Numerics for Bioinformaticians, Semester 1
Lecture 15

Vikram Sunkara, Max von Kleist

February 14, 2017

Chapter 1

Lecture 1

1.1 Random Variable

Definition 1. *The state space is the set of all possible outcomes a random variable can take. We denote it by Ω (Greek: capital omega). The state space can be discrete or continuous. Discrete in the sense that the set is countable; and continuous implying that the set is uncountable.*

Example 1. *In a coin toss, the state space is heads or tails. Then $\Omega = \{H, T\}$. If you were to toss two coins, then the state space would be, $\Omega = \{HH, HT, TH, TT\}$.*

Definition 2. *A probability distribution is a function which maps elements of Ω to $[0, 1]$. We denote it by p (lowercase p). It has the property that*

$$\int_{\Omega} p(x)dx = 1,$$

for a continuous state space, and

$$\sum_{x \in \Omega} p(x) = 1,$$

for a discrete state space.

Definition 3. *Poisson Distribution: The state space $\Omega = \mathbb{N}_0$. For $x \in \Omega$,*

$$p(x) := \frac{\lambda^x e^{-\lambda}}{x!},$$

with $\lambda > 0$ being a free parameter.

Definition 4. *Exponential Distribution: The state space $\Omega = (0, \infty)$. For $x \in \Omega$,*

$$p(x) := \lambda e^{-\lambda x},$$

with $\lambda > 0$ being a free parameter.

Definition 5. *Gaussian Distribution: The state space $\Omega = \mathbb{R}$. For $x \in \Omega$,*

$$p(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are free parameters.

Definition 6. *A random variable is a selected state from the state space, where the probability of the state occurring is given by an underlying probability distribution. For a given state space Ω and probability distribution p, we say a random variable X is distributed according to p. We denote this as $X \sim p$.*

Definition 7. *Let $X \sim p$ over a state space Ω. We define the cumulative distribution of $x \in \Omega$ to be*

$$P(x) = p(X \leq x).$$

If $\Omega = \mathbb{N}$, then

$$P(x) = \sum_{n=1}^{x} p(x).$$

Similarly if $\Omega = \mathbb{R}$, then

$$P(x) = \int_{-\infty}^{x} p(\omega) d\omega.$$

We denote the cumulative distribution by P (capital p).

Exercise 1. *Exponential Distribution: The state space $\Omega = (0, \infty)$. For $x \in \Omega$,*

$$P(x) = \int_0^x p(y) dy = 1 - e^{-\lambda x},$$

with $\lambda > 0$ being a free parameter.

Definition 8. *Let $X \sim p$ over the state space Ω. We define the expectation of X, which we denote as $\mathbb{E}[X]$, to be*

$$\mathbb{E}[X] := \sum_{x \in \Omega} x p(x),$$

for a discrete state space and

$$\mathbb{E}[X] := \int_{\Omega} x p(x) dx,$$

for a continuous state space.

Example 2. *Poisson Distribution: The state space $\Omega = \mathbb{N}_0$. Let X be distributed according to a Poisson distribution for a parameter $\lambda > 0$. Then*

$$\mathbb{E}[X] = \lambda.$$

Exercise 2. *Gaussian Distribution: The state space $\Omega = \mathbb{R}$. Let X be distributed according to a Gaussian distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Then*

$$\mathbb{E}[X] = \mu.$$

Definition 9. *Let $X \sim p$ over the state space Ω. We define the variance of X, which we denote as $\mathbb{V}[X]$, to be*

$$\mathbb{V}[X] := \sum_{x \in \Omega} (x - \mathbb{E}[X])^2 p(x),$$

for a discrete state space and

$$\mathbb{V}[X] := \int_{\Omega} (x - \mathbb{E}[X])^2 p(x) dx,$$

for a continuous state space.

Exercise 3. *Poisson Distribution: The state space $\Omega = \mathbb{N}_0$. Let X be distributed according to a Poisson distribution for a parameter $\lambda > 0$. Then*

$$\mathbb{V}[X] = \lambda.$$

Example 3. *Gaussian Distribution: The state space $\Omega = \mathbb{R}$. Let X be distributed according to a Gaussian distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$. Then*

$$\mathbb{V}[X] = \sigma^2.$$

Definition 10. *Let X and Y be two random variables defined over Ω_X, Ω_Y, respectively. We define the covariance of X and Y, which we denote as $\mathbb{C}[X, Y]$, to be*

$$\mathbb{C}[X, Y] := \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} (x - \mathbb{E}[X])(y - \mathbb{E}[Y]) p(x, y),$$

for a discrete state space and

$$\mathbb{C}[X, Y] := \int_{\Omega_X} \int_{\Omega_Y} (x - \mathbb{E}[X])(y - \mathbb{E}[Y]) p(x, y) dx dy,$$

for a continuous state space. The term $p(x, y)$ is the joint probability distribution. It states the probability of x and y occurring.

1.2 Sampling

We draw random variables from p by first drawing from a uniform random number, τ between $[0, 1)$. Then our sample from p is the state $x \in \Omega$ such that

$$P(x) = \tau.$$

In practice, multiple samples of a distribution are generated, we denote them with a superscript index. For example, if we draw 10 samples from a distribution p, we denote it $x^{(i)} \sim p$, for $i = 1, \ldots, 10$.

Exercise 4. *Sampling the Exponential Distribution: For a given $\lambda > 0$ the exponential distribution over $\Omega = [0, \infty)$ is given by*

$$p(x) := \lambda e^{-\lambda x}.$$

The cumulative distribution is given by

$$\begin{aligned} P(x) &:= \int_0^x p(y)dy, \\ &= \int_0^x \lambda e^{-\lambda x}, \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Let $\tau \sim U(0, 1)$. Then to sample the exponential distribution we let

$$\tau = 1 - e^{-\lambda x},$$

and solve for x. This then reduces to

$$x = \frac{1}{\lambda} log\left(\frac{1}{\tau}\right).$$

In many cases, the probability distribution itself cannot be calculated, hence, we design methods which decompose the random variables into smaller parts which can be sampled.

1.3 Processes

A simple definition of a stochastic process is that it is a random variable distributed according to a probability distribution which is evolving through time.

Definition 11. *Let Ω be the state space. Let T be a finite subset of $[0, \infty)$. If we have a family of probability distributions indexed by time $\{p_t : t \in T\}$. Then the collection of all random variables $X_t \sim p_t$ is called a stochastic process.*

Definition 12. *We define $\mathcal{P}(\lambda t)$ to be the Poisson Process at time $t > 0$ with rate $\lambda > 0$. That is, for $x \in \Omega$,*

$$p(\mathcal{P}(\lambda t) = x; t) := \frac{(\lambda t)^x e^{-\lambda t}}{x!}.$$

Remark 1. *We denote the probability of observing a stochastic process at a state x at a particular time t by $p(x; t)$.*

Lemma 1. *For $t > 0$ and $\lambda > 0$. The Poisson Process has the following properties*

1. $\mathbb{E}[\mathcal{P}(\lambda t)] = \lambda t$,
2. $p(\mathcal{P}(\lambda t) = 1; t) = \lambda t e^{-\lambda t}$,
3. $p(\mathcal{P}(\lambda t) > 0; t) = 1 - e^{-\lambda t}$.

To compute a trajectory of this process we need to sample twice. First for the time at which an event occurs, then second to pick which event occurs. **When** and **What**.

Exercise 5. *Birth Process: Let $\Omega = \mathbb{N}_0$.*

$$R_1 : \emptyset \xrightarrow{0.5} A.$$

$$X_t = X_0 + \mathcal{P}(0.5t). \tag{1.1}$$

Note that we do not yet know how to sample from a Poisson Distribution. However, we do know a single event occurring is exponentially distributed according to Lemma 1. Hence, we keep sampling the next event to occur until we reach our time point. For example:
Let $X_0 = 10$. We want a sample for, X_{10}, time equal to 10.

- *Step 1: Sample $\tau \in U(0, 1)$. Then solve for $t^{(1)}$ such that $\tau = 1 - e^{-\lambda t^{(1)}}$.*
- *Step 2: Now you have that at $X_{t^{(1)}} = 10 + 1 = 11$. (Because you have just sampled for the time of one event occurring).*
- *Step 3: If $t^{(1)} \neq 10$. We need to sample again. Repeat Step 1 to compute t_2.*
- *Step 4: Now you have that at $X_{t^{(1)}+t^{(2)}} = 11 + 1$. (Because you have just sampled for the time of one event occurring).*
- *Step 5: If $t^{(1)} + t^{(2)} \neq 10$. We need to sample again. Repeat Step 1 to compute $t^{(3)}$.*
- *Repeat until you reach your final time point or pass it.*

The state you are in when time crosses 10 is a sample of the pull process X_{10} (2.1).

Chapter 2

Lecture 2

2.1 Recap

Lemma 2. *For $t > 0$ and $\lambda > 0$. The Poisson Process has the following properties*

1. $\mathbb{E}[\mathcal{P}(\lambda t)] = \lambda t$,
2. $p(\mathcal{P}(\lambda t) = 1; t) = \lambda t e^{-\lambda t}$,
3. $p(\mathcal{P}(\lambda t) > 0; t) = 1 - e^{-\lambda t}$.

To compute a trajectory of this process we need to sample twice. First for the time at which an event occurs, then second to pick which event occurs. **When** and **What**.

Exercise 6. *Birth Process: Let $\Omega = \mathbb{N}_0$.*

$$R_1 : \emptyset \xrightarrow{0.5} A.$$

$$X_t = X_0 + \mathcal{P}\left(0.5t\right). \tag{2.1}$$

$X_{t^{(1)}} = X_0 + 1$, $X_{t^1 + t^{(2)}} = X_{t^{(1)}} + 1$... so on and so forth. Where $t^{(1)}, t^{(2)} \sim \lambda e^{-\lambda t}$, exponentially distributed.

Example 4. *Death Process. Let $\Omega = \{0, \ldots, X_0\}$.*

$$R_1 : A \xrightarrow{\beta} \emptyset.$$

$$X_t = X_0 - \mathcal{P}\left(\int_0^t \beta X_s ds\right).$$

2.2 Multiple Reactions

Theorem 1. *The sum of two independent Poisson processes is a Poisson process,*

$$\mathcal{P}(\lambda_1 t) + \mathcal{P}(\lambda_2 t) = \mathcal{P}((\lambda_1 + \lambda_2)t),$$

with $\lambda_1, \lambda_2 > 0$.

Proof. The probability distribution of the sum of two independent random variables is a convolution of the individual probability distributions. Let

$$Z_t = \mathcal{P}(\lambda_1 t) + \mathcal{P}(\lambda_2 t),$$

For the probability distribution of Z_t we convolute the distributions of $\mathcal{P}(\lambda_1 t)$ and $\mathcal{P}(\lambda_2 t)$;

$$p(Z_t = z) = \sum_{x=0}^{z} p(\mathcal{P}(\lambda_1 t) = x) p(\mathcal{P}(\lambda_2 t) = z - x).$$

Then it follows,

$$\begin{aligned} p(Z_t = z) &= \sum_{x=0}^{z} \frac{(\lambda_1 t)^x e^{-\lambda_1 t}}{x!} \frac{(\lambda_2 t)^{z-x} e^{-\lambda_2 t}}{(z-x)!}, \\ &= \frac{e^{-(\lambda_1 t + \lambda_2 t)}}{z!} \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} (\lambda_1 t)^x (\lambda_2 t)^{z-x}, \\ &= \frac{(\lambda_1 t + \lambda_2 t)^z e^{-(\lambda_1 t + \lambda_2 t)}}{z!}, \end{aligned}$$

using the Binomial formula. Hence, Z_t has the probability distribution of $\mathcal{P}((\lambda_1 + \lambda_2)t)$ □

Example 5. *Birth–Death Process. Let* $\Omega = \mathbb{N}_0$.

$$R_1 : \emptyset \xrightarrow{\alpha} A, \qquad R_2 : A \xrightarrow{\beta} \emptyset.$$

$$X_t = X_0 + \mathcal{P}(\alpha t) - \mathcal{P}\left(\int_0^t \beta X_s ds\right).$$

In this example, we have two reactions changing the system. We can use Theorem 1 to work out the time at which an event occurs. Then we need to choose which event occurred. In the case above, a birth or a death reaction could have occurred. Let us consider a single step:

When

Let

$$N_t := \mathcal{P}(\alpha t) + \mathcal{P}\left(\int_0^t \beta X_s ds\right),$$

be the process which counts how many events occurred by time t. Using Theorem 1, we can deduce that,

$$N_t = \mathcal{P}\left(\alpha t + \int_0^t \beta X_s ds\right).$$

Now we wish to sample for a time at which one reaction occurs. That is $t^{(1)}$ such that $N_{t^{(1)}} = 1$. Since N_t is a Poisson Process, Lemma 1 tells us that the first firing time is exponentially distributed. In our case,

$$\lambda t = \alpha t + \int_0^t \beta X_s ds,$$

where λ is the rate for the process N_t fires. Since we are trying to compute the first time at which an event occurs, the process $X_\bullet$ has not changed, it is still X_0. Hence, substituting this into the integral gives us

$$\lambda t = \alpha t + \beta X_0 t.$$

Then $\lambda = \alpha + \beta X_0$. Using Exercise 4, a sample for the time of a reaction firing is

$$t^{(1)} = \frac{1}{\alpha + \beta X_0} log\left(\frac{1}{\tau}\right),$$

where $\tau \sim U(0,1)$.

What

We now know that at time $t^{(1)}$ an event will occur. However, we do not know which one. Since X_t is being changed by two random processes, we do not know whether the reaction that occurred was

$$\mathcal{P}(\alpha t) \text{ or } \mathcal{P}\left(\int_0^t \beta X_s ds\right).$$

Using Lemma 1,

$$\begin{aligned} p(R_1 \text{ firing }; t^{(1)}) &:= \alpha t^{(1)} e^{-\alpha t^{(1)}}, \\ &\approx \alpha t^{(1)}. \end{aligned}$$

Similarly,

$$\begin{aligned} p(R_2 \text{ firing } ; t^{(1)}) &:= \left(\int_0^{t^{(1)}} \beta X_s ds \right) e^{-\int_0^{t^{(1)}} \beta X_s ds}, \\ &\approx \beta X_0 t^{(1)}. \end{aligned}$$

Then we have the following probability distribution:

Reaction	R_1	R_2
$p(\text{ reaction } R_\bullet)$	$\frac{\alpha t^{(1)}}{\alpha t^{(1)} + \beta X_0 t^{(1)}}$	$\frac{\beta X_0 t^{(1)}}{\alpha t^{(1)} + \beta X_0 t^{(1)}}$
	$\frac{\alpha}{\alpha + \beta X_0}$	$\frac{\beta X_0}{\alpha + \beta X_0}$

Now we simply sample from this distribution to find out which reaction fired. Let $\tau \sim U(0,1)$,

$$r^{(1)} = \begin{cases} 1, & \text{if } 0 < \tau \leq \frac{\alpha}{\alpha + \beta X_0} \\ 2, & \text{if } \frac{\alpha}{\alpha + \beta X_0} < \tau \leq 1. \end{cases}$$

Putting them together, a single step would look like $X_{t^{(1)}} = X_0 + S_{r^{(1)}}$. We can then progressively keep evolving the process forward until we compute a sample for our required time step.

