



STK4051/STK9051

Computational Statistics

Riccardo De Bin

`debin@math.uio.no`

Outline of the lecture

- Introduction
- Review
 - Notation
 - Taylor's Theorem
 - Likelihood Inference
 - Bayesian Inference
 - Statistical Limit Theory
 - Markov Chains (later in the course)
- Optimization and Solving non-linear Equations
 - Optimization
 - Univariate Problems
 - Bisection method
 - Convergence criteria

Introduction to the course: Content

This course:

- **textbook** “Computational Statistics”, by Givens & Hoeting (2013);
- chapters we will (try to) **cover**: <https://www.uio.no/studier/emner/matnat/math/STK4051/v26/pensum.html>;
- we may **not follow the book's order**, the plan will be posted step by step at <https://www.uio.no/studier/emner/matnat/math/STK4051/v26/lectures.html>;
- the **exercises** for the following week will be posted at <https://www.uio.no/studier/emner/matnat/math/STK4051/v26/exercises.html>;
- programming language: the book prefers **R** (I also do), but it is not mandatory.

Introduction to the course: Organization

We meet twice a week:

- **Mondays** from 14:15 to 16:00 at VB Aud. 4: **exercises**;
- **Tuesdays** from 14:15 to 16:00 at VB Aud. 4: **theory**;
- 15 minutes break between the hours.

Mandatory assignments/exam:

- **mandatory assignments** through Canvas;
 - ▶ first: deliver to pass;
 - ▶ second: evaluate programming competences;
- **exam**: 1st of June from 15:00 to 19:00 at Silurveien 2 Sal 3D;

Review: Mathematical Notation

I will try to use the **book's notation** as much as I can:

- a lower-case letter denotes a variable: x ;
- a lower-case bold letter denotes a vector: $\mathbf{x} = (x_1, \dots, x_p)$;
- an upper-case letter denotes a constant: M ;
- an upper-case bold letter a matrix: $\mathbf{M} = (\mathbf{x}_1 \dots \mathbf{x}_n)^T$;
 - $\mathbf{M}^T$ denotes the transpose of $\mathbf{M}$.
- the other notation is pretty standard: identity matrix, gradient, indicator function, ...

PS: there will be many typos, feel free to point them out.

PSII: in the first chapter the book ignores its own notation (.-.-') and uses a classical statistical one. Let us do the same...

Review: Taylor's Theorem

Orders of convergence: “big oh” $\mathcal{O}()$ and “little oh” $\mathcal{o}()$.

Consider a sequence $\{x_n\}$:

- $\{x_n\}$ is $\mathcal{O}(n^\alpha)$ if $\forall n \in \mathcal{N}, \left| \frac{x_n}{n^\alpha} \right| < A$;
 - ▶ e.g.: $(-1)^n$ is $\mathcal{O}(1)$; $\log(1 + 1/n)$ is $\mathcal{O}(n^{-1})$.
- $\{x_n\}$ is $\mathcal{o}(n^\alpha)$ if $\lim_{n \rightarrow \infty} \frac{x_n}{n^\alpha} = 0$;
 - ▶ e.g.: $\sin(1/n)$ is $\mathcal{o}(n^{-1/2})$.

Note:

- $\mathcal{O}(n) \not\Rightarrow \mathcal{o}(n)$ (sequence may converge to a value $\neq 0$);
- $\mathcal{o}(n) \Rightarrow \mathcal{O}(n)$;
- $\mathcal{O}(n) \Rightarrow \mathcal{O}(n^2)$ and so on: the smallest characterises the best.

Review: Taylor's Theorem

Using functions, as in the book: consider two functions $f(z)$ and $g(z)$ defined on a common interval, with a point z_0 belonging to the interval and $g(z) \neq 0$ in a neighbour of z_0 .

Then:

- $f(z) = \mathcal{O}(g(z))$ if $\exists M : \lim_{n \rightarrow \infty} \left| \frac{f(z)}{g(z)} \right| < M$ for $z \rightarrow z_0$;
 - ▶ e.g.: $(n+1)/(3n^2)$ is $\mathcal{O}(n^{-1})$ (understood for $n \rightarrow \infty$).
- $f(z) = \mathcal{o}(g(z))$ if $\lim_{z \rightarrow z_0} \frac{f(z)}{g(z)} = 0$;
 - ▶ e.g.: $f(x_0 + h) - f(x_0) = hf'(x_0) + \mathcal{o}(h)$ as $h \rightarrow 0$, if f is differentiable in x_0 .

