



Machine Learning Appendix

Model Assessment and Selection

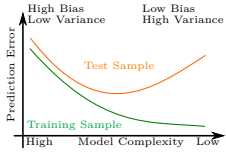
Definition 1.1 Statistical Inference: Is the process of deducing properties of an underlying probability distribution by mere analysis of data.

Definition 1.2 Model Selection:
Is the process of selecting a model f from a given or chosen class of models \mathcal{F}

Definition 1.3 Hyperparameter Tuning: Is the process of choosing the hyperparameters θ of a given model $f \in \mathcal{F}$

Definition 1.4 Model Assessment/Evaluation: Is the process of evaluating the performance of a model.

Definition 1.5 Overfitting:
Describes the result of training/fitting a model f to closely to the training data $\mathcal{Z}^{\text{train}}$.
That is, we are producing overly complicated model by fitting the model to the noise of the training set.
Consequences: the model will generalize poorly as the test set $\mathcal{Z}^{\text{test}}$ will not have not the same noise
 \Rightarrow big test error.



Definition 1.6 Training Set $\mathcal{Z}^{\text{train}}$: Is the part of the data that is used in order to train the model i.e. part of data which is used in order to update the weight according to the loss.

Definition 1.7 Validation Set \mathcal{Z}^{val} : Is the part of the data that is used in order to evaluate different hyperparamters.

Definition 1.8 Test Set $\mathcal{Z}^{\text{test}}$: Is part of the data that is used in order to test the performance of our model.

Move next section to statistical perspective section or merge it somehow

1.1. Core Problem of Statistical Inference

We assume that our data is generated by some probability distribution:

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \stackrel{\text{i.i.d.}}{\sim} f$$

and we want to calculate the expectation of some statistic e.g. the expected loss:

$$\mathcal{R}(f) = \int \int_{\mathcal{X} \times \mathcal{Y}} f(\mathbf{x}, y) l(y, f(\mathbf{x})) \, \mathrm{d}y \, \mathrm{d}\mathbf{x}$$

Problem: we do not know $f(\mathbf{x}, y) \Rightarrow$ can only estimate the **empirical risk** of this statistic:

$$\hat{\mathcal{R}}(f) := \frac{1}{n} \sum_{i=1}^n l(y_i, f(\mathbf{x}_i))$$

Questions

- ① How far is the true risk $\mathcal{R}(f)$ from the empirical risk $\hat{\mathcal{R}}(f)$, for a given f
- ② Given a chosen hypothesis class \mathcal{F} . How far is the minimizer of the true cost way from the minimizer of the empirical cost

$$f^*(\mathbf{x}) \in \arg \min_{f \in \mathcal{F}} \mathcal{R}(f) \quad \text{vs.} \quad \hat{f}(\mathbf{x}) \in \arg \min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f)$$

1.2. Empirical Risk Minimization

- 1. For a chosen set of function classes \mathcal{F} minimize the empirical risk/loss:

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{\mathcal{R}} \left(f, \mathcal{Z}^{tr} \right) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n l \left(y_i, f(\mathbf{x}_i) \right)$$

- 2. Determine the best parameter θ^* by using the validation set for evaluation:

$$\hat{\theta} \left(\mathcal{Z}^{val} \right) \in \arg \min_{\theta: \hat{f}_{\theta} \in \mathcal{F}_{\theta}} \hat{R} \left(\hat{f}_{\theta} \left(\mathcal{Z}^{tr} \right), \mathcal{Z}^{val} \right)$$

- 3. Use the tests set in order to test the model:

$$\hat{\mathcal{R}} \left(\hat{f}_{\hat{\theta} \left(\mathcal{Z}^{val} \right)} \left(\mathcal{Z}^{tr} \right), \mathcal{Z}^{test} \right)$$

Note: overfitting to the validation set

Tuning the configuration/hyperparameters of the model based on its performance on the validation set can result in overfitting to the validation set, even though your model is never directly trained on it \Rightarrow split the data into a test and training and validation set.

2. Cross Validation

Definition 1.9 Cross Validation: Is a model validation/assessment techniques for assessing how the results of a statistical analysis (model) will generalize to an independent data set.

2.1. Validation Set Approach

Definition 1.10 Hold out/Validation Set:

add

2.2. Leave-One-Out Cross Validation (LOOCV)

2.3. K-Fold Cross Validation

A Statistical Perspective

1. Information Theory

1.1. Information Content

Definition 2.1 Information (Claude Elwood Shannon): Information is the resolution of uncertainty.

Amount of Information

The information gained of by the realization of a coin tossed n -times should equal to the sum of the information of tossing a coin once n -times:

$$I(\mathbf{p}_1 \cdots \mathbf{p}_0 \cdots \mathbf{p}_1) = I(\mathbf{p}_1) + I(\mathbf{p}_0) + \cdots + I(\mathbf{p}_n)$$

\Rightarrow can use the logarithm to satisfy this

Definition 2.2 Surprise/Self-Information/-Content: Is a measure of the information of a realization x of a random variable $X \sim \mathbf{p}$:

$$I_X(x) = \log\left(\frac{1}{\mathbf{p}(X=x)}\right) = -\log \mathbf{p}(X=x) \quad (2.1)$$

Corollary 2.1 Units of the Shannon Entropy:

The Shannon entropy can be defined for different logarithms

	log	units
\cong units:	Base 2	Bits/Shannons
	Natural	Nats
	Base 10	Dits/Bans

Explanation 2.1. An uncertain event is much more informative than an expected/certain event:

$$\text{surprise/inf. content} = \begin{cases} \text{big} & \text{if } \mathbf{p}_X(x) \text{ unlikely} \\ \text{small} & \text{if } \mathbf{p}_X(x) \text{ likely} \end{cases}$$

Information content deals with a single event if we want to quantify the amount of uncertainty/information of an probability distribution we need to take the expectation over the information content^[def. 2.2]:

Definition 2.3 Shannon Entropy example 2.3: Is the expected amount of information of a random variable $X \sim \mathbf{p}$:

$$\begin{aligned} H(\mathbf{p}) &= \mathbb{E}_X[I_X(x)] = \mathbb{E}_X\left[\log \frac{1}{\mathbf{p}_X(x)}\right] = -\mathbb{E}_X[\log \mathbf{p}_X(x)] \\ &= -\sum_{i=1}^n \mathbf{p}(x_i) \log \mathbf{p}(x_i) \end{aligned} \quad (2.2)$$

Definition 2.4 Differential/Continuous entropy: Is the continuous version of the Shannon entropy^[def. 2.3]:

$$H(\mathbf{p}) = \int_{x \sim \mathbf{p}} -f(x) \log f(x) \, dx \quad (2.3)$$

Notes

The Shannon entropy is maximized for uniform distributions

1.2. Cross Entropy

Definition 2.5 Cross Entropy proof 4: Lets say a model follows a true distribution $X \sim \mathbf{p}$ but we model X as with a different distribution $X \sim \mathbf{q}$. The cross entropy between \mathbf{p} and \mathbf{q} measure the average amount of information/bits needed to model an outcome $x \sim \mathbf{p}$ with X :

$$H(\mathbf{p}, \mathbf{q}) = \mathbb{E}_{x \sim \mathbf{p}} \left[\log \left(\frac{1}{\mathbf{q}(x)} \right) \right] = -\mathbb{E}_{x \sim \mathbf{p}} [\log \mathbf{q}(x)] \quad (2.4)$$

$$= H(\mathbf{p}) + D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \quad (2.5)$$

Corollary 2.2 Kullback-Leibler Divergence: $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q})$ measures the extra price (bits) we need to pay for using \mathbf{q} .

1.3. Kullback-Leibler (KL) divergence

If we want to measure how different two distributions \mathbf{q} and \mathbf{p} over the same random variable X are we can define another measure.

Definition 2.6

Kullback–Leibler divergence. examples 2.4 and 2.5 /**Relative Entropy from \mathbf{p} to \mathbf{q} :** Given two probability distributions \mathbf{p}, \mathbf{q} of a random variable X . The Kullback–Leibler divergence is defined to be:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = \mathbb{E}_{x \sim \mathbf{p}} \left[\log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} \right] = \mathbb{E}_{x \sim \mathbf{p}} [\log \mathbf{p}(x) - \log \mathbf{q}(x)] \quad (2.6)$$

and measures how far away a distribution \mathbf{q} is from a another distribution \mathbf{p} .

Explanation 2.2.

- \mathbf{p} decides where we put the mass if $\mathbf{p}(x)$ is zero we do not care about $\mathbf{q}(x)$.
- $\mathbf{p}(x)/\mathbf{q}(x)$ determines how big the difference between the distributions is.

Intuition

The KL-divergence helps us to measure just how much information we lose when we choose an approximation.

Property 2.1 Non-Symmetric:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \neq D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) \quad \forall \mathbf{p}, \mathbf{q} \quad (2.7)$$

Property 2.2:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \geq 0 \quad (2.8)$$

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = 0 \iff \mathbf{p}(x) = \mathbf{q}(x) \forall x \in \mathcal{X} \quad (2.9)$$

Note

The KL-divergence is not a real distance measure as $\text{KL}(\mathbf{p} \parallel \mathbf{q}) \neq \text{KL}(\mathbf{q} \parallel \mathbf{p})$

2. Supervised Learning

Recall: goal of supervised learning

Given: training data:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$$

find a hypothesis $h: \mathcal{X} \rightarrow \mathcal{Y}$ e.g.

- Linear Regression:** $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$
- Linear Classification:** $h(\mathbf{x}) = \text{sing}(\mathbf{w}^\top \mathbf{x})$
- Kernel Regression:** $h(\mathbf{x}) = \sum_{i=1}^n \alpha_i \mathbf{k}(\mathbf{x}_i, \mathbf{x})$
- Neural Networks** (single hidden layer):
 $h(\mathbf{x}) = \sum_{i=1}^n \mathbf{w}'_i \phi(\mathbf{w}^\top_i \mathbf{x})$

s.t. we minimize prediction error/empirical risk^[def. 2.11].

Fundamental assumption

The data is generated *i.i.d.* from some unknown probability distribution:

$$(\mathbf{x}_i, y_i) \sim \mathbf{p}_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}_i, y_i)$$

Note

The distribution $\mathbf{p}_{\mathcal{X}, \mathcal{Y}}$ is dedicated by nature and may be highly complex (not smooth, multimodal, ...).

2.1. Generalization Error

Definition 2.7

Generalization/Prediction Error (Risk): Is defined as the expected value of a loss function l of a given predictor h , for data drawn from a distribution $\mathbf{p}_{\mathcal{X}, \mathcal{Y}}$.

$$\begin{aligned} R_{\mathbf{p}}(h) &= \mathbb{E}_{(\mathbf{x}, y) \sim \mathbf{p}} [l(y; h(\mathbf{x}))] = \int_{\mathcal{D}} \mathbf{p}(\mathbf{x}, y) l(y; h(\mathbf{x})) \, d\mathbf{x} \, dy \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{p}(\mathbf{x}, y) l(y, h(\mathbf{x})) \, d\mathbf{x} \, dy \\ &\stackrel{??}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} l(y, h(\mathbf{x})) \mathbf{p}(y|\mathbf{x}) \mathbf{p}(\mathbf{x}) \, d\mathbf{x} \, dy \end{aligned} \quad (2.10)$$

Interpretation

Is a measure of how accurately an algorithm is able to predict outcome values for future/unseen/test data.

Definition 2.8 Expected Conditional Risk: If we only know a certain \mathbf{x} but not the distribution of those measurements ($\mathbf{x} \sim \mathbf{p}_{\mathcal{X}}(\mathbf{x})$), we can still calculate the expected risk given/conditioned on the known measurement \mathbf{x} :

$$\mathcal{R}_{\mathbf{p}}(h, \mathbf{x}) = \int_{\mathcal{Y}} l(y, h(\mathbf{x})) \mathbf{p}(y|\mathbf{x}) \, dy$$

Note: [def. 2.7] \iff [def. 2.8]

$$R_{\mathbf{p}}(h) = \mathbb{E}_{\mathbf{x} \sim \mathbf{p}} [R_{\mathbf{p}}(h, \mathbf{x})] = \int_{\mathcal{X}} \mathbf{p}(\mathbf{x}) R_{\mathbf{p}}(h, \mathbf{x}) \, d\mathbf{x} \quad (2.11)$$

Definition 2.9 Expected Risk Minimizer (TRM) h^* : Is the model h that minimizes the total expected risk:

$$h^* \in \arg \min_{h \in \mathcal{C}} \mathcal{R}(h) \quad (2.12)$$

Problem

In practice we do neither know the distribution $\mathbf{p}_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}, y)$, nor $\mathbf{p}_{\mathcal{X}}(\mathbf{x})$ or $\mathbf{p}_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x})$ (otherwise we would already know the solution).

But: even though we do not know the distribution of $\mathbf{p}_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}, y)$ we can still sample from it in order to define an empirical risk.

2.2. Empirical Risk

Definition 2.10 Empirical Risk: Is the the average of a loss function of an estimator h over a finite set of data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ drawn from $\mathbf{p}_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}, y)$:

$$\hat{\mathcal{R}}_n(h) = \frac{1}{n} \sum_{i=1}^n l(h(\mathbf{x}_i), y_i)$$

Note

- $\hat{\mathcal{R}}_n(f) \neq \mathbb{E}_{\mathcal{X}, \mathcal{Y}}[l(f(\mathbf{x}), y)]$.
- We hope that $\lim_{n \rightarrow \infty} \hat{\mathcal{R}}_n(f) = \mathcal{R}(f)$.

Definition 2.11 Empirical Risk Minimizer (ERM) \hat{h} : Is the model h that minimizes the total empirical risk:

$$\hat{h} \in \arg \min_{h \in \mathcal{C}} \hat{\mathcal{R}}(h) \quad (2.13)$$

2.2.1. Bayes Optimal Predictor

Objective

Given data generated *i.i.d.* from an distribution

$\mathbf{p}_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}_i, y_i)$.

Goal: find the function/predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes the expected risk^[def. 2.7] i.e. we want to find the expected risk minimizer^([def. 2.9]).

Definition 2.12

Bayes' optimal predictor for the L2-Loss:

Assuming: i.i.d. generated data by $(\mathbf{x}_i, y_i) \sim \mathbf{p}(\mathcal{X}, \mathcal{Y})$.

Considering: the least squares risk:

$$R_{\mathbf{p}}(h) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathbf{p}} [(y - h(\mathbf{x}))^2]$$

The best hypothesis/predictor h^* minimizing $R(h)$ is given by **conditional mean/expectation** of the data:

$$h^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \quad (2.14)$$

see

Proof. proof: defn: bayesOptPredictor

Notes

- The optimal predictor may not be unique as even for a fixed \mathbf{X} we may sample different Y , that is if we observe a \mathbf{x} multiple times we may still get different y values.
- Our model/prediction is unique, can only predict a specific y .
Hence even if our model fits exactly the data generating process $\mathbf{X} = \mathbf{x}$ we may still obtain different y 's because due to random/independent measurement noise/errors that the optimal bayes predictor still makes.

2.3. How to make use of this in Practice

In Practice

Problem: we do not know the real distribution $\mathbf{p}_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x})$, which we need in order to find the bayes optimal predictor according to eq. (2.14).

Idea:

- Use artificial data/density estimator $\hat{\mathbf{p}}(\mathcal{Y}|\mathcal{X})$ in order to estimate $\mathbb{E}[\mathcal{Y}|\mathcal{X} = \mathbf{x}]$
- Predict a test point \mathbf{x} by:
$$\hat{y} = \hat{\mathbb{E}}[\mathcal{Y}|\mathcal{X} = \mathbf{x}] = \int \hat{\mathbf{p}}(y|\mathbf{X} = \mathbf{x}) y \, dy$$

Common approach: $\mathbf{p}(\mathcal{X}, \mathcal{Y})$ may be some very complex (non-smooth, ...) distribution \Rightarrow need to make some assumptions in order to approximate $\mathbf{p}(\mathcal{X}, \mathcal{Y})$ by $\hat{\mathbf{p}}(\mathcal{X}, \mathcal{Y})$ **Idea:** choose parametric form $\hat{\mathbf{p}}(Y|\mathbf{X}, \theta) = \hat{\mathbf{p}}_{\theta}(Y|\mathbf{X})$ and then optimize the parameter θ which results in the so called maximum likelihood estimation section 1.

Definition 2.13 Statistical Inference: Goal of Inference

- What is a good guess of the parameters of my model?
- How do I quantify my uncertainty in the guess?

3. Estimators

Definition 2.14 (Sample) Statistic: A statistic is a measurable function f that assigns a **single** value F to a sample of random variables:
$$\mathbf{X} = \{X_1, \dots, X_n\}$$
$$f: \mathbb{R}^n \mapsto \mathbb{R} \quad F = f(X_1, \dots, X_n)$$

E.g. F could be the mean, variance, ...

Note

The function itself is independent of the sample's distribution; that is, the function can be stated before realization of the data.

Definition 2.15 Statistical/Population Parameter:

Is a parameter defining a family of probability distributions see example 2.1

Definition 2.16 (Point) Estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$:

Given: n -samples $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathbf{X}$ an estimator
$$\hat{\theta} = \hat{h}(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (2.15)$$

is a statistic/random variable used to estimate a true (population) parameter θ ^[def. 2.15] see also example 2.2.

Note

The other kind of estimators are interval estimators which do not calculate a statistic **but** an interval of plausible values of an unknown population parameter θ .

The most prevalent forms of interval estimation are:

- Confidence intervals (frequentist method).
- Credible intervals (Bayesian method).

4. Proofs

Proof. [def. 2.5]

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim \mathbf{q}} \left[\log \left(\frac{1}{\mathbf{p}(x)} \right) \right] &= \mathbb{E}_{\mathbf{x} \sim \mathbf{q}} \left[\log \left(\frac{\mathbf{q}(x)}{\mathbf{p}(x)} \right) + \log \left(\frac{1}{\mathbf{q}(x)} \right) \right] \\ &= H(\mathbf{p}) + D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \end{aligned}$$

□

Proof.

[def. 2.12]:

$$\begin{aligned} \min_h R(h) &= \min_h \mathbb{E}_{(\mathbf{x}, y) \sim \mathbf{p}} [(y - h(\mathbf{x}))^2] \\ &\stackrel{??}{=} \min_h \mathbb{E}_{\mathbf{x} \sim \mathbf{p}} \mathbb{E}_{y \sim \mathbf{p}_{Y|\mathcal{X}}} \left[\mathbb{E}_{y \sim \mathbf{p}_{Y|\mathcal{X}}} \left[(y - h(\mathbf{x}))^2 \mid \mathbf{x} \right] \right] \\ &\stackrel{\heartsuit}{=} \mathbb{E}_{\mathbf{x} \sim \mathbf{p}} \mathcal{R}_{\mathbf{p}} \left(\min_h \underbrace{\mathbb{E}_{y \sim \mathbf{p}_{Y|\mathcal{X}}} \left[(y - h(\mathbf{x}))^2 \mid \mathbf{x} \right]}_{\mathcal{R}_{\mathbf{p}}(h, \mathbf{x})} \right) \end{aligned}$$

Now let's minimize the conditional executed risk:

$h^*(\mathbf{x}) = \arg \min_h \mathbb{E}_{y \sim \mathbf{p}_{Y|\mathcal{X}}} \left[(y - h(\mathbf{x}))^2 \mid \mathbf{x} \right]$

(2.16)

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{d}{dh^*} \mathcal{R}_{\mathbf{p}}(h^*, \mathbf{x}) = \frac{d}{dh^*} \int (y - h^*)^2 \mathbf{p}(y|x) dy \\ &= \int \frac{d}{dh^*} (y - h^*)^2 \mathbf{p}(y|x) dy = \int 2(y - h^*) \mathbf{p}(y|x) dy \\ &= -2h^* \underbrace{\int \mathbf{p}(y|x) dy}_{=1} + 2 \underbrace{\int y \mathbf{p}(y|x) dy}_{\mathbb{E}_Y[Y|X=x]} \end{aligned}$$

Notes: ♡

Since we can pick $h(\mathbf{x}_i)$ independently from $h(\mathbf{x}_j)$.

Note

$$\begin{aligned} \mathbb{E}[X] \mathbb{E}[Y|X] &= \int_X \mathbf{p}_X(x) dx \int_Y \mathbf{p}(y|x) dy \\ &= \int_X \int_Y \mathbf{p}_X(x) \mathbf{p}(y|x) xy dx dy = \mathbb{E}[X, Y] \end{aligned}$$

5. Examples

Example 2.1 :

Normal distribution has two population parameters: the mean μ and the variance σ^2 .

Example 2.2 Various kind of estimators:

- Best linear unbiased estimator (BLUE).
- Minimum-variance mean-unbiased estimator (MVUE): minimizes the risk (expected loss) of the squared-error loss-function.
- Minimum mean squared error (MMSE).
- Maximum likelihood estimator (MLE): is given by the least squares solution (minimum squared error), assuming that the noise is i.i.d. Gaussian with constant variance and will be considered in the next section.

Example 2.3 Entropy of a Gaussian:

$$\begin{aligned} H(\mathcal{N}(\mu, \Sigma)) &= \frac{1}{2} \ln |2\pi e \Sigma| \\ &\stackrel{\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)}{=} \frac{1}{2} \ln |2\pi e| + \frac{1}{2} \sum_{i=1}^d \ln \sigma_i^2 \end{aligned}$$

Example 2.4 KL Divergence of Gaussians:

Given two Gaussian distributions:

$$\mathbf{p} = \mathcal{N}(\mu_{\mathbf{p}}, \Sigma_{\mathbf{p}}) \quad q = \mathcal{N}(\mu_q, \Sigma_q) \quad \text{it holds}$$

$$\begin{aligned} D_{\text{KL}}(\mathbf{p} \parallel q) &= \frac{1}{2} \left(\text{tr} \left(\Sigma_q^{-1} \Sigma_{\mathbf{p}} \right) + (\mu_q - \mu_{\mathbf{p}})^{\top} \Sigma_q^{-1} (\mu_q - \mu_{\mathbf{p}}) \right. \\ &\quad \left. - d + \ln \left(\frac{|\Sigma_q|}{|\Sigma_{\mathbf{p}}|} \right) \right) \end{aligned}$$

Example 2.5 KL Divergence of Gaussians:

$$\mathbf{p} = \mathcal{N}(\mu_{\mathbf{p}}, \text{diag}(\sigma_1^2, \dots, \sigma_d^2)) \quad q = \mathcal{N}(\mathbf{0}, I) \quad \text{it holds}$$

$$D_{\text{KL}}(\mathbf{p} \parallel q) = \frac{1}{2} \sum_{i=1}^d \left(\sigma_i^2 + \mu_i^2 - 1 - \ln \sigma_i^2 \right)$$

Model Parameter Estimation

1. Maximum Likelihood Estimation

1.1. Likelihood Function

Is a method for estimating the parameters θ of a model that agree best with observed data $\{x_1, \dots, x_n\}$. Let: $\theta = (\theta_1 \dots \theta_k)^\top \in \Theta \subset \mathbb{R}^k$ vector of unknown model parameters.

Consider: a probability density/mass function $f_X(x; \theta)$

Definition 3.1 Likelihood Function $\mathcal{L}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}_+$: Let $\mathbf{X} = \{x_i\}_{i=1}^n$ be a random sample of i.i.d. data points drawn from an unknown probability distribution $x_i \sim p_X$. The likelihood function gives the likelihood/probability of the joint probability of the data $\{x_1, \dots, x_n\}$ given a fixed set of model parameters θ :

$$\mathcal{L}_n(\theta|\mathbf{X}) = \mathcal{L}_n(\theta; \mathbf{X}) = f(\mathbf{X}|\theta) = f(\mathbf{X}; \theta) \quad (3.1)$$

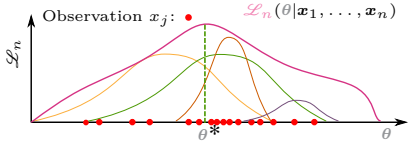


Figure 1: Possible Likelihood function in pink. Overlaid: possible candidate functions for Gaussian model explaining the observations.

Likelihood function is not a pdf

The likelihood function by default not a probability density function and may not even be differentiable. However if it is, then it may be normalized to one.

Corollary 3.1 i.i.d. data: If the n-data points of our sample are i.i.d. then the likelihood function can be decomposed into a product of n-terms:

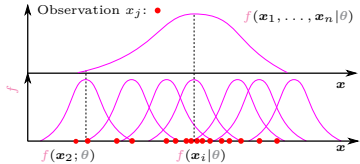


Figure 2: Bottom: probability distributions of the different data points x_i given a fixed θ for a Gaussian distribution. Top: joint probability distribution of the i.i.d. data points $\{x_i\}_{i=1}^n$ given a fixed θ

$$f(x_1, \dots, x_n | \theta) \stackrel{\text{i.i.d.}}{=} \prod_{i=1}^n f(x_i | \theta)$$

Notation

- The probability density $f(\mathbf{X}|\theta)$ is considered for a fixed θ and thus as a function of the samples.
- The likelihood function on the other hand is considered as a function over parameter values θ for a fixed sample $\{x_i\}_{i=1}^n$ and thus written as $\mathcal{L}_n(\theta|\mathbf{X})$.
- Often the colon symbol ; is written instead of the *is given* symbol | in order to indicate that θ resp. \mathbf{X} is a parameter and not a random variable.

1.2. Maximum Likelihood Estimation (MLE)

Let $f_\theta(x)$ be the probability of an i.i.d. sample x for a given model.

Goal: find θ of a given model that maximizes the joint probability/likelihood of the observed data $\{x_1, \dots, x_n\}$? \iff maximum likelihood estimator θ^* .

Definition 3.2 Log Likelihood Function $l_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}$:

$$l_n(\theta|\mathbf{X}) = \log \mathcal{L}_n(\theta|\mathbf{X}) = \log f(\mathbf{X}|\theta) \quad (3.2)$$

Corollary 3.2 i.i.d. data: Differentiating the product of n-terms with the help of the chain rule leads often to complex terms. As a result one usually prefers maximizing the log (especially for exponential terms), as it does not change the *argmax*eq. (11.51):

$$\log f(x_1, \dots, x_n | \theta) \stackrel{\text{i.i.d.}}{=} \log \left(\prod_{i=1}^n f(x_i | \theta) \right) = \sum_{i=1}^n \log f(x_i | \theta)$$

Definition 3.3 Maximum Likelihood Estimator θ^* : Is the estimator $\theta^* \in \Theta$ that maximizes the likelihood of the model/predictor:

$$\theta^* = \arg \max_{\theta \in \Theta} \mathcal{L}_n(\theta; \mathbf{x}) \quad \text{or} \quad \theta^* = \arg \max_{\theta \in \Theta} l_n(\theta; \mathbf{x}) \quad (3.3)$$

1.3. Maximization vs. Minimization

For optimization problems we minimize by convention. The logarithm is a concave function^[def. 11.21] \cap , thus if we calculate the extremal point we will obtain a maximum. If we want to calculate a minimum instead (i.e. in order to be compatible with some computer algorithm) we can convert the function into a convex function^{section 3} \cup by multiplying it by minus one and consider it as a loss function instead of a likelihood.

Definition 3.4 Negative Log-likelihood $-l_n(\theta|\mathbf{X})$:

$$\theta^* = \arg \max_{\theta \in \Theta} l_n(\theta|\mathbf{X}) = \arg \min_{\theta \in \Theta} -l_n(\theta|\mathbf{X}) \quad (3.4)$$

1.4. Conditional Maximum Likelihood Estimation

Maximum likelihood estimation can also be used for conditional distributions.

Assume the labels y_i are drawn i.i.d. from an unknown true conditional probability distribution $f_{Y|X}$ and we are given a data set $\mathbf{Z} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n$.

Now we want to find the parameters $\theta = (\theta_1 \dots \theta_k)^\top \in \Theta \subset \mathbb{R}^k$ of a hypothesis $\hat{f}_{Y|X}$ that agree best with the given data \mathbf{Z} .

Note

For simplicity we omit the hat of our model $\hat{f}_{Y|X}$ and simply assume that our data is generated by some data generating probability distribution.

Definition 3.5 Conditional (log) likelihood function:

Models the likelihood of a model with parameters θ given the data $\mathbf{Z} = \{x_i, y_i\}_{i=1}^n$
 $\mathcal{L}_n(\theta|Y, \mathbf{X}) = \mathcal{L}_n(\theta; Y, \mathbf{X}) = f(Y|\mathbf{X}, \theta) = f(Y|\mathbf{X}; \theta)$

2. Maximum a posteriori estimation (MAP)

Idea

We have seen (??), that trading/increasing a bit of bias can lead to a big reduction of variance of the generalization error. We also know that the least squares MLE is unbiased (??).

Thus the question arises if we can introduce a bit of bias into the MLE in turn of decreasing the variance?
 \Rightarrow use Bayes rule (??) to introduce a bias into our model via. a **Prior** distribution.

2.1. Prior Distribution

Definition 3.6 Prior (Distribution) $\pi(\theta) = p(\theta)$:

Assumes: that the model parameters θ are no longer constant but random variables distributed according to a prior distribution that models some prior belief/bias that we have about the model:

$$\theta \sim \pi(\theta) = p(\theta) \quad (3.5)$$

Notes

In this section we use the terms model parameters θ and model as synonymous, as the model is fully described by its population parameters (^[def. 2.15] θ).

Corollary 3.3 The prior is independent of the data: The prior $p(\theta)$ models a prior belief/bias and is thus independent of the data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$:

$$p(\theta|\mathbf{X}) = p(\theta) \quad (3.6)$$

Definition 3.7 Hyperparameters $p_\lambda(\theta)$: In most cases the prior distribution are parameterized that is the pdf $\pi(\theta|\lambda)$ depends on a set of parameters λ . The parameters of the prior distribution, are called hyperparameters and are supplied due to believe/prior knowledge (and do not depend on the data) see example 3.1

2.2. Posterior Distribution

Definition 3.8 Posterior Distribution $p(\theta|\text{data})$: The posterior distribution $p(\theta|\text{data})$ is a probability distribution that describes the relationship of a unknown parameter θ a posterior/after observing evidence of a random quantity \mathbf{Z} that is in a relation with θ :

$$p(\theta|\text{data}) = p(\theta|\mathbf{Z}) \quad (3.7)$$

Definition 3.9

Posterior Distribution and Bayes Theorem:

Using Bayes theorem 19.3 we can write the posterior distribution as a product of the *likelihood*^[def. 3.1] weighted with our *prior*^[def. 3.6] and normalized by the *evidence* $\mathbf{Z} = \{\mathbf{X}, \mathbf{y}\}$ s.t. we obtain a real probability distribution:

$$p(\theta|\text{data}) = p(\theta|\mathbf{Z}) = \frac{p(\mathbf{Z}|\theta) \cdot p_\lambda(\theta)}{p(\mathbf{Z})} \quad (3.8)$$

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Normalization}} \quad (3.9)$$

$$p(\theta|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\theta, \mathbf{X}) \cdot p_\lambda(\theta)}{p(\mathbf{y}|\mathbf{X})} \quad (3.10)$$

see proof section 3

2.2.1. Maximization – MAP

We do not care about the full posterior probability distribution as in Bayesian Inference (section 9). We only want to find a point estimator ?? θ^* that maximizes the posterior distribution.

2.2.2. Maximization

Definition 3.10

Maximum a-Posteriori Estimates (MAP):

Is model/parameters θ that maximize the posterior probability distribution:

$$\theta_{\text{MAP}}^* = \arg \max_{\theta} p(\theta|\mathbf{X}, \mathbf{y}) \quad (3.11)$$

Log-MAP estimator:

$$\theta^* = \arg \max_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \quad (3.12)$$

$$= \arg \max_{\theta} \left\{ \frac{p(\mathbf{y}|\mathbf{X}, \theta) \cdot p_\lambda(\theta)}{p(\mathbf{y}|\mathbf{X})} \right\}$$

$$\stackrel{\text{eq. (11.48)}}{\propto} \arg \max_{\theta} \{p(\mathbf{y}|\theta, \mathbf{X}) \cdot p_\lambda(\theta)\}$$

Corollary 3.4 Negative Log MAP:

$$\theta^* = \arg \max_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \quad (3.13)$$

$$= \arg \min_{\theta} - \log \overbrace{p(\theta)}^{\text{Prior}} - \log \overbrace{p(\mathbf{y}|\theta, \mathbf{X})}^{\text{Likelihood}} + \log \underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{not depending on } \theta}$$

3. Proofs

Proof. 3.10:

$$\begin{aligned} p(\mathbf{X}, \mathbf{y}, \theta) &= \left\{ \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \right\} \\ \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} &= \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \\ &= \frac{p(\mathbf{y}|\mathbf{X}, \theta)p(\theta|\mathbf{X})p(\mathbf{X})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \\ &\stackrel{\text{eq. (3.6)}}{=} \frac{p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)p(\mathbf{X})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \\ &\Rightarrow \underline{p(\theta|\mathbf{X}, \mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)p(\mathbf{X})}{p(\mathbf{y}|\mathbf{X})p(\mathbf{X})} \end{aligned}$$

□

Note

This can also be derived by using the normal Bayes rule but additionally condition everything on \mathbf{X} (where the prior is independent on \mathbf{X})

4. Examples

Example 3.1 Hyperparameters Gaussian Prior:

$$f_\lambda(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta - \mu)^2}{2\sigma^2}\right)$$

with the hyperparameter $\lambda = (\mu \ \sigma^2)^\top$.

Regression

5. Linear/Ordinary Least Squares (OLS)

Definition 3.11 Linear Regression: Refers to regression that is linear w.r.t. to the parameter vector w (but not necessarily the data):

y = phi(x)^T w (3.14)

6. MLE with linear Model & Gaussian Noise

6.1. MLE for conditional linear Gaussians

Questions: what is P(Y|X) if we assume a relationship of the form: We can use the MLE to estimate the parameters theta_R^k of a model/distribution h s.t.

y approx h(X; theta) iff y = h(X; theta) + epsilon
X: set of explicative variables. epsilon: noise/error term.

