# Deep Learning

*LM Computer Science & Data Science*

*2nd semester - 6 CFU*

*Nicolò Navarin & Alessandro Sperduti*

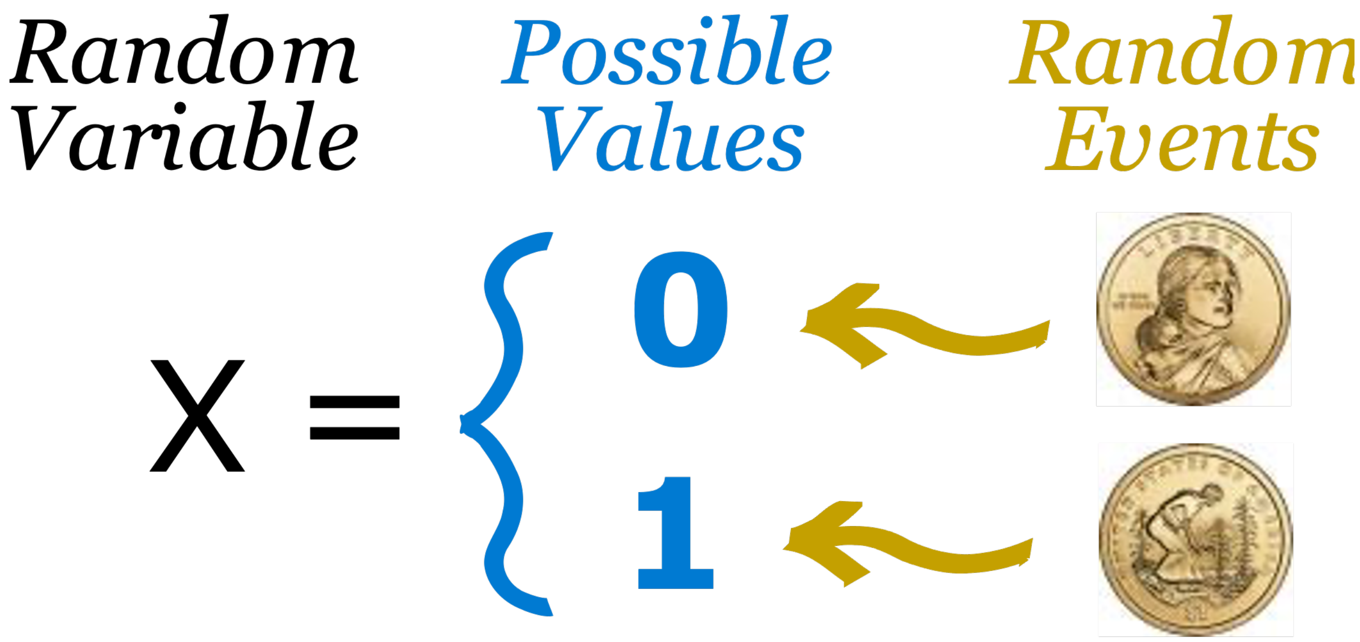
# Probability / Information theory

A primer on probability and information theory (chapter 3)

Maximum Likelihood estimation (section 5.5)

# Probability

- **Random variable**: a variable that can take different values randomly
- Example: Tossing a coin: we could get Heads or Tails.
    - **Heads=0** and **Tails=1**
    - **In each experiment,** Random Variable x can be either 0 or 1



*Random Variable*

*Possible Values*

*Random Events*

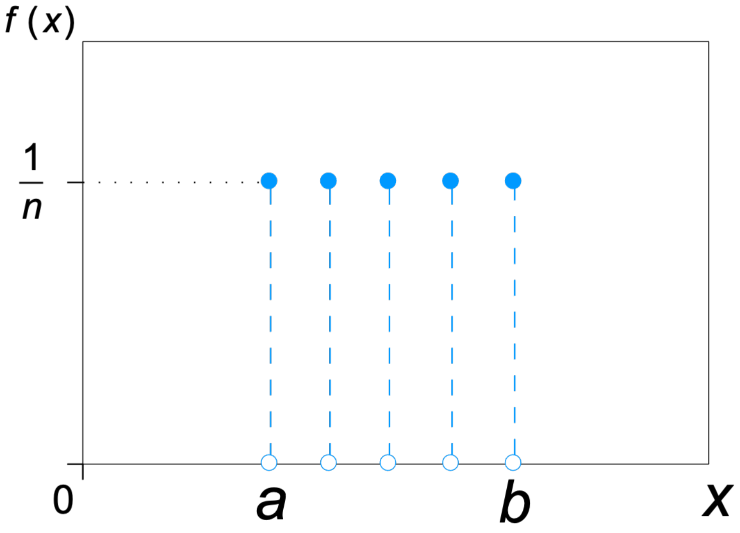
$$X = \begin{cases} 0 \\ 1 \end{cases}$$