Remark 2. *Note that the order of the reactions does not matter. Since we are drawing from a uniform distribution to choose which reaction occurs, the only thing to keep in mind is that the labels of the stoichiometry and the propensities are consistent.*

2.3 Stochastic Simulation Algorithm (SSA or Gillespie)

Table 2.1: SSA parameters

x_0	Initial population configuration of the system.
t_0	Starting time step.
t_{final}	Final time step for the system.
N_r	Number of reactions.
N_s	Number of species.
a_i	$i = 1, \ldots, N_r$ the propensity functions.
$\mathbf{S}_i$	$i = 1, \ldots, N_r$ the stoichiometric vectors.
x	The realised state at time t_{final}.

Algorithm 1: Stochastic Simulation Algorithm

input : $x_0, t_0, t_{final}, a_1, \ldots, a_{N_r}, \mathbf{S}_1, \ldots, \mathbf{S}_{N_r}$.
output: x

begin
  $x \leftarrow x_0$
  $t \leftarrow t_0$
  while $t < t_{final}$ **do**
    $a_0 \leftarrow \sum_{i=1}^{N_r} a_i(x)$
    if $a_0 == 0$ **then**
      $t \leftarrow t_{final}$
      `break`
    end
    $\tau_1, \tau_2 \leftarrow$ `uniform` $(0, 1)$
    $\triangle t \leftarrow \frac{1}{a_0} log(\frac{1}{\tau_1})$
    if $t + \triangle t > t_{final}$ **then**
      $t \leftarrow t_f$
      `break`
    end
    $t \leftarrow t + \triangle t$
    choose j such that $\sum_{i=1}^{j-1} a_i(x) < \alpha_0 \tau_2 \leq \sum_{i=j}^{N_s} a_i(x)$.
    $x \leftarrow x + \mathbf{S}_j$
  end
  `return` x
end

Chapter 3

Lecture 3

3.1 Recap. I.

When we use the stochastic simulation algorithm (SSA or Gillespie algorithm), we sample from the Poisson process, i.e. $x \sim X$. Consequently, we can generate a series of n independent realisations $x^{(1)}, \ldots, x^{(n)} \sim X$.

3.2 Law of large numbers and convergence

Definition 13. *Let $x^{(1)}, \ldots, x^{(n)}$ denote a sequence of independent identically distributed random variables with finite expectation value $\mu = \mathbb{E}[x^{(j)}]$ and positive variance, $\sigma^2 = \mathbb{V}[x^{(j)}]$.*
The sample mean and the sample variance are:

$$\bar{x}^{(n)} = \frac{1}{n}\sum_{j=1}^{n} x^{(j)}, \qquad \hat{x}^{(n)} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x^{(j)} - \bar{x}^{(n)}\right)^2.$$

Remark 3. *In our context, let $x^{(1)}, \ldots, x^{(n)}$ be the state of our system in a set of realisations $1, \ldots, n$ of a Poisson process at some time t, i.e. $x^{(1)}, \ldots, x^{(n)} \sim X_t$. Then, $\bar{x}^{(n)}$ and $\hat{x}^{(n)}$ are the empirical estimates (mean, variance) of the process at a particular time.*

Proposition 1 (strong law of large numbers)**.** *For x with finite mean $\mu = \mathbb{E}[x^{(j)}]$ and positive variance $\sigma^2 = \mathbb{V}[x^{(j)}]$, if $x^{(1)}, \ldots, x^{(n)}$ denotes a sequence of independent identically distributed random variables used to compute the sample mean $\bar{x}^{(n)}$ and variance $\hat{x}^{(n)}$, we have*

$$P\big(\lim_{n\to\infty} |\bar{x}^{(n)} - \mu| = 0\big) = 1.$$

and

$$P\big(\lim_{n\to\infty} |\hat{x}^{(n)} - \sigma^2| = 0\big) = 1.$$

Proposition 2 (weak law of large numbers)**.** *For x with finite mean $\mu = \mathbb{E}[x^{(j)}]$ and positive variance $\sigma^2 = \mathbb{V}[x^{(j)}]$, if $x^{(1)}, \ldots, x^{(n)}$ denotes a sequence of independent identically distributed random variables used to compute the sample mean $\bar{x}^{(n)}$ and variance $\hat{x}_{(n)}$.*

$$P(|\bar{x}^{(n)} - \mu| > \varepsilon) \xrightarrow[\infty]{n} 0 \ \ \forall\, \varepsilon > 0$$

Remark 4. *The weak law of large numbers says that for every sufficiently large fixed n, the sample mean $\bar{x}^{(n)}$ is likely to be near the expectation value μ.*

Question: How fast will it converge? How often do we need to sample?

Proposition 3 (convergence of a sample statistic)**.** *Under the assumptions and definitions above, the standard deviation of the sample mean $\bar{x}^{(n)}$ expressed as*

$$\varepsilon = \sqrt{\mathbb{E}[|\bar{x}^{(n)} - \mu|^2]}, \quad \textit{(standard deviation of the sample mean)}$$

shrinks at the order of $\mathcal{O}\left(1/\sqrt{(}n)\right)$ with the number of samplings/realizations n.

Proof.

$$\begin{aligned}
\mathbb{E}[|\bar{x}^{(n)} - \mu|^2] = \mathbb{V}[\bar{x}^{(n)}] = &\ \mathbb{V}\left[\tfrac{1}{n}\textstyle\sum_{j=1}^{n} x^{(j)}\right] && \text{(Definition 9 \& 8)} \\
= &\ \tfrac{1}{n^2}\textstyle\sum_{j=1}^{n} \mathbb{V}[x^{(j)}] && \text{(variance of a lin. comb. of indep. random var.)} \\
= &\ \tfrac{1}{n^2} \cdot n \cdot \mathbb{V}[X] && (x^{(j)} \text{ i.d.; } x^{(j)} \sim X) \\
= &\ \tfrac{1}{n}\mathbb{V}[X] \\
\Rightarrow &\ \varepsilon = \left(\tfrac{1}{\sqrt{n}}\right) \cdot c
\end{aligned}$$

where c is a constant (the standard deviation of the Poisson process). □

Remark 5. *To double the accuracy, the number of realisations n has to be quadrupled.*

Example 6 ((linear) SIR model)**.** *You are given the following epidemiologic model of a small but virulent outbreak (linear version of the susceptible-infected-recovered (SIR) model). Stoichiometric matrix S:*

	R_3	R_4	R_5	R_6
x_2	1	-1	-1	0
x_3	0	0	1	-1
x_4	0	1	0	0

and propensities (reaction rates): $a_3 \dots a_6$.

$$\begin{aligned} a_3 &= \frac{\lambda}{\delta} \cdot x_2 \cdot \beta \\ a_4 &= x_2 \cdot 3 \cdot 10^7 \cdot \delta \\ a_5 &= x_2 \cdot k_r \\ a_6 &= x_3 \cdot \delta \end{aligned}$$

where the number of susceptible individuals stays constant at $x_1(t) = \frac{\lambda}{\delta}$. x_2 *are the number of infected individuals,* x_3 *are the number of individuals that recovered and are subsequently resistant to infection and* x_4 *are the number of individuals that died from the infection. Parameter values are* $\lambda = 1 \cdot 10^{-4}$, $\delta = 1 \cdot 10^{-8}$, $\beta = 5 \cdot 10^{-5}$, $k_r = 0.3$ *and the initial state is* $x_1(t_0) = \frac{\lambda}{\delta}$, $x_2(t_0) = 5$, $x_3(t_0) = 0$, $x_4(t_0) = 0$.

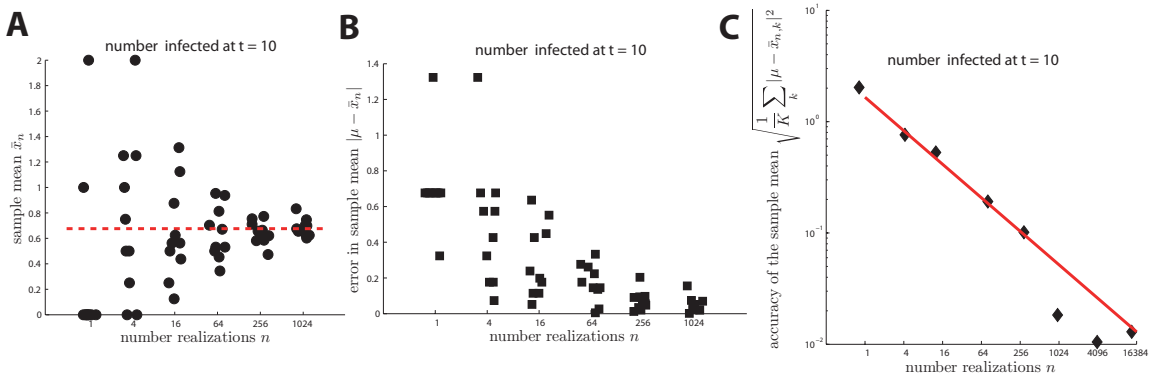


Figure 3.1: A: Each dot represents an empirical estimate of the sample mean calculated from a batch of n realisations (x-axis). B: Corresponding error in the sample mean $|\bar{x}^{(n)} - \mu|$ for a known true expectation value μ. C: Trajectory of the average error in the sample mean reduces approx. by half when the number of realisations is quadrupled.

3.3 The expectation value of a poisson process

Recap II: Reaction rates and their order.

The single reaction j

$$\underbrace{\sum_i m_i x_i}_{educts} \xrightarrow{a_j(x,k)} \underbrace{\sum_i n_i x_i}_{products}$$

order	reaction	propensity/rate	unit of k
0^{th}	$\emptyset \xrightarrow{k_\emptyset} ...$	$a_0(k) = k_\emptyset$	mole/time
1^{st}	$X_i \xrightarrow{k_1} ...$	$a_1(X,k) = k_1 \cdot X_i$	1/time
2^{nd}	$X_i + X_\ell \xrightarrow{k_2} ...$	$a_2(X,k) = k_2 \cdot X_i \cdot X_\ell$	1/(time · mole)

Table 3.1: Simple reactions and their order.

has reaction rate $a_j(x,k) = k_j \cdot \prod_i x_i^{m_i}(t)....$ (elementary/simple reactions), yielding the *#reactions* $\times$ 1 *reaction rate vector* $\mathbf{a}(x,k) = \{a_j(x,k)\}$

Recap III: A linear system.
A linear system may only involve reactions of order 0 and 1.

3.3.1 Linear case.

Example 7 (Birth–Death Process)**.** *Let* $\Omega = \mathbb{N}_0$.

$$R_1 : \emptyset \xrightarrow{\alpha} x, \qquad\qquad R_2 : x \xrightarrow{\beta} \emptyset.$$

with stoichiometric matrix S *and propensity function vector* $r = \{r_i\}$

	R_1	R_2
x	1	-1

$$\begin{aligned} a_1 &= \alpha \\ a_2 &= x \cdot \beta. \end{aligned}$$

The Poisson process formulation looks like:

$$X_t = X_0 + \mathcal{P}(\alpha t) - \mathcal{P}\left(\int_0^t \beta X_s ds\right).$$

where we write X_t *for* $X(t)$ *and* X_0 *instead of* $X(t_0)$ *for the initial state.*

Derivation

$$\begin{aligned} X_t &= X_0 + \mathcal{P}(\alpha t) - \mathcal{P}\left(\int_0^t \beta X_s ds\right) \\ \mathbb{E}[X_t] &= X_0 + \mathbb{E}\Big[\mathcal{P}(\alpha t)\Big] - \mathbb{E}\left[\mathcal{P}\left(\int_0^t \beta X_s ds\right)\right] \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X_t] &= \alpha - \beta \cdot \mathbb{E}[X_t] \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X_t] &= S \cdot \mathbf{a}(\mathbb{E}[X_t], \ldots), \end{aligned}$$

Remark 6. *For linear systems, the ODE-system describes how the expectation value of the Poisson process $\mu = \mathbb{E}[X]$ evolves in time.*

Example 8 ((linear) SIR model)**.** *Consider the (linear) SIR model from above. The ODE solution (green) alongside the sample mean of the process (dashed black line) and the range of prediction contained by $\pm$ 1 standard deviation (dashed red lines) are shown below.*

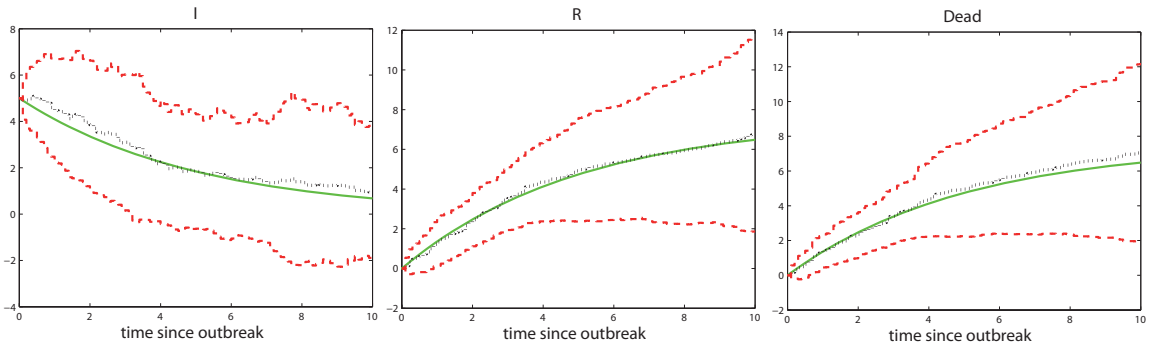


Figure 3.2: From left to right: Number of infected, recovered and individuals that died from the infection.