Review: Taylor's Theorem

Taylor's theorem: polynomial approximation of a function f around a given point x_0 .

Consider $f : \mathbb{R} \rightarrow \mathbb{R}$, n -times differentiable at $x_0 \in \mathbb{R}$. Then

$$f(x) = \sum_{i=0}^n \frac{f^{(i)}(x_0)}{i!} (x - x_0)^i + \mathcal{O}(|x - x_0|^n), \quad x \rightarrow x_0$$

There are explicit forms for $R_n = f(x) - \sum_{i=0}^n \frac{f^{(i)}(x_0)}{i!} (x - x_0)^i$. Suppose f has finite $(n+1)$ th derivatives on (a, b) and continuous n th derivatives in $[a, b]$. Then, for any $x_0 \neq x$ in $[a, b]$,

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{(n+1)} \quad \text{for } \xi \text{ between } x \text{ and } x_0$$

and $R_n = \mathcal{O}(|x - x_0|^n)$ as $x \rightarrow x_0$.

Review: Likelihood Inference

Consider $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. random vectors, each with density $f(\mathbf{x}|\theta)$, θ a p -dimensional parameter. The **likelihood function** is

$$L(\theta) = \prod_{i=1}^n f(\mathbf{x}_i|\theta).$$

The maximum likelihood **estimate** $\hat{\theta}$ is the value, given the data, of the maximum likelihood **estimator**, $\hat{\theta}$, the **maximizer** of $L(\theta)$.

Typically it is easier to work with the **log-likelihood function**,

$$\ell(\theta) = \log L(\theta).$$

Maximizing the log-likelihood corresponds to solve the system

$$\ell_{\theta}(\theta) = 0,$$

where $\ell_{\theta}(\theta) = \left(\frac{d\ell(\theta)}{d\theta_1}, \dots, \frac{d\ell(\theta)}{d\theta_p} \right)$ is the **score function**.

Review: Likelihood Inference

The score function usually satisfies the **Bartlett's identities**:

- $E[\ell_{\theta}(\theta)] = 0$;
- $E[\ell_{\theta}(\theta)\ell_{\theta}(\theta)^T] = -E[\ell_{\theta\theta}(\theta)]$

with $-\ell_{\theta\theta}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}$ being the **observed Fisher information** and its expected value $E[-\ell_{\theta\theta}(\theta)]$ called **expected Fisher information**. Both matrices are $p \times p$.

Under regularity conditions,

$$\hat{\theta} \sim N_p(\theta_0, \mathbf{I}(\theta_0)^{-1}) \text{ as } n \rightarrow \infty$$

where the maximum likelihood estimate is plugged in to substitute the true value of the parameter θ_0 .

Review: Bayesian Inference

In **Bayesian inference** the likelihood function is only a part of the game. Inference is based on the **posterior distribution** of θ ,

$$f(\theta|\mathbf{x}) = \frac{f(\theta)L(\theta|\mathbf{x})}{\int f(\theta)L(\theta|\mathbf{x}) d\theta}$$

where

- $f(\theta)$ is the **prior distribution**;
- $\int f(\theta)L(\theta|\mathbf{x}) d\theta$ is a **normalization constant**, often hard to compute $\rightarrow$ part of this course is related to it;
- under regularity conditions, the distribution of the posterior mode tends to $N_p(\theta_0, \mathbf{I}(\theta_0)^{-1})$ as $n \rightarrow \infty$.