# Probability Distributions

- Probability Distribution: A description of how likely a random variable $\mathrm{x}$ (or a set of random variables) is to take each of its possible states

- Discrete variables -> Probability Mass Function

    - Domain of $\mathrm{P}$ is the set of all possible states of $\mathrm{x}$ ($k$ different values)
    - $\forall x \in \mathrm{x}\ 0 \leq P(\mathrm{x} = x) \leq 1$
    - $\sum_{x \in \mathrm{x}} P(x) = 1$



- E.g. Uniform distribution $\forall_{x \in \mathrm{x}} P(\mathrm{x} = x) = \frac{1}{k}$

# Probability Distributions

- **Joint probability**: probability distribution over many variables $P(\mathrm{x} = x, \mathrm{y} = y)$ or $P(x, y)$
- **Marginalization (sum rule):**

$$\forall x \in \mathrm{x}\; P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

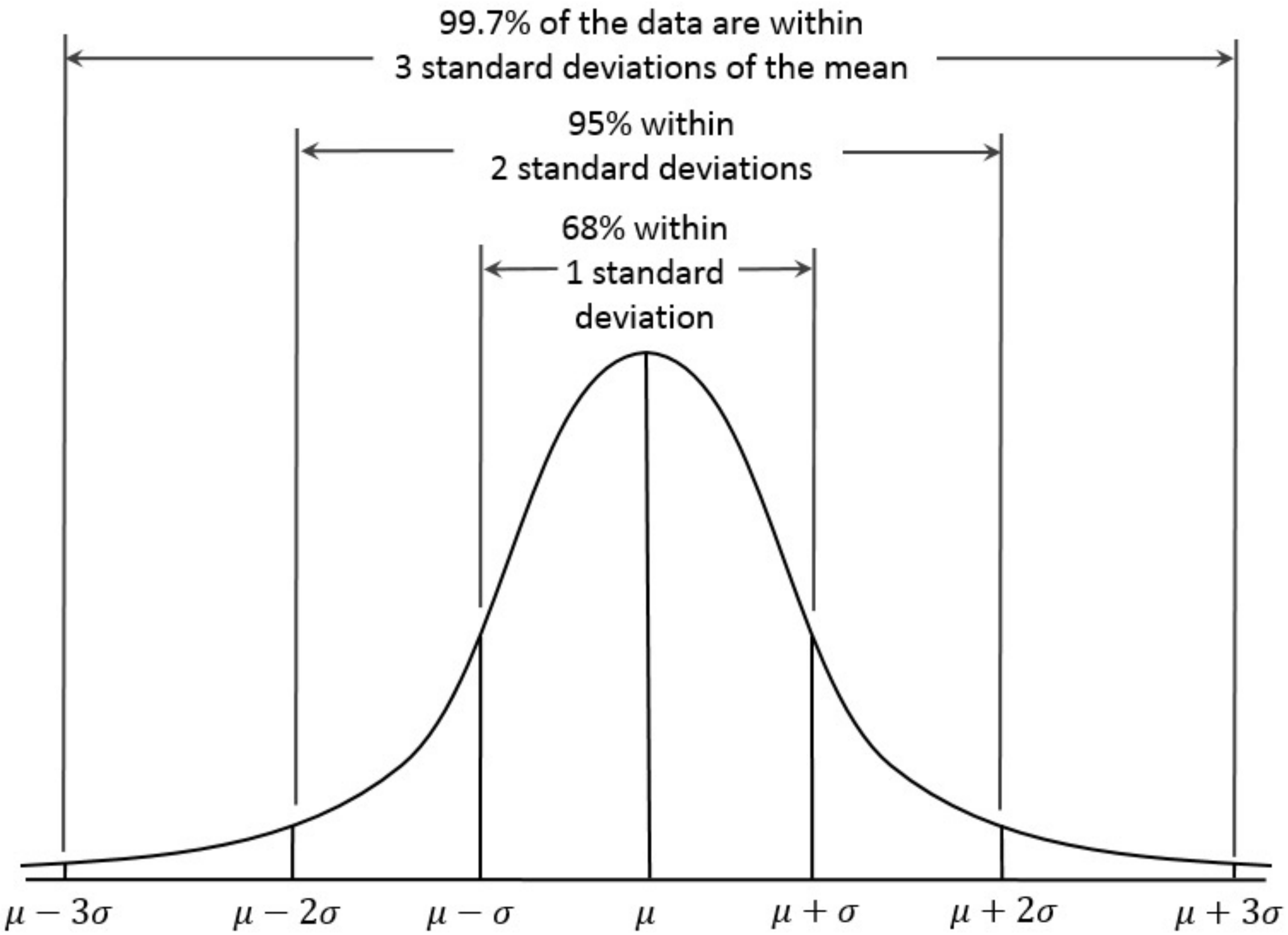
- When dealing with **continuous variables**, a Probability distribution is described by a Probability Density Function (PDF)
  - Domain of $p$ is the set of all possible states of $x$
  - $\forall x \in \mathrm{x}\; P(x) \geq 0$
  - $\int p(x)dx = 1$
- E.g. Gaussian distribution

