Statistical Theory and Modeling (ST2601)
Differentiation and Optimization

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Course introduction**
- **Functions**
- **The derivative**
- **Optimization of functions**

Course introduction

Structure

- **12+1 Lectures** with concepts and theory (Mattias)
- **8 Exercises** with problem solving (Fasna + Ralf)
- **3 Computer labs** for a two-part **home assignment** (Ralf)
- **Jour sessions** for support (Fasna + Ralf)
- Open **Zoom jour sessions** every Thursday (Mattias)

Information sources

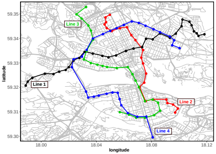
- **Course webpage** at https://statisticssu.github.io/STM/ with reading instructions, slides, exercises, assignment and other material.
- **Athena platform** only for: student hand-ins, messages and recorded lectures. Last minute messages on Athena. Download **It's learning app**.

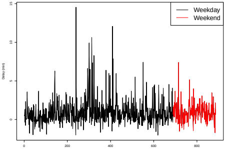
Course introduction

- Aim: Learn what you need for the Bayesian Learning course.
- Some (frequentist) concepts will be missing. By design.
- Examination:
  - Exam, 6 credits (pen and paper, with computer available)
  - Home assignment, 1.5 credits (groups of 3 students)

Why probabilistic models in data science and AI?

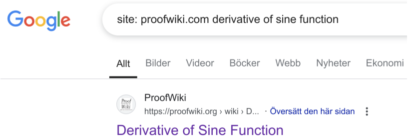
- **Uncertainty quantification**
  - **Point predictions**, **best guess**.
  - **Interval predictions**, **range guess**.
  - **Predictive distributions**, probability for **extremes**.
- **Decisions** under uncertainty - need probabilities conditional on data. **Bayes**. Deep learning's second wave.
- Probability and Statistics are **prerequisties for AI**. [Deep Learning Book]
- **Generative AI**.
- **Principled approach to data analysis**.

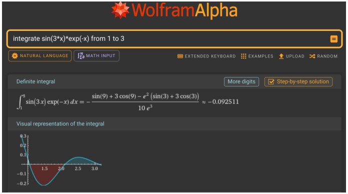


Mathematics

- Some mathematics is needed for statistics.
- Calculations needed for grounding concepts.
- For proofs, see for example: Calculus - a long-form text
- The internet is also helpful:

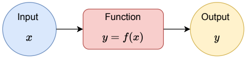


- **ChatGPTs** are great companions. But never trust them!
- Wolfram Alpha is great (or Mathematica)



Functions

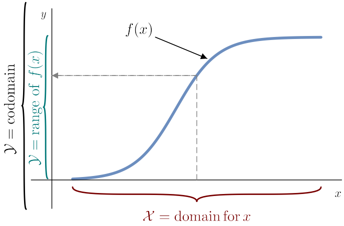
- **Function**: maps an input x to a (unique) output y.



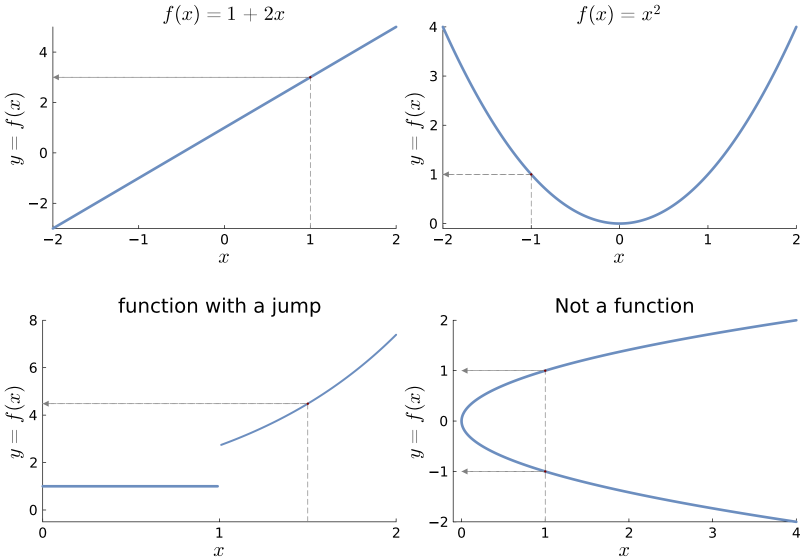

```
f <- function(x){
  # Do something to x
  y = x^2
  return(y)
}
```

```
x = 2
y = f(x) # y is now 4
```