3.3.2 Note.

Knowledge about the expected number of particles $\mathbb{E}[X]$ or any higher moments (variance, skewness, kurtosis) of the Poisson process *may not* provide essential answers to questions of interest. E.g. in the above mentioned example (the SIR model), no answer is given to the question *'what is the probability that the infection is eliminated at $t = 10$?'*. In essence, this accounts for all processes which have multiple modes of interest (i.e. multi-stable systems), more below. Similarly, information of particular paths of the underlying process may not be given.

3.3.3 Non-linear case.

Example 9 (Non-linear SIR)**.** *You are given the following epidemiologic model of a virulent outbreak (also called susceptible-infected-recovered (SIR) model). Stoichiometric matrix S:*

	R_1	R_2	R_3	R_4	R_5	R_6
x_1	1	-1	-1	0	0	0
x_2	0	0	1	-1	-1	0
x_3	0	0	0	0	1	-1
x_4	0	0	0	1	0	0

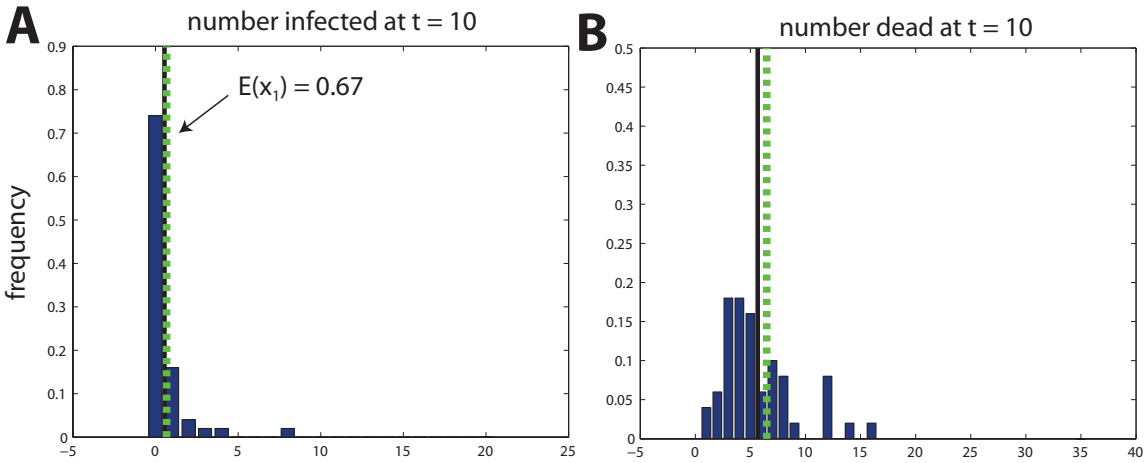


Figure 3.3: A: Number of infected and B: Number of individuals that died from the infection at $t = 10$. Bars represent the empirical distribution for $n = 30$ stochastic realisations with the empirical means illustrated by vertical black lines. The corresponding ODE solution is depicted by a dashed green vertical line.

and propensities (reaction rates): $a_1 \dots a_6$.

$$
\begin{aligned}
a_1 &= \lambda && (3.1)\\
a_2 &= x_1 \cdot \delta && (3.2)\\
a_3 &= x_1 \cdot x_2 \cdot \beta && (3.3)\\
a_4 &= x_2 \cdot 3 \cdot 10^7 \cdot \delta && (3.4)\\
a_5 &= x_2 \cdot k_r && (3.5)\\
a_6 &= x_3 \cdot \delta && (3.6)
\end{aligned}
$$

where x_1 denotes the number of susceptible individuals, x_2 are the number of infected individuals, x_3 are the number of individuals that recovered and are subsequently resistant to infection and x_4 are the number of individuals that died from the infection.
The Poisson process formulation for x_2 looks like:

$$
\begin{aligned}
X_{2,t} &= X_{2,t_0} + \mathcal{P}\left(\int_0^t \beta \cdot X_{2,s} \cdot X_{1,s} ds\right) - \mathcal{P}\left(\int_0^t (k_r + \delta \cdot 10^7) X_{2,s}\, ds\right)\\
\mathbb{E}[X_{2,t}] &= X_{2,t_0} + \mathbb{E}\left[\mathcal{P}\left(\int_0^t \beta \cdot X_{2,s} \cdot X_{1,s} ds\right)\right] - \mathbb{E}\left[\mathcal{P}\left(\int_0^t (k_r + \delta \cdot 10^7) X_{2,s}\, ds\right)\right]\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X_{2,t}] &= \beta \cdot \mathbb{E}\left[X_{2,t} \cdot X_{1,t}\right] - (k_r + \delta \cdot 10^7) \cdot \mathbb{E}[X_{2,t}]\\
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X_{2,t}] &\neq \beta \cdot \mathbb{E}\left[X_{2,t}\right] \cdot \mathbb{E}\left[X_{1,t}\right] - (k_r + \delta \cdot 10^7) \cdot \mathbb{E}[X_{2,t}],
\end{aligned}
$$

in general. However, if the covariance $\mathbb{C}$ between considered variables (here: X_1 and X_2) is zero, then the ODE is accurate. Also, if $\mathbb{E}[X_{2,t} \cdot X_{1,t}] \approx \mathbb{E}[X_{2,t}] \cdot \mathbb{E}[X_{1,t}]$, then the ODE approximation will be appropriate. The latter case often occurs when $[X_t] \to \infty$ (large particle number limit).

3.3.4 An example where the ODE fails.

Example 10 (Slögl model)**.** *Denote the following model*

$$R_1 : x_2 + 2x_1 \xrightarrow{k_1} 3x_1, \qquad R_2 : 3x_1 \xrightarrow{k_2} x_2 + 2x_1. \tag{3.7}$$

$$R_3 : x_1 \xrightarrow{k_3} x_3, \qquad R_4 : x_3 \xrightarrow{k_4} x_1, \tag{3.8}$$

with propensities (stochastic reaction rates): $a_1 \dots a_4$.

$$a_1 = k_1 \cdot x_2 \cdot x_1(x_1 - 1)/2 \tag{3.9}$$

$$a_2 = k_2 \cdot x_1(x_1 - 1)(x_1 - 2)/6 \tag{3.10}$$

$$a_3 = k_3 \cdot x_1 \tag{3.11}$$

$$a_4 = k_4 \cdot x_3 \tag{3.12}$$

with parameter values are $1.5{\cdot}10^{-7}$, $1.66667{\cdot}10^{-5}$, $1{\cdot}10^{-3}$ and 3.5 for $k_1, \dots, k_4$ respectively. The initial values of x_2 and x_3 are 10^5 and $2 \cdot 10^5$ respectively. The initial value for x_1 is $x_1(t_0) = 250$. Displayed are also ODE solutions for $x_1(t_0) = 247$ and $x_1(t_0) = 248$. The stoichiometric matrix S is given by:

	R_1	R_2	R_3	R_4
x_1	1	-1	1	-1
x_2	-1	1	0	0
x_3	0	0	-1	1

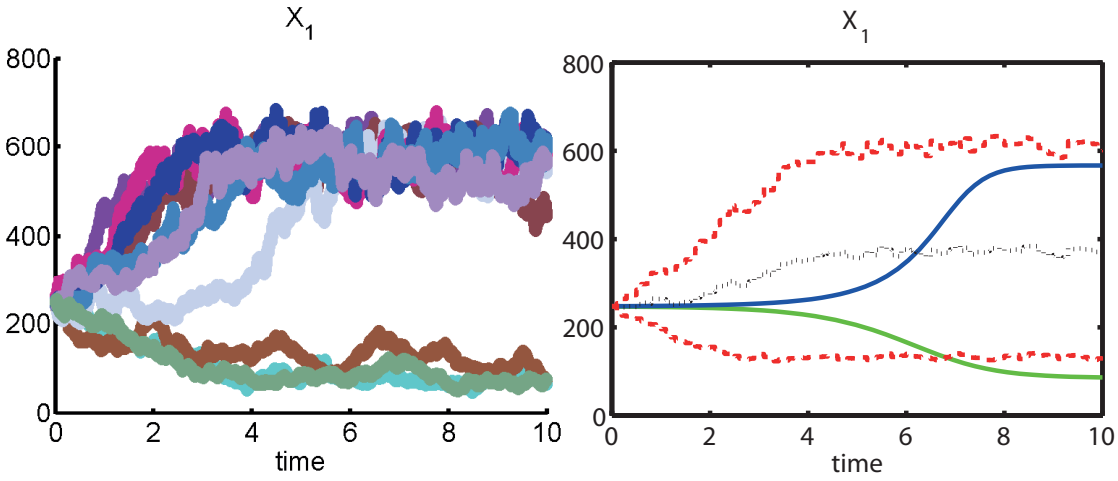


Figure 3.4: A: Stochastic trajectories of the Sloegl model. B: Sample mean (black dashed lines) and area covered by the mean $\pm$ one standard deviation, based on $n = 10$ trajectories, together with the solution of the corresponding ODE with $x_1(t_0) = 247$ and $x_1(t_0) = 248$ (green and blue lines).

Chapter 4

Integration I

Definition 14. *Let f be a non-negative function on the interval $[a,b] \subset \mathbb{R}$. Then the Riemann Sum is defined to be*

$$\int_a^b f(x)dx := \lim_{h\to 0} \sum_{n=1}^{N} f(a+(n-1)h)\cdot h,$$

where $N = (b-a)/h$ are the number of sub intervals of length h in $[a,b]$. The following are some simple properties of the Riemann sum:

- *Change in the direction of integration,*

$$\int_a^b f(x)dx = -\int_b^a f(x)dx.$$

- *Additivity, for $a < b < c$*

$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx.$$

- *If $f(x) \in \mathbb{R}$, then*

$$\int_a^b f(x)dx \leq \int_a^b |f(x)|dx.$$

Theorem 2 (Fundamental Theorem of Calculus)**.** *Let $f : [a,b] \to \mathbb{R}$ be a continuous function from a closed set $[a,b] \subset \mathbb{R}$ into $\mathbb{R}$. Define*

$$F(x) := \int_a^x f(y)dy, \text{ with } y \in [a,x].$$

If F is uniformly continuous and differentiable, then

$$\frac{dF(x)}{dx} = f(x), \text{ or alternatively, } \frac{d}{dx}\int_a^x f(y)dy = f(x).$$

We refer to F as the antiderivative of f.

Example 11. *Well known Indefinite Integrals*

1. *Exponential: $\lambda \neq 0$,*
$$\int e^{\lambda x}dx = \frac{1}{\lambda}e^{\lambda x} + c.$$
2. *Power Rule: $n \neq -1$,*
$$\int x^n dx = \frac{x^{n+1}}{n+1} + c.$$
3. *Power Rule: $n = -1$ and x is strictly positive or strictly negative,*
$$\int \frac{1}{x}dx = ln(|x|) + c.$$

In all cases c is some constant.

Theorem 3. *Taylor Expansion*

Let f be an infinitely differentiable function from $\mathbb{R}$ into $\mathbb{R}$. Then the Taylor series of $f(x)$ centred around the point $a \in \mathbb{R}$ is given by,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \cdots . \tag{4.1}$$

Where n is the refereeing to the order of the derivative. For $a = x$ and $\Delta x > 0$,

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x)}{2!}(\Delta x)^2 + \cdots + \frac{f^{(n)}(x)}{n!}(\Delta x)^n + \cdots . \tag{4.2}$$

Remark 7. *Note that for the purpose of this course we are only interested in functions where the Taylor series does converge.*

Example 12. *The Taylor series of e^x centred at $a = 0$ is given by,*

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

Proposition 4. *Let $f : \mathbb{R} \to \mathbb{R}$ be infinitely differentiable and $\Delta x > 0$,*

$$\int_0^{\Delta x} f(x)dx = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)\Delta x^{n+1}}{(n+1)!}.$$

Furthermore, for $m \in \mathbb{N}$,

$$\left| \int_0^{\Delta x} f(x)dx - \sum_{n=0}^{m-1} \frac{f^{(n)}(0)\Delta x^{n+1}}{(n+1)!} \right| = O(\Delta x^{m+1}).$$

We refer to $\sum_{n=0}^{m-1} \frac{f^{(n)}(0)\Delta x^{n+1}}{(n+1)!}$ as the $O(\Delta x^{m+1})$ Taylor approximation of $\int_0^{\Delta x} f(x)dx$.

Proof. We take the Taylor series of f centred at $a = 0$ then reduce.

$$\begin{aligned} \int_0^{\Delta x} f(x)dx &= \int_0^{\Delta x} \sum_{n=0}^{\infty} \frac{f^{(n)}(0)x^n}{n!} dx \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} \int_0^{\Delta x} x^n dx \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)\Delta x^{n+1}}{(n+1)!} \end{aligned}$$

□

Exercise 7. *Using Proposition 4 show that,*

$$\int_0^{\Delta x} e^x dx = e^{\Delta x} - 1.$$

Exercise 8. *Compute $O(0.1^4)$ Taylor approximation of $\int_0^{0.1} e^{x^2} dx$.*

Example 13. *Rectangle Rule: Let f be infinitely differentiable; $x \in \mathbb{R}$ and $\Delta x > 0$,*

$$\int_x^{x+\Delta x} f(y)dy = f(x)\Delta x + O(\Delta x^2).$$

Example 14. *Trapezoidal Rule: Let f be infinitely differentiable; $x \in \mathbb{R}$ and $\Delta x > 0$,*

$$\int_x^{x+\Delta x} f(y)dy = \frac{f(x+\Delta x) + f(x)}{2} \cdot \Delta x + O(\Delta x^3).$$

Using Proposition 4 and centering the Taylor expansion around x, we get,

$$
\begin{aligned}
\int_x^{x+\Delta x} f(y)dy &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)\Delta x^{n+1}}{(n+1)!}, \\
&= f(x)\Delta x + \frac{f'(x)\Delta x^2}{2!} + \sum_{n=2}^{\infty} \frac{f^{(n)}(x)\Delta x^{n+1}}{(n+1)!}, \\
&= \frac{f(x)}{2}\Delta x + \left(\frac{f(x)\Delta x}{2} + \frac{f'(x)\Delta x^2}{2!} + \sum_{n=2}^{\infty} \frac{f^{(n)}(x)\Delta x^{n+1}}{(n+1)!} \right), \\
&= \frac{f(x)}{2}\Delta x + \underbrace{\left(f(x) + f'(x)\Delta x + \sum_{n=2}^{\infty} \frac{f^{(n)}(x)\Delta x^n}{(n)!} \right)}_{f(x+\Delta x)} \frac{\Delta x}{2} \\
&\qquad + \underbrace{\left(\sum_{n=2}^{\infty} \frac{f^{(n)}(x)\Delta x^n}{(n+1)!} - \sum_{n=2}^{\infty} \frac{f^{(n)}(x)\Delta x^n}{(n)!} \right) \cdot \frac{\Delta x}{2}}_{O(\Delta x^3)}, \\
&= \frac{f(x) + f(x+\Delta x)}{2} \cdot \Delta x + O(\Delta x^3).
\end{aligned}
$$

Remark 8. *Notice that the classical numerical integration methods are just the application of the Taylor expansion. But the Taylor expansion has a heavy assumption on all the derivatives existing and well behaving. What do you suppose could happen if we were to integrate functions where these assumptions no longer hold?*

Chapter 5

Integration II

5.1 Ordinary Differential Equation (ODE)

Definition 15. *A general first-order differential equation for a function $y : \mathbb{R} \to \mathbb{R}$ is written as,*

$$\frac{dy}{dx} = f(x, y), \tag{5.1}$$

where $f(x, y)$ is a mapping from $\mathbb{R}^2 \to \mathbb{R}$.