# Gaussian distribution

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\,\sigma^2}(x-\mu)^2\right)$$

Parametrized by

PDF of Gaussian distribution



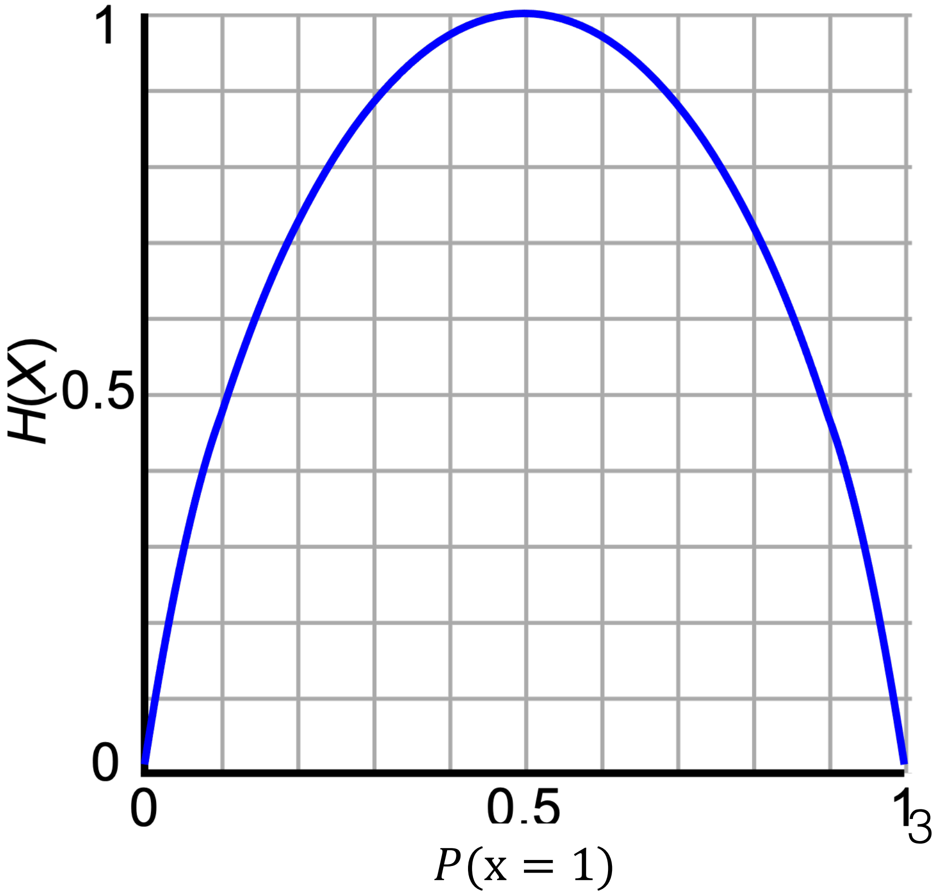
# Entropy

- Shannon Entropy (discrete variable)

$$H(\mathrm{x}) = -\mathbb{E}_{x\sim P(\mathrm{x})}[\log P(x)]$$

- Expected amount of (self-)information in an event drawn from distribution P
- Lower bound on the number of bits needed on average to encode a symbol drawn from that distribution

Entropy H(*X*) of a coin flip, measured in bits, graphed versus the bias of the coin P(*x* = 1), where *x* = 1 represents a result of heads.