Review: Statistical Limit Theory

Review some concepts:

- **Convergence** of X_n :
 - ▶ in **probability**: $X_n \xrightarrow{p} X$ if $\lim_{n \rightarrow \infty} P[|X_n - X| < \epsilon] = 1, \forall \epsilon$;
 - ▶ **almost surely**: $X_n \xrightarrow{as} X$ if $P[\lim_{n \rightarrow \infty} |X_n - X| < \epsilon] = 1, \forall \epsilon$;
 - ▶ in **distribution**: $X_n \xrightarrow{d} X$ if $\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x), \forall x$.
- **laws of the large numbers** ($\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, with x_i i.i.d.):
 - ▶ **weak**: $\bar{X}_n \xrightarrow{p} \mu = E[X_i]$ if $E[|X_i|] < \infty$;
 - ▶ **strong**: $\bar{X}_n \xrightarrow{as} \mu = E[X_i]$ if $E[|X_i|] < \infty$;
 - ▶ both assumes that $Var[X_i] < \infty$.
- **central limit theorem** (as presented in the book):
 - ▶ $X_1, \dots, X_n$ i.i.d. r.v. with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, and $E[\exp\{tX_i\}]$ exists in a neighbourhood of $t = 0$. Then $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0; 1)$ as $n \rightarrow \infty$.

Optimization: general

Optimization means to solve the problem

$$\arg \min_{x \in D} f(x)$$

for x . Here:

- f is called the **objective function**;
- x is the **variable** (usually a vector);
- the domain D is the **feasible set**.

Depending on the **cardinality** of D :

- D **continuous**, dense, and uncountable $\rightarrow$ Chapter 2;
- D **discrete** and countable $\rightarrow$ Chapter 3;
- D may (or may not) define **constraints** on the variables.

Optimization: general

Many problems in statistical inference can be seen as optimization:

- minimize **deviations** between predicted and observed values,
 - ▶ e.g., least squares;
 - ▶ except for simple case, the algorithms are iterative;
 - ▶ independent of the data generating process.
- **maximizing the likelihood**,
 - ▶ **maximization is just minimization of $-f(x)$** ;
 - ▶ dependent on the data generating process;
 - ▶ often the likelihood cannot be maximised analytically.

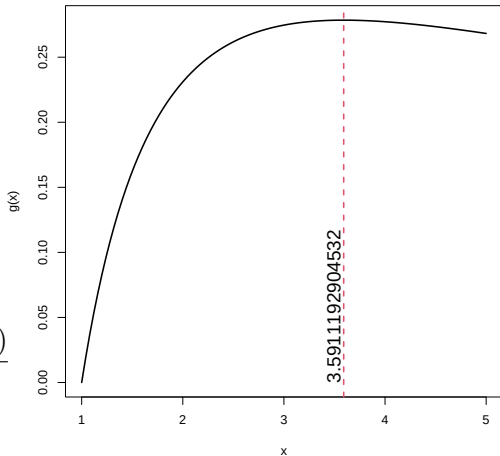
Univariate Problems: Example

We will use for illustration the example of the book. The task is to find the $\arg \max_x$ of the function:

$$g(x) = \frac{\log(x)}{1+x}$$

- no analytical solution
- not so far from 3
- 3 as starting point
- derivative

$$g'(x) = \frac{1 + 1/x - \log(x)}{(1+x)^2}$$



Univariate Problems: Bisection method

Bolzano's Theorem

If:

- $g'(x)$ is continuous on $[a_0, b_0]$;
- $g'(a_0)g'(b_0) \leq 0$;

Then:

- $\exists x^* \in [a_0, b_0] : g'(x^*) = 0$.

Bisection method: shrink the interval $[a_t, b_t] \supset [a_{t+1}, b_{t+1}] \supset \dots$
in which the point of interest lies.

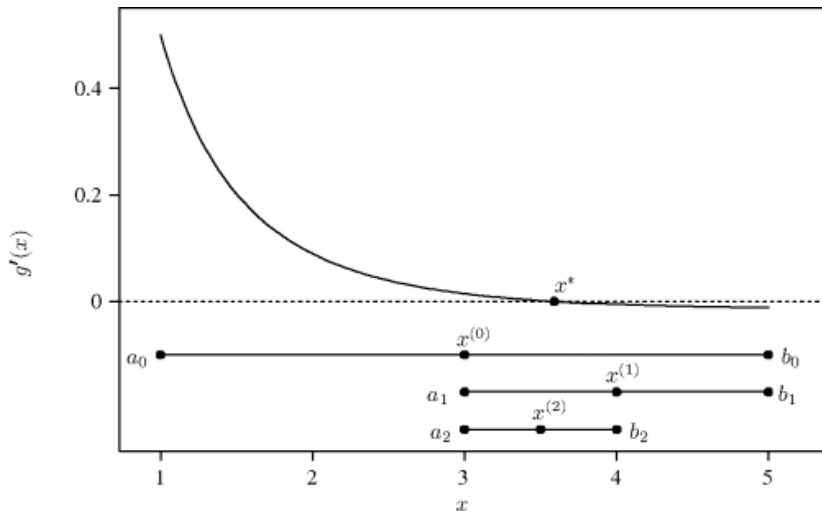
Given $x^{(t)} = (a_t + b_t)/2$. The new interval is