Lemma 3.1 : The conditional distribution D of Y given X is equivalent to the unconditional distribution of the noise epsilon:
P(Y|X) ~ D iff epsilon ~ D

Example: Conditional linear Gaussian

Assume: a linear model h(x) = w^T x and Gaussian noise epsilon ~ N(0, sigma^2)

With E[epsilon] = 0 and yi = w^T xi + epsilon, as well as ?? it follows:
y ~ P(Y = y|X = x, theta) ~ N(mu = h(x), sigma^2)
with: theta = (w^T sigma^T)^T in R^(n+1)

Hence Y is distributed as a linear transformation of the X variable plus some Gaussian noise epsilon: yi ~ N(w^T xi, sigma^2) => Conditional linear Gaussian.

if we consider an i.i.d. sample {yi, xi}^n_{i=1}, the corresponding conditional (log-)likelihood is defined to be:

Ln(Y|X, theta) = P-hat(Y1, ..., Yn | x1, ..., xn, theta)
= product_{i=1}^n P-hat(Y|X|xi, theta) = product_{i=1}^n N(w^T xi, sigma^2)
= product_{i=1}^n 1/(sqrt(sigma^2) * 2 * pi) * exp(-(yi - w^T xi)^2 / (2 * sigma^2))
= (sigma^2 * 2 * pi)^(-n/2) * exp(-1/(2 * sigma^2) * sum_{i=1}^n (yi - w^T xi)^2)

ln(Y|X, theta) = -n/2 * ln(sigma^2) - n/2 * ln(2 * pi) - 1/(2 * sigma^2) * sum_{i=1}^n (yi - w^T xi)^2

theta* = arg max_{w in R^d, sigma^2 in R+} ln(Y|X, theta)

partial ln(Y|X, theta) / partial theta = (partial ln(Y|X, theta) / partial w1, ..., partial ln(Y|X, theta) / partial wd, partial ln(Y|X, theta) / partial sigma^2)^T = (0d, 0)^T

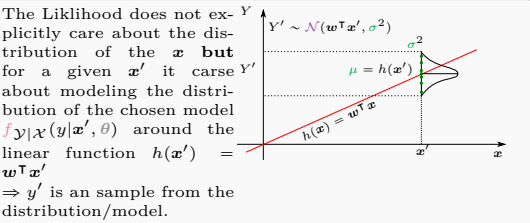
partial ln(Y|X, theta) / partial w = 1/sigma^2 * sum_{i=1}^n xi (yi - w^T xi) = 0 in R^d
= (sum_{i=1}^n xi xi^T) * w = sum_{i=1}^n xi yi

partial ln(Y|X, theta) / partial sigma^2 = -n/(2 * sigma^2) + 1/(2 * sigma^4) * sum_{i=1}^n (yi - w^T xi)^2 = 0

theta* = (w*, sigma*^2) = ((sum_{i=1}^n xi xi^T)^-1 * (sum_{i=1}^n xi yi), 1/n * sum_{i=1}^n (yi - w*^T xi)^2) (3.15)

Note

- The mean mu of the normal distribution follows from: E[w^T xi + epsilon] = E[w^T xi] + E[epsilon] = w^T xi
const = 0
- The noise epsilon must have zero mean, otherwise it wouldn't be random anymore.
- The optimal function h*(x) determines the mean mu.
- We can also minimize: theta* = arg max_theta P(Y|X, theta) = arg min_theta -P-hat(Y|X, theta)



6.2. Conditional MLE equals Least Squares

Assuming that the noise is i.i.d. Gaussian with constant variance sigma, that is theta = (w, sigma^2)^T and considering the negative log likelihood in order to minimize arg max_alpha = -arg min_alpha:

-ln(w) = -product_{i=1}^n ln(N(w^T xi, sigma^2)) = -n/2 * ln(2 * pi * sigma^2) + sum_{i=1}^n (yi - w^T xi)^2 / (2 * sigma^2)

arg max_w ln(w) iff arg min_w -ln(w)

arg min_w 1/sigma^2 * sum_{i=1}^n (yi - w^T xi)^2 = arg min_w sum_{i=1}^n (yi - w^T xi)^2 (3.16)

Thus Least squares regression equals Conditional MLE with a linear model + Gaussian noise. Maximizing Likelihood iff Minimizing least squares

Corollary 3.5 : The Maximum Likelihood Estimate (MLE) for i.i.d. Gaussian noise (and general models) is given by the squared loss/Least squares solution, assuming that the variance is constant.

Heuristics for [def. 2.12]

Consider a sample {y1, ..., yn} ~ i.i.d. N(mu, sigma^2)

partial ln(y|x, theta) / partial mu = 1/sigma^2 * sum_{i=1}^n (yi - mu) = 0

partial ln(y|x, theta) / partial sigma^2 = -n/(2 * sigma^2) + 1/(2 * sigma^4) * sum_{i=1}^n (yi - mu)^2 = 0

theta* = (mu*, sigma*^2) = ((1/n * sum_{i=1}^n yi, 1/n * sum_{i=1}^n (yi - yi_bar)^2)

So, the optimal MLE correspond to the empirical mean and the variance.

Note

partial w^T x / partial w = x, partial x^T w / partial w = x

6.3. MLE for general conditional Gaussians

Suppose we do not just want to fit linear functions but a general class of models Hsp := {h : X -> R} e.g. neural networks, kernel functions,...

Given: data D = {(x1, y1), ..., (xn, yn)} The MLE for general models h and i.i.d. Gaussian noise:

h ~ P-hat(Y|X) (Y = y|X = x, theta) = N(y|h*(x), sigma^2)

Is given by the least squares solution:

h* = arg min_{h in H} sum_{i=1}^n (yi - h(xi))^2

E.g. for linear models H = {h(x) = w^T x with parameter w}

Other distributions

If we use other distributions instead of Gaussian noise, we obtain other loss functions e.g. L1-Norm for Poisson Distribution. => if we know something about the distribution of the data we know which loss function we should chose.

7. Gaussian MAP

Classification

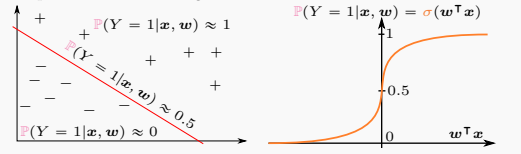
8. Logistic Regression

Bern(y; sigma(w^T x, sigma^2))

Idea: in order to classify dichotomies we use a distribution that maps probabilities to a binary values 0/1 => Bernoulli Distribution [def. 2.122].

Problem: we need to convert/translate distance w^T x into probability in order to use a bernouli distribution.

Idea: use a sigmoidal function to convert distances z := w^T x into probabilities => Logistic Function [def. 3.12].



8.1. Logistic Function

Definition 3.12 Sigmoid/Logistic Function:

sigma(z) = 1 / (1 + e^-z) = 1 / (1 + e^neg. dist. from deci. boundary) (3.18)

Explanation 3.1 (Sigmoid/Logistic Function).

sigma(z) = { 0 if z large, 1 if z large, 0.5 if z = 0 }

8.2. Logistic Regression

Definition 3.13 Logistic Regression:

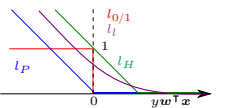
models the likelihood of the output y as a Bernoulli Distribution [def. 2.122] y ~ Bern(p), where the probability p is given by the Sigmoid function [def. 3.12] of a linear regression:

p(y|x, w) = Bern(sigma(w^T x)) = { 1/(1+e^-w^T x) if y = +1, 1/(1+e^-w^T x) if y = -1 }
section 9: 1 / (1 + exp(-y * w^T x)) = sigma(-y * w^T x) (3.19)

8.2.1. Maximum Likelihood Estimate

Definition 3.14 Logistic Loss l_l proof 9: Is the objective we want to minimize when performing mle [def. 3.3] for a logistic regression likelihood and incurs higher cost for samples closer to the decision boundary:

l_l(w; x, y) := log(1 + exp(-y * w^T x)) (3.20)
alpha log(1 + e^z) = { z for large z, 0 for small z }



Corollary 3.6 MLE for Logistic Regression:

ln(w) = sum_{i=1}^n l_l = sum_{i=1}^n log(1 + exp(-yi * w^T xi)) (3.21)

Stochastic Gradient Descent

The logistic loss l_l is a convex function. Thus we can use convex optimization techniques s.a. SGD in order to minimize the objective corollary 3.6.

Definition 3.15 proof 9

Logistic Loss Gradient
nabla_w l_l(w) = P(Y = -y|x, w) * (-y * x)
= 1 / (1 + exp(yw^T x)) * (-y * x) (3.22)

Explanation 3.2.

nabla_w l_l(w) = P(Y = -y|x, w) * (-y * x) * alpha nabla_w l_H(w)
The logistic loss l_l is equal to the hinge loss l_h but weighted by the probability of being in the wrong class P(Y = -1|x, w). Thus the more likely we are in the wrong class the bigger the step we take:

P(Y = -y|y-hat = w^T x) = { up take big step, down take small step }

Algorithm 3.1 Vanilla SGD for Logistic Regression:

Initialize: w
1: for 1, 2, ..., T do
2: Pick (x, y) unif. at random from data D
3: P(Y = -y|x, w) = 1 / (1 + exp(-y * w^T x)) = sigma(y * w^T x)
 compute prob. of misclassif. with cur. model
4: w = w + eta * y * x * sigma(y * w^T x)
5: end for

Making Predictions

Given an optimal parameter vector w-hat found by algorithm 3.1 we can predict the output of a new label by eq. (3.19):

P(y|x, w-hat) = 1 / (1 + exp(-y * w-hat^T x)) (3.23)

Drawback

Logistic regression, does not tell us anything about the likelihood P(Y = 1|x') of a point, thus it will not be able to detect outliers, as it will assign a very high probability to all correctly classified points, far from the decision boundary.

8.2.2. Maximum a-Posteriori Estimates

adapt

8.3. Logistic regression and regularization

Adding Priors to Logistic Likelihood

• **L2 (Gaussian prior):**

arg min_w sum_{i=1}^n log(1 + exp(-yi * w^T xi)) + lambda ||w||_2^2

• **L1 (Laplace prior):**

arg min_w sum_{i=1}^n log(1 + exp(-yi * w^T xi)) + lambda ||w||_1

• **Generalized:**

w-hat = arg min_w sum_{i=1}^n log(1 + exp(-yi * w^T xi)) + lambda C(w)
= arg max_w P(w|X, Y)

8.4. SGD for L2-gregularized logistic regression

Initialize: \boldsymbol{w}

1: **for** $1, 2, \dots, T$ **do**

2: Pick (\boldsymbol{x}, y) unif. at randomn from data \mathcal{D}

3: $\hat{\mathbb{P}}(Y = -y|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{(1+\exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}))}$

▷ compute prob. of misclassif. with cur. model

4: $\boldsymbol{w} = \boldsymbol{w}(1 - 2\lambda \eta_t) + \eta_t y \boldsymbol{x}$ $\hat{\mathbb{P}}(Y = -y|\boldsymbol{x}, \boldsymbol{w})$

5: **end for**

Thus: \boldsymbol{w} is pulled/shrunken towards zero, depending on the regularization parameter $\lambda > 0$

9. Proofs

Proof. [def. 3.13] We need to only proof the second expression, as the first one is fulfilled anyway:

$$1 - \frac{1}{1 + e^z} = \frac{1 + e^z}{1 + e^z} - \frac{1}{1 + e^z} = \frac{e^z + 1 - 1}{1 + e^z} = \frac{e^z}{e^z + 1}$$

$$= \frac{1}{1 + e^{-z}}$$

□

Proof. [def. 3.14]

$$l_n(\boldsymbol{w}) = \arg \max_{\boldsymbol{w}} \mathbb{P}(y_{1:n}|\boldsymbol{x}_{1:n}, \boldsymbol{w}) = \arg \min_{\boldsymbol{w}} -\log \mathbb{P}(Y|\boldsymbol{X}, \boldsymbol{w})$$

$$\stackrel{\text{i.i.d.}}{=} \arg \min_{\boldsymbol{w}} \sum_{i=1}^n -\log \mathbb{P}(y_i|\boldsymbol{x}_i, \boldsymbol{w})$$

$$\stackrel{\text{eq. (3.19)}}{=} -\log \frac{1}{1 + \exp(-y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i)}$$

$$= \log(1 + \exp(-y_i \cdot \boldsymbol{w}^\top \boldsymbol{x}_i)) =: l_l(\boldsymbol{w})$$

□

Proof. [def. 3.15]

$$\nabla_{\boldsymbol{w}} l_l(\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}} \log(1 + \exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}))$$

$$\stackrel{\text{C.R.}}{=} \frac{1}{(1 + \exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}))} \frac{\partial}{\partial \boldsymbol{w}} (1 + \exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}))$$

$$\stackrel{\text{C.R.}}{=} \frac{1}{(1 + \exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}))} \exp(-y \cdot \boldsymbol{w}^\top \boldsymbol{x}) \cdot (-y \boldsymbol{x})$$

$$= \frac{e^{-z} \cdot (-y \boldsymbol{x})}{(1 + e^{-z})} = \frac{-y \boldsymbol{x}}{e^z (1 + e^{-z})} = \frac{-y \boldsymbol{x}}{(e^z + e^{-z} + z)}$$

$$= \frac{1}{\exp(y \cdot \boldsymbol{w}^\top \boldsymbol{x}) + 1} \cdot (-y \boldsymbol{x})$$

$$\stackrel{\text{eq. (3.19)}}{=} \hat{\mathbb{P}}(Y = -y|\boldsymbol{x}, \boldsymbol{w}) \cdot (-y \boldsymbol{x})$$

□

Bayesian Inference/Modeling

Definition 3.16 Bayesian Inference: So far we only really looked at point estimators/estimates^[def. 21.47]. But what if we are interested not only into the most likely value but also want to have a notion of the uncertainty of our prediction? Bayesian inference refers to statistical inference^[def. 2.13], where uncertainty in inferences is quantified using probability. Thus we usually obtain a distribution over our parameters and not a single point estimates \Rightarrow can deduce statistical properties of parameters from their distributions.

Definition 3.17 $p(w|y, X)/p(w|D)$
Posterior Probability Distribution:
 ① Specify the prior $p_\lambda(w)$
 ② Specify the likelihood $p(y|w, X)/p(D|w)$
 ③ Calculate the evidence $p(y|X)/p(D)$
 ④ Calculate the posterior distribution $\mathbb{P}(w|y, X)/p(w|D)$

$$p(w|y, X) = \frac{p(y|w, X) \cdot p_\lambda(w)}{p(y|X)} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Normalization}}$$

Definition 3.18 $p(y|X)/p(D)$
Marginal Likelihood [see proof 4]: is the normalization constant that makes sure that the posterior distribution^[def. 3.17] is an true probability distribution:

$$p(y|X) = \int p(y|w, X) \cdot p_\lambda(w) dw = \int \text{Likelihood} \cdot \text{Prior} dw \quad (3.24)$$

Note
 It is called marginal likelihood as we marginalize over all possible parameter values.

Definition 3.19 $p(f_*|x_*, X, y)/p(f_*|y)$
Posterior Predictive Distribution: is the distribution of a real process f (i.e. $f(x) = x^\top w$) given:
 • new observation(s) x_*
 • the posterior distribution^[def. 3.17] of the observed data $D = \{X, y\}$
 • The likelihood of a real process f_*

$$p(f_*|x_*, X, y) = \int p(f_*|x_*, w) \cdot p(w|X, y) dw \quad (3.25)$$

it is calculated by weighting the likelihood^[def. 3.1] of the new observation x_* with the posterior of the observed data and averaging over all parameter values w see section 4.
 \Rightarrow obtain a distribution not depending on w .

Note f vs. y
 • Usually f denotes the model i.e.:
 $f(x) = x^\top w$ or $f(x) = \phi(x)^\top w$
 and y the model plus the noise $y = f(x) + \epsilon$.
 • Sometime people also write only: $p(y_*|x_*, X, y)$

10. Types of Uncertainty

Definition 3.20 Epistemic/Systematic Uncertainty:
 Is the uncertainty that is due to finite sub sample of the data and thus about the true data-generating probability distribution. The epistemic noise will decrease the more data we have.

Definition 3.21 Aleatoric/Statistical Uncertainty:
 Is the uncertainty of an underlying random process/model. The aleatoric uncertainty stems from the fact that we create random process models. If we run our *trained* model multiple times with the *same* input X data we will end up with different outcomes \hat{y} .
 The aleatoric noise is *irreducible* as it is an underlying part of probabilistic models.

Bayesian Filtering

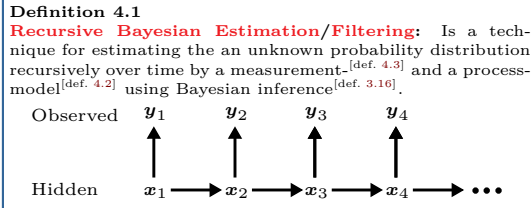


Figure 4: This problem corresponds to a *hidden Markov model* (HMM)??
 $x_t = (x_{t,1} \quad \dots \quad x_{t,n}) \quad y_t = (y_{t,1} \quad \dots \quad y_{t,m})$

Note
 Comes from the idea that spam can be filtered out by the probability of certain words.

Definition 4.2 $x_{t+1} \sim p(x_t|x_{t-1})$
Process/Motion/Dynamic Model: is a model q of how our system state x_t evolves and is usually fraught with some uncertainty.

Corollary 4.1 Markov Property $x_t \perp x_{1:t-2}|x_{t-1}$: The process models^[def. 4.2] is Markovian^[def. 22.10] i.e. the current state depends only on the previous state:

$$p(x_t|x_{1:t-1}) = p(x_t|x_{t-1}) \quad (4.1)$$

Definition 4.3 $y_t \sim p(y_t|x_t)$
Measurement/Sensor-Model/Likelihood: is a model q that maps observations/sensor measurements of our model y_t to the model state x_t

Corollary 4.2 $y_t \perp y_{1:t-1}x_{1:t-1}|x_t$
Conditional Independent Measurements: The measurements y_t are conditionally independent of the previous observations $y_{1:t-1}$ given the current state x_t :

$$p(y_t|y_{1:t-1}, x_t) = p(y_t|x_t) \quad (4.2)$$

Goal
 We want to combine the process model^[def. 4.2] and the measurement model^[def. 4.3] in a recursive way to obtain a good estimate of our model state:

$$\left. \begin{array}{l} p(x_t|x_{t-1}) \\ p(y_t|x_t) \end{array} \right\} p(x_t|y_{1:t}) \xrightarrow[\text{recursion rule}]{p(x_{t+1}|y_{1:t})} p(x_{t+1}|y_{1:t+1})$$

Definition 4.4 Chapman-Kolmogorov eq. $p(x_t|y_{1:t-1})$
Prior Update/Prediction Step [see section 4]:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1}) dx_{t-1} \quad (4.3)$$
Prior Distribution:

$$p(x_0|y_{0-1}) = p(x_0) = p_0 \quad (4.4)$$

Definition 4.5 $p(x_t|y_{1:t})$
Posterior Distribution/Update Step [see section 4]:

$$p(x_t|y_{1:t}) = \frac{1}{Z_t} p(y_t|x_t) p(x_t|y_{1:t-1}) \quad (4.5)$$

Definition 4.6 Normalization [see proof 4]:

$$Z_t = p(y_t|y_{1:t-1}) = \int p(y_t|x_t)p(x_t|y_{1:t-1}) dx_t \quad (4.6)$$

Algorithm 4.1 Optimal Bayesian Filtering:

```

1: Input:  $p(x_0)$ 
2: while Stopping Criterion not full-filled do
3:   Prediction Step:

$$p(x_t|y_{1:t}) = \frac{1}{Z_t} p(y_t|x_t)p(x_t|y_{1:t-1})$$

4:   Update Step:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1}) dx_{t-1}$$

   with:

$$Z_t = \int p(y_t|x_t)p(x_t|y_{1:t-1}) dx_t$$

5: end while
  
```

Corollary 4.3 [see proof 4]
Joint Probability Distribution of (HMM): we can also calculate the joint probability distribution of the (HMM):

$$p(x_{1:t}, y_{1:t}) = p(x_1)p(y_1|x_1) \prod_{i=2}^t p(x_i|x_{i-1})p(y_i|x_i) \quad (4.7)$$

Example 4.1 Types of Bayesian Filtering:
 • **Kalman Filter:** assumes a *linear* system, q, h are linear and Gaussian noise v, w .
 • **Extended Kalman Filter:** assumes a *non-linear* system, q, h are non-linear and Gaussian noise v, w .
 • **Particle Filter:** assumes a *non-linear* system q, h are non-linear and Non-Gaussian noise v, w , especially multi-modal distributions.

1. Kalman Filters

Definition 4.7 Kalman Filter Assumptions: Assumes a *linear*^[def. 11.13] process model^[def. 4.2], q with Gaussian model-noise v and a linear measurement model^[def. 4.3] h with Gaussian process-noise w .

Definition 4.8 Kalman Filter Model:
Process Model (4.8)

$$x^{(k)} = A^{(k-1)}x^{(k-1)} + u^{(k-1)} + v[k-1] \quad \text{with}$$

$$x^{(0)} \sim \mathcal{N}(x_0, P_0) \quad \text{and} \quad v^{(k)} \sim \mathcal{N}(0, Q^{(k)}) \quad (4.9)$$

Measurement Model

$$z^{(k)} = H^{(k)}x^{(k)} + w^{(k-1)} \quad \text{with} \quad w^{(k)} \sim \mathcal{N}(0, R^{(k)})$$

and define:

$$\hat{x}_p^{(k)} := \mathbb{E}[x_p^{(k)}] \quad \text{and} \quad P_p^{(k)} := \mathbb{V}\left[x_p^{(k)}\right] \quad (4.10)$$

$$\hat{x}_m^{(k)} := \mathbb{E}[x_m^{(k)}] \quad \text{and} \quad P_m^{(k)} := \mathbb{V}\left[x_m^{(k)}\right] \quad (4.11)$$

Note
 The CRVs $x_0, \{v(\cdot)\}, \{w(\cdot)\}$ are mutually independent.

old Kalman algorithm (in slides Joseph Form 1 think)

Gaussian Processes (GP)

1. Gaussian Process Regression

1.1. Gaussian Linear Regression

Given
 ① Linear Model with Gaussian Noise:

$$f(x) = w^\top x \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2 I) \quad (5.1)$$

$$y = f(x) + \epsilon$$
 \Rightarrow Gaussian Likelihood: $p(y|X, w) = \mathcal{N}(Xw, \sigma_n^2 I)$
 ② Gaussian Prior: $p(w) = \mathcal{N}(0, \Sigma_p)$

Sought
 ① Posterior Distribution: $p(w|y, X)$
 ② Posterior Predictive Distribution: $p(f_*|x_*, X, y)$

Definition 5.1 $p(w|y, X) = \mathcal{N}(\bar{w}, \Sigma_w^{-1})$
Posterior Distribution proof 4:

$$\mu_w = \frac{1}{\sigma_n^2} \Sigma_w^{-1} X y \quad \Sigma_w = \frac{1}{\sigma_n^2} X X^\top + \Sigma_p^{-1}$$

Note
 We could also use a prior with non-zero mean $p(w) = \mathcal{N}(\mu, \Sigma_p)$ but by convention w.o.l.g. we use zero mean see ??.

Definition 5.2 $p(f_*|x_*, X, y) = \mathcal{N}(\mu_*, \Sigma_*)$
Posterior Predictive Distribution proof 4:

$$\mu_* = \frac{1}{\sigma_n^2} x_*^\top \Sigma_w^{-1} X y \quad \Sigma_* = x_*^\top \Sigma_w^{-1} x_* \quad (5.2)$$

1.2. Kernelized Gaussian Linear Regression

Definition 5.3 Posterior Predictive Distribution:

$$p(f_*|x_*, X, y) = \mathcal{N}(\mu_*, \Sigma_*) \quad (5.3)$$

$$\mu_* \quad (5.4)$$

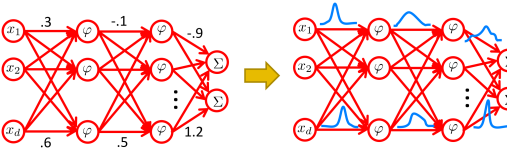
2. Model Selection

2.1. Marginal Likelihood

Approximate Inference

Problem In statistical inference we often want to calculate integrals of probability distributions i.e. • Expectations $\mathbb{E}_{X \sim p}[g(X)] = \int g(x)p(x) dx$ • Normalization constants: $p(\theta y) = \frac{1}{Z} p(\theta, y) \quad Z = \int p(y \theta)p(\theta) d\theta$ $= \int p(\theta) \prod_{i=1}^n p(y_i x_i, \theta) d\theta$ For non-linear distributions this integrals are in general intractable which may be due to the fact that there exist no analytic form of the distribution we want to integrate or highly dimensional latent spaces that prohibits numerical integration (curse of dimensionality).	
Definition 6.1 Approximate Inference: Is the procedure of finding an probability distribution q that approximates a true probability distribution p as well as possible.	
1. Variational Inference	
Definition 6.2 Bayes Variational Inference: Given an unnormalized (posterior) probability distribution: $p(\theta y) = \frac{1}{Z} p(\theta, y) \quad (6.1)$ seeks an <i>approximate</i> probability distribution q_λ , that is parameterized by a <i>variational parameter</i> λ and approximates $p(\theta y)$ well.	
Definition 6.3 Variational Family of Distributions Q: a set of probability distributions Q that is parameterized by the same <i>variational parameter</i> λ is called a variational family.	
1.1. Laplace Approximation	
Definition 6.4 example 6.1/proofs 4.4,4.4 Laplace Approximation: Tries to approximate a desired probability distribution $p(\theta \mathcal{D})$ by a Gaussian probability distribution: $Q = \{q_\lambda(\theta) = \mathcal{N}(\lambda)\} = \mathcal{N}(\mu, \Sigma) \quad (6.2)$ the distribution is given by: $q(\theta) = c \cdot \mathcal{N}(\theta; \lambda_1, \lambda_2) \quad (6.3)$ with $\lambda_1 = \hat{\theta} = \arg \max_{\theta} p(\theta y)$ $\lambda_2 = \Sigma = H^{-1}(\hat{\theta}) = -\nabla \nabla_{\theta} \log p(\hat{\theta} y)$	
Corollary 6.1 : Taylor approximation of a function $p(\theta y) \in \mathcal{C}^k$ around its mode $\hat{\theta}$ naturally induces a Gaussian approximation. See proofs 4.4,4.4	
2. Black Box Stochastic Variational Inference	
The most common way of finding q_λ is by minimizing the KL-divergence ^(def. 2.6) between our approximate distribution q and our true posterior p : $q^* \in \arg \min_{q \in Q} \text{KL}(q(\theta) \parallel p(\theta y)) = \arg \min_{\lambda \in \mathbb{R}^d} \text{KL}(q_\lambda(\theta) \parallel p(\theta y))$	
Note Usually we want to minimize $\text{KL}(p(\theta y) \parallel q(\theta))$ but this is often infeasible s.t. we only minimize $\text{KL}(q(\theta) \parallel p(\theta y))$	
Definition 6.5 ELBO-Optimization Problem proof 4: $q_\lambda^* \in \arg \min_{\lambda: q_\lambda \in Q} \text{KL}(q_\lambda(\theta) \parallel p(\theta y))$ $= \arg \max_{\lambda: q_\lambda \in Q} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y, \theta)] + H(q_\lambda) \quad (6.4)$ $= \arg \max_{\lambda: q_\lambda \in Q} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] - \text{KL}(q_\lambda(\theta) \parallel p(\theta)) \quad (6.5)$ $:= \arg \max_{\lambda: q_\lambda \in Q} \text{ELBO}(\lambda) \quad (6.6)$	

Attention: Sometimes people write simply p for the posterior and $p(\cdot)$ for prior.	
Explanation 6.1. • eq. (6.4): <ul style="list-style-type: none"> prefer uncertain approximations i.e. we maximize $H(q)$ that jointly make the joint posterior likely • eq. (6.6): Expected likelihood of our posterior over q minus a regularization term that makes sure that we are not too far away from the prior.	
2.1. Expected Lower Bound of Evidence (ELBO)	
Definition 6.6 example 6.2/proof 4 Expected Lower Bound of Evidence (ELBO): The evidence lower bound is a bound on the log prior: $\text{ELBO}(q_\lambda) \leq \log p(y) \quad (6.7)$	
2.1.1. Maximizing The ELBO	
Definition 6.7 Gradient of the ELBO Loss: $\nabla_{\lambda} L(\lambda) = \nabla_{\lambda} \text{ELBO}(\lambda) \quad (6.8)$ $= \nabla_{\lambda} \left[\mathbb{E}_{\theta \sim q_\lambda} [\log p(y, \theta)] + H(q_\lambda) \right]$ $= \nabla_{\lambda} \left[\mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] - \text{KL}(q_\lambda(\theta) \parallel p(\theta)) \right]$ $= \nabla_{\lambda} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] - \nabla_{\lambda} \text{KL}(q_\lambda(\theta) \parallel p(\theta))$	
Problem In order to use SGD we need to evaluate the gradient of the loss: $\nabla_{\lambda} \mathbb{E} [l(\theta; \mathbf{x})] = \mathbb{E} [\nabla_{\mathbf{x} \sim p} l(\theta; \mathbf{x})] = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x} \sim p} l(\theta; \mathbf{x}_i)$ however in eq. (6.8) only second term can be derived easily. For the first term we cannot move the gradient inside the expectation as the expectations depends on the parameter w.r.t. which we differentiate: $\nabla_{\lambda} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] = \frac{\partial}{\partial \lambda} \int q_\lambda \log p(y \theta) d\theta$ Solutions: • Score Gradients • Reparameterization Trick: reparameterize a function s.t. it depends on another parameter and reformulate it s.t. it still returns the same value.	
2.2. The Reparameterization Trick	
Principle 6.1 proof 4 Reparameterization Trick: Let ϕ some base distribution from which we can sample and assume there exist an invertible function g s.t. $\theta = g(\epsilon, \lambda)$ then we can write θ in terms of a new distribution parameterized by $\epsilon \sim \phi(\epsilon)$: $\theta \sim q(\theta \lambda) = \phi(\epsilon) \nabla_{\epsilon} g(\epsilon; \lambda) ^{-1} \quad (6.9)$ we can then write by the law of the unconscious statistician law 19.6: $\mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] = \mathbb{E}_{\epsilon \sim \phi} [\log p(y g(\epsilon; \lambda))] \quad (6.10)$ \Rightarrow the expectations does not longer depend on λ and we can pull in the gradient! $\nabla_{\lambda} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] = \nabla_{\epsilon} \mathbb{E}_{\theta \sim \phi} [\log p(y g(\epsilon; \lambda))] \quad (6.11)$ $= \mathbb{E}_{\epsilon \sim \phi} [\nabla_{\lambda} \log p(y g(\epsilon; \lambda))] \quad (6.12)$	
Definition 6.8 example 6.3 Reparameterized ELBO Gradient ^[def. 6.7] By using the reparameterization trick principle 6.1 we can write the gradient of the ELBO as: $\nabla_{\lambda} L(\lambda) = \nabla_{\lambda} \text{ELBO}(\lambda) \quad (6.13)$ $= \nabla_{\lambda} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)] - \nabla_{\lambda} \text{KL}(q_\lambda(\theta) \parallel p(\theta))$ $= \mathbb{E}_{\epsilon \sim \phi} [\nabla_{\lambda} \log p(y g(\epsilon; \lambda))] - \nabla_{\lambda} \text{KL}(q_\lambda(\theta) \parallel p(\theta))$	
Corollary 6.2 Reparameterized ELBO for Gaussians: $\nabla_{\lambda} L(\lambda) = \nabla_{\lambda} \text{ELBO}(\lambda) \quad (6.14)$ $= \nabla_{\lambda} \mathbb{E}_{\theta \sim q_\lambda} [\log p(y \theta)]$ $= \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} [\nabla_{\mathbf{C}, \mu} \log p(y \mathbf{C}\epsilon + \mu)]$ $= \nabla_{\mathbf{C}, \mu} (\mathbf{q}(\mathbf{C}, \mu) \parallel \pi(\theta)) \quad (6.15)$	

3. Bayesian Neural Networks (BNN)	
Definition 6.9 Bayesian Neural Networks (BNN): ① Model the prior over our weights $\theta = [W^0 \dots W^L]$ by a neural network: $\theta \sim p_\lambda(\theta) = F \quad \text{with} \quad F = F^L \circ \dots \circ F^1$ $F^l = \varphi \circ \bar{F}^l = \varphi(W^l x + b^l)$ for each weight $w_{k,j}^{(0)}$ of input x_j with weight on the hidden variable $z_k^{(0)}$ with $a_i^{(0)} = \varphi(z_i^{(0)})$ it follows: $w_{k,j}^{(0)} = p_w(\lambda_{k,j}) \stackrel{\text{i.e.}}{=} \mathcal{N}(\mu_{k,j}, \sigma_{k,j}^2)$ 	
Figure 5	
② The parameters of likelihood function are modeled by the output of the network: $p(y F(\theta, \mathbf{X})) \quad \text{see example 6.4} \quad (6.16)$	
Note Recall for normal Bayesian Linear regression we had:	
Problem All the weights of the prior $p_\lambda(\theta) = F$ are correlated in some complex way see Figure 5. Thus even if the prior and likelihood are simple, the posterior will be not. \Rightarrow need to approximate the posterior $p(\theta \mathbf{y}, \mathbf{X})$ i.e. by fitting a Gaussian distribution to each weight of the posterior neural network.	
3.0.1. MAP estimates for BNN	
Definition 6.10 BNN MAP Estimate: We need to do a forward pass for each \mathbf{x}_i in order to obtain $\mu(\mathbf{x}_i; \theta)$ and $\sigma(\mathbf{x}_i; \theta)^2$: $\theta^* = \arg \max_{\theta} \{p(\theta \mathbf{X}, \mathbf{y})\} \stackrel{\text{eq. (3.13)}}{=} \arg \min_{\theta} \lambda \ \theta\ _2^2$ $= \sum_{i=1}^n \left(\frac{1}{2\sigma(\mathbf{x}_i; \theta)^2} \ y_i - \mu(\mathbf{x}_i; \theta)\ ^2 + \frac{1}{2} \log \sigma(\mathbf{x}_i; \theta)^2 \right)$	
Explanation 6.2. ^[def. 6.10] • $\frac{1}{2} \log \sigma(\mathbf{x}_i; \theta)^2$: tries to force neural network to predict small uncertainty • $\frac{1}{2\sigma(\mathbf{x}_i; \theta)^2} \ y_i - \mu(\mathbf{x}_i; \theta)\ ^2$: tries to force neural network to predict accurately but if this is not possible for certain data points the network can attenuate the loss to a larger variance.	
Definition 6.11 proof 4 MAP Gradient of BNN: $\theta_{t+1} = \theta_t (1 - 2\lambda \eta_t) - \eta_t \nabla \sum_{i=1}^n \log p(y_i x_i, \theta) \quad (6.17)$	
Note • The gradients of the objective eq. (6.20) can be calculated using auto-differentiation techniques e.g. Pytorch or Tensorflow. • The BNN MAP estimate fails to predict epistemic uncertainty ^[def. 3.20] \iff it is overconfident in regions where we haven not even seen any data. \Rightarrow need to use Bayesian approach to approximate posterior distribution.	
3.1. Variational Inference For BNN	