Definition 16. *Linear ODE An ODE is linear if it is of the form,*

$$\frac{dy}{dx} = f(x) \cdot y + g(x),$$

where $f(x), g(x) : \mathbb{R} \to \mathbb{R}$. If $g(x) =$ constant for all x, then we say the ODE is homogenous. Otherwise, the ODE is called an inhomogeneous ODE. If $f(x)$ is constant and $g(x) = 0$ for all x, then the ODE is said to be an autonomous ODE.

Example 15. *What would be the classification of the derivative of the expectation of the birth death process?*

$$\frac{d\mathbb{E}[X_t]}{dt} = \alpha - \beta \cdot \mathbb{E}[X_t]$$

5.2 Separable Equations

A way of solving simple ODEs is by the Separation of Variables. Let us consider the following ODE,

$$\frac{dy}{dx} = f(x)g(y).$$

We can simply solve this by gathering the like variables on each side of the equality,

$$\frac{dy}{dx}\frac{1}{g(y)} = f(x).$$

Then we can move dx to the right hand side, which is equivalent to taking the indefinite integral of the respective variables,

$$\int g(y)dy = \int f(x)dx.$$

The indefinite integral generate multiple possible solutions to the equation. Hence an initial value for y has to prescribed to reduce the possible solutions to one. When solving an ODE and an initial value is given, this is called the initial value problem (IVP).

Remark 9. *The existence of solutions of ODEs is out of the scope of this course. However, it is very important to note that there are a set of conditions, if satisfied, we know the existence of a solution to the IVP. One must consider if their particular ODE satisfies these conditions before proposing a solution, even if it a numerical solution.*

Example 16. *Find the solution to the following ODE*

$$\frac{dy}{dx} = -2xy, \text{ where } y(0) = 1.$$

Example 17. *Consider the Death process*

$$X_t = 10 - \mathcal{P}\left(\int_0^t 2 \cdot X_s ds\right).$$

We can derive that

$$\frac{d\mathbb{E}[X_t]}{dt} = -2 \cdot \mathbb{E}[X_t].$$

Solve the above ODE for the analytical solution of the expectation of the death process.

5.3 System of ODES

In practise, we are seldom working with a single ODE. Rather, we multiple ODEs where the solution of one is influencing the change of derivative in another. When considering such reactions, we group all odes into a vector and try to solve them all simultaneously as a system of ODES.

Example 18. *Let $x(t), y(t) \in \mathbb{R}$ and $t > 0$,*

$$\frac{dx(t)}{dt} = -y(t)x(t) + 2$$
$$\frac{dy(t)}{dt} = y(t)x(t)$$

Let $v(t) = (x(t), y(t)) \in \mathbb{R}^2$. Then our system of ODEs would be

$$\frac{dv(t)}{dt} = F(v(t)),$$

where $F(v(t)) = (-y(t)x(t) + 2, y(t)x(t))$. How would you classify the above ODE?

Definition 17. *Let $v(t) \in \mathbb{R}^N$ be a $N \in \mathbb{N}$ dimensional vector evolving through time $t > 0$. Let $F : \mathbb{R}^N \to \mathbb{R}^N$ be a mapping. Then*

$$\frac{dv(t)}{dt} = F(v(t))$$

is a system of ordinary differential equations. If F is linear, then F is a matrix of dimension $N \times N$, we write this as

$$\frac{dv(t)}{dt} = Av(t).$$

Example 19. *Consider the following system:*

$$a \xrightarrow{\lambda_1} b \tag{5.2}$$

$$b \xrightarrow{\lambda_2} b + b \tag{5.3}$$

$$b \xrightarrow{\lambda_3} a \tag{5.4}$$

$$\emptyset \xrightarrow{\lambda_4} a \tag{5.5}$$

$$a \xrightarrow{\lambda_5} \emptyset \tag{5.6}$$

The ODEs which solve for the evolution of the expectation of this system are given by:

$$\begin{bmatrix} \frac{da(t)}{dt} \\ \frac{db(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\lambda_1 - \lambda_5 & \lambda_3 \\ \lambda_1 & \lambda_2 - \lambda_3 \end{bmatrix} \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} + \begin{bmatrix} \lambda_4 \\ 0 \end{bmatrix}.$$

What would be the solution of this ODE?

Chapter 6

Numerics of ODES I

6.1 Solutions for a Linear Inhomogeneous ODE

We will use the method of integrating factor to solve explicitly some simple linear inhomogeneous ODEs. Let us consider solving the following ODE,

$$\frac{dy(x)}{dx} + p(x)y(x) = q(x),$$

with initial condition $y(0) = y_0$ and $x \in [x_0, x_f]$.

We introduce the integration factor

$$\phi(x) := \int p(x)dx. \tag{6.1}$$

Given $p(x)$ is smooth and $\phi(x)$ is uniformly continuous, we know that

$$\frac{d\phi(x)}{dx} = p(x). \tag{6.2}$$

To solve for y we multiple $e^{\phi(x)}$ to both sides of our ODE,

$$e^{\phi(x)}\left[\frac{dy(x)}{dx} + p(x)y(x)\right] = e^{\phi(x)}q(x). \tag{6.3}$$

Remark 10.

$$e^{\phi(x)}\left[\frac{dy(x)}{dx} + p(x)y(x)\right] = \frac{d}{dx}\left[e^{\phi(x)}y(x)\right]. \tag{6.4}$$

Using the remark above, we can reduce (6.3) to

$$\frac{d}{dx}\left[e^{\phi(x)}y(x)\right] = e^{\phi(x)}q(x). \tag{6.5}$$

Integrating both side with respect to x over the interval $[x_0, x]$ gives,

$$e^{\phi(x)}y(x) = \int_{x_0}^{x} e^{\phi(z)}q(z)dz + C.$$

Moving the exponential the right hand side gives us,

$$y(x) = e^{-\phi(x)} \int_{x_0}^{x} e^{\phi(z)}q(z)dz + e^{-\phi(x)}C.$$

Example 20. *Consider the Birth–Death process,*

$$\emptyset \xrightarrow{\lambda} X, \quad X \xrightarrow{\beta} \emptyset.$$

We have shown earlier that the derivative of the expectation is given by

$$\frac{d\mathbb{E}[X_t]}{dt} = \alpha - \beta \cdot \mathbb{E}[X_t].$$

Let X_0 be the initial population. Show that the solution of the ODE is given by

$$\mathbb{E}[X_t] = \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right) + e^{-\beta t}X_0.$$

6.2 Numerical Methods for Solving ODEs

We seldom find analytical solutions for ODEs solved in practise. Hence we have to resort to using numerical approximations. Here in we consider the solutions of the following generalised system of first ODE,

$$\frac{dv(t)}{dt} = F(v(t), t),$$

where $F : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ and $v(0), v(t) \in \mathbb{R}^N$.

Notation 1. *We split the time interval $T = [0, t_{final}]$ into subintervals of length h. We denote the approximation of $v(n \cdot h)$ by u_n. Similarly let $t_n := n \times h$.*

Example 21 (Explicit Euler Method)**.**

$$u_{n+1} = u_n + hF(u_n, t_n)$$

$$\sum_n |u_n - v(h \cdot n)| = O(h^1)$$

Example 22 (Implicit Euler Method)**.**

$$u_{n+1} = u_n + hF(u_{n+1}, t_{n+1})$$

$$\sum_n |u_n - v(h \cdot n)| = O(h^1)$$

Example 23 (Midpoint Method)**.**

$$u_{n+1} = u_n + hF\left(u_n + \frac{h}{2}F(u_n, t_n), t_n + \frac{h}{2}\right)$$

$$\sum_n |u_n - v(h \cdot n)| = O(h^2)$$

Example 24 (Order 4 Runge-Kutta Method)**.**

$$\begin{aligned} k_1 &= hF(u_n, t_n) && (6.6)\\ k_2 &= hF\left(u_n + \frac{k_1}{2}, t_n + \frac{h}{2}\right) && (6.7)\\ k_3 &= hF\left(u_n + \frac{k_2}{2}, t_n + \frac{h}{2}\right) && (6.8)\\ k_4 &= hF\left(u_n + k_3, t_{n+1}\right) && (6.9)\\ u_{n+1} &= u_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} && (6.10) \end{aligned}$$

$$\sum_n |u_n - v(h \cdot n)| = O(h^4)$$

Chapter 7

Numerics of ODES II

7.1 Order 4 Runge-Kutta Method

Example 25 (Order 4 Runge-Kutta Method)**.**

$$\begin{aligned} k_1 &= hF(u_n, t_n) && (7.1)\\ k_2 &= hF\left(u_n + \frac{k_1}{2}, t_n + \frac{h}{2}\right) && (7.2)\\ k_3 &= hF\left(u_n + \frac{k_2}{2}, t_n + \frac{h}{2}\right) && (7.3)\\ k_4 &= hF\left(u_n + k_3, t_{n+1}\right) && (7.4)\\ u_{n+1} &= u_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} && (7.5) \end{aligned}$$

$$\sum_n |u_n - v(h \cdot n)| = O(h^4)$$

7.2 Stability

Definition 18. *A numerical IVP solver is said to be stable if small perturbations of the initial condition do not cause large deviations between the numerical approximation and the true solution. Given the true solutions exists and is bounded.*

Let us consider the stability of our numerical solvers for a simple autonomous ODE,

$$\frac{dy(t)}{dt} = \lambda y(t), \text{ with } t \in [0, T], y(0) = y_0.$$

We know that,

$$y(t) = e^{\lambda t} y_0.$$

We define our perturbation to be $\delta > 0$, then the propagation of $y_0 + \delta$ is,

$$y_\delta(t) = (y_0 + \delta)e^{\lambda t}.$$

We can see that if λ is greater than 0, then the deviation between y_δ and y get larger with time. On the other hand, if λ is less than 0, then $\|y_\delta - y\| \leq \|\delta\|$. We can say for the analytical solution, we have stability when $\lambda < 0$ and instability when $\lambda > 0$. Similarly, when considering our numerical approximations, which have a step size ((h) as the free parameter, we can ask the question, for which values of h is our approximation scheme stable or unstable?

Example 26 (Stability of Explicit Euler)**.** *Let us consider $\lambda < 0$, so the analytical solution is stable. How let us investigate the condition under which the approximation is also stable.*

$$\begin{aligned} v_{n+1} &= v_n + h\lambda v_n, \\ &= (1 + \lambda h)v_n. \end{aligned}$$

If we consider taking N Forward Euler steps, we have that,

$$v_N = (1 + \lambda h)^N v_0.$$

We can see that to stabilise the approximation we would need the condition that $|(1+\lambda h)| < 1$. This implies that for the Forward Euler to give us stability we need that $h < \frac{-2}{\lambda}$.

Example 27 (Stability of Implicit Euler)**.** *A single step of the Implicit Euler gives us that*

$$\begin{aligned} v_{n+1} &= v_n + h\lambda v_{n+1}, \\ &= \left(\frac{1}{1 - \lambda h}\right) v_n. \end{aligned}$$

If we consider N steps of the Implicit Euler, we have that,

$$v_N = \left(\frac{1}{1 - \lambda h}\right)^N v_0.$$

We can see here, that $\left(\frac{1}{1-\lambda h}\right)$ is less than one for all $\lambda < 0$ and $h > 0$. So the Implicit Euler is stable for all choices of h.

We can see the difference between the Explicit and the Implicit Euler clearly with respect to stability. The two methods have similar order of convergence with respect to step size, however, if the derivative of the ODE is changing rapidly, the Implicit Euler can take much larger step sizes compared to the Forward Euler.

Definition 19. *A numerical method for solving IVPs is said to be A–stable if its region of stability includes the entire complex half–plane with negative real part.*

Theorem 4. *If a linear numerical method is A–stable, then it must be an implicit method. Furthermore, the order of the method is at most* 2.

7.3 Stiffness

The definition of stiffness has not yet been standardised as different fields interpret different phenomenon as stiffness. However, for our purpose, we refer to a system being stiff if the time step we are having to choose is many orders of magnitude smaller than our time horizon.

Chapter 8

Dynamics & stability analysis for ODEs

Recap: Let us consider the following system of first-order *autonomous* ODEs,

$$\frac{dv(t)}{dt} = F(v(t)),$$

where $F : \mathbb{R}^N \to \mathbb{R}^N$ is a mapping and $v(0), v(t) \in \mathbb{R}^N$ and $v(t) = (x_1(t), \ldots, x_n(t)) \in \mathbb{R}^N$ is an $N \in \mathbb{N}$ dimensional vector evolving through time $t > 0$.

Remark 11. *At any time t, the state can be represented by the point $v(t) = (x_1(t), \ldots, x_n(t))$. If we draw the sequence of points through which the system passes as it evolves, we will draw out a continuous curve called a trajectory or orbit.*

8.1 Phase plane analysis

Suppose you wanted to analyze the properties of your dynamical system, that is to sketch all possible trajectories. The system of ODEs (above) give us the magnitude and direction of change of our variables at any point in the phase space. In other words, $F(v(t)) = (f_1(v(t)), \ldots, f_n(v(t)))$ is as a kind of velocity vector for our dynamical system at the point $v(t)$.
If we draw velocity vectors in our state space, we get a *vector field* corresponding to our dynamical system. This vector field tells us how each possible trajectory (starting from anywhere is state space) may look like.