# Kullback-Leibler divergence and Cross Entropy

- Let's consider two probability distributions $P(x)$ and $Q(x)$

- Measure how different they are

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$$

  - It is not a true distance because it is not symmetric

- Cross Entropy

$$H(P,Q) = H(P) + D_{KL}(P \parallel Q) = - \mathbb{E}_{x\sim P}[\log Q(x)]$$

Note: minimizing CE of P w.r.t. Q is equivalent to minimize KL divergence between P and Q (if P is given, H(P) and $\mathbb{E}_{x\sim P}[\log P(x)]$ are constants)

# Maximum likelihood estimation

- Principled way to derive estimators (models)
- Consider n examples $Tr = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^n\}$ drawn i.i.d. from $p_{data}(\boldsymbol{x})$
- Let us consider a family of parametric probability distributions (models) $p_{model}(\boldsymbol{x}; \boldsymbol{\theta})$.
  - $p_{model}(\boldsymbol{x}; \boldsymbol{\theta})$ maps a point $\boldsymbol{x}$ to a real number estimating $p_{data}(\boldsymbol{x})$
  - Maximum Likelihood estimation for $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_{ML} = arg\max_{\boldsymbol{\theta}} p_{model}(Tr; \boldsymbol{\theta}) = arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p_{model}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Assumption: independence

# ..A side note on maximum likelihood

- ML is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution

- MAP and maximum likelihood approach makes predictions using a single point estimate of $\boldsymbol{\theta}$

- the Bayesian approach is to make predictions using a full probability distribution over $\boldsymbol{\theta}$

# ..A side note on maximum likelihood

Given a new instance $x$, what is the most probable *classification* ?

- $h_{MAP}(x)$, in general, is not the most probable classification!

Example: let's consider:

- three possibile hypoteses:

$$P(h_1|D) = .4, \; P(h_2|D) = .3, \; P(h_3|D) = .3$$

- given a new instance $x$,

$$h_1(x) = +, \; h_2(x) = -, \; h_3(x) = -$$

- what is the most probable classification for $x$?

Bayes optimal classifier! (not covered in this course)

# Maximum likelihood estimation

- Taking the product of many probabilities is numerically unstable. We can apply the log and the arg max does not change

$$\boldsymbol{\theta}_{ML} = arg\max_{\boldsymbol{\theta}} \sum_{1=1}^{n} \log p_{model}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

- We can equivalently divide by $n$ to express ML as an expectation over training data

$$\boldsymbol{\theta}_{ML} = arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x} \sim \hat{p}_{data}} [\log p_{model}(\boldsymbol{x} \quad ; \boldsymbol{\theta})]$$

- ML minimizes the dissimilarity between $\hat{p}_{data}$ and $p_{model}$, measured by the KL divergence (actually cross entropy, see next slide)

# ML estimation as KL divergence

$$D_{KL}(\hat{p}_{data} \parallel p_{model}) =$$
$$\mathbb{E}_{\boldsymbol{x}\sim\hat{p}_{data}}[\log \hat{p}_{data}(\boldsymbol{x}) - \log p_{model}(\boldsymbol{x};\boldsymbol{\theta})]$$

- The term on the left does not depend on the model.
- To minimize the KL, we need only to minimize