4. Proofs	
Proof. Definition 3.19: $p(f_* x_*, X, y) = \frac{p(f_*, x_*, X, y)}{p(x_*, X, y)}$ $= \frac{\int p(f_*, x_*, X, y, w) dw}{p(x_*, X, y)}$ eq. (19.19) $\frac{\int p(f_* x_*, X, y, w) p(x_*, X, y, w) dw}{p(x_*, X, y)}$ eq. (19.19) $\frac{\int p(f_* x_*, X, y, w) p(w x_*, X, y) p(x_*, X, y) dw}{p(x_*, X, y)}$ $= \int p(f_* x_*, X, y, w) p(w x_*, X, y) dw$ $\stackrel{\clubsuit}{=} \int p(f_* x_*, w) p(w X, y) dw$	
Note ♣ • f_* is independent of $\mathcal{D} = \{X, y\}$ given the fixed parameter w . • w does only depend on the observed data $\mathcal{D} = \{X, y\}$ and not the unseen data x_* .	
Proof. Definition 3.18: $p(y X) = \int p(y, w X) dw = \int p(y w, X) p(w X) dw$ eq. (3.6) $\int p(y w, X) p(w) dw$	
Proof. Definition 4.4: $p(x_t, x_{t-1} y_{1:t_1}) \stackrel{\text{eq. (19.19)}}{=} p(x_t x_{t-1}, y_{1:t_1}) p(x_{t-1} y_{1:t_1})$ $\stackrel{\text{independ.}}{=} p(x_t x_{t-1}) p(x_{t-1} y_{1:t_1})$ marginalization/integration over x_{t-1} gives the desired result.	
Proof. Definition 4.5: $p(x_t, y_t y_{1:t-1}) \stackrel{\text{eq. (19.23)}}{=} \begin{cases} p(x_t y_t, y_{1:t-1}) p(y_t y_{1:t-1}) \\ p(y_t x_t, y_{1:t-1}) p(x_t y_{1:t-1}) \end{cases}$ $\stackrel{\text{corollary 4.2}}{=} p(y_t x_t)$ from which follows immediately eq. (4.5).	
Proof. Definition 4.6: $p(y_t y_{1:t-1}) = \int p(y_t, x_t y_{1:t-1}) dx_t$ $= \int p(y_t x_t, y_{1:t-1}) p(x_t y_{1:t-1}) dx_t$ $\stackrel{\text{corollary 4.2}}{=} \int p(y_t x_t) p(x_t y_{1:t-1}) dx_t$	
Proof. corollary 4.3: $p(x_{1:t}, y_{1:t}) \stackrel{\text{eq. (19.19)}}{=} p(y_{1:t} x_{1:t}) p(x_{1:t})$ $\stackrel{\text{law 19.2}}{=} p(y_{1:t} x_{1:t}) p(x_t x_{1:t-1}) \dots p(x_2 x_1) p(x_1)$ $\stackrel{\text{eq. (4.1)}}{=} p(y_{1:t} x_{1:t}) \left(p(x_1) \prod_{i=2}^t p(x_i x_{i-1}) \right)$ $\stackrel{\text{law 19.2}}{=} \left(p(y_1 x_1) \dots p(y_t x_t) \right) \left(p(x_1) \prod_{i=2}^t p(x_i x_{i-1}) \right)$ $\stackrel{\text{corollary 4.2}}{=} \left(p(y_1 x_1) p(x_1) \right) \prod_{i=2}^t p(y_i x_i) p(x_i x_{i-1})$	

Proof. ^[def. 5.1]

$$\begin{aligned} & \mathbf{p}(w|\mathcal{D}) \propto \mathbf{p}(\mathcal{D}|w) \mathbf{p}(w) \\ & \propto \exp\left(-\frac{1}{2} \frac{1}{\sigma_n^2} (\mathbf{y} - \mathbf{X}w)^\top (\mathbf{y} - \mathbf{X}w)\right) \exp\left(-\frac{1}{2} w^\top \Sigma^{-1} w\right) \\ & \propto \exp\left\{-\frac{1}{2} \frac{1}{\sigma_n^2} \left(\mathbf{y}^\top \mathbf{y} - 2w^\top \mathbf{X}^\top \mathbf{y} + w^\top \mathbf{X}^\top \mathbf{X} w + \sigma_n^2 w^\top \Sigma^{-1} w\right)\right\} \\ & \propto \exp\left\{-\frac{1}{2} \frac{1}{\sigma_n^2} \left(\mathbf{y}^\top \mathbf{y} - 2w^\top \mathbf{X}^\top \mathbf{y} + \underbrace{w^\top (\mathbf{X}^\top \mathbf{X}^\top + \sigma_n^2 \Sigma^{-1}) w}_{\text{.....}}\right)\right\} \end{aligned}$$

We know that a Gaussian $\mathcal{N}(w|\bar{w}, \Sigma_w^{-1})$ should look like:

$$\begin{aligned} & \mathbf{p}(w|\mathcal{D}) \propto \exp\left(-\frac{1}{2} (w - \bar{w})^\top \Sigma_w (w - \bar{w})\right) \\ & \propto \exp\left(-\frac{1}{2} \left(\underbrace{w^\top \Sigma_w w}_{\text{.....}} - 2w^\top \Sigma_w \bar{w} + \bar{w}^\top \Sigma_w \bar{w}\right)\right) \end{aligned}$$

Σ_w follows directly $\Sigma_w = \sigma_n^{-2} \mathbf{X} \mathbf{X}^\top + \Sigma_p$

\bar{w} follows from $2w^\top \mathbf{X}^\top \mathbf{y} = 2w^\top \Sigma_w \bar{w} \Rightarrow \bar{w} = \Sigma_w^{-1} \mathbf{X}^\top \mathbf{y}$. \square

Proof. ^[def. 5.2] \square

Proof. ^[def. 6.4] In a Bayesian setting we are usually interested in maximizing the log prior/likelihood:

$$\mathcal{L}_n(\theta) = \log(\mathbf{p}(\theta|y)) = (\log \text{Prior} + \log \text{Likelihood})$$

we now approximate $\mathcal{L}_n(\theta)$ by a Taylor approximation around its maximum $\hat{\theta}$:

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\hat{\theta}) + \frac{1}{2} \frac{\partial^2 \mathcal{L}_n}{\partial \theta^2} \Big|_{\hat{\theta}} (\theta - \hat{\theta})^2 + \mathcal{O}\left((\theta - \hat{\theta})^3\right)$$

we can no derive the distribution:

$$\begin{aligned} \mathbf{p}(\theta|y) & \approx \exp(\mathcal{L}_n(\theta)) = \exp(\log \mathbf{p}(\theta|y)) \\ & = \mathbf{p}(\hat{\theta}) \exp\left(\frac{1}{2} \frac{\partial^2 \mathcal{L}_n}{\partial \theta^2} \Big|_{\hat{\theta}}\right) \\ & = \sqrt{2\pi\sigma^2} \mathbf{p}(\hat{\theta}) \mathcal{N}\left(\theta; \hat{\theta}, \sigma\right) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \mathcal{N}\left(\theta; \hat{\theta}, \sigma\right) \end{aligned}$$

\square

Notes

- the derivative of the maximum must be zero by definition $\frac{\partial \mathcal{L}_n}{\partial \theta} \Big|_{\hat{\theta}} = 0$
- we approximate the normalization constant $\frac{1}{2}$ by $\sqrt{2\pi\sigma^2} \mathbf{p}(\hat{\theta})$.

Proof. ^[def. 6.4] 2D:

$$\begin{aligned} & \nabla \mathcal{L}_n(\theta) = \nabla \mathcal{L}_n(\theta_1, \theta_2) = 0 \\ & \mathcal{L}_n(\theta) = \mathcal{L}_n(\hat{\theta}) + \frac{1}{2} \left(A(\theta_1 - \hat{\theta}_1)^2 + B(\theta_2 - \hat{\theta}_2)^2 + C(\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2) \right) \\ & \mathcal{L}_n(\theta) = \mathcal{L}_n(\hat{\theta}) + \left(\theta - \hat{\theta} \right)^\top H(\hat{\theta}) (\theta - \hat{\theta}) \\ & = \mathcal{L}_n(\hat{\theta}) + \frac{1}{2} Q(\theta) \\ & A = \frac{\partial^2 \mathcal{L}_n}{\partial \theta^2} \Big|_{\hat{\theta}} \quad B = \frac{\partial^2 \mathcal{L}_n}{\partial \theta^2} \Big|_{\hat{\theta}} \quad C = \frac{\partial^2 \mathcal{L}_n}{\partial \theta_1 \partial \theta_2} \Big|_{\hat{\theta}} \\ & H = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \quad \Sigma = H^{-1}(\hat{\theta}) \end{aligned}$$

\square

Proof. ^[def. 6.4] k -dimensional:

$$\begin{aligned} & \mathcal{L}_n(\theta) \approx \mathcal{L}_n(\hat{\theta}) + (\theta - \hat{\theta})^\top \nabla \mathcal{L}_n(\hat{\theta}) (\theta - \hat{\theta}) \\ & H(\theta) = \nabla \nabla^\top \mathcal{L}_n(\theta) \quad \Sigma = H^{-1}(\hat{\theta}) \\ & \mathbf{p}(\theta|y) = \sqrt{(2\pi)^n \det(\Sigma)} \mathbf{p}(\hat{\theta}) \mathcal{N}\left(\theta; \hat{\theta}, \Sigma\right) \\ & \approx c \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \mathcal{N}\left(\theta; \hat{\theta}, \Sigma\right) \end{aligned}$$

\square

Proof. ^[def. 6.5]

$$\begin{aligned} & q^* \in \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\theta) \parallel \mathbf{p}(\theta|y)) \\ & \mathbf{p}(\theta|y) = \frac{1}{Z} \mathbf{p}(\theta, y) \\ & = \arg \min_q \mathbb{E}_{\theta \sim q} \left[\log \frac{q(\theta)}{\frac{1}{Z} \mathbf{p}(\theta, y)} \right] \\ & = \arg \min_q \mathbb{E}_{\theta \sim q} \left[\log q(\theta) - \log \frac{1}{Z} - \log \mathbf{p}(\theta, y) \right] \\ & = \arg \min_q \mathbb{E}_{\theta \sim q} \left[-\underbrace{\log q(\theta)}_{H(q)} + \underbrace{\mathbb{E}_{\theta \sim q} [\log Z]}_{\text{.....}} - \mathbb{E}_{\theta \sim q} [\log \mathbf{p}(\theta, y)] \right] \\ & = \arg \max_q \mathbb{E}_{\theta \sim q} [\log \mathbf{p}(\theta, y)] + H(q) \\ & = \arg \max_q \mathbb{E}_{\theta \sim q} [\log \mathbf{p}(\theta|y) + \log \mathbf{p}(\theta) - \log q(\theta)] \\ & = \arg \max_q \mathbb{E}_{\theta \sim q} [\log \mathbf{p}(\theta|y)] + \text{KL}(q(\theta) \parallel \mathbf{p}(\theta)) \end{aligned}$$

\square

Proof. ^[def. 6.6]

$$\begin{aligned} \log \mathbf{p}(y) & = \log \int \mathbf{p}(y, \theta) d\theta = \log \int \mathbf{p}(y|\theta) \mathbf{p}(\theta) d\theta \\ & = \log \int \mathbf{p}(y|\theta) \frac{\mathbf{p}(\theta)}{q_\lambda(\theta)} q_\lambda(\theta) d\theta \\ & = \log \mathbb{E}_{\theta \sim q_\lambda} \left[\mathbf{p}(y|\theta) \frac{\mathbf{p}(\theta)}{q_\lambda(\theta)} \right] \\ & \stackrel{\text{eq. (19.54)}}{\geq} \mathbb{E}_{\theta \sim q_\lambda} \left[\log \left(\mathbf{p}(y|\theta) \frac{\mathbf{p}(\theta)}{q_\lambda(\theta)} \right) \right] \\ & = \mathbb{E}_{\theta \sim q_\lambda} \left[\log \mathbf{p}(y|\theta) - \log \frac{\mathbf{p}(\theta)}{q_\lambda(\theta)} \right] \\ & = \mathbb{E}_{\theta \sim q_\lambda} [\log \mathbf{p}(y|\theta)] - \text{KL}(q_\lambda \parallel \mathbf{p}(\cdot)) \end{aligned}$$

\square

Proof. principle 6.1 Let:

$$\begin{aligned} \epsilon \sim \phi(\epsilon) \quad X \sim f_X \\ \theta = g(\epsilon; \lambda) \quad \text{correspond to} \quad \mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \end{aligned}$$

then it follows immediately with eq. (19.46):

$$\begin{aligned} \theta \sim q_\lambda(\theta) & = q(\theta|\lambda) = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx}(g^{-1}(y)) \right|} \\ & = \phi(\epsilon) |\nabla_\epsilon g(\epsilon; \lambda)|^{-1} \end{aligned}$$

\Rightarrow parameterized in terms of ϵ \square

Proof. ^[def. 6.11]

$$\begin{aligned} \theta_{t+1} & = \theta_t - \eta_t \left(\nabla \log \mathbf{p}(\theta) - \nabla \sum_{i=1}^n \log \mathbf{p}(y_i | \mathbf{x}_i, \theta) \right) \\ & = \theta_t - \eta_t \left(2\lambda \theta_t - \nabla \sum_{i=1}^n \log \mathbf{p}(y_i | \mathbf{x}_i, \theta) \right) \\ & = \theta_t (1 - 2\lambda \eta_t) - \eta_t \nabla \sum_{i=1}^n \log \mathbf{p}(y_i | \mathbf{x}_i, \theta) \end{aligned}$$

\square

5. Examples

Example 6.1 Laplace Approximation Logistic Regression Likelihood + Gaussian Prior:

Example 6.2 ELBO Bayesian Logistic Regression: Suppose:

$$Q = \text{diag. Gaussians} \Rightarrow \lambda = \begin{bmatrix} \mu_{1:d} & \sigma_{1:d}^2 \end{bmatrix} \in \mathbb{R}^{2d}$$

$$\mathbf{p}(\theta) = \mathcal{N}(0, I)$$

Then it follows for the terms of the ELBO:

$$\begin{aligned} \text{KL}(q_\lambda \parallel \mathbf{p}(\theta)) & = \frac{1}{2} \sum_{i=1}^d \left(\mu_i^2 + \sigma_i^2 - 1 - \ln \sigma_i^2 \right) \\ \mathbb{E}_{\theta \sim q_\lambda} [\mathbf{p}(y|\theta)] & = \mathbb{E}_{\theta \sim q_\lambda} \left[\sum_{i=1}^n \log \mathbf{p}(y_i | \theta, \mathbf{x}_i) \right] \\ & = \mathbb{E}_{\theta \sim q_\lambda} \left[- \sum_{i=1}^n \log (1 + \exp(-y_i \theta^\top \mathbf{x}_i)) \right] \end{aligned}$$

Example 6.3 ELBO Gradient Gaussian: Suppose:

$$\begin{aligned} \theta \sim q(\theta|\lambda) & = \mathcal{N}(\theta; \mu, \Sigma) \Rightarrow \lambda = \begin{bmatrix} \mu & \Sigma \end{bmatrix} \\ \epsilon \sim \phi(\epsilon) & = \mathcal{N}(\epsilon; 0, I) \end{aligned}$$

we can reparameterize using principle 6.1 by using:

$$\theta \sim g(\epsilon, \lambda) = C\epsilon + \mu \quad \text{with} \quad C : \quad CC^\top = \Sigma$$

from this it follows: (C is the Cholesky factor of Σ)

$$g^{-1}(\theta, \lambda) = \epsilon = C^{-1}(\theta - \mu) \quad \frac{\partial g(\epsilon; \lambda)}{\partial \epsilon} = C$$

from this it follows:

$$\begin{aligned} q(\theta|\lambda) & = \frac{\phi(\epsilon)}{\left| \frac{dg(\epsilon; \theta)}{d\epsilon}(g^{-1}(\theta)) \right|} = \phi(\epsilon) |C|^{-1} \\ & \iff \phi(\epsilon) = q(\theta|\lambda) |C| \end{aligned}$$

we can then write the reparameterized expectation part of the gradient of the ELBO as:

$$\begin{aligned} \nabla_\lambda L(\lambda)_1 & = \nabla_\lambda \mathbb{E}_{\epsilon \sim \phi} [\log \mathbf{p}(y|g(\epsilon; \lambda))] \\ & = \nabla_{C, \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} [\log \mathbf{p}(y|C\epsilon + \mu)] \\ & \stackrel{\text{i.i.d.}}{=} \nabla_{C, \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[\sum_{i=1}^n \log \mathbf{p}(y_i | C\epsilon + \mu, \mathbf{x}_i) \right] \\ & = \nabla_{C, \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[n \frac{1}{n} \sum_{i=1}^n \log \mathbf{p}(y_i | C\epsilon + \mu, \mathbf{x}_i) \right] \\ & = \nabla_{C, \mu} n \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} \left[\mathbb{E}_{i \sim \mathcal{U}(\{1, n\})} \log \mathbf{p}(y_i | C\epsilon + \mu, \mathbf{x}_i) \right] \\ & \text{Draw a mini batch } \left\{ \epsilon_1^{(1)}, \dots, \epsilon_m^{(m)} \right\}_{j=1}^m \sim \mathcal{U}(\{1, n\}) \\ & = n \frac{1}{m} \sum_{j=1}^m \nabla_{C, \mu} \log \mathbf{p}(y_j | C\epsilon + \mu, \mathbf{x}_j) \end{aligned}$$

$$\begin{aligned} \nabla_\lambda L(\lambda) & = \nabla_\lambda \text{ELBO}(\lambda) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I)} [\nabla_{C, \mu} \log \mathbf{p}(y|C\epsilon + \mu)] \\ & \quad - \nabla_{C, \mu} (qC, \mu \parallel \pi(\theta)) \end{aligned}$$

Example 6.4 BNN Likelihood Function Examples:

$$\mathbf{p}(y|\mathbf{X}, \theta) = \begin{cases} \mathcal{N}(y; F(\mathbf{X}, \theta), \sigma^2) \\ \mathcal{N}(y; F(\mathbf{X}, \theta)_1, \exp F(\mathbf{X}, \theta)_1) \end{cases}$$

Kernels

Given objects we cannot assume that they are vectors/can be represented as vectors in feature space.
Hence it is also not guaranteed that those objects can be added and multiplied by scalars.
Question: then how can we define a more general notion of similarity?

Definition 7.1 Similarity Measure $\text{sim}(A, B)$: A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects.
No single definition of a similarity measure exists but often they are defined in terms of the inverse of distance metrics and they take on large values for similar objects and either zero or a negative value for very dissimilar objects.

Definition 7.2 Dissimilarity Measure $\text{dissim}(A, B)$: Is a measure of how dissimilar objects are, rather than how similar they are.
Thus it takes the largest values for objects that are really far apart from another.
Dissimilarities are often chosen as the squared norm of two difference vectors:

$$\|x - y\|^2 = x^T x + y^T y - 2x^T y \quad \forall x, y \in \mathbb{R}^d \quad (7.1)$$
$$\text{dissim}(x, y) = \text{sim}(x, x) + \text{sim}(y, y) - 2\text{dissim}(x, y)$$

Attention
It is better to rely on similarity measures instead of dissimilarity measures. Dissimilarities are often not adequate from a modeling point of view, because for objects that are really dissimilar/far from each other, we usually have the biggest problem to estimate their distance.
E.g. for a bag of words it is easy to determine similar words, but it is hard to estimate which words are most dissimilar. For normed vectors the only information of a dissimilarity defined as in eq. (7.1) becomes

$$2x^T y = 2\text{dissim}(x, y)$$

Definition 7.3 Feature Map ϕ : is a mapping $\phi: \mathcal{X} \mapsto \mathcal{Y}$ that takes an input $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and maps it into another feature space $\mathcal{Y} \subseteq \mathbb{R}^D$.

Note
Such feature maps can lead to an exponential number of terms i.e. for a polynomial feature map, with monomials of degree up to p and feature vectors of dimension $x \in \mathbb{R}^d$ we obtain a feature space of size:

$$D = \dim(\mathcal{Y}) = \binom{p+d}{d} = \mathcal{O}(d^p) \quad (7.2)$$

when using the polynomial kernel^[def. 7.10], this can be reduced to the order d .

Definition 7.4 Kernel k: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the data space. A map $k: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called kernel if there exists an inner product space^[def. 14.18] called **feature space** $(\mathcal{Y}, \langle \cdot, \cdot \rangle_{\mathcal{Y}})$ and a map $\phi: \mathcal{X} \mapsto \mathcal{Y}$ s.t.

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{Y}} \quad \forall x, y \in \mathcal{X} \quad (7.3)$$

Corollary 7.1 Kernels and similarity: Kernels are defined in terms of inner product spaces and hence they have a notion of similarity between their arguments.

Example
Let $k(x, y) := x^T A y$ **thus** the kernel measures the similarity between x and y by the inner product $x^T y$ weighted by the matrix A .
Corollary 7.2 Kernels and distance: Let $k(x, y)$ be a measure of similarity between x and y then k induces a dissimilarity/distance between x and y defined as the difference between the self-similarities $k(x, x) + k(y, y)$ and the cross-similarities $k(x, y)$:

$$\text{dissimilarity}(x, y) := k(x, x) + k(y, y) - 2k(x, y)$$

Note
The factor 2 is required to ensure that $d(x, x) = 0$.

1. The Gram Matrix

Definition 7.5 Kernel (Gram) Matrix:
Given: a mapping $\phi: \mathbb{R}^d \mapsto \mathbb{R}^D$ and a corresponding kernel function $k: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ with $\mathcal{X} \subseteq \mathbb{R}^d$.
Let S be any finite subset of data $S = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$. Then the kernel matrix $K: \mathbb{R}^{n \times n}$ is defined by:

$$K = \phi(X)\phi(X)^T = (\phi(x_1), \dots, \phi(x_n))(\phi(x_1), \dots, \phi(x_n))^T$$
$$= \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix} = \begin{pmatrix} \phi(x_1)^T \phi(x_1) & \dots & \phi(x_1)^T \phi(x_n) \\ \vdots & \ddots & \vdots \\ \phi(x_n)^T \phi(x_1) & \dots & \phi(x_n)^T \phi(x_n) \end{pmatrix}$$
$$K_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

Corollary 7.3 $V \Lambda V^T$
Kernel Eigenvector Decomposition:
For any symmetric matrix (Gram matrix $K(x_i, x_j)_{i,j=1}^n$) there exists an eigenvector decomposition:

$$K = V \Lambda V^T \quad (7.4)$$

V : orthogonal matrix of eigenvectors $(v_{t,i})_{i=1}^n$
 Λ : diagonal matrix of eigenvalues λ_i
Assuming all eigenvalues λ_t are non-negative, we can calculate the mapping:

$$\phi: x_i \mapsto \left(\sqrt{\lambda_t} v_{t,i} \right)_{t=1}^n \in \mathbb{R}^n, \quad i = 1, \dots, n \quad (7.5)$$

which allows us to define the Kernel K as:

$$\phi^T(x_i) \phi(x_j) = \sum_{t=1}^n \lambda_t v_{t,i} v_{t,j} = (V \Lambda V^T)_{i,j} = K(x_i, x_j) \quad (7.6)$$

1.1. Necessary Properties

Property 7.1 Inner Product Space:
 k must be an *inner product* of a suitable space \mathcal{Y} .

Property 7.2 Symmetry: k/K must be symmetric:
 $k(x, y) = k(y, x) = \phi(x)^T \phi(y) = \phi(y)^T \phi(x) \quad \forall x, y \in \mathcal{X}$

Property 7.3 Non-negative Eigenvalues/p.s.d.s Form:
Let $S = \{x_1, \dots, x_n\}$ be an n -set of a *finite* input space \mathcal{Y} . A kernel k must induce a *p.s.d. symmetric* kernel matrix k for any possible $S \subseteq \mathcal{X}$ see section 1.
 \iff all eigenvalues of the kernel gram matrix K for *finite* \mathcal{Y} must be non-negative corollary 14.2.

Notes
• The extension to infinite dimensional Hilbert Spaces might also include a non-negative weighting/eigenvalues:
$$\langle \phi(x), \phi(z) \rangle = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(z)$$

• In order to be able to use a kernel, we need to verify that the kernel is **p.s.d.** for all n -vectors $\mathcal{X} = \{x_1, \dots, x_n\}$, as well as for future unseen values.

2. Mercers Theorem

Theorem 7.1 Mercers Theorem: Let \mathcal{X} be a compact subset of \mathbb{R}^n and $k: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ a **kernel function**.
Then one can expand k in a uniformly convergent series of bounded functions ϕ s.t.

$$k(x, x') = \sum_{i=1}^{\infty} \lambda \phi_i(x) \phi_i(x') \quad (7.7)$$

Theorem 7.2 General Mercers Theorem: Let Ω be a compact subset of \mathbb{R}^n . Suppose k is a general continuous symmetric function such that the integral operator:

$$T_k: L_2(X) \mapsto L_2(X) \quad (T_k f)(\cdot) = \int_{\Omega} k(\cdot, x) f(x) dx \quad (7.8)$$

is **positive**, that is it satisfies:

$$\int_{\Omega \times \Omega} k(x, z) f(x) f(z) dx dz > 0 \quad \forall f \in L_2(\Omega)$$

Then we can expand $k(x, z)$ in a uniformly convergent series in terms of T_k 's eigen-functions $\phi_j \in L_2(\Omega)$, with $\|\phi_j\|_{L_2} = 1$ and **positive** associated eigenvalues $\lambda_j > 0$.

Note
All kernels satisfying Mercers conditions describe an inner product in a high dimensional space.
 \implies can replace the inner product by the kernel function.

Check if \mathbb{R} or \mathbb{R}^+ as in script

3. The Kernel Trick

Definition 7.6 Kernel Trick: If a kernel has an analytic form we do no longer need to calculate:
• the function mapping $x \mapsto \phi(x)$ and
• the inner product $\phi(x)^T \phi(y)$ explicitly but simply use the formula for the kernel:

$$\phi(x)^T \phi(x) = k(x, x) \quad (7.9)$$

see examples 7.1 and 7.2

Note
• Possible to operate in any n -dimensional function space, efficiently.
• ϕ not necessary anymore.
• Complexity independent of the functions space.

4. Types of Kernels

4.1. Stationary Kernels

Definition 7.7 Stationary Kernel: A stationary kernel is a kernel that only considers vector differences:

$$k(x, y) = k(x - y) \quad (7.10)$$

see example 7.3

4.2. Isotropic Kernels

Definition 7.8 Isotropic Kernel: A isotropic kernel is a kernel that only considers distance differences:

$$k(x, y) = k(\|x - y\|_2) \quad (7.11)$$

Corollary 7.4 :
Isotropic \rightarrow Stationary

5. Important Kernels on \mathbb{R}^d

5.1. The Linear Kernel

Definition 7.9 Linear/String Kernel:

$$k(x, y) = x^T y \quad (7.12)$$

5.2. The Polynomial Kernel

Definition 7.10 Polynomial Kernel: represents all monomials^[def. 11.2] of degree up to m

$$k(x, y) = (1 + x^T y)^m \quad (7.13)$$

5.3. The Sigmoid Kernel

Definition 7.11 Sigmoid/tanh Kernel:

$$k(x, y) = \tanh(\kappa x^T y - b) \quad (7.14)$$

5.4. The Exponential Kernel

Definition 7.12 Exponential Kernel: is a continuous kernel that is non-differential $k \in C^0$:

$$k(x, y) = \exp\left(-\frac{\|x - y\|_1}{\theta}\right) \quad (7.15)$$

$\theta \in \mathbb{R}$: corresponds to a threshold.

5.5. The Gaussian Kernel

Definition 7.13 Gaussian/Squared Exp. Kernel/Radial Basis Functions (RBF):
Is an infinite dimensional smooth kernel $k \in C^\infty$ with some useful properties

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\theta^2}\right) \approx \begin{cases} 1 & \text{if } x \text{ and } y \text{ close} \\ 0 & \text{if } x \text{ and } y \text{ far away} \end{cases} \quad (7.16)$$

Explanation 7.1 (Threshold θ). $2\theta \in \mathbb{R}$ corresponds to a threshold that determines how close input values need to be in order to be considered similar:

$$k = \exp\left(-\frac{\text{dist}^2}{2\theta^2}\right) \approx \begin{cases} 1 & \iff \text{sim} & \text{if dist} \ll \theta \\ 0 & \iff \text{dissim} & \text{if dist} \gg \theta \end{cases}$$

or in other words how much we believe in our data i.e. for smaller length scale we do trust our data less and the admissible functions vary much more.

Note
If we chose h small, all data points not close to h will be 0/discarded \iff data points are considered as independent. Length of all vectors in **feature space** is one $k(x, x) = e^0 = 1$.
Thus: Data points in input space are projected onto a high-(infinite)-dimensional sphere in feature space.
Classification: Cutting with hyperplanes through the sphere. **How to choose h :** good heuristics, take median of the distance all points but better is cross validation.

5.6. The Matern Kernel

When looking at actual data/sample paths the smoothness of the Gaussian kernel^[def. 7.13] is often a too strong assumption that does not model reality the same holds true for the non-smoothness of the exponential kernel^[def. 7.12]. A solution to this dilemma is the Matern kernel.

Definition 7.14 Matern Kernel: is a kernel which allows you to specify the level of smoothness $k \in C^{[\nu]}$ by a positive parameter ν :

$$k(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \|x - y\|_2}{\rho} \right)^\nu \mathcal{K}_\nu \left(\frac{\sqrt{2\nu} \|x - y\|_2}{\rho} \right) \quad (7.17)$$

$\nu, \rho \in \mathbb{R}_+$
 \mathcal{K}_ν modified Bessel function of the second kind

6. Kernel Engineering

Often linear and even non-linear simple kernels are not sufficient to solve certain problems, especially for pairwise problems i.e. user & product, exon & intron, ...
Composite kernels can be the solution to such problems.

6.1. Closure Properties/Composite Rules

Suppose we have two kernels:

$$k_1: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \quad k_2: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$$

defined on the data space $\mathcal{X} \subseteq \mathbb{R}^d$. Then we may define using Composite Rules:

$$k(x, x') = k_1(x, x') + k_2(x, x') \quad (7.18)$$
$$k(x, x') = k_1(x, x') \cdot k_2(x, x') \quad (7.19)$$
$$k(x, x') = \alpha k_1(x, x') \quad \alpha \in \mathbb{R}_+ \quad (7.20)$$
$$k(x, x') = f(x) f(x') \quad (7.21)$$
$$k(x, x') = k_3(\phi(x), \phi(x')) \quad (7.22)$$
$$k(x, x') = p(k(x, x')) \quad (7.23)$$
$$k(x, x') = \exp(k(x, x')) \quad (7.24)$$

Where

$f: \mathcal{X} \mapsto \mathbb{R}$	a real valued function
$p: \mathcal{X} \mapsto \mathbb{R}^e$	the explicit mapping
ϕ	a polynomial with pos. coefficients
k_3	a Kernel over $\mathbb{R}^e \times \mathbb{R}^e$

Proofs

Proof. Property 7.3 The kernel matrix is positive-semidefinite:
Let $\phi: \mathcal{X} \mapsto \mathbb{R}^d$ and $\Phi = [\phi(x_1) \dots \phi(x_n)]^T \in \mathbb{R}^{d \times n}$.
Thus: $K = \Phi^T \Phi \in \mathbb{R}^{n \times n}$.
 $v^T K v = v^T \Phi^T \Phi v = (\Phi v)^T \Phi v = \|\Phi v\|_2^2 \geq 0$

Examples

Example 7.1 Calculating the Kernel by hand:

Let :

$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$\phi(\boldsymbol{x}) \mapsto \{x_1^2, x_2^2, \sqrt{2}x_1, x_2\}$
 $\phi : \mathbb{R}^{d=2} \mapsto \mathbb{R}^{D=3}$

We can now have a decision boundary in this 3-D feature space \mathcal{Y} of ϕ as:

$$\beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 \sqrt{2}x_1 x_2 = 0$$
$$\left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}) \right\rangle$$
$$= \left\langle \left\{ x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}, x_{i2} \right\}, \left\{ x_{j1}^2, x_{j2}^2, \sqrt{2}x_{j1}, x_{j2} \right\} \right\rangle$$
$$= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2x_{i1} x_{i2} x_{j1} x_{j2}$$

Operation Count:

- 2 · 3 operations to map \boldsymbol{x}_i and \boldsymbol{x}_j into the 3D space \mathcal{Y} .
- Calculating an inner product of $\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}) \rangle$ with 3 additional operations.

Example 7.2

Calculating the Kernel using the Kernel Trick:

$$\left\langle \phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}) \right\rangle = \underbrace{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^2}_{:= \mathbf{k}(\boldsymbol{x}_i, \boldsymbol{x}_j)} = \langle \{x_{i1}, x_{i2}\}, \{x_{i1}, x_{i2}\} \rangle^2$$
$$= (x_{i1} x_{i2} + x_{j1} x_{j2})^2$$
$$= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2x_{i1} x_{i2} x_{j1} x_{j2}$$

Operation Count:

- 2 multiplicaitons of $x_{i1} x_{j1}$ and $x_{i2} x_{j2}$.
- 1 operation for taking the square of a scalar.

Conclusion

The Kernel trick needed only 3 in comparison to 9 operations.

Example 7.3 Stationary Kernels:

$$\mathbf{k}(\boldsymbol{x}, \boldsymbol{y}) = \exp \left(\frac{(\boldsymbol{x} - \boldsymbol{y})^\top M (\boldsymbol{x} - \boldsymbol{y})}{h^2} \right)$$

is a stationary but not an isotropic kernel.

Math Appendix

Logic
Set Theory

Definition 9.1 Set $A = \{1, 3, 2\}$:
is a well-defined group of distinct items that are considered as an object in its own right. The arrangement/order of the objects does not matter but each member of the set must be unique.

Definition 9.2 Empty Set $\{\}$ / \emptyset :
is the unique set having no elements/cardinality^[def. 9.4] zero.

Definition 9.3 Multiset/Bag: Is a set-like object in which multiplicity matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 9.4 Cardinality $|S|$: Is the number of elements that are contained in a set.

Definition 9.5 The Power Set $\mathcal{P}(S)/2^S$: The power set of any set S is the set of all subsets of S, including the empty set and S itself. The cardinality of the power set is 2^S is equal to $2^{|S|}$.

Definition 9.6 Closure: A set is *closed* under an operation Ω if performance of that operations onto members of the set always produces a member of that set.

1. Number Sets

1.1. The Real Numbers \mathbb{R}
1.1.1. Intervals

Definition 9.7 Closed Interval $[a, b]$:
The closed interval of a and b is the set of all real numbers that are within a and b , including a and b :
 $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$ (9.1)

Definition 9.8 Open Interval (a, b) :
The open interval of a and b is the set of all real numbers that are within a and b :
 $(a, b) = \{x \in \mathbb{R} \mid a < x \leq b\}$ (9.2)

1.2. The Rational Numbers \mathbb{Q}

Example 9.1 Power Set/Cardinality of $S = \{x, y, z\}$:
The subsets of S are:
 $\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}$
and hence the power set of S is $\mathcal{P}(S) = \{\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

Sequences&Series

Definition 10.1 Index Set: Is a set^[def. 9.1] A , whose members are labels to another set S . In other words its members index member of another set. An index set is build by enumerating the members of S using a function f s.t.
 $f : A \mapsto S \qquad A \in \mathbb{N}$ (10.1)

Definition 10.2 Sequence $(a_n)_{n \in A}$:
is an by an index set A *enumerated* multiset^[def. 9.3] (repetitions are allowed) of objects in which *order does matter*.

Definition 10.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 10.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \dots)$.
 $t_n = t_0 + nd \quad d$:difference between two terms (10.2)

1.2. Geometric Sequence

Definition 10.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \dots)$.
 $t_n = t_0 \cdot r^n \qquad r$:ratio between two terms (10.3)

Calculus and Analysis

1. Building Blocks of Analysis

1.1. Polynomials

Definition 11.1 Polynomial: A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:
$$\mathcal{P}_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1} + a_nx^n \quad (11.1)$$

Corollary 11.1 Degree n-of a Polynomial $\deg(\mathcal{P}_n)$: the *degree* of the polynomial is the highest exponent of the variable x , among all non-zero coefficients $a_i \neq 0$.

Definition 11.2 Monomial: Is a polynomial with only one term.

Definition 11.3 Quadratic Formula: $ax^2 + bx + c = 0$ or in reduced form:
 $x^2 + px + q = 0$ with $p = b/a$ and $q = c/a$

Definition 11.4 Discriminant: $\delta = b^2 - 4ac$

Definition 11.5 Solution to ^[def. 11.3]:
$$x_{\pm} = \frac{-b \pm \sqrt{\delta}}{2a} \quad \text{or} \quad x_{\pm} = \frac{1}{2} \left(-p \pm \sqrt{p^2 - 4q} \right)$$

Theorem 11.1
Fist Fundamental Theorem of Calculus: Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined $\forall x \in [a, b]$ by:

$$F(X) = \int_a^x f(t) dt \quad (11.2)$$

Then it follows:
$$F'(x) = f(x) \quad \forall x \in (a, b) \quad (11.3)$$

Theorem 11.2
Second Fundamental Theorem of Calculus: Let f be a real-valued function on a closed interval $[a, b]$ and F an antiderivative of f in $[a, b]$: $F'(x) = f(x)$, then it follows if f is Riemann integrable on $[a, b]$:

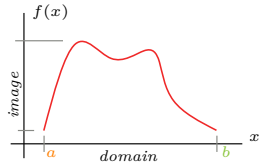
$$\int_a^b f(t) dt = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t) dt = F(x) \quad (11.4)$$

Definition 11.6 Domain of a function $\text{dom}(\cdot)$:
Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the set of all possible input values \mathcal{X} is called the domain of $f - \text{dom}(f)$.

Definition 11.7
Codomain/target set of a function $\text{codom}(\cdot)$:
Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the codaomain of that function is the set \mathcal{Y} into which all of the output of the function is constrained to fall.

Definition 11.8 Image (Range) of a function: $f[\cdot]$
Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the image of that function is the set to which the function can actually map:
$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \quad (11.5)$$

Evaluating the function f at each element of a given subset A of its domain $\text{dom}(f)$ produces a set called the *image* of A under (or through) f . The image is thus a subset of a function's codomain.



Definition 11.9 Inverse Image/Preimage $f^{-1}(\cdot)$:
Let $f : X \mapsto Y$ be a function, and A a subset set of its codomain Y .
Then the preimage of A under f is the set of all elements of the domain X , that map to elements in A under f :
$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \quad (11.6)$$

Example 11.1 :
Given $f : \mathbb{R} \rightarrow \mathbb{R}$
defined by $f : x \mapsto x^2 \iff f(x) = x^2$
 $\text{dom}(f) = \mathbb{R}$, $\text{codom}(f) = \mathbb{R}$ but its image is $f[\mathbb{R}] = \mathbb{R}_+$.

Image (Range) of a subset

The image of a subset $A \subseteq \mathcal{X}$ under f is the subset $f[A] \subseteq \mathcal{Y}$ defined by:

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \quad (11.7)$$

Note: Range

The term range is ambiguous as it may refer to the image or the codomain, depending on the definition. However, modern usage almost always uses range to mean image.

Definition 11.10 (strictly) Increasing Functions:
A function f is called **monotonically increasing/ increasing/non-decreasing** if:

$$x \leq y \iff f(x) \leq f(y) \quad \forall x, y \in \text{dom}(f) \quad (11.8)$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \quad \forall x, y \in \text{dom}(f) \quad (11.9)$$

Definition 11.11 (strictly) Decreasing Functions:
A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geq y \iff f(x) \geq f(y) \quad \forall x, y \in \text{dom}(f) \quad (11.10)$$

And **strictly decreasing** if:

$$x > y \iff f(x) > f(y) \quad \forall x, y \in \text{dom}(f) \quad (11.11)$$

Definition 11.12 Monotonic Function: A function f is called monotonic iff either f is **increasing** or **decreasing**.

Definition 11.13 Linear Function:
A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:
$$L(\mathbf{x} + \mathbf{y}) = L(\mathbf{x}) + L(\mathbf{y})$$
$$L(\alpha \mathbf{x}) = \alpha L(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

Corollary 11.2 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:
$$\frac{d}{dx} (af(x) + bg(x)) = a \frac{d}{dx} f(x) + b \frac{d}{dx} g(x) \quad a, b \in \mathbb{R} \quad (11.12)$$

Definition 11.14 Quadratic Function:
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(x) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \quad (11.13)$$

2. Continuity and Smoothness

Definition 11.15 Continuous Function:
Definition 11.16 Smoothness of a Function
tcblackC^k: **Given** a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, the function is said to be of class k if it is differentiable up to order k and continuous, on its entire domain:
$$f \in C^k(\mathcal{X}) \iff \exists f', f'', \dots, f^{(k)} \text{ continuous} \quad (11.14)$$

Note

- The class C^0 consists of all continuous functions.
- P.w. continuous \neq continuous.
- A function of that is k times differentiable must at least be of class C^{k-1} .
- $C^m(\mathcal{X}) \subset C^{m-1}, \dots, C^1 \subset C^0$
- Continuity is implied by the differentiability of all **derivatives** of up to order $k - 1$.

Corollary 11.3 Smooth Function C^∞ : Is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that has derivatives infinitely many times differentiable.
$$f \in C^\infty(\mathcal{X}) \iff f', f'', \dots, f^{(\infty)} \quad (11.15)$$

Corollary 11.4 Continuously Differentiable Function C^1 : Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ of the class must satisfy:
$$f \in C^1(\mathcal{X}) \iff f' \text{ continuous} \quad (11.16)$$

Often functions are not differentiable but we still want to state something about the rate of change of a function \Rightarrow hence we need a weaker notion of differentiability.

Definition 11.17 Lipschitz Continuity: A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Contant L :
$$|f(x) - f(y)| \leq L \|x - y\|_2^2 \quad \forall x, y, \quad L > 0 \quad (11.17)$$

Note

This property is useful as it allows us to conclude that a small perturbation in the input (i.e. of an algorithm) will result in small changes of the output \Rightarrow tells us something about robustness.

Definition 11.18 Lipschitz Continuous Gradient:
A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *L-Lipschitz continuous gradient* if it satisfies:
$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \forall x, y \in \text{dom}(f), \quad L > 0 \quad (11.18)$$

if $f \in C^2$, this is equivalent to:
$$\nabla^2 f(x) \leq LI \quad \forall x \in \text{dom}(f), \quad L > 0 \quad (11.19)$$

Lemma 11.1 Descent Lemma: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (11.18) over its domain, then it holds that:

$$|f(x) - f(y) - \nabla f(y)^T(x - y)| \leq \frac{L}{2} \|x - y\|^2 \quad (11.20)$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian ^(def. 12.5) of f is uniformly upper bounded by L

Proof. lemma 11.1 for C^1 functions:
Let $g(t) \equiv f(y + t(x - y))$ from the FToC (theorem 11.2) we know that:

$$\int_0^1 g'(t) dt = g(1) - g(0) = f(x) - f(y)$$

It then follows from the reverse:
$$|f(x) - f(y) - \nabla f(y)^T(x - y)|$$

$$\begin{aligned} & \stackrel{\text{Chain. R}}{\underset{\text{FTOC}}{=}} \left| \int_0^1 \nabla f(y + t(x - y))^T(x - y) dt - \nabla f(y)^T(x - y) \right| \\ &= \left| \int_0^1 (\nabla f(y + t(x - y)) - \nabla f(y))^T(x - y) dt \right| \\ &= \left| \int_0^1 (\nabla f(y + t(x - y)) - \nabla f(y))^T(x - y) dt \right| \\ & \stackrel{\text{C.S.}}{\leq} \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| dt \\ & \stackrel{\text{eq. (11.18)}}{=} \int_0^1 L \|y + t(x - y) - y\| \cdot \|x - y\| dt \\ &= \left| L \|x - y\|^2 \int_0^1 t dt \right| = \frac{L}{2} \|x - y\|^2 \end{aligned}$$

Proof. lemma 11.1 for C^2 functions:

$$f(y) \stackrel{\text{Taylor}}{=} f(x) + \nabla f(x)^T(y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(z) (y - x)$$

Now we plug in $\nabla^2 f(x)$ and recover eq. (11.21):

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2} (y - x)^T L (y - x)$$

□

Definition 11.19 L-Smoothness: A L -smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) + \frac{L}{2} \|x - y\|^2$$

with $\forall x, y \in \text{dom}(f), \quad L > 0 \quad (11.21)$

If f is a twice differentiable this is equivalent to:
$$\nabla^2 f(x) \leq LI \quad L > 0 \quad (11.22)$$

Theorem 11.3

L-Smoothness of convex functions:
A *convex* and L-Smooth function ^(def. 11.19) has a Lipschitz continuous gradient (eq. (11.18)) thus it holds that:

$$f(x) \leq f(y) + \nabla f(y)^T(x - y) \leq \frac{L}{2} \|x - y\|^2 \quad (11.23)$$

Proof. theorem 11.3:

With the definition of convexity for a differentiable function (eq. (11.26)) it follows

$$f(x) - f(y) + \nabla f(y)^T(x - y) \geq 0$$

$$\Rightarrow |f(x) - f(y) + \nabla f(y)^T(x - y)|$$

$$\stackrel{\text{if eq. (11.26)}}{=} f(x) - f(y) + \nabla f(y)^T(x - y)$$

with lemma 11.1 and ^[def. 11.19] it follows theorem 11.3 □

Corollary 11.5 : L -smoothnes is a weaker condition than L -Lipschitz continuous gradients

3. Convexity

Read stuff about uniqueness and so on again in NPDE/or NUM CSE and add proofs

Definition 11.20 Convex Functions:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if it satisfies:
$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad \forall x, y \in \text{dom}(f) \quad \forall \lambda \in [0, 1] \quad (11.24)$$

include figure from tika/convexity

Definition 11.21 Concave Functions:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is concave if it satisfies:
$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \quad \forall x, y \in \text{dom}(f) \quad \forall \lambda \in [0, 1] \quad (11.25)$$

Corollary 11.6 Convexity \rightarrow global minimima: Convexity implies that all local minima (if they exist) are global minima.

Definition 11.22 Stricly Convex Functions:

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is **strictly** convex if it satisfies:
$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \quad \forall x, y \in \text{dom}(f) \quad \forall \lambda \in [0, 1]$$

add plot

If f is a differentiable function this is equivalent to:
$$f(x) \geq f(y) + \nabla f(y)^T(x - y) \quad \forall x, y \in \text{dom}(f) \quad (11.26)$$

If f is a twice differentiable function this is equivalent to:
$$\nabla^2 f(x) \geq 0 \quad \forall x, y \in \text{dom}(f) \quad (11.27)$$

Intuition

- Convexity implies that a function f is bound by/below a linear interpolation from x to y and strong convexity that f is strictly bound/below.
- eq. (11.26) implies that $f(x)$ is above the tangent $f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \text{dom}(f)$
- ?? implies that $f(x)$ is flat or curved upwards

Corollary 11.7 Strict Convexity \rightarrow Uniqueness:
 Strict convexity implies a unique minimizer \iff at most one global minimum.

Corollary 11.8 : A twice differentiable function of one variable $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval \mathcal{X} :

$$f''(x) \geq 0 \quad \forall x \in \mathcal{X} \quad (11.28)$$

Definition 11.23 μ -Strong Convexity:
 Let \mathcal{X} be a Banach space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. A function $f : \mathcal{X} \rightarrow \mathbb{R}$ is called strongly convex iff the following equation holds:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{t(1-t)}{2} \mu \|x - y\|^2 \quad \forall x, y \in \mathcal{X}, \quad t \in [0, 1], \quad \mu > 0$$

If $f \in \mathcal{C}^1 \iff f$ is differentiable, this is equivalent to:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \quad (11.29)$$

If $f \in \mathcal{C}^2 \iff f$ is twice differentiable, this is equivalent to:

$$\nabla^2 f(x) \geq \mu \mathbf{I} \quad \forall x, y \in \mathcal{X} \quad \mu > 0 \quad (11.30)$$

Corollary 11.9 Strong Convexity implies Strict Convexity:
<https://math.stackexchange.com/question/2090991/proof-for-strongly-convex-function-is-strictly-convex>

Property 11.1:

$$f(y) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|_2^2 \quad (11.31)$$

Intuition
 Strong convexity implies that a function f is lower bounded by its second order (quadratic) approximation, rather than only its first order (linear) approximation.

Size of μ
 The parameter μ specifies how strongly the bounding quadratic function/approximation is.

Proof. eq. (11.30) analogously to **Proof** eq. (11.22) \square

Note
 If f is twice differentiable then the smallest eigenvalue of the Hessian (^[def. 12.5]) of f is uniformly lower bounded by μ .
 Hence strong convexity can be considered as the analogous to smoothness

Example 11.2 Quadratic Function: A quadratic function eq. (11.13) is convex if:

$$\nabla_{\mathbf{x}}^2 \text{eq. (11.13)} = \mathbf{A} \geq 0 \quad (11.32)$$

Corollary 11.10 :
 Strong convexity \Rightarrow Strict convexity \Rightarrow Convexity

3.1. Properties that preserve convexity

Property 11.2 Non-negative weighted Sums: Let f be a convex function then $g(x)$ is convex as well:

$$g(x) = \sum_{i=1}^n \alpha_i f_i(x) \quad \forall \alpha_j > 0$$

Property 11.3 Composition of Affine Mappings: Let f be a convex function then $g(x)$ is convex as well:

$$g(x) = f(\mathbf{Ax} + \mathbf{b})$$

Property 11.4 Pointwise Maxima: Let f be a convex function then $g(x)$ is convex as well:

$$g(x) = \max_i \{f_i(x)\}$$

Functions

Even Functions: have rotational symmetry with respect to the origin.
 \Rightarrow **Geometrically:** its graph remains unchanged after reflection about the y-axis.

$$f(-x) = f(x) \quad (11.33)$$

Odd Functions: are symmetric w.r.t. to the y-axis.
 \Rightarrow **Geometrically:** its graph remains unchanged after rotation of 180 degrees about the origin.

$$f(-x) = -f(x) \quad (11.34)$$

Theorem 11.4 Rules:
 Let f be even and f odd respectively.
 $g =: f \cdot f$ is even $g =: f \cdot f$ is even
 $g =: f \cdot f$ is odd the same holds for division

Examples

Even: $\cos x, |x|, \mathbf{c}, x^2, x^4, \dots \exp(-x^2/2)$.
Odd: $\sin x, \tan x, x, x^3, x^5, \dots$

x-Shift: $f(x - \mathbf{c}) \Rightarrow$ shift to the right
 $f(x + \mathbf{c}) \Rightarrow$ shift to the left
y-Shift: $f(x) \pm \mathbf{c} \Rightarrow$ shift up/down

Proof. eq. (11.35) $f(x_n - \mathbf{c})$ we take the x -value at x_n but take the y -value at $x_o := x_n - \mathbf{c} \Rightarrow$ we shift the function to x_n . \square

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \quad (11.37)$$

Euler's Identity

$$e^{\pm i} = -1 \quad (11.38)$$

Note

$$e^n = 1 \iff n = i2\pi k, \quad k \in \mathbb{N} \quad (11.39)$$

Corollary 11.11 Every norm is a convex function: By using definition (^[def. 11.20]) and the triangular inequality it follows (with the exception of the L0-norm):

$$\|\lambda x + (1 - \lambda)y\| \leq \lambda \|x\| + (1 - \lambda) \|y\|$$

3.2. Taylor Expansion

Definition 11.24 Taylor Expansion:

$$T_n(x) = \sum_{i=0}^n \frac{1}{i!} f^{(i)}(x_0) \cdot (x - x_0)^{(i)} \quad (11.40)$$

$$= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \mathcal{O}(x^3) \quad (11.41)$$

Definition 11.25 Incremental Taylor:
Goal: evaluate $T_n(x)$ (eq. (11.41)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^n \frac{h^i}{i!} f^{(i)}(x_0) i^{-1} \quad (11.42)$$

$$= f(x_0) \pm h f'(x_0) + \frac{h^2}{2} f''(x_0) \pm f'''(x_0)(h)^3 + \mathcal{O}(h^4)$$

Note
 If we chose Δx small enough it is sufficient to look only at the first two terms.

Definition 11.26 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\mathbf{x} \in X, f : X \mapsto \mathbb{R}$ and $f \in \mathcal{C}^2$ then it holds that

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top H(\mathbf{x} - \mathbf{x}_0) \quad (11.43)$$

Definition 11.27 Argmax: The argmax of a function defined on a set D is given by:

$$\arg \max_{x \in D} f(x) = \{x | f(x) \geq f(y), \forall y \in D\} \quad (11.44)$$

Definition 11.28 Argmin: The argmin of a function defined on a set D is given by:

$$\arg \min_{x \in D} f(x) = \{x | f(x) \leq f(y), \forall y \in D\} \quad (11.45)$$

Corollary 11.12 Relationship $\arg \min \leftrightarrow \arg \max$:

$$\arg \min_{x \in D} f(x) = \arg \max_{x \in D} -f(x) \quad (11.46)$$

Property 11.5 Argmax Identities:

- Shifting:** $\forall \lambda \text{ const} \quad \arg \max f(x) = \arg \max f(x) + \lambda \quad (11.47)$
- Positive Scaling:** $\forall \lambda > 0 \text{ const} \quad \arg \max f(x) = \arg \max \lambda f(x) \quad (11.48)$
- Negative Scaling:** $\forall \lambda < 0 \text{ const} \quad \arg \max f(x) = \arg \min \lambda f(x) \quad (11.49)$
- Positive Functions:** $\forall \arg \max f(x) > 0, \forall x \in \text{dom}(f)$
 $\arg \max f(x) = \arg \min \frac{1}{f(x)} \quad (11.50)$
- Stricly Monotonic Functions:** for all strictly monotonic increasing functions (^[def. 11.10]) g it holds that:
 $\arg \max g(f(x)) = \arg \max f(x) \quad (11.51)$

Definition 11.29 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \max f(x) \quad (11.52)$$

Definition 11.30 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \min f(x) \quad (11.53)$$

Corollary 11.13 Relationship $\min \leftrightarrow \max$:

$$\min_{x \in D} f(x) = - \max_{x \in D} -f(x) \quad (11.54)$$

Property 11.6 Max Identities:

- Shifting:** $\forall \lambda \text{ const} \quad \max \{f(x) + \lambda\} = \lambda + \max f(x) \quad (11.55)$
- Positive Scaling:** $\forall \lambda > 0 \text{ const} \quad \max \lambda f(x) = \lambda \max f(x) \quad (11.56)$
- Negative Scaling:** $\forall \lambda < 0 \text{ const} \quad \max \lambda f(x) = \lambda \min f(x) \quad (11.57)$
- Positive Functions:** $\forall \arg \max f(x) > 0, \forall x \in \text{dom}(f) \quad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \quad (11.58)$
- Stricly Monotonic Functions:** for all strictly monotonic increasing functions (^[def. 11.10]) g it holds that:
 $\max g(f(x)) = g(\max f(x)) \quad (11.59)$

Definition 11.31 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x \in D} f(x) = \{y | y \geq f(x), \forall x \in D\} = \min_{y | y \geq f(x), \forall x \in D} y \quad (11.60)$$

and is the smallest value y that is equal or greater $f(x)$ for any $x \iff$ smallest upper bound.

Definition 11.32 Infimum: The infimum of a function defined on a set D is given by:

$$\inf_{x \in D} f(x) = \{y | y \leq f(x), \forall x \in D\} = \max_{y | y \leq f(x), \forall x \in D} y \quad (11.61)$$

and is the biggest value y that is equal or smaller $f(x)$ for any $x \iff$ largest lower bound.

Corollary 11.14 Relationship $\sup \leftrightarrow \inf$:

$$\inf_{x \in D} f(x) = - \sup_{x \in D} -f(x) \quad (11.62)$$

Note
 The supremum/infimum is necessary to handle unbound function that seem to converge and for which the max/min does not exist as the argmax/argmin may be empty.
 E.g. consider $-e^x/e^x$ for which the max/min converges toward 0 but will never reached s.t. we can always choose a bigger $x \Rightarrow$ there exists no argmax/argmin \Rightarrow need to bound the functions from above/below \iff infimum/supremum.

Definition 11.33 Time-invariant system (TIS): A function f is called time-invariant, if shifting the input in time leads to the same output shifted in time by the same amount.

$$y(t) = f(x(t), t) \xrightarrow{\text{time-invariance}} y(t - \tau) = f(x(t - \tau), t) \quad \forall \tau \quad (11.63)$$

Definition 11.34 Inverse Function $g = f^{-1}$:
 A function g is the inverse function of the function $f : A \subset \mathbb{R} \rightarrow B \subset \mathbb{R}$ if

$$f(g(x)) = x \quad \forall x \in \text{dom}(g) \quad (11.64)$$

and

$$g(f(u)) = u \quad \forall u \in \text{dom}(f) \quad (11.65)$$

Property 11.7
Reflective Property of Inverse Functions: f contains (a, b) if and only if f^{-1} contains (b, a) .
 The line $y = x$ is a symmetry line for f and f^{-1} .

Theorem 11.5 The Existence of an Inverse Function:
 A function has an inverse function if and only if it is one-to-one.

Corollary 11.15 Inverse functions and strict monotonicity: If a function f is **strictly monotonic** (^[def. 11.12]) on its entire domain, then it is one-to-one and therefore has an inverse function.

4. Special Functions

4.1. The Gamma Function

Definition 11.35 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (??) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad \Re(z) > 0 \quad (11.66)$$

$$\Gamma(n) \xleftrightarrow{n \in \mathbb{N}} \Gamma(n) = (n-1)!$$

Differential Calculus

Definition 12.1 Critical/Stationary Point: Given a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that is differentiable at a point \mathbf{x}_0 then it is called a **critical point** if the functions derivative vanishes at that point:

f'(x0) = 0 ⇔ ∇xf(x0) = 0

Definition 12.2 Second Derivative ∂² / ∂xi ∂xj :

Corollary 12.1 Second Derivative Test f : ℝ → ℝ: Suppose f : ℝ → ℝ is twice differentiable at a stationary point x [def. 12.1] then it follows that:

- f''(x) > 0 ⇔ f'(x + ϵ) > 0 slope points uphill f'(x - ϵ) < 0 slope points downhill f(x) is a local minimum
- f''(x) < 0 ⇔ f'(x + ϵ) > 0 slope points downhill f'(x - ϵ) < 0 slope points uphill f(x) is a local maximum

ϵ > 0 sufficiently small enough

Definition 12.3 Gradient: Given f : n → ℝ its gradient is defined as: grad_x(f) = ∇xf := [∂f/∂x1 ∂f/∂x2 ... ∂f/∂xn] (12.1)

Definition 12.4 Jacobi Matrix: Given a vector valued function f : ℝ^n → ℝ^m its derivative/Jacobian is defined as:

J(f(x)) = Jf(x) = Df = ∂f/∂x(x) = ∂(f1,...,fm)/∂(x1,...,xn)(x) = [∂f1/∂x1(x) ∂f1/∂x2(x) ... ∂f1/∂xn(x) ∂f2/∂x1(x) ∂f2/∂x2(x) ... ∂f2/∂xn(x) ... ∂fm/∂x1(x) ∂fm/∂x2(x) ... ∂fm/∂xn(x)] (12.2)

Theorem 12.1 Symmetry of second derivatives/Schwartz's Theorem: Given a continuous and twice differentiable function f : ℝ^n → ℝ then its second order partial derivatives commute:

∂/∂xi ∂f/∂xj = ∂/∂xj ∂f/∂xi

Definition 12.5 Hessian Matrix: Given a function f : ℝ → ℝ^n its Hessian ∈ ℝ^{n×n} is defined as:

H(f)(x) = Hf(x) = J(∇f(x))^T (12.3) [∂²f/∂x²1(x) ∂²f/∂x1∂x2(x) ... ∂²f/∂x1∂xn(x) ∂²f/∂x2∂x1(x) ∂²f/∂x²2(x) ... ∂²f/∂x2∂xn(x) ... ∂²f/∂xn∂x1(x) ∂²f/∂xn∂x2(x) ... ∂²f/∂xn²(x)]

and it corresponds to the Jacobian of the Gradient. Due to the differentiability and theorem 12.1 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 12.2 Eigenvector basis of the Hessian: Due to the fact that the Hessian is real and symmetric we can decompose it into a set of real eigenvalues and an orthogonal basis of eigenvectors {(λ1,v1),..., (λn,vn)}. Not let d be a directional unit vector then the second derivative in that direction is given by:

d^T Hd ⇔ d^T ∑_{i=1}^n λi vi ⇔ if d=vj d^T λj vj

- The eigenvectors that have smaller angle with d have bigger weight/eigenvalues
- The minimum/maximum eigenvalue determines the minimum/maximum second derivative

Corollary 12.3 Second Derivative Test f : ℝ^n → ℝ: Suppose f : ℝ^n → ℝ is twice differentiable at a stationary point x [def. 12.1] then it follows that:

- If H is p.d ⇔ ∀λi > 0 ∈ H → f(x) is a local min.
- If H is n.d ⇔ ∀λi < 0 ∈ H → f(x) is a local max.
- If ∃λi > 0 ∈ H and ∃λi < 0 ∈ H then x is a local maximum in one cross section of f but a local minimum in another
- If ∃λi = 0 ∈ H and all other eigenvalues have the same sign the test is inclusive as it is inconclusive in the cross section corresponding to the zero eigenvalue.

Note

If H is positive definite for a minima x* of a quadratic function f then this point must be a global minimum of that function.

Integral Calculus

Theorem 13.1 Important Integral Properties:

Addition $\int\limits_a^b f(x) \, dx = \int\limits_a^c f(x) \, dx + \int\limits_c^b f(x) \, dx$ (13.1)

Reflection $\int\limits_a^b f(x) \, dx = - \int\limits_b^a f(x) \, dx$ (13.2)

Translation $\int\limits_a^b f(x) \, dx \stackrel{u:=x\pm c}{=} \int\limits_{a\pm c}^{b\pm c} f(x \mp c) \, dx$ (13.3)

f **Odd** $\int\limits_{-a}^a f(x) \, dx = 0$ (13.4)

f **Even** $\int\limits_{-a}^a f(x) \, dx = 2 \int\limits_0^a f(x) \, dx$ (13.5)

Proof. eqs. (13.4) and (13.5)

$$\begin{aligned} I &:= \int\limits_{-a}^a f(x) \, dx = \int\limits_{-a}^0 f(x) \, dx + \int\limits_0^a f(x) \, dx \\ &\stackrel{t=-x}{dt=-dx} = - \int\limits_a^0 f(-x) \, dx + \int\limits_0^a f(x) \, dx \\ &= \int\limits_0^a f(-x) + f(x) \, dx = \begin{cases} 0 & \text{if } f \text{ odd} \\ 2I & \text{if } f \text{ even} \end{cases} \end{aligned}$$

□

Linear Algebra

Given a matrix $A \in \mathbb{K}^{m,n}$

Rank:
of a matrix is the dimension of the vector space generated (or spanned) by its columns/rows.

Span/Linear Hull: $\text{span}(v_1, v_2, \dots, v_n) = \{ \lambda_1 v_1, \lambda_2 v_2, \dots, \lambda_n v_n \} = \{ v \mid v = \sum_{i=1}^n \lambda_i v_i, \lambda_i \in \mathbb{R} \}$

Is the set of vectors tha can be expressed as a linear combination of the vectors v_1, \dots, v_n .
Note these vectors may be linearly independent.
Generatring Set: Is the set of vectors which span the \mathbb{R}^n that is: $\text{span}(v_1, \dots, v_m) = \mathbb{R}^n$.
e.g. $(4, 0)^T, (0, 5)^T$ span the \mathbb{R}^n .
Basis \mathfrak{B} : A lin. indep. generating set of the \mathbb{R}^n is called basis of the \mathbb{R}^n .
The unit vectors e_1, \dots, e_n build a standard basis of the \mathbb{R}^n

Vector Space
Image/Range: $\mathfrak{R}(A) := \{ Ax \mid x \in \mathbb{K}^n \} \subset \mathbb{K}^n$
Null-Space/Kernel: $\mathfrak{N} := \{ z \in \mathbb{K}^n \mid Az = 0 \}$
Dimension theorem:

Theorem 14.1 Rank-Nullity theorem: For any $A \in \mathbb{Q}^{m \times n}$
 $n = \dim(\mathfrak{N}[A]) + \dim(\mathfrak{R}[A])$

From orthogonality it follows $x \in \mathfrak{R}(A), y \in \mathfrak{N}(A) \Rightarrow x^\top y = 0$.

1. Transformations

1.1. Affine Transformations

Definition 14.1 Affine Transfromation/Map:
Let $x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ then:
 $Y = Ax + b$

(14.1)

is called an affine transformation of x .

2. Eigenvalues and Vectors

Formula 14.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix A its eigenvalues can be calculated by:

$$\{\lambda_1, \lambda_2\} \in \frac{\text{tr}(A) \pm \sqrt{\text{tr}(A)^2 - 4 \det(A)}}{2}$$

with $\text{tr}(A) = a + d \quad \det(A) = ad - bc$

3. Special Kind of Vectors

Definition 14.2 Orthogonal Vectors: Let \mathcal{V} be an inner-product space^[def. 14.18]. A set of vectors $\{u_1, \dots, u_n, \dots\} \in \mathcal{V}$ is called *orthogonal* iff:

$$\langle u_i, u_j \rangle = 0 \quad \forall i \neq j$$

(14.3)

Definition 14.3 Orthonormal Vectors: Let \mathcal{V} be an inner-product space^[def. 14.18]. A set of vectors $\{u_1, \dots, u_n, \dots\} \in \mathcal{V}$ is called *orthonormal* iff:

$$\langle u_i, u_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \forall i, j$$

(14.4)

4. Special Kind of Matrices

Definition 14.4 Orthogonal Matrix: A real valued square matrix $Q \in \mathbb{R}^{n \times n}$ is said to be orthogonal if its row vectors (and respectively its column vectors) build an orthonormal basis:

$$\langle q_i, q_j \rangle = \delta_{ij} \quad \text{and} \quad \langle q_i, q_j \rangle = \delta_{ij} \quad (14.5)$$

This is exactly true if the inverse of Q equals its transpose:

$$Q^{-1} = Q^\top \iff QQ^\top = Q^\top Q = I \quad (14.6)$$

Definition 14.5 Unitary/Hermitian Matrices:

$$A = A^H \quad (14.7)$$

5. Block Partitioned Matrices

Definition 14.6 Block Partitioned Matrix:
A matrix $M \in \mathbb{R}^{k+l, k+l}$ can be partitioned into a *block partitioned matrix*:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad A \in \mathbb{R}^{k,k}, B \in \mathbb{R}^{k,l}, C \in \mathbb{R}^{l,k}, D \in \mathbb{R}^{l,l}$$

(14.8)

Definition 14.7 Block Partitioned Linear System:
A linear system $Mx = b$ with $M \in \mathbb{R}^{k+l, k+l}$ and $x, b \in \mathbb{R}^{k+l}$ can be partitioned into a *block partitioned system*:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad A \in \mathbb{R}^{k,k}, B \in \mathbb{R}^{k,l}, C \in \mathbb{R}^{l,k}, D \in \mathbb{R}^{l,l}$$

$$x_1, b_1 \in \mathbb{R}^k, x_2, b_2 \in \mathbb{R}^l \quad (14.9)$$

5.1. Schur Complement

Definition 14.8 Schur Complement: Given a block partitioned matrix^[def. 14.6] $M \in \mathbb{R}^{k+l, k+l}$ its Schur complements are given by:

$$S_A = D - CA^{-1}B \quad S_D = A - BD^{-1}C \quad (14.10)$$

5.2. Inverse of Block Partitioned Matrix

Definition 14.9 Inverse of a Block Partitioned Matrix:
Given a block partitioned matrix^[def. 14.6] $M \in \mathbb{R}^{k+l, k+l}$ its inverse M^{-1} can be partitioned as well:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \quad M^{-1} = \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{bmatrix} \quad (14.11)$$

$$\tilde{A} = A^{-1} + A^{-1}BS_A CA^{-1} \quad \tilde{D} = S_A$$

$$\tilde{B} = -A^{-1}BS_A \quad \tilde{C} = -S_A CA^{-1}$$

where $S_A = D - CA^{-1}B$ is the Schur complement of A .

5.3. Properties of Matrices

5.3.1. Square Root of p.s.d. Matrices

Definition 14.10 Square Root:

6. Decompositions

6.0.1. Eigendecomposition

Definition 14.11 Eigendecomposition $A = Q\Lambda Q^{-1}$:

6.0.2. Cholesky Decomposition

7. Spaces and Measures

Definition 14.12 Bilinear Form/Functional:
Is a mapping $a : \mathcal{V} \times \mathcal{V} \mapsto F$ on a field of scalars $F \subseteq \mathbb{K}, K = \mathbb{R}$ or \mathbb{C} that satisfies:

$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$

$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$

$$\forall u, v, w \in \mathcal{V}, \quad \forall \alpha, \beta \in \mathbb{K}$$

Thus: a is linear w.r.t. each argument.

Definition 14.13 Symmetric bilinear form: A bilinear form a on \mathcal{V} is symmetric if and only if:

$$a(u, v) = a(v, u) \quad \forall u, v \in \mathcal{V}$$

Definition 14.14 Positive (semi) definite bilinear form:
A symmetric bilinear form a on a vector space \mathcal{V} over a field F is **positive defintie** if and only if:

$$a(u, u) > 0 \quad \forall u \in \mathcal{V} \setminus \{0\}$$

And **positive semidefinte** $\iff \geq$ (14.13)

Corollary 14.1 Matrix induced Bilinear Form:
For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any *sym-metric* matrix $A \in \mathbb{R}^{n \times n}$ induces a **bilinear form**:

$$a(x, x') = x^\top A x' = (Ax')^\top x$$

Definition 14.15 Positive (semi) definite Matrix $>$:
A matrix $A \in \mathbb{R}^{n \times n}$ is **positive defintie** if and only if:

$$x^\top A x > 0 \iff A > \quad \forall x \in \mathbb{R}^n \setminus \{0\} \quad (14.14)$$

And **positive semidefinte** $\iff \geq$ (14.15)

Corollary 14.2

Eigenvalues of positive (semi) definite matrix:
A positive definite matrix is a *symmetric matrix* where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive*.

$$\forall \lambda_i \in \text{eigenv}(A) > 0 \quad (14.16)$$

And **positive semidefinte** $\iff \geq$ (14.17)

Proof. corollary 14.2 (for real matrices):
Let v be an eigenvector of A then it follows:

$$0 < \text{corollary 14.2} \quad v^\top A v = v^\top \lambda v = \|v\| \lambda$$

\square

Corollary 14.3 Positive Definiteness and Determinant:
The determinant of a positive definite matrix is always positive. Thus a positive definite matrix is always *nonsingular*

Definition 14.16 Negative (semi) definite Matrix $<$:
A matrix $A \in \mathbb{R}^{n \times n}$ is **negative defintie** if and only if:

$$x^\top A x < 0 \iff A < 0 \quad \forall x \in \mathbb{R}^n \setminus \{0\} \quad (14.18)$$

And **negative semidefinte** $\iff \leq$ (14.19)

Theorem 14.2 Sylvester's criterion: Let A be *symmet-ric/Hermitian* matrix and denote by $A^{(k)}$ the $k \times k$ upper left sub-matrix of A .

Then it holds that:

- $A > 0 \iff \det(A^{(k)}) > 0 \quad k = 1, \dots, n$ (14.20)
- $A < 0 \iff (-1)^k \det(A^{(k)}) > 0 \quad k = 1, \dots, n$ (14.21)

- A is indefinite if the first $\det(A^{(k)})$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive (A can be anything of the previous three) if the first $\det(A^{(k)})$ that breaks both patterns is 0.

8. Inner Products

Definition 14.17 Inner Product: Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on \mathcal{V} is a map:

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \mapsto F \subseteq \mathbb{K} \quad K = \mathbb{R} \text{ or } \mathbb{C} \quad (14.22)$$

that satisfies:

$$\forall x, y, z \in \mathcal{V}, \quad \alpha, \beta \in F$$

- (Conjugate) Stmmetry:** $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
- Linearity** in the first argument:
 $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
- Positive-definiteness:**
 $\langle x, x \rangle \geq 0 : x = 0 \iff \langle x, x \rangle = 0$

Definition 14.18 Inner Product Space $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$: Let $F \in \mathbb{K}$ be a field of scalars.
An inner product space \mathcal{V} is a vetor space over a field F together with an an **inner product** $\langle \cdot, \cdot \rangle_{\mathcal{V}}$.

Corollary 14.4 Inner product \rightarrow S.p.d. Bilinear Form:
Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on \mathcal{V} is a positive definite symmetric bilinear form on \mathcal{V} .

Example: scalar prodct

Let $a(u, v) = u^\top I v$ then the standard scalar product can be defined in terms of a bilinear form vice versa the standard scalar product induces a bilinear form.

Note

Inner products must be positive definite by defintion $\langle x, x \rangle \geq 0$, whereas bilinear forms must not.

Definition 14.19 Norm $\|\cdot\|_{\mathcal{Y}}$:
A norm measures the **size** of its argument.
Formally let \mathcal{Y} be a vector space over a field F , a norm on \mathcal{Y} is a map:

$$\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \mapsto \mathbb{R}_+ \quad (14.23)$$

that satisfies: $\forall x, y \in \mathcal{Y}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R} \text{ or } \mathbb{C}$

- Definiteness:** $\|x\|_{\mathcal{Y}} = 0 \iff x = 0$.
- Homogenity:** $\|\alpha x\|_{\mathcal{Y}} = |\alpha| \|x\|_{\mathcal{Y}}$
- Triangular Inequality:** $\|x + y\|_{\mathcal{Y}} \leq \|x\|_{\mathcal{Y}} + \|y\|_{\mathcal{Y}}$

Meaning: Triangular Inequality

States that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side.

Corollary 14.5 Reverse Triangular Inequality:

$$-\|x - y\|_{\mathcal{Y}} \leq \|x\|_{\mathcal{Y}} - \|y\|_{\mathcal{Y}} \leq \|x - y\|_{\mathcal{Y}}$$

resp. $|\|x\|_{\mathcal{Y}} - \|y\|_{\mathcal{Y}}| \leq \|x - y\|_{\mathcal{Y}}$

Semi-norm

add

Corollary 14.6 Normed vector space: Is a vector space \mathcal{V} over a field F , on which a norm $\|\cdot\|_{\mathcal{Y}}$ can be defined.

Corollary 14.7 Inner product induced norm $\langle \cdot, \cdot \rangle_{\mathcal{Y}} \rightarrow \|\cdot\|_{\mathcal{Y}}$: Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ induces a norm of the form:

$$\|x\|_{\mathcal{Y}} = \sqrt{\langle x, x \rangle} \quad x \in \mathcal{Y}$$

Thus We can define function spaces by their associated norm $(\mathcal{V}, \|\cdot\|_{\mathcal{Y}})$ and inner product spaces lead to normed vector spaces and vice versa.

Corollary 14.8 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{V} \times \mathcal{V} \mapsto F$ induces an energy norm:

$$\|x\|_a := (a(x, x))^{\frac{1}{2}} = \sqrt{a(x, x)} \quad x \in \mathcal{V}$$

Definition 14.20 Distance Function/Measure: Is measuring the **distance** between two things.

Formally: on a set S is a mapping:

$$d(\cdot, \cdot) : S \times S \mapsto \mathbb{R}_+$$

that satisfies:

$$\forall x, y, z \in S$$

- ?**: $d(x, x) = 0$
- Symmetry:** $d(x, y) = d(y, x)$
- Triangular Identi**y: $d(x, z) \leq d(x, y) + d(y, z)$

Definition 14.21 Metric: Is a distance measure that additionally satisfies:

$$\forall x, y \in S$$

identity of indiscernibles : $d(x, y) = 0 \iff x = y$

Corollary 14.9 Metric \rightarrow Norm: Every norm $\|\cdot\|_{\mathcal{Y}}$ on a vector space \mathcal{V} over a field F induces a metric by:

$$d(x, y) = \|x - y\|_{\mathcal{Y}} \quad \forall x, y \in \mathcal{Y}$$

metric induced by norms additionally satisfy: $\forall x, y \in \mathcal{V}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R} \text{ or } \mathbb{C}$

- Homogenity/Scaling:** $d(\alpha x, \alpha y)_{\mathcal{Y}} = |\alpha| d(x, y)_{\mathcal{Y}}$
- Translatational Invariance:** $d(x + \alpha, y + \alpha) = d(x, y)$

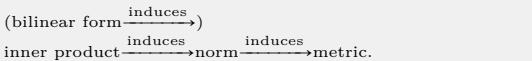
Conversely not every metric induces a norm **but** if a metric d on a vector space \mathcal{V} satisfies the properties then it induces a norm of the form:

$$\|x\|_{\mathcal{Y}} := d(x, 0)_{\mathcal{Y}}$$

Note

Similarity measure is a much weaker notion than a metric as triangular inequality does not have to hold.
Hence: If a is similar to b and b is similar to c it does not imply that a is similar to c .

Note



9. Vector Algebra

9.1. Planes

<https://math.stackexchange.com/questions/1485509/show-that-two-planes-are-parallel-and-find-the-distance-between-them>

10. Derivatives

$$\begin{aligned}\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{b}) = \mathbf{b} \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) &= 2\mathbf{x} \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{A} \mathbf{x}) &= (\mathbf{A}^\top + \mathbf{A})\mathbf{x} \qquad \frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{A} \mathbf{x}) = \mathbf{A}^\top \mathbf{b} \\ \frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top \mathbf{X} \mathbf{b}) &= \mathbf{c} \mathbf{b}^\top \qquad \frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|_2} \\ \frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x}\|_2^2) &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x} \qquad \frac{\partial}{\partial \mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X} \\ \frac{\partial}{\partial \mathbf{x}}\|\mathbf{x}\|_1 &= \frac{\mathbf{x}}{|\mathbf{x}|} \\ \frac{\partial}{\partial \mathbf{x}}(\|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2^2) &= 2(\mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b}) \qquad \frac{\partial}{\partial \mathbf{X}}(|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1} \\ \frac{\partial}{\partial x}(\mathbf{Y}^{-1}) &= -\mathbf{Y}^{-1} \frac{\partial \mathbf{Y}}{\partial x} \mathbf{Y}^{-1}\end{aligned}$$

11. Proofs

Proof. [def. 14.9]

$$MM^{-1} = \begin{bmatrix} I_{k,k} & \mathbf{0}_{k,l} \\ \mathbf{0}_{l,k} & I_{l,l} \end{bmatrix} \tag{14.24}$$

□

Geometry

Corollary 15.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The **affine transformation** of $\phi : \hat{\Omega} \rightarrow \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d - c}{b - a} (x - a) + c \tag{15.1}$$

Proof. **corollary 15.1** By ^[def. 14.1] we want a function $f : [a, b] \rightarrow [c, d]$ that satisfies:

$f(a) = c$ **and** $f(b) = d$

additionally $f(x)$ has to be a linear function (^[def. 11.13]), that is the output scales the same way as the input scales.

Thus it follows:
$$\frac{d - c}{b - a} = \frac{f(x) - f(a)}{x - a} \quad \iff \quad f(x) = \frac{d - c}{b - a} (x - a) + c$$

Trigonometry

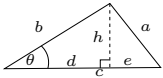
Law 15.1 Law of Cosine: relates the side of a triangle to the cosine of its angles.

$$a^2 = b^2 + c^2 - 2bc \cos \theta_{b,c} \tag{15.2}$$

More general for vectors it holds:

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos \theta_{\boldsymbol{x},\boldsymbol{y}} \tag{15.3}$$

Proof. eq. (15.2):
We know: $\sin \theta = \frac{h}{b} \Rightarrow \underline{h} \quad \text{and} \quad \cos \theta = \frac{d}{b} \Rightarrow d$
Thus $\underline{e} = c - d = c - b \cos \theta \Rightarrow a^2 = \underline{e}^2 + \underline{h}^2 \Rightarrow a$ \square



Proof. eq. (15.3):
$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{y}\|^2 &= (\boldsymbol{x} - \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y}) \\ &= \boldsymbol{x} \cdot \boldsymbol{x} - 2\boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{y} \cdot \boldsymbol{y} \\ &= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2(\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos \theta) \end{aligned}$$

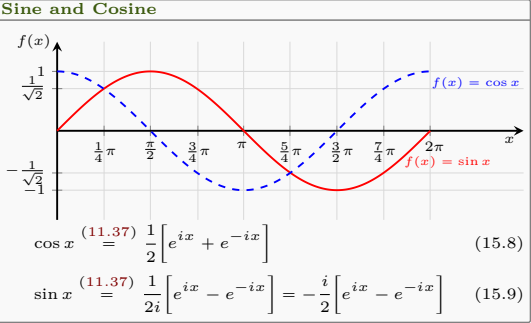
Law 15.2 Pythagorean theorem: special case of ?? for right triangle:

$$a^2 = b^2 + c^2 \tag{15.4}$$

Formula 15.1 Euler's Formula:
$$e^{\pm i x} = \cos x \pm i \sin x \tag{15.5}$$

Formula 15.2 Euler's Identity:
$$e^{\pm i} = -1 \tag{15.6}$$

Note
$$e^n = 1 \Leftrightarrow n = i 2 \pi k, \quad k \in \mathbb{N} \tag{15.7}$$



Sinh and Cosh

$$\cosh x \stackrel{(11.37)}{=} \frac{1}{2} \left[e^x + e^{-x} \right] = \cos(i x) \tag{15.10}$$
$$\sinh x \stackrel{(11.37)}{=} \frac{1}{2} \left[e^x - e^{-x} \right] = -i \sin(i x) \tag{15.11}$$

Note

$$e^x = \cosh x + \sinh x \quad e^{-x} = \cosh x - \sinh x \tag{15.12}$$

Note

- $\cosh x$ is strictly positive.
- $\sinh x = 0$ has a unique root at $x = 0$.

Theorem 15.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin \alpha \cos \beta \pm \cos \alpha \sin \beta \tag{15.13}$$
$$\cos(\alpha \pm \beta) = \cos \alpha \cos \beta \mp \sin \alpha \sin \beta \tag{15.14}$$

Werner Formulas

$$\sin \alpha \cos \beta = \frac{1}{2} \left[\sin(\alpha + \beta) + \sin(\alpha - \beta) \right] \tag{15.15}$$
$$\sin \alpha \sin \beta = \frac{1}{2} \left[\cos(\alpha - \beta) - \cos(\alpha + \beta) \right] \tag{15.16}$$
$$\cos \alpha \cos \beta = \frac{1}{2} \left[\cos(\alpha + \beta) + \cos(\alpha - \beta) \right] \tag{15.17}$$

Note

Using theorem 15.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos \alpha \quad \text{and} \quad \sin(\alpha \pm \pi) = -\sin \alpha \tag{15.18}$$

Topology

Numerics

1. Machine Arithmetic's

1.1. Machine Numbers

Definition 17.1 Institute of Electrical and Electronics Engineers (IEEE): Is a engineering associations that defines a standard on how computers should treat machine numbers in order to have certain guarantees.

Definition 17.2 Machine/Floating Point Numbers \mathbb{F} : Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{F} \subset \mathbb{R}$

1.1.1. Floating Point Arithmetic's $x\tilde{\Omega}y = \mathfrak{fl}(x\Omega y)$

Corollary 17.1 Closure: Machine numbers \mathbb{F} are not *closed*^[def. 9.6] under basic arithmetic operations:
 $\mathbb{F} \Omega \mathbb{F} \mapsto \not\mathbb{F} \quad \Omega = \{+, -, *, /\}$ (17.1)

Note
Corollary 17.1 provides a problem as the computer can only represent floating point number \mathbb{F} .

Definition 17.3 Floating Point Operation $\tilde{\Omega}$:
Is a basic arithmetic operation that obtains a number $x \in \mathbb{F}$ by applying a function rd:
 $\mathbb{F} \tilde{\Omega} \mathbb{F} \mapsto \mathbb{F} \quad \tilde{\Omega} := \text{rd} \circ \Omega$
 $\Omega = \{+, -, *, /\}$ (17.2)

Definition 17.4 Rounding Function rd:
Given a real number $x \in \mathbb{R}$ the rounding function replaces it by the nearest machine number $\tilde{x} \in \mathbb{F}$. If this is ambiguous (there are two possibilities), then it takes the larger one:
 $\text{rd} : \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases}$ (17.3)

Consequence
Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.
Axiom 17.1 Axiom of Round off Analysis:
Let $x, y \in \mathbb{F}$ be (normalized) floats and assume that $x\tilde{\Omega}y \in \mathbb{F}$ (i.e. no over/underflow). Then it holds that:
 $x\tilde{\Omega}y = (x\Omega y)(1 + \delta) \quad \Omega = \{+, -, *, /\}$
 $\tilde{f}(x) = f(x)(1 + \delta) \quad f \in \{\exp, \sin, \cos, \log, \dots\}$ (17.4)
with $|\delta| < \text{EPS}$

Explanation 17.1 (axiom 17.1). *gives us a guarantee that for any two floating point numbers $x, y \in \mathbb{F}$, any operation involving them will give a floating point result which is within a factor of $1 + \delta$ of the true result $x\Omega y$.*

Remark

Definition 17.5 Overflow: Result is bigger then the biggest representable floating point number.

Definition 17.6 Underflow: Result is smaller then the smaller representable floating point number i.e. to close to zero.

1.2. Roundoff Errors Log-Sum-Exp Trick

The sum exponential trick is at trick that helps to calculate the log-sum-exponential in a robust way by avoiding over/underflow. The log-sum-exponential^[def. 17.7] is an expression that arises frequently in machine learning i.e. for the cross entropy loss or for calculating the evidence of a posterior prediction.
The root of the problem is that we need to calculate the exponential $\exp(x)$, this comes with two different problems:
• If x is large (i.e. 89 for single precision floats) then $\exp(x)$ will lead to overflow
• If x is very negative $\exp(x)$ will lead to underflow/0. This is not necessarily a problem but if $\exp(x)$ occurs in the denominator or the logarithm for example this is catastrophic.

Definition 17.7 Log sum Exponential:
 $\text{LogSumExp}(x_1, \dots, x_n) := \log \left(\sum_{i=1}^n e^{x_i} \right)$ (17.5)

Formula 17.1 Log-Sum-Exp Trick:
 $\log \left(\sum_{i=1}^n e^{x_i} \right) = a + \log \sum_{i=1}^n e^{x_i - a} \quad a := \max_{i \in \{1, \dots, n\}} x_i$ (17.6)

Explanation 17.2 (formula 17.1). *The value a can be any real value but for robustness one usually chooses the max s.t.*
• The leading digits are preserved by pulling out the maximum a
• Inside the log only zero or negative numbers are exponentiated, so there can be no overflow.
• If there is underflow inside the log we know that at least the leading digits have been returned by the max.

Proof.
$$\begin{aligned} \text{LSE} &= \log \left(\sum_{i=1}^n e^{x_i} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} e^a \right) \\ &= \log \left(e^a \sum_{i=1}^n e^{x_i - a} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} \right) + \log(e^a) \\ &= \log \left(\sum_{i=1}^n e^{x_i - a} \right) + a \end{aligned}$$

Definition 17.8 Partition Π :
Given an interval $[0, T]$ a sequence of values $0 < t_0 < \dots < t_n < T$ is called a partition $\Pi(t_0, \dots, t_n)$ of this interval.

1.3. Convergence for iterative methods

Definition 17.9 Linear/Exponential Convergence: A sequence $\{x^{(k)}\}_k \in \mathbb{R}^n$ converges linearly to x^* if in the asymptotic limit $k \rightarrow \infty$ it satisfies:
 $\left\| x^{k+1} - x^* \right\| \leq \rho \left\| x^{(k)} - x^* \right\| \quad \rho \in (0, 1), \forall k \in \mathbb{N}_0$ (17.7)

Exponential Convergence
Linear convergence is sometimes called exponential convergence. This is due to the fact that:
1. We often have expressions of the form:
 $\left\| x^{k+1} - x^* \right\| \leq \underbrace{(1 - \alpha)}_{:= \rho} \left\| x^{(k)} - x^* \right\|$
2. and that $(1 - \alpha) = \exp(-\alpha)$ from which follows that:
eq. (17.8) $\iff \left\| x^{k+1} - x^* \right\| \leq e^{-\alpha} \left\| x^{(k)} - x^* \right\|$

Definition 17.10 Rate of Convergence: Is a way to measure the rate of convergence of a sequence $\{x^{(k)}\}_k \in \mathbb{R}^n$ to a value to x^* . Let $\rho \in [0, 1]$ be the *rate of convergence* and define:

$$\lim_{k \mapsto \infty} \frac{\left\| x^{k+1} - x^* \right\|}{\left\| x^{(k)} - x^* \right\|} = \rho \quad (17.8)$$

- $\rho = 1 \iff$ **Sublinear Rate** i.e. slower than linear
- $\rho \in (0, 1) \iff$ **Linear Rate**
- $\rho = 0 \iff$ **Superlinear Rate** i.e. faster then linear

Definition 17.11 Convergence of order p : In order to distinguish *superlinear convergence* we define the order of convergence.
A sequence $\{x^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \dots\}$ to x^* if it satisfies:
$$\lim_{k \mapsto \infty} \frac{\left\| x^{k+1} - x^* \right\|}{\left\| x^{(k)} - x^* \right\|^p} = C \quad C < 1 \quad (17.9)$$

Does this even exist/check if this is true

Definition 17.12 Exponential Convergence: A sequence $\{x^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate ρ to x^* if in the asymptotic limit $k \rightarrow \infty$ it satisfies:
 $\left\| x^{k+1} - x^* \right\| \leq \rho^k \left\| x^{(k)} - x^* \right\| \quad \rho < 1 \quad (17.10)$

1.4. Convention for discrization methods 2. Numerical Quadrature

Definition 17.13 Order of a Quadrature Rule: The order of a quadrature rule $\mathcal{Q}_n : C^0([a, b]) \rightarrow \mathbb{R}$ is defined as:
 $\text{order}(\mathcal{Q}_n) := \max \left\{ n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \int_a^b p(t) dt \quad \forall p \in \mathcal{P}_n \right\} + 1$ (17.11)

Thus it is the maximal degree+1 of polynomials (of degree maximal degree) \mathcal{P} maximal degree for which the quadrature rule yields exact results.

Note
Is a quality measure for quadrature rules.

2.1. Composite Quadrature

Definition 17.14 Composite Quadrature:
Given a mesh $\mathcal{M} = \{a = x_0 < x_1 < \dots < x_m = b\}$ apply a Q.R. \mathcal{Q}_n to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \dots, m \triangleq \text{p.w.}$ Quadrature:
$$\int_a^b f(t) dt = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t) dt = \sum_{j=1}^m \mathcal{Q}_n(f|_{I_j}) \quad (17.12)$$

Lemma 17.1 Error of Composite quadrature Rules:
Given a function $f \in C^k([a, b])$ with integration domain:

$$\sum_{i=1}^m h_i = |b - a| \quad \text{for } \mathcal{M} = \{x_j\}_{j=1}^m$$

Let: $h_{\mathcal{M}} = \max_j |x_j, x_{j-1}|$ be the **mesh-width**
Assume an equal number of quadrature nodes for each interval $I_j = [x_{j-1}, x_j]$ of the mesh \mathcal{M} i.e. $n_j = n$.
Then the error of a quadrature rule $\mathcal{Q}_n(f)$ of order q is given by:
$$\epsilon_n(f) = \mathcal{O} \left(n^{-\min\{k, q\}} \right) = \mathcal{O} \left(h_{\mathcal{M}}^{\min\{k, q\}} \right) \quad \text{for } n \rightarrow \infty$$

corollary 11.3 $\mathcal{O} \left(n^{-q} \right) = \mathcal{O} \left(h_{\mathcal{M}}^q \right) \quad \text{with } h_{\mathcal{M}} = \frac{1}{n}$ (17.13)

Definition 17.15 Complexity W : Is the number of function evaluations \triangleq number of quadrature points.
 $W(\mathcal{Q}(f)_n) = \#f\text{-eval} \triangleq n$ (17.14)

Lemma 17.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :
$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \iff \epsilon_n(f) = cn^{-q} \quad c \in \mathbb{R}_+$$

the error complexity is **algebraic** (??) and is given by:
$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}(\sqrt[q]{\epsilon_n}) \quad (17.15)$$

Proof. lemma 17.2: **Assume:** we want to reduce the error by a factor of ϵ_n by increasing the number of quadrature points $n_{\text{new}} = a \cdot n_{\text{old}}$.
Question: what is the additional effort ($\#f\text{-eval}$) needed in order to achieve this reduction in error?
$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \implies n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \quad (17.16)$$

Optimization

Definition 18.1 Fist Order Method: A first-order method is an algorithm that chooses the k -th iterate in
 $w_0 + \text{span}\{\nabla f(w_0), \dots, \nabla f(w_{k-1})\} \quad \forall k = 1, 2, \dots$ (18.1)

Note
Gradient descent is a first order method

1. Lagrangian Optimization Theory

Add: derivation of lagrange function

Definition 18.2 (Primal) Constraint Optimization:
Given an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:
$$\begin{aligned} &\min_{w \in \Omega} f(w) \\ \text{s.t.} \quad &g_i(w) \leq 0 \quad 1 \leq i \leq k \\ &h_j(w) = 0 \quad 1 \leq j \leq m \end{aligned}$$

Definition 18.3 Lagrange Function:
$$\mathcal{L}(w, \alpha, \beta) := f(w) + \alpha g(w) + \beta h(w) \quad (18.2)$$

Extremal Conditions
$$\begin{aligned} \nabla \mathcal{L}(w) &\stackrel{!}{=} 0 && \text{Extremal point } w^* \\ \frac{\partial}{\partial \beta} \mathcal{L}(w) &= h(w) \stackrel{!}{=} 0 && \text{Constraint satisfaction} \end{aligned}$$

For the inequality constraints $g(w) \leq 0$ we distinguish two situations:
Case I : $g(w^*) < 0$ switch const. off
Case II : $g(w^*) \geq 0$ optimize using active eq. constr.
$$\frac{\partial}{\partial \alpha} \mathcal{L}(w) = g(w) \stackrel{!}{=} 0 \quad \text{Constraint satisfaction}$$

Definition 18.4 Lagrangian Dual Problem: Is given by:
Find $\max_{w \in \Omega} \theta(\alpha, \beta) = \inf_{w \in \Omega} \mathcal{L}(w, \alpha, \beta)$
s.t. $\alpha_i \geq 0 \quad 1 \leq i \leq k$

Solution Strategy
1. Find the extremal point w^* of $\mathcal{L}(w, \alpha, \beta)$:
$$\frac{\partial \mathcal{L}}{\partial w} \Big|_{w=w^*} \stackrel{!}{=} 0 \quad (18.3)$$

2. Insert w^* into \mathcal{L} and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\frac{\partial \theta}{\partial \beta} \Big|_{\beta=\beta^*} \stackrel{!}{=} 0 \quad (18.4)$$

3. Calculate the solution $w^*(\beta^*)$ of the constraint minimization problem.

Value of the Problem
Value of the problem: the value $\theta(\alpha^*, \beta^*)$ is called the value of problem (α^*, β^*) .

Theorem 18.1 Upper Bound Dual Cost: Let $w \in \Omega$ be a feasible solution of the primal problem^[def. 18.2] and (α, β) a feasible solution of the respective dual problem^[def. 18.4].
Then it holds that:
$$f(w) \geq \theta(\alpha, \beta) \quad (18.5)$$

Proof.
$$\begin{aligned} \theta(\alpha, \beta) &= \inf_{u \in \Omega} \mathcal{L}(u, \alpha, \beta) \leq \mathcal{L}(w, \alpha, \beta) \\ &= f(w) + \sum_{i=1}^k \underbrace{\alpha_i}_{\geq 0} g_i(w) + \sum_{j=1}^m \underbrace{\beta_j}_{\leq 0} \underbrace{h_j(w)}_{=0} \\ &\leq f(w) \end{aligned}$$

Corollary 18.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:
$$\sup \{\theta(\alpha, \beta) : \alpha \geq 0\} \leq \inf \{f(w) : g(w) \leq 0, h(w) = 0\} \quad (18.6)$$

Theorem 18.2 Optimality: The triple (w^*, α^*, β^*) is a saddle point of the Lagrangian function for the primal problem, if and only if its components are optimal solutions of the primal and dual problems and if there is no duality gap, that is, the primal and dual problems having the same value:
$$f(w^*) = \theta(\alpha^*, \beta^*) \quad (18.7)$$

Definition 18.5 Convex Optimization: Given: a **convex function** f and a **convex set** S solve:

$$\begin{aligned} \min_{\mathbf{x} \in S} f(\mathbf{x}) \end{aligned} \tag{18.8}$$

Often S is specified using linear inequalities:

e.g. $S = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{A}\mathbf{x} \leq \mathbf{b} \right\}$

Theorem 18.3 Strong Duality: Given an convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w} \in \Omega} f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 & 1 \leq i \leq k \\ & h_j(\mathbf{w}) = 0 & 1 \leq j \leq m \end{aligned}$$

where g_i, h_i can be written as affine functions: $y(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b}$.

Then it holds that the **duality gap** is zero and we obtain an optimal solution.

Theorem 18.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

$$\begin{aligned} \min_{\mathbf{w} \in \Omega} f(\mathbf{w}) \\ \text{s.t.} \quad & g_i(\mathbf{w}) \leq 0 & 1 \leq i \leq k \\ & h_j(\mathbf{w}) = 0 & 1 \leq j \leq m \end{aligned}$$

with $f \in C^1$ convex and g_i, h_i affine.

Necessary and sufficient conditions for a normal point \mathbf{w}^* to be an optimum are the existence of α^*, β^* s.t.:

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha, \beta)}{\partial \mathbf{w}} \stackrel{!}{=} 0 \qquad \frac{\partial \mathcal{L}(\mathbf{w}^*, \alpha, \beta)}{\partial \beta} \stackrel{!}{=} 0 \tag{18.9}$$

under the conditions that:

- $\forall i_1, \dots, k \quad \alpha_i^* g_i(\mathbf{w}^*) = 0$, s.t.:
 - Inactive Constraint: $g_i(\mathbf{w}^*) < 0 \rightarrow \alpha_i = 0$.
 - Active Constraint: $g_i(\mathbf{w}^*) \nless 0 \rightarrow \alpha_i \geq 0 \quad \text{s.t.} \quad \alpha_i^* g_i(\mathbf{w}^*) = 0$

Consequence

We may become very sparce problems, if a lot of constraints are not actice $\iff \alpha_i = 0$.

Only a few points, for which $\alpha_i > 0$ may affact the decision surface.

Stochastics

Definition 18.6 Stochastics: Is a collective term for the areas of <i>probability theory</i> and <i>statistics</i> .
Definition 18.7 Statistics: Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.
Definition 18.8 Probability: Is concerned with the quantification of the uncertainty of random experiments by use of statistical models. Hence it is the opposite of statistics.
Definition 18.9 Probability: Probability is the measure of the likelihood that an event will occur in a Random Experiment. Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty.
<div>Improve these definitions, maybe ask on quora/hilo</div> Note: Stochastics vs. Stochastic Stochastics is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is an <i>adjective</i> , describing that a certain phenomena is governed by uncertainty i.e. a process.
Probability Theory
Definition 19.1 Probability Space $W = \{\Omega, \mathcal{F}, \mathbb{P}\}$: Is the unique triple $\{\Omega, \mathcal{F}, \mathbb{P}\}$, where Ω is its sample space, \mathcal{F} is its σ -algebra of events, and \mathbb{P} its probability measure.
Definition 19.2 Sample Space Ω: Is the set of all possible outcomes (elementary events corollary 19.5) of an experiment see example 19.1
Definition 19.3 Event A : An “event” is a subset of the sample space Ω and is a property which can be observed to hold or not to hold <i>after</i> the experiment is done (example 19.2). Mathematically speaking not every subset of Ω is an event and has an associated probability. Only those subsets of Ω that are part of the corresponding σ -algebra \mathcal{F} are events and have their assigned probability.
Corollary 19.1 : If the outcome ω of an experiment is in the subset A , then the event A is said to “have ocured”.
Corollary 19.2 Complement Set A^C : is the contrary event of A .
Corollary 19.3 The Union Set $A \cup B$: Let A, B be to evenest. The event “ A or B ” is interpreted as the union of both.
Corollary 19.4 The Intersection Set $A \cap B$: Let A, B be to evenest. The event “ A and B ” is interpreted as the intersection of both.
Corollary 19.5 The Elementary Event ω : Is a “singleton”, i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω .
Corollary 19.6 The Sure Event Ω : Is equal to the sample space as it contains all possible elementary events.
Corollary 19.7 The Impossible Event \emptyset : The impossible event i.e. nothing is happening is denoted by the empty set.
Definition 19.4 The Family of All Events $\mathcal{A}/2^\Omega$: The set of all subset of the sample space Ω called family of all events is given by the power set of the sample space $\mathcal{A} = 2^\Omega$ (for finite sample spaces).

Definition 19.5 Probability $\mathbb{P}(A)$: Is a number associated with every A , that measures the likelihood of the event to be realized “a priori”. The bigger the number the more likely the event will happen. 1. $0 \leq \mathbb{P}(A) \leq 1$ 2. $\mathbb{P}(\Omega) = 1$ 3. If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
Note We can think of the probability of an event A as the limit of the "frequency" of repeated experiments: $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \frac{\delta(A)}{n} \quad \text{where} \quad \delta(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$
0.1. Sigma Algebras
Definition 19.6 Sigma Algebra σ: A set \mathcal{F} of subsets of Ω is called a σ -algebra on Ω if the following properties apply <ul style="list-style-type: none">$\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$If $A \in \mathcal{F}$ then $\Omega \setminus A = A^C \in \mathcal{F}$: The complementary subset of A is also in Ω.For all $A_i \in \mathcal{F} : \bigcup_{i=1} A_i \in \mathcal{F}$ See example 19.3.
Corollary 19.8 \mathcal{F}_{\min}: $\mathcal{F} = \{\emptyset, \Omega\}$ is the simplest σ -algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.
Corollary 19.9 \mathcal{F}_{\max}: $\mathcal{F} = 2^\Omega$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.
Definition 19.7 Measurable Space (Ω, \mathcal{F}) : Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.
Corollary 19.10 \mathcal{F}-measurable Event: The elements $A_i \in \mathcal{F}$ are called <i>measurable sets</i> or <i>\mathcal{F}-measurable</i> .
Interpretation The σ -algebra represents all of possible events of the experiment that we can detect. Thus we call the sets in \mathcal{F} measurable sets/events. The sigma algebra is the mathematical construct that tells us how much information we obtain once we conduct some experiment.
Definition 19.8 Sigma Algebra generated by a subset of Ω $\sigma(\mathcal{C})$: Let \mathcal{C} be a class of subsets of Ω . The σ -algebra generated by \mathcal{C} , denoted by $\sigma(\mathcal{C})$, is the <i>smallest</i> sigma algebra \mathcal{F} that included all elements of \mathcal{C} see example 19.4.
Definition 19.9 Borel σ-algebra $\mathcal{B}(\mathbb{R})$: The Borel σ -algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ -algebra containing all open intervals in \mathbb{R} . The sets in contained in $\mathcal{B}(\mathbb{R})$ are called Borel sets. The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward. For all real numbers $a, b \in \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ contains various sets see example 19.5.
Why do we need Borel Sets So far we only looked at atomic events ω , with the help of sigma algebras we are now able to measure continuous events s.a. $[0, 1]$.
Corollary 19.11 : The Borel σ -algebra of \mathbb{R} is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ (\mathbb{Q} =rationals). See proof section 13.
Definition 19.10 (\mathbb{P})-trivial Sigma Algebra: is a σ -algebra \mathcal{F} for which each event has a probability of zero or one: $\mathbb{P}(A) \in \{0, 1\} \quad \forall A \in \mathcal{F} \quad (19.1)$

Interpretation A trivial sigma algebra means that all events are almost surely constant and that there exist no non-trivial information. An example of a trivial sigma algebra is $\mathcal{F}_{\min} = \{\Omega, \emptyset\}$.
0.2. Measures
Definition 19.11 Measure μ : A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map: $\mu : \mathcal{F} \mapsto [0, \infty]$ for which holds: <ul style="list-style-type: none">$\mu(\emptyset) = 0$countable additivity [def. 19.12]
Definition 19.12 Countable/σ-Additive Function: Given a function μ defined on a σ -algebra \mathcal{F} . The function μ is said to be countable additive if for every countable sequence of pairwise disjoint elements $(F_i)_{i \geq 1}$ of \mathcal{F} it holds that: $\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \emptyset \quad \forall j \neq k \quad (19.3)$
Corollary 19.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds: $\mu(F \cup G) = \mu(F) + \mu(G) \iff F \cap G = \emptyset \quad (19.4)$
Intuition If we take two event that cannot occur simultaneously, then the probability that at least one vent occurs is just the sum of the measure (probabilities) of the original events.
Definition 19.13 Equivalent Measures $\mu \sim \nu$: Let μ and ν be two measures defined on a measurable space [def. 19.7] (Ω, \mathcal{F}) . The two measures are said to be equivalent if it holds that: $\mu(A) > 0 \iff \nu(A) > 0 \quad \forall A \subseteq \mathcal{F} \quad (19.5)$ this is equivalent to μ and ν having equivalent null sets: $\mathcal{N}_\mu = \mathcal{N}_\nu \quad \begin{matrix} \mathcal{N}_\mu = \{A \in \mathcal{A} \mu(A) = 0\} \\ \mathcal{N}_\nu = \{A \in \mathcal{A} \nu(A) = 0\} \end{matrix} \quad (19.6)$ see example 19.6
Definition 19.14 Measure Space $\{\mathcal{F}, \Omega, \mu\}$: The triplet of sample space, sigma algebra and a measure is called a measure space.
Definition 19.15 Lebesgue Measure on \mathcal{B} λ : Is the measure defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns the measure of each interval to be its length: $\lambda([a, b]) = b - a \quad (19.7)$
Corollary 19.13 Lebesgue Measure of Atomitics: <ul style="list-style-type: none">The Lebesgue measure of a set containing only one point must be zero: $\lambda(\{a\}) = 0 \quad (19.8)$The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2, \dots, a_n\}$ must be zero: $\lambda(A) + \sum_{i=1}^n \lambda(\{a_i\}) = 0 \quad (19.9)$The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2, \dots\}$ can be either zero, positive and finite or infinite.
0.3. Probability/Kolomogorov's Axioms 1931
One problem we are still having is the range of μ , by standardizing the measure we obtain a well defined measure of events.
Axiom 19.1 Non-negativity: The probability of an event is a non-negative real number: If $A \in \mathcal{F}$ then $\mathbb{P}(A) \geq 0 \quad (19.10)$

Axiom 19.2 Unitaarity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one: The certain event $\mathbb{P}(\Omega) = 1 \quad (19.11)$
Axiom 19.3 σ-additivity: If $A_1, A_2, A_3, \dots \in \mathcal{F}$ are mutually disjoint, then: $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad (19.12)$
Corollary 19.14 : As a consequence of this it follows: $\mathbb{P}(\emptyset) = 0 \quad (19.13)$
Corollary 19.15 Complementary Probability: $\mathbb{P}(A^C) = 1 - \mathbb{P}(A) \quad \text{with} \quad A^C = \Omega - A \quad (19.14)$
Definition 19.16 Probability Measure \mathbb{P} : a probability measure is function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ defined on a σ -algebra \mathcal{F} of a sample space Ω that satisfies the probability axioms.
1. Conditional Probability
Definition 19.17 Conditional Probability: Let A, B be events, with $\mathbb{P}(B) \neq 0$. Then the conditional probability of the event A given B is defined as: $\mathbb{P}(A B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \mathbb{P}(B) \neq 0 \quad (19.15)$
2. Independent Events
Theorem 19.1 Independent Events: Let A, B be two events. A and B are said to be independent iff: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \quad \begin{matrix} \mathbb{P}(A B) = \mathbb{P}(A), & \mathbb{P}(B) > 0 \\ \mathbb{P}(B A) = \mathbb{P}(B), & \mathbb{P}(A) > 0 \end{matrix} \quad (19.16)$
Note The requirement of no impossible events follows from [def. 19.17]
Corollary 19.16 Pairwise Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>pairwise independent</i> if every pair of events is independent: $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \mathbb{P}(A_j) \quad i \neq j, \quad \forall i, j \in \mathcal{A} \quad (19.17)$
Corollary 19.17 Mutal Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>mutal independent</i> if every event A_j is independent of any intersection of the other events: $\mathbb{P}\left(\bigcap_{i=1}^k B_i\right) = \prod_{i=1}^k \mathbb{P}(B_i) \quad \begin{matrix} \forall \{B_i\}_{i=1}^k \subseteq \{A_i\}_{i=1}^n \\ k \leq n, \quad \{A_i\}_{i=1}^n \in \mathcal{A} \end{matrix} \quad (19.18)$
3. Product Rule
Law 19.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by: $\mathbb{P}(A \cap B) = \mathbb{P}(B A)\mathbb{P}(A) = \mathbb{P}(A B)\mathbb{P}(B) \quad (19.19)$
Law 19.2 Generalized Product Rule/Chain Rule: is the generalization of the product rule?? to n events $\{A_i\}_{i=1}^n$ $\mathbb{P}\left(\bigcap_{i=1}^k E_i\right) = \prod_{k=1}^n \mathbb{P}\left(E_k \middle \bigcap_{i=1}^{k-1} E_i\right) = \quad (19.20)$ $= \mathbb{P}(E_n E_{n-1} \cap \dots \cap E_1) \cdot \mathbb{P}(E_{n-1} E_{n-2} \cap \dots \cap E_1) \dots \dots \mathbb{P}(E_3 E_2 \cap E_1) \mathbb{P}(E_2 E_1) \mathbb{P}(E_1)$

4. Law of Total Probability

Definition 19.18 Complete Event Field: A complete event field $\{A_i : i \in I \subseteq \mathbb{N}\}$ is a countable or finite partition of Ω that is the partitions $\{A_i : i \in I \subseteq \mathbb{N}\}$ are a *disjoint union* the sample space:

$$\bigcup_{i \in I} A_i = \Omega \quad A_i \cap A_j = \emptyset \quad i \neq j, \forall i, j \in I \quad (19.21)$$

Theorem 19.2

Law of Total Probability/Partition Equation: Let $\{A_i : i \in I\}$ be a complete event field^[def. 19.18] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \quad (19.22)$$

5. Bayes Theorem

Law 19.3 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \mathbb{P}(B) > 0 \quad (19.23)$$

follows directly from eq. (19.19).

Theorem 19.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field^[def. 19.18] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \quad (19.24)$$

proof section 13

Distributions on \mathbb{R}

6.1. Distribution Function

Definition 19.19 Distribution Function of \mathbb{P} F : The distribution function F induced by a probability \mathbb{P} on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((-\infty, x]) \quad (19.25)$$

Theorem 19.4 : A function F is the distribution function of a (unique) probability on $(\mathbb{R}, \mathcal{B})$ iff:

- F is non-decreasing
- F is right continuous
- $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$

Corollary 19.18 : A probability \mathbb{P} is uniquely determined by a distribution function F . That is if there exist another probability \mathbb{Q} s.t.

$$G(x) = \mathbb{Q}((-\infty, x])$$

and if $F = G$ then it follows $\mathbb{P} = \mathbb{Q}$.

6.2. Random Variables

A random variable X is a quantity that is not a variable in the classical sense but a variable with respect to the outcome of an experiment. Thus it is actually not a variable but a function/map.

Its value is determined in two steps:

- The outcome of an experiment is a random quantity $\omega \in \Omega$
- The outcome ω determines (possibly various) quantities of interests \iff random variables

Thus a random variable X , defined on a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ is a mapping from Ω into another space \mathcal{E} , usually $\mathcal{E} = \mathbb{R}$ or $\mathcal{E} = \mathbb{R}^n$:

$$X : \Omega \mapsto \mathcal{E} \quad \omega \mapsto X(\omega)$$

Let now $E \in \mathcal{E}$ be a quantity of interest, in order to quantify its probability we need to map it back to the original sample space Ω :

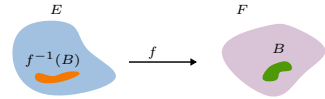
Probability for an event in Ω

$$\mathbb{P}_X(E) = \mathbb{P}(\{\omega : X(\omega) \in E\}) = \mathbb{P}(X \in E) = \mathbb{P}(X^{-1}(E))$$

Probability for an event in E

Definition 19.20 \mathcal{E} -measurable function: Let (E, \mathcal{E}) and (F, \mathcal{F}) be two measurable spaces. A function $f : E \mapsto F$ is called measurable (relative to \mathcal{E} and \mathcal{F}) if

$$\forall B \in \mathcal{F} : \quad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \quad (19.26)$$



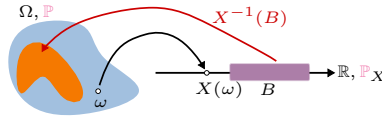
Interpretation

The pre-image^[def. 11.9] of B under f i.e. $f^{-1}(B)$ maps all values of the target space F back to the sample space \mathcal{E} (for all possible $B \in \mathcal{F}$).

Definition 19.21 Random Variable: A real-valued random variable (vector) X , defined on a probability space $\{\Omega, \mathcal{E}, \mathbb{P}\}$ is an \mathcal{E} -measurable function mapping, if it maps its sample space Ω into a target space (F, \mathcal{F}) :

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \quad (19.27)$$

Since X is \mathcal{E} -measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 19.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ -algebra \mathcal{B} of \mathbb{R} .

Corollary 19.20 Random Variables of Borel Sets: Given that we work with Borel σ -algebras then the definition of a random variable is equivalent to (due to corollary 19.11):

$$X^{-1}(B) = X^{-1}((-\infty, a]) = \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \quad (19.28)$$

Definition 19.22

Realization of a Random Variable $x = X(\omega)$: Is the value of a random variable that is actually observed after an experiment has been conducted. In order to avoid confusion lower case letters are used to indicate actual observations/realization of a random variable.

Corollary 19.21 Indicator Functions

$I_A(\omega)$: An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \quad (19.29)$$

We know that a probability measure \mathbb{P} on \mathbb{R} is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 19.22 : Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let (E, \mathcal{E}) and arbitrary measurable space. Let X be a real value function on E .

Then it holds that X is measurable if and only if $\{X \leq a\} = \{\omega : X(\omega) \leq a\} = X^{-1}((-\infty, a]) \in \mathcal{E}$, each $a \in \mathbb{R}$ or $\{X < a\} \in \mathcal{E}$.

Explanation 19.1 (corollary 19.22). A random variable is a function that is measurable if and only if its distribution function is defined.

6.3. The Law of Random Variables

Definition 19.23 Law/Distribution of X $\mathcal{L}(X)$: Let X be a r.v. on $\{\Omega, \mathcal{F}, \mathbb{P}\}$, with values in (E, \mathcal{E}) , then the distribution/law of X is defined as:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \quad (19.30)$$

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \quad \forall B \in \mathcal{E}$$

Note

- Sometimes \mathbb{P}^X is also called the *image* of \mathbb{P} by X
- The law can also be written as:
 $\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)) = (\mathbb{P} \circ X^{-1})(B)$

Theorem 19.5 : The law/distribution of X is a probability measure \mathbb{P} on (E, \mathcal{E}) .

Definition 19.24

(Cumulative) Distribution Function F_X : Given a real-valued r.v. then its cumulative distribution function is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x) \quad (19.31)$$

Corollary 19.23 : The distribution of \mathbb{P}^X of a real valued r.v. is entirely characterized by its cumulative distribution function F_X ^[def. 19.31].

Property 19.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \quad (19.32)$$

Property 19.2: Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad (19.33)$$

6.4. Probability Density Function

Definition 19.25 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 19.26 Probability Density Function: Let X be a r.v. with associated cdf F_X . If F_X is continuously integrable for all $x \in \mathbb{R}$ then X has a probability density f_X defined by:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy \quad (19.34)$$

or alternatively:

$$f_X(x) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x \leq X \leq x + \epsilon)}{\epsilon} \quad (19.35)$$

Corollary 19.24 $\mathbb{P}(X = b) = 0$, $\forall b \in \mathbb{R}$:

$$\mathbb{P}(X = b) = \lim_{a \rightarrow b} \mathbb{P}(a < X \leq b) = \lim_{a \rightarrow b} \int_a^b f(x) dx = 0 \quad (19.36)$$

Corollary 19.25 corollary 19.24: From corollary 19.24 it follows that the exact borders are not necessary:

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$$

Corollary 19.26 :

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (19.37)$$

Notes

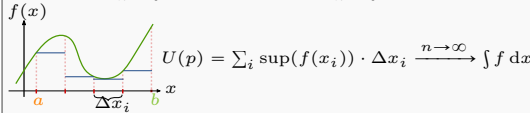
- Often the cumulative distribution function is referred to as "cdf" or simply *distribution function*.
- Often the probability density function is referred to as "pdf" or simply *density*.

6.5. Lebesgue Integration

Problems of Riemann Integration

- Difficult to extend to higher dimensions – general domains of definitions $f : \Omega \mapsto \mathbb{R}$
- Depends on continuity
- Integration of limit processes require strong uniform convergence in order to integrate limit processes

$$\lim_{n \rightarrow \infty} \int f(x) dx \stackrel{\text{str. u.c.}}{=} \int \lim_{n \rightarrow \infty} f(x) dx$$

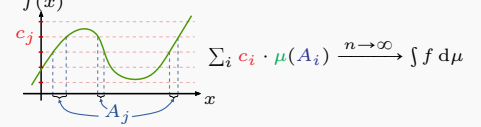


Idea

Partition domain by function values of equal size i.e. values that lie within the same sets/have the same value A_j build up the partitions w.r.t. to the variable x .

Problem: we do not know how big those sets/partitions on the x -axis will be.

Solution: we can use the measure μ of our measure space $\{\Omega, \mathcal{A}, \mu\}$ in order to obtain the size of our sets $A_j \Rightarrow$ we do not have to care anymore about discontinuities, as we can measure the size of our sets using our measure.



Definition 19.27 Lebesgue Integral:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n c_i \mu(A_i) = \int_{\Omega} f d\mu \quad f(x) \approx c_i \quad \forall x \in A_i \quad (19.38)$$

Definition 19.28

Simple Functions (Random Variables): A r.v. X is called simple if it takes on only a finite number of values and hence can be written in the form:

$$X = \sum_{i=1}^n a_i \mathbb{I}_{A_i} \quad a_i \in \mathbb{R} \quad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \quad (19.39)$$

6.6. Independent Random Variables

We have seen that two events A and B are independent if knowledge that B has occurred does not change the probability that A will occur theorem 19.1.

For two random variables X, Y we want to know if knowledge of Y leaves the probability of X , to take on certain values unchanged.

Definition 19.29 Independent Random Variables:

Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leq x | Y \leq y) = \mathbb{P}(X \leq x) \quad \forall x, y \in \mathbb{R} \quad (19.40)$$

which amounts to:

$$F_{X,Y}(x, y) = \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(X \leq x, Y \leq y) = F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R} \quad (19.41)$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \quad \forall A, B \in \mathcal{B} \quad (19.42)$$

Note

If the joint distribution $F_{X,Y}(x, y)$ can be factorized into two functions of x and y then X and Y are independent.

Definition 19.30

Independent Identically Distributed:

7. Product Rule

Law 19.4 Product Rule: Let X, Y be two random variables then their jo

Law 19.5

Generalized Product Rule/Chain Rule:

8. Change Of Variables Formula

Formula 19.1

(Scalar Discret) Change of Variables: Let X be a discret rv $X \in \mathcal{X}$ with pmf \mathbb{P}_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic^[def. 11.12] function.

Let: $\mathcal{X}_y = x_i$ be the set of all $x_i \in \mathcal{X}$ s.t. $y = g(x_i)$.

Then the pmf of Y is given by:

$$\mathbb{P}_Y(y) = \sum_{x_i \in \mathcal{X}_y} \mathbb{P}_X(x_i) = \sum_{x \in \mathcal{Y} : g(x) = y} \mathbb{P}_X(x) \quad (19.43)$$

see proof section 13

Formula 19.2
(Scalar Continuous) Change of Variables:
 Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic^[def. 11.12] function.
 Define a new r.v. Y as

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \tag{19.44}$$
 then the pdf of Y is given by:

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X(x) \left| \frac{d}{dy} (g^{-1}(y)) \right| \tag{19.45}$$

$$= f_X(x) \frac{1}{\left| \frac{dy}{dx} \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx}(g^{-1}(y)) \right|} \tag{19.46}$$

Formula 19.3
(Continuous) Change of Variables:
 Let $X = \{X_1, \dots, X_n\} \sim f_X$ be a continuous random vector and let g be an arbitrary strictly monotonic^[def. 11.12] function $g: \mathbb{R}^n \mapsto \mathbb{R}^m$

 Define a new r.v. Y as

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \tag{19.47}$$
 and let $h(x) := g(x)^{-1}$ then the pdf of Y is given by:

$$\begin{aligned} f_Y(y) &= f_X(x_1, \dots, x_n) \cdot |J| \\ &= f_X(h_1(y), \dots, h_n(y)) \cdot |J| \\ &= f_X(y) |\det D_x h(x)| \Big|_{x=y} \\ &= f_X(g^{-1}(y)) \left| \det \left(\frac{\partial g}{\partial x} \right) \right|^{-1} \end{aligned} \tag{19.48}$$
 where $J = \det Dh$ is the Jacobian^[def. 12.4].
 See also proof section 13 and example 19.8

Note
 A monotonic function is required in order to satisfy inevitability.

Probability Distributions on \mathbb{R}^n

10. Joint Distribution

Definition 19.31
Joint (Cumulative) Distribution Function F_X :
 Let $X = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n , then its *cumulative distribution function* is defined as:

$$\begin{aligned} F_X(x) &= \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x) \\ &= \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \end{aligned} \tag{19.49}$$

Definition 19.32 Joint Probability Distribution:
 Let $X = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n with associated cdf F_X . If F_X is continuously integrable for all $x \in \mathbb{R}$ then X has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_X(y_1, \dots, y_n) dy_1 dy_n \tag{19.50}$$

or alternatively:

$$f_X(x) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x_1 \leq X_1 \leq x_1 + \epsilon, \dots, x_n \leq X_n \leq x_n + \epsilon)}{\epsilon} \tag{19.51}$$

10.1. Marginal Distribution

Definition 19.33 Marginal Distribution:

11. The Expectation

Definition 19.34 Expectation:

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) \mathbb{P}(d\omega) = \int_{\Omega} X d\mathbb{P} \tag{19.52}$$

Corollary 19.27 Expectation of simple r.v.:
 If X is a simple^[def. 19.28] r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \tag{19.53}$$

11.1. The Jensen Inequality

Theorem 19.6 Jensen Inequality: Let X be a random variable and g some function, then it holds:

$$\begin{aligned} g(\mathbb{E}[X]) &\geq \mathbb{E}[g(X)] && \text{if } g \text{ is convex}^{\text{[def. 11.20]}} \\ g(\mathbb{E}[X]) &\leq \mathbb{E}[g(X)] && g \text{ is concave}^{\text{[def. 11.21]}} \end{aligned} \tag{19.54}$$

11.2. Law of the Unconscious Statistician

Law 19.6 Law of the Unconscious Statistician:
 Let $X \in \mathcal{X}, Y \in \mathcal{Y}$ be random variables where Y is defined as:

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$$
 then the expectation of Y can be calculated in terms of X :

$$\mathbb{E}_Y[y] = \mathbb{E}_X g(x) \tag{19.55}$$

Consequence

Hence if we \mathbf{p}_X we do not have to first calculate \mathbf{p}_Y in order to calculate $\mathbb{E}_Y[y]$.

12. Moment Generating Function (MGF)

Definition 19.35 Moment of Random Variable: The i -th moment of a random variable X is defined as (if it exists):

$$m_i := \mathbb{E}[X^i] \tag{19.56}$$

Definition 19.36
Moment Generating Function (MGF):

$$\psi_X(t) = \mathbb{E}\left[e^{tX}\right] \quad t \in \mathbb{R} \tag{19.57}$$

Corollary 19.28 Sum of MGF: The moment generating function of a sum of n independent variables $(X_j)_{1 \leq j \leq n}$ is the product of the moment generating functions of the components:

$$\psi_{S_n}(t) = \psi_{X_1}(t) \cdots \psi_{X_n}(t) \quad S_n := X_1 + \ldots X_n \tag{19.58}$$

Corollary 19.29 : The i -th moment of a random variable is the i -th derivative of its associated moment generating function evaluated zero:

$$\mathbb{E}[X^i] = \psi_X^{(i)}(0) \tag{19.59}$$

13. The Characteristic Function

Transforming probability distributions using the Fourier transform is a handy tool in probability in order to obtain properties or solve problems in another space before transforming them back.

Definition 19.37
Fourier Transformed Probability Measure:

$$\hat{\mu} = \int e^{i\langle u, x \rangle} \mu(dx) \tag{19.60}$$

Corollary 19.30 : As $e^{i\langle u, x \rangle}$ can be rewritten using formulaeqs. (15.5) and (15.6) it follows:

$$\hat{\mu} = \int \cos(\langle u, x \rangle) \mu(dx) + i \int \sin(\langle u, x \rangle) \mu(dx) \tag{19.61}$$
 where $x \mapsto \cos(\langle x, u \rangle)$ and $x \mapsto \sin(\langle x, u \rangle)$ are both bounded and Borel i.e. Lebesgue integrable.

Definition 19.38 Characteristic Function φ_X : Let X be an \mathbb{R}^n -valued random variable. Its characteristic function φ_X is defined on \mathbb{R}^n as:

$$\begin{aligned} \varphi_X(u) &= \int e^{i\langle u, x \rangle} \mathbb{P}^X(dx) = \widehat{\mathbb{P}^X}(u) \\ &= \mathbb{E}\left[e^{i\langle u, x \rangle}\right] \end{aligned} \tag{19.62}$$

$$= \mathbb{E}\left[e^{i\langle u, x \rangle}\right] \tag{19.63}$$

Corollary 19.31 : The characteristic function φ_X of a distribution always exists as it is equal to the Fourier transform of the probability measure, which always exists.

Note

This is an advantage over the moment generating function.

Theorem 19.7 : Let μ be a probability measure on \mathbb{R}^n . Then $\hat{\mu}$ is a bounded continuous function with $\hat{\mu}(0) = 1$.

Theorem 19.8 Uniqueness Theorem: The Fourier Transform $\hat{\mu}$ of a probability measure μ on \mathbb{R}^n characterizes μ . That is, if two probability measures on \mathbb{R}^n admit the same Fourier transform, they are equal.

add proof

Corollary 19.32 : Let $X = (X_1, \dots, X_n)$ be an \mathbb{R}^n -valued random variable. Then the real valued r.v.'s $(X_j)_{1 \leq j \leq n}$ are independent if and only if:

$$\varphi_X(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \tag{19.64}$$

Proofs

Proof. corollary 19.11: Let \mathcal{C} denote all open intervals. Since every open set in \mathbb{R} is the countable union of open intervals^[def. 9.8], it holds that $\sigma(\mathcal{C})$ is the Borel σ -algebra of \mathbb{R} .
 Let \mathcal{D} denote all intervals of the form $(-\infty, a]$, $a \in \mathbb{Q}$.
 Let $a, b \in \mathcal{C}$, and let

- $(a_n)_{n>1}$ be a sequence of rationals *decreasing* to a and
- $(b_n)_{n>1}$ be a sequence of rationals *increasing strictly* to b

 $(a, b) = \cup_{n=1}^{\infty} (a_n, b_n] = \cup_{n=1}^{\infty} ((-\infty, b_n] \cap (-\infty, a_n]^C)$
 Thus $\mathcal{C} \subset \sigma(\mathcal{D})$, whence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$ **but** as each element of \mathcal{D} is a closed subset, $\sigma(\mathcal{D})$ must also be contained in the Borel sets \mathcal{B} with

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{D}) \subset \mathcal{B}$$

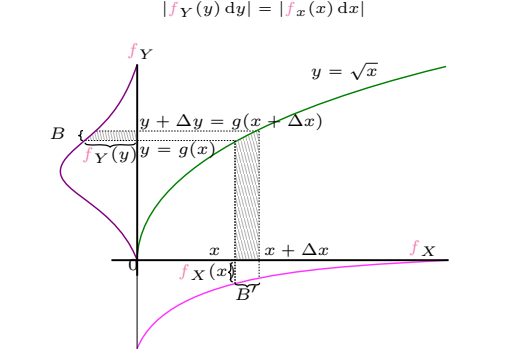
Proof. theorem 19.3 Plug eq. (19.22) into the denominator and ?? into the numerator and then use^[def. 19.17]:

$$\frac{\mathbb{P}(B|A_j) \mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i) \mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

Proof. formula 19.1:

$$Y = g(X) \quad \Longleftrightarrow \quad \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = \mathbf{p}_Y(y)$$

Proof. formula 19.2 (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:



Proof. formula 19.2 from CDF:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \begin{cases} \mathbb{P}(X \leq g^{-1}(y)) & \text{if } g \text{ is increas.} \\ \mathbb{P}(X \geq g^{-1}(y)) & \text{if } g \text{ is decreas.} \end{cases}$$
 If g is monotonically increasing:

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$
 If g is monotonically decreasing:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$

Proof. formula 19.2: Let $B = [x, x + \Delta x]$ and $B' = [y, y + \Delta y] = [g(x), g(x + \Delta x)]$ we know that the probability of equal events is equal:

$$y = g(x) \quad \Rightarrow \quad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$
 Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t) dt \xrightarrow{\Delta x \rightarrow 0} |\Delta x \cdot f_X(x)|$$
 For y we use Taylor (??)

$$\begin{aligned} g(x + \Delta x) &\stackrel{\text{eq. (11.41)}}{=} g(x) + \frac{dg}{dx} \Delta y \quad \text{for } \Delta x \rightarrow 0 \\ &= y + \Delta y \quad \text{with } \Delta y := \frac{dg}{dx} \cdot \Delta x \end{aligned} \tag{19.65}$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\begin{aligned} \mathbb{P}(X \in B') &= \int_y^{y+\Delta y} f_Y(t) dt \xrightarrow{\Delta y \rightarrow 0} |\Delta y \cdot f_Y(y)| \\ &= \left| \frac{dg}{dx}(x) \Delta x \cdot f_Y(y) \right| \end{aligned}$$

Now we simply need to related the surface of the two pdfs:

$$B = [x, x + \Delta x] \stackrel{\text{same surfaces}}{\propto} [y, y + \Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\stackrel{\Delta y \rightarrow 0}{\Longleftrightarrow} |f_Y(y) \cdot \Delta y| = \left| f_Y(y) \cdot \frac{dg}{dx}(x) \Delta x \right| = |f_X(x) \cdot \Delta x|$$

$$f_Y(y) \cdot \left| \frac{dg}{dx}(x) \right| |\Delta x| = f_X(x) \cdot |\Delta x|$$

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

Examples

Example 19.1 :

- Toss of a coin (with head and tail): $\Omega = \{H, T\}$.
- Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$
- A cubic die: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
- The positive integers: $\Omega = \{1, 2, 3, \dots\}$
- The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$

Example 19.2 :

- Head in coin toss $A = \{H\}$
- Odd number in die roll: $A = \{\omega_1, \omega_3, \omega_5, \}$
- The integers smaller five: $A = \{1, 2, 3, 4\}$

Example 19.3 : If the sample space is a die toss $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, the sample space may be that we are only told whether an even or odd number has been rolled:

$$\mathcal{F} = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

Example 19.4 : If we are only interested in the subset-set $\mathcal{A} \in \Omega$ of our experiment, then we can look at the corresponding generating σ -algebra $\sigma(\mathcal{A}) = \{\emptyset, \mathcal{A}, \mathcal{A}^C, \Omega\}$.

Example 19.5 :

- open half-lines: $(-\infty, a)$ and (a, ∞) ,
- union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
- closed interval: $[a, b] = (-\infty, \cup a) \cup (b, \infty)$,
- closed half-lines:
 $(-\infty, a] = \bigcup_{n=1}^{\infty} [a - n, a]$ and $[a, \infty) = \bigcup_{n=1}^{\infty} [a, a + n]$,
- half-open and half-closed $(a, b] = (-\infty, b] \cup (a, \infty)$,
- every set containing only one real number:
 $\{a\} = \bigcap_{n=1}^{\infty} (a - \frac{1}{n}, a + \frac{1}{n})$,
- every set containing finitely many real numbers:
 $\{a_1, \dots, a_n\} = \bigcup_{k=1}^n a_k$.

Example 19.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\} \qquad \mathbb{P}(\{1, 2, 3\}) = \{2/3, 1/6, 1/6\}$$

$$\qquad \qquad \qquad \mathbb{P}(\{1, 2, 3\}) = \{1/3, 1/3, 1/3\}$$

Example 19.7 :

add example fat book p.1286

add example prob th book 4

Example 19.8 formula 19.2: Let $X, Y \overset{\text{ind.}}{\sim} \mathcal{N}(0, 1)$.
Question: proof that:

$$U = X + Y \qquad V = X - 1$$
are indepent and normally distributed:

$$h(u, v) = \begin{cases} h_1(u, v) = \frac{u+v}{2} \\ h_2(u, v) = \frac{u-v}{2} \end{cases} \qquad J = \det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$

$$\begin{aligned} f_{U,V} &= f_{X,Y}(\underline{x}, \underline{y}) \cdot \frac{1}{2} \\ &\stackrel{\text{indp.}}{=} f_X(x) \cdot f_Y(y) \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\left\{ \left(\frac{u+v}{2} \right)^2 + \left(\frac{u-v}{2} \right)^2 \right\} / 2} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{v^2}{4}} \end{aligned}$$

Thus U, V are independent r.v. distributed as $\mathcal{N}(0, 2)$.

Combinatorics

0.1. Permutations

Definition 20.1 Permutation $n!$: Given a set^[def. 9.1] S of n distinct objects, into how many distinct sequences/orders can we arrange/permutate those distinct objects

$$P(S) = n! \iff P(S) = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1 \qquad (20.1)$$

If there exists multiple n_j objects of the same kind within S with $j \in 1, \dots, n-1$ then we need to divide by those permutations:

$$P(S) = \frac{n!}{n_1! \cdot \dots \cdot n_k} \qquad \text{s.t.} \qquad \sum_{i=1}^k n_i \leq n \qquad (20.2)$$

Note

This is because the sequence/order does not change if we exchange objects of the same kind (e.g. red ball by red ball).

Statistics

Delete/Move the following stuff appropriately

The probability that a discret random variable x is equal to some value $\bar{x} \in \mathcal{X}$ is:

$$\mathbb{P}_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapt

Definition 21.1 Almost Surely (a.s.): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \iff \omega \text{ happens a.s.} \qquad (21.1)$$

Definition 21.2 Probability Mass Function (PMF):

Definition 21.3 Discrete Random Variable (DVR): The set of possible values \bar{x} of \mathcal{X} is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \dots, 8\} \qquad \mathcal{X} = \mathbb{N} \qquad (21.2)$$

Definition 21.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \rightarrow [0, \infty)$ that satisfies:

Non-negativity: $f(x) \geq 0, \quad \forall x \in \mathbb{R}^n \qquad (21.3)$

Normalization: $\int_{-\infty}^{\infty} f(x) \, dx \stackrel{!}{=} 1 \qquad (21.4)$

Must be integrable

Note: why do we need probability density functions

A continuous random variable X can realise an infinite count of real number values within its support B (as there are an infinitude of points in a line segment).
Thus we have an infinitude of values whose sum of probabilities must equal one.
Thus these probabilities must each be zero otherwise we would obtain a probability of ∞ . As we can not work with zero probabilities we use the next best thing, infinitesimal probabilities (defined as a limit).
We say they are almost surely equal to zero:

$$\mathbb{P}(X = x) = 0 \qquad \text{a.s.}$$

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 21.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf ^[def. 21.4] f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x) \, dx \qquad (21.6)$$

Property 21.1 Zero Probability: If X is a continuous rrv ^[def. 21.5], then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \qquad (21.7)$$

Property 21.2 Open vs. Closed Intervals: For any real numbers a and b , with $a < b$ it holds:

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) \\ &= \mathbb{P}(a < X < b) \end{aligned} \qquad (21.8)$$

\iff including or not the bounds of an interval does not modify the probability of a continuous rrv.

Note

Changing the value of a function at finitely many points has no effect on the value of a definite integral.

Corollary 21.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_x(x) \, dx$$

Proof. Property 21.1:

$$\begin{aligned} \mathbb{P}(X = a) &= \lim_{\Delta x \rightarrow 0} \mathbb{P}(X \in [a, a + \Delta x]) \\ &= \lim_{\Delta x \rightarrow 0} \int_a^{a+\Delta x} f_X(x) \, dx = 0 \end{aligned}$$

□

Proof. Property 21.2:

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) \\ &= \mathbb{P}(a < X < b) = \int_a^b f_X(x) \, dx \end{aligned}$$

□

Definition 21.6 Support of a probability density function: The support of the density of a pdf $f_X(\cdot)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\text{supp}((f_X) := \{x \in \mathcal{X} | f(x) > 0\} \qquad (21.9)$$

Note: this is not a rigorous definition.

Theorem 21.1 RVs are defined by a PDFs: A probability density function f_X completely determines the distribution of a continuous real-valued random variable X .

Corollary 21.2 Identically Distributed: From theorem 21.1 it follows that to RV X and Y that have exactly the same pdf follow the same distribution.
We say X and Y are **identically distributed**.

0.1. Cumulative Distribution Fuction

Definition 21.7 Cumulative distribution function (CDF): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The (cumulative) distribution function of a real-valued random variable X is the function given by:

$$F_X(x) = \mathbb{P}(X \leq x) \qquad \forall x \in \mathbb{R}$$

Property 21.3:

Monotonically Increasing $x \leq y \iff F_X(x) \leq F_X(y) \quad \forall x, y \in \mathbb{R}$

$$\qquad \qquad \qquad (21.10)$$

Upper Limit $\lim_{x \rightarrow \infty} F_X(x) = 1 \qquad (21.11)$

Lower Limit $\lim_{x \rightarrow -\infty} F_X(x) = 0 \qquad (21.12)$

Definition 21.8 CDF of a discret rv X : Let X be discret rv with pdf \mathbb{P}_X , then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{t=-\infty}^x \mathbb{P}_X(t)$$

Definition 21.9 CDF of a continuous rv X : Let X be continuous rv with pdf f_X , then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t) \, dt \iff \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 21.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X , then it holds that:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) \qquad (21.13)$$

Proof. ^[def. 21.9]:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^x f_X(t) \, dt$$

□

Proof. lemma 21.1:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$$

or by the fundamental theorem of calculus (theorem 11.2):

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t) \, dt = \int_a^b \frac{\partial F_X(t)}{\partial t} \, dt = [F_X(t)]_a^b$$

□

Theorem 21.2 A continuous rv is fully characterized by its CDF: A cumulative distribution function completely determines the distribution of a continuous real-valued random variable.

1. Key figures

1.1. The Expectation

Definition 21.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x} \in \mathcal{X}} \bar{x} \mathbb{P}_x(\bar{x}) \qquad (21.14)$$

Definition 21.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{x} \in \mathcal{X}} \bar{x} f_x(\bar{x}) \, d\bar{x} \qquad (21.15)$$

Law 21.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \qquad (21.16)$$

Property 21.4 Translation and scaling: If $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$ are random vectors, and $a, b, c \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + bX + cY] = a + b\mathbb{E}[X] + c\mathbb{E}[Y] \qquad (21.17)$$

Thus \mathbb{E} is a linear operator ^[def. 11.13]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 21.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \qquad (21.18)$$

Property 21.5 Matrix×Expectation: If $X \in \mathbb{R}^n$ is a randomn vector and $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[AXB] = A\mathbb{E}[(XB)] = A\mathbb{E}[X]B \qquad (21.19)$$

Proof. eq. (21.24):

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{P}_{X,Y}(x, y) xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} \mathbb{P}_X(x) x \sum_{y \in \mathcal{Y}} \mathbb{P}_Y(y) y = \mathbb{E}[X] \mathbb{E}[Y] \end{aligned}$$

□

Definition 21.12 Autocorrelation/Crosscorrelation $\gamma(t_1, t_2)$: Describes the covariance ^[def. 21.16] between the two values of a stochastic process $(X_t)_{t \in T}$ at different time points t_1 and t_2 .

$$\gamma(t_1, t_2) = \text{Cov}[X_{t_1}, X_{t_2}] = \mathbb{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})] \qquad (21.20)$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \text{Cov}[X_t, X_t] \stackrel{\text{eq. (21.35)}}{=} \mathbb{V}[X_t] \qquad (21.21)$$

Notes

- Hence** the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- Given** a random time dependent variable $x(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $x(t - \tau)$ and the original function $x(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is non-zero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

Definition 21.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x} \in \mathcal{X}} \bar{x} p_x(\bar{x}) \quad (21.22)$$

Definition 21.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{x} \in \mathcal{X}} \bar{x} f_x(\bar{x}) d\bar{x} \quad (21.23)$$

Law 21.2 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \quad (21.24)$$

Property 21.6 Translation and scaling: If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, c \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \quad (21.25)$$

Thus \mathbb{E} is a **linear** operator^[def. 11.13].

Property 21.7 Affine Transformation of the Expectation:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mu + b \quad (21.26)$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 21.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \quad (21.27)$$

Property 21.8 Matrix×Expectation: If $\mathbf{X} \in \mathbb{R}^n$ is a random vector and $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B} \quad (21.28)$$

Proof. eq. (21.24):

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} p_X(x) x \sum_{y \in \mathcal{Y}} p_Y(y) y = \mathbb{E}[X] \mathbb{E}[Y] \end{aligned}$$

□

2.2. The Variance

Definition 21.15 Variance $\mathbb{V}[X]$: The variance of a random variable X is the expected value of the squared deviation from the expectation of X ($\mu = \mathbb{E}[X]$). It is a measure of how much the actual values of a random variable X fluctuate around its executed value $\mathbb{E}[X]$ and is defined by:

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] \stackrel{\text{see section 3}}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (21.29)$$

2.2.1. Properties

Property 21.9 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic \Rightarrow we have no uncertainty \Rightarrow no variance:

$$\mathbb{V}[a] = 0 \quad \text{with} \quad a \in \mathbb{R} \quad (21.30)$$

see shift and scaling for proof section 3

Property 21.10 Shifting and Scaling:

$$\mathbb{V}[a + bX] = a^2 \sigma^2 \quad \text{with} \quad a \in \mathbb{R} \quad (21.31)$$

see section 3

Property 21.11

Affine Transformation of the Variance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbb{V}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^\top \quad (21.32)$$

see section 3.

Definition 21.16 Covariance: The Covariance is a measure of how much two or more random variables vary **linearly** with each other.

$$\begin{aligned} \text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \quad (21.33)$$

see section 3

Definition 21.17 Covariance Matrix: The variance of a k -dimensional random vector $\mathbf{X} = (X_1 \dots X_k)$ is given by a p.s.d. eq. (14.15) matrix called Covariance Matrix.

The Covariance is a measure of how much two or more random variables vary **linearly** with each other and the Variance on the diagonal is again a measure of how much a variable varies:

$$\begin{aligned} \mathbb{V}[\mathbf{X}] &:= \Sigma(\mathbf{X}) := \text{Cov}[\mathbf{X}, \mathbf{X}] := \\ &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top \in [-\infty, \infty] \end{aligned} \quad (21.34)$$

$$\begin{aligned} &= \begin{bmatrix} \mathbb{V}[X_1] & \dots & \text{Cov}[X_1, X_k] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_k, X_1] & \dots & \mathbb{V}[X_k] \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \dots & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \dots & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)] \end{bmatrix} \end{aligned}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\text{Cov}[X, X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \quad (21.35)$$

[add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/](http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/)

Property 21.12 Translation and Scaling:

$$\text{Cov}(a + bX, c + dY) = bd\text{Cov}(X, Y) \quad (21.36)$$

Property 21.13

Affine Transformation of the Covariance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\text{Cov}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^\top = \mathbf{A}\Sigma(\mathbf{X})\mathbf{A}^\top \quad (21.37)$$

Definition 21.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned} \text{Corr}[\mathbf{X}] &:= \frac{\text{Cov}[\mathbf{X}]}{\sigma_{X_1} \dots \sigma_{X_k}} \in [-1, 1] \\ &= \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases} \end{aligned} \quad (21.38)$$

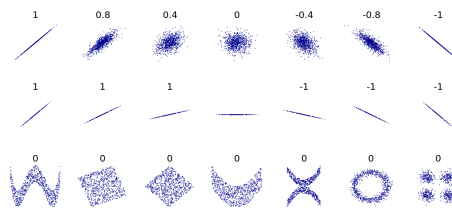


Figure 6: Several sets of (x, y) points, with their correlation coefficient

Law 21.3 Translation and Scaling:

$$\text{Corr}(a + bX, c + dY) = \text{sign}(b)\text{sign}(d)\text{Cov}(X, Y) \quad (21.39)$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 6), **but** not the slope of that relationship (middle row fig. 6) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 6 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0$ implies that there does not exist a **linear** relationship between the random variables X and Y .

Difference Covariance&Correlation

- Variance is affected by scaling and covariance not ?? and law 21.3.
- Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 21.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (??) is zero:

$$\begin{aligned} \text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &\stackrel{\text{eq. (21.24)}}{=} \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0 \end{aligned}$$

Zero covariance/correlation \nRightarrow independence

$$\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0 \nRightarrow p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

For example: let $X \sim \mathcal{U}([-1, 1])$ and let $Y = X^2$.

- Clearly X and Y are **dependent**
- But** the covariance/correlation between X and Y is non-zero:

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, X^2) = \mathbb{E}[X \cdot X^2] - \mathbb{E}[X]\mathbb{E}[X^2] \\ &= \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] \stackrel{\text{eq. (21.61)}}{=} 0 - 0 \cdot \mathbb{E}[X^2] \\ &\stackrel{\text{eq. (21.50)}}{=} 0 \end{aligned}$$
 \Rightarrow the relationship between Y and X must be non-linear.

Definition 21.19 Quantile: Are specific values q_α in the range^[def. 11.8] of a random variable X that are defined as the value for which the cumulative probability is less then $\alpha \in (0, 1)$:

$$q_\alpha : \mathbb{P}(X \leq x) = F_X(q_\alpha) = \alpha \xrightarrow{F \text{ invert.}} q_\alpha = F_X^{-1}(\alpha) \quad (21.40)$$

[add figures](#)

3. Proofs

Proof. eq. (21.29)

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &\stackrel{\text{Property 21.6}}{=} \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2 \end{aligned}$$

□

Proof. Property 21.10

$$\begin{aligned} \mathbb{V}[a + bX] &= \mathbb{E}[(a + bX - \mathbb{E}[a + bX])^2] \\ &= \mathbb{E}[(a + bX - a - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[(bX - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[b^2(X - \mathbb{E}[X])^2] \\ &= b^2\mathbb{E}[(X - \mathbb{E}[X])^2] = b^2\sigma^2 \end{aligned}$$

□

Proof. Property 21.11

$$\begin{aligned} \mathbb{V}(\mathbf{A}\mathbf{X} + b) &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^2] + 0 = \\ &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^\top] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}]))^\top] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{A}^\top] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \mathbf{A}^\top = \mathbf{A}\mathbb{V}[\mathbf{X}]\mathbf{A}^\top \end{aligned}$$

□

Proof. eq. (21.33)

$$\begin{aligned} \text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

□

Discrete Distributions

Definition 21.20 Multivariate Distribution: the variate refers to the number of input variables i.e. a m-variate distribution has m-input variables whereas a uni-variate distribution has only one.

Dimensional vs. Multivariate

The dimension refers to the number of dimensions we need to embed the function. If the variables of a function are independent than the dimension is the same as the number of inputs but the number of input variables can also be less.

4.1. Bernoulli Distribution Bern(p)

Definition 21.21 Bernoulli Trial: Is a random experiment with exactly two possible outcomes, success (1) and failure (0), in which the probability of success/failure is constant in every trial i.e. independent trials.

Definition 21.22 Bernoulli Distribution $X \sim \text{Bern}(\mathbf{p})$: X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter \mathbf{p} that signifies the success probability:

$$\mathbf{p}(x; \mathbf{p}) = \begin{cases} \mathbf{p} & \text{for } x = 1 \\ 1 - \mathbf{p} & \text{for } x = 0 \end{cases} \iff \begin{cases} \mathbb{P}(X = 1) = \mathbf{p} \\ \mathbb{P}(X = 0) = 1 - \mathbf{p} \end{cases}$$
$$= \mathbf{p}^x \cdot (1 - \mathbf{p})^{1-x} \quad \text{for } x \in \{0, 1\}$$

$$\mathbb{E}[X] = \mathbf{p} \quad (21.41) \quad \mathbb{V}[X] = \mathbf{p}(1 - \mathbf{p}) \quad (21.42)$$

4.2. Binomial Distribution B(n, p)

Definition 21.23 Binomial Distribution: Models the probability of exactly X success given a fixed number n -Bernoulli experiments^[def. 21.21], where the probability of success of a single experiment is given by \mathbf{p} :

$$\mathbf{p}(x) = \binom{n}{x} \mathbf{p}^x (1 - \mathbf{p})^{n-x} \quad \begin{array}{l} n : \text{nb. of repetitions} \\ x : \text{nb. of successes} \\ \mathbf{p} : \text{probability of success} \end{array}$$
$$\mathbb{E}[X] = n\mathbf{p} \quad (21.43) \quad \mathbb{V}[X] = n\mathbf{p}(1 - \mathbf{p}) \quad (21.44)$$

Note: Binomial Coefficient

The Binomial Coefficient corresponds to the permutation of two classes and not the variations as it seems from the formula.

Lets consider a box of n balls consisting of black and white balls. If we want to know the probability of drawing first x white and then $n - x$ black balls we can simply calculate:

$$\underbrace{(\mathbf{p} \cdots \mathbf{p})}_{x\text{-times}} \cdot \underbrace{(q \cdots q)}_{n-x\text{-times}} = \mathbf{p}^x q^{n-x}$$

But there exists obviously further realization $X = x$, that correspond to permutations of the n -drawn balls.

There exist two classes of $n_1 = x$ -white and $n_2 = (n - x)$ black balls s.t.

$$P(n; n_1, n_2) = \frac{n!}{x!(n - x)!} = \binom{n}{x}$$

4.3. Geometric Distribution Geom(p)

Definition 21.24 Geometric Distribution $\text{Geom}(\mathbf{p})$: Models the probability of the number X of Bernoulli trials^[def. 21.21] until the first success

$$\mathbf{p}(x) = \mathbf{p}(1 - \mathbf{p})^{x-1} \quad \begin{array}{l} x : \text{nb. of repetitions until first success} \\ \mathbf{p} : \text{success probability of single Bernoulli experiment} \end{array}$$

$$F(x) = \sum_{i=1}^x \mathbf{p}(1 - \mathbf{p})^{i-1} \stackrel{??}{=} 1 - (1 - \mathbf{p})^x$$

$$\mathbb{E}[X] = \frac{1}{\mathbf{p}} \quad (21.45) \quad \mathbb{V}[X] = \frac{1 - \mathbf{p}}{\mathbf{p}^2} \quad (21.46)$$

Notes

- $\mathbb{E}[X]$ is the mean waiting time until the first success
- the number of trials x in order to have at least one success with a probability of $\mathbf{p}(x)$:

$$x \geq \frac{\mathbf{p}(x)}{1 - \mathbf{p}}$$

- $\log(1 - \mathbf{p}) \approx -\mathbf{p}$ for small \mathbf{p}

4.4. Poisson Distribution Pois(λ)

Definition 21.25 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geq 0}$.

It expresses the probability of a given number of events X occurring in a fixed interval if those events occur independently of the time since the last event.

$$\mathbf{p}(x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad \begin{array}{l} \lambda > 0 \\ x \in \mathbb{Z}_{\geq 0} \end{array} \quad (21.47)$$

Event Rate λ : describes the average number of events in a single interval.

$$\mathbb{E}[X] = \lambda \quad (21.48) \quad \mathbb{V}[X] = \lambda \quad (21.49)$$

Continuous Distributions

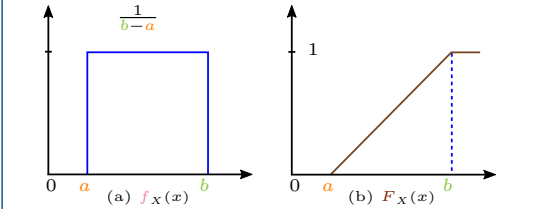
5.1. Uniform Distribution U(a, b)

Definition 21.26 Uniform Distribution $\mathcal{U}(a, b)$: Is probability distribution, where all intervals of the same length on the distribution's support^[def. 21.6] $\text{supp}(\mathcal{U}[a, b]) = [a, b]$ are equally probable/likely.

$$\mathbf{f}(x) = \frac{1}{b - a} \mathbb{1}_{x \in [a; b]} = \begin{cases} \frac{1}{b - a} = \text{const} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases} \quad (21.50)$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x - a}{b - a} & \text{if } a \leq x \leq b \\ 1 & x > b \end{cases} \quad (21.51)$$

$$\mathbb{E}[X] = \frac{a + b}{2} \quad \mathbb{V}(X) = \frac{(b - a)^2}{12} \quad (21.52)$$



5.2. Exponential Distribution exp(λ)

Definition 21.27 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution^[def. 21.24].

It describes the probability $\mathbf{f}(x; \lambda)$ that a continuous Poisson process (i.e., a process in which events occur continuously and independently at a constant average rate) will succeed/change state after a time interval x .

$$\mathbf{f}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & x < 0 \end{cases} \quad (21.53)$$

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & x < 0 \end{cases} \quad (21.54)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \mathbb{V}(X) = \frac{1}{\lambda^2} \quad (21.55)$$

5.3. Laplace Distribution

Definition 21.28 Laplace Distribution:

Laplace Distribution
$$\mathbf{f}(x; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right) \quad (21.56)$$

5.4. The Normal Distribution N(μ, σ)

Definition 21.29 Normal Distribution $X \sim \mathcal{N}(\mu, \sigma^2)$: Is a symmetric distribution where the population parameters μ, σ^2 are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X] = \mu \quad \mathbb{V}(X) = \sigma^2 \quad (21.57)$$

$$\mathbf{f}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\} \quad (21.58)$$

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right\} du \quad (21.59)$$

$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \quad (21.60)$$

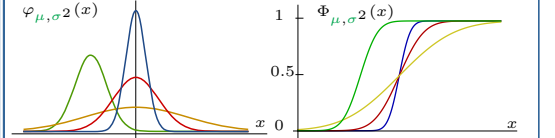


Figure 8: $\mu = 0, \sigma^2 = 0.2$ $\mu = 0, \sigma^2 = 1.0$ $\mu = 0, \sigma^2 = 5.0$ $\mu = -2, \sigma^2 = 0.5$

Property 21.14: $\mathbb{P}_X(\mu - \sigma \leq x \leq \mu + \sigma) = 0.66$

Property 21.15: $\mathbb{P}_X(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.95$

5.5. The Standard Normal distribution N(0, 1)

Historic Problem: the cumulative distribution eq. (21.59) does not have an analytical solution and numerical integration was not always computationally so easy. So how should people calculate the probability of x falling into certain ranges $\mathbb{P}(x \in [a, b])$?

Solution: use a standardized form/set of parameters (by convention) $\mathcal{N}_{0,1}$ and tabulate many different values for its cumulative distribution $\phi(x)$ s.t. we can transform all families of Normal Distributions into the standardized version $\mathcal{N}(\mu, \sigma^2) \xrightarrow{z} \mathcal{N}(0, 1)$ and look up the value in its table.

Definition 21.30

Standard Normal Distribution $X \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}[X] = 0 \quad \mathbb{V}(X) = 1 \quad (21.61)$$

$$\mathbf{f}(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (21.62)$$

$$F(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}u^2} du \quad (21.63)$$

$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty$

$$\psi_X(u) = e^{-\frac{u^2}{2}} \quad \varphi_X(u) = e^{-\frac{u^2}{2}} \quad (21.64)$$

Corollary 21.3

Standard Normal Distribution Notation: As the standard normal distribution is so commonly used people often use the letter Z in order to denote its the *standard* normal distribution and its α -quantile^[def. 21.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \quad \alpha \in (0, 1) \quad (21.65)$$

5.5.1. Calculating Probabilities

Property 21.16 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leq z) = \Phi(z) \quad (21.66)$$

$$\mathbb{P}(Z \leq -z) = \Phi(-z) = 1 - \Phi(z) \quad (21.67)$$

$$\mathbb{P}(-a \leq Z \leq b) = \Phi(b) - \Phi(-a) = \Phi(b) - (1 - \Phi(a))$$

$\stackrel{a=b=z}{=} 2\Phi(z) - 1 \quad (21.68)$

5.5.2. Linear Transformations of Normal Dist.

Proposition 21.1 Linear Transformation proof 1:
Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the linear transformed r.v. Y given by the affine transformation $Y = a + bX$ with $a \in \mathbb{R}, b \in \mathbb{R}_+$ follows:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \quad (21.69)$$

Proposition 21.2 Standardization: Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then there exists a linear transformation $Z = a + bX$ s.t. Z is a standard normally distributed random variable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow[Z \sim \mathcal{N}(0, 1)]{Z = \frac{X - \mu}{\sigma}} \quad (21.70)$$

section 1

Note

If we know how many standard deviations our distribution is away from our target value then we can characterize it fully by the standard normal distribution.

Proposition 21.3 Standardization of the CDF: Let $F_X(X)$ be the cumulative distribution function of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function $\Phi_Z(z)$ of the standardized random normal variable $Z \sim \mathcal{N}(0, 1)$ is related to $F_X(X)$ by:

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (21.71)$$

section 1

6. The Multivariate Normal distribution

Definition 21.31 Multivariate Normal/Gaussian:

An \mathbb{R}^n -valued random variable $\mathbf{X} = (X_1 \dots X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma : \mathcal{L}\left(\sum_{i=1}^n \alpha_i X_i\right) = \mathcal{N}(\mu, \sigma^2) \quad \forall \alpha_i \in \mathbb{R} \quad (21.72)$$

(possible degenerated $\mathcal{N}(0, 0)$ for $\forall \alpha_j = 0$)

Note

- Joint vs. multivariate:** a joint normal distribution can be a multivariate normal distribution or a product of univariate normal distributions **but**
- Multivariate refers to the number of variables that are placed as inputs to a function.

Definition 21.32

Multivariate Normal distribution $X \sim \mathcal{N}_k(\mu, \Sigma)$:

A k -dimensional random vector

$$\mathbf{X} = (X_1 \dots X_n)^\top \quad \text{with} \quad \mu = (\mathbb{E}[x_1] \dots \mathbb{E}[x_k])^\top$$

and $k \times k$ **p.s.d.** covariance matrix:

$$\Sigma := \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top] = [\text{Cov}[x_i, x_j], 1 \leq i, j \leq k]$$

follows a k -dim multivariate normal/Gaussian distribution if its law^[def. 19.23] satisfies:

$$f_{\mathbf{X}}(X_1, \dots, X_k) = \mathcal{N}(\mu, \Sigma) \quad (21.73)$$

$$= \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{X} - \mu)^\top \Sigma^{-1}(\mathbf{X} - \mu)\right)$$

Normalisation

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{i\mathbf{u}^\top \mu - \frac{1}{2}\mathbf{u}^\top \Sigma \mathbf{u}\right\} \quad (21.74)$$

6.1. Joint Gaussian Distributions

Definition 21.33 Jointly Gaussian Random Variables:

Two random variables X, Y both scalars or vectors, are said to be **jointly Gaussian** if the joint vector random variable $\mathbf{Z} = [X \ Y]^\top$ is again a GRV.

Property 21.17

Joint Independent Gaussian Random Variables: Let X_1, \dots, X_n be \mathbb{R} -valued independent random variables with laws $\mathcal{N}(\mu_i, \sigma_i^2)$. Then the law of $\mathbf{X} = (X_1 \dots X_n)$ is a (multivariate) Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \quad (21.75)$$

Corollary 21.4 Quadratic Form:

If \mathbf{x} and \mathbf{y} are both independent GRVs

$$\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x) \quad \mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$$

then they are jointly Gaussian^[def. 21.33] given by:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \quad (21.76)$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\mathbf{x} - \mu_x)^\top \Sigma_x^{-1}(\mathbf{x} - \mu_x) + (\mathbf{y} - \mu_y)^\top \Sigma_y^{-1}(\mathbf{y} - \mu_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}\left[(\mathbf{x} - \mu_x)^\top \quad (\mathbf{y} - \mu_y)^\top\right] \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_y^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mu_x \\ \mathbf{y} - \mu_y \end{bmatrix}\right)$$

$$\cong \exp\left(-\frac{1}{2}(\mathbf{z} - \mu_z)^\top \Sigma_z^{-1}(\mathbf{z} - \mu_z)\right)$$

Property 21.18

Marginal Distribution of Multivariate Gaussian: Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. The k -variate marginal distribution of the Gaussian indexed by a subset of the variables:

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad \mathbf{i}_j \in V \quad (21.77)$$

is given by:

$$\mathbf{X} = (X_{i_1} \dots X_{i_k})^\top \sim \mathcal{N}(\mu_A, \Sigma_{AA}) \quad (21.78)$$

$$\Sigma = \begin{bmatrix} \sigma_{i_1, i_1}^2 & \dots & \sigma_{i_1, i_k}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{i_k, i_1}^2 & \dots & \sigma_{i_k, i_k}^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

6.2. Conditional Gaussian Distributions

Property 21.19 Conditional Gaussian Distribution: Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. Suppose we take two disjoint subsets of V :

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad \mathbf{B} = \{j_1, \dots, j_m\} \quad \mathbf{i}_l, \mathbf{j}_{l'} \in V$$

then the conditional distribution of the random vector \mathbf{X}_A , conditioned on \mathbf{X}_B given by $p(\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B)$ is:

$$\mathbf{X}_A = (X_{i_1} \dots X_{i_k})^\top \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B}) \quad (21.79)$$

$$\begin{aligned} \mu_{A|B} &= \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1}(\mathbf{x}_B - \mu_B) \\ \Sigma_{A|B} &= \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA} \end{aligned}$$

Note

Can be proofed using the matrix inversion lemma but is a very tedious computation.

maybe add something

Corollary 21.5

Conditional Distribution of Joint Gaussian's: Let \mathbf{X} and \mathbf{Y} be jointly Gaussian random vectors:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix}\right) \quad (21.80)$$

then the *marginal* distribution of \mathbf{x} conditioned on \mathbf{y} can be written as:

$$X \sim \mathcal{N}(\mu_{X|Y}, \Sigma_{X|Y})$$

$$\begin{aligned} \mu_{X|Y} &= \mu_X + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mu_Y) \\ \Sigma_{X|Y} &= \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top \end{aligned} \quad (21.81)$$

add proofs

6.3. Transformations

Property 21.20 Multiples of Gaussian's

AX: Let $\mathbf{X} = (X_1 \dots X_n)^\top \sim \mathcal{N}(\mu, \Sigma)$ be an \mathbb{R}^n valued Gaussian and let $\mathbf{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \in \mathbb{R}^d \quad \mathbf{Y} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^\top) \quad (21.82)$$

Property 21.21 Affine Transformation of GRVs: Let $\mathbf{y} \in \mathbb{R}^n$ be GRV, $\mathbf{A} \in \mathbb{R}^{d \times n}$, $\mathbf{b} \in \mathbb{R}^d$ and let \mathbf{x} be defined by the affine transformation^[def. 14.1]:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b} \quad \mathbf{A} \in \mathbb{R}^{d \times n}, \mathbf{b} \in \mathbb{R}^d$$

Then \mathbf{x} is a GRV (see Section 1).

Property 21.22 Linear Combination of jointly GRVs: Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ two jointly GRVs, and let \mathbf{z} be defined as:

$$\mathbf{z} = \mathbf{A}_x \mathbf{x} + \mathbf{A}_y \mathbf{y} \quad \mathbf{A}_x \in \mathbb{R}^{d \times n}, \mathbf{A}_y \in \mathbb{R}^{d \times m}$$

Then \mathbf{z} is GRV (see Section 1).

Definition 21.34 Gaussian Noise: Is statistical noise having a probability density function (PDF) equal to that of the normal/Gaussian distribution.

6.4. Gamma Distribution

$$\Gamma(x, \alpha, \beta)$$

Definition 21.35 Gamma Distribution $X \sim \Gamma(x, \alpha, \beta)$: Is a widely used distribution that is related to the exponential distribution, Erlang distribution, and chi-squared distribution as well as Normal distribution:

$$f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & x \leq 0 \end{cases} \quad (21.83)$$

$$\Gamma(\alpha) \stackrel{\text{eq. (11.66)}}{=} \int_0^\infty t^{\alpha-1} e^{-t} dt \quad (21.84)$$

with

$$\alpha, \beta \in \mathbb{R}_{>0}$$

7. Student's t-distribution

Definition 21.36 Student' t-distribution:

add

7.1. Delta Distribution

Definition 21.37 The delta function $\delta(\mathbf{x})$:

The delta/dirac function $\delta(\mathbf{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$$

for any integrable function f on \mathbb{R} .

Or alternatively by:

$$\delta(\mathbf{x} - \mathbf{x}_0) = \lim_{\sigma \rightarrow 0} \mathcal{N}(\mathbf{x} | \mathbf{x}_0, \sigma) \quad (21.85)$$

$$\approx \infty \mathbb{1}_{\{\mathbf{x} = \mathbf{x}_0\}} \quad (21.86)$$

Property 21.23 Properties of δ :

- Normalization:** The delta function integrates to 1:

$$\int_{\mathbb{R}} \delta(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}} \delta(\mathbf{x}) \cdot c_1(\mathbf{x}) d\mathbf{x} = c_1(0) = 1$$

where $c_1(\mathbf{x}) = 1$ is the constant function of value 1.

- Shifting:**

$$\int_{\mathbb{R}} \delta(\mathbf{x} - \mathbf{x}_0) f(\mathbf{x}) d\mathbf{x} = f(\mathbf{x}_0) \quad (21.87)$$

- Symmetry:**

$$\int_{\mathbb{R}} \delta(-\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$$

- Scaling:**

$$\int_{\mathbb{R}} \delta(\alpha \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \frac{1}{|\alpha|} f(0)$$

Note

- In mathematical terms δ is not a function but a **generalized function**.
- We may regard $\delta(\mathbf{x} - \mathbf{x}_0)$ as a density with all its probability mass centered at the single point \mathbf{x}_0 .
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normal distribution eq. (21.85) would be a non-differentiable/discrete form of the dirac measure.

Proofs

Proof. proposition 21.1: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_Y(y) \stackrel{y \geq 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \leq \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

$$F_Y(y) \stackrel{y \leq 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \geq \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right)$$

Differentiating both expressions w.r.t. y leads to:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{b} \frac{dF_X\left(\frac{y-a}{b}\right)}{dy} \\ \frac{1}{-b} \frac{dF_X\left(\frac{y-a}{b}\right)}{dy} \end{cases} = \frac{1}{|b|} f_X(x) \left(\frac{y-a}{b}\right)$$

eq. (21.69)).

in order to prove that $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$ we simply plug f_X in the previous expression:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma}|b|} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\} = \frac{1}{\sqrt{2\pi\sigma}|b|} \exp\left\{-\frac{1}{2}\left(\frac{y - (a + b\mu)}{\sigma|b|}\right)^2\right\}$$

□

Proof. proposition 21.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{std}{std} X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

eq. (21.69) $\mathcal{N}(a\mu + b, a^2\sigma^2) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$

□

Proof. proposition 21.3: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leq x) \stackrel{-\mu}{=} \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

□

Proof. Property 21.21 scalar case

$$\text{Let } \mathbf{y} \sim p(\mathbf{y}) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \text{ and}$$

define $\mathbf{x} = a\mathbf{y} + \mathbf{b}$ $a \in \mathbb{R}_+, b \in \mathbb{R}$

Using the Change of variables formula it follows:

$$\begin{aligned} p_x(\bar{x}) &\stackrel{??}{=} \frac{p_y(\bar{y})}{\left|\frac{d\mathbf{x}}{d\mathbf{y}}\right|} \stackrel{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a} \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{\bar{x}-b}{a} - \mu\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi a^2 \mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2}(\bar{x} - b - a\mu)^2\right) \end{aligned}$$

$$\text{Hence} \quad \mathbf{x} \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$$

□

Note

We can also verify that we have calculated the right mean and variance by:

$$\begin{aligned} \mathbb{E}[x] &= \mathbb{E}[a\mathbf{y} + b] = a\mathbb{E}[y] + b = a\mu + b \\ \mathbb{V}[x] &= \mathbb{V}[a\mathbf{y} + b] = a^2\mathbb{V}[y] = a^2\sigma^2 \end{aligned}$$

Proof. ??

$$\begin{aligned}
 \mathbb{P}_{\mathbf{X}}(\mathbf{u}) &= \prod_i^n \mathbb{P}_{X_i}(u_i) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \\
 \varphi_{\mathbf{X}}(\mathbf{u}) &= \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_nu_n^2\right\} \\
 &= \exp\left\{i\sum_i^n u_n\mu_n - \frac{1}{2}\sum_i^n \sigma_nu_n^2\right\} = \exp\left\{i\mathbf{u}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{u} \boldsymbol{\Sigma} \mathbf{u}\right\}
 \end{aligned}$$

□

Proof. Property 21.22

From Property 21.21 it follows immediately that \mathbf{z} is GRV

$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ with:

$\mathbf{z} = \mathbf{A}\boldsymbol{\xi}$ with $\mathbf{A} = \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}$ and $\boldsymbol{\xi} = \begin{pmatrix} \mathbf{x} & \mathbf{y} \end{pmatrix}$

Knowing that \mathbf{z} is a GRV it is sufficient to calculate $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ in order to characterize its distribution:

$$\begin{aligned}
 \mathbb{E}[\mathbf{z}] &= \mathbb{E}[\mathbf{A}_x \mathbf{x} + \mathbf{A}_y \mathbf{y}] = \mathbf{A}_x \boldsymbol{\mu}_x + \mathbf{A}_y \boldsymbol{\mu}_y \\
 \mathbb{V}[\mathbf{z}] &= \mathbb{V}[\mathbf{A}\boldsymbol{\xi}] \stackrel{??}{=} \mathbf{A} \mathbb{V}[\boldsymbol{\xi}] \mathbf{A}^\top \\
 &= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[\mathbf{x}] & \text{Cov}[\mathbf{x}, \mathbf{y}] \\ \text{Cov}[\mathbf{y}, \mathbf{x}] & \mathbb{V}[\mathbf{y}] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}^\top \\
 &= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[\mathbf{x}] & \text{Cov}[\mathbf{x}, \mathbf{y}] \\ \text{Cov}[\mathbf{y}, \mathbf{x}] & \mathbb{V}[\mathbf{y}] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x^\top \\ \mathbf{A}_y^\top \end{bmatrix} \\
 &= \mathbf{A}_x \mathbb{V}[\mathbf{x}] \mathbf{A}_x^\top + \mathbf{A}_y \mathbb{V}[\mathbf{y}] \mathbf{A}_y^\top \\
 &\quad + \underbrace{\mathbf{A}_y \text{Cov}[\mathbf{y}, \mathbf{x}] \mathbf{A}_x^\top}_{=0 \text{by independence}} + \underbrace{\mathbf{A}_x \text{Cov}[\mathbf{x}, \mathbf{y}] \mathbf{A}_y^\top}_{=0 \text{by independence}} \\
 &= \mathbf{A}_x \boldsymbol{\Sigma}_x \mathbf{A}_x^\top + \mathbf{A}_y \boldsymbol{\Sigma}_y \mathbf{A}_y^\top
 \end{aligned}$$

□

Note

Can also be proofed by using the normal definition of [def. 21.15] and tedious computations.

8. Sampling Random Numbers

Most math libraries have uniform **random number generator (RNG)** i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a, b]$ (eq. (21.50)). Furthermore repeated calls to these RNG are independent, that is:

$$\begin{aligned} \mathbb{P}_{U_1, U_2}(u_1, u_2) &\stackrel{??}{=} \mathbb{P}_{U_1}(u_1) \cdot \mathbb{P}_{U_2}(u_2) \\ &= \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Question: using samples $\{u_1, \dots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discrete or continuous PDFs?

9. Inverse-transform Technique

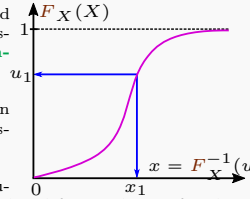
Idea

Can make use of section 1 and the fact that CDF are increasing functions (def. 11.10). **Advantage:**

- Simple to implement
- All discrete distributions can be generated via inverse-transform technique

Drawback:

- Not all continuous distributions can be integrated/have closed form solution for their CDF.
E.g. Normal, Gamma, Beta-distribution.



9.1. Continuous Case

Definition 21.38 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1, u_2, \dots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X :

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (21.88)$$

2. Set $F_X(X) \stackrel{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0, 1]$.
3. Invert this equation/find the inverse $F_X^{-1}(U)$ i.e. solve:

$$U = F_X(X) = F_X\left(\underbrace{F_X^{-1}(U)}_X\right) \quad (21.89)$$

4. Plug in the uniformly distributed rn:

$$x_i = F_X^{-1}(u_i) \quad \text{s.t.} \quad x_i \sim f_X \quad (21.90)$$

Definition 21.39 Multiple Continuous Variable:

Given: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (21.91)$$

2. Use [def. 21.40] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 21.40] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof. [def. 21.40]:

Claim: if U is a uniform rv on $[0, 1]$ then $F_X^{-1}(U)$ has F_X as its CDF.

Assume that F_X is strictly increasing (def. 11.10). Then for any $u \in [0, 1]$ there must exist a **unique** x s.t. $F_X(x) = u$.

Thus F_X must be invertible and we may write $x = F_X^{-1}(u)$.

Now let a arbitrary:

$$F_X(a) = \mathbb{P}(\underline{x} \leq a) = \mathbb{P}(F_X^{-1}(U) \leq a)$$

Since F_X is strictly increasing:

$$\begin{aligned} \mathbb{P}(F_X^{-1}(U) \leq a) &= \mathbb{P}(U \leq F_X(a)) \\ &\stackrel{\text{eq. (21.50)}}{=} \int_0^{F_X(a)} 1 dt = F_X(a) \end{aligned}$$

□

Note

Strictly speaking we may not assume that a CDF is **strictly** increasing but we as all CDFs are weakly increasing (def. 11.10) we may always define an auxiliary function by its infimum:

$$\hat{F}_X^{-1} := \inf \{x | F_X(x) \geq 0\} \quad u \in [0, 1] \quad (21.92)$$

9.2. Discret Case

Idea

Given: a desired $U \sim \mathcal{U}[0, 1]$ and uniformly distributed rn $\{u_1, u_2, \dots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:

$$\begin{aligned} k-1 < U \leq k &\iff F_X(x_{k-1}) < u \leq F_X(x_k) \\ \sum_{i=1}^{k-1} & \iff \end{aligned} \quad (21.93)$$

and return x_k .

Definition 21.40 One Discret Variable:

1. Compute the CDF of p_X (def. 21.8)

$$F_X(x) = \sum_{t=-\infty}^x p_X(t) \quad (21.94)$$

2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i ($\hat{=}$ inversion) s.t.:

$$F_X(x_{k(i)-1}) < u_i \leq F_X(x_{k(i)}) \quad \forall u_i \quad (21.95)$$

Proof. ??: First of all notice that we can always solve for an unique x_k .

Ask: why are Discret CRV always strictly increasing/unique?

Given a fixed x_k determine the values of u for which:

$$F_X(x_{k-1}) < u \leq F_X(x_k) \quad (21.96)$$

Now observe that:

$$\begin{aligned} u &\leq F_X(x_k) = F_X(x_{k-1}) + p_X(x_k) \\ &\Rightarrow F_X(x_{k-1}) < u \leq F_X(x_{k-1}) + p_X(x_k) \end{aligned}$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\begin{aligned} \mathbb{P}(U \in [F_X(x_{k-1}), F_X(x_k)]) &= \int_{F_X(x_{k-1})}^{F_X(x_k)} p_U(t) dt \\ &= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1 dt = F_X(x_k) - F_X(x_{k-1}) = p_X(x_k) \end{aligned}$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X . □

Definition 21.41

Multiple Continuous Variables (Option 1):

Given: a pdf of multiple rvs $p_{X,Y}$:

1. Use the product rule (??) in order to decompose $p_{X,Y}$:

$$p_{X,Y} = p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y) \quad (21.97)$$

2. Use ?? to first get a rv for y of $Y \sim p_Y(y)$.
3. Then with this fixed y use ?? again to get a value for x of $X \sim p_{X|Y}(x|y)$.

Definition 21.42

Multiple Continuous Variables (Option 2):

Note: this only works if \mathcal{X} and \mathcal{Y} are finite.

Given: a pdf of multiple rvs $p_{X,Y}$ let $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in \mathcal{X} and \mathcal{Y} .

Define $p_Z(1) = p_{X,Y}(1, 1), p_Z(2) = p_{X,Y}(1, 2), \dots, p_Z(N_x \cdot N_y) = p_{X,Y}(N_x, N_y)$

Then simply apply ?? to the auxillary pdf p_Z

1. Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (21.98)$$

2. Use [def. 21.40] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 21.40] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

See examples see comment in code text

10. Descriptive Statistics

10.1. Population Parameters

Definition 21.43 Population/Statistical Parameter: Are parameters defining families of probability distributions and thus characteristics of population following such distributions i.e. the normal distribution has two parameters $\{\mu, \sigma^2\}$

Definition 21.44 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \tag{21.99}$$

Definition 21.45 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \tag{21.100}$$

Note

The population variance and mean are equally to the mean derived from the true distribution of the population.

10.2. Sample Estimates

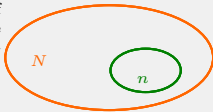
Definition 21.46 (Sample) Statistic: A statistic is a measurable function f that assigns a **single** value F to a sample of random variables or population:

$$f: \mathbb{R}^n \mapsto \mathbb{R} \qquad F = f(X_1, \dots, X_n)$$

E.g. F could be the mean, variance,...

Note

The function itself is independent of the sample's distribution; that is, the function can be stated before realization of the data.



Definition 21.47 (Point) Estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$:
Given: n-samples $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathbf{X}$ an estimator
 $\hat{\theta} = h(\mathbf{x}_1, \dots, \mathbf{x}_n)$ (21.101)

is a statistic/randomn variable used to estimate a true (population) parameter θ ^[def. 21.43].

Note

The other kind of estimators are interval estimators which do not calculate a statistic **but** an interval of plausible values of an unknown population parameter θ .

The most prevalent forms of interval estimation are:

- Confidence intervals (frequentist method).
- Credible intervals (Bayesian method).

Definition 21.48 Degrees of freedom of a Statistic: Is the number of values in the final calculation of a statistic that are free to vary.

10.2.1. Empirical Mean

Definition 21.49 Sample/Empirical Mean \bar{x} :
The sample mean is an estimate/statistic of the population mean^[def. 21.44] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i \tag{21.102}$$

Corollary 21.6 Expectation: The sample mean estimator is unbiased (see section 14):

$$\mathbb{E}[\hat{\mu}_X] = \mu \tag{21.103}$$

Corollary 21.7 Variance: For the variance of the sample mean estimator it holds (see section 14):

$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n} \sigma_X^2 \tag{21.104}$$

10.2.2. Empirical Variance

Definition 21.50 Biased Sample Variance: The sample mean is an estimate/statistic of the population variance^[def. 21.45] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \tag{21.105}$$

Definition 21.51 (Unbiased) Sample Variance:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \tag{21.106}$$

see section 14

Definition 21.52 Bessel's Correction: The factor $\frac{n}{n-1}$ (21.107)

as multiplying the uncorrected population varianceeq. (21.105) by this term yields an unbiased estimated of the variance (not the standard deviation). The reason for this is that are

Attention: Usually only unbiased variance is used and also sometimes denoted by s_n^2

Proof.



11. Statistical Tests

Definition 21.53 Null Hypothesis: A Null Hypothesis H_0 is usually a commonly accepted fact/view/base hypothesis that researchers try to nullify or disprove.

$$H_0 : \theta = \theta_0 \tag{21.108}$$

Definition 21.54 Alternative Hypothesis: The Alternative Hypothesis H_A/H_1 is the opposite of the Null Hypotheses/contradicts it and is what we try to test against the Null Hypothesis.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{21.109}$$

Definition 21.55 Testing Parameters:

Given: a parameter θ that we want to test.

Let Θ be the set of all possible values that θ can achieve.

We now split Θ in two disjunct sets Θ_0 and Θ_1 .

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \Theta_0 \cap \Theta_1 = \emptyset$$

$$\text{Null Hypothesis} \qquad H_0 : \theta \in \Theta_0 \tag{21.110}$$

$$\text{Alternative Hypothesis} \qquad H_A : \theta \in \Theta_1 \tag{21.111}$$

11.1. Type I&II Errors

Definition 21.56 Type I Error: Is the rejection of a Null Hypothesis, even-tough its true (also known as a "false positive").

Definition 21.57 Type II Error: Is the acceptance of a Null Hypothesis, even-tough its false (also known as a "false negative").

Decision	H_0 true	H_0 false	
Accept	TN	Type II (FN)	
Reject	Type I (FP)	TP	

Definition 21.58 Critical Value c: Value from which on the Null-hypothesis H_0 gets rejected.

Definition 21.59 Statistical significance α : A study's defined significance level, denoted α , is the **probability** of the study rejecting the null hypothesis, given that the null hypothesis were true (Type I Error).

Definition 21.60 Critical Region K_α : Is the set of all values that causes us to reject the Null Hypothesis in favor for the Alternative Hypothesis H_A .

The Critical region is usually chosen s.t. we incur a Type I Error with probability less than α .

$$K_\alpha \in \Theta : \mathbb{P}(\text{Type I Error}) \leq \alpha \tag{21.112}$$

$$\begin{aligned} &\mathbb{P}(c_2 \leq X \leq c_1) \leq \alpha && \text{two-sided} \\ \text{or } &\mathbb{P}(c_2 \leq X) \leq \frac{\alpha}{2} && \text{and } \mathbb{P}(X \leq c_1) \leq \frac{\alpha}{2} \\ &\mathbb{P}(c_2 \leq X) \leq \alpha && \text{one-sided} \\ &\mathbb{P}(X \leq c_1) \leq \alpha && \text{one-sided} \end{aligned}$$

Definition 21.61 Acceptance Region: Is the region where we accept the null hypothesis H_0 .

Note

see example 21.3.

11.2. Normally Distributed Data

Let us consider a sample of $\{x_i\}_{i=1}^n$ i.i.d. observations, that follow a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

11.2.1. Z-Test σ known

11.2.2. t-Test σ unknown

12. Inferential Statistics

Goal of Inference

- ① What is a good guess of the parameters of my model?
- ② How do I quantify my uncertainty in the guess?

13. Examples

Example 21.1 ??: Let x be uniformly distributed on $[0, 1]$ (def. 21.26) with pmf $p_X(x)$ then it follows:
 $\frac{dy}{dx} = \frac{1}{p_Y(y)} \Rightarrow dx = dy p_Y(y) \Rightarrow x = \int_{-\infty}^y p_Y(t) dt = F_Y(x)$

Example 21.2 ??: Let

add <https://www.youtube.com/watch?v=WUUbTVIRagg>

Example 21.3 Binomialtest:
Given: a manufacturer claims that a maximum of 10% of its delivered components are substandard goods.
In a sample of size $n = 20$ we find $x = 5$ goods that do not fulfill the standard and are skeptical that the what the manufacture claims is true, so we want to test:
 $H_0 : p = p_0 = 0.1$ vs. $H_A : p > 0.1$

We model the number of number of defective goods using the binomial distribution^(def. 21.23)

$$X \sim \mathcal{B}(n, p), n = 20 \quad \mathbb{P}(X \geq x) = \sum_{k=x}^n \binom{n}{k} p^k (1-p)^{n-k}$$

from this we find:
 $\mathbb{P}_{p_0}(X \geq 4) = 1 - \mathbb{P}_{p_0}(X \leq 3) = 0.13$
 $\mathbb{P}_{p_0}(X \geq 4) = 1 - \mathbb{P}_{p_0}(X \leq 3) = 0.04 \leq \alpha$
thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
 \Rightarrow throw away null hypothesis for the 5% niveau in favor to the alternative.
 \Rightarrow the 5% significance niveau is given by $K = \{5, 6, \dots, 20\}$

Note

If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x-1)$

14. Proofs

Proof. corollary 21.6:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\underbrace{\mu + \dots + \mu}_{1, \dots, n}\right]$$

Proof. corollary 21.7:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \stackrel{\text{Property 21.10}}{=} \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n x_i\right]$$
$$\frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n} \sigma^2$$

Proof. definition 21.51:

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_X^2] &= \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2)\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2n\bar{x} \cdot n\bar{x} + n\bar{x}^2\right] \\ &= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n \mathbb{E}[x_i^2] - n\mathbb{E}[\bar{x}^2]\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\mathbb{E}[\bar{x}^2]\right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\ &= \frac{1}{n-1} \left[(n\sigma^2 + n\mu^2) - (\sigma^2 + n\mu^2)\right] \\ &= \frac{1}{n-1} [n\sigma^2 - \sigma^2] = \frac{1}{n-1} [(n-1)\sigma^2] = \sigma^2 \end{aligned}$$

□

Stochastic Calculus

Stochastic Processes

Definition 22.1 Random/Stochastic Process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$: is a collection of random variables on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The index set \mathcal{T} is usually representing time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \dots\}$. Therefore, the random process X can be written as a function:

$$X : \mathbb{R} \times \Omega \mapsto \mathbb{R} \iff (t, \omega) \mapsto X(t, \omega) \quad (22.1)$$

Definition 22.2 Sample path/Trajectory/Realization: Is the *stochastic/noise signal* $r(\cdot, \omega)$ on the index set \mathcal{T} , that we obtain be sampling ω from Ω .

Notation
Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$

Definition 22.3 Filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$: A collection $\{\mathcal{F}_t\}_{t \geq 0}$ of sub σ -algebras $\{\mathcal{F}_t\}_{t \geq 0} \in \mathcal{F}$ is called filtration if is *increasing*:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \quad \forall s \leq t \quad (22.2)$$

Definition 22.4 Adapted Process: A stochastic process $\{X_t : 0 \leq t \leq \infty\}$ is called adapted to a filtration \mathbb{F} if, X_t is $\{\mathcal{F}_{t-1}\}$ -measurable, i.e. the value of X_t is known at time $t - 1$.

Definition 22.5 Predictable Process: A stochastic process $\{X_t : 0 \leq t \leq \infty\}$ is called predictable w.r.t. a filtration \mathbb{F} if, X_t is $\{\mathcal{F}_{t-1}\}$ -measurable, i.e. the value of X_t is known at time $t - 1$.

Note
The price of a stock will usually be adapted since date k prices are known at date k .
On the other hand the interest rate of a bank account is usually already known at the beginning $k - 1$, s.t. the interest rate r_t ought to be \mathcal{F}_{k-1} measurable, i.e. the process $r = (r_k)_{k=1, \dots, T}$ should be predictable.

Definition 22.6 Filtered Probability Space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$: A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ is called a *filtered probability space*.

Corollary 22.1 : The amount of information of an adapted random process is increasing see example 22.1.

Definition 22.7 Martingales: A stochastic process $X(t)$ is a martingale on a *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ if the following conditions hold:

① Given $s \leq t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value:
$$\forall s \leq t \quad \mathbb{E}[X(t) | \mathcal{F}_s] = X(s) \quad \text{a.s.} \quad (22.3)$$

② The expectation is finite:
$$\mathbb{E}[|X(t)|] < \infty \quad \forall t \geq 0 \quad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geq 0} \text{ adapted} \quad (22.4)$$

Interpretation

- For any \mathcal{F}_s -adapted process the best prediction of $X(t)$ is the currently known value $X(s)$ i.e. if $\mathcal{F}_s = \mathcal{F}_{t-1}$ then the best prediction is $X(t - 1)$
- A martingale models fair games of limited information.

Definition 22.8 Auto Covariance $\gamma(t_2 - t_1)$: Describes the covariance^[def. 21.16] between two values of a stochastic process $(X_t)_{t \in \mathcal{T}}$ at different time points t_1 and t_2 .

$$\gamma(t_1, t_2) = \text{Cov}[X_{t_1}, X_{t_2}] = \mathbb{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})] \quad (22.5)$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \text{Cov}[X_t, X_t] \stackrel{\text{eq. (21.35)}}{=} \mathbb{V}[X_t] \quad (22.6)$$

Notes

- Hence the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- Given a random time dependent variable $\mathbf{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $\mathbf{x}(t - \tau)$ and the original function $\mathbf{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random, then the autocorrelation is non-zero.
- The auto covariance is maximized/most similar for no translation $\tau = 0$ at all.

Definition 22.9 Auto Correlation $\rho(t_2 - t_1)$: Is the scaled version of the auto-covariance^[def. 22.8]:

$$\begin{aligned} \rho(t_2 - t_1) &= \text{Corr}[X_{t_1}, X_{t_2}] \\ &= \frac{\text{Cov}[X_{t_1}, X_{t_2}]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} \end{aligned} \quad (22.7)$$

1. Different kinds of Processes

1.1. Markov Process

Definition 22.10 Markov Process: A continuous-time stochastic process $X(t), t \in T$, is called a Markov process if for any finite parameter set $\{t_i : t_i < t_{i+1}\} \in T$ it holds:

$$\mathbb{P}(X(t_{n+1}) \in B | X(t_1), \dots, X(t_n)) = \mathbb{P}(X(t_{n+1}) \in B | X(t_n))$$

it thus follows for the *transition probability* – the probability of $X(t)$ lying in the set B at time t , given the value x of the process at time s :

$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \quad 0 \leq s < t \quad (22.8)$$

Interpretation
In order to predict the future only the current/last value counts.

Corollary 22.2 Transition Density: The transition probability of a continuous distribution \mathbf{p} can be calculated via:

$$\mathbb{P}(s, x, t, B) = \int_B \mathbf{p}(s, x, t, y) dy \quad (22.9)$$

1.2. Gaussian Process

Definition 22.11 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:

$$X(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)) \quad \forall t \in T \quad (22.10)$$

1.3. Diffusions

Definition 22.12 Diffusion: Is a Markov Process^[def. 22.10] for which it holds that:

$$\begin{aligned} \mu(t, X(t)) &= \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[X(t + \Delta t) - X(t) | X(t)] \quad (22.11) \\ \sigma^2(t, X(t)) &= \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[(X(t + \Delta t) - X(t))^2 | X(t)] \quad (22.12) \end{aligned}$$

See ??/eq. (22.12) for simple proof of eq. (22.11)/??.

- $\mu(t, X(t))$ is called **drift**
- $\sigma^2(t, X(t))$ is called **diffusion coefficient**

Interpretation
There exist not discontinuities for the trajectories.

1.4. Brownian Motion/Wiener Process

Definition 22.13 d-dim standard Brownian Motion/Wiener Process: Is an \mathbb{R}^d valued *stochastic process*^[def. 22.1] $(W_t)_{t \in \mathcal{T}}$ starting at $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies:

- ① **Normal Independent Increments:** the increments are *normally distributed independent random variables*:

$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, (t_i - t_{i-1}) \mathbb{1}_{d \times d}) \quad \forall i \in \{1, \dots, T\} \quad (22.13)$$
- ② **Stationary increments:** $W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$
- ③ **Continuity:** for a.e. $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous

$$\lim_{t \rightarrow 0} \frac{\mathbb{P}(|W(t + \Delta t) - W(t)| \geq \delta)}{\Delta t} = 0 \quad \forall \delta > 0 \quad (22.14)$$
- ④ **Start**

$$W(0) := W_0 = 0 \quad \text{a.s.} \quad (22.15)$$

Notation

- In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wiener process.
- However in some sources the Wiener process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 22.3 $W_t \sim \mathcal{N}(0, \sigma)$: The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law

$$\begin{aligned} \mathbb{E}[W(t)] &= \mu = 0 \quad (22.16) \\ \mathbb{V}[W(t)] &= \mathbb{E}[W^2(t)] = \sigma^2 = t \quad (22.17) \end{aligned}$$

See section 5

1.4.1. Properties of the Wiener Process

Property 22.1 Non-Differentiable Trajectories: The sample paths of a Brownian motion are not differentiable:

$$\begin{aligned} \frac{dW(t)}{dt} &= \lim_{t \rightarrow 0} \mathbb{E} \left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t} \right)^2 \right] \\ &= \lim_{t \rightarrow 0} \frac{\mathbb{E}[W(t + \Delta t) - W(t)]}{\Delta t} = \lim_{t \rightarrow 0} \frac{\sigma^2}{\Delta t} = \infty \end{aligned}$$

$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
 $\xrightarrow{\text{solution}}$ Ito Calculus see section 23.

Property 22.2 Auto covariance Function: The auto-covariance^[def. 22.8] for a Wiener process

$$\mathbb{E}[(W(t) - \mu(t))(W(t') - \mu(t'))] = \min(t, t') \quad (22.18)$$

Property 22.3: A standard Brownian motion is a **Quadratic Variation**

Definition 22.14 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$LV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_{\Pi}-1} |f(x_{i+1}) - f(x_i)| \quad (22.19)$$

$$\mathcal{S} = \{\Pi\{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } \stackrel{[\text{def. 17.8}]}{\text{of}} [a, b]\}$$

it is a measure of the (one dimensional) length of a function w.r.t. to the y-axis, when moving alone the function. Hence it is a measure of the variation of a function w.r.t. to the y-axis.

Definition 22.15 Total Quadratic Variation/“sum of squares”: The total quadratic variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$QV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_{\Pi}-1} |f(x_{i+1}) - f(x_i)|^2 \quad (22.20)$$

$$\mathcal{S} = \{\Pi\{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } \stackrel{[\text{def. 17.8}]}{\text{of}} [a, b]\}$$

Corollary 22.4 Bounded (quadratic) Variation: The (quadratic) variation^[def. 22.14] of a function is bounded if it is finite:

$$\exists M \in \mathbb{R}_+ : LV_{[a,b]}(f) \leq M \iff (QV_{[a,b]}(f) \leq M) \quad \forall \Pi \in \mathcal{S} \quad (22.21)$$

Theorem 22.1 Variation of Wiener Process: Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:

$$\mathbb{P}(\omega : LV(W(\omega)) < \infty) = 0 \quad (22.22)$$

Theorem 22.2 Quadratic Variation of standard Brownian Motion: The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:

$$\lim_{N \rightarrow \infty} \sum_{k=1}^N \left[W\left(k \frac{T}{N}\right) - W\left((k-1) \frac{T}{N}\right) \right]^2 = T$$

with probability 1

$$(22.23)$$

See ??

Corollary 22.5 : theorem 22.2 can also be written as:

$$(dW(t))^2 = dt \quad (22.24)$$

1.4.2. Lévy’s Characterization of BM

Theorem 22.3 d-dim standard BM/Wiener Process by Paul Lévy: An \mathbb{R}^d valued *adapted stochastic process*^[def’x. 22.1, 22.3] $(W_t)_{t \in \mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}_+}$, that satisfies:

- ① **Start**

$$W(0) := W_0 = 0 \quad \text{a.s.} \quad (22.25)$$
- ② **Continuous Martingale:** W_t is an a.s. *continuous martingale*^[def. 22.7] w.r.t. the filtration $(\mathcal{F}_t)_{t \in \mathcal{T}}$ under \mathbb{P} .
- ③ **Quadratic Variation:** $W_t^2 - t$ is also an martingale $\iff QV(W_t) = t$

$$(22.26)$$

is a standard Brownian motion^[def. 22.20]. Proof see section 5

Further Stochastic Processes

1.4.3. White Noise

understand script and add

Definition 22.16 Discrete-time white noise: Is a random signal $\{\epsilon_t\}_{t \in T_{\text{discret}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * [k]] = 0 \quad \forall k \in T_{\text{discret}} \quad (22.27)$$
- Zero autocorrelation^[def. 22.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * [k], \epsilon * [k + n]) = \mathbb{E}[\epsilon * [k] \epsilon * [k + n]^T] = \mathbb{V}[\epsilon * [k]] \delta_{\text{discret}}[n] \quad \forall k, n \in T_{\text{discret}} \quad (22.28)$$

$$\text{With} \quad \delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$$

See proofs

Definition 22.17 Continuous-time white noise: Is a random signal $(\epsilon_t)_{t \in T_{\text{continuous}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * (t)] = 0 \quad \forall t \in T_{\text{continuous}} \quad (22.29)$$
- Zero autocorrelation^[def. 22.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * (t), \epsilon * (t + \tau)) = \mathbb{E}[\epsilon * (t) \epsilon * (t + \tau)^T] \stackrel{\text{eq. (21.86)}}{=} \mathbb{V}[\epsilon * (t)] \delta(t - \tau) = \begin{cases} \mathbb{V}[\epsilon * (t)] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases} \quad (22.30)$$

$$\forall t, \tau \in T_{\text{continuous}} \quad (22.31)$$

Definition 22.18 Homoscedastic Noise: Is constant for all observations/time-steps:

$$\forall [\epsilon_t] = \sigma^2 \quad \forall t = 1, \dots, T \quad (22.32)$$

Definition 22.19 Heteroscedastic Noise: Is noise that can vary with each observation/time-step:

$$\forall [\epsilon_t] = \sigma(t)^2 \quad \forall t = 1, \dots, T \quad (22.33)$$

1.4.4. Generalized Brownian Motion

Definition 22.20 Brownian Motion: Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 22.13], and define:

$$X_t = \mu t + \sigma W_t \quad t \in \mathbb{R}_+ \quad \begin{array}{l} \mu \in \mathbb{R} : \text{drift parameter} \\ \sigma \in \mathbb{R}_+ : \text{scale parameter} \end{array} \quad (22.34)$$

then $\{X_t\}_{t \in \mathbb{R}_+}$ is normally distributed with mean μt and variance $t\sigma^2$ $X_t \sim \mathcal{N}(\mu t, \sigma^2 t)$.

Theorem 22.4 Normally Distributed Increments: If $W(T)$ is a Brownian motion, then $W(t) - W(0)$ is a normal random variable with mean μt and variance $\sigma^2 t$, where $\mu, \sigma \in \mathbb{R}$. From this it follows that $W(t)$ is distributed as:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x - \mu t)^2}{2\sigma^2 t}\right\} \quad (22.35)$$

Corollary 22.6 : More generally we may define the process: $t \mapsto f(t) + \sigma W_t$ (22.36) which corresponds to a noisy version of f .

Corollary 22.7 Brownian Motion as a Solution of an SDE: A stochastic process X_t follows a BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} dX(t) &= \mu dt + \sigma dW(t) & (22.37) \\ X(0) &= 0 & (22.38) \end{aligned}$$

1.4.5. Geometric Brownian Motion (GBM)

For many processes $X(t)$ it holds that:

- there exists an (exponential) growth
- that the values may not be negative $X(t) \in \mathbb{R}_+$

Definition 22.21 Geometric Brownian Motion: Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 22.13] the exponential transform:

$$X(t) = \exp(W(t)) = \exp(\mu t + \sigma W(t)) \quad t \in \mathbb{R}_+ \quad (22.39)$$

is called geometric Brownian motion

Corollary 22.8 Log-normal Returns: For a geometric BM we obtain log-normal returns:

$$\ln\left(\frac{S_t}{S_0}\right) = \mu t + \sigma W(t) \iff \mu t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \quad (22.40)$$

meaning that the mean and the variance of the process (stock) log-returns grow over time linearly.

Corollary 22.9 Geometric BM as a Solution of an SDE: A stochastic process X_t follows a geometric BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} dX(t) &= X(t) (\mu dt + \sigma dW(t)) & (22.41) \\ &= \mu X(t) dt + \sigma X(t) dW(t) & (22.42) \\ X(0) &= 0 \end{aligned}$$

1.4.6. Locally Brownian Motion

Definition 22.22 Locally Brownian Motion: Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 22.13] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = \mu(X(t), t) dt + \sigma(X(t), t) dW(t) \quad (22.43)$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

1.4.7. Ornstein-Uhlenbeck Process

Definition 22.23 Ornstein-Uhlenbeck Process: Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 22.13] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = -aX(t) dt + b\sigma dW(t) \quad a > 0 \quad (22.44)$$

1.5. Poisson Processes

Definition 22.24 Rare/Extreme Events: Are events that lead to discontinuous in stochastic processes.

Problem

A Brownian motion is not sufficient as model in order to describe extreme events s.a. crashes in financial market time series. Need a model that can describe such discontinuities/jumps.

Definition 22.25 Poisson Process: A Poisson Process with rate $\lambda \in \mathbb{R}_{\geq 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$, having a discrete state space $N = \{0, 1, 2, \dots\}$ and satisfies:

1. $X_0 = 0$
2. The increments follow a Poisson distribution^[def. 21.25]:

$$\mathbb{P}((X_t - X_s) = k) = \frac{\lambda(t-s)}{k!} e^{-\lambda(t-s)} \quad 0 \leq s < t < \infty \quad \forall k \in \mathbb{N}$$
3. No correlation of (non-overlapping) increments:
 $\forall t_0 < t_1 < \dots < t_n$: the increments are independent
 $X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}} \quad (22.45)$

Interpretation

A Poisson Process is a continuous-time process with discrete, positive realizations in $\mathbb{N}_{\geq 0}$

Corollary 22.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}(X_{(t+\Delta t)} - X_t \neq 0) = \lambda \Delta t + o(\Delta t^2) \quad t \text{ small i.e. } t \rightarrow 0 \quad (22.46)$$

1. Thus the probability of an event happening during Δt is proportional to time period and the rate λ
2. The probability of two or more events to happen during Δt is of order $o(\Delta t^2)$ and thus extremely small (as Δt is small).

Definition 22.26 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$dX_t = \lim_{\Delta t \rightarrow dt} (X_{(t+\Delta t)} - X_t) \quad (22.47)$$

Property 22.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\begin{aligned} \mathbb{P}(dX_t = 0) &= 1 - \lambda & (22.48) \\ \mathbb{P}(|dX_t| = 1) &= \lambda & (22.49) \end{aligned}$$

Proofs

Proof. eq. (22.11):

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)] \\ &\stackrel{\text{induction}}{=} \mathbb{E}[x_{n-1}] = \dots = \mathbb{E}[x(0)] = 0 \end{aligned}$$

Thus in expectation the particles goes nowhere. \square

Proof. eq. (22.12):

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)^2] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)^2\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta]^2 \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2] \\ &\stackrel{\text{ind.}}{=} \mathbb{E}[x_{n-1}^2] + \delta^2 = \mathbb{E}[x_{n-2}^2] + 2\delta^2 = \dots \\ &= \mathbb{E}[x(0)^2] + n\delta^2 = n\delta^2 \end{aligned}$$

as $n = \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x}$ it follows:

$$\sigma^2 = \mathbb{E}[x^2(n)] - \mathbb{E}[x(n)]^2 = \mathbb{E}[x^2(n)] = \frac{\delta^2}{\Delta x} t \quad (22.50)$$

Thus in expectation the particles goes nowhere. \square

Proof. eq. (22.30):

$$\begin{aligned} \gamma(\epsilon * [k], \epsilon * [k+n]) &= \text{Cov}[\epsilon * [k], \epsilon * [k+1]] \\ &= \mathbb{E}[(\epsilon * [k] - \mathbb{E}[\epsilon * [k]]) (\epsilon * [k+n] - \mathbb{E}[\epsilon * [k+n]])] \\ &\stackrel{\text{eq. (22.27)}}{=} \mathbb{E}[(\epsilon * [k]) (\epsilon * [k+n])] \end{aligned} \quad \square$$

Proof. corollary 22.3:

Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the increment over the interval $[s-s, t-s] = [0, t-s]$

$$\begin{aligned} \text{Thus} \quad B_t - B_s &\sim B_{t-s} - B_0 \\ \text{but as } B_0 &\text{ is a.s. zero by definition eq. (22.15) it follows:} \\ B_t - B_s &\sim B_{t-s} \quad B_{t-s} \sim \mathcal{N}(0, t-s) \end{aligned} \quad \square$$

Proof. corollary 22.3:

$$\begin{aligned} W(t) &= W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t) \\ \Rightarrow \quad \mathbb{E}[X] &= 0 \quad \mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = t \end{aligned} \quad \square$$

Proof. theorem 22.2:

$$\begin{aligned} \sum_{k=0}^{N-1} [W(t_k) - W(t_{k-1})]^2 & \quad t_k = k \frac{T}{N} \\ &= \sum_{k=0}^{N-1} X_k^2 \quad X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right) \\ &= \sum_{k=0}^{N-1} Y_k = n \left(\frac{1}{n} \sum_{k=0}^{N-1} Y_k\right) \quad \mathbb{E}[Y_k] = \frac{T}{N} \\ &\stackrel{\text{S.L.L.N.}}{=} n \frac{T}{n} = T \end{aligned} \quad \square$$

Proof. theorem 22.3 ②:

1. first we need to show eq. (22.3): $\mathbb{E}[W_t | \mathcal{F}_s] = W_s$
Due to the fact that W_t is \mathcal{F}_t measurable i.e. $W_t \in \mathcal{F}_t$ we know that:

$$\begin{aligned} \mathbb{E}[W_t | \mathcal{F}_t] &= W_t & (22.51) \\ \mathbb{E}[W_t | \mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s | \mathcal{F}] \\ &= \mathbb{E}[W_t - W_s | \mathcal{F}_s] + \mathbb{E}[W_s | \mathcal{F}_s] \\ &\stackrel{\text{eq. (22.51)}}{=} \mathbb{E}[W_t - W_s] + W_s \\ &\stackrel{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s \end{aligned}$$

2. second we need to show eq. (22.4): $\mathbb{E}[|X(t)|] < \infty$

$$\mathbb{E}[|W(t)|]^2 \stackrel{??}{\leq} \mathbb{E}[|W(t)|^2] = \mathbb{E}[W^2(t)] = t < \infty$$

Proof. theorem 22.3 ③: $W_t^2 - t$ is a martingale?
Using the binomial formula we can write and adding $W_s - W_s$:

$$W_t^2 = (W_t - W_s)^2 + 2W_s(W_t - W_s) + W_s^2$$

using the expectation:

$$\begin{aligned} \mathbb{E}[W_t^2 | \mathcal{F}_s] &= \mathbb{E}[(W_t - W_s)^2 | \mathcal{F}_s] + \mathbb{E}[2W_s(W_t - W_s) | \mathcal{F}_s] \\ &\quad + \mathbb{E}[W_s^2 | \mathcal{F}_s] \\ &\stackrel{\text{eq. (22.51)}}{=} \mathbb{E}[(W_t - W_s)^2] + 2W_s \mathbb{E}[(W_t - W_s)] + W_s^2 \\ &\stackrel{\text{eq. (22.17)}}{=} \mathbb{V}[W_t - W_s] + 0 + W_s^2 \\ &\quad t - s + W_s^2 \end{aligned}$$

from this it follows that:

$$\mathbb{E}[W_t^2 - t | \mathcal{F}_s] = W_s^2 - s \quad \square$$

understand why $\mathbb{E}[(W_t - W_s)^2 | \mathcal{F}] = \mathbb{E}[(W_t - W_s)^2]$

Examples

Example 22.1 :

Suppose we have a sample space of four elements: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. At time zero, we do not have any information about which ω has been chosen. At time $T/2$ we know whether we have $\{\omega_1, \omega_2\}$ or $\{\omega_3, \omega_4\}$. At time T , we have full information.

$$\mathcal{F} = \begin{cases} \{\emptyset, \Omega\} & t \in [0, T/2) \\ \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^\Omega & t = T \end{cases} \quad (22.52)$$

Thus, \mathcal{F}_0 represents initial information whereas \mathcal{F}_∞ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$.

Ito Calculus