Remark 12. *This tells us something about the properties of our autonomous differential equations, rather than about their time evolution, but it helps a great deal constructing useful models.*

Example 28 (Birth-death process)**.** *Let us consider a simple one-dimensional birth-death process.*

$$x \xrightarrow{k} x + x \tag{8.1}$$

$$x \xrightarrow{\delta} \emptyset \tag{8.2}$$

where k and δ denote birth- and death rates respectively. The corresponding ODE is given by

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}t}x \quad = \quad (k - \delta)x \tag{8.3}$$

and has solution $x(t) = x(t_0) \cdot e^{(k-\delta)t}$.

Example 29 (Predator-Prey model.)**.** *Let us consider the predator-prey model.*

$$x_1 \xrightarrow{k_0} x_1 + x_1 \tag{8.4}$$

$$x_1 + x_2 \xrightarrow{k_1} x_2 \tag{8.5}$$

$$x_1 + x_2 \xrightarrow{k_2} x_2 + x_2 \tag{8.6}$$

$$x_2 \xrightarrow{\delta} \emptyset \tag{8.7}$$

where x_1, x_2 denotes the population of prey and predators (x_2 eats x_1). Parameters are 0.3, 0.01, 0.01 and 0.3 for k_0, k_1, k_2 and δ_2. The corresponding ODE is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}x_1 \quad = \quad x_1(k_0 - x_2 \cdot k_1 - x_2 \cdot k_2) \tag{8.8}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_2 \quad = \quad x_2(x_1 \cdot k_2 - \delta) \tag{8.9}$$

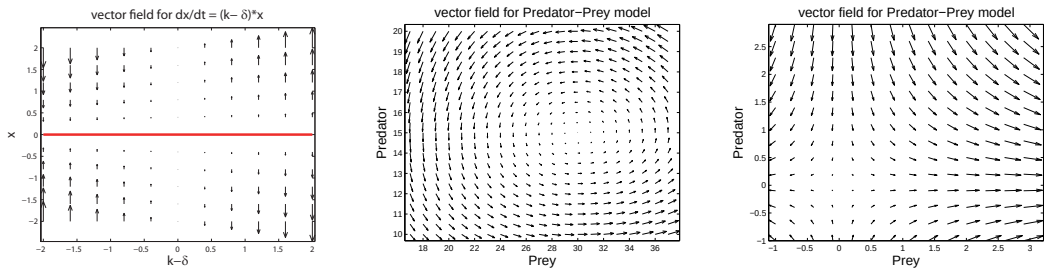


Figure 8.1: A: Vector field for the birth death process. B-C: Parts of the vector field for the Predator-Prey model around the fixed points.

8.2 Fixed points of ODEs

A fixed point is a point in state space that does not change upon application of a mapping. In line with the notation above, this implies

$$0 = F(v^*(t)),$$

where $v^*(t) = (x_1^*(t), \ldots, x_n^*(t))$.
Example 28:
In the birth-death example we have

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}t}x \quad = \quad (k - \delta)x, \tag{8.10}$$

hence $x^* = 0$ is a fixed point.
Note: Not all fixed points are stable. Whether a fixed point is stable or unstable determines the properties of the dynamical system.
Simple Example (Birth-death process, example 29):

$$f'(x) = (k - \delta) \tag{8.11}$$

If $f'(x^*) > 0$, the fixed point is unstable. If $f'(x^*) < 0$, the fixed point is stable.

Remark 13. *Note if we use the high dimensional formulation. The eigenvalue λ^* of the corresponding fixed point x^* is $f'(x^*)$. Since, we are only with real valued functions, we only have positive or negative eigenvalues.*

8.3 Formal method for linear stability analysis

Recap: Let f be an infinitely differentiable function from $\mathbb{R}$ into $\mathbb{R}$. Then the Taylor series of $f(x)$ centred around the point $a \in \mathbb{R}$ is given by,

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n + \cdots . \tag{8.12}$$

Where n is the referring to the order of the derivative.
Suppose $v^*(t) = (x_1^*(t), \ldots, x_n^*(t))$ is a fixed point, such that

$$0 = F(v^*(t)),$$

Lets make a first-order Taylor approximation of F around v^* (a linear approximation).

$$\frac{dv(t)}{dt} \quad = \quad F(v^*) + F'(v - v^*) + \ldots \tag{8.13}$$

$$\quad = \quad F'(v - v^*) + \ldots \tag{8.14}$$

where the derivative is to be interpreted as the **Jacobian matrix**. Note, our state vector $v = (x_1, \ldots, x_n)$ and the components of our rate vector are $F(\cdot) = (f_1(\cdot), \ldots, f_n(\cdot))$. Thus,

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \cdots \frac{\partial f_1}{\partial x_n} \\ \vdots \quad \vdots \\ \frac{\partial f_n}{\partial x_1} \cdots \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{8.15}$$

evaluated at v^* gives $\mathbf{J}^*$. Next, solve for the **eigenvalues** of this matrix, i.e. the roots of the **characteristic polynomial**.

$$|\lambda \mathbf{I} - \mathbf{J}^*| = 0.$$

Depending on the sign of the eigenvalues and on whether they are real- or complex numbers, conclusions about the characteristics of the fixed point can be made, see Table below.

Eigenvalues		Type	Stability
$\lambda_1, \lambda_2 \in \mathbb{R}$	$\lambda_1 < \lambda_2 < 0$	nodal sink (node)	a.-stable
	$\lambda_2 < \lambda_1 > 0$	nodal source	unstable
$\lambda_1, \lambda_2 \in \mathbb{R}$	$\lambda_1 < 0$, $\lambda_2 > 0$	saddle node	unstable
$\lambda_1, \lambda_2 \in \mathbb{R}$	$\lambda_1 = \lambda_2 < 0$	star point or improper node	a.-stable
	$\lambda_1 = \lambda_2 > 0$	star point or improper node	unstable
$\lambda_1, \bar{\lambda_2} \in \mathbb{C}$	$\Re(\lambda_{1,2}) = 0$	center	stable
$\lambda_1, \bar{\lambda_2} \in \mathbb{C}$	$\Re(\lambda_{1,2}) < 0$	spiral point	a.-stable
	$\Re(\lambda_{1,2}) > 0$	spiral point	unstable

Table 8.1: Fixed point classification for a 2-dimensional ODE system.

Workflow

1. determine all fixed points by solving $0 = F(v^*)$, for v^*.
2. compute the Jacobian matrix J
3. find all eigenvalues λ (the roots of the the characteristic polynomial) by solving $|\lambda \mathbf{I} - \mathbf{J}^*| = 0$ for λ.
4. consolidate the Table above to classify the fixed points.

Example 29, Predator-Prey model

1. Fixed points: $v^* = (0, 0)$; $v^+ = \left(\frac{\delta}{k_2}, \frac{k_0}{k_1 + k_2}\right) = (30, 15)$
2. $$J = \begin{bmatrix} k_0 - x_2(k_1 + k_2) & -x_1(k_1 + k_2) \\ x_2 \cdot k_2 & x_1 \cdot k_2 - \delta \end{bmatrix}$$

$$\Rightarrow J^* = \begin{bmatrix} 0.3 & 0 \\ 0 & -0.3 \end{bmatrix}$$

$$\Rightarrow J^+ = \begin{bmatrix} 0 & -0.15 \\ 0.15 & 0 \end{bmatrix}$$

3. $|\lambda \mathbf{I} - \mathbf{J}^*| = 0 = (\lambda - k_0)(\lambda + \delta) \Rightarrow \lambda_1 = -\delta = -0.3, \lambda_2 = k_0 = 0.3.$
   $|\lambda \mathbf{I} - \mathbf{J}^+| = 0 = \lambda^2 + (0.15)^2 \Rightarrow \lambda_{1/2} = 0 \pm \sqrt{-(0.15)^2}.$

4. v^* is a saddle node, v^+ is a stable center node (compare Fig. 8.1B–C).

Chapter 9

Inverse problems I

Recap: 'Forward problems', so far:

1. Stochastic processes $X_t \sim p_t$: simulation algorithm, statistical convergence, Poisson processes $\leftrightarrow$ ODEs.
   Example:
   $$X_t = X_0 + \mathcal{P}(X_t, \theta, t)$$
   where $X_t \in \mathbb{N}^n$ is an n-dimensional set of random variables and $\theta \in \mathbb{R}^m$ is an m-dimensional parameter set, i.e.
   $$\theta_{\mathcal{M}} = (\theta_{\mathcal{M},1} \dots \theta_{\mathcal{M},m})$$

2. ODEs: Integration, analytical- and numerical solution, analysis of dynamics and stability
   $$v(t) = \int F\big(v(t), \theta, t\big)$$
   where $F : \mathbb{R}^N \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^N$ is a mapping/an ODE nodel and $v(0), v(t) \in \mathbb{R}^N$ and $v(t) = (x_1(t), \dots, x_n(t)) \in \mathbb{R}^N$ is an $N \in \mathbb{N}$ dimensional vector evolving through time $t > 0$ and $\theta \in \mathbb{R}^m$ is an m-dimensional parameter set for the mapping F.

Note that all approaches so far answer related questions, namely

$$\text{known causality} \longrightarrow \text{prediction} \tag{9.1}$$

i.e. essentially one needs to know a **function**/model of **parameters** θ to obtain predictions. This type of problem is called the 'normal'- or 'forward problem'. Subsequently, we will define 'causality' in terms of

- a **structural model** $\mathcal{M}$

- model **parameters** $\theta_{\mathcal{M}}$

Subsequently, we will write θ instead of $\theta_{\mathcal{M}}$ as it should be clear now that parameters depend on the model (i.e. cannot be compared across different models). Note that the 'forward problem' assumes we know some underlying causality (defined above) from which we make our predictions. However, in real life situations the *causality* underlying an observation is unknown and not directly measurable. Finding the causal factors underlying an observation may be of key interest in many real-life situations.

9.1 Definition of an inverse problem.

An 'inverse problem' is the attempt of inferring causal factors $(\mathcal{M}, \theta)$ from a set of observations z that produced them. Mathematically speaking,

$$z \to (\mathcal{M}, \theta) \tag{9.2}$$

where $z = \{(t_i, y_i)\}$ denotes a set of observations with $t_i \in \mathbb{R}$ and $y_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. We have

$$\mathcal{M}(t, \theta) : \mathbb{R} \times \mathbb{R}^m \mapsto \mathbb{R}^d.$$

where d denotes the number of measured dimensions/species.

Goal

The goal is to estimate parameters θ (and models $\mathcal{M}$) such that the accordance with the data z is maximized. Remember:

> *If you would be a real seeker after truth, it is necessary that [...] you doubt, as far as possible, all things.* (R. Descartes)

Example 30 (Phylogenetic Inference)**.** *A good example of an inverse problem is phylogenetic reconstruction: The goal is to reconstruct the ancestry of genomic sequences. The data (genomic information, time of sampling) does not directly measure the ancestry of the genomic sequences, however it contains a great deal of information about it.*
Central to inferring the ancestry is to come up with a model of how sequences change over time $\mathcal{M}$ (substitution matrices, model of rate variability across the genome), how geneologies are passed on (assumption of tree-like ancestry etc..) and to fit its parameters θ (e.g. substitution rates, internal nodes, branch length). When all of this maximizes the accordance with the data, an optimal phylogenetic tree is obtained, under the modelling assumptions made.

9.2 Modelling aspects

Modelling is at the core of inverse problems and essentially one needs to be fully aware of the assumptions made (and how they may affect the outcome of the inverse problem). In this course we will touch upon the following *modelling*-related topics.

(a) Modelling measurements

(b) Parameter boundaries

(c) Constraints

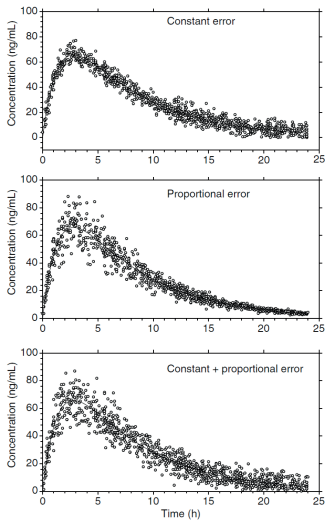
(d) (Prior information on parameters)

(e) (Population models)

(f) (Model selection)

9.2.1 Modelling measurements

Data is imperfect. Depending on the measurement procedure different errors may occur and assumption on the errors strongly affect the 'data accordance' in any particular example. Thus, error assumptions crucially influence the outcome of any inference procedure. We will in the following assume that time t can be measured exactly, while y can only be measured up to a certain precision.



Some error models

constant/additive error $y_i = x_{i|\mathcal{M},\theta} + \eta_i$
(scale-independent variance)

proportional error $y_i = x_{i|\mathcal{M},\theta}(1 + \eta_i)$
(constant coeff. of variation)

additive + proportional error $y_i = x_{i|\mathcal{M},\theta}(1 + \eta_{1,i}) + \eta_{2,i}$

where η_i is a random variable. y_i denotes the ith measurement and $x_{i|\mathcal{M},\theta}$ denotes the corresponding model prediction.

Example 31 (Direct measurement)**.** *For most direct measurements, given that the instruments have been calibrated appropriately, the measurement error may be assumed to be additive and scale-independent within the dynamic range of the assay.*

Example 32 (Complex measurement procedure)**.** *Assume the concentration of viruses is measured the following way: (i) Viral DNA is extracted from a defined volume of blood. (ii) This DNA is amplified using PCR (exponential amplification) for a specific number of PCR cycles and (iii) the resultant DNA is quantified. Since each virus has only a specific*

number of DNA, this is a good proxy for the number of viruses per unit volume.
Now, step (i) introduces an additive error (volume not exact, or distribution not completely even in the liquid). This additive error is amplified in step (ii), yielding a proportional error with regard to the amount of DNA before measurement. The measurement itself (step iii) will introduce an additive error. All in all, this may result in an additive + proportional measurement error.

Note: The random variable η_i is typically assumed to come from a normal distribution centered around zero, i.e. $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$. This can be motivated by the central limit theorem (and obviously only holds when many measurements exist). In a sparse and outlier-confounded data set, it may be advantageous to assume a different type of error distribution (one that has larger tails).

Note: Data may be transformed in many ways to (log transformation, weighing) to make errors comparable across different measurement scales (and thus enabling the fitting).

