first and last integral are the same, because the last integral can be computed for $r \to -r$ and maintains its value

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\int r\frac{\partial}{\partial x}\left(\hat{K}(x,r)p(x,t)\right)dr + \frac{1}{2}\int r^2\frac{\partial^2}{\partial x^2}\left(\hat{K}(x,r)p(x,t)\right)dr \\ &= -\frac{\partial}{\partial x}\left(p(x,t)\int r\left(\hat{K}(x,r)\right)dr\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(p(x,t)\int r^2\left(\hat{K}(x,r)\right)dr\right)\end{aligned}$$

Now we can introduce the so called jump moments

$$a_n(x) = \int r^n\left(\hat{K}(x,r)\right)dr$$

and finally write the Fokker-Planck equation as an approximation of the master equation

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(a_1(x)p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(a_2(x)p(x,t)\right).$$

With the jump moments we also have a way to compute the the necessary drift and diffusion functions in terms of the rate kernel function $K(x,y)$.

The exact approximation with all terms of the Taylor expansion is known as Kramer-Moyer-Expansion. It is equivalent to the Master-Equation.

$$\frac{\partial p(x,t)}{\partial t} = \sum_{n=1}^{\infty}\frac{(-1)^n}{n!}\frac{\partial^n}{\partial x^n}\left(a_n(x)p(x,t)\right)$$

9.3 Stochastic Integrals

The SDE above

$$\frac{dX_t}{dt} = u\left(t, X_t\right) + v\left(t, X_t\right) R_t$$

is discretized for time steps t_j

$$\frac{X_{j+1} - X_j}{t_{j+1} - t_j} = u(t_j, X_j) + v(t_j, X_j) R_j$$

and thus

$$X_{j+1} = X_j + u(t_j, X_j)\Delta t + v(t_j, X_j) R_j \Delta t$$

Since the R_j cannot be independent due to non continuous paths we have to replace $R_j \Delta t$ by increments $\Delta W_j = W_{t_{j+1}} - W_{t_j}$. And we want

these to be

1. independent for a finite number of time steps, so that $\Delta W_{j-1}, \ldots, \Delta W_1$ are independent increment,s
2. stationary, so independent of movement in time $t \to t + \tau$ and the distribution of $W_t - W_s$ depends only on $t - s$
3. and $\mathbb{E}\left(W_t - W_s\right) = 0$

This is only fulfilled for the Wiener process, which is a stochastic process $W_t : \Omega \to \mathbb{R}$ with

1. $W_0 = 0$ (almost surely)
2. $(W_t)_{t \geq 0}$ has independent increments
3. $W_t - W_s \sim \mathcal{N}(0, t - s),\ \forall 0 \leq s < t$

Then we compute

$$\mathbb{E}(W_t) = \mathbb{E}(W_t - W_0) + \mathbb{E}(W_0) = 0$$

and

$$\mathbb{E}((W_t - W_s)^2) = Var(W_t - W_s) + \mathbb{E}((W_t - W_s))^2 = Var(W_t - W_s) = t - s$$

$$\mathbb{E}\left(W_t \cdot W_s\right) = \min\left\{s, t\right\}$$

With the Extension-theorem of Kolmogorov we can prove, that such a process exists and the existence of a process, that is almost surely continuous follows from the Continuity-theorem of Kolmogorov.

9.3.1 Ito-Integral

We derive the definition for the Stochastic Integral only for a set of elementary function. It can then be shown, that for most stochastic process exists a suitable series of elementary function that converge to it and the Ito-Integral can be defined as the limit of the integral for the series of elementary functions.

The elementary functions have the form of time-wise constant random variable $e_j(\omega)$ and points $t_j = j \cdot 2^{-n}$

$$\phi(t, \omega) = \sum_{j \geq 0} e_j(\omega) \cdot \chi_{[t_j, t_{j+1})}(t)$$

for fixed spacing $n \in \mathbb{N}$. And we define a stochastic integral for these functions by

$$\int_a^b \phi(t, \omega) dW_t(\omega) = \sum_{j \geq 0} e_j(\omega) \left(W_{t_{j+1}} - W_{t_j}\right)$$

and assume, that a and b are in accordance with the time grid induced by n. Otherwise the points where W_t is evaluated have to be adapted at the end-points accordingly.