- input = **variable**/argument.
- output = **function value**. y or $f(x)$.
- **Domain** $x \in \mathcal{X}$, **codomain** $y \in \mathcal{Y}$ and **range** (image).



Example functions



The exponential function

- **Exponential function** with base b

$$f(x) = b^x$$

- Compound interest example with 5% interest

$$\text{money after } x \text{ years in bank} = 100 \cdot 1.05^x$$

- **Power function** with power p

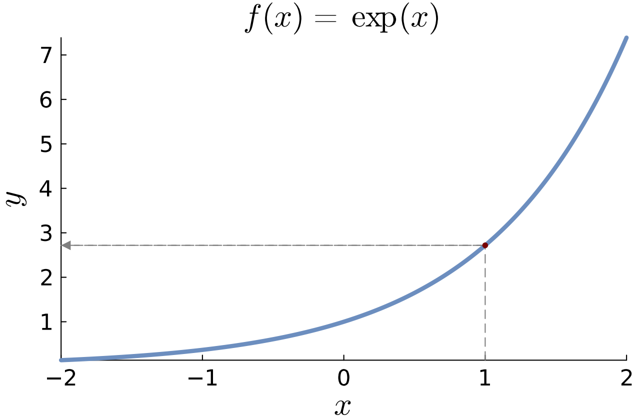
$$f(x) = x^p$$

- **Natural exponential function** with base $e \approx 2.71828$

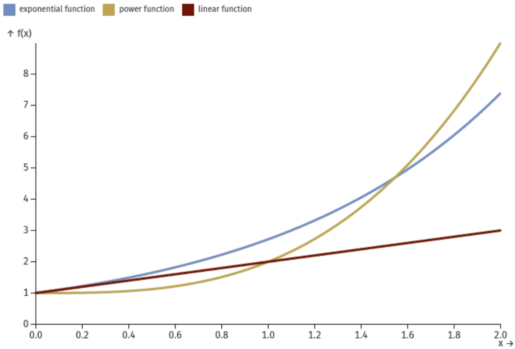
$$f(x) = e^x$$

- Often written as $\exp(x)$.

The exponential function



Exponential function



Properties of exponential numbers

Rules for exponents

$$a^n a^m = a^{n+m}$$

$$(a^n)^m = a^{nm}$$

$$\frac{a^n}{a^m} = a^{n-m}$$

$$a^{-n} = \frac{1}{a^n}$$

$$(ab)^n = a^n b^n$$

$$a^0 = 1$$

$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n}$$

$$\sqrt{a} = a^{1/2}$$

The logarithm function

- **Logarithm** with base 10:

$$\log_{10}(1000) = 3 \iff 1000 = 10^3$$

- Logarithm with base 2:

$$\log_2(256) = 8 \iff 256 = 2^8$$

- Logarithm with base $e \approx 2.71828$:

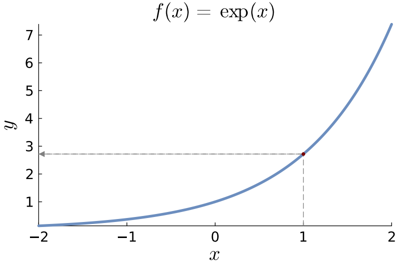
$$\log_e(256) \approx 5.54517 \iff e^{5.54517} \approx 256$$