9.2.2 Likelihood of data

The conditional probability of observing the data y given the model $\mathcal{M}$ and its parameters θ, assuming measurement errors as above is called the **likelihood** of the data $\mathcal{L}_y$.

$$\mathcal{L}_y = p(y|\theta, \mathcal{M}) = \prod p(y_i|\theta, \mathcal{M})$$

since $p(y_i|\theta) < 1$, it is numerically more convenient to work with the (negative) log-likelihood instead, i.e.

$$\ell_y = -\log\ p(y|\theta) = -\sum \log\ p(y_i|\theta)$$

Note: Minimizing ℓ_y maximizes $\mathcal{L}_y$.

Exercise 9 (Likelihood)**.** *Maximize the likelihood of the single data point y_0, for model prediction x_0, assuming an additive error*

$$y_i = x_{i|\mathcal{M},\theta} + \eta_i$$

where $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$ with $\sigma = 1$.
Solution:

$$\mathcal{L}_y = \frac{1}{\sqrt{2 \cdot \sigma^2 \pi}} e^{-\frac{(y_0 - x_0)^2}{2\sigma^2}} \tag{9.3}$$

$$\Rightarrow \ell_y = \log(\sqrt{2 \cdot \sigma^2 \pi}) + \frac{(y_0 - x_0)^2}{2\sigma^2} \quad \textit{(neg. log-likelihood)} \tag{9.4}$$

$$\Rightarrow \frac{d\ell_y}{dx_0} = -(2y_0 + 2x_0) \cdot (2\sigma^2)^{-1} \tag{9.5}$$

$$\Rightarrow x_0 = y_0 \quad \textit{(Gaussian distr. is unimodal)} \tag{9.6}$$

the likelihood of y_0 is maximized when it coincides with the model predictions.

Exercise 10 (maximum likelihood estimate I)**.** *Consider the structural model $\mathcal{M}$:*

$$x_{i|k} = k \cdot t_i + x_0$$

with $x_0 = 1$, assuming an additive error

$$y_i = x_{i|k} + \eta_i$$

where $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$ with $\sigma = 1$.
You are given the data point $z = (t, y) = (2, 5)$. What is the maximum likelihood estimate of k?

Exercise 11 (maximum likelihood estimate II)**.** *Consider the structural model $\mathcal{M}$:*

$$x_{i|k} = k \cdot t_i + x_0$$

with $x_0 = 1$, assuming an additive error

$$y_i = x_{i|k} + \eta_i$$

where $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$.
You are given the data points $z = (t_i, y_i)$ with $t = 1, 2$ and $y = 2.5, 5.5$. What is the maximum likelihood estimate of k?

9.2.3 Boundaries and Constraints

The parameter search space is $\mathbb{R}^m$, where m defines the number of distinct parameters for model $\mathcal{M}$. In order to allow efficient parameter search it may be reasonable to contain the search space $\mathcal{K} \subset \mathbb{R}^m$. This can be done by utilizing some prior/extended knowledge in terms of boundaries and relations (constraints) about the system being studied.

Parameter boundaries
Parameter boundaries define the parameter search space, i.e.

$$LB \leq \theta \leq UB,$$

where LB and UB are sets of lower- and upper bounds respectively.

Remark 14. *Unless specified, the typical parameter search space is $\mathbb{R}^m$, including all negative numbers. Quite often the parameters have to be strictly positive.*

Constraints
Constraints define additional relations that decrease (constrain) the size of the parameter search space. These can come in a set of non-linear or linear functions.

Implementation	
bound	transformation
$\tilde{\theta} > 0$	$\tilde{\theta} = \exp(\theta)$
$LB \leq \tilde{\theta} \leq UB$	$\tilde{\theta} = LB + \frac{UB-LB}{2}(1 + \sin\theta)$
$\tilde{\theta} \leq UB$	$\tilde{\theta} = UB + (1 - \sqrt{1+\theta^2})$
$LB \leq \tilde{\theta}$	$\tilde{\theta} = LB - (1 - \sqrt{1+\theta^2})$

Table 9.1: $\tilde{\theta}$ bounded parameter, θ unbounded parameter.

Example 33. *Consider the linear program*

$$\mu = \min_{\theta} Y \cdot \theta \tag{9.7}$$

$$w.r.t. \qquad \mathcal{Z} \cdot \theta \geq \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{9.8}$$

where Y is an $1 \times m$ matrix and $\mathcal{Z}$ is a $n \times m$ matrix and here θ is a $m \times 1$ dimensional vector.

Example 34. *If your constraints refer to some equality (equality constraint), e.g. $\sum p_i = 1$, where p_i is a probability, lagrangian multiplier techniques can be used to deduce useful solutions to the problem.*

Example 35. *Assume you want to obtain an optimal parameter set of the predator-prey model, such that the accordance with some data set is maximized. At the same time the model should have a stable center fixed point. Remembering the fixed point analysis, you could add the corresponding conditions of the fixed point analysis as constraints.*

9.2.4 Prior information on parameters

In maximum likelihood based approaches to determining a unique parameter set θ^*, these can be initial parameter guesses/starting parameters θ_0,but also constraints & parameter bounds and regularizations.
In Bayes'ian approaches the *prior* is usually a initial multivariate probability distribution (pdf) for the location of parameters $p(\theta)$ with $\theta \in \Sigma$ and $\Sigma = \mathbb{R}^m$, unless otherwise specified (bounds and constraints).

9.2.5 Population Approaches (optimal)

Individual Parameters; Example: Viral kinetics

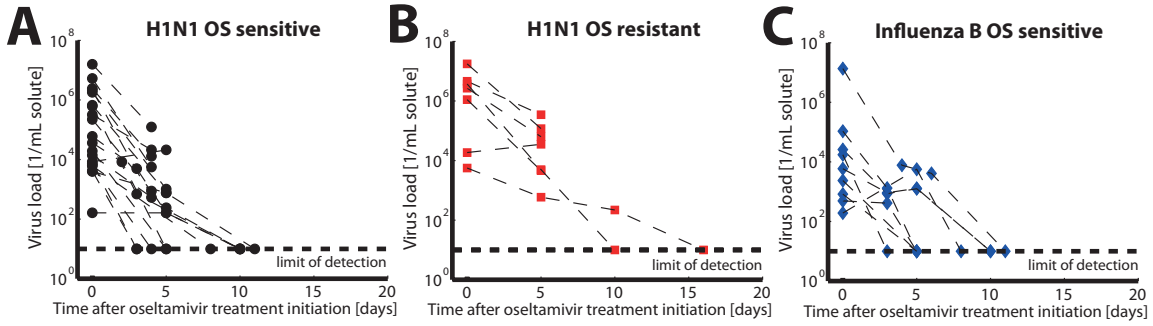


Figure 9.1: Virus load decay in influenza infected children after application of oseltamivir (tamiflu).

Model:

$$\begin{aligned} y_{ij} &= x_{ij|\mathcal{M},\theta_j} + \eta_{ij} \\ x_{ij|\mathcal{M},\theta_j} &= x_{0j} \cdot e^{-t_i \cdot k_j} \end{aligned}$$

with $\theta = \{(x_{0j}, k_j)\}$

- pro: *unbiased* statistics of population heterogeneity.
- contra: only possible if the number of data points per patient allow to infer dynamics (dense sampling/little "noise").
- usually *sparse* sampling.

Population Estimation Methods

- (naive) averaging
  $\bar{y}(t_i) = \frac{1}{n_j} \sum_{j=1}^{n_j} y_j(t_i)$
- (naive) pooling
  $\theta^* = \arg\min_{\theta} \sum_j \sum_i \left(\frac{x(t_i|\theta) - y_j(t_i)}{\omega_{ij}} \right)^2$
  $\vdots$
- Nonlinear Mixed Effects Approach

$$\begin{aligned} y_{ij} &= x_{ij|\mathcal{M},\theta_j} + \eta_{ij} \\ \theta_j &= \mathbb{E}(\theta) + \epsilon_j \end{aligned}$$

  $\Rightarrow$ estimates "typical" parameters/population characteristics $\mathbb{E}(\theta)$ (also called hyper-parameters) in empirical Bayes' method. Expensive to solve ... e.g. EM algorithm.

9.2.6 Ill-posed problems

A problem is well-posed if (Jacques Hadamard, 1902):

- A solution exists
- The solution is unique
- The solution's behavior changes continuously with the initial conditions.

A problem that is not well-posed is called ill-posed. Ill-posed problems need reformulation for numerical treatment.

9.2.7 Ill-conditioned problems, regularisation & model selection

A well-posed problem may still be ill-conditioned, meaning that

- a small error in the initial data can result in much larger errors in the answers.

I.e. Small changes in the model $\mathcal{M}$ or the observations z lead to large differences in the solution θ. Ill-conditioned problems have a large condition number.

Exercise 12 (Ill-conditioned)**.**

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix}}_{coeff.\ matrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 4 \\ 7.999 \end{bmatrix}}_{sol.\ vector} \tag{9.9}$$

with $\theta_1^* = 2$, $\theta_2^* = 1$. Now perturb the coefficient matrix on the left side or the solution on the right side and recompute θ.
Relative condition number:

$$||J_\theta^*|| \cdot \left(\frac{||\theta^*||}{||y||} \right)^{-1} \tag{9.10}$$

where J_θ^* is the Jacobian matrix (first-order approximation of the solution of the inverse problem around its optimum with respect to changes in the data y).

Exercise 13 (well-conditioned)**.** *Now, let*

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix} \tag{9.11}$$

with $\theta_1^* = 2$, $\theta_2^* = 1$.

Problem ill-posed

- parameter identifiability.
  - structural (often ill-posed)
  - practical
- regularization: embedding of an extra term into the performance criterium that is not related to prediction accuracy, but enables a meaningful parameter estimation.

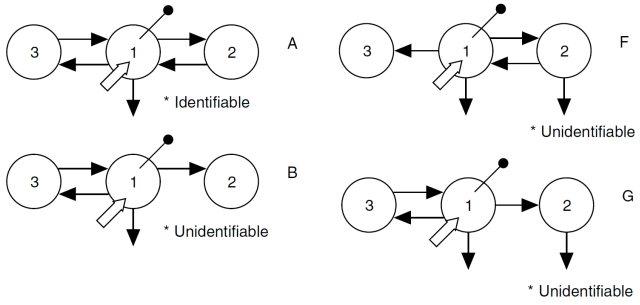


Figure 9.2: Examples of structurally non-identifiable pharmacokinetic models, taken from Bonate P.L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation* (2006), Springer.

Regularizations

$$\text{model accordance} + w \cdot \mathcal{R} \tag{9.12}$$

where $0 \leq w \leq \infty$ is a regularization weight.

L_0	minimize number of non-zero parameters	$\mathcal{R}_j = \begin{cases} k \text{ if } \theta_j \neq 0 \\ 0 \; else \end{cases}$	non-convex
L_1	minimize absolute deviation from 0	$\mathcal{R} = \sum_j \lvert\theta_j\rvert$	convex, technically optimizes L_0, lasso
L_2	minimize large deviations from 0	$\mathcal{R} = \sum_j (\theta_j)^2$	convex, Tikhonov regularization/ridge regression
$\vdots$	...	...	...
L_∞	penalizes the maximum	$\mathcal{R} = \max(\theta)$	non-convex

Table 9.2: Some regularizations. Combinations are possible (e.g. *elastic net* regularization etc...)

Chapter 10

Inverse problems II

Recap: Modelling inverse problems

1. Modelling measurement errors
   → Likelihood function $\mathcal{L}_y$

   $$\mathcal{L}_y = p(y|\theta, \mathcal{M}) = \prod p(y_i|\theta, \mathcal{M})$$

   since $p(y_i|\theta) < 1$, it is numerically more convenient to work with the (negative) log-likelihood instead, i.e.

   $$\ell_y = -\log\ p(y|\theta) = -\sum \log\ p(y_i|\theta)$$

   Example:

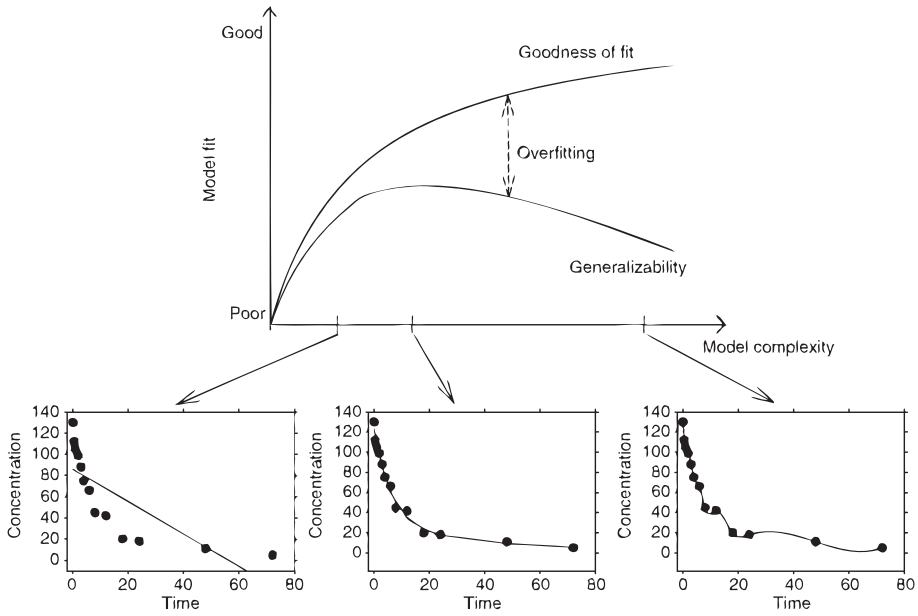
   $$y_i = x_i|\mathcal{M}, \theta + \eta_i \ \text{ with } \ \eta_i \sim \mathcal{N}(0, \sigma_i^2)$$

   $$\Rightarrow p(y_i|\theta) = \frac{1}{\sqrt{2\sigma_i^2\pi}} \exp\left[-\frac{(x_{i|\mathcal{M},\theta} - y_i)^2}{2\sigma_i^2}\right]$$

2. Parameter boundaries & constraints

3. Ill-posed & ill-conditioned problems

4. Regularisation: Can enable parameter estimation for ill-posed (over parameterised) problems
   Most frequently used: L_1 & L_2 regularization



10.0.8 Model selection

Structural model

Model has to be simple, yet sufficient. Balance between goodness-of-fit and generalisability.
Overfitting = fitting to noise.