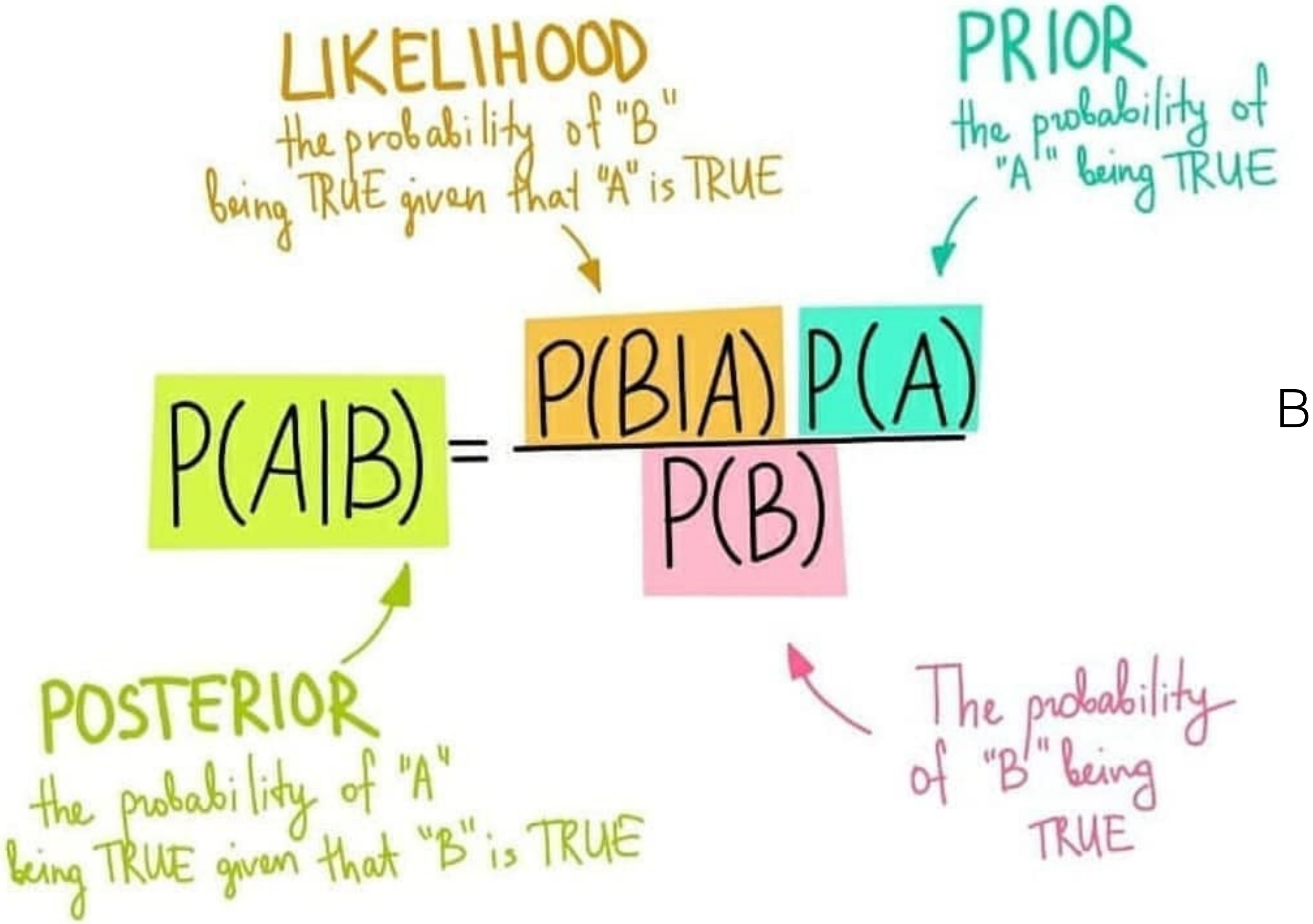
$$\arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{\boldsymbol{x}\sim\hat{p}_{data}}[\log p_{model}(\boldsymbol{x};\boldsymbol{\theta})]$$

- That is the same equation of ML in previous slide
- It also corresponds to minimizing the **cross-entropy** between the two distributions (5 slides back)

# Conditional Probability

- Probability of an event, **given that some other event has happened.**

$$P(a,b) = \mathrm{P(a|b)P(b)}$$ Chain rule of probability


LIKELIHOOD
the probability of "B" being TRUE given that "A" is TRUE
PRIOR
the probability of "A" being TRUE
P(A|B) = P(B|A) P(A) / P(B)
POSTERIOR
the probability of "A" being TRUE given that "B" is TRUE
The probability of "B" being TRUE

Bayes' rule

# Conditional log likelihood

- We can use ML to estimate a **conditional** probability $P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ to predict $\boldsymbol{y}$ given $\boldsymbol{x}$ (supervised learning)

$$\boldsymbol{\theta}_{\boldsymbol{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y}\ |\boldsymbol{X};\boldsymbol{\theta})$$

- If examples are i.i.d.

$$\boldsymbol{\theta}_{\boldsymbol{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log P\left(\boldsymbol{y}^{(i)}\middle|\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right)$$

# Why MSE? Linear regression as ML

- Let’s think about the model as producing a conditional distribution $p(y\,|\boldsymbol{x})$