Choose for example

$$e_j(\omega) := W_{(1-\tau)t_j + \tau t_{j+1}}(\omega)$$

and we compute the integral to

$$I_\tau = \sum_{j \geq 0} W_{(1-\tau)t_j + \tau t_{j+1}} \left(W_{t_{j+1}} - W_{t_j}\right)$$

with expectation

$$\mathbb{E}(I_\tau) = \tau(b - a)$$

Strangely, the expectation of the integral depends on the position of evaluation. We define the stochastic integral by

$$I_\tau[f](\omega) = \sum_{j\geq 0} f((1-\tau)t_j + \tau t_{j+1}, \omega)\left(W_{t_{j+1}}(\omega) - W_{t_j}(\omega)\right)$$

and distinguish the two cases

- $\tau = 0$: The left-side of the intervals are used and the integral is called Ito-Integral. The corresponds to the case where only information from the past and present is used to compute information a specific time and so is the resulting integrated stochastic process with Expectation zero. This is the mathematical more consistent way to construct a new stochastic process by integration.
- $0 < \tau \leq 1$: Now the expectation is not vanishing and the integral uses "future" information. For $\tau = \frac{1}{2}$ it is called the Stratonovich Integral has simpler transformation rules and a more reasonable physical interpretation:
  Each path in an SDE that is almost continuous can be approximated by smooth functions, which can be inserted into the SDE and then solved as a deterministic DEQ. The solution converges to a random process in t and can be identified with the Stratonovich integral. Thus it is the correct description for random forces in SDE and not necessarily randomness with other origin, that is mimicked by a random process.

Both integrals are linear and can be split as ordinary integrals can.

For Ito-Integral there exists a transformation rule. If a SDE is given by

$$dX_t = u\,dt + v\,dW_t$$

and there is a function $Y_t := g(t, X_t)$ with

$$g : [0, \infty) \times \mathbb{R} \to \mathbb{R}$$

then

$$dY_t = \left(\frac{\partial g}{\partial t} + u\frac{\partial g}{\partial x} + \frac{1}{2}v^2\frac{\partial^2 g}{\partial x^2}\right)dt + v\frac{\partial g}{\partial x}dW_t$$

is also a stochastic process. From this transformation law follows the Fokker-Planck-Equation. This means, that the Fokker-Planck Equation can be regarded as an Ito-Process of the above form.

9.3.2 Stratonovich Integral

For the Stratonovich Integral the solution is similar (denoted by $\circ$)

$$dX_t = u\,dt + v \circ dW_t$$

is transformed by

$$dY_t = \left(\frac{\partial g}{\partial t} + u\frac{\partial g}{\partial x}\right) dt + v\frac{\partial g}{\partial x} \circ dW_t$$

The above Stratonovich SDE will not transform into the Fokker-Planck-Equation, but there exists also a transformation between both integration ways, which then leads to an SDE, that is equivalent to the Fokker-Planck Equation. If $h : \mathbb{R} \to \mathbb{R}$ is a smooth function, then

$$h(W_t) \circ dW_t = \frac{1}{2}h(W_t)\frac{\partial h}{\partial x}(W_t)dt + h(W_t)dW_t$$

holds. This leads to the following relations between the different transport equations and SDEs

$$\begin{array}{cc}
P\left(j, t_{n+1} \mid i, t_{n-1}, x_{n-2}; t_{n-2}; \ldots\right) = P\left(j, t_{n+1} \mid i, t_{n-1}\right) & \text{(Markov Process)} \\
\downarrow & \downarrow \\
p_i^{(t+1)} = \sum_j p_j^{(t)} P(j \to i) & \text{(Chapman-Kolmogorov)} \\
\downarrow & \downarrow \\
\frac{\partial p(x,t)}{\partial t} = \int \left(K(y,x)p(y,t) - K(x,y)p(x,t)\right) dy & \text{(Master-Equation)} \\
\downarrow & \downarrow \\
\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(u(x,t)p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(v(x,t)^2 p(x,t)\right) & \text{(Fokker-Planck)} \\
\Updownarrow & \Updownarrow \\
dX_t = u\,dt + v\,dW_t & \text{(Ito Calculus)} \\
\Updownarrow & \Updownarrow \\
dX_t = \left(u - \frac{1}{2}v\frac{\partial}{\partial x}v\right) dt + v \circ dW_t & \text{(Stratonovich Calculus)}
\end{array}$$

Chapter 10

Discretization

When working with state-continuous Markov processes in the computer, the state space must be discretized. This is effectively transforming the state-continuous Markov process into a state-discrete Markov process, i.e. a Markov chain, thus allowing all Markov chain analysis and estimation tools above to be applied. Importantly, it must be ensured that this Markov chain is still a good approximation of the original process, i.e. that the discretization error is small. This is in detail worked out in the paper

M. Sarich, F. Noé and C. Schütte: "On the approximation error of Markov state models". Multiscale Model. Simul. (2010)

Bibliography

[1] Nina Singhal Hinrichs and Vijay S Pande. Calculation of the distribution of eigenvalues and eigenvectors in markovian state models for molecular dynamics. *The Journal of Chemical Physics*, 126(24):244101, Jan 2007.

[2] Roland Pulch. Stochastische differentialgleichungen: Theorie und numerik. *Lec. Notes*, pages 1–115, Jan 2006.

[3] Nina Singhal and Vijay S Pande. Error analysis and efficient sampling in markovian state models for molecular dynamics. *J. Chem. Phys*, 123(20):204909, Nov 2005.

[4] William C Swope, Jed W Pitera, and Frank Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *J. Phys. Chem. B*, 108:6571–6581, Jan 2004.

[5] N G van Kampen. *Stochastic Processes in Physics and Chemistry*. Number Third Edition. 2006.