How to assess generalisability: cross-validation methods, bootstrap.

Model comparison

- likelihood ratios
- AIC: Akaike Information Criterion

$$\mathrm{AIC} = 2m - 2m \log\left(\mathcal{L}_y^*\right)$$

  -select model with lowest AIC

- BIC: Baysian Information Criterion

$$\mathrm{BIC} = \log(n) \cdot m - 2m \log\left(\mathcal{L}_y^*\right)$$

  -select model with lowest BIC

Above, m denotes the number of free parameters in the model under investigation and $\mathcal{L}_y^*$ denotes the maximum likelihood estimate (at the optimal parameter set θ^*). The parameter n denotes the number of data points.

Further reading: e.g. Bonate P.L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation* (2006), Springer.

Error model

visual/graphical

1. plot predicted value vs. errors e:
   $\Rightarrow$ there should be no systematic trend

2. histogram of errors
   $\Rightarrow$ should be roughly normal distributed

3. plot observation/data vs. prediction (visual predictive check, VPC)
   $\Rightarrow$ data should be evenly spread around the line of unity.

Regression analysis

- concordance coefficient (Lin, L.I. (1989). *Biometrics* 45: 255).
  $\Rightarrow$ measure of deviation from the diagonal (line of unity)

10.1 Frameworks

Two frameworks are apparent: Likelihood-based approaches and Bayes'ian approaches. Both frameworks can be used to derive point estimates/an optimal parameter set θ^*. The derived optimal parameters are called *maximum likelihood-*, or a *maximum/mean a posteriori estimate*, depending on the approach. Moreover, it is possibly to deduce an entire distribution of parameters from the Bayes'ian approach, which also tells much about the **certainty/uncertainty of parameter estimates**.

10.1.1 Likelihood-based approaches

The *objective* in a *maximum likelihood*-based parameter estimation approach is

$$\theta^* = \min_{\theta} \ell_y + w \cdot \mathcal{R} \tag{10.1}$$

$$s.t. \quad LB \leq \theta \leq UB \text{ (parameter boundaries)} \tag{10.2}$$

$$s.t. \quad \mathcal{F}(\theta) \leq 0,\ \mathcal{G}(\theta) = 0 \text{ (constraints)} \tag{10.3}$$

where θ^* denotes the maximum likelihood parameter estimate and $\mathcal{F}$, $\mathcal{G}$ are inequality and equality constraints respectively. ℓ_y denotes the negative log-likelihood, w is a regularisation weight and $\mathcal{R}$ denotes the regularisation.

Relation to least-squares minimization

Assume an i.i.d. additive gaussian error model, i.e.

$$y_i = x_{i|\mathcal{M},\theta} + \eta_i$$

with $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\, i$. It follows that

$$\begin{aligned}
\mathcal{L}_y &= \prod_{i=1}^{N} \frac{1}{\sqrt{2 \cdot \sigma^2 \pi}} e^{-\frac{(y_i - x_i)^2}{2\sigma^2}} && (10.4)\\
&= \frac{1}{(2 \cdot \sigma^2 \pi)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i)^2\right) && (10.5)\\
\Rightarrow \ell_y &= \underbrace{-\log\left(\frac{1}{(2\pi\sigma^2)^{N/2}}\right)}_{C_2} + \underbrace{\frac{1}{2\sigma^2}}_{C_2} \sum_{i=1}^{N} (y_i - x_i)^2 && (10.6)\\
&\propto \sum_{i=1}^{N} (y_i - x_i)^2 && (10.7)
\end{aligned}$$

which is the **ordinary least-squares** criterium and ℓ_y denotes the negative log likelihood. Note that above, we used the short-hand notation x_i instead of $x_{i|\mathcal{M},\theta}$.

Exercise 14. *Assume a proportional gaussian error model, i.e.*

$$y_i = x_{i|\mathcal{M},\theta}(1 + \eta_i)$$

with $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\, i$. This can be re-written as

$$y_i = x_{i|\mathcal{M},\theta} + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}(0, x_i \cdot \sigma^2)$.

10.1.2 Bayes'ian

Bayes' theorem states that

$$\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta) \cdot p(\theta)}{p(y)} && (10.8)\\
&\propto p(y|\theta) \cdot p(\theta) && (10.9)\\
&= \mathcal{L}_y \cdot p(\theta) && (10.10)
\end{aligned}$$

where $p(\theta|y)$ is the posterior (probability of the parameter after seeing the data) and $p(\theta)$ is the prior belief of the parameter distribution and $p(y) = \int p(y|\theta) \cdot p(\theta)dp$ is called the evidence (a normalization constant).
Maximum a posteriori estimate (MAP):

$$\theta^+ = \max_\theta p(\theta|y) \tag{10.11}$$

In practice, **boundaries, constraints and regularisation** may be **enforced by** appropriate **prior selection** in sampling-based approaches.

Example 36 (L_2 regularisation)**.** *Choose the prior $p(\theta)$ to be gaussian centered around 0.*

$$p(\theta|y) \propto \mathcal{L}_y \cdot p(\theta) = \mathcal{L}_y \cdot \prod_j^m \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[-\frac{(0-\theta_j)^2}{2\sigma^2}\right] \tag{10.12}$$

$$\Rightarrow -\log\left(\mathcal{L}_y \cdot p(\theta)\right) \propto -\ell_y + w\sum_j (\theta_j)^2 \tag{10.13}$$

Exercise 15 (L_1 regularisation)**.** *Show that the Bays'ian prior in L_1 regularisation corresponds to a 0-centered Laplace distribution.*

Chapter 11

Numerical optimization

Recap: Model selection

A. structural model

- model has to be simple, yet sufficient
  tradeoff: goodness of fit $\leftrightarrow$ generisability
  too complex model (too many parameters) = fitting to noise = not generalisable
- As a result, the model may be *descriptive*, not mechanistic

B. measurement error model

- diagnostic plots

C. measures for model checking (likelihood-ratio test, AIC, BIC)

Recap: Inference frameworks & philosophies

1. Maximum likelihood (ML) based approaches

   Goal: Deduce an optimal parameter set θ^*

   When the measurement error is *gaussian*, the maximum likelihood estimate is the solution of a least-squares problem.
   Examples:

   a) measurement error η_i is additive and *gaussian* with identical variance for all data points, i.e. $\eta_i \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$

   $$\Rightarrow -\log\left(\mathcal{L}_y(\theta)\right) \propto \sum_i (x_i - y_i)^2 = g(\theta) \text{ (ordinary least squares)}$$

   where we used x_i as a short form for $x_{i|\mathcal{M},\theta}$ and where $g(\theta)$ is called the objective function.

b) measurement error η_i is additive and *gaussian* with distinct variance for each data point, i.e. $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$

$$\Rightarrow -\log\left(\mathcal{L}_y(\theta)\right) \propto \sum_i \frac{(x_i - y_i)^2}{\sigma_i^2} = g(\theta) \text{ (weighted least squares)}$$

c) measurement error ε_i is proportional and *gaussian* with identical variance, i.e. $y_i = x_i(1+\varepsilon)$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)\ \forall\ i$. This can be transformed into an additive error model $y_i = x_i + \eta_i$ with $\eta_i \sim \mathcal{N}(0, x_i \cdot \sigma^2)$

$$\Rightarrow -\log\left(\mathcal{L}_y(\theta)\right) \propto \sum_i \frac{(x_i - y_i)^2}{x_i} = g(\theta) \text{ (weighted least squares)}$$

2. Bayes'ian inference

Goal: Deduce a parameter distribution/posterior $p(\theta^*|y)$

- utilize prior information = probability that the parameter is located somewhere $p(\theta)$
  example: regularisation as prior information.

11.1 Algorithms for solving maximum likelihood based inference

derivative-free	gradient-based	stochastic
simplex	steepest descent Gauss-Newton Levenberg-Marquart trust region	genetic algorithm Metropolis-Hastings simulated annealing
	expectation-maximization (EM)	
general, not very accurate	require 1st (2nd) derivative to exist, converge fast, locally	some asymptotically find the global optima, slow, inaccurate
no guarantee that **global** optimum has been found		

Table 11.1: Examples of algorithms frequently used to solve ML-based inference problems.

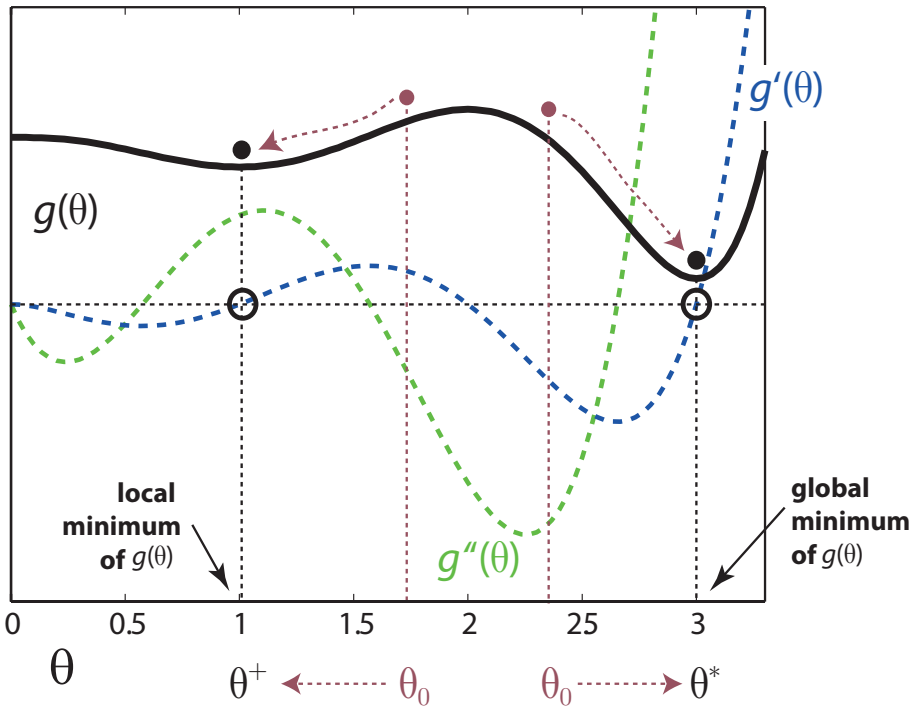


Figure 11.1: An exemplary objective function $g(\theta)$ (solid black line) for a one-parameter model and its first and second derivative in (blue and green dashed lines). The horizontal thin dashed line marks 0 and the black circles mark the values of (θ^+, θ^*) that satisfy the conditions for a local optimum. Depending on where the initial parameter guess θ_0 lies (vertical red dashed lines), the algorithm will either converge to a local- θ^+ or a global optimum θ^*.

11.1.1 Sufficient conditions for a (local) optima

For a single variable case (one parameter), sufficient conditions for a (local) optimum in $g(\theta)$ are:

$$g'(\theta) = 0 \tag{11.1}$$
$$g''(\theta) > 0 \tag{11.2}$$

where $g'(\theta)$ and $g''(\theta)$ are the first- and second derivative of the objective function with respect to the parameter θ.
Fig. 11.1 shows that there is no way for any gradient based algorithm to determine whether an optimum is local or global. For this reason, in practice, one performs a a multi-start optimisation, i.e. repeating the process for random choices of the initial parameter(s) θ_0.

11.1.2 Newton's method

We briefly discuss the Newton method to motivate the Gauss-Newton method for optimizing a multi-variable objective function.
Consider an objective function $g(\theta)$ only depending on a single parameter, i.e. $\theta \in \mathbb{R}^m$ with $m = 1$. Now, perform a second order Taylor expansion of g around θ.

$$g(\theta + \Delta\theta) \approx g(\theta) + g'(\theta)\Delta\theta + \frac{g''(\theta)}{2!}(\Delta\theta)^2$$

We want to find $\Delta\theta$, such that $g'(\theta + \Delta\theta) = 0$ (a stationary point).

$$\begin{aligned} 0 &= \frac{d}{d\Delta\theta}\left(g(\theta + \Delta\theta) \approx g(\theta) + g'(\theta)\Delta\theta + \frac{g''(\theta)}{2}(\Delta\theta)^2\right) && (11.3)\\ &= g'(\theta) + g''(\theta)\Delta\theta && (11.4)\\ \Rightarrow\ \Delta\theta &= -\frac{g'(\theta)}{g''(\theta)}. && (11.5) \end{aligned}$$

Implementation into an algorithm to find an optimal parameter:

$$\theta^{s+1} = \theta^s + \Delta\theta = \theta^s - \frac{g'(\theta^s)}{g''(\theta^s)},$$

where θ^s is the current parameter in step s of the algorithm and θ^{s+1} is a new parameter proposal.
The generalisation of Newton's method to a multi parameter setting, i.e. $\theta \in \mathbb{R}^m$ with $m \geq 1$ is

$$\theta^{s+1} = \theta^s + H^{-1} \cdot \nabla g(\theta^s),$$

where $\nabla g(\theta^s) = J^T W(x - y)$ is the gradient of the objective function with $n \times m$ Jacobian matrix $J_{ij} = \left[\frac{\partial x_i}{\partial \theta_j}\right]$ and $W = \begin{pmatrix} \ddots & & 0 \\ & w_i & \\ 0 & & \ddots \end{pmatrix}$ being an $n \times n$ matrix with the least squares weights on its diagonal and zeros everywhere else and $x, y \in \mathbb{R}^n$ are n-dimensional vectors of the model predictions and the corresponding data points. H denotes the Hessian matrix with $H_{jk} = \left[\frac{\partial^2 g}{\partial\theta_j \partial\theta_k}\right]$.
The Gauss-Newton algorithm makes the approximation $H \approx J^T W J$ (more below).