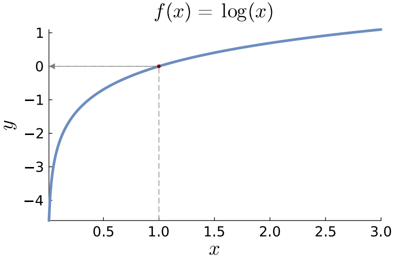
- Logarithm is the **inverse function** to the exponential function

$$\log_e(e^x) = x$$

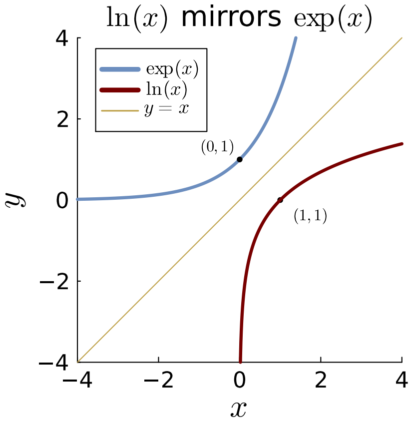
- We often write $\ln(x)$ or just $\log(x)$ when using base e.

Logarithms are inverses to exponentials





Logarithm is inverse to exponential



Properties of logarithms

Rules for logarithms

$$\ln(e) = 1 \qquad \log(1) = 0$$

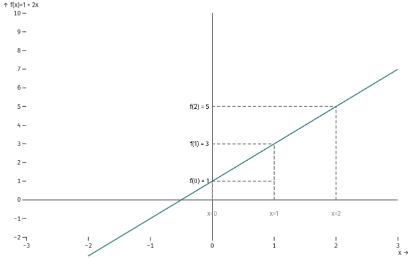
$$\ln(x \cdot y) = \ln x + \ln y \qquad \log\left(\frac{x}{y}\right) = \ln x - \ln y$$

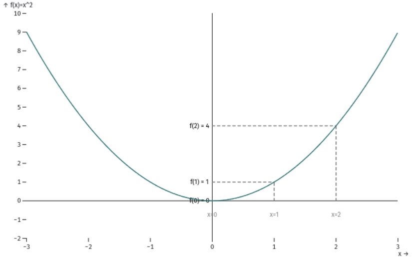
$$\ln x^y = y \ln x \qquad \ln e^y = y \ln e = y$$

- Logarithms turn products into sums (of logs).
- Logarithms 'pull down exponents'.

Rate of change of function

- **How fast does a function change** when x changes from a to $a + h$?
- **Linear function** $c + bx$. Rate of change is always b, for any x-value.
- **Non-linear function** $y = f(x)$. Rate of change depends on x.

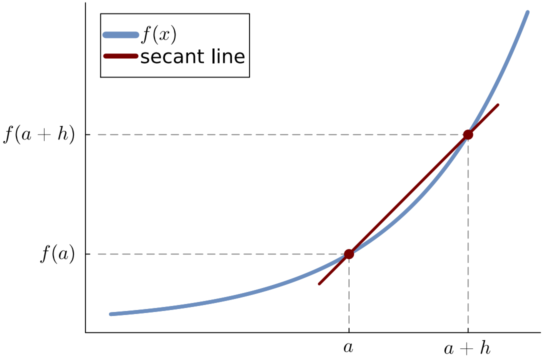




Average rate of change of function

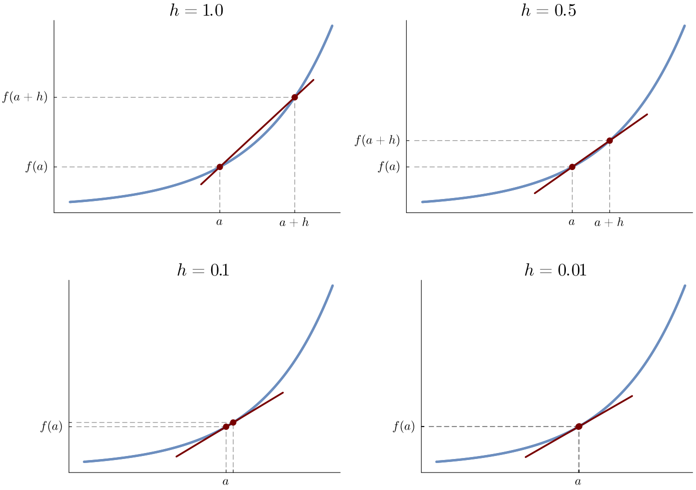
- **Average rate of change** of a function $y = f(x)$