Model prediction

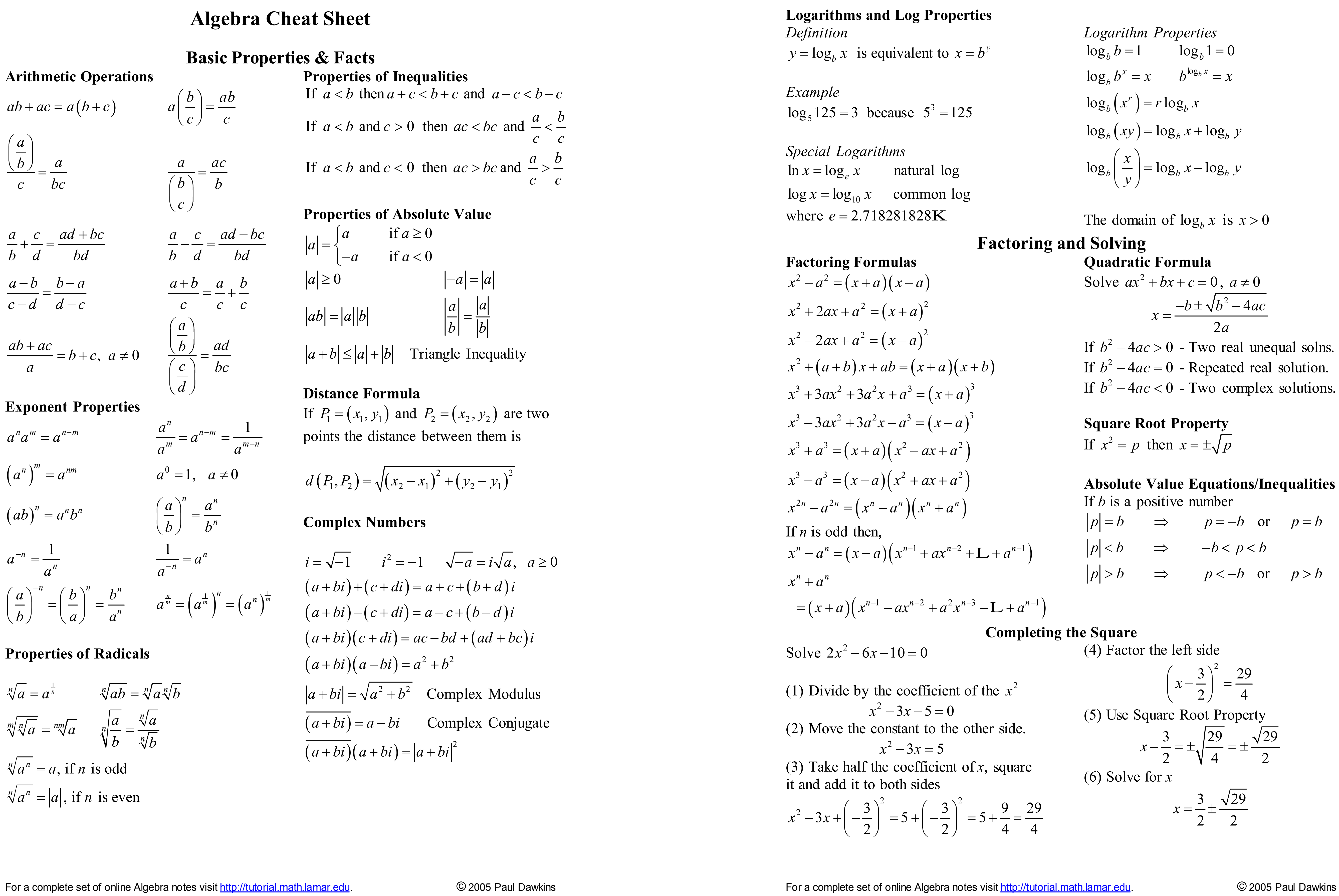
- We define $p(y\,|\boldsymbol{x}) = \mathcal{N}(y;\ \hat{y}(\boldsymbol{x};\boldsymbol{\theta}), \sigma^2)$
- Our model produces $\hat{y}(\boldsymbol{x};\boldsymbol{\theta})$, the mean of a Gaussian distribution
- For i.i.d. examples,

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p\left(y^{(i)}\middle|\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right)$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(\sqrt{\frac{1}{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\,\sigma^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right)\right)$$

# Warning!

- We will use logarithm properties. Check Algebra cheat sheet if some of the rules applied in the next slide are not clear!