$$[a_{t+1}, b_{t+1}] = \begin{cases} [a_t, x^{(t)}] & \text{if } g'(a_t)g'(x^{(t)}) \leq 0 \\ [x^{(t)}, b_t] & \text{if } g'(a_t)g'(x^{(t)}) > 0 \end{cases}$$

and $x^{(t+1)} = (a_{t+1} + b_{t+1})/2$.

Note: if g has more than one root, the methods find **only one**.

Univariate Problems: Bisection method



(Givens & Hoeting, 2013, figure 2.2)

Univariate Problems: Convergence criteria

When to **stop**?

- we cannot run the procedure infinitely;
- we need a **stopping rule**:
 - ▶ define a **stopping criterion**;
 - ▶ **at each iteration check** if it is met.
- the procedure stops if there is convergence or no hope for it.

What to base the criterion on?

- monitoring $g'(x)$, e.g. close to 0, is dangerous (flat functions);
- better consider the **difference between x^t and $x^{(t+1)}$** .

Univariate Problems: Convergence criteria

Absolute convergence criterion stops when

$$\left| x^{(t+1)} - x^{(t)} \right| < \epsilon$$

Relative convergence criterion stops when

$$\frac{\left| x^{(t+1)} - x^{(t)} \right|}{\left| x^{(t)} \right|} < \epsilon \text{ or } \frac{\left| x^{(t+1)} - x^{(t)} \right|}{\left| x^{(t)} \right| + \omega} < \epsilon$$

Note:

- preference between the two depends on the problem on hand;
- ω is a small term added to **avoid** denominators **too close** to 0;
- the relative criterion corrects for the **scale of x** .

Univariate Problems: Convergence criteria for bisection

Since in the **bisection method**

$$b_{t+1} - a_{t+1} = 2^{-(t+1)}(b_0 - a_0)$$

the absolute convergence criterion is achieved when

$$t > \log_2\{(b_0 - a_0)/\epsilon\} - 1.$$

So, to achieve **an extra decimal place of precision** ($\epsilon_{\text{new}} = \epsilon/10$), one needs to run the algorithm for **further $\log_2 10 \approx 3.32$ iterations**.

Univariate Problems: Bisection method

Bisection method in **algorithmic form**:

0. Set $t = 0$, and find an interval $[a_0, b_0]$ in which a solution lies.
1. Set $x^{(t)} = (a_t + b_t)/2$.
2. Update the interval
$$[a_{t+1}, b_{t+1}] = \begin{cases} [a_t, x^{(t)}] & \text{if } g'(a_t)g'(x^{(t)}) \leq 0 \\ [x^{(t)}, b_t] & \text{if } g'(a_t)g'(x^{(t)}) > 0 \end{cases}$$
3. if $|x^{(t)} - x^{(t-1)}| < \epsilon$, return $x^{(t)}$,
otherwise, set $t = t + 1$ and go back to step 1.

Note:

- in step 1. it may be better to use $x^{(t)} = a_t + (b_t - a_t)/2$;
- the outcome **depends on the starting interval** $[a_0, b_0]$;
- it finds **a solution**, not necessary the best solution.

Univariate Problems: Bisection method

Summary:

- very **easy approach** (to understand and to implement);
- a solution is **always** found (may not be the global optimum);
- the solution always remains within a **known interval**;
- after T steps, the length of that interval is 2^{-T} times its initial length (the **error** of the approximation is **of order 2^{-T}**);
- each iteration **improves** accuracy (1 digit, ≈ 3.32 iterations);
- it converges **linearly** (the ratios of the lengths of successive intervals is constant, see later).

References |

GIVENS, G. H. & HOETING, J. A. (2013). *Computational Statistics*. Wiley.