11.1.3 Gauss-Newton method

Preliminaries

Let us consider the general form of a least squares objective function

$$g(\theta) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{w^{-1}}$$

with

- $w^{-1} = 1$ for additive gaussian error with equal variance for all data points $\eta_i \sim \mathcal{N}(0, \sigma^2)$
- $w^{-1} = \sigma_i$ for additive gaussian error with distinct variance for each data point $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$
- $w^{-1} = x_i$ in case of a gaussian proportional error of the form $\eta_i \sim \mathcal{N}(0, \sigma^2 \cdot x_i)$.

$$\begin{aligned} g(\theta) &= \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{w^{-1}} \\ &= (x-y)^T W (x-y) \\ &= x^T W x - 2x^T W y + y^T W y \end{aligned} \tag{11.6}$$

Step-by-step derivation

The method makes a linear approximation to the model function x (first order Taylor approximation) around a parameter set θ:

$$x(\theta + \Delta\theta) \approx x(\theta) + J\Delta\theta,$$

where the Jacobian J is an $n \times m$ matrix with entries $J_{ij} = \left[\frac{\partial x_i}{\partial \theta_j}\right]$, as above.
Now, substitute the above equation into eq. (11.6).

$$\begin{aligned} g(\theta + \Delta\theta) &= (x + J\Delta\theta)^T W (x + J\Delta\theta) - 2(x + J\Delta\theta)^T W y + y^T W y && (11.7) \\ &= x^T W (x + J\Delta\theta) + (J\Delta\theta)^T W (x + J\Delta\theta) - 2x^T W y \\ &\quad -2(J\Delta\theta)^T W y + y^T W y && (11.8) \\ &= x^T W x + x^T W J\Delta\theta + (J\Delta\theta)^T W x + (J\Delta\theta)^T W (J\Delta\theta) \\ &\quad -2x^T W y - 2(J\Delta\theta)^T W y + y^T W y && (11.9) \\ &= x^T W x + (J\Delta\theta)^T W (J\Delta\theta) \\ &\quad -2x^T W y + y^T W y + 2(x-y)^T W J\Delta\theta && (11.10) \end{aligned}$$

where we used $(J\Delta\theta)^T W x = x^T W J\Delta\theta$ and $2(J\Delta\theta)^T W y = 2y^T W (J\Delta\theta)$.
The goal is to find $\Delta\theta$, such that $\frac{\partial g(\theta+\Delta\theta)}{\partial \Delta\theta} = 0$. Taking the partial derivatives and setting the left side to zero, we get

$$\begin{aligned} 0 &= 2(J\Delta\theta)^T W J + 2(x-y)^T W J && (11.11) \\ &= \Delta\theta^T J^T W J + (x-y)^T W J && (11.12) \\ \Rightarrow J^T W J \Delta\theta &= -J^T W (x-y) && (11.13) \\ \Delta\theta &= -\left(J^T W J\right)^{-1} J^T W (x-y). && (11.14) \end{aligned}$$

The parameter update in the Gauss-Newton method is therefore

$$\theta^{s+1} = \theta^s - \left(J^T W J\right)^{-1} J^T W (x - y). \tag{11.15}$$

Exercise 16 (Algorithm)**.** *Implement the Gauss-Newton method to estimate the parameters k_1 and k_2 of the model*

$$x_i = k_1 \cdot e^{k_2 \cdot t_i}$$

with $t = (1, 2.5, 5, 7.5, 10)^T$ and $y = (2.98, 4.41, 8.44, 16.17, 30.97)^T$, assuming an additive gaussian measurement error $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$ with $\sigma^2 = (2.27, 2.26, 1.14, 1.70, 0.23)$. Stop the algorithm, if a maximum of 30 updates have been exceeded, or the parameters change by less than $\epsilon = 10^{-8}$, i.e. $||\theta_j^{s+1} - \theta_j^s||_1 = \sum_{j=1}^m |\theta_j^{s+1} - \theta_j^s| < \epsilon$. Let your initial parameter guess be $\theta_0 = (9.38, 0.96)^T$.

Chapter 12

Bayesian Inference

12.1 Non-Identifiability

Let us try to find the maximum likelihood of an observation which comes from a Poisson distribution.

Exercise 17. *Let $\lambda, \delta \in \mathbb{R}_+$ be the parameters which we want to estimate. Let $y \in \mathbb{N}$ be the data which we wish to fit and $t > 0$ the time at which the data was observed. Let us consider the model ,*

$$x_{\lambda,\delta}(t) = \frac{\lambda}{\delta}(1 - e^{-\delta t}).$$

We choose the likelihood of the observe y at time t by our given model to be

$$\mathcal{L}((y,t)|(\lambda,\delta)) = \frac{(x_{\lambda,\delta}(t))^y e^{-x_{\lambda,\delta}(t)}}{y!}.$$

Find the λ, δ which maximise the likelihood of observing the data y at time t.

Proof. To find the maximum likelihood, we take the log of the likelihood and find the λ and δ which set the derivative to zero. For ease of notation, we denote $x_{\lambda,\delta}(t)$ as $x(t)$.

$$\log(\mathcal{L}((y,t)|(\lambda,\delta))) = y\log(x(t)) - x(t) - \log(y!).$$

Taking the derivative with respect to λ gives us

$$\begin{aligned} \frac{\partial \log(\mathcal{L}((y,t)|(\lambda,\delta)))}{\partial \lambda} &= y\frac{1}{x(t)}\frac{\partial x(t)}{\partial \lambda} - \frac{\partial x(t)}{\partial \lambda}, \\ &= \frac{\partial x(t)}{\partial \lambda}\left(\frac{y}{x(t)} - 1\right). \end{aligned} \tag{12.1}$$

Similarly,

$$\frac{\partial \log(\mathcal{L}((y,t)|(\lambda,\delta)))}{\partial \delta} = \frac{\partial x(t)}{\partial \delta}\left(\frac{y}{x(t)} - 1\right).$$

Hence, the condition for which the derivative is zero in both cases is:

$$\frac{y}{x(t)} - 1 = 0, \tag{12.2}$$

$$x(t) = y, \tag{12.3}$$

$$\frac{\lambda}{\delta}(1 - e^{-\delta t}) = y. \tag{12.4}$$

So we have two unknown and one equation to solve. We can choose either λ or δ to be our free parameter, then

$$(\lambda, \delta) = \left(\frac{y\delta}{1 - e^{-\delta t}}, \delta \right),$$

are all points which maximise the likelihood for observing y at time t. □

Remark 15. *If we were to run a maximisation algorithm what would we find?*

12.2 Bayesian Inference

The following names are important to keep in mind. Let Σ be the parameter space. Let Ω be the data space.

- Parameter (θ) : A vector or scaler in Σ.
- Data (y) : A set of points from the data space Ω.
- Prior (ν) : A distribution over the parameter space Σ.
- Likelihood ($\mathcal{L}$): A conditional probability distribution, which gives the probability of observing the data given a parameter.
- Evidence ($\mathcal{E}$): A distribution over the data space Ω.
- Posterior (π): The conditional probability distribution, which gives the probability of observing the parameter, given the data.

Using Bayes' Theorem we can deduce that

$$\pi(\theta|y) = \frac{\nu(\theta)\mathcal{L}(y|\theta)}{\mathcal{E}(y)}. \tag{12.5}$$

12.2.1 Prior

The posterior is the a probability distribution which describes the probability of seeing the parameter given the data. Then like in the case of the likelihood, one can ask the question: what is the maximum a posteriori estimate given some data?.

Similarly with the likelihood, we take the log of the posterior distribution and reduce,

$$\max_{\theta} \log(\pi(\theta|y)) = \max_{\theta} \left(\log(\nu(\theta)) + \log(\mathcal{L}(y|\theta)) - \log(\mu(y))\right). \tag{12.6}$$

Taking the derivative with respect to θ gives us,

$$\frac{\partial \log(\pi(\theta|y))}{\partial \theta} = \underbrace{\frac{\partial \log(\nu(\theta))}{\partial \theta}}_{??} + \underbrace{\frac{\partial \log(\mathcal{L}(y|\theta))}{\partial \theta}}_{likelihood}.$$

Let us consider a few different priors and observe what happens to the above expression.

Case 1: Let $\nu(\theta) = \frac{1}{2b} e^{-\frac{|\theta - a|}{b}}$ be the Laplace distribution with a and b being the distribution parameters. If we consider the log of this function, then we get,

$$\log(\nu(\theta)) = -\log(2b) - \frac{|\theta - a|}{b}.$$

Substituting this into (12.6), gives us,

$$\log(\pi(\theta|y)) = \frac{|\theta - a|}{b} + \log(\mathcal{L}(y|\theta)) + C.$$

Let us consider that we have data $y_1, \ldots, y_N$ and a model prediction $x_1, \ldots, x_N$ which are a function of the parameters. Then if we consider the likelihood to be Gaussian around x_i with variance σ^2, then the above equation reduces to

$$\log(\pi(\theta|y)) = \frac{|\theta - a|}{b} + \sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma} + C.$$

If we choose $h = 0$, we can see that the maximum a posteriori problem is a least squares problem with an ℓ_1 regulariser.

Case 2: Let $\nu(\theta) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(\theta - a)^2}{2b}}$ be the Gaussian distribution with a and b being the distribution parameters. If we take the log of the Gaussian distribution, we get that,

$$\log(\nu(\theta)) = -\frac{1}{2}\log(2\pi b) - \frac{(\theta - a)^2}{2b}.$$

Substituting this into (12.6), gives us,

$$\log(\pi(\theta|y)) = -\frac{(\theta - a)^2}{2b} + \log(\mathcal{L}(y|\theta)) + C.$$

Hence, when we choose a Gaussian prior distribution and set $a = 0$, it is equivalent to solving a maximum likelihood method with an ℓ_2 regulariser.

It is easy to see in this context that the prior is used as a way to make the parameters more identifiable. They have a natural translation across into the classical least squares setting. However, we can see that the notion of prior is very general. From here on in, we will only consider solving for the posterior distribution and not just point estimates.

Chapter 13

Posterior Exploration

In the Bayesian setting we are interested posterior distribution, or statistics of it, rather than just point estimates. In practise, the posterior distributions have many modes and simply considering estimates gives only part of the whole picture. There are two strategies for empirically constructing the posterior distribution: Likelihood based methods and Likelihood Free Methods. The names give away the critical characteristic of the methods. We will explore one method from each class to gain some general institution.

13.1 MCMC

Markov Chain Monte Carlo (MCMC) is a great numerical scheme for constructing approximations using a random walker strategy. In this section, we are interested in the theory behind MCMC so that we can understand how to use it for constructing accurate posteriors.

Let x, y be two states in the state space Ω. Let X_t be a Markov Chain with $T(x \to y)$ being the transition probability of going from x to y. For finite state space, we can think of T as a matrix. Let us assume that X_t is well behaved (irreducibility, aperiodicity) and has a stationary distribution pi. That is,

$$X_\infty \sim \pi, \text{ and } \sum_x T(x \to y)\pi(x) = \pi(y).$$

If we think of T as a matrix, then it is similar to saying $\pi T = \pi$. π is the stationary distribution. This tell us that, no matter where we start, if we keep sampling the transition probability, we will end up with a sample of π. Let $x^{(1)} \sim X_1,\ x^{(2)} \sim X_2, \ldots, x^{(?)} \sim \pi$.

The theory tells us that if any distribution we want to sample, is a stationary distribution of some Markov chain, then we can transitions of the Markov chain and "eventually", the state of the chain is a sample of our distribution. Now the question is, can we construct such a Markov chain for any given distribution which I wish to sample?

13.1.1 Detailed Balance

We say the transitional probability satisfies detailed balance when for any $x, y \in \Omega$,

$$T(x \to y)\pi(x) = T(y \to x)\pi(y).$$

13.2 Metropolis–Hastings

Metropolis-Hastings proposed if we want to compute samples of a distribution π, then we can constructor a transition probability using $Q(x, y) = \mathcal{N}(y, \sigma)$, $\sigma > 0$. Choose $\tau \in U(0, 1)$, we transition from x to y if

$$\tau > \min\left(1, \frac{\pi(y)Q(x, y)}{\pi(x)Q(y, x)}\right).$$

We can test to see if choosing our transition probability really us our target probability distribution as the stationary distribution.

Proof. Let us considering transition from x to y. Then using detailed balance, we can deduce that,

$$\begin{aligned} T(x \to y)\pi(x) &= \pi(x)Q(y, x)\min\left(1, \frac{\pi(y)Q(x, y)}{\pi(x)Q(y, x)}\right), && (13.1)\\ &= \min\left(\pi(x)Q(y, x), \pi(y)Q(x, y)\right), && (13.2)\\ &= \pi(y)Q(x, y)\min\left(1, \frac{\pi(x)Q(y, x)}{\pi(y)Q(x, y)}\right), && (13.3)\\ &= \pi(y)T(y \to x). && (13.4) \end{aligned}$$

To prove that π is a stationary distribution, we have that,

$$\begin{aligned} \sum_x T(x \to y)\pi(x) &= \sum_x \pi(y)T(y \to x), \text{ using (13.4)}\\ &= \pi(y)\sum_x T(y \to x),\\ &= \pi(y). \end{aligned}$$

Hence π is a stationary distribution for the Metropolis–Hastings algorithm. □

Remark 16. *What is an appropriate σ?*

What does this mean in terms of our posterior? Since, we want samples of the posterior, we simply substitute this into the Metropolis–Hastings algorithm. Consider the transition from the parameter θ_1 to θ_2, with $Q(\cdot, \cdot)$ as the transition probability distributed $\mathcal{N}(\cdot, \sigma)$. Then the acceptance criterion looks like this,

$$\tau > \min\left(1, \frac{\pi(\theta_2|y)Q(\theta_1,\theta_2)}{\pi(\theta_1|y)Q(\theta_2,\theta_1)}\right),$$

By subbing in the definitions (12.5),

$$\frac{\pi(\theta_2|y)}{\pi(\theta_1|y)} = \frac{\frac{\nu(\theta_2)\mathcal{L}(y|\theta_2)}{\mathcal{E}(y)}}{\frac{\nu(\theta_1)\mathcal{L}(y|\theta_1)}{\mathcal{E}(y)}}.$$

Since the data is the same for both, the evidence term disappears and we get that,

$$\frac{\pi(\theta_2|y)}{\pi(\theta_1|y)} = \frac{\nu(\theta_2)\mathcal{L}(y|\theta_2)}{\nu(\theta_1)\mathcal{L}(y|\theta_1)}$$

Then a Metropolis–Hastings step for sampling the posterior distribution looks like

$$\tau > \min\left(1, \frac{\nu(\theta_2)\mathcal{L}(y|\theta_2)Q(\theta_1,\theta_2)}{\nu(\theta_1)\mathcal{L}(y|\theta_1)Q(\theta_2,\theta_1)}\right) \tag{13.5}$$

13.3 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a likelihood free method. It means that you do not explicitly compute the likelihood, but rather take samples from it and compare those simulated data with the real data. We simply compute trajectories for many different parameters and keep the parameters which reproduced, through simulations, a proportion of the data. It was shown that accuracy of the posterior converges asymptotically with the number of simulations run for each parameter choice.