# Algebra Cheat Sheet

## Basic Properties & Facts

### Arithmetic Operations

$ab+ac=a(b+c)$ $\qquad a\left(\frac{b}{c}\right)=\frac{ab}{c}$

$\frac{\left(\frac{a}{b}\right)}{c}=\frac{a}{bc}$ $\qquad \frac{a}{\left(\frac{b}{c}\right)}=\frac{ac}{b}$

$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}$ $\qquad \frac{a}{b}-\frac{c}{d}=\frac{ad-bc}{bd}$

$\frac{a-b}{c-d}=\frac{b-a}{d-c}$ $\qquad \frac{a+b}{c}=\frac{a}{c}+\frac{b}{c}$

$\frac{ab+ac}{a}=b+c,\ a\neq 0$ $\qquad \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)}=\frac{ad}{bc}$

### Exponent Properties

$a^n a^m=a^{n+m}$ $\qquad \frac{a^n}{a^m}=a^{n-m}=\frac{1}{a^{m-n}}$

$\left(a^n\right)^m=a^{nm}$ $\qquad a^0=1,\ a\neq 0$

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

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

$\left(\frac{a}{b}\right)^{-n}=\left(\frac{b}{a}\right)^n=\frac{b^n}{a^n}$ $\qquad a^{\frac{n}{m}}=\left(a^{\frac{1}{m}}\right)^n=\left(a^n\right)^{\frac{1}{m}}$

### Properties of Radicals

$\sqrt[n]{a}=a^{\frac{1}{n}}$ $\qquad \sqrt[n]{ab}=\sqrt[n]{a}\sqrt[n]{b}$

$\sqrt[m]{\sqrt[n]{a}}=\sqrt[nm]{a}$ $\qquad \sqrt[n]{\frac{a}{b}}=\frac{\sqrt[n]{a}}{\sqrt[n]{b}}$

$\sqrt[n]{a^n}=a$, if $n$ is odd

$\sqrt[n]{a^n}=|a|$, if $n$ is even

### Properties of Inequalities

If $a<b$ then $a+c<b+c$ and $a-c<b-c$

If $a<b$ and $c>0$ then $ac<bc$ and $\frac{a}{c}<\frac{b}{c}$

If $a<b$ and $c<0$ then $ac>bc$ and $\frac{a}{c}>\frac{b}{c}$

### Properties of Absolute Value

$$|a|=\begin{cases} a & \text{if } a\geq 0\\ -a & \text{if } a<0\end{cases}$$

$|a|\geq 0$ $\qquad |-a|=|a|$

$|ab|=|a||b|$ $\qquad \left|\frac{a}{b}\right|=\frac{|a|}{|b|}$

$|a+b|\leq|a|+|b|$ Triangle Inequality

### Distance Formula

If $P_1=(x_1,y_1)$ and $P_2=(x_2,y_2)$ are two points the distance between them is

$$d(P_1,P_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

### Complex Numbers

$i=\sqrt{-1}$ $\qquad i^2=-1$ $\qquad \sqrt{-a}=i\sqrt{a},\ a\geq 0$

$(a+bi)+(c+di)=a+c+(b+d)i$

$(a+bi)-(c+di)=a-c+(b-d)i$

$(a+bi)(c+di)=ac-bd+(ad+bc)i$

$(a+bi)(a-bi)=a^2+b^2$

$|a+bi|=\sqrt{a^2+b^2}$ Complex Modulus

$\overline{(a+bi)}=a-bi$ Complex Conjugate

$\overline{(a+bi)}(a+bi)=|a+bi|^2$

For a complete set of online Algebra notes visit http://tutorial.math.lamar.edu.

