# Probabilistic Artificial Intelligence

## Active Learning

Here we are interested in choosing the next input point $\mathbf{x}$ that some expert should label $y$. **Goal**: we want to choose the observations that provides us with the biggest gain of information/reduction in uncertainty.

**Definition 1.1 Active Learning:** Is to actively choose the most information samples in order to reduce the amount of samples we need to label.

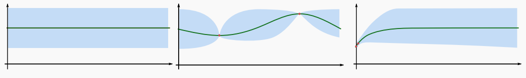
**Definition 1.2 Utility Function F:**
Is a function that provide a ranking to judge uncertain situations.

## 1. Uncertainty Sampling for Regression

### 1.1. Maximizing the Information Gain

Let $f$ be a unkown function that we can evaluate with $D = \texttt{dom}(f)$. Let $S$ be a subset of points $S \subseteq D$ that we can choose make noisy observations $y_S$ of $f$ in order to maximze the *information gain*[def. 4.1]: [example 1.1]

$$\boxed{F(S) := H(f) - H(f|y_s) \overset{\text{eq. (4.17)}}{=} I(f; y_s)} \tag{1.1}$$



(a) Prior (b) High Infogain (c) Low Infogain

**Definition 1.3 Optimal Set of labels:**

$$\left\{x_1, \dots, x_{|S|}\right\} = \underset{S\subseteq D, |S|\leqslant T}{\arg\max} \quad F(S) \tag{1.2}$$

**Problem**: $F(S)$ is NP-hard to optimize.
**Idea**: optimize greedily only the next point.

**Definition 1.4** [proof 1.1]
**Greedy Mutual Information Maximization Objective:**
Only consider the next point that maximizes the mutual information and not all at once:

$$\begin{aligned} x_{t+1} &= \underset{x\in D}{\arg\max}\, F\left(S_t \cup \{x\}\right) \\ &= \underset{x\in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \end{aligned} \tag{1.3}$$

**Corollary 1.1** [proof 1.2]
**Homoscedactic Gaussian:**

$$x_t = \underset{x\in D}{\arg\max}\, \sigma^2_{t-1}(x) \tag{1.4}$$

this can then be maximized.
Let $A_t = \{x_1, \dots, x_t\}$ then it follows:

$$\sigma_t^2(x) = \mathbf{k}(x,x) - \mathbf{k}_{x,A_t}\left(\mathcal{K}_{A_t,A_t} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{x,A_t} \tag{1.5}$$

**Algorithm 1.1 Greedy Uncertainty Sampling:**

```
    Given: $S_t := \{x_1, \dots, x_t\}$
1:  for $t+1 \dots, T$ do
        $x_{t+1} = \underset{x\in D}{\arg\max}\, F\left(S_t \cup \{x\}\right)$
        $= \underset{x\in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f)$
2:  end for
```

**Corollary 1.2 Diminishing Returns Property:**
Mutal information satisifies modular submodularity (Property 4.9)
⇒ adding a label/memasurement for some data point can only increase information:

$$\begin{aligned} F\left(A \cup \{x\}\right) - F\left(A\right) &\geqslant F\left(B \cup \{X\}\right) - F(B) \\ H\left(y_x|y_A\right) - H\left(y_X|f\right) &\geqslant H\left(y_x|y_B\right) - H\left(y_X|f\right) \\ \iff \qquad H\left(y_x|y_A\right) &\geqslant H\left(y_x|y_A\right) \end{aligned}$$

**Note**

For Gaussians processes the utitlity $F$ does only the depend on the set of observations we require but not on the actual observations/labels. This is because the entropy for Guassian depends only on the covariance matrix and not the actual measurments.

**Corollary 1.3 Constant Factor Approximation:** algorithm 1.1 provides a constant factor approximation of eq. (1.1):

$$F(S_T) \leqslant \underbrace{\left(1 - \frac{1}{\mathrm{e}}\right)}_{\approx .63} \underset{S\subseteq D, |S|\leqslant T}{\max} \quad F(S) \tag{1.6}$$

**Note**

There exist other objectives then entropy reduction/mutual information in order to quantify uncertainty but they are usually more expesnive but may offer other advantages.

### 1.2. Heteroscedastic Case

So far we considered homoscedastic noise[def. 35.22] but sometimes we may have heteroscedasctic[def. 35.23] noise $\sigma_n(x) \iff$ different locations may have different noise i.e. to different sensors.
**Problem**: in the heteroscedas case the most uncertain outcomes are no longer necessarily the most informative.

**Corollary 1.4** [proof 1.3]
**Heteroscedastic Gaussian:**

$$x_t = \underset{x\in D}{\arg\max} \frac{\text{epistemic uncertainty}}{\text{aleatoric uncertainty}} = \underset{x\in D}{\arg\max} \frac{\sigma_f^2(x)}{\sigma_n^2(x)} \tag{1.7}$$

this can then be maximized.
Let $A_t = \{x_1, \dots, x_t\}$ then it follows:

$$\sigma_t^2(x) = \mathbf{k}(x,x) - \mathbf{k}_{x,A_t}\left(\mathcal{K}_{A_t,A_t} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{x,A_t} \tag{1.8}$$

## 2. Uncertainty Sampling for Classification

we now want to choose get/obtain labels for those samples that we are most unusre/uncertain about.
⇒ maximize the entropy in order to select the next label.

**Definition 1.5 Greedy Mutual Entropy Maximization:** select the next point that maximizes the entropy over the label distribution:

$$\begin{aligned} x_{t+1} &= \underset{x\in D}{\arg\max}\, H(Y|x, x_{1:t}, y_{1:t}) = \underset{x\in D}{\arg\max} && (1.9) \\ &= \underset{x\in D}{\arg\max} -\mathrm{p}(y|x, x_{1:t}, y_{1:t}) && (1.10) \\ &= \underset{x\in D}{\arg\max} -\sum_y \mathrm{p}(y|x, x_{1:t}, y_{1:t}) && (1.11) \end{aligned}$$

**Notes**

The posterior $\mathrm{p}(y|x, x_{1:t}, y_{1:t})$ is usually intractable but we can using approximate inference section 9 methods:

- Approximate Inference section 1
- Markov Chain Monte Carlos section 2

### 2.1. Heteroscedastic Case

So far we considered homoscedastic noise[def. 35.22] but sometimes we may have heteroscedasctic[def. 35.23] noise $\sigma_n(x) \iff$ different locations may have different noise i.e. to different sensors.
**Problem**: in the heteroscedas case the most uncertain labels are no longer necessarily the most informative.

#### 2.1.1. Informative Sampling for Classification

**Definition 1.6** [proof 1.4]
**Bayesian active learning by disagreement (BALD):**

$$\begin{aligned} x_{t+1} &= \underset{\hat{x}\in D}{\arg\max}\, I\left(\theta; \hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right) \qquad (1.12) \\ &= \underset{\hat{x}\in D}{\arg\max}\, H\left(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right) - \mathbb{E}_{\theta\sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H\left(\hat{y}, \hat{x}, \theta\right)\right] \end{aligned}$$

**Explanation 1.1.**

① $H\left(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right)$:
*is the entropy of the predictive posterior distribution*[def. 6.14]*, approximate using approximate inference section 9.*

② $\mathbb{E}_{\theta\sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H\left(\hat{y}, \hat{x}, \theta\right)\right]$:
*is the conditional Entropy over the labels by drawing $\theta$ from the posterior distribution and averagin over them.*

## 3. Examples

**Example 1.1 Gaussian Information Gain:**

$$F(S) \overset{\text{example 4.8}}{=} \frac{1}{2}\log\left|\mathbf{I} + \sigma^{-2}\mathcal{K}_S\right|$$

## 4. Proofs

**Proof 1.1.** [def. 1.4]

$$\begin{aligned} x_{t+1} &= \underset{x\in D}{\arg\max}\, F\left(S_t \cup \{x\}\right) \overset{eq.\ (22.58)}{=} \underset{x\in D}{\arg\max}\, F\left(S_t \cup \{x\}\right) - F\left(S_t\right) \\ &= \underset{x\in D}{\arg\max}\, I(f; y_{S_t+x}) - I(f; y_{S_t}) \\ &= \underset{x\in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{eq.\ (4.17)}{=} \underset{x\in D}{\arg\max}\, H\left(y_{S_t+x}\right) - H\left(y_{S_t+x}|f\right) - H\left(y_{S_t}\right) + H\left(y_{S_t}|f\right) \\ &= \underset{x\in D}{\arg\max}\, \underline{H\left(y_{S_t}, x\right)} - H\left(y_{S_t+x}|f\right) \underline{-H\left(y_{S_t}\right)} + H\left(y_{S_t}|f\right) \\ &\overset{\underline{eq.\ (4.7)}}{=} \underset{x\in D}{\arg\max}\, \underline{H\left(y_x|y_{S_t}\right)} - \underline{H\left(y_{S_t+x}|f\right)} + H\left(y_{S_t}|f\right) \\ &\overset{\clubsuit}{=} H(y_x|y_{S_t}) - H(y_x|f) \end{aligned}$$

**Note ♣**

$$\begin{aligned} \underline{H\left(y_{S_t} \cup x|f\right)} &\overset{\text{eq. (4.7)}}{=} H\left(y_{s_t}|f\right) + H\left(y_x|f, y_{S_t}\right) \\ &= H\left(y_{s_t}|f\right) + H\left(y_x|f\right) \end{aligned}$$

**Proof 1.2.** [cor. 1.1]

$$\begin{aligned} y &= f(x) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned} \quad \Rightarrow \quad \mathrm{p}(y|x, f) = \mathcal{N}\left(f(x), \sigma_n^2\right)$$

$$\begin{aligned} x_{t+1} &= \underset{x\in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{eq.\ (4.24)}{=} \underset{x\in D}{\arg\max}\, \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2 x|S_t - \frac{1}{2}\ln(2\pi\mathrm{e})\sigma_n^2 \\ &\overset{eq.\ (22.58)}{=} \sigma^2_{x|S_t} \end{aligned}$$

*Thus if we define* $\sigma^2_{t-1}(x) = \sigma^2_{x|x_{1:t-1}}$ *it follows:*

$$x_t = \underset{x\in D}{\arg\max}\, \sigma^2_{t-1}(x) \tag{1.13}$$

**Proof 1.3.** [cor. 1.4]

$$\begin{aligned} y &= f(x) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned} \quad \Rightarrow \quad \mathrm{p}(y|x, f) = \mathcal{N}\left(f(x), \sigma_n^2(x)\right)$$

$$\begin{aligned} x_{t+1} &= \underset{x\in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{eq.\ (4.24)}{=} \underset{x\in D}{\arg\max}\, \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2 x|S_t - \frac{1}{2}\ln(2\pi\mathrm{e})\sigma_n^2(x) \\ &\overset{eq.\ (22.58)}{=} \underset{x\in D}{\arg\max}\, \ln\frac{\sigma_f^2(x)}{\sigma_n^2(x)} \overset{eq.\ (22.70)}{=} \underset{x\in D}{\arg\max}\, \frac{\sigma_f^2(x)}{\sigma_n^2(x)} \end{aligned}$$

**Proof 1.4.** [def. 1.6]

$$\begin{aligned} I\left(\theta; \hat{y}|x_{1:t}, y_{1:t}\right) &\overset{eq.\ (4.17)}{=} H\left(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right) - H\left(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t}\right) \\ &\overset{eq.\ (4.6)}{=} H\left(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right) \\ &\quad - \mathbb{E}_{\hat{y}|\theta\sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[\log \mathrm{p}_{\hat{y}|\theta}\left(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t}\right)\right] \\ &\overset{eq.\ (4.2)}{=} H\left(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}\right) \\ &\quad - \mathbb{E}_{\hat{y}|\theta\sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H\left(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t}\right)\right] \end{aligned}$$

# Bayesian Optimizaton

In section 1 we tried to maximize our information gain about an unknown function $f$.
While While sequentially optimizing eqs. (1.3) and (1.4) is a provably good way to explore $f$ globally, it is not well suited for function value optimization, where we only care about maximizing our knowledge about the maxima.

**Given**

- set of possible inputs $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
- unknown (black-box) function/oracle:
$$f \in \mathcal{F} \qquad\qquad f : D \mapsto \mathbb{R} \tag{2.1}$$
that is expensive but from which we can draw noisy observations:
$$y_t = f(\mathbf{x}_t) + \epsilon \tag{2.2}$$

**Goal**

Adaptively choose inputs $\mathbf{x}_1, \dots, \mathbf{x}_T \in D$ that maximize the performance/function/sum of rewards:

$$\sum_{t=1}^{T} f(\mathbf{x}_t) \tag{2.3}$$

⇒ need a measure of performance i.e. cumulative regret [def. 2.3] as we can only draw point samples from $f$.

**Definition 2.1**
**Action Set** $\mathcal{A} = \{a_1, \dots, a_n\}$:
Is the set of possible actions from which we can choose at each step.

**Corollary 2.1 :** If we want to maximize a function $f$, then its just the set of possible inputs $\mathcal{A} = D$

**Definition 2.2**
**Optimizing Agent/ Decision Making Policy:**
Is a policy on how to choose an action $a \in \mathcal{A}$ based on a objective/utility function [def. 1.2]

**Definition 2.3 (Cumulative) Regret for a fixed $f$:** Is defined as the the cumulative loss we suffer in comparison to taking the optimal value $\mathbf{x}^*$ if we had full knowlede of $f$.

$$\begin{aligned} R_T := \sum_{t=1}^{T} \left( \max_{\mathbf{x}\in D} f(\mathbf{x}) - f(\mathbf{x}_t) \right) = \sum_{t=1}^{T} r_t \\ = T \max_{\mathbf{x}\in D} f(\mathbf{x}) - \sum_{t=1}^{T} f(\mathbf{x}_t) \end{aligned} \tag{2.4}$$

$r_t$: instantaneous regret

**Definition 2.4 (Time) Average Regret:**

$$\frac{R_T}{T} = \frac{1}{T}\sum_{t=1}^{T} r_t = \frac{1}{T}\sum_{t=1}^{T} \left( f(\mathbf{x}^*) - f(\mathbf{x}_t) \right) \tag{2.5}$$

**Definition 2.5**
**No/Sublinear Regret Algorithms:**

$$\lim_{T\to\infty} \underbrace{\frac{R_T}{T}}_{\text{Average regret}} = 0 \qquad \begin{matrix} R_T = o(T) \\ \dfrac{R_T}{T} = o(1) \end{matrix} \qquad \forall \text{sequences } 1, \dots, T \tag{2.6}$$

**Explanation 2.1.** *Due to more information the instantaneous regret decreases over time and we obtain no regret in average.*

**Definition 2.6**
**Pure Exploration/Follow the Leader Policy:** Take the action with the current maximum empirical mean payoff.

**Algorithm 2.1 Epsilon Greedy Algorithm:**

```
Set: $\epsilon_t = \mathcal{O}\left(\frac{1}{t}\right)$
for $t = 1, \dots, T$ do
    With probability $\epsilon_t$ explore unif. at randomn:
        $a_{t+1} = \mathcal{U}(a_1, \dots |\mathcal{A}|)$    (2.7)
    With probability $1 - \epsilon_t$ take action with highest known empirical mean payoff:
        $a_{t+1} = \arg\max_{a\in\mathcal{A}} \hat{\mu}_{a,T} \qquad \hat{\mu}_{a,T} = \frac{1}{n_{a,T}} \sum_{s=1}^{T} \mathbb{1}_{\{a_s = a\}} v_{a,s}$    (2.8)
end for
```

**Problem**

This policy is a first good try but can easily get stuck at local optima. A better way would be not to sample randomly but take into account the uncertainty.

## 1. Optimistic Bayesian Optimization

**Problem**

Picking the nex point greedily by maximizing the mean payoff [def. 2.6]

$$x_t = \arg\max_{x\in D} \mu_{t-1}(x) \tag{2.9}$$

of the posterior distribution tends to lead to local optima.

**Assumption**

If the true function $f$ is within the confidence bounds of our posterior distribution:

$$f(x) \in (\mu(x) - \beta\sigma, \mu(x) + \beta\sigma)$$

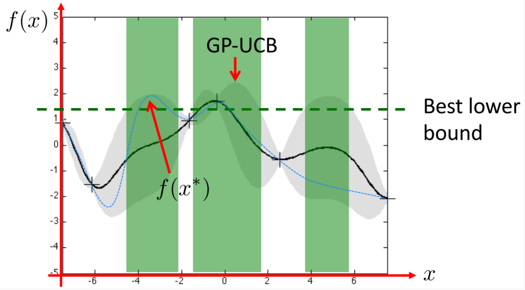
then it follows that:

$$f\left(\mathbf{x}^*\right) \geqslant \max \mu(x) - \beta\sigma(x) \tag{2.10}$$

this implies that we should focus only on certain regions. Because if the best predicted value of a point $\mu(x') + \sigma(x')$ is less then the *best lower confidence bound*eq. (2.10) then the maximum cannot be at $x'$:

$$f\left(\mathbf{x}^*\right) \geqslant \max \mu(x) - \beta\sigma(x) \geqslant \mu(x') + \sigma(x') \tag{2.11}$$

add own figure



This idea can be utilized in various ways:

- GP-UCB section 1
- Thompson Sampling section 2

### 1.1. Gaussian Process-UCB

**Principle 2.1 Optimization in the phase of uncertainty:** Pick the action that has the highest upper confidence bound (UCP).

**Explanation 2.2** (principle 2.1). *We do not pick the action that maximizes our current estimate $\mu(x)$ but the most optimistic one.*
*If the guess is wrong optimism will fade quickly but if the guess is right we will maximize our utility will decreasing uncertainty.*

**Definition 2.7 GP-UCB:**

$$\mathbf{x}_t = \arg\max_{x\in D} \mu_{t-1}(\mathbf{x}) + \beta_t \sigma_{t-1}(\mathbf{x}) \tag{2.12}$$

**Explanation 2.3** (Definition 2.7).

- *$\beta_t \to \infty$ recover uncertainty sampling*
- *$\beta_t = 0$ recover greedy algorithm*

#### 1.1.1. Maximizing the UCB

The GP-UCB [def. 2.7] is usually a non-convex function. Thus in order to maximize this objective we need to use:

- *Lipschitz Optimization* (in low dimension)
- Use gradient descent based on multiple random initialization (in high dimension)

#### 1.1.2. Guarantees on the regret

**Theorem 2.1 Bayesian Regret of GP-UCB:** assuming the true function $f$ follows a Gaussian Process $f \sim \mathcal{GP}$ then it holds that for a suitable choice of $\beta_t$ (needs to slowly decay with const$\cdot \log t$):

$$\frac{1}{T}\sum_{t=1}^{T} \left[ f(\mathbf{x}^*) - f(\mathbf{x}_t) \right] = \mathcal{O}\left(\frac{\gamma_T}{T}\right) \qquad T: \; \#\text{of samples}$$

with $\qquad \gamma_T = \max_{|S|\leqslant T} I(f; y_S)$

**Explanation 2.4** ($\gamma_T$). *The regret depends on how much information we can gain in $T$ steps.*

**Corollary 2.2 Linear Kernel:**
For a linear kernel [def. 11.9] it holds:

$$\gamma_T = \mathcal{O}(d \log T) \tag{2.13}$$

**Corollary 2.3 Squared Exponential Kernel:**
For a squared exponential kernel [def. 11.13] it holds:

$$\gamma_T = \mathcal{O}\left((\log T)^{d+1}\right) \tag{2.14}$$

**Corollary 2.4 Matern Kernel $\nu > 2$:**
For a linear kernel [def. 11.14] it holds:

$$\gamma_T = \mathcal{O}\left(T^{\frac{d(d+1)}{2\nu + d(d+1)}}\right) \tag{2.15}$$

**Note: Reproducing Kernel Hilbert Space (RKHS)**

There exists also a frequentists regret of GP-UCB which only assumes that $f$ is part of a hilbert space and overinflates the confidence bounds in order to obtain good estimates.

### 1.2. Thompson Sampling

**Definition 2.8 Thompson Sampling:** Draw a function $\tilde{f}$ from the posterior and maximize it:

$$\tilde{f} \sim p(f|\mathbf{x}_{1:n}, \mathbf{y}_{1:t}) \qquad \mathbf{x}_{t+1} \in \arg\max_{\mathbf{x}\in D} \tilde{f}(\mathbf{x}) \tag{2.16}$$

**Explanation 2.5** (Definition 2.8). *The randomness in $\tilde{f}$ helps to trade of exploration vs. exploitation.*

# Machine Learning Appendix

## Model Assessment and Selection

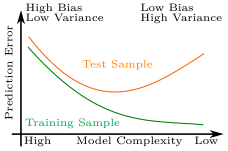
**Definition 3.1 Statistical Inference:** Is the process of deducing properties of an underlying probability distribution by mere analysis of data.

**Definition 3.2**
**Model Selection:**
Is the process of selecting a model $f$ from a given or chosen class of models $\mathcal{F}$

**Definition 3.3 Hyperparameter Tuning:** Is the process of choosing the hyperparameters $\theta$ of a given model $f \in \mathcal{F}$

**Definition 3.4 Model Assessment/Evaluation:** Is the process of evaluating the performance of a model.

**Definition 3.5 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
⇒ big test error.

**Definition 3.6 Labeled Data** $\mathcal{D}/\mathcal{Z}$:

$$\mathcal{D} := \left\{ (\mathbf{x}_j, \mathbf{y}_j) \;\middle|\; \mathbf{x}_j \in \mathcal{X}, \mathbf{y}_j \in \mathcal{Y} \right\}$$

**Definition 3.7 Training Set $\mathcal{Z}^{\text{train}} \subset \mathcal{Z}$:** 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 3.8 Validation Set $\mathcal{Z}^{\text{val}} \subset \mathcal{Z}$:** Is the part of the data that is used in order to evaluate different hyperparamters.

**Definition 3.9 Test Set $\mathcal{Z}^{\text{test}} \subset \mathcal{Z}$:** 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:

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

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

$$\mathcal{R}(f) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(\mathbf{x}, y) l(y, f(\mathbf{x}))\, dy\, dx$$

**Problem**: we do not know $f(\mathbf{x}, y)$ ⇒ can only estimate the empricial 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 $\hat{R}(f)$, 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

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

### 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 $\theta^*$ by using the validation set for evaluation:
$$\underline{\hat{\theta}\left(\mathcal{Z}^{val}\right)} \in \arg\min_{\theta : \hat{f}_\theta \in \mathcal{F}_\theta} \hat{R}\left(\hat{f}_\theta\left(\mathcal{Z}^{tr}\right), \mathcal{Z}^{val}\right)$$
3. Use the tests set in order to test the model:
$$\hat{\mathcal{R}}\left(\hat{f}_{\underline{\hat{\theta}(\mathcal{Z}^{val})}}\left(\mathcal{Z}^{tr}\right), \mathcal{Z}^{test}\right)$$

**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 ⇒ split the data into a test and training and validation set.

## 2. Cross Validation

**Definition 3.10 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 3.11 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 4.1 Information** (***Claude Elwood Shannon***)**:**
Information is the resolution of uncertainty.

**Amount of Information**

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

$$I\left(\mathrm{p}_0 \cdot \mathrm{p}_1 \cdots \mathrm{p}_n\right) = I(\mathrm{p}_0) + I(\mathrm{p}_1) + \cdots + I(\mathrm{p}_n)$$

⇒ can use the logarithm to satisfy this

**Definition 4.2 Surprise/Self-Information/-Content:** Is a measure of the information of a realization $x$ of a random variable $X \sim \mathrm{p}$:

$$I_X(x) = \log\left(\frac{1}{\mathrm{p}(X = x)}\right) = -\log \mathrm{p}(X = x) \tag{4.1}$$

**Explanation 4.1** (Definition 4.2). $I(A)$ *measures the number of possibilities for an event* $A$ *to occur in bits:*

$$I(A) = \log_2(\#\textit{possibilities for } A \textit{ to happen})$$

**Corollary 4.1 Units of the Shannon Entropy:**
The Shannon entropy can be defined for different logarithms

≙ units:

| log | units |
|---|---|
| Base 2 | Bits/Shannons |
| Natural | Nats |
| Base 10 | Dits/Bans |

**Explanation 4.2.** *An uncertain event is much more informative than an expected/certain event:*

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

### 1.2. Entropy

Information content deals with a single event. If we want to quantify the amount of uncertainty/information of a probability distribution, we need to take the expectation over the information content[def. 4.2]:

**Definition 4.3 Shannon Entropy** example 4.3**:**
Is the expected amount of information of a random variable $X \sim \mathrm{p}$:

$$H(\mathrm{p}) = \mathbb{E}_X[I_X(x)] = \mathbb{E}_X\left[\log \frac{1}{\mathrm{p}_X(x)}\right] = -\mathbb{E}_X[\log \mathrm{p}_X(x)]$$
$$= -\sum_{i=1}^{n} \mathrm{p}(x_i) \log \mathrm{p}(x_i) \tag{4.2}$$

**Definition 4.4 Differential/Continuous entropy:** Is the continuous version of the Shannon entropy[def. 4.3]:

$$H(\mathrm{p}) = \int_{x\sim\mathrm{p}} -f(x) \log f(x)\,\mathrm{d}x \tag{4.3}$$

**Notes**

- The Shannon entropy is maximized for uniform distributions
- People somtimes write $H(X)$ instead of $H(\mathrm{p})$ with the understanding that $\mathrm{p}$ is the distribution of $\mathrm{p}$.

**Property 4.1 Non negativity:**
Entropy is always non-negative:

$$H(X) \geqslant 0 \qquad \text{if } X \text{ is deterministic} \qquad H(X) = 0 \tag{4.4}$$

### 1.2.1. Conditonal Entropy

**Proposition 4.1 Conditioned Entropy** $H(Y|X = x)$**:**
Let $X$ and $Y$ be two random variables with a condititional pdf $\mathrm{p}_{X|Y}$. The entropy of $Y$ conditioned on $X$ taking a certain value $x$ is given as:

$$H(Y|X = x) = \mathbb{E}_{Y|X=x}\left[\log \frac{1}{\mathrm{p}_{y|X}(Y|X = x)}\right]$$
$$= -\mathbb{E}_{Y|X=x}\left[\log \mathrm{p}_{Y|X}(y|X = x)\right] \tag{4.5}$$

**Definition 4.5** proof 4.3
**Conditional Entropy** $H(Y|X)$**:**
Is the amount of information need to determine $Y$ if we arleady know $X$ and is given by averagin $H(Y|X = x)$ over $X$.

$$H(Y|X) = [\mathbb{E}_X H(Y|X = x)] = -\mathbb{E}_{X,Y}\left[\log \frac{\mathrm{p}(x, y)}{\mathrm{p}(x)}\right] \tag{4.6}$$
$$= \mathbb{E}_{X,Y}\left[\log \frac{\mathrm{p}(x)}{\mathrm{p}(x, y)}\right]$$

**Definition 4.6** proof 4.4
**Chain Rule for Entropy:**

$$H(Y|X) = H(X, Y) - H(X)$$
$$H(X|Y) = H(X, Y) - H(Y) \tag{4.7}$$

**Property 4.2 Monotonicity:**
Information/conditioning reduces the entropy
⇒ *Information never hurts.*

$$H(X|Y) \geqslant H(X) \tag{4.8}$$

**Corollary 4.2 From eq. (4.17):**

$$H(X, Y) \leqslant H(X) + H(Y) \tag{4.9}$$

### 1.3. Cross Entropy

**Definition 4.7 Cross Entropy** proof 4.1**:**
Lets say a model follows a true distribution $X \sim \mathrm{p}$ but we model $X$ as with a different distribution $X \sim q$. The cross entropy between $\mathrm{p}$ and $q$ measure the average amount of information/bits needed to model an outcome $x \sim \mathrm{p}$ with $X$:

$$H(\mathrm{p}, q) = \mathbb{E}_{x\sim\mathrm{p}}\left[\log\left(\frac{1}{q(x)}\right)\right] \tag{4.10}$$
$$= -\mathbb{E}_{x\sim\mathrm{p}}[\log q(x)] \tag{4.11}$$
$$= H(\mathrm{p}) + D_{\mathrm{KL}}(\mathrm{p} \parallel q) \tag{4.12}$$

**Corollary 4.3 Kullback-Leibler Divergence:** $D_{\mathrm{KL}}(\mathrm{p} \parallel q)$ measures the extra price (bits) we need to pay for using $q$.

### 1.4. Kullback-Leibler (KL) divergence

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

**Definition 4.8**
**Kullback–Leibler divergence.** examples 4.4 and 4.7
**/Relative Entropy from** $\mathrm{p}$ **to** $q$**:** Given two probability distributions $\mathrm{p}$, $q$ of a random variable $X$. The Kullback–Leibler divergence is defined to be:

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) = \mathbb{E}_{x\sim\mathrm{p}}\left[\log \frac{\mathrm{p}(x)}{q(x)}\right] = \mathbb{E}_{x\sim\mathbb{P}}[\log \mathrm{p}(x) - \log q(x)] \tag{4.13}$$

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

**Explanation 4.3.**

- $\mathrm{p}$ *decides where we put the mass if* $\mathrm{p}(x)$ *is zero we do not care about* $q(x)$.
- $\mathrm{p}(x)/q(x)$ *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 4.3 Non-Symmetric:**

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) \neq D_{\mathrm{KL}}(q \parallel \mathrm{p}) \qquad \forall \mathrm{p}, q \tag{4.14}$$

**Property 4.4:**

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) \geqslant 0 \tag{4.15}$$
$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) = 0 \quad \Longleftrightarrow \quad \mathrm{p}(x) = q(x) \forall x \in \mathcal{X} \tag{4.16}$$

**Note**

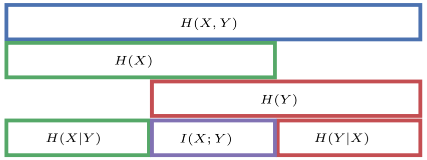
The KL-divergence is not a real distance measure as $\mathrm{KL}(\mathbb{P} \parallel Q) \neq \mathrm{KL}(Q \parallel \mathbb{P})$

**Corollary 4.4 Lower Bound on the Cross Entropy:** The entropy provides a lower bound on the cross entropy, which follows directly eq. (4.16). from

### 1.5. Mutual Information

**Definition 4.9** **example 4.8**
**Mutual Information/Information Gain:** Let $X$ and $Y$ be two random variables with a joint probability distribution. The mutal information of $X$ and $Y$ is the reduction in uncertainty in $X$ if we know $Y$ and vice versa.

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.17}$$
$$= H(X) + H(Y) - H(X, Y)$$
$$= D_{\mathrm{KL}}\left(\mathrm{p}_{X,Y} \parallel \mathrm{p}_X \mathrm{p}_Y\right)$$



**Explanation 4.4** (Definition 4.9).

$$I(X; Y) = \begin{cases} \textit{big} & \textit{if } X \textit{ and } Y \textit{ are highly dependent} \\ 0 & \textit{if } X \textit{ and } Y \textit{ are independent} \end{cases} \tag{4.18}$$

**Property 4.5 Symmetry:**

$$I(X; Y) = I(Y, X)$$

**Property 4.6 Positiveness:**

$$I(X; Y) \geqslant 0 \qquad \text{if } X \perp\!\!\!\perp Y \qquad I(X; Y) = 0 \tag{4.19}$$

**Property 4.7:**

$$I(X; Y) \leqslant H(X) \qquad I(X; Y) \leqslant H(Y) \tag{4.20}$$

**Property 4.8 Self-Information:**

$$H(X) = I(X; X)$$

**Property 4.9 Montone Submodularity:** Mutual information is monotone submodular[def. 19.12]:

$$H(X, z) - H(x) \geqslant H(Y, z) - H(Y) \tag{4.21}$$
$$\overset{\text{[def. 4.6]}}{\Longleftrightarrow} \qquad H(z|X) \geqslant H(x|Y) \tag{4.22}$$

## 2. Proofs

**Proof 4.1.** [def. 4.7]

$$\mathbb{E}_{x\sim q}\left[\log\left(\frac{1}{\mathrm{p}(x)}\right)\right] = \mathbb{E}_{x\sim q}\left[\log\left(\frac{q(x)}{\mathrm{p}(x)}\right) + \log\left(\frac{1}{q(x)}\right)\right] = H(\mathrm{p}) + D_{KL}(\mathrm{p}\parallel q)$$

**Proof 4.2.** [def. 4.15]:

$$\begin{aligned}\min_h R(h) &= \min_h \mathbb{E}_{(\mathbf{x},y)\sim \mathrm{p}}[(y-h(\mathbf{x}))^2]\\ &\overset{??}{=} \min_h \mathbb{E}_{\mathbf{x}\sim \mathrm{p}_{\mathcal{X}}}\left[\mathbb{E}_{\mathbf{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y-h(\mathbf{x}))^2\,|\mathbf{x}\right]\right]\\ &\overset{\heartsuit}{=} \mathbb{E}_{\mathbf{x}\sim \mathrm{p}_{\mathcal{X}}}\left[\min_h(\mathbf{x})\,\underbrace{\mathbb{E}_{\mathbf{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y-h(\mathbf{x}))^2\,|\mathbf{x}\right]}_{\mathcal{R}_{\mathrm{p}}(h,\mathbf{x})\ ([\text{def. } 4.11])}\right]\end{aligned}$$

*Now lets minimize the conditional executed risk:*

$$h^*(\mathbf{x}) = \arg\min_h \mathbb{E}_{\mathbf{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y-h(\mathbf{x}))^2\,|\mathbf{x}\right] \tag{4.23}$$

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

**Notes:** ♡

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

**Note**

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

**Proof 4.3.** *Definition 4.5*

$$\begin{aligned}\mathbb{E}_X[H(Y|X=x)] &= \sum_{x\in\mathcal{X}}\mathrm{p}(x)\sum_{y\in\mathcal{Y}}\mathrm{p}(y|x)\log\mathrm{p}(y|x)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathrm{p}(x)\mathrm{p}(y|x)\log\mathrm{p}(y|x)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathrm{p}(x,y)\log\mathrm{p}(y|x)\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathrm{p}(x,y)\log\left(\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)}\right)\end{aligned}$$

**Proof 4.4.** [def. 4.6] *We start from eq. (4.6):*

$$\begin{aligned}H(Y|X) &= -\mathbb{E}_{X,Y}\left[\log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)}\right]\\ &= -\sum_{x,y}\mathrm{p}(x,y)\log\mathrm{p}(x,y) + \sum_x\mathrm{p}(x)\log\frac{1}{\mathrm{p}(X)}\\ &= H(X,Y) - H(X)\end{aligned}$$

**Proof 4.5.** *example 4.4*

$$\begin{aligned}KL(\mathrm{p}||q) =&\,\mathbb{E}_{\mathrm{p}}[\log(\mathrm{p}) - \log(q)]\\ =&\,\mathbb{E}_{\mathrm{p}}\Big[\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_{\mathrm{p}}|} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})^{\intercal}\Sigma_{\mathrm{p}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})\\ &+ \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_q)^{\intercal}\Sigma_q^{-1}(\mathbf{x}-\boldsymbol{\mu}_q)\Big]\\ =&\,\frac{1}{2}\mathbb{E}_{\mathrm{p}}\left[\log\frac{|\Sigma_q|}{|\Sigma_{\mathrm{p}}|}\right] - \frac{1}{2}\mathbb{E}_{\mathrm{p}}\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})^{\intercal}\Sigma_{\mathrm{p}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})\right]\\ &+ \frac{1}{2}\mathbb{E}_{\mathrm{p}}\left[(\mathbf{x}-\boldsymbol{\mu}_q)^{\intercal}\Sigma_q^{-1}(\mathbf{x}-\boldsymbol{\mu}_q)\right]\\ =&\,\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_{\mathrm{p}}|} - \frac{1}{2}\underline{\mathbb{E}_{\mathrm{p}}\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})^{\intercal}\Sigma_{\mathrm{p}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})\right]}\\ &+ \frac{1}{2}\underline{\mathbb{E}_{\mathrm{p}}\left[(\mathbf{x}-\boldsymbol{\mu}_q)^{\intercal}\Sigma_q^{-1}(\mathbf{x}-\boldsymbol{\mu}_q)\right]}\end{aligned}$$

$$\begin{aligned}\underline{\mathbb{E}_{\mathrm{p}}[a]} &\overset{tr(\mathbb{R})=\mathbb{R}}{=} \mathbb{E}_{\mathrm{p}}\left[tr\left\{(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})^{\intercal}\Sigma_{\mathrm{p}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})\right\}\right]\\ &\overset{eq.\ (25.53)}{=} \mathbb{E}_{\mathrm{p}}\left[tr\left\{(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})(\mathbf{x}-\boldsymbol{\mu}_{\mathrm{p}})^{\intercal}\Sigma_{\mathrm{p}}^{-1}\right\}\right]\\ &= \mathbb{E}_{\mathrm{p}}\left[tr\left\{\Sigma_{\mathrm{p}}\Sigma_{\mathrm{p}}^{-1}\right\}\right]\\ &\overset{eq.\ (25.53)}{=} \mathbb{E}_{\mathrm{p}}[tr\{\mathbf{I}_d\}] = \mathbb{E}_{\mathrm{p}}[d] = d\\ \underline{\mathbb{E}_{\mathrm{p}}[b]} &\overset{eq.\ (31.54)}{=} (\boldsymbol{\mu}_{\mathrm{p}}-\boldsymbol{\mu}_q)^{\intercal}\Sigma_q^{-1}(\boldsymbol{\mu}_{\mathrm{p}}-\boldsymbol{\mu}_q) + tr\left\{\Sigma_q^{-1}\Sigma_{\mathrm{p}}\right\}\end{aligned}$$

## 3. Examples

**Example 4.1 :** Normal distribution has two population parameters: the mean $\mu$ and the variance $\sigma^2$.

**Example 4.2 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 4.3 Entropy of a Gaussian:**

$$\begin{aligned}H(\mathcal{N}(\mu,\Sigma)) &= \frac{1}{2}\ln|2\pi e\Sigma| \overset{\text{eq. }(25.54)}{=} \frac{1}{2}\ln\left((2\pi e)^d|\Sigma|\right)\\ &= \frac{d}{2}\ln(2\pi e)^d + \log|\Sigma| \qquad (4.24)\\ &\overset{\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)}{=} \frac{1}{2}\ln|2\pi e| + \cancel{\frac{1}{2}}\sum_{i=1}^{d}\ln\sigma_i^{\cancel{2}}\end{aligned}$$

**Example 4.4** proof 4.5
**KL Divergence of Gaussians:**
Given two Gaussian distributions:

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

$$D_{\mathrm{KL}}(\mathrm{p}\parallel q) = \\ = \frac{\mathrm{tr}\left(\Sigma_q^{-1}\Sigma_{\mathrm{p}}\right) + (\mu_q-\mu_{\mathrm{p}})^{\intercal}\Sigma_q^{-1}(\mu_q-\mu_{\mathrm{p}}) - d + \ln\left(\frac{|\Sigma_q|}{|\Sigma_{\mathrm{p}}|}\right)}{2}$$

**Example 4.5 KL Divergence of Scalar Gaussians:**

$$\theta\sim q(\theta|\boldsymbol{\lambda}) = \mathcal{N}\left(\mu_q,\sigma_q^2\right) \qquad \boldsymbol{\lambda} = [\mu_q \quad \sigma_q]$$

$$\mathrm{p} = \mathcal{N}\left(\mu_{\mathrm{p}},\sigma_{\mathrm{p}}^2\right)$$

$$D_{\mathrm{KL}}(\mathrm{p}\parallel q) = \frac{1}{2}\left(\frac{\sigma_{\mathrm{p}}^2}{\sigma_q^2}(\mu_q-\mu_{\mathrm{p}})^2\sigma_q^{-2} - 1 + \log\left(\frac{\sigma_q^2}{\sigma_{\mathrm{p}}^2}\right)\right)$$

**Example 4.6 KL Divergence of Diag. Gaussians:**

$$\theta\sim q(\theta|\boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{\mu}_q,\mathrm{diag}\left(\sigma_1^2,\ldots,\sigma_d^2\right)\right) \qquad \boldsymbol{\lambda} = [\boldsymbol{\mu}_{1:d} \quad \sigma_{1:d}]$$

$$\mathrm{p} = \mathcal{N}\left(\mu_{\mathrm{p}},\mathrm{diag}\left(\sigma_1^2,\ldots,\sigma_d^2\right)\right)$$

**Example 4.7 KL Divergence of Gaussians:**

$$\mathrm{p} = \mathcal{N}\left(\mu_{\mathrm{p}},\mathrm{diag}\left(\sigma_1^2,\ldots,\sigma_d^2\right)\right) \qquad q = \mathcal{N}(\mathbf{0},\mathbf{I}) \qquad \text{it holds}$$

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

**Example 4.8 Gaussian Mutal Information:**

Given $\quad X\sim\mathcal{N}(\mu,\Sigma) \quad Y = X+\epsilon \quad \epsilon\sim\mathcal{N}(0,\sigma\mathbf{I})$

$$\begin{aligned}I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(\epsilon)\\ &\overset{\text{eq. }(4.24)}{=} \frac{1}{2}\ln(2\pi e)^d|\Sigma+\sigma^2\mathbf{I}| - \frac{1}{2}\ln(2\pi e)^d|\sigma^2\mathbf{I}|\\ &= \frac{1}{2}\ln\frac{\cancel{(2\pi e)^d}\,\cancel{\sigma^2}|\sigma^{-2}\Sigma+\mathbf{I}|}{\cancel{(2\pi e)^d}|\cancel{\sigma^2}\mathbf{I}|}\\ &= \frac{1}{2}\ln|\mathbf{I}+\sigma^{-2}\Sigma|\end{aligned}$$

# Supervised Learning

$$\mathcal{D}\xrightarrow[\text{Learning method}]{\text{Model Fitting}}\left(\mathcal{X}\xrightarrow{c}\mathcal{Y}\right)\xrightarrow[\text{of data } \mathbf{x} \text{ without label}]{\text{Prediction}}\hat{\mathbf{y}}$$

**Recall:** goal of supervised learning

**Given**: training data:

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

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

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

**s.t.** we minimize prediction error/empirical risk [def. 4.14].

**Fundamental assumption**

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

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

**Note**

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

### 3.1. Generalization Error

**Definition 4.10**
**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 $\mathrm{p}_{\mathcal{X},\mathcal{Y}}$.

$$\begin{aligned}R_{\mathrm{p}}(h) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathrm{p}}[l(y;h(\mathbf{x}))] = \int_{\mathcal{D}}\mathrm{p}(\mathbf{x},y)l(y;h(\mathbf{x}))\,\mathrm{d}\mathbf{x}\,\mathrm{d}y\\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}}\mathrm{p}(\mathbf{x},y)l(y,h(\mathbf{x}))\,\mathrm{d}\mathbf{x}\,\mathrm{d}y\\ &\overset{??}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}}l(y,h(\mathbf{x}))\mathrm{p}(y|\mathbf{x})\mathrm{p}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}y \qquad (4.25)\end{aligned}$$

**Interpretation**

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

**Definition 4.11 Expected Conditional Risk:** If we only know a certain $\mathbf{x}$ but not the distribution of those measurements ($\mathbf{x}\sim\mathrm{p}_{\mathcal{X}}(\mathbf{x})$), we can still calculate the expected risk given/conditioned on the known measurement $\mathbf{x}$:

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

**Note:** [def. 4.10] $\iff$ [def. 4.11]

$$R_{\mathrm{p}}(h) = \mathbb{E}_{\mathbf{x}\sim\mathrm{p}}[R_{\mathrm{p}}(h,\mathbf{x})] = \int_{\mathcal{X}}\mathrm{p}(\mathbf{x})R_{\mathrm{p}}(h,\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{4.26}$$

**Definition 4.12 Expected Risk Minimizer (TRM) $h^*$:** Is the model $h$ that minimizes the total expected risk:

$$h^*\in\arg\min_{h\in\mathcal{C}}\mathcal{R}(h) \tag{4.27}$$

**Problem**

In practice we do neither know the distribution $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\mathbf{x},y)$, nor $\mathrm{p}_{\mathcal{X}}(\mathbf{x})$ or $\mathrm{p}_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x})$ (otherwise we would already know the solution).
**But**: even though we do not know the distribution of $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\mathbf{x},y)$ we can still sample from it in order to define an empirical risk.

### 3.2. Empirical Risk

**Definition 4.13 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 $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\mathbf{x},y)$:

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

**Note**

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

**Definition 4.14 Empirical Risk Minimizer (ERM) $\hat{h}$:** Is the model $h$ that minimizes the total empirical risk:

$$\hat{h}\in\arg\min_{h\in\mathcal{C}}\hat{\mathcal{R}}(h) \tag{4.28}$$

**Objective**

**Given** data generated *i.i.d.* from an distribution $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\mathbf{x}_i,y_i)$.
**Goal**: find the function/predictor $h:\mathcal{X}\mapsto\mathcal{Y}$ that minimizes the expected risk [def. 4.10] i.e. we want to find the expected risk minimizer ([def. 4.12]).

**Definition 4.15**
**Bayes' optimal predictor for the L2-Loss:**
**Assuming**: i.i.d. generated data by $(\mathbf{x}_i,y_i)\sim\mathrm{p}(\mathcal{X},\mathcal{Y})$.
**Considering**: the least squares risk:

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

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

$$h^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X}=\mathbf{x}] \tag{4.29}$$

see

**Proof 4.6: .** *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 obsereve 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 genrating process $\mathbf{X}=\mathbf{x}$ we may still obtain different $y$'s because due to randomn/independent measurment noise/errors that the optimal bayes predictor still makes.

#### 3.2.1. Bayes Optimal Predictor

## 3.3. How to make use of this in Practice

**In Practice**

**Problem**: we do not know the real distribution $p_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x})$, which we need in order to find the bayes optimal predictor according to eq. (4.29).

**Idea**:

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

$$\hat{y} = \hat{\mathbb{E}}[\mathcal{Y}|\mathcal{X} = \mathbf{x}] = \int \hat{p}(y|\mathbf{X} = \mathbf{x}) y \, dy$$

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

**Definition 4.16 Statistical Inference:** 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?

# 4. Estimators

**Definition 4.17 (Sample) Statistic:** A statistc is a measuarble function $f$ that assigns a **single** value $F$ to a sample of random variables: $\mathbf{X} = \{X_1, \dots, X_n\}$

$$f : \mathbb{R}^n \mapsto \mathbb{R} \qquad F = f(X_1, \dots, X_n)$$

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 4.18 Statistical/Population Parameter:**
Is a parameter defining a family of probabilty distributions see example 4.1

**Definition 4.19 (Point) Estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$:**
**Given**: n-samples $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathbf{X}$ an estimator

$$\hat{\theta} = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \tag{4.30}$$

is a statistic/randomn variable used to estimate a true (population) parameter $\theta$[def. 4.18] see also example 4.2.

**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 $\theta$.
The most prevalent forms of interval estimation are:

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

# Regression

**Definition 5.1**
**Explanatory-/Indep.-/Predi.-/Variables/Covariates x:**
Are the input variable(s) that we want to relate to the response variable(s)[def. 5.2].

**Definition 5.2**
**Response-/Dependent-/Variable(s)** **y:**
Are the output quantities that we are interested in.

**Definition 5.3 Regression:** Is the process of finding a possible relationship via some coefficients $\beta$ between *response-variables* $\mathbf{x}$ and a *predictor-variable*(s) $\mathbf{y}$ up to some error $\epsilon$:

$$\mathbf{y} = f(\mathbf{x}, \beta) + \epsilon \tag{5.1}$$

**Note**

The term regression comes from the latin term "regressus" and means "to go back" to something. Historically the term was introduced by Galton, who discovered that given an outlier point, further observations will regress back to the mean. In particular he discovered that children of very tall/small people tend to be a smaller/larger.

**Definition 5.4 Linear Regression:** Refers to regression that is linear w.r.t. to the parameter vector $\beta$ (but not necessarily the data):

$$\mathbf{y} = \beta^{\intercal}\phi(\mathbf{x}) + \epsilon \tag{5.2}$$

**Linearity**

Linearity is w.r.t. the coefficients $\beta_j$.

Thus a model with transformed non-linear predictor[def. 5.1] variables is still called *linear*.

**Definition 5.5 Residual** **r:**
Let us consider $n$ observations $\{x_i, y_i\}_{i=1}^n$. The residual (error) is the deviation of the observed values from the predicted values:

$$r_i := e_i = y_i - \hat{y}_i = y_i - \hat{\beta}^{\intercal}\mathbf{x}_i \qquad i = 1, \ldots, n \tag{5.3}$$

## Simple (linear) regression (SLR)

**Definition 5.6** [example 5.1]
**Simple Linear Regression:** Is a *linear regression*[def. 5.7] with only one explanatory variable[def. 5.1]:

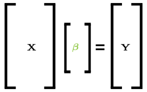
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{5.4}$$

## Multiple (linear) regression (MLR)

**Definition 5.7 Multiple Linear Regression:**
Is a linear regression model with multiple $\{\beta_j\}_{j=1}^p$ explanatory[def. 5.1] variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i = \beta^{\intercal} x_i + \epsilon_i \qquad i = 1, \ldots, n$$



$$\mathbf{y} = \mathbf{X}\beta$$

**Design Matrix:**
$\mathbf{X} \in \mathbb{R}^{n,(p+1)}$
$\mathbf{y} \in \mathbb{R}^n$
$\beta \in \mathbb{R}^{p+1}$ (5.5)

add offset inclusion and fix dimension either p+1 or p' or something

**Note**

Eq. 5.7 is usually an over-determined system of linear equations i.e. we have more observations then predictor variables.

**Multiple vs. Multivariate lin. Reg.**

Multivariate linear regression is simply linear regression with multiple response variables and thus nothing else but a set of simple linear regression models that have the same types of explanatory variables.

**Definition 5.8** [example 5.2]
**Simple Linear Quadratic Regression:** Is a *linear regression*[def. 5.7] with two explanatory variables[def. 5.1] written as:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i \qquad i = 1, \ldots, n \tag{5.6}$$

### 0.0.1. Existence

**Corollary 5.1 Existence:**

$$\exists\beta : \quad \begin{matrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p \end{matrix} = \begin{matrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{matrix} \tag{5.7}$$

$$\iff \quad \mathbf{y} \in \mathfrak{R}(\mathbf{X}) \tag{5.8}$$

## 1. Linear/Ordinary Least Squares (OLS)

**Problem**: for an over determined system $n > p$ (usually) $\nexists\mathbf{y} \in \mathfrak{R}(\mathbf{X})$ (in particular given round off errors) s.t. there exists no parameter vector $\beta$ that solves [def. 5.7].
**Idea**: try to find the next best solution by minimizing the residual(s)[def. 5.5].

**Definition 5.9 Residual Sum of Squares:**
Is the sum of residuals[def. 5.5]:

$$\mathbf{RSS}(\beta) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 \tag{5.9}$$

**Definition 5.10 Least Squares Regression** **lsq(X, y):**
Minimizes the residual sum of squares:

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \operatorname*{arg\,min}_{\mathbf{u}\in\mathfrak{R}(\mathbf{X})} \|\mathbf{y} - \mathbf{u}\|_2^2 \tag{5.10}$$

$$= \operatorname*{arg\,min}_{\beta} \|\mathbf{r}\|_2^2 = \sum_{i=1}^{n} \left( \sum_{j=1}^{p} x_{ij}\beta_j - y_i \right)^2 = \mathbf{RSS}(\beta)$$

**Alternative Formulation**

Sometimes people write eq. (5.10) as $\frac{1}{2} \operatorname{arg\,min}_{\beta} \|\mathbf{r}\|_2^2$ which leads to the same solutioneq. (22.59).

## 2. Maximum Likelihood Estimate

### Ridge MLE

**Proposition 5.1**
**Assumptions for Linear Regression Model:**

1. The $\{\mathbf{x}_i\}_{i=1}^n$ are deterministic and measured without errors.
2. The variance of the error terms is *homoscedastic*[def. 35.22]:
$$\mathbb{V}[\epsilon_i] = \sigma^2 \qquad \forall i \tag{5.11}$$
3. The errors are uncorrelated:
$$\operatorname{Cov}[\epsilon_i, \epsilon_j] = 0 \qquad \forall i \neq j \tag{5.12}$$
4. The errors are jointly normally distributed with mean 0 and constant variance $\sigma^2$:
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i = 1, \ldots, n \quad \iff \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n) \tag{5.13}$$

**Definition 5.11** [proof 5.1]
**Simple Linear Regression Log-Likelihood:**

**Assume**: a linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$
with Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$

**With**: $\mu = \mathbb{E}_\epsilon[\mathbf{y}] = \mathbb{E}_\epsilon[\mathbf{X}\beta + \epsilon] = \mathbf{X}\beta + 0$
$\mathbb{V}_\epsilon[\mathbf{y}] = \mathbb{V}_\epsilon[\mathbf{X}\beta + \epsilon] = 0 + \mathbb{V}[\epsilon] = \mathbf{I}\sigma^2$

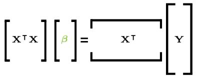
**Thus**: $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{I}\sigma^2)$ $\qquad Y_i|\mathbf{X} \sim \mathcal{N}(\mathbf{x}_i^{\intercal}\beta, \sigma^2)$
with: $\theta = (\beta^{\intercal} \quad \sigma)^{\intercal} \in \mathbb{R}^{p+1}$

$$\mathrm{l}_n(\mathbf{y}|\mathbf{X}, \theta) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^{\intercal}\mathbf{x}_i)^2 = -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2 \tag{5.14}$$

$$\theta^* \in \operatorname*{arg\,max}_{\theta\in\mathbb{R}^{p+1}} \mathrm{l}_n(\mathbf{y}|\mathbf{X}, \theta) = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{p+1}} -\mathrm{l}_n(\mathbf{y}|\mathbf{X}, \theta) \tag{5.15}$$

### 2.1. The Normal Equation

**Definition 5.12** [proof 5.3]
**The Normal Equations:**
Is the equation we need to solve in order to solveeq. (5.10) or equivalently eq. (5.15) and is no longer an over determined system:

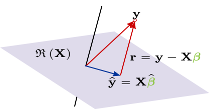


$$\mathbf{X}^{\intercal}\mathbf{X}\hat{\beta} = \mathbf{X}^{\intercal}\mathbf{Y}$$

$\mathbf{X}^{\intercal}\mathbf{X} \in \mathbb{R}^{p\times p}$
$\beta \in \mathbb{R}^p$
$\mathbf{X}^{\intercal} \in \mathbb{R}^{p\times n}$
$\mathbf{Y} \in \mathbb{R}^n$
(5.16)

**Geometric Interpretation**

**Corollary 5.2** [proof 5.4]
**Geometric Interpretation:**



We want to find $\operatorname{arg\,min}_{\beta\in\mathbb{R}^n} \|\mathbf{X}\beta - \mathbf{y}\|_2^2$ which is equal to finding:

$$\operatorname*{arg\,min}_{\hat{\mathbf{y}}\in\{\mathbf{X}\beta:\beta\in\mathbb{R}^n\}=\mathfrak{R}(\mathbf{X})} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$$

but this minimum is equal to the orthogonal projection[def. 25.19] of $\mathbf{y}$ onto $\mathfrak{R}(\mathbf{X})$.

#### 2.1.1. The Least Squares Solution $\hat{\beta} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{Y}$

**Proposition 5.2 Least Squares Solution:**

$$\hat{\beta} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y} := \mathbf{X}^{\dagger}\mathbf{y} \tag{5.17}$$

**Note**

$\mathbf{X}^{\dagger}$ is the Moore-Penrose pseudo-inverse of the matrix $\mathbf{X}$.

add MPPI to linear algebra appendix

#### 2.1.2. Solving The Normal Equation

**Cholesky Decomposition**

**Corollary 5.3 Computational Complexity:** $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^d$ with $n$, the number of observations and $d$, the number of equations/feautres/dimension of the problem.
**Assume**: $d \leqslant n$, that is we have an overdetermined system, more equations than unkowns.

1. Compute regular matrix (Matrix Product): $\mathbf{C} := \mathbf{X}^{\intercal}\mathbf{X} \mathrel{\hat{=}} \mathcal{O}(n \cdot d^2)$.
2. Compute the r.h.s. vector (Matrix-Vector): $\mathbf{c} := \mathbf{X}^{\intercal}\mathbf{y} \in \mathbb{R}^d \mathrel{\hat{=}} \mathcal{O}(nd)$.
3. Solve s.p.d. LSE via. Cholesky decomposition: $\mathbf{C}\mathbf{w} = \mathbf{c} \mathrel{\hat{=}} \mathcal{O}(d^3)$.

Thus the total cost amounts to $\mathcal{O}(d^3 + nd^2)$.

**Note: s.p.d. C and cholesky decomposition**

**Assume**: $\mathbf{X}$ has a trivial kernel $\iff$ $\mathbf{X}^{\intercal}\mathbf{X}$ is invertible.

1. Symmetric: a transposed matrix times itself is symmetric $\Rightarrow$ $\mathbf{C}$ is symmetric.
2. Posistive definite:
$$\mathbf{w}^{\intercal}\mathbf{C}\mathbf{w} = \mathbf{w}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{w} = \|\mathbf{X}\mathbf{w}\|^2 \underset{\text{has trivial kernel}}{>} 0 \qquad \forall\mathbf{w} \neq 0$$

**QR Decomposition**

#### 2.1.3. Simple Linear Regression Solution

**Definition 5.13** [proof 5.3]
**Linear Regression Solution:**

$$\hat{\beta} = \underbrace{(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}}_{\mathbf{X}^{\dagger}}\mathbf{y} \quad \text{with} \quad \begin{aligned} \Sigma^2 &= (\mathbf{X}^{\intercal}\mathbf{X})^{-1} \\ \mathrm{P} &= \mathbf{X}^{\intercal}\mathbf{y} \end{aligned} \tag{5.18}$$

$\Sigma^2$ : Varianece-Covar. M. $\quad$ P : Inp./Oup. Covariance

**Moore-Penrose pseudo-inverse**: $\mathbf{X}^{\dagger}$ with $\mathbf{X}^{\dagger}\mathbf{X} = \mathbf{I}$ (5.19)

#### 2.1.4. Making Predictions

**Definition 5.14** $\quad \mathbf{P}/\mathbf{H} = \mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal} : y \mapsto \hat{y}$
**Hat/Projection Matrix:**
Is the matrix that projects the $y$ onto the $\hat{y}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y} =: \mathbf{P}\mathbf{y} \tag{5.20}$$

**Uniqueness**

**Theorem 5.1 :** Let $\mathbf{A} \in \mathbb{R}^{p,p}$, $p \geqslant p$ then it holds that:

$$\mathtt{N}(\mathbf{A}) = \mathtt{N}(\mathbf{A}^{\intercal}\mathbf{A}) \qquad \mathfrak{R}(\mathbf{A}^{\intercal}) = \mathfrak{R}(\mathbf{A}^{\intercal}\mathbf{A}) \tag{5.21}$$

add rest from formulary

**Theorem 5.2 Full-Rank Condition** **F.R.C.:**
Equation 5.12 has a unique least squares solution given by:

$$\hat{\beta} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{Y} \tag{5.22}$$

$$\iff \quad \mathtt{N}(\mathbf{X}) = \{0\} \quad \iff \quad \operatorname{rank}(\mathbf{X}) = p \qquad p \geqslant p \tag{5.23}$$

add delete next section

# 3. MLE with linear Model & Gaussian Noise

## 3.1. MLE for conditional linear Gaussians

**Questions**: what is $\mathbb{P}(Y|X)$ if we assume a relationship of the form: We can use the MLE to estimate the parameters $\theta\,\mathbb{R}^k$ of a model/distribution $h$ s.t.

$$\mathbf{y} \approx h(\mathbf{X};\theta) \iff \mathbf{y} = h(\mathbf{X};\theta) + \epsilon$$

$\mathbf{X}$: set of explicative variables. $\epsilon$: noise/error term.

**Lemma 5.1 :** The conditional distribution $D$ of $Y$ given $\mathbf{X}$ is equivilant to the unconditional distribution of the noise $\epsilon$:

$$\mathbb{P}(Y|\mathbf{X}) \sim D \iff \epsilon \sim D$$

### Example: Conditional linear Gaussian

**Assume**: a linear model $h(\mathbf{x}) = \mathbf{w}^\intercal\mathbf{x}$

and Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

With $\mathbb{E}[\epsilon] = 0$ and $y_i = \mathbf{w}^\intercal\mathbf{x} + \epsilon$, as well as **??** it follows:

$$y \sim \hat{\mathrm{p}}(Y = y|\mathbf{X} = \mathbf{x}, \theta) \sim \mathcal{N}(\mu = h(\mathbf{x}), \sigma^2)$$

with: $\theta = \begin{pmatrix}\mathbf{w}^\intercal & \sigma\end{pmatrix}^\intercal \in \mathbb{R}^{n+1}$

**Hence** $Y$ is distributed as a linear transformation of the $\mathbf{X}$ variable plus some Gaussian noise $\epsilon$: $y_i \sim \mathcal{N}(\mathbf{w}^\intercal\mathbf{x}_i, \sigma^2) \Rightarrow$ Conditional linear Gaussian.
if we consider an i.i.d. sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$, the corresponding conditional (log-)likelihood is defined to be:

$$\begin{aligned}
\mathscr{L}_n(Y|\mathbf{X}, \theta) &= \hat{\mathrm{p}}(y_1, \ldots, y_n|\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta) \\
&\overset{\text{i.i.d.}}{\underset{\mathbf{??}}{=}} \prod_{i=1}^n \hat{\mathrm{p}}_{Y|\mathbf{X}}(y_i|\mathbf{x}_i, \theta) = \prod_{i=1}^n \mathcal{N}(\mathbf{w}^\intercal\mathbf{x}_i, \sigma^2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2}{2\sigma^2}\right) \\
&= \left(\sigma^2 2\pi\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2\right)
\end{aligned}$$

$$\boxed{\mathrm{l}_n(Y|\mathbf{X}, \theta) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2}$$

$$\theta^* = \underset{\mathbf{w}\in\mathbb{R}^d, \sigma^2\in\mathbb{R}_+}{\arg\max}\ \mathrm{l}_n(Y|\mathbf{X}, \theta)$$

$$\frac{\partial\mathrm{l}_n(Y|\mathbf{X}, \theta)}{\partial\theta} = \begin{pmatrix} \frac{\partial\mathrm{l}_n(Y|\mathbf{X},\theta)}{\partial w_1} \\ \vdots \\ \frac{\partial\mathrm{l}_n(Y|\mathbf{X},\theta)}{\partial w_d} \\ \frac{\partial\mathrm{l}_n(Y|\mathbf{X},\theta)}{\partial\sigma^2} \end{pmatrix} \overset{!}{=} \begin{pmatrix}\mathbf{0}_d \\ 0\end{pmatrix}$$

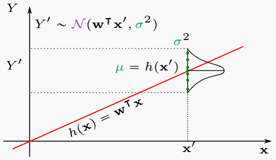
$$\begin{aligned}
\frac{\partial\mathrm{l}_n(Y|\mathbf{X}, \theta)}{\partial\mathbf{w}} &= \frac{1}{\sigma^2}\sum_{i=1}^n \mathbf{x}_i\left(y_i - \mathbf{w}^\intercal\mathbf{x}_i\right) = \mathbf{0} \in \mathbb{R}^d \\
&= \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\intercal\right)\mathbf{w} = \sum_{i=1}^n \mathbf{x}_i y_i \\
\frac{\partial\mathrm{l}_n(Y|\mathbf{X}, \theta)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2 = 0
\end{aligned}$$

$$\boxed{\theta^* = \begin{pmatrix}\mathbf{w}_* \\ \sigma_*^2\end{pmatrix} = \begin{pmatrix}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\intercal\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i y_i\right) \\ \frac{1}{n}\sum_{i=1}^n \left(y_i - \mathbf{w}_*^\intercal\mathbf{x}_i\right)^2\end{pmatrix}} \tag{5.24}$$

### Note

- The mean $\mu$ of the normal distribution follows from:
$$\mathbb{E}\left[\mathbf{w}^\intercal\mathbf{x}_i + \epsilon_i\right] = \mathbb{E}[\underbrace{\mathbf{w}^\intercal\mathbf{x}_i}_{\text{const}}] + \underbrace{\mathbb{E}[\epsilon_i]}_{=0} = \mathbf{w}^\intercal\mathbf{x}_i$$
- The noise $\epsilon$ must have zero mean, otherwise it wouldn't be randomn anymore.
- The optimal function $h^*(\mathbf{x})$ determines the mean $\mu$.
- We can also minimize:
$$\theta^* = \underset{\theta}{\arg\max}\ \hat{\mathrm{p}}(Y|\mathbf{X}, \theta) = \underset{\theta}{\arg\min} -\hat{\mathrm{p}}(Y|\mathbf{X}, \theta)$$

The Liklihood does not explicitly care about the distribution of the $\mathbf{x}$ **but** for a given $\mathbf{x}'$ it carse about modeling the distribution of the chosen model $f_{\mathcal{Y}|\mathcal{X}}(y|\mathbf{x}', \theta)$ around the linear function $h(\mathbf{x}') = \mathbf{w}^\intercal\mathbf{x}'$
$\Rightarrow y'$ is an sample from the distribution/model.



$Y$
$Y' \sim \mathcal{N}(\mathbf{w}^\intercal\mathbf{x}', \sigma^2)$
$\sigma^2$
$Y'$
$\mu = h(\mathbf{x}')$
$h(\mathbf{x}) = \mathbf{w}^\intercal\mathbf{x}$
$\mathbf{x}'$
$\mathbf{x}$

## 3.2. Conditional MLE ≙ Least Squares

**Assuming** that the noise is i.i.d. Gaussian with *constant* variance $\sigma$, that is $\theta = \begin{pmatrix}\mathbf{w} & \cancel{\sigma^2}\end{pmatrix}^\intercal$
and considering the negative log likelihood in order to minimize $\arg\max\alpha = -\arg\min\alpha$:

$$-\mathrm{l}_n(\mathbf{w}) = -\prod_{i=1}^n \ln\mathcal{N}(\mathbf{w}^\intercal\mathbf{x}_i, \sigma^2) = \frac{n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^n \frac{(y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2}{2\sigma^2}$$

$$\underset{\mathbf{w}}{\arg\max}\ \mathrm{l}_n(\mathbf{w}) \iff \underset{\mathbf{w}}{\arg\min} -\mathrm{l}_n(\mathbf{w})$$

$$\underset{\mathbf{w}}{\arg\min}\ \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2 = \underset{\mathbf{w}}{\arg\min}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2 \tag{5.25}$$

**Thus** Least squares regression equals Conditional MLE with a linear model + Gaussian noise.
Maximizing Liklihood $\iff$ Minimizing least squares

**Corollary 5.4 :** 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. 4.15]

**Consider** a sample $\{y_1, \ldots, y_n\} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned}
\frac{\partial\mathrm{l}_n(y|\mathbf{x}, \theta)}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mu) \overset{!}{=} 0 \\
\frac{\partial\mathrm{l}_n(y|\mathbf{x}, \theta)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \mu)^2 \overset{!}{=} 0
\end{aligned}$$

$$\boxed{\theta^* = \begin{pmatrix}\mu_* \\ \sigma_*^2\end{pmatrix} = \begin{pmatrix}\frac{1}{n}\sum_{i=1}^n y_i \\ \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_i)^2\end{pmatrix}} \tag{5.26}$$

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

### Note

$$\frac{\partial\mathbf{w}^\intercal\mathbf{x}}{\partial\mathbf{w}} = \frac{\partial\mathbf{x}^\intercal\mathbf{w}}{\partial\mathbf{w}} = \mathbf{x}$$

## 3.3. MLE for general conditional Gaussians

**Suppose** we do not just want to fit linear functions but a gerneal class of models $Hsp := \{h : \mathcal{X} \mapsto \mathbb{R}\}$ e.g. neural networks, kernel functions,...
**Given**: data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ The MLE for general models $h$ and i.i.d. Gaussian noise:

$$h \sim \hat{\mathrm{p}}_{Y|\mathbf{X}}(Y = y|\mathbf{X} = \mathbf{x}, \theta) = \mathcal{N}(y|h^*(\mathbf{x}), \sigma^2)$$

Is given by the least squares solution:

$$h^* = \underset{h\in\mathcal{H}}{\arg\min}\sum_{i=1}^n (y_i - h(\mathbf{x}_i))^2$$

E.g. for linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\intercal\mathbf{x}$ with parameter $\mathbf{w}\}$

### Other distributions

If we use other distributions instead of Guassian noise, we obtain other loss functions e.g. L1-Norm for **Poission Distribution**.
$\Rightarrow$ if we know somthing about the distribution of the data we know which loss fucntion we should chose.

### Ridge Max Prior

### Prior

**Assume**: prior $\mathbb{P}(\beta|\mathbf{\Sigma})$ on the model parameter $\beta$ is gaussian as well and depends on the hyperparameter ([def. 6.7]) $\mathbf{\Sigma}$ (≙ co-variance matrix):

$$\begin{aligned}
\beta &\sim \mathrm{p}^{\text{Ridge}}(\beta|\mathbf{\Sigma}) = \mathcal{N}(\beta|0, \mathbf{\Sigma}) \\
&\overset{\text{[def. 32.33]}}{=} (2\pi)^{-\frac{d+1}{2}}\det(\mathbf{\Sigma})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\beta^\intercal\mathbf{\Sigma}^{-1}\beta\right)
\end{aligned}$$

$$\mathrm{l}_n(\beta|\mathbf{\Sigma}) = -\frac{1}{2}\ln\det(\mathbf{\Sigma})^{-1} - \frac{d+1}{2}\ln 2\pi - \frac{1}{2}\beta\mathbf{\Sigma}^{-1}\beta \tag{5.27}$$

eventually add change of variabels formula and check if factor n is missing

### Max Prior

$$\begin{aligned}
\beta^* &\in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\ \mathrm{l}_n(\beta|\mathbf{\Sigma}) \\
&= \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max} -\frac{1}{2}\ln\det(\mathbf{\Sigma})^{-1} - \frac{d+1}{2}\ln 2\pi - \frac{1}{2}\beta\mathbf{\Sigma}^{-1}\beta \\
0 &\overset{!}{=} \frac{\partial}{\partial\beta^*}\mathrm{l}_n(\beta^*|\mathbf{\Sigma}) = -\frac{\partial}{\partial\beta^*}\beta^*\mathbf{\Sigma}^{-1}\beta^* \overset{\text{eq. (5.36)}}{=} -2\mathbf{\Sigma}^{-1}\beta^* \\
\beta^* &\in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\log\mathrm{p}(\beta|\mathbf{\Sigma}) = \underset{\beta\in\mathbb{R}^{d+1}}{\arg\min} -\mathrm{l}_n(\beta|\mathbf{\Sigma}) = 2\mathbf{\Sigma}^{-1}\beta^*
\end{aligned} \tag{5.28}$$

### Log-MAP

$$\begin{aligned}
\beta^* &\in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\ \mathbb{P}(\beta|\mathbf{X}, \mathbf{y}) \\
&= \underset{\beta\in\mathbb{R}^{d+1}}{\arg\min} -\log\overbrace{\mathbb{P}(\beta|\mathbf{\Sigma})}^{\text{eq. (5.28)}} - \log\overbrace{\mathbb{P}(\mathbf{X}, \mathbf{y}|\beta)}^{??} \\
&= \mathbf{\Sigma}^{-1}\beta^* - \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{y} + \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{X}\beta^* = 0 \\
\iff &(\mathbf{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X}\sigma^{-2})\beta^* = \sigma^{-2}\mathbf{X}^\intercal\mathbf{y} \\
&(\sigma^2\mathbf{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X})\hat{\beta} = \mathbf{X}^\intercal\mathbf{y} \\
\hat{\beta}^{\text{MAP}} &= \left(\sigma^2\mathbf{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}
\end{aligned}$$

**Definition 5.15 Ridge MAP:** For ridge regression we assume that the noise of the prior is uncorrelated/diagonal i.e.

$$\mathbf{\Sigma}^{-1} = \mathbf{I}\sigma^{-2} \quad \textbf{and let} \quad \mathbf{\Lambda} := \sigma^2\mathbf{\Sigma}^{-1} = \mathbf{I}\frac{\sigma^2}{\sigma^2} \tag{5.29}$$

which leads to:

$$\hat{\beta}^{\text{MAP}} = (\mathbf{\Lambda} + \mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \quad \text{with} \quad \mathbf{\Lambda} = \mathbf{I}\lambda = \mathbf{I}\frac{\sigma^2}{\sigma^2} \tag{5.30}$$

**Definition 5.16 Regularization:** Regularization is the process of introducing additional information/bias in order to solve an ill-posed problem or to prevent overfitting.
(It is not feature selection)

**Definition 5.17 Tikhonov regularization:** Commonly used method of regularization of ill-posed problems.

$$\|\mathbf{X}\beta - \mathbf{y}\|^2 + \|\mathbf{\Gamma}\beta\|^2 \tag{5.31}$$

$\Gamma$: Tikhonov matrix in many cases, this matrix is chosen as $\Gamma = \alpha\mathbf{I}$ giving preference to solutions with smaller norms; this is known as **Ridge/L2 regularization**.

### Gaussian Prior/Liklihood MAP inference

$$\begin{aligned}
\hat{\beta}^{\text{Ridge}} &= \underset{\beta}{\arg\min}\left\{\underbrace{(\mathbf{y} - \mathbf{X}\beta)^\intercal(\mathbf{y} - \mathbf{X}\beta)}_{\text{data term}} + \underbrace{\beta^\intercal\mathbf{\Lambda}\beta}_{\text{regularizer/penalty}}\right\} \\
&= \underset{\beta}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \beta^\intercal\mathbf{\Lambda}\beta\right\} \\
&\overset{\text{eq. (5.29)}}{=} \underset{\beta}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2\right\} \\
&= \underset{\beta}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\sum_{i=1}^d \beta_i^2\right\}
\end{aligned}$$

- $\|\mathbf{y} - \mathbf{X}\beta\|^2$ is forced to be small so that we find a weight vector $\beta$ that matches the data as close as possible:
$$y_i = \beta_i\mathbf{x}_i + \epsilon_i \qquad \text{s.t.} \qquad \sum_{i=1}^n \epsilon_i \text{ small}$$
In other words we want to fit the data well.
- $\beta^\intercal\Lambda\beta \overset{\text{ridge}}{=} \lambda\|\beta\|^2$ says chose a model with a small magnitude $\|\beta\|^2$.
**Thus** the smaller $\lambda$ the bigger can the data faith fullness term be $\|\mathbf{y} - \mathbf{X}\beta\|^2$.

### Note

The intercept $\beta_0$ in the regularizer term has to be left out.
Penalization of the intercept would make the procedure depend on the origin chosen for $y$.
**Thus** we actually have (for data with non-zero mean):

$$\beta^* = \underset{\beta\in\mathbb{R}^d}{\arg\min}\left\{\|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + +\lambda\sum_{i=1}^d \beta_i^2\right\}$$

### Note: SVD

Using SVD one can show that ridge regression shrinks first the eigenvectors with minimum explanatory variance.
**Hence** L2/Ridge regression can be used to estimate the predictor importance and penalize predictors that are not important (have small explanatory variance).

### Note: no feature selection

The coefficients in a ridge will go to zero as $\lambda$ increases but will no become zero (as long as $\lambda \neq \infty$)!
They are fit in a restricted fashion controlled by the shrinkage penalty $\lambda$.

$$\text{dofs}(\lambda) = \begin{cases} d & \text{if } \lambda = 0 \text{ (no regularization)} \\ \to 0 & \text{if } \lambda \to \infty \end{cases} \tag{5.32}$$

$\Rightarrow$ Ridge cannot be used for variable selection since it retains all the predictors

Balance of $\lambda = \frac{\sigma^2}{\sigma^2}$ controls the tradeoff between simplicity and data faith fullness because:

① $\lambda \xrightarrow[\sigma\downarrow]{\sigma\uparrow} \infty$: $\|\beta\|^2$ must be minimized:

- $\sigma\uparrow$: model does not need to match data so perfectly as we have more noise in our data/observations $\iff$ bigger errors (recall $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$).
- $\sigma\downarrow$: prior has smaller variance, thus our prior knowledge of the model is pretty exact/important (recall $\beta \sim \mathcal{N}(\beta|0, \mathbf{I}\sigma)$)

② $\lambda \xrightarrow[\sigma\uparrow]{\sigma\downarrow} 0$: $\|\mathbf{y} - \mathbf{X}\beta\|^2$ must be minimized: model must match data perfectly

- $\sigma\downarrow$: model does need to match perfectly, our observation/data has small variance/is well defined $\iff$ do not allow big errors (recall $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$).
- $\sigma\uparrow$: our knowledge about the model is pretty vague (recall $\beta \sim \mathcal{N}(\beta|0, \mathbf{I}\sigma)$)

### Note

- Often $\Lambda^{-1} = \mathbb{1} \in \mathbb{R}^{d+1\times d+1}$
- $\Lambda$ is symmetric and diagonal.
- $(d+1)$ dimension as we included offset into $\beta$.

### Heuristic Map Inference

A really large weight vector $\beta$ will result in amplifying noise/larger variance/fluctuations ≙ overfitting.
This is because the complexity of the estimate increases with the magnitude of the parameter as it becomes easier to fit complex noise.

### Ill-posed problem/Invertability and Ridge

Another advantage of Ridge regression is that, even if $\mathbf{X}^\intercal\mathbf{X}$ in eq. (5.30) is not invertible/regular/has not full rank.
Then $(\mathbf{X}^\intercal\mathbf{X} + \mathbf{\Lambda})$ will still be invertible/well posed.
This was the original reason for L2/Ridge Regression.

### MAP ≙ Ridge

$$\arg\max_{\mathbf{w}} \mathbb{P}(\mathbf{w}|\mathbf{x}, y) = \arg\min_{\mathbf{w}} \lambda\|\mathbf{w}\|^2 + \sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2$$

MAP with a linear model and Gaussian noise equals classical ridge regression **??**.

$$\underbrace{\arg\min_{\mathbf{w}} \lambda\|\mathbf{w}\|^2 + \sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2}_{\text{Ridge Regression}} \equiv \underbrace{\arg\max_{\mathbf{w}} \mathbb{P}(\mathbf{w}) \prod_{i=1}^n \mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w})}_{\text{MAP}}$$

**Thus** if we know our data $\beta$, $\sigma$ we can chose $\lambda$ statistically and do not need cross-validation.

## Generalization

Regularized estimation can often be understood as MAP inference:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^n l(\mathbf{w}^\intercal\mathbf{x}_i; \mathbf{x}_i, y_i) + C(\mathbf{w}) =$$
$$= \arg\max_{\mathbf{w}} \prod_{i=1}^n \mathbb{P}(\mathbf{w})\mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w}) = \arg\max_{\mathbf{w}} \mathbb{P}(\mathbf{w}|\text{data})$$

**with** $\quad C(\mathbf{w}) = -\log \mathbb{P}(\mathbf{w})$
$\quad l(\mathbf{w}^\intercal\mathbf{x}_i; \mathbf{x}_i, y_i) = -\log \mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w})$

## Priors

### 3.4. Laplace Prior≙Lasso/L1-regularization

### Intro

**Question**: what if $d \gg n$ e.g.

- bag of words with $d$ =nb. of words≫ $n$nb. of documents.
- Genome analysis $d$ =nb. of genes≫ $n$ patients.

**Problem**: we have more unkowns/parameters than observations ⇒ no unique solution. **e.g.**: Trying to fit 1 data point with polynomimal of degree 12.

**Question**: can we somehow still find a good solution if
$n = \mathcal{O}(\ln d) \iff$ exp. more dim. than observations

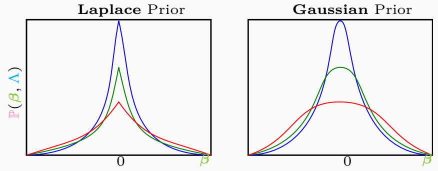
**Idea**: If most of the dimensions are irrelevant for the problem, then we can find a good (sparse) solution ≙ feature selection/dimensionality reduction.

**Given**: Laplacian model prior $\beta \sim \mathrm{p}(\beta|\Lambda)$:

$$\mathbb{P}^{\text{Lasso}}(\beta|\Lambda) \overset{\text{eq. (32.58)}}{=} \frac{\Lambda}{2} e^{(-\Lambda|\beta|)} = \prod_{j=1}^d \frac{\lambda_j}{2} e^{-\lambda_j|\beta_j|}$$

**With** $\quad \Lambda^{-1} := \Sigma$ hyperparameter/covariance matrix

This leads to a L1 regularized model:



**Thus**: laplace priors gives sparesness, higher liklihood to get value at $\beta = 0$.

$$-\ln \mathbb{P}(\beta|\Lambda) = \sum_{j=1}^d \lambda_j|\beta_j| - d\ln\frac{\lambda_j}{2} \tag{5.33}$$

### Laplacian MAP Prior Inference

$$\begin{aligned}\beta^* &= \arg\min_{\beta\in\mathbb{R}^d} \left\{\|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda\|\beta\|_1\right\}\\ &= \arg\min_{\beta\in\mathbb{R}^d} \left\{\|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda\sum_{i=1}^d |\beta_i|\right\}\end{aligned} \tag{5.34}$$

$|\beta|_i$ does not change $\beta_i$ while $\beta_i^2$ becomes very small for values∈ $(0, 1)$ thus when minimizing the L2 error $\|betac\|^2 \to 0$ **but** not $\beta_i$ while for L1 regularization will actually have to set $\beta_i$ values to zero for large enough $\lambda$.

### Advantage

Combines advantages of Ridge regression (convex function/optimization) and L0-regression (sparse and easy to interpret solution).

### Difference L1& L2 penalties

Typically ridge or L2 penalties are much better for minimizing prediction error rather than L1 penalties. The reason for this is that when two predictors are highly correlated, L1 regularizer will simply pick one of the two predictors. In contrast, the L2 regularizer will keep both of them and jointly shrink the corresponding coefficients a little bit. Thus, while the L1 penalty can certainly reduce overfitting, you may also experience a loss in predictive power.

### Notes

The unconstrained convex (see [cor. 22.11]) optimization problem eq. (5.34) is not differentiable at $\beta_i = 0$ and **thus** has no closed form solution as the L2 problem ⇒ quadratic programming.

### 3.5. Sparsness Priors/L0-regularization

$$-\ln \mathbb{P}(\beta|s) = s\sum_{j=1}^d \mathbb{1}_{\beta_j\neq 0} = s\sum_{j=1}^d \cdot \begin{cases}1 & \text{if } \beta_j \neq 0\\ 0 & \text{otherwise}\end{cases} \tag{5.35}$$

⇒ measure for the number of possible non-zero dimesnions/-parameters in $\beta$.

### Advantage

- Leads always to sparse solution.
- Indicates/Explains model well as we only get a few non-zero parameters that determine/characterize the model.

### Drawback

Non-convex, non-differentiable problem ⇒ computationally difficult combinatorics.

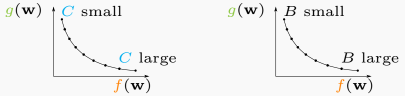
### Scalarization vs. Constrained Optimization

Their are two equivilant ways of trading:

- $g(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$: the data term and
- $f(\beta)$: the Regularizer.

**Scalariztion** $\quad \min_{\mathbf{w}} f(\beta) + Cg(\beta)$

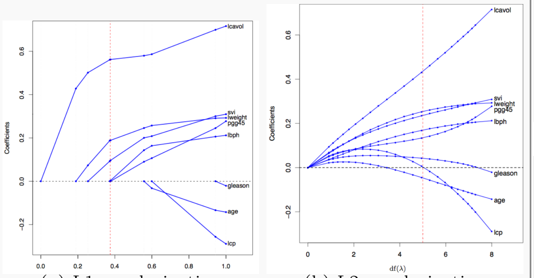
**Constraint opt.** $\quad \min_{\mathbf{w}} g(\beta)$ s.t. $f(\mathbf{w}) \leqslant B$



### Note

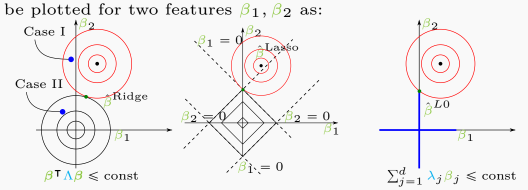
**Scalarization** and **constrained optimzization** gives the same curves $\iff$ $f, g$ are both convex functions.
**This** is not necessarily for the same values of $C$ and $B$ **but** their exisits always a relationship $C = u(B)$ s.t. this is true.

### Comparison of priors



(a) L1-regularization (b) L2-regularization

The constraint formulation of the optimization problems can be plotted for two features $\beta_1, \beta_2$ as:



Case I
Case II
$\beta_2$
$\hat\beta^{\text{Ridge}}$
$\beta_1$
$\beta^\intercal\Lambda\beta \leqslant$ const
$\beta_2$
$\beta_1 = 0$
$\hat\beta^{\text{Lasso}}$
$\beta_2 = 0$
$\beta_2 = 0$
$\beta_1$
$\beta_1 = 0$
$\beta_2$
$\hat\beta^{L0}$
$\beta_1$
$\sum_{j=1}^d \lambda_j\beta_j \leqslant$ const

- **Ridge Regression/L2-regression**: if the leasts squares error solution satisfies the constraint, we are fine (Case II), otherwise we do violated the constraint $\beta_1^2 + \beta_2^2 \leqslant$ const (Case I).
- **Lasso/L1-regression**: Here the constraint equals $|\beta_1| + |\beta| \leqslant$ const and leads to polyhedron. Most of the time we obtain a sparse solution ≙ corrner, due to the fact that corner regions increas much faster in volume, as the mixed regions (sparseness increases with number of dimensions).
- **Sparsness prior/L0-regression**: Leads to a super spiky geometry ⇒ always leads to a sparse solution.

## Liklihoods

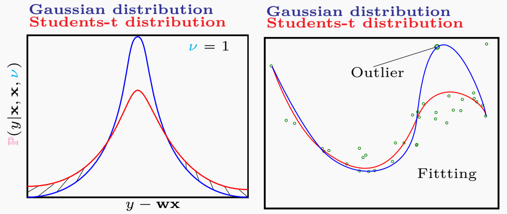
### 3.6. Student‘s-t likelihood loss function

Students-t Distribution:

$$f(y|\mathbf{x}, \mathbf{w}, \nu, \sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu\sigma^2\Gamma\left(\frac{\nu}{2}\right)}}\left(1 + \frac{(y - \mathbf{w}^\intercal\mathbf{x})^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

$\nu$: determines speed of decay.
**Problem** L2/squared loss functions lead to estimates that are sensitive to outliers, that is because something that is far away, from the expected value, will be increased/influences the model very much.

- For Gaussian noise: outliers are very unlikly and thus will have a big influence on the model.
- For Students-t noise: noise, outliers are not as unlikly as for Gaussian noise and thus will not have that much of an influence on the model.



**Speed of Decay**: $\mathbb{P}(|y - \mathbf{w}^\intercal\mathbf{x}| > t)$ probability of having a outlier/derivation of larger than t, for linear regression.

**Students-t** $\quad \mathbb{P}(|y - \mathbf{w}^\intercal\mathbf{x}| > t) = \mathcal{O}(t^{-\alpha}) \quad \alpha > 0$
(Polynomial decay)

**Gaussian** $\quad \mathbb{P}(|y - \mathbf{w}^\intercal\mathbf{x}| > t) = \mathcal{O}(\exp^{-\alpha t}) \quad \alpha > 0$
(Exponential decay)

⇒ Students-t distribution decays less fast then the Gaussian distribution and **thus** has heavier tails/tailmasses and does not get so easily influenced by noise.
**Thus** if we know that our model contains outliers/noise, we should use student's t distribution.

## 4. Proofs

**Proof 5.1** (5.11). *From eq. (5.12) it follows that the response variables are uncorrelated given the explanatory variables* Cov $[Y_i, Y_j|\mathbf{X}] = 0$. *Hence we have i.i.d. samples with a corresponding conditional (log-)likelihood given by:*

$$\begin{aligned}\mathscr{L}_n(\mathbf{y}|\mathbf{X}, \theta) &\overset{i.i.d.}{=} \prod_{i=1}^n \mathrm{p}(\mathbf{x}_i, y_i|\theta) = \prod_{i=1}^n \mathcal{N}(\beta^\intercal\mathbf{x}_i, \sigma^2)\\ &= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(y_i - \beta^\intercal\mathbf{x}_i)^2}{2\sigma^2}\right)\\ &= \left(\sigma^2 2\pi\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\intercal\mathbf{x}_i)^2\right)\end{aligned}$$

$$\mathrm{l}_n(\mathbf{y}|\mathbf{X}, \theta) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\intercal\mathbf{x}_i)^2$$

**Proof 5.2** (Definition 5.13).

$$\begin{aligned}\beta^* &\in \arg\min_{\beta\in\mathbb{R}^p} -\mathrm{l}_n(\mathbf{y}|\mathbf{X}, \theta)\\ &= \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^\intercal\mathbf{x}_i)^2\\ &= \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\intercal(\mathbf{y} - \mathbf{X}\beta)\\ &= \arg\min_{\beta\in\mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)^\intercal(\mathbf{y} - \mathbf{X}\beta)\\ &\overset{\star}{\iff} \left(-2\mathbf{y}^\intercal\mathbf{X} + 2\mathbf{X}^\intercal\mathbf{X}\beta^*\right) \overset{!}{=} 0\end{aligned}$$

$$\Rightarrow \mathbf{X}^\intercal\mathbf{X}\beta^* = \mathbf{X}^\intercal\mathbf{y}$$

add to proof and linear algrba section

**Note:** ★

$$\begin{aligned} &(\mathbf{y}-\mathbf{X}\beta)^\intercal(\mathbf{y}-\mathbf{X}\beta) \\ &= \mathbf{y}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\beta + (\mathbf{X}\beta)^\intercal\mathbf{y} - (\mathbf{X}\beta)^\intercal(\mathbf{X}\beta) \\ &= \mathbf{y}^\intercal\mathbf{y} - 2\mathbf{y}^\intercal\mathbf{X}\beta + \beta^\intercal\mathbf{X}^\intercal(\mathbf{X}\beta) \end{aligned}$$

$$\frac{\partial}{\partial\mathbf{x}}\mathbf{M}\mathbf{x} = \mathbf{M} \qquad \text{and} \qquad \frac{\partial}{\partial\mathbf{x}}\mathbf{x}^\intercal\mathbf{M}\mathbf{x} = (\mathbf{M}+\mathbf{M}^\intercal)\mathbf{x} \tag{5.36}$$

If we let $\mathbf{M} = \mathbf{X}^\intercal\mathbf{X}$ then it follows:

$$\frac{\partial}{\partial\beta}\beta^\intercal\mathbf{X}^\intercal(\mathbf{X}\beta) = (\mathbf{X}^\intercal\mathbf{X} + (\mathbf{X}^\intercal\mathbf{X})^\intercal)\beta = 2\mathbf{X}^\intercal\mathbf{X}\beta$$

**Thus**

$$0 = \frac{\partial}{\partial\beta}(\mathbf{y}-\mathbf{X}\beta)^\intercal(\mathbf{y}-\mathbf{X}\beta) = 2\mathbf{X}^\intercal(\mathbf{X}\beta-\mathbf{y}) \tag{5.37}$$

combine proofs

**Proof 5.3.** *[def. 5.12]*

$$\begin{aligned} \operatorname{lsq}(\mathbf{X},\mathbf{y}) &= (\mathbf{y}-\mathbf{X}\beta)^\intercal(\mathbf{y}-\mathbf{X}\beta) \\ &= \mathbf{y}^\intercal\mathbf{y} - \mathbf{y}^\intercal\mathbf{X}\beta + (\mathbf{X}\beta)^\intercal\mathbf{y} - (\mathbf{X}\beta)^\intercal(\mathbf{X}\beta) \\ &= \mathbf{y}^\intercal\mathbf{y} - 2\mathbf{y}^\intercal\mathbf{X}\beta + \beta^\intercal\mathbf{X}^\intercal(\mathbf{X}\beta) \end{aligned}$$

$$0 = \frac{\partial}{\partial\beta}\operatorname{lsq}(\mathbf{X},\mathbf{y}) = 2\mathbf{X}^\intercal\mathbf{X}\beta - 2\mathbf{X}^\intercal\mathbf{y} = 2\mathbf{X}^\intercal(\mathbf{X}\beta-\mathbf{y})$$

**Note**

$$\frac{\partial}{\partial\beta}\beta^\intercal\mathbf{X}^\intercal(\mathbf{X}\beta) \overset{\text{eq. }(25.130)}{=} (\mathbf{X}^\intercal\mathbf{X} + (\mathbf{X}^\intercal\mathbf{X})^\intercal)\beta = 2\mathbf{X}^\intercal\mathbf{X}\beta$$

**Proof 5.4.** *Corollary 5.2*

$$\begin{aligned} &(\mathbf{X}\beta-\mathbf{y}) && \perp \Re(\mathbf{X}) \\ \iff &(\mathbf{X}\beta)^\intercal(\mathbf{X}\beta-\mathbf{y}) = \mathbf{0} && \forall\beta\in\mathbb{R}^{m} \\ \iff &\mathbf{X}^\intercal(\mathbf{X}\beta-\mathbf{y}) = \mathbf{0} \end{aligned}$$

*where* $\mathbf{X} = \{\mathbf{x}_{:,1},\ldots,\mathbf{x}_{:,m}\}$ *is the "basis" of the Range space:*

$$(\mathbf{X}\beta-\mathbf{y})^\intercal\mathbf{x}_{:,j} = \mathbf{0} \qquad \forall j = 1,\ldots,m$$

## 5. Examples

**Example 5.1 Simple Linear Regression:**

$$p = 2 \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

**Example 5.2 Simple Linear Quadratic Regression:**

$$p = 3 \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_2 \end{pmatrix}$$

# Classification

## 6. Intro

**Definition 5.18 Training Data** $\mathcal{D}$:

$$\mathcal{D} := \left\{(\mathbf{x}_i, y_i) \quad \middle| \quad \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathcal{Y} := \{c_1, \dots, c_K\}\right\}$$

**Definition 5.19 Classifier** $c$:
Is a mapping that maps the features into classes:

$$c : \mathcal{X} \to \mathcal{Y} \tag{5.38}$$

### 6.1. Types of Classification

**Definition 5.20 Dichotomy:**
Given a set $\mathcal{S} = \{s_1, \dots, s_N\}$ a dichotomy is partition of the set $\mathcal{S}$ into two subsets $A$, $A^c$ that satisfy:

- *Collectively/jointly exhaustiveness*:

$$S = A \cup A^C \tag{5.39}$$

- *Mutual exclusivity*:

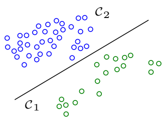
$$s \in A \qquad \Longrightarrow \qquad s \notin A^C \qquad \forall s \in S \tag{5.40}$$

**Explanation 5.1** (title). *Nothing can belong simultaneously to both parts $A$ and $A^c$.*

**Definition 5.21 Binaray Classification:**

Is a classification problem where the labels are binary:

$$\mathcal{Y} = \{c_1, c_2\} = \{-1, 1\} \tag{5.41}$$



multiclass

### 6.2. Encodings

#### 6.2.1. One Hot Encoding

**Definition 5.22 One-hot encoding/representation:**
Is the representation/encoding of the $K$ categories $\{c_1, \dots, c_K\}\}$ by a *sparse vectors*[def. 25.66] with one non-zero entry, where the index $j$ of the non-zero entry indicates the class $c_j$:

$$\mathbb{B}^n = \left\{\mathbf{y} \in \{0,1\}^n : \mathbf{y}^\intercal\mathbf{y} = \sum_{i=1}^n \mathbf{y} = 1\right\}$$

s.t. $\qquad \mathbf{y}_i = \mathbf{e}_j \qquad \Longleftrightarrow \qquad \mathbf{y}_i = c_j$

#### 6.2.2. Soft vs. Hard Labels

**Definition 5.23 Hard Labels/Targets:** Are observations $y \in \mathcal{Y}$ that are consider as true observations. We can encode them using a one hot encoding[def. 5.22]:

$$y = c_k \qquad \Longrightarrow \qquad y = \mathbf{e}_k \tag{5.42}$$

**Definition 5.24 Soft Labels/Targets:** Are observations $y \in \mathcal{Y}$ that are consider as noisy observations or probabilities $\mathrm{p}$. We can encode them using a probabilistic vector[def. 25.67]:

$$y = [\mathrm{p}_1 \cdots\cdots\cdots \mathrm{p}_K]^\intercal \tag{5.43}$$

**Corollary 5.5 Hard labels as special case:** If we consider hard targets[def. 5.23] as events with probability one then we can think of them as a special case of the soft labels.

## 7. Binary Classification $\{-1, 1\}$

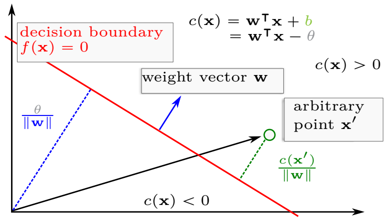
### 7.1. Linear Classification

**Definition 5.25 Linear Dichotomy:**

**Definition 5.26 Linear Classifier:** A linear classifier is a classifier $c$ that assigns labels $\hat{y}$ to samples $\mathbf{x}_i$ using a *linear decision boundary/hyperplane*[def. 25.12]:

$$\hat{y} = c(\mathbf{x}_i) = \begin{cases} c_1 \in \mathcal{H}^+ & \text{if } \mathbf{w}^\intercal\mathbf{x} > \theta \\ c_2 \in \mathcal{H}^- & \text{if } \mathbf{w}^\intercal\mathbf{x} < \theta \end{cases} \tag{5.44}$$

**Explanation 5.2** (Definition 5.26).


decision boundary
$f(\mathbf{x}) = 0$
$c(\mathbf{x}) = \mathbf{w}^\intercal\mathbf{x} + b$
$= \mathbf{w}^\intercal\mathbf{x} - \theta$
$c(\mathbf{x}) > 0$
weight vector $\mathbf{w}$
$\frac{\theta}{\|\mathbf{w}\|}$
arbitrary point $\mathbf{x}'$
$\frac{c(\mathbf{x}')}{\|\mathbf{w}\|}$
$c(\mathbf{x}) < 0$

- The $b \in \mathbb{R}$ corresponds to the offset of the decision surface from the origin, otherwise the decision surface would have to pass through the origin.
- $\mathbf{w} \in \mathbb{R}^d$ is the normal unit vector of the decision surface. Its components $\{w_j\}_{j=1}^d$ correspond to the importance of each feature/dimension.

**Explanation 5.3** (Threshold $\theta$ vs. Bias $b$). *The offset is called bias if it is considered as part of the classifier $\mathbf{w}^\intercal\mathbf{x} + b$ and as threshold if it is considered to be part of the hyperplane $\theta = -b$, but its just a matter of definition.*

**Definition 5.27 (Normalized) Classification Criterion:**

$$\tilde{\mathbf{w}}^\intercal\mathbf{x} = \mathbf{w}^\intercal\mathbf{x}y > 0 \qquad \forall\,(\mathbf{x}, y) \in \mathcal{D} \tag{5.45}$$

**Definition 5.28 Linear Separable Data set:** A data set is *linearly separable* if there exists a separating hyperplane $\mathcal{H}$ s.t. each label can be assigned correctly:

$$\hat{y} := c(\mathbf{x}) = y \qquad \forall(\mathbf{x}, \mathbf{y}) \in \mathcal{D} \tag{5.46}$$

#### 7.1.1. Normalization

**Proposition 5.3 Including the Offset:** In order to simplify notation the offset is usually included into the parameter vector:

$$\mathbf{w} \leftarrow \begin{pmatrix}\mathbf{w}\\ b\end{pmatrix} \qquad \mathbf{x} \leftarrow \begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix}$$

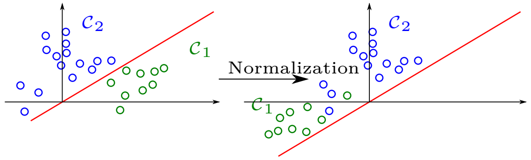
$$\Rightarrow \qquad \mathbf{w}^\intercal\mathbf{x} = \begin{pmatrix}\mathbf{w}^\intercal & b\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix} = \mathbf{w}^\intercal\mathbf{x} + b$$

**Proposition 5.4 Uniform Classification Criterion:**
In order to avoid the case distinction in the classification criterion of eq. (5.44) we may transform the input samples by:

$$\tilde{\mathbf{x}} = \begin{cases} \mathbf{x} & \text{if } \mathbf{w}^\intercal\mathbf{x} > \theta \\ -\mathbf{x} & \text{if } \mathbf{w}^\intercal\mathbf{x} < \theta \end{cases} \tag{5.47}$$

**Explanation 5.4** (proposition 5.4).

We transform the input s.t. the separating hyper-plane puts all labels on the same "positive" side $\mathbf{w}^\intercal\mathbf{x} > 0$.

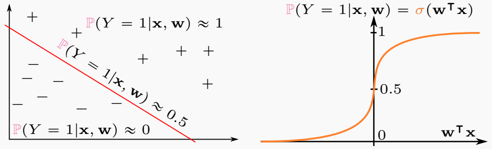


**Corollary 5.6 :** How can we achieve this in practice?
If $\mathcal{Y} = \{-1, 1\}$ then we can simply multiply with the label $y_i$:

$$\left.\begin{aligned} \mathbf{w}^\intercal\mathbf{x} > 0 \qquad \forall y = +1 \\ \mathbf{w}^\intercal\mathbf{x} < 0 \qquad \forall y = -1 \end{aligned}\right\} \qquad \Longleftrightarrow \qquad \mathbf{w}^\intercal\mathbf{x} \cdot \hat{y} > 0 \qquad \forall y$$

## 8. Logistic Regression $\mathrm{Bern}(y; \sigma(\mathbf{w}^\intercal\mathbf{x}, \sigma^2))$

**Idea**: in order to classify dichotomies[def. 5.20] we use a distribution that maps probabilities to a binary values 0/1
$\Rightarrow$ *Bernoulli Distribution*[def. 32.22].
**Problem**: we need to convert/translate distance $\mathbf{w}^\intercal\mathbf{x}$ into probability in order to use a bernouli distribution.
**Idea**: use a sigmoidal function to convert distances $\mathbf{z} := \mathbf{w}^\intercal\mathbf{x}$ into probabilities $\Rightarrow$ *Logistic Function*[def. 5.29].



### 8.1. Logistic Function

**Definition 5.29 Sigmoid/Logistic Function:**

$$\sigma(z) = \frac{1}{1 + \mathrm{e}^{-z}} = \frac{1}{1 + \mathrm{e}^{\text{neg. dist. from deci. boundary}}} \tag{5.48}$$

**Explanation 5.5** (Sigmoid/Logistic Function).

$$\sigma(z) = \begin{cases} 0 & & -z \textit{ large} \\ 1 & \textit{if} & z \textit{ large} \\ 0.5 & & z = 0 \end{cases}$$

### 8.2. Logistic Regression

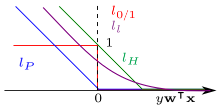
**Definition 5.30 Logistic Regression:**
models the likelihood of the output $y$ as a Bernoulli Distribution[def. 32.22] $y \sim \mathrm{Bern}(\mathrm{p})$, where the probability $\mathrm{p}$ is given by the Sigmoid function[def. 5.29] of a linear regression:

$$\mathrm{p}(y|\mathbf{x}, \mathbf{w}) = \mathrm{Bern}\left(\sigma(\mathbf{w}^\intercal\mathbf{x})\right) = \begin{cases} \frac{1}{1+\mathrm{e}^{-\mathbf{w}^\intercal\mathbf{x}}} & \text{if } y = +1 \\ 1 - \frac{1}{1+\mathrm{e}^{-\mathbf{w}^\intercal\mathbf{x}}} & \text{if } y = -1 \end{cases}$$

$$\overset{??\ 5.5}{=} \frac{1}{1 + \exp(-y \cdot \mathbf{w}^\intercal\mathbf{x})} = \sigma\left(-y \cdot \mathbf{w}^\intercal\mathbf{x}\right) \tag{5.49}$$

#### 8.2.1. Maximum Likelihood Estimate

**Definition 5.31 Logistic Loss $l_l$** proof 5.6:
Is the objective we want to minimize when performing mle[def. 6.3] for a logistic regression likelihood and incurs higher cost for samples closer to the decision boundary:



$$l_l(\mathbf{w}; \mathbf{x}, y) := \log\left(1 + \exp(-y \cdot \mathbf{w}^\intercal\mathbf{x})\right) \tag{5.50}$$

$$\propto \log(1 + \mathrm{e}^z) = \begin{cases} z & \text{for large } z \\ 0 & \text{for small } z \end{cases}$$

**Corollary 5.7 MLE for Logistic Regression:**

$$\mathrm{l}_n(\mathbf{w}) = \sum_{i=1}^n l_l = \sum_{i=1}^n \log\left(1 + \exp(-y_i \cdot \mathbf{w}^\intercal\mathbf{x}_i)\right) \tag{5.51}$$

**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 [cor. 5.7].

**Definition 5.32 Logistic Loss Gradient** proof 5.7 $\nabla_{\mathbf{w}} l_l(\mathbf{w})$:

$$\nabla_{\mathbf{w}} l_l(\mathbf{w}) = \mathbb{P}(Y = -y|\mathbf{x}, \mathbf{w}) \cdot (-y\mathbf{x})$$

$$= \frac{1}{1 + \exp(y\mathbf{w}^\intercal\mathbf{x})} \cdot (-y\mathbf{x}) \tag{5.52}$$

**Explanation 5.6.**

$$\nabla_{\mathbf{w}} l_l(\mathbf{w}) = \mathbb{P}(Y = -y|\mathbf{x}, \mathbf{w}) \cdot (-y\mathbf{x}) \propto \nabla_{w} l_H(\mathbf{w})$$

*The logistic loss $l_l$ is equal to the hinge loss $l_h$ but weighted by the probability of beeing in the wrong class $\mathbb{P}(Y = -1|\mathbf{x}, \mathbf{w})$. Thus the more likely we are in the wrong class the bigger the step we take:*

$$\mathbb{P}(Y = -y|\hat{y} = \mathbf{w}^\intercal\mathbf{x}) = \begin{cases} \uparrow & \textit{take big step} \\ \downarrow & \textit{take small step} \end{cases}$$

**Algorithm 5.1 Vanilla SGD for Logistic Regression:**

```
    Initalize: $\mathbf{w}$
for $1, 2 \dots, T$ do
    Pick $(\mathbf{x}, y)$ unif. at randomn from data $\mathcal{D}$
    $\mathbb{P}(Y = -y|\mathbf{x}, \mathbf{w}) = \frac{1}{(1+\exp(-y\cdot\mathbf{w}^\intercal\mathbf{x}))} = \sigma(y \cdot \mathbf{w}^\intercal\mathbf{x})$
            ▷ compute prob. of misclassif. with cur. model
    $\mathbf{w} = \mathbf{w} + \eta_t y\mathbf{x}\sigma(y \cdot \mathbf{w}^\intercal\mathbf{x})$
end for
```

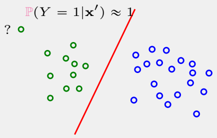
**Making Predictions**

Given an optimal parameter vector $\hat{\mathbf{w}}$ found by algorithm 5.1 we can predict the output of a new label by eq. (5.49):

$$\mathbb{P}(y|\mathbf{x}, \hat{\mathbf{w}}) = \frac{1}{1 + \exp\left(-y\hat{\mathbf{w}}^\intercal\mathbf{x}\right)} \tag{5.53}$$

**Drawback**

Logistic regression, does not tell us anything about the liklihood $\mathrm{p}(\mathbf{x})$ of a point, thus it will not be able to detect outliers, as it will assign a very high probability to all correctly classfied points, far from the decsion boundary.



#### 8.2.2. Maximum a-Posteriori Estimates

adapt

### 8.3. Logistic regression and regularization

**Adding Priors to Logistic Liklihood**

- **L2 (Gaussian prior):**

$$\arg\min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp(-y_i\mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda\|\mathbf{w}\|_2^2$$

- **L1 (Laplace prior):**

$$\arg\min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp(-y_i\mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda\|\mathbf{w}\|_1$$

- **Generalized:**

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^n \log\left(1 + \exp(-y_i\mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda C(\mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \mathbb{P}(\mathbf{w}|\mathbf{X}, Y)$$

### 8.4. SGD for L2-gregularized logistic regression

```
    Initalize: $\mathbf{w}$
for $1, 2 \dots, T$ do
    Pick $(\mathbf{x}, y)$ unif. at randomn from data $\mathcal{D}$
    $\hat{\mathbb{P}}(Y = -y|\mathbf{x}, \mathbf{w}) = \frac{1}{(1+\exp(-y\cdot\mathbf{w}^\intercal\mathbf{x}))}$
            ▷ compute prob. of misclassif. with cur. model
    $\mathbf{w} = \mathbf{w}(1 - 2\lambda\eta_t) + \eta_t y\mathbf{x}\hat{\mathbb{P}}(Y = -y|\mathbf{x}, \mathbf{w})$
end for
```

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

## 9. Proofs

**Proof 5.5.** *[def. 5.30] We need to only proof the second expression, as the first one is fulfilled anyway:*

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

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

**Proof 5.6.** *[def. 5.31]*

$$
\begin{aligned}
\mathrm{l}_n(\mathbf{w}) &= \arg\max_{\mathbf{w}} \mathrm{p}(y_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}) = \arg\min_{\mathbf{w}} -\log \mathrm{p}(Y|\mathbf{X}, \mathbf{w}) \\
&\overset{i.i.d.}{=} \arg\min_{\mathbf{w}} \sum_{i=1}^{n} -\log \mathrm{p}(y_i|\mathbf{x}_i, \mathbf{w}) \\
&\overset{eq.\ (5.49)}{=} -\log \frac{1}{1+\exp(-y_i \cdot \mathbf{w}^\intercal \mathbf{x}_i)} \\
&= \log\left(1+\exp(-y_i \cdot \mathbf{w}^\intercal \mathbf{x}_i)\right) =: l_l(\mathbf{w})
\end{aligned}
$$

**Proof 5.7.** *[def. 5.32]*

$$
\begin{aligned}
\nabla_{\mathbf{w}} l_l(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \log\left(1+\exp(-y \cdot \mathbf{w}^\intercal \mathbf{x})\right) \\
&\overset{C.R.}{=} \frac{1}{(1+\exp(-y \cdot \mathbf{w}^\intercal \mathbf{x}))} \frac{\partial}{\partial \mathbf{w}} \left(1+\exp(-y \cdot \mathbf{w}^\intercal \mathbf{x})\right) \\
&\overset{C.R.}{=} \frac{1}{(1+\exp(-y \cdot \mathbf{w}^\intercal \mathbf{x}))} \exp(-y \cdot \mathbf{w}^\intercal \mathbf{x}) \cdot (-y\mathbf{x}) \\
&= \frac{e^{-z} \cdot (-yx)}{(1+e^{-z})} = \frac{-yx}{e^{z}(1+e^{-z})} = \frac{-yx}{(e^{z}+e^{-z+z})} \\
&= \frac{1}{\exp(y \cdot \mathbf{w}^\intercal \mathbf{x})+1} \cdot (-yx) \\
&\overset{eq.\ (5.49)}{=} \hat{\mathbb{P}}(Y=-y|\mathbf{x}, \mathbf{w}) \cdot (-y\mathbf{x})
\end{aligned}
$$

# Model Parameter Estimation

## 1. Maximum Likelihood Estimation

### 1.1. Likelihood Function

Is a method for estimating the parameters $\theta$ of a model that agree best with observed data $\{x_1,\dots,x_n\}$. **Let**: $\theta = (\theta_1 \ \dots \ \theta_k)^\intercal \in \Theta\,\mathbb{R}^k$ vector of unknown model parameters.
**Consider**: a probability density/mass function $f_X(\mathbf{x};\theta)$

**Definition 6.1 Likelihood Function** $\mathscr{L}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}_+$**:**
Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ be a random sample of i.i.d. data points drawn from an unknown probability distribution $\mathbf{x}_i \sim p_{\mathcal{X}}$.
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 $\theta$:

$$\mathscr{L}_n(\theta|\mathbf{X}) = \mathscr{L}_n(\theta;\mathbf{X}) = f(\mathbf{X}|\theta) = f(\mathbf{X};\theta) \tag{6.1}$$

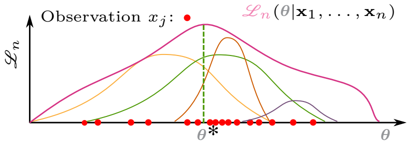


Figure 6: Possible Likelihood function in pink.
Overlayed: possible candidate functions for Gaussian model explaining the observations.

**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 6.1 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:

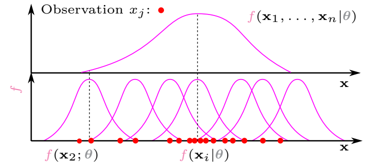


Figure 7: Bottom: probability distributions of the different data points $\mathbf{x}_i$ given a fixed $\theta$ for a Gaussian distribution
Top: joint probability distribution of the i.i.d. data points $\{\mathbf{x}_i\}_{i=1}^n$ given a fixed $\theta$

$$f(\mathbf{x}_1,\dots,\mathbf{x}_n|\,\theta) \overset{\text{i.i.d.}}{=} \prod_{i=1}^n f(\mathbf{x}_i|\theta)$$

**Notation**

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

### 1.2. Maximum Likelihood Estimation (MLE)

Let $f_\theta(\mathbf{x})$ be the probability of an i.i.d. sample $\mathbf{x}$ for a given model.
Goal: find $\theta$ of a given model that maximizes the joint probability/likelihood of the observed data $\{x_1,\dots,x_n\}$? $\iff$ maximum likelihood estimator $\theta^*$.

**Definition 6.2 Log Likelihood Function** $\mathrm{l}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}$**:**

$$\mathrm{l}_n(\theta|\mathbf{X}) = \log \mathscr{L}_n(\theta|\mathbf{X}) = \log f(\mathbf{X}|\theta) \tag{6.2}$$

**Corollary 6.2 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. (22.62):

$$\log f(\mathbf{x}_1,\dots,\mathbf{x}_n|\,\theta) \overset{\text{i.i.d.}}{=} \log\left(\prod_{i=1}^n f(\mathbf{x}_i|\theta)\right) = \sum_{i=1}^n \log f(\mathbf{x}_i|\theta)$$

**Definition 6.3 Maximum Likelihood Estimator** $\theta^*$**:**
Is the estimator $\theta^* \in \Theta$ that maximizes the likelihood of the model/predictor:

$$\theta^* = \arg\max_{\theta\in\Theta} \mathscr{L}_n(\theta;\mathbf{x}) \quad \text{or} \quad \theta^* = \arg\max_{\theta\in\Theta} \mathrm{l}_n(\theta;\mathbf{x}) \tag{6.3}$$

### 1.3. Maximization vs. Minimization

For optimization problems we minimize by convention.
The logarithm is a concave function[def. 22.26] $\cap$, thus if we calculate the extremal point we will obtain a maximum.
If we want to calculate a mimimum instead (i.e. in order to be compatible with some computer algorithm) we can convert the function into a convex functionsection 4 $\cup$ by multiplying it by minus one and consider it as a loss function instead of a likelihood.

**Definition 6.4 Negative Log-likelihood** $-\mathrm{l}_n(\theta|\mathbf{X})$**:**

$$\theta^* = \arg\max_{\theta\in\Theta} \mathrm{l}_n(\theta|\mathbf{X}) = \arg\min_{\theta\in\Theta} -\mathrm{l}_n(\theta|\mathbf{X}) \tag{6.4}$$

### 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 a unknown true conditional probability distribution $f_{Y|X}$ and we are given a data set $\mathbf{Z} = \left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\right\}_{i=1}^n$.
Now we want to find *the parameters* $\theta = (\theta_1 \ \dots \ \theta_k)^\intercal \in \Theta\,\mathbb{R}^k$ of a hypothesis $\hat{f}_{Y|X}$ that agree best with the given data $\mathcal{Z}$.

**Note**

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

**Definition 6.5 Conditional (log) likelihood function:**
Models the liklihood of a model with parameters $\theta$ given the data $\mathbf{Z} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

$$\mathscr{L}_n(\theta|Y,\mathbf{X}) = \mathscr{L}_n(\theta;Y,\mathbf{X}) = f(Y|\mathbf{X},\theta) = f(Y|\mathbf{X};\theta)$$

## 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?
⇒ use Bayes rule (**??**) to introduce a bias into our model via. a Prior distribution.

### 2.1. Prior Distribution

**Definition 6.6 Prior (Distribution)** $\pi(\theta) = p(\theta)$**:**
**Assumes**: that the model parameters $\theta$ 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:

$$\theta \sim \pi(\theta) = p(\theta) \tag{6.5}$$

**Notes**

In this section we use the terms model parameters $\theta$ and model as synonymous, as the model is fully described by its population parameters ([def. 4.18]) $\theta$.

**Corollary 6.3 The prior is independent of the data:** The prior $p(\theta)$ models a prior belief/bias and is thus independent of the data $\mathcal{D} = \{\mathcal{X},\mathcal{Y}\}$:

$$p(\theta|\mathbf{X}) = p(\theta) \tag{6.6}$$

**Definition 6.7 Hyperparameters** $p_\lambda(\theta)$**:**
In most cases the prior distribution are parameterized that is the pdf $\pi(\theta|\boldsymbol{\lambda})$ depends on a set of parameters $\boldsymbol{\lambda}$.
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 6.1

### 2.2. Posterior Distribution

**Definition 6.8 Posterior Distribution** $p(\theta|\mathbf{data})$**:**
The posterior distribution $p(\theta|\text{data})$ is a probability distribution that describes the relationship of a unknown *parameter* $\theta$ a posterior/after observing evidence of a random quantity $\mathbf{Z}$ that is in a relation with $\theta$:

$$p(\theta|\mathbf{data}) = p(\theta|\mathbf{Z}) \tag{6.7}$$

**Definition 6.9**
**Posterior Distribution and Bayes Theorem:**
Using Bayes theorem 31.3 we can write the posterior distribution as a product of the *likelihood*[def. 6.1] weighted with our *prior*[def. 6.6] and normalized by the *evidence* $\mathbf{Z} = \{\mathbf{X},\mathbf{y}\}$ s.t. we obtain a real probability distribution:

$$p(\theta|\mathbf{data}) = p(\theta|\mathbf{Z}) = \frac{p(\mathbf{Z}|\theta)\cdot p_\lambda(\theta)}{p(\mathbf{Z})} \tag{6.8}$$

$$\text{Posterior} = \frac{\text{Liklihood}\cdot\text{Prior}}{\text{Normalization}} \tag{6.9}$$

$$\boxed{p(\theta|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\theta,\mathbf{X})\cdot p_\lambda(\theta)}{p(\mathbf{y}|\mathbf{X})}} \tag{6.10}$$

see proof **??** 6.1

#### 2.2.1. Maximization –MAP

We do not care about the full posterior probability distribution as in Bayesian Inference (section 4). We only want to find a point estimator **??** $\theta^*$ that maximizes the posterior distribution.

#### 2.2.2. Maximization

**Definition 6.10**
**Maximum a-Posteriori Estimates (MAP):**
Is model/parameters $\theta$ that maximize the posterior probability distribution:

$$\boxed{\theta^*_{\text{MAP}} = \arg\max_\theta \mathbb{P}(\theta|\mathbf{X},\mathbf{y})} \tag{6.11}$$

**Log-MAP estimator:**

$$\begin{aligned} \theta^* &= \arg\max_\theta \{p(\theta|\mathbf{X},\mathbf{y})\} \qquad (6.12)\\ &= \arg\max_\theta \left\{\frac{p(\mathbf{y}|\mathbf{X},\theta)\cdot p_\lambda(\theta)}{p(\mathbf{y}|\mathbf{X})}\right\}\\ &\overset{\text{eq. (22.59)}}{\propto} \arg\max_\theta \left\{p(\mathbf{y}|\theta,\mathbf{X})\cdot p_\lambda(\theta)\right\} \end{aligned}$$

**Corollary 6.4 Negative Log MAP:**

$$\begin{aligned} \theta^* &= \arg\max_\theta \{p(\theta|\mathbf{X},\mathbf{y})\} \qquad (6.13)\\ &= \arg\min_\theta -\log \overbrace{p(\theta)}^{\text{Prior}} - \log \overbrace{p(\mathbf{y}|\theta,\mathbf{X})}^{\text{Likelihood}} + \underbrace{\cancel{\log p(\mathbf{y}|\mathbf{X})}}_{\text{not depending on }\theta} \end{aligned}$$

## 3. Proofs

**Proof 6.1.** *6.10:*

$$p(\mathbf{X},\mathbf{y},\theta) = \begin{cases} p(\theta|\mathbf{X},\mathbf{y})p(\mathbf{X},\mathbf{y}) \\ p(\mathbf{y}|\mathbf{X},\theta)p(\mathbf{X},\theta) \end{cases}$$

$$\begin{aligned} p(\theta|\mathbf{X},\mathbf{y})p(\mathbf{X},\mathbf{y}) &= p(\theta|\mathbf{X},\mathbf{y})p(\mathbf{y}|\mathbf{X})p(\mathbf{X})\\ p(\mathbf{y}|\mathbf{X},\theta)p(\mathbf{X},\theta) &= p(\mathbf{y}|\mathbf{X},\theta)p(\mathbf{X},\theta)\\ &= p(\mathbf{y}|\mathbf{X},\theta)p(\theta|\mathbf{X})p(\mathbf{X})\\ &\overset{eq.\ (6.6)}{=} p(\mathbf{y}|\mathbf{X},\theta)p(\theta)p(\mathbf{X}) \end{aligned}$$

$$\Rightarrow p(\theta|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\theta)p(\theta)\cancel{p(\mathbf{X})}}{p(\mathbf{y}|\mathbf{X})\cancel{p(\mathbf{X})}}$$

**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 6.1 Hyperparameters Gaussian Prior:**

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

with the hyperparameter $\boldsymbol{\lambda} = (\mu \ \ \sigma^2)^\intercal$.

## Bayesian Inference/Modeling

**Definition 6.11 Bayesian Inference:** So far we only really looked at point estimators/estimates[def. 34.8].
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. 4.16], where uncertainty in inferences is quantified using probability. Thus we usually obtain a distribution over our parameters and not a single point estimates
⇒ can deduce statistical properties of parameters from their distributions.

**Definition 6.12** $\mathrm{p}(\mathrm{w}|\mathbf{y},\mathbf{X})/\mathrm{p}(\mathrm{w}|\mathcal{D})$
**Posterior Probability Distribution:**
1. Specify the prior $\mathrm{p}_\lambda(\mathrm{w})$
2. Specify the likelihood $\mathrm{p}(\mathbf{y}|\mathrm{w},\mathbf{X})/\mathrm{p}(\mathcal{D}|\mathrm{w})$
3. Calculate the evidence $\mathrm{p}(\mathbf{y}|\mathbf{X})/\mathrm{p}(\mathcal{D})$
4. Calculate the posterior distribution $\mathbb{P}(\mathrm{w}|\mathbf{y},\mathbf{X})/\mathrm{p}(\mathrm{w}|\mathcal{D})$

$$\mathrm{p}(\mathrm{w}|\mathbf{y},\mathbf{X}) = \frac{\mathrm{p}(\mathbf{y}|\mathrm{w},\mathbf{X})\cdot \mathrm{p}_\lambda(\mathrm{w})}{\mathrm{p}(\mathbf{y}|\mathbf{X})} = \frac{\text{Liklihood}\cdot\text{Prior}}{\text{Normalization}}$$

**Definition 6.13** $\mathrm{p}(\mathbf{y}|\mathbf{X})/\mathrm{p}(\mathcal{D})$
**Marginal Likelihood** [see proof 10.2]:
is the normalization constant that makes sure that the posterior distribution[def. 6.12] is an true probability distribution:

$$\mathrm{p}(\mathbf{y}|\mathbf{X}) = \int \mathrm{p}(\mathbf{y}|\mathrm{w},\mathbf{X})\cdot \mathrm{p}_\lambda(\mathrm{w})\,\mathrm{d}\mathrm{w} = \int \text{Likelihood}\cdot\text{Prior}\,\mathrm{d}\mathrm{w} \tag{6.14}$$

**Note**

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

**Definition 6.14** $\mathrm{p}(\mathbf{f}_*|\mathbf{x}_*,\mathbf{X},\mathbf{y})/\mathrm{p}(\mathbf{f}_*|\mathbf{y})$ [see proof 10.1]
**Posterior Predictive Distribution:**
is the distribution of a real process $\mathbf{f}$ (i.e. $f(x)=\mathbf{x}^\intercal\mathrm{w}$) given:
- new observation(s) $\mathbf{x}_*$
- the posterior distribution[def. 6.12] of the observed data $\mathcal{D}=\{\mathbf{X},\mathbf{y}\}$
- The likelihood of a real process $\mathbf{f}_*$

$$\mathrm{p}(\mathbf{f}_*|\mathbf{x}_*,\mathbf{X},\mathbf{y}) = \int \mathrm{p}(\mathbf{f}_*|\mathbf{x}_*,\mathrm{w})\cdot \mathrm{p}(\mathrm{w}|\mathbf{X},\mathbf{y})\,\mathrm{d}\mathrm{w} \tag{6.15}$$

it is calculated by weighting the likelihood[def. 6.1] of the new observation $\mathbf{x}_*$ with the posterior of the observed data and averaging over all parameter values $\mathrm{w}$.
⇒ obtain a distribution not depending on $\mathrm{w}$.

**Note f vs. y**

- Usually $\mathbf{f}$ denotes the model i.e.:
  $$\mathbf{f}(\mathbf{x}) = \mathbf{x}^\intercal\mathrm{w} \qquad \text{or} \qquad \mathbf{f}(\mathbf{x}) = \phi(\mathbf{x})^\intercal\mathrm{w}$$
  and $\mathbf{y}$ the model plus the noise $\mathbf{y}=\mathbf{f}(\mathbf{x})+\epsilon$.
- Sometime people also write only: $\mathrm{p}(\mathbf{y}_*|\mathbf{x}_*,\mathbf{X},\mathbf{y})$

### 5. Types of Uncertainty

**Definition 6.15 Epistemic/Systematic Uncertainty:**
Is the uncertainty that is due to things that one could in principle know but does not i.e. only having a finite sub sample of the data. The epistemic noise will decrease the more data we have.

**Definition 6.16 Aleatoric/Statistical Uncertainty:**
Is the uncertainty of an underlying random process/model.
The aleatroic uncertainty stems from the fact that we are create random process models. If we run our *trained* model multiple times with *the same* input $\mathbf{X}$ 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

**Definition 7.1**
**Recursive Bayesian Estimation/Filtering:** Is a technique for estimating the an unknown probability distribution recursively over time by a measurement-[def. 7.3] and a process-model[def. 7.2] using Bayesian inference[def. 6.11].

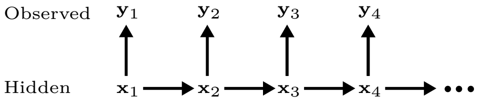


Figure 8: This problem corresponds to a *hidden Markov model (HMM)*[def. 14.1]

$$\mathbf{x}_t = (x_{t,1} \quad \cdots \quad x_{t,n}) \qquad \mathbf{y}_t = (y_{t,1} \quad \cdots \quad y_{t,m})$$

**Note**

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

**Definition 7.2** $\mathbf{x}_{t+1} \sim \mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1})$
**Process/Motion/Dynamic Model:** is a model $q$ of how our system state $\mathbf{x}_t$ evolves and is usually fraught with some uncertainty.

**Corollary 7.1 Markov Property** $\mathbf{x}_t \perp\!\!\!\perp \mathbf{x}_{1:t-2}|\mathbf{x}_{t-1}$**:** The process models[def. 7.2] is Markovian[def. 35.14] i.e. the current state depends only on the previous state:

$$\mathrm{p}(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = \mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{7.1}$$

**Definition 7.3** $\mathbf{y}_t \sim \underline{\mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)}$
**Measurement/Sensor–Model/Likelihood:** is a model $h$ that maps observations/sensor measurements of our model $\mathbf{y}_t$ to the model state $\mathbf{x}_t$

**Corollary 7.2** $\mathbf{y}_t \perp\!\!\!\perp \mathbf{y}_{1:t-1}\mathbf{x}_{1:t-1}|\mathbf{x}_t$
**Conditional Independent Measurements:** The measurements $\mathbf{y}_t$ are conditionally independent of the previous observations $\mathbf{y}_{1:t-1}$ given the current state $\mathbf{x}_t$:

$$\mathrm{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1},\mathbf{x}_t) = \mathrm{p}(\mathbf{y}_t|\mathbf{x}_t) \tag{7.2}$$

**Goal**

We want to combine the process model[def. 7.2] and the measurement model[def. 7.3] in a recursive way to obtain a good estimate of our model state:

$$\left.\begin{matrix}\mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1}) \\ \mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)\end{matrix}\right\} \mathrm{p}(\mathbf{x}_t|y_{1:t}) \xrightarrow[\text{recursion rule}]{\mathrm{p}(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} \mathrm{p}(\mathbf{x}_{t+1}|y_{1:t+1})$$

**Definition 7.4 Chapman-Kolmogorov eq.** $\underline{\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1})}$
**Prior Update/Prediction Step** [proof 10.3]:

$$\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int \mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1})\mathrm{p}(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_{t-1} \tag{7.3}$$

**Prior Distribution:**

$$\mathrm{p}(\mathbf{x}_0|\mathbf{y}_{0-1}) = \mathrm{p}(\mathbf{x}_0) = \mathrm{p}_0 \tag{7.4}$$

**Definition 7.5** $\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t})$
**Posterior Distribution/Update Step** [proof 10.4]:

$$\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{Z_t}\underline{\mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)}\,\underline{\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1})} \tag{7.5}$$

**Definition 7.6 Normalization** [see proof 10.5]:

$$Z_t = \mathrm{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int \mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_t \tag{7.6}$$

**Algorithm 7.1 Optimal Bayesian Filtering:**

```
1: Input:    $\mathrm{p}(\mathbf{x}_0)$
2: while Stopping Criterion not full-filed do
3:     Prediction Step:
           $\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{Z_t}\mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1})$
4:     Update Step:
           $\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int \mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1})\mathrm{p}(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_{t-1}$
           with:
           $Z_t = \int \mathrm{p}(\mathbf{y}_t|\mathbf{x}_t)\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_t$
5: end while
```

**Corollary 7.3** [proof 10.6]
**Joint Probability Distribution of (HMM):** we can also calculate the joint probability distribution of the (HMM):

$$\mathrm{p}(\mathbf{x}_{1:t},\mathbf{y}_{1:t}) = \mathrm{p}(\mathbf{x}_1)\mathrm{p}(\mathbf{y}_1|\mathbf{x}_1)\prod_{i=2}^{t}\mathrm{p}(\mathbf{x}_i|\mathbf{x}_{i-1})\mathrm{p}(\mathbf{y}_i|\mathbf{x}_i) \tag{7.7}$$

**Example 7.1 Types of Bayesian Filtering:**
- **Kalman Filter**: assumes a *linear* system, $q$, $h$ are linear and Gaussian noise $\mathbf{v},\mathbf{w}$.
- **Extended Kalman Filter**: assumes a *non-linear* system, $q$, $h$ are non-linear and Gaussian noise $\mathbf{v},\mathbf{w}$.
- **Particle Filter**: assumes a *non-linear* system $q$, $h$ are non-linear and Non-Gaussian noise $\mathbf{v},\mathbf{w}$, especially multi-modal distributions.

### 1. Kalman Filters

**Definition 7.7 Kalman Filter Assumptions:** Assumes a *linear*[def. 22.17] process model[def. 7.2], $\mathbf{q}$ with Gaussian model-noise $\mathbf{v}$ and a linear measurement model[def. 7.3] $\mathbf{h}$ with Gaussian process-noise $\mathbf{w}$.

**Definition 7.8 Kalman Filter Model:**
**Process Model** (7.8)

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

$$\mathbf{x}^{(0)} \sim \mathcal{N}(\mathbf{x}_0, P_0) \qquad \text{and} \qquad \mathbf{v}^{(k)} \sim \mathcal{N}(0, Q^{(k)})$$

**Measuremnt Model** (7.9)

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

**and** define:

$$\hat{x}_p^{(k)} := \mathbb{E}[\mathbf{x}_p^{(k)}] \qquad \textbf{and} \qquad P_p^{(k)} := \mathbb{V}\left[\mathbf{x}_p^{(k)}\right] \tag{7.10}$$

$$\hat{x}_m^{(k)} := \mathbb{E}[\mathbf{x}_m^{(k)}] \qquad \textbf{and} \qquad P_m^{(k)} := \mathbb{V}\left[\mathbf{x}_m^{(k)}\right] \tag{7.11}$$

**Note**

The CRVs $\mathbf{x}_0$, $\{\mathbf{v}(\cdot)\}$ , $\{\mathbf{w}(\cdot)\}$ are mutually independent.

add Kalman algorithm (in slides Joesph Form I think)

# Gaussian Processes (GP)

## 1. Gaussian Process Regression

add complexitit On3 due to inverse

### 1.1. Gaussian Linear Regression

**Given**

① Linear Model with Gaussian Noise:

$$f(\mathbf{x}) = \mathrm{w}^\intercal \mathbf{x} \qquad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2 \mathbf{I}\right) \tag{8.1}$$
$$\mathbf{y} = f(\mathbf{x}) + \epsilon$$

⇒ Gaussian Likelihood: $\mathrm{p}\left(\mathbf{y}|\mathbf{X}, \mathrm{w}\right) = \mathcal{N}\left(\mathbf{X}\mathrm{w}, \sigma_n^2 \mathbf{I}\right)$

② Gaussian Prior: $\mathrm{p}(\mathrm{w}) = \mathcal{N}\left(\mathbf{0}, \Sigma_p\right)$

**Sought**

① Posterior Distribution: $\mathrm{p}(\mathrm{w}|\mathbf{y}, \mathbf{X})$

② Posterior Predictive Distribution: $\mathrm{p}(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y})$

**Definition 8.1** $\mathrm{p}(\mathrm{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\bar{\mathrm{w}}, \Sigma_{\mathrm{w}}^{-1}\right)$
**Posterior Distribution** proof 10.7:

$$\mu_{\mathrm{w}} = \frac{1}{\sigma_n^2}\Sigma_{\mathrm{w}}^{-1}\mathbf{X}\mathbf{y} \qquad \Sigma_{\mathrm{w}} = \frac{1}{\sigma_n^2}\mathbf{X}\mathbf{X}^\intercal + \Sigma_p^{-1}$$

**Note**

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

**Definition 8.2** $\mathrm{p}(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mu_*, \Sigma_*\right)$
**Posterior Predictive Distribution** proof 10.8:

$$\mu_* = \frac{1}{\sigma^2}\mathbf{x}_*^\intercal \Sigma_{\mathrm{w}}^{-1}\mathbf{X}\mathbf{y} \qquad \Sigma_* = \mathbf{x}_*^\intercal \Sigma_{\mathrm{w}}^{-1}\mathbf{x}_* \tag{8.2}$$

### 1.2. Kernelized Gaussian Linear Regression

**Definition 8.3 Posterior Predictive Distribution:**

$$\mathrm{p}(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mu_*, \Sigma_*\right) \tag{8.3}$$
$$\mu_* \tag{8.4}$$

**Definition 8.4 Gaussian Process:**

## 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

$$\mathbb{E}_{X\sim \mathrm{p}}\left[g(X)\right] = \int g(x)\mathrm{p}(x)\,\mathrm{d}x$$

- Normalization constants:

$$\mathrm{p}(\theta|y) = \frac{1}{Z}\mathrm{p}(\theta, y) \qquad Z = \int \mathrm{p}(y|\theta)\mathrm{p}(\theta)\,\mathrm{d}\theta$$
$$= \frac{\mathrm{p}(y|\theta)\mathrm{p}(\theta)}{Z} \qquad = \int \mathrm{p}(\theta)\prod_{i=1}^{n}\mathrm{p}(y_i|\mathbf{x}_i, \theta)\,\mathrm{d}\theta$$

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 9.1 Approximate Inference:** Is the procedure of finding an probability distribution $q$ that approximates a true probability distribution $\mathrm{p}$ as well as possible.

## 1. Variational Inference

**Definition 9.2 Bayes Variational Inference:** Given an unormalized (posterior) probability distribution:

$$\mathrm{p}(\theta|y) = \frac{1}{Z}\mathrm{p}(\theta, y) \tag{9.1}$$

seeks an *approximate* probability distribution $q_\lambda$, that is parameterized by a *variational parameter* $\lambda$ and approximates $\mathrm{p}(\theta|\mathbf{y})$ well.

**Definition 9.3 Variational Family of Distributions $Q$:** a set of probability distributions $Q$ that is parameterized by the same *variational parameter* $\lambda$ is called a variational familiy.

### 1.1. Laplace Approximation

**Definition 9.4** [example 10.1], [proof 10.9,10.10,10.11]
**Laplace Approximation:** Tries to approximate a desired probability distribution $\mathrm{p}(\theta|\mathcal{D})$ by a Gaussian probability distribution:

$$Q = \{q_\lambda(\theta) = \mathcal{N}(\lambda)\} = \mathcal{N}(\mu, \Sigma)\} \tag{9.2}$$

the distribution is given by:

$$q(\theta) = c\cdot\mathcal{N}(\theta; \lambda_1, \lambda_2) \tag{9.3}$$

with
$$\lambda_1 = \widehat{\theta} = \arg\max_{\theta}\mathrm{p}(\theta|y)$$
$$\lambda_2 = \Sigma = H^{-1}\left(\widehat{\theta}\right) = -\nabla\nabla_\theta \log \mathrm{p}(\widehat{\theta}|y)$$

**Corollary 9.1 :** Taylor approximation of a function $\mathrm{p}(\theta|y) \in \mathcal{C}^k$ around its mode $\widehat{\theta}$ naturally induces a Gaussian approximation. See proofs 10.9,10.10,10.11

### 1.2. Black Box Stochastic Variational Inference

The most common way of finding $q_\lambda$ is by minimizing the KL-divergence[def. 4.8] between our approximate distribution $q$ and our true posterior $\mathrm{p}$:

$$q^* \in \arg\min_{q\in Q}\mathrm{KL}\left(q(\theta)\,\|\,\mathrm{p}(\theta|y)\right) = \arg\min_{\lambda\in\mathbb{R}^d}\mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta|y))$$

**Note**

Usually we want to minimize $\mathrm{KL}\left(\mathrm{p}(\theta|y)\,\|\,q(\theta)\right)$ but this is often infeasible s.t. we only minimize $\mathrm{KL}\left(q(\theta)\,\|\,\mathrm{p}(\theta|y)\right)$

**Definition 9.5 ELBO-Optimization Problem** [proof 10.12]:

$$\begin{aligned}
q_\lambda^* \in\ & \arg\min_{\{\lambda: q_\lambda\in Q\}}\ \mathrm{KL}\left(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta|y)\right) \\
=\ & \arg\max_{\{\lambda: q_\lambda\in Q\}}\ \mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y, \theta)\right] + H(q_\lambda) && (9.4)\\
=\ & \arg\max_{\{\lambda: q_\lambda\in Q\}}\ \mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta)) && (9.5)\\
:=\ & \arg\max_{\{\lambda: q_\lambda\in Q\}}\ \mathrm{ELBO}(\lambda) && (9.6)
\end{aligned}$$

**Attention:** Sometimes people write simply $\mathrm{p}$ for the posterior and $\mathrm{p}(\cdot)$ for prior.

**Explanation 9.1.**

- *eq. (9.4):*
  - *prefer uncertain approximations i.e. we maximize $H(q)$*
  - *that jointly make the joint posterior likely*
- *eq. (9.6): Expected likelihood of our posterior over $q$ minus a regularization term that makes sure that we are not too far away from the prior.*

### 1.3. Expected Lower Bound of Evidence (ELBO)

**Definition 9.6** example 10.2/proof 10.13
**Expected Lower Bound of Evidence (ELBO):**
The evidence lower bound is a bound on the log prior:

$$\mathrm{ELBO}\left(q_\lambda\right) \leqslant \log \mathrm{p}(y) \tag{9.7}$$

#### 1.3.1. Maximizing The ELBO

**Definition 9.7 Gradient of the ELBO Loss:**

$$\begin{aligned}
\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.8)\\
&= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y, \theta)\right] + H(q_\lambda)\right]\\
&= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta))\right]\\
&= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta))
\end{aligned}$$

**Problem**

In order to use SGD we need to evaluate the gradient of the loss:

$$\nabla_\lambda\mathbb{E}\left[l(\theta;\mathbf{x})\right] = \mathbb{E}\left[\nabla_{\mathbf{x}\sim\mathrm{p}}l(\theta;\mathbf{x})\right] = \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}\sim\mathrm{p}}l(\theta;\mathbf{x})$$

however in eq. (9.8) 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:

$$\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] = \frac{\partial}{\partial\lambda}\int q_\lambda \log \mathrm{p}(y|\theta)\,\mathrm{d}\theta$$

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.

### 1.4. The Reparameterization Trick

**Principle 9.1** proof 10.14
**Reparameterization Trick:** Let $\phi$ 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 $\theta$ in terms of a new distribution parameterized by $\epsilon \sim \phi(\epsilon)$:

$$\theta \sim q(\theta|\lambda) = \phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1} \tag{9.9}$$

we can then write by the law of the unconscious statistician law 31.6:

$$\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] = \mathbb{E}_{\epsilon\sim\phi}\left[\log \mathrm{p}(y|g(\epsilon;\lambda))\right] \tag{9.10}$$

⇒ the expectations does not longer depend on $\lambda$ and we can pull in the gradient!

$$\begin{aligned}
\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] &= \nabla_\epsilon\mathbb{E}_{\theta\sim\phi}\left[\log \mathrm{p}(y|g(\epsilon;\lambda))\right] && (9.11)\\
&= \mathbb{E}_{\epsilon\sim\phi}\left[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))\right] && (9.12)
\end{aligned}$$

**Definition 9.8** example 10.3
**Reparameterized ELBO Gradient**[def. 9.7]**:**
By using the reparameterization trick principle 9.1 we can write the gradient of the ELBO as:

$$\begin{aligned}
\nabla_\lambda L(\lambda) &= \nabla_\lambda\mathrm{ELBO}(\lambda) && (9.13)\\
&= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta))\\
&= \mathbb{E}_{\epsilon\sim\phi}\left[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))\right] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta))
\end{aligned}$$

**Corollary 9.2** proof 10.3
**Reparameterized ELBO for Gaussians:**

$$\begin{aligned}
\nabla_\lambda L(\lambda) &= \nabla_\lambda\mathrm{ELBO}(\lambda) && (9.14)\\
&= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,\mathrm{p}(\theta))\\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\nabla_{\mathbf{C},\mu}\log \mathrm{p}(y|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu})\right]\\
&\qquad - \nabla_{\mathbf{C},\mu}\mathrm{KL}\left(q_{\mathbf{C},\boldsymbol{\mu}}\,\|\,\mathrm{p}(\theta)\right)\\
&\approx \frac{n}{m}\sum_{j=1}^{m}\nabla_{\mathbf{C},\mu}\log \mathrm{p}\left(y_{i_j}|\mathbf{C}\boldsymbol{\epsilon}^j + \boldsymbol{\mu}, \mathbf{x}_{i_j}\right)\\
&\qquad - \nabla_{\mathbf{C},\mu}\mathrm{KL}\left(q_{\mathbf{C},\boldsymbol{\mu}}\,\|\,\mathrm{p}(\theta)\right)
\end{aligned}$$

## 2. Markov Chain Monte Carlos Methods

**Definition 9.9**
**Markov Chain Monte Carlo (MCMC) Methods:**

# Bayesian Neural Networks (BNN)

**Definition 10.1 Bayesian Neural Networks (BNN):**

① Model the prior over our weights $\theta = \left[\mathbf{W}^0 \cdots\cdots \mathbf{W}^L\right]$ by a neural network:

$$\theta \sim p_\lambda(\theta) = \mathbf{F} \quad \text{with} \quad \begin{aligned} \mathbf{F} &= \mathbf{F}^L \circ \cdots \circ \mathbf{F}^1 \\ \mathbf{F}^l &= \varphi \circ \bar{\mathbf{F}}^l = \varphi\left(\mathbf{W}^l \mathbf{x} + b^l\right) \end{aligned}$$

for each weight $w_{k,j}^{(0)}$ of input $x_j$ with weight on the hidden variable $z_k^{(0)}$ with $a_i^0 = \varphi\left\{\mathbf{z}_i^{(0)}\right\}$ it follows:

$$w_{k,j}^{(0)} = p_w\left(\lambda_{k,j}\right) \overset{\text{i.e.}}{=} \mathcal{N}\left(\mu_{k,j}, \sigma_{k,j}^2\right)$$

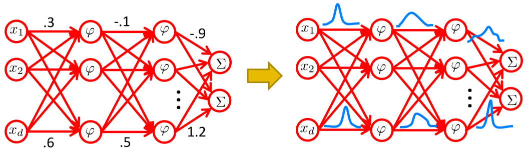


Figure 9

② The parameters of likelihood function are modeled by the output of the network:

$$p(y|F(\theta, \mathbf{X})) \qquad \text{see example 10.4} \qquad (10.1)$$

**Note**

Recall for normal Bayesian Linear regression we had:

**Problem**

All the weights of the prior $p_\lambda(\theta) = \mathbf{F}$ are correlated in some complex way see Figure 9. Thus even if the prior and likelihood are simple, the posterior will be not. ⇒ 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.

### 0.0.1. MAP estimates for BNN

**Definition 10.2 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 $\sigma(\mathbf{x}_i; \theta)^2$:

$$\theta^* = \underset{\theta}{\arg\max}\{p(\theta|\mathbf{X}, \mathbf{y})\} \overset{\text{eq. (6.13)}}{=} \underset{\theta}{\arg\min}\, \lambda\|\theta\|_2^2 - \sum_{i=1}^n \left(\frac{1}{2\sigma(\mathbf{x}_i;\theta)^2}\|y_i - \mu(\mathbf{x}_i;\theta)\|^2 + \frac{1}{2}\log\sigma(\mathbf{x}_i;\theta)^2\right)$$

**Explanation 10.1.** [def. 10.2]

- $\frac{1}{2}\log\sigma(\mathbf{x}_i;\theta)^2$: *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 10.3** proof 10.15
**MAP Gradient of BNN:**

$$\theta_{t+1} = \theta_t(1 - 2\lambda\eta_t) - \eta_t\nabla\sum_{i=1}^n \log p(y_i|\mathbf{x}_i, \theta) \qquad (10.2)$$

**Note**

- The gradients of the objective eq. (10.2) can be calculated using auto-differentiation techniques e.g. Pytorch or Tensorflow.
- The BNN MAP estimate fails to predict epistemic uncertainty[def. 6.15] $\iff$ it is overconfident in regions where we haven not even seen any data.
  ⇒ need to use Bayesian approach to approximate posterior distribution.

### 0.1. Variational Inference For BNN

We use the objective eq. (9.14) as loss in order to perform back propagation.

### 0.2. Making Predictions

**Proposition 10.1 Title:**

## 1. Proofs

**Proof 10.1.** *Definition 6.14:*

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \frac{p(\mathbf{f}_*, \mathbf{x}_*, \mathbf{X}, \mathbf{y})}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &= \frac{\int p(\mathbf{f}_*, \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})\,d\mathrm{w}}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &\overset{\text{eq. (31.19)}}{=} \frac{\int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})\,d\mathrm{w}}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &\overset{\text{eq. (31.19)}}{=} \frac{\int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathrm{w}|\mathbf{x}_*, \mathbf{X}, \mathbf{y})\cancel{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})}\,d\mathrm{w}}{\cancel{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})}} \\ &= \int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathrm{w}|\mathbf{x}_*, \mathbf{X}, \mathbf{y})\,d\mathrm{w} \\ &\overset{\clubsuit}{=} \int p(\mathbf{f}_*|\mathbf{x}_*, \mathrm{w})p(\mathrm{w}|\mathbf{X}, \mathbf{y})\,d\mathrm{w} \end{aligned}$$

**Note ♣**

- $\mathbf{f}_*$ is independent of $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ given the fixed parameter $\mathrm{w}$.
- $\mathrm{w}$ does only depend on the observed data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ and not the unseen data $\mathbf{x}_*$.

**Proof 10.2.** *Definition 6.13:*

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}, \mathrm{w}|\mathbf{X})\,d\mathrm{w} = \int p(\mathbf{y}|\mathrm{w}, \mathbf{X})p(\mathrm{w}|\mathbf{X})\,d\mathrm{w} \\ &\overset{\text{eq. (6.6)}}{=} \int p(\mathbf{y}|\mathrm{w}, \mathbf{X})p(\mathrm{w})\,d\mathrm{w} \end{aligned}$$

**Proof 10.3.** *Definition 7.4:*

$$\begin{aligned} p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) &\overset{\text{eq. (31.19)}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t_1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) \\ &\overset{\text{independ.}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) \end{aligned}$$

*marginalization/integration over* $\mathbf{x}_{t-1}$ *gives the desired result.*

**Proof 10.4.** *Definition 7.5:*

$$p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}_{1:t-1}) \overset{\text{eq. (31.23)}}{=} \begin{cases} p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1})p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \\ p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \end{cases}$$

$$p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}) \overset{\text{[cor. 7.2]}}{=} p(\mathbf{y}_t|\mathbf{x}_t)$$

*from which follows immediately eq. (7.5).*

**Proof 10.5.** *Definition 7.6:*

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) &= \int p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_{1:t-1})\,d\mathbf{x}_t \\ &= \int p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,d\mathbf{x}_t \\ &\overset{\text{[cor. 7.2]}}{=} \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,d\mathbf{x}_t \end{aligned}$$

**Proof 10.6.** [cor. 7.3]:

$$\begin{aligned} p(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) &\overset{\text{eq. (31.19)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})p(\mathbf{x}_{1:t}) \\ &\overset{\text{law 31.2}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})p(\mathbf{x}_t|\mathbf{x}_{t-1:0})\cdots p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_1) \\ &\overset{\text{eq. (7.1)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})\left(p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1})\right) \\ &\overset{\substack{\text{law 31.2} \\ \text{[cor. 7.2]}}}{=} \left(\underline{p(\mathbf{y}_1|\mathbf{x}_1)}\cdots p(\mathbf{y}_t|\mathbf{x}_t)\right)\left(p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1})\right) \\ &= \underline{p(\mathbf{y}_1|\mathbf{x}_1)}p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\mathbf{x}_{i-1}) \end{aligned}$$

**Proof 10.7.** [def. 8.1]

$$\begin{aligned} &p(\mathrm{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathrm{w})p(\mathrm{w}) \\ &\propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}(\mathbf{y} - \mathbf{X}\mathrm{w})^\intercal(\mathbf{y} - \mathbf{X}\mathrm{w})\right)\exp\left(-\frac{1}{2}\mathrm{w}^\intercal\Sigma^{-1}\mathrm{w}\right) \\ &\propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}\left(\mathbf{y}^\intercal\mathbf{y} - 2\mathrm{w}^\intercal\mathbf{X}^\intercal\mathbf{y} + \mathrm{w}^\intercal\mathbf{X}^\intercal\mathbf{X}^\intercal\mathrm{w} + \sigma_n^2\mathrm{w}^\intercal\Sigma^{-1}\mathrm{w}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}\left(\mathbf{y}^\intercal\mathbf{y} - \underline{2\mathrm{w}^\intercal\mathbf{X}^\intercal\mathbf{y}} + \underline{\mathrm{w}^\intercal\left(\mathbf{X}^\intercal\mathbf{X}^\intercal + \sigma_n^2\Sigma^{-1}\right)\mathrm{w}}\right)\right\} \end{aligned}$$

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

$$\begin{aligned} p(\mathrm{w}|\mathcal{D}) &\propto \exp\left(-\frac{1}{2}(\mathrm{w} - \bar{\mathrm{w}})^\intercal\Sigma_{\mathrm{w}}(\mathrm{w} - \bar{\mathrm{w}})\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\underline{\mathrm{w}^\intercal}\Sigma_{\mathrm{w}}\underline{\mathrm{w}} - \underline{2\mathrm{w}^\intercal\Sigma_{\mathrm{w}}\bar{\mathrm{w}}} + \bar{\mathrm{w}}^\intercal\Sigma_{\mathrm{w}}\bar{\mathrm{w}}\right)\right) \end{aligned}$$

$\Sigma_{\mathrm{w}}$ *follows directly* $\Sigma_{\mathrm{w}} = \sigma_n^{-2}\mathbf{X}\mathbf{X}^\intercal + \Sigma_p$

$\bar{\mathrm{w}}$ *follows from* $\underline{2\mathrm{w}^\intercal\mathbf{X}^\intercal\mathbf{y}} = \underline{2\mathrm{w}^\intercal\Sigma_{\mathrm{w}}\bar{\mathrm{w}}} \Rightarrow \bar{\mathrm{w}} = \Sigma_{\mathrm{w}}^{-1}\mathbf{X}^\intercal\mathbf{y}$.

**Proof 10.8.** [def. 8.2]

**Proof 10.9.** [def. 9.4] *In a Bayesian setting we are usually interested in maximizing the log prior+likelihood:*

$$\mathscr{L}_n(\theta) = \log(p(\theta|y)) = (\log \textit{Prior} + \log \textit{Likelihood})$$

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

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

*we can no derive the distribution:*

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

**Notes**

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

**Proof 10.10.** [def. 9.4] *2D:*

$$\nabla\mathscr{L}_n(\theta) = \nabla\mathscr{L}_n(\theta_1, \theta_2) = 0$$

$$\mathscr{L}_n(\theta) = \mathscr{L}_n\left(\hat{\theta}\right) + \frac{1}{2}\Big(A(\theta_1 - \hat{\theta}_1)^2 + B(\theta_2 - \hat{\theta}_2)^2 + C(\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)\Big)$$

$$\begin{aligned} \mathscr{L}_n(\theta) &= \mathscr{L}_n\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right)^\intercal H\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right) \\ &= \mathscr{L}_n\left(\hat{\theta}\right) + \frac{1}{2}Q(\theta) \end{aligned}$$

$$A = \frac{\partial^2\mathscr{L}_n}{\partial\theta^2}\Big|_{\hat{\theta}} \qquad B = \frac{\partial^2\mathscr{L}_n}{\partial\theta^2}\Big|_{\hat{\theta}} \qquad C = \frac{\partial^2\mathscr{L}_n}{\partial\theta_1\partial\theta_2}\Big|_{\hat{\theta}}$$

$$H = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \qquad \Sigma = H^{-1}\left(\hat{\theta}\right)$$

**Proof 10.11.** [def. 9.4] $k$*-dimensional:*

$$\mathscr{L}_n(\theta) \approx \mathscr{L}_n\left(\hat{\theta}\right) + \left(\theta - \hat{\theta}\right)^\intercal\nabla\nabla^\intercal\mathscr{L}_n\left(\hat{\theta}\right)\left(\theta - \hat{\theta}\right)$$

$$H(\theta) = \nabla\nabla^\intercal\mathscr{L}_n(\theta) \qquad \Sigma = H^{-1}\left(\hat{\theta}\right)$$

$$\begin{aligned} p(\theta|y) &= \sqrt{(2\pi)^n\det(\Sigma)}p\left(\hat{\theta}\right)\mathcal{N}\left(\theta; \hat{\theta}, \Sigma\right) \\ &\approx c\frac{1}{\sqrt{(2\pi)^n\det(\Sigma)}}\mathcal{N}\left(\theta; \hat{\theta}, \Sigma\right) \end{aligned}$$

**Proof 10.12.** *[def. 9.5]*

$$\begin{aligned}
q^{*} &\in \underset{q\in Q}{\arg\min}\, KL\left(q(\theta) \parallel \mathrm{p}(\theta|y)\right)\\
&\overset{\mathrm{p}(\theta|y)=\frac{1}{Z}\mathrm{p}(\theta,y)}{=} \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q}\left[\log \frac{q(\theta)}{\frac{1}{Z}\mathrm{p}(\theta,y)}\right]\\
&= \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q}\left[\log q(\theta) - \log\frac{1}{Z} - \log \mathrm{p}(\theta,y)\right]\\
&= \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q} - \underbrace{[-\log q(\theta)]}_{H(q)} + \cancel{\mathbb{E}_{\theta\sim q}[\log Z]}\\
&\qquad - \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta,y)\right]\\
&= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta,y)\right] + H(q)\\
&= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta|y) + \log \mathrm{p}(\theta) - \log q(\theta)\right]\\
&= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta|y)\right] + KL\left(q(\theta) \parallel \mathrm{p}(\theta)\right)
\end{aligned}$$

**Proof 10.13.** *[def. 9.6]*

$$\begin{aligned}
\log \mathrm{p}(y) &= \log\int \mathrm{p}(y,\theta)\,\mathrm{d}\theta = \log\int \mathrm{p}(y|\theta)\mathrm{p}(\theta)\,\mathrm{d}\theta\\
&= \log\int \mathrm{p}(y|\theta)\frac{\mathrm{p}(\theta)}{q_{\lambda}(\theta)}q_{\lambda}(\theta)\,\mathrm{d}\theta\\
&= \log \mathbb{E}_{\theta\sim q_{\lambda}}\left[\mathrm{p}(y|\theta)\frac{\mathrm{p}(\theta)}{q_{\lambda}(\theta)}\right]\\
&\overset{eq.\ (31.55)}{\geqslant} \mathbb{E}_{\theta\sim q_{\lambda}}\left[\log\left(\mathrm{p}(y|\theta)\frac{\mathrm{p}(\theta)}{q_{\lambda}(\theta)}\right)\right]\\
&= \mathbb{E}_{\theta\sim q_{\lambda}}\left[\log \mathrm{p}(y|\theta) - \log\frac{\mathrm{p}(\theta)}{q_{\lambda}(\theta)}\right]\\
&= \mathbb{E}_{\theta\sim q_{\lambda}}\left[\log \mathrm{p}(y|\theta)\right] - KL\left(q_{\lambda} \parallel \mathrm{p}(\cdot)\right)
\end{aligned}$$

**Proof 10.14.** *principle 9.1 Let:*

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

*then it follows immediately with* **??***:*

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

$\Rightarrow$ *parameterized in terms of* $\epsilon$

**Proof 10.15.** *[def. 10.3]*

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

## 2. Examples

**Example 10.1 Laplace Approximation Logistic Regression Likelihood + Gaussian Prior:**

**Example 10.2 ELBO Bayesian Logistic Regression:** Suppose:

$$\begin{aligned}
Q &= \text{diag. Gaussians} \qquad \Rightarrow \qquad \lambda = \begin{bmatrix}\mu_{1:d} & \sigma^2_{1:d}\end{bmatrix} \in \mathbb{R}^{2d}\\
\mathrm{p}(\theta) &= \mathcal{N}(0,\mathbf{I})
\end{aligned}$$

Then it follows for the terms of the ELBO:

$$\mathrm{KL}(q_{\lambda} \parallel \mathrm{p}(\theta)) = \frac{1}{2}\sum_{i=1}^{d}\left(\mu_i^2 + \sigma_i^2 - 1 - \ln\sigma_i^2\right)$$

$$\begin{aligned}
\mathbb{E}_{\theta\sim q_{\lambda}}\left[\mathrm{p}(y|\theta)\right] &= \mathbb{E}_{\theta\sim q_{\lambda}}\left[\sum_{i=1}^{n}\log \mathrm{p}(y_i|\theta,\mathbf{x}_i)\right]\\
&= \mathbb{E}_{\theta\sim q_{\lambda}}\left[-\sum_{i=1}^{n}\log\left(1 + \exp\left(-y_i\theta^{\intercal}\mathbf{x}_i\right)\right)\right]
\end{aligned}$$

**Example 10.3 ELBO Gradient Gaussian:** Suppose:

$$\begin{aligned}
\theta &\sim q(\theta|\lambda) = \mathcal{N}(\theta;\boldsymbol{\mu},\Sigma) \qquad \Rightarrow \qquad \lambda = [\boldsymbol{\mu} \quad \Sigma]\\
\boldsymbol{\epsilon} &\sim \phi(\epsilon) = \mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})
\end{aligned}$$

we can reparameterize using principle 9.1 by using:

$$\theta \sim g(\boldsymbol{\epsilon},\lambda) = \mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu} \qquad \text{with} \qquad \mathbf{C}: \quad \mathbf{C}\mathbf{C}^{\intercal} = \Sigma$$

from this it follows: (**C** is the Cholesky factor of $\Sigma$)

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

from this it follows:

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

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

$$\begin{aligned}
\nabla_{\lambda}L(\lambda)_1 &= \nabla_{\lambda}\mathbb{E}_{\epsilon\sim\phi}\left[\log \mathrm{p}\left(y|g(\epsilon;\lambda)\right)\right]\\
&= \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\log \mathrm{p}(y|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu})\right]\\
&\overset{\text{i.i.d.}}{=} \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\sum_{i=1}^{n}\log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]\\
&= \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[n\frac{1}{n}\sum_{i=1}^{n}\log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]\\
&= \nabla_{\mathbf{C},\mu}n\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbb{E}_{i\sim\mathcal{U}(\{1,n\})}\log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]\\
&\qquad \text{Draw a mini batch} \begin{cases}\epsilon^{(1)},\dots,\epsilon^{(m)}\\ j_1,\dots,j_m \sim \mathcal{U}(\{1,n\})\end{cases}\\
&= n\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{C},\mu}\log \mathrm{p}\left(y_j|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_j\right)
\end{aligned}$$

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

**Example 10.4 BNN Likelihood Function Examples:**

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

# 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 11.1 Similarity Measure $\mathrm{sim}(A, B)$:** 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 11.2 Dissimilarity Measure $\mathrm{disssim}(A, B)$:** 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 sqaured norm of two difference vectors:

$$\|\mathbf{x}-\mathbf{y}\|^2 = \mathbf{x}^\intercal\mathbf{x} + \mathbf{y}^\intercal\mathbf{y} - 2\mathbf{x}^\intercal\mathbf{y} \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{11.1}$$

$$\mathrm{dissim}(\mathbf{x}, \mathbf{y}) = \mathrm{sim}(\mathbf{x}, \mathbf{x}) + \mathrm{sim}(\mathbf{y}, \mathbf{y}) - 2\mathrm{dissim}(\mathbf{x}, \mathbf{y})$$

**Attention**

It is better to rely on similarity measures instead of dissimilarity measures. Dissimilarities are often not adequat 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. (11.1) becomes $\quad 2\mathbf{x}^\intercal\mathbf{y} = 2\mathrm{dissim}(\mathbf{x}, \mathbf{y})$

**Definition 11.3 Feature Map $\phi$:** is a mapping $\phi : \mathcal{X} \mapsto \mathcal{V}$ that takes an input $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and maps it into another feature space $\mathcal{V} \subseteq \mathbb{R}^D$.

**Note**

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

$$D = \dim(\mathcal{V}) = \binom{p+d}{d} = \mathcal{O}\left(d^p\right) \tag{11.2}$$

when using the polynomial kernel[def. 11.10], this can be reduced to the order $d$.

**Definition 11.4 Kernel $\mathbf{k}$: Let $\mathcal{X} \subseteq \mathbb{R}^d$** be the data space. A map $\mathbf{k} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called kernel if their exists an inner product space[def. 25.74] called **feature space** $(\mathcal{V}, \langle\cdot,\cdot\rangle_{\mathcal{V}})$ and a map $\phi : \mathcal{X} \mapsto \mathcal{V}$ s.t.

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \langle\phi(\mathbf{x}), \phi(\mathbf{y})\rangle_{\mathcal{V}} \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \tag{11.3}$$

**Corollary 11.1 Kernels and similarity:** Kernels are defined in terms of inner product spaces and hence the have a notion of similarity between its arguments.

**Example**

**Let $\mathbf{k}(\mathbf{x}, \mathbf{y}) := \mathbf{x}^\intercal\mathbf{A}\mathbf{y}$ thus** the kernel measures the similarity between $\mathbf{x}$ and $\mathbf{y}$ by the inner product $\mathbf{x}^\intercal\mathbf{y}$ weighted by the matrix $\mathbf{A}$.

**Corollary 11.2 Kernels and distance: Let $\mathbf{k}(\mathbf{x}, \mathbf{y})$** be a measure of similarity between $\mathbf{x}$ and $\mathbf{y}$ then $\mathbf{k}$ induces a dissimilarity/distance between $\mathbf{x}$ and $\mathbf{y}$ defined as the difference betweend the self-similarities $\mathbf{k}(\mathbf{x}, \mathbf{x}) + \mathbf{k}(\mathbf{y}, \mathbf{y})$ and the cross-similarities $\mathbf{k}(\mathbf{x}, \mathbf{y})$:

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

**Note**

The factor 2 is required to ensure that $d(\mathbf{x}, \mathbf{x}) = 0$.

## 1. The Gram Matrix

**Definition 11.5 Kernel (Gram) Matrix:**
**Given**: a mapping $\phi : \mathbb{R}^d \mapsto \mathbb{R}^D$ and a corresponding kernel function $\mathbf{k} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ with $\mathcal{X} \subseteq \mathbb{R}^d$.
**Let** $S$ be any finite subset of data $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \mathcal{X}$.
Then the kernel matrix $\mathcal{K} :\in \mathbb{R}^{n\times n}$ is defined by:

$$\mathcal{K} = \phi(\mathbf{X})\phi(\mathbf{X}^\intercal) = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))(\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))^\intercal$$

$$= \begin{pmatrix} \mathbf{k}(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \phi(\mathbf{x}_1)^\intercal\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_1)^\intercal\phi(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_n)^\intercal\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_n)^\intercal\phi(\mathbf{x}_n) \end{pmatrix}$$

$$\mathcal{K}_{ij} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\intercal\phi(\mathbf{x}_j)$$

**Corollary 11.3** $\quad \mathbf{V}\Lambda\mathbf{V}^\intercal$
**Kernel Eigenvector Decomposition:**
For any symmetric matrix (Gram matrix $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)|_{i,j=1}^n$) there exists an eigenvector decomposition:

$$\mathcal{K} = \mathbf{V}\Lambda\mathbf{V}^\intercal \tag{11.4}$$

$\mathbf{V}$ : orthogonal matrix of eigenvectors $(\mathbf{v}_{t,i})|_{i=1}^n$
$\Lambda$ : diagonal matrix of eigenvalues $\lambda_i$

**Assuming** all eigenvalues $\lambda_t$ are non-negative, we can calculate the mapping:

$$\phi : \mathbf{x}_i \mapsto \left(\sqrt{\lambda_t}\mathbf{v}_{t,i}\right)_{t=1}^n \in \mathbb{R}^n, \qquad i = 1, \ldots, n \tag{11.5}$$

which allows us to define the Kernel $\mathcal{K}$ as:

$$\phi^\intercal(\mathbf{x}_i)\phi(\mathbf{x}_j) = \sum_{t=1}^n \lambda_t\mathbf{v}_{t,i}\mathbf{v}_{t,j} = \left(\mathbf{V}\Lambda\mathbf{V}^\intercal\right)_{i,j} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{11.6}$$

### 1.1. Necessary Properties

**Property 11.1 Inner Product Space:**
$\mathbf{k}$ must be an *inner product* of a suitable space $\mathcal{V}$.

**Property 11.2 Symmetry: $\mathbf{k}/\mathcal{K}$** must be symmetric:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}(\mathbf{y}, \mathbf{x}) = \phi(\mathbf{x})^\intercal\phi(\mathbf{y}) = \phi(\mathbf{y})^\intercal\phi(\mathbf{x}) \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$$

**Property 11.3 Non-negative Eigenvalues/p.s.d.s Form:**
Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be an $n$-set of a *finite* input space $\mathcal{V}$.
A kernel $\mathbf{k}$ must induces a *p.s.d. symmetric* kernel matrix $\mathbf{k}$ for any possible $S \subseteq \mathcal{X}$ see **??** 11.1.
$\iff$ all eigenvalues of the kernel gram matrix $\mathcal{K}$ for *finite* $\mathcal{V}$ must be non-negative **??** 25.2.

**Notes**

- The extension to infinite dimensional Hilbert Spaces might also include a non-negative weighting/eigenvalues:

$$\langle\phi(\mathbf{x}), \phi(\mathbf{z})\rangle = \sum_{i=1}^\infty \lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{z})$$

- In order to be able to use a kernel, we need to verify that the kernel is p.s.d. for all n-vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, as well as for future unseen values.

## 2. Mercers Theorem

**Theorem 11.1 Mercers Theorem:** Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$ and $\mathbf{k} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ a kernel function.
**Then** one can expand $\mathbf{k}$ in a uniformly convergent series of bounded functions $\phi$ s.t.

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^\infty \lambda\phi(\mathbf{x})\phi(\mathbf{x}') \tag{11.7}$$

**Theorem 11.2 General Mercers Theorem:** Let $\Omega$ be a compact subset of $R^n$. Suppose $\mathbf{k}$ is a gernal continuous symmetric function such that the integral operator:

$$T_{\mathbf{k}} : L_2(\mathbf{X}) \mapsto L_2(\mathbf{X}) \qquad (T_{\mathbf{k}}f)(\cdot) = \int_\Omega \mathbf{k}(\cdot, \mathbf{x})f(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{11.8}$$

is positve, that is it satisfies:

$$\int_{\Omega\times\Omega} \mathbf{k}(\mathbf{x}, \mathbf{z})f(\mathbf{x})f(\mathbf{z})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z} > 0 \qquad \forall f \in L_2(\Omega)$$

**Then** we can expand $\mathbf{k}(\mathbf{x}, \mathbf{z})$ in a uniformly convergent series in terms of $T_{\mathcal{K}}$'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 condtions describe an inner product in a high dimensional space.
$\Rightarrow$ can replace the inner product by the kernel function.

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

## 3. The Kernel Trick

**Definition 11.6 Kernel Trick:** If a kernel has an analytic form we do no longer need to calculate:

- the function mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$ and
- the inner product $\phi(\mathbf{x})^\intercal\phi(\mathbf{y})$

explicitly but simply us the formula for the kernel:

$$\phi(\mathbf{x})^\intercal\phi(\mathbf{x}) = \mathbf{k}(\mathbf{x}, \mathbf{y}) \tag{11.9}$$

see examples 11.1 and 11.2

**Note**

- Possible to operate in any n-dimenional function space, efficiently.
- $\phi$ not necessary anymore.
- Complexity independent of the functions space.

## 4. Types of Kernels

### 4.1. Stationary Kernels

**Definition 11.7 Stationary Kernel:** A stationary kernel is a kernel that only considers vector differences:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}(\mathbf{x} - \mathbf{y}) \tag{11.10}$$

see example example 11.3

### 4.2. Isotropic Kernels

**Definition 11.8 Isotropic Kernel:** A isotropic kernel is a kernel that only considers distance differences:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}\left(\|\mathbf{x} - \mathbf{y}\|_2\right) \tag{11.11}$$

**Corollary 11.4 :**
Isotropic → Stationary

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

### 5.1. The Linear Kernel

**Definition 11.9 Linear/String Kernel:**

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal\mathbf{y} \tag{11.12}$$

### 5.2. The Polynomial Kernel

**Definition 11.10 Polynomial Kernel:**
represents all monomials[def. 22.5] of degree up to $m$

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \left(1 + \mathbf{x}^\intercal\mathbf{y}\right)^m \tag{11.13}$$

### 5.3. The Sigmoid Kernel

**Definition 11.11 Sigmoid/tanh Kernel:**

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \tanh \kappa\mathbf{x}^\intercal\mathbf{y} - b \tag{11.14}$$

### 5.4. The Exponential Kernel

**Definition 11.12 Exponential Kernel:**
is an continuous kernel that is non-differential $\mathbf{k} \in \mathcal{C}^0$:

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_1}{\theta}\right) \tag{11.15}$$

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

### 5.5. The Gaussian Kernel

**Definition 11.13 Gaussian/Squared Exp. Kernel/ Radial Basis Functions (RBF):**
Is an inifite dimensional smooth kernel $\mathbf{k} \in \mathcal{C}^\infty$ with some usefull properties

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\theta^2}\right) \approx \begin{cases} 1 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ close} \\ 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ far away} \end{cases} \tag{11.16}$$

**Explanation 11.1** (Threshold $\theta$). $2\theta \in \mathbb{R}$ *corresponds to a threshold that determines how close input values need to be in order to be considered similar:*

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

*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 admitable functions vary much more.*

**Note**

If we chose $h$ small, all data points not close to $h$ will be 0/discared $\iff$ data points are considered as independent.
Length of all vectors in feature space is one $\mathbf{k}(\mathbf{x}, \mathbf{x}) = e^0 = 1$.
**Thus**: Data points in input space are projected onto a high-(infintie-)dimensional sphere in feature space.
**Classification**: Cutting with hyperplances through the sphere. **How to chose $h$**: 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. 11.13] is often a too strong assumption that does not model reality the same holds true for the non-smoothness of the exponential kernel[def. 11.12]. A solution to this dilemma is the Matern kernel.

**Definition 11.14 Matern Kernel:** is a kernel which allows you to specify the level of smoothness $\mathbf{k} \in \mathcal{C}^{\lfloor\nu\rfloor}$ by a positive parameter $\nu$:

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

$\nu, \rho \in \mathbb{R}_+ \qquad \nu$ : Smoothness $\qquad \rho$ : Le

(11.17)

$\mathcal{K}_\nu$ 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:

$$\mathbf{k}_1 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \qquad \mathbf{k}_2 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$$

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

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_1(\mathbf{x}, \mathbf{x}') + \mathbf{k}_2(\mathbf{x}, \mathbf{x}') \tag{11.18}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_1(\mathbf{x}, \mathbf{x}') \cdot \mathbf{k}_2(\mathbf{x}, \mathbf{x}') \tag{11.19}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \alpha\,\mathbf{k}_1(\mathbf{x}, \mathbf{x}') \qquad \alpha \in \mathbb{R}_+ \tag{11.20}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})f(\mathbf{x}') \tag{11.21}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) \tag{11.22}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = p\left(\mathbf{k}(\mathbf{x}, \mathbf{x}')\right) \tag{11.23}$$

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \exp\left(\mathbf{k}(\mathbf{x}, \mathbf{x}')\right) \tag{11.24}$$

**Where** $\quad f : \mathcal{X} \mapsto \mathbb{R} \quad$ a real valued function
$\phi : \mathcal{X} \mapsto \mathbb{R}^e \quad$ the explicit mapping
$p \quad$ a polynomial with pos. coefficients
$\mathbf{k}_3 \quad$ a Kernel over $\mathbb{R}^e \times \mathbb{R}^e$

# Proofs

**Proof 11.1.** *Property 11.3* **The kernel matrix is positive-semidefinite**:
**Let** $\phi : \mathcal{X} \mapsto \mathbb{R}^d$ *and* $\mathbf{\Phi} = [\phi(\mathbf{x}_1) \quad \ldots \quad \phi(\mathbf{x}_n)]^\intercal \in \mathbb{R}^{d\times n}$.
**Thus**: $\mathcal{K} = \mathbf{\Phi}^\intercal\mathbf{\Phi} \in \mathbb{R}^{n\times n}$.

$$\mathbf{v}^\intercal\mathcal{K}\mathbf{v} = \mathbf{v}^\intercal\mathbf{\Phi}^\intercal\mathbf{\Phi}\mathbf{v} = (\mathbf{\Phi}\mathbf{v})^T\mathbf{\Phi}\mathbf{v} = \|\mathbf{\Phi}\mathbf{v}\|_2^2 \geqslant 0$$

## Examples

**Example 11.1 Calculating the Kernel by hand:**

**Let** : $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\quad \phi(\mathbf{x}) \mapsto \{x_1^2, x_2^2, \sqrt{2}x_1, x_2\}$

$\phi : \mathbb{R}^{d=2} \mapsto \mathbb{R}^{D=3}$

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

$$\beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 \sqrt{2} x_1 x_2 = 0$$

$$\begin{aligned} &\left\langle \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) \right\rangle \\ &= \left\langle \left\{ x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}, x_{i2} \right\}, \left\{ x_{j1}^2, x_{j2}^2, \sqrt{2}x_{j1}, x_{j2} \right\} \right\rangle \\ &= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} \end{aligned}$$

**Operation Count**:

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

**Example 11.2**
**Calculating the Kernel using the Kernel Trick:**

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

**Operation Count**:

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

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

**Example 11.3 Stationary Kernels:**

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

is a stationary but not an isotropic kernel.

# Time Series

## State Space Models

**Definition 12.1 State Variables** $\mathbf{x}$:
Is the smallest set of variables $\{x_1, \dots, x_n\}$ that are fully capable of describing the state of our system which is usually *hidden* and not directly observable.

**Definition 12.2 State Space** $\mathcal{X}$:
Is the $n$-dimensional space spanned by the state variables??:

$$\mathbf{x} = \left[x_1 \cdots\cdots x_n\right]^{\mathsf{T}} \in \mathcal{S} \subseteq \mathbb{R}^n \tag{12.1}$$

**Definition 12.3**
**Input/Control Variables** $\mathbf{u} \in \mathcal{A}$:
Are a variables $\mathbf{u}$ of the *transition model*[def. 12.5] that influence the propagation of to the state variables $\mathbf{x}$.

**Definition 12.4** $\mathbf{y} \in \mathcal{O}$
**Output/Measurment Variables/State Observations:**
Are a variables $\mathbf{y}$ that are directly related to the state space $\mathbf{x}$ and are usually observable by us.

**Definition 12.5 Transition Model** $f$:
Describes the transition of the state $\mathbf{x}$ over time.

**Definition 12.6**
**Measurment/Output/Observation Model** $h$:
Describes the mapping of the state $\mathbf{x}$ onto the output $\mathbf{y}$.

**Definition 12.7 (Discrete) State Space Model:**

$$\mathbf{x}^{k+1} = f(t, \mathbf{x}^k, \mathbf{u}^k) \qquad t = 1, \dots, K \tag{12.2}$$

$$\mathbf{y}^k = h(t, \mathbf{x}^k, \mathbf{u}^k) \tag{12.3}$$

## Markov Models

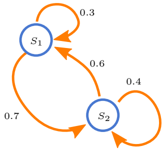
**Definition 13.1 States** $\mathcal{S} = \{s_1, \dots, s_n\}$:
A state $s_i$ encodes all information of the current configuration of a system.

**Definition 13.2**
**Markovian Property/Memorylessness:**
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration $(\mathcal{F}_s, s \in I)$, for some index set[def. 20.1]; and let $(S, \mathcal{S})$ be a measurable space[def. 31.7].
A $(S, \mathcal{S}))$-valued stochastic process $X = \{X_t : \Omega \to S\}_{t\in I}$ adapted to the filtration is said to possess the Markov property if:

$$\mathbb{P}\left(X_t \in A | \mathcal{F}_s\right) = \mathbb{P}\left(X_t \in A | X_s\right) \qquad \begin{array}{l} \forall A \in \mathcal{S} \\ s, t \in I \quad \text{s.t. } s < t \end{array} \tag{13.1}$$

### 1. Markov Chains

**Definition 13.3 Markov Chain:**
Is a sequence of random variables $\{X_i\}_{i\in\mathcal{T}}$[def. 35.3] that processes the markovian property[def. 13.2] i.e. each state $X_t$ depend only on the previous state $X_{t-1}$:



$$\mathbb{P}(X_t = x | X_{t-1} = x_{t-1}, \dots, X_1 = x_1) = \mathbb{P}(X_t = x | X_{t-1} = x_{t-1})$$

**Definition 13.4 Initial Distribution** $\mathbf{q}_0$: Describes the initial distribution of states:

$$q_0\left(s_i\right) = \mathbb{P}\left(X_0 = s_i\right) \qquad \forall s_i \in S$$

$$\iff \quad \mathbf{q}_0 = \left[q_0\left(s_1\right) \quad \cdots \quad q_0\left(s_n\right)\right] \tag{13.2}$$

**Definition 13.5 Transition Probability** $\mathrm{p}_{ji}(t)$:
is the probability of a random variable $X_t$ in state $s_i$ to transition into state $s_j$:

$$\mathrm{p}_{ij}(t) = \mathbb{P}\left(X_{t+1} = s_j | X_t = s_i\right) \qquad \forall s_i, s_j \in S \tag{13.3}$$

**Definition 13.6** $n^{\text{th}}$ **Transition Probability** $\mathrm{p}_{ji}^{(n)}(t)$:
denotes the probability of reaching state $s_j$ from state $s_i$ in $n$ steps:

$$\mathrm{p}_{ij}^{(n)}(t) = \mathbb{P}\left(X_{t+n} = s_j | X_t = s_i\right) \qquad \forall s_i, s_j \in S \tag{13.4}$$

**Definition 13.7 Transition Matrix** $\mathrm{P}(t)$:
The transition probabilities eq. (13.4) can be represented by a *row-stochastic matrix*?? $\mathrm{P}(t)$ where the $i^{th}$ row represents the transition probabilities for the $i^{th}$ state $s_i$ i.e.

$$\begin{array}{cc} & \textbf{To } j \\ \textbf{From } i & \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

**Corollary 13.1 Row stochastic matrices and Graphs:**
Row stochastic matrices?? represent graphs where the outgoing edges must sum to one:

$$\sum \delta^+\left(s_i\right) = 1 \tag{13.5}$$

#### 1.1. Simulating Markov Chains

**Corollary 13.2** proof 13.1
**Realization of a Markov Chain:**

$$\mathbb{P}\left(X_0 = x_0, \dots, X_N = x_N\right) = q_0\left(x_1\right) \sum_{n=1}^{N} \mathrm{p}_{n-1,n}(t)$$

**Algorithm 13.1 Forward Sampling:**

```
    Input:    $\mathbf{q}(\mathbf{x}_0)$ and $\mathrm{P}$
    Output:   $\mathbb{P}\left(X_{0:N}\right)$
    Sample $x_0 \sim \mathbb{P}\left(X_0\right)$
    for $j = 1, \dots, n$ do
        $x_j \sim \mathbb