$$\frac{\Delta y}{\Delta x} = \frac{f(a+h) - f(a)}{h}$$



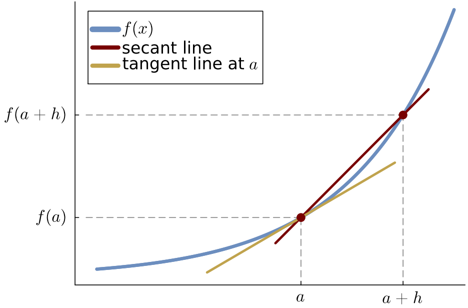
The derivative

- The **derivative** is the average rate of change as $h \to 0$.
- **Instantaneous rate of change**

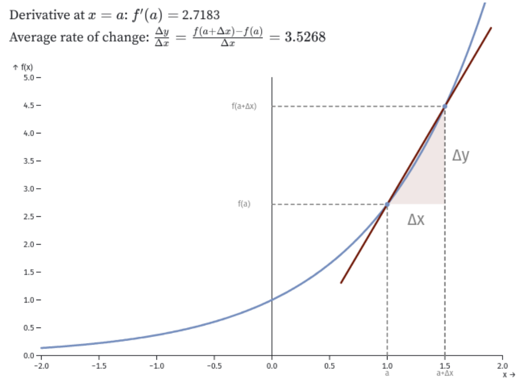


The derivative

- Differentiable at $x = a$: the secant line converges to the **tangent line** as $h \to 0$



Differentiation



Derivative - definition

Definition. *The **derivative** of a function $f(x)$ at $x = a$ is*

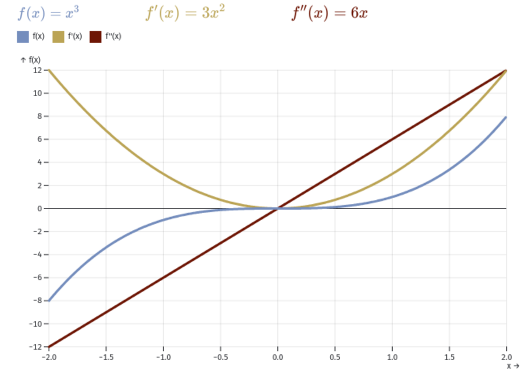
$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

provided that the limit exists.
*If the limit exists we say that $f(x)$ is **differentiable** at $x = a$.*

- The derivative $f'(x)$ is function of x.
- Evaluating $f'(a)$ for some a gives the derivative at $x = a$.
- Alternative notation for the derivative

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) \qquad \text{or} \qquad \frac{\mathrm{d}f(x)}{\mathrm{d}x}$$

Function and its derivatives



Derivatives elementary functions

Derivatives of elementary functions

$$\frac{\mathrm{d}}{\mathrm{d}x}a = 0 \text{ for constant } a$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(a + bx) = b$$

$$\frac{\mathrm{d}}{\mathrm{d}x}x^p = px^{p-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}e^x = e^x$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\ln(x) = \frac{1}{x}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}a^x = a^x \ln(a)$$

Derivative of exponential function

Source: Paula_S_15 on r/mathmemes

Derivatives for combined functions

Derivative of a combination of differentiable functions

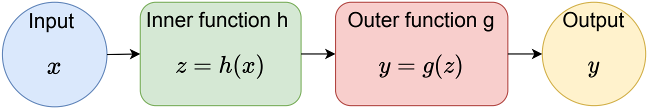
Constant rule	$\frac{\mathrm{d}}{\mathrm{d}x} a = 0$ for constant a
Scaling rule	$\frac{\mathrm{d}}{\mathrm{d}x}\left(a \cdot f(x)\right) = a \cdot f'(x)$ for constant a
Sum rule	$\frac{\mathrm{d}}{\mathrm{d}x}\left(f(x) + g(x)\right) = f'(x) + g'(x)$
Product rule	$\frac{\mathrm{d}}{\mathrm{d}x}\left(f(x)g(x)\right) = f'(x)g(x) + f(x)g'(x)$
Quotient rule	$\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
Reciprocal rule	$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{g(x)} = -\frac{g'(x)}{(g(x))^2}$
Chain rule	$\frac{\mathrm{d}}{\mathrm{d}x} g(h(x)) = g'(h(x)) \cdot h'(x)$

Composite functions? Dude, tell me in code!