### Logarithms and Log Properties

*Definition*

$y=\log_b x$ is equivalent to $x=b^y$

*Example*

$\log_5 125=3$ because $5^3=125$

*Special Logarithms*

$\ln x=\log_e x$ natural log

$\log x=\log_{10} x$ common log

where $e=2.718281828\mathbf{K}$

*Logarithm Properties*

$\log_b b=1$ $\qquad \log_b 1=0$

$\log_b b^x=x$ $\qquad b^{\log_b x}=x$

$\log_b\left(x^r\right)=r\log_b x$

$\log_b(xy)=\log_b x+\log_b y$

$\log_b\left(\frac{x}{y}\right)=\log_b x-\log_b y$

The domain of $\log_b x$ is $x>0$

## Factoring and Solving

### Factoring Formulas

$x^2-a^2=(x+a)(x-a)$

$x^2+2ax+a^2=(x+a)^2$

$x^2-2ax+a^2=(x-a)^2$

$x^2+(a+b)x+ab=(x+a)(x+b)$

$x^3+3ax^2+3a^2x+a^3=(x+a)^3$

$x^3-3ax^2+3a^2x-a^3=(x-a)^3$

$x^3+a^3=(x+a)\left(x^2-ax+a^2\right)$

$x^3-a^3=(x-a)\left(x^2+ax+a^2\right)$

$x^{2n}-a^{2n}=\left(x^n-a^n\right)\left(x^n+a^n\right)$

If $n$ is odd then,

$x^n-a^n=(x-a)\left(x^{n-1}+ax^{n-2}+\mathbf{L}+a^{n-1}\right)$

$x^n+a^n$

$=(x+a)\left(x^{n-1}-ax^{n-2}+a^2x^{n-3}-\mathbf{L}+a^{n-1}\right)$

### Quadratic Formula

Solve $ax^2+bx+c=0$, $a\neq 0$

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

If $b^2-4ac>0$ - Two real unequal solns.

If $b^2-4ac=0$ - Repeated real solution.

If $b^2-4ac<0$ - Two complex solutions.

### Square Root Property

If $x^2=p$ then $x=\pm\sqrt{p}$

### Absolute Value Equations/Inequalities

If $b$ is a positive number

$|p|=b \quad \Rightarrow \quad p=-b$ or $p=b$

$|p|<b \quad \Rightarrow \quad -b<p<b$

$|p|>b \quad \Rightarrow \quad p<-b$ or $p>b$

### Completing the Square

Solve $2x^2-6x-10=0$

(1) Divide by the coefficient of the $x^2$

$$x^2-3x-5=0$$

(2) Move the constant to the other side.

$$x^2-3x=5$$

(3) Take half the coefficient of $x$, square it and add it to both sides

$$x^2-3x+\left(-\frac{3}{2}\right)^2=5+\left(-\frac{3}{2}\right)^2=5+\frac{9}{4}=\frac{29}{4}$$

(4) Factor the left side

$$\left(x-\frac{3}{2}\right)^2=\frac{29}{4}$$

(5) Use Square Root Property

$$x-\frac{3}{2}=\pm\sqrt{\frac{29}{4}}=\pm\frac{\sqrt{29}}{2}$$

(6) Solve for $x$

$$x=\frac{3}{2}\pm\frac{\sqrt{29}}{2}$$

For a complete set of online Algebra notes visit http://tutorial.math.lamar.edu.

# Why MSE? Linear regression as ML

Log product rule

$$\theta_{ML} = \underset{\boldsymbol{\theta}}{\arg\max} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log exp\left(-\frac{1}{2\,\sigma^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right)$$

Log quotient rule

$$= \underset{\boldsymbol{\theta}}{\arg\max} \textstyle\sum_{i=1}^{n} \log(1) - \log(\sqrt{2\pi\sigma^2}) + \log exp\left(-\frac{1}{2\,\sigma^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right)$$

Log power rule

$$= \underset{\boldsymbol{\theta}}{\arg\max} \textstyle\sum_{i=1}^{n} \log(1) - \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\,\sigma^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \cancel{\log(e)}$$

Natural logarithm and Log power rule

$$= \underset{\boldsymbol{\theta}}{\arg\max} \textstyle\sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\,\sigma^2}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

Algebra

$$= \underset{\boldsymbol{\theta}}{\arg\max} -\frac{\boldsymbol{n}}{\boldsymbol{2}}\log(2\pi\sigma^2) + \textstyle\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{\left(y^{(i)}-\hat{y}^{(i)}\right)^2}{\sigma^2}\right)$$

$$= \underset{\boldsymbol{\theta}}{\arg\max} -\frac{\boldsymbol{n}}{\boldsymbol{2}}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

# Comparing ML with MSE

$$\theta_{ML} = \arg\max_{\boldsymbol{\theta}} \underbrace{-\frac{\boldsymbol{n}}{\boldsymbol{2}}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}}_{\text{Does not depend on } \boldsymbol{\theta}} \sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$$\theta_{MSE} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

- The two functions give the same $\boldsymbol{\theta}$!
- ML estimator is, asymptotically in the number of examples, the best (single) estimator.
- With a small number of examples, regularization strategies to reduce the variance (dedicated chapter).