- **Math**

$$f(x) = g(h(x))$$

- **Flow chart**



- **Code**

```
# Inner function
h <- function(x){
  z = log(x)
  return(z)
}

# Outer function
g <- function(z){
  y = z^2
  return(y)
}

g(h(2)) # log(2) = 0.6931472 followed by (0.6931472)^2
```

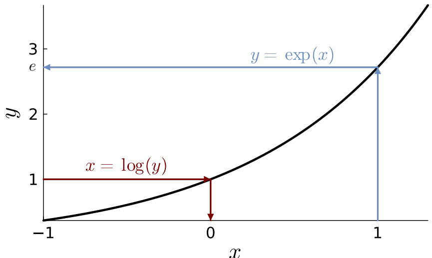
Inverse functions

- **Bijective function** (one-to-one and onto):
  - maps distinct x to distinct y (one-to-one)
  - its range is the whole codomain $\mathcal{Y}$ (onto)
- Bijective function $y = f(x)$ has an **inverse function** $x = f^{-1}(y)$ such that

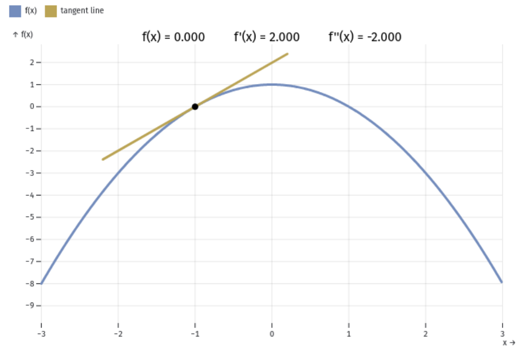
$$f^{-1}(f(x)) = x$$

- Inverse functions goes 'backwards on f' from y down to x.

$\ln(x)$ is $\exp(x)$ backwards



Function optimization



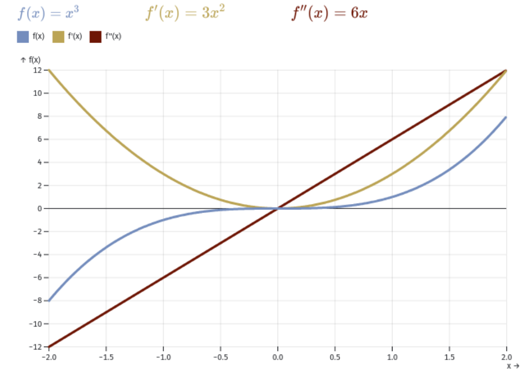
Second order derivative

- Recall: the derivative $f'(x)$ is itself a function of x.
- The **second order derivative** $f''(x)$ is the derivative of $f'(x)$

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

- $f''(x)$ measures **how fast the derivative changes**.
    - can evaluated $f''(a)$ at any $x = a$ or
    - considered as a function of x.
- Example: $f(x) = x^3$. $f'(x) = 3x^2$. $f''(x) = 6x$.
- $f(2) = 2^3 = 8$. $f'(2) = 3 \cdot 2^2 = 12$. $f''(2) = 12$.

Function and its second derivatives



Three uses of second order derivatives

- Second derivative test in **function optimization**
  - $f'(x_{\text{cand}}) = 0$ and $f''(x_{\text{cand}}) < 0$ then is a (local) maximum.
  - $f'(x_{\text{cand}}) = 0$ and $f''(x_{\text{cand}}) > 0$ then is a (local) minimum.
  - $f'(x_{\text{cand}}) = 0$ and $f''(x_{\text{cand}}) = 0$ then test is inconclusive
- $f''(x_{\max})$ measures **how peaked** $f(x)$ is at $x_{\max}$ (or min).
- **Function approximation** (second order Taylor).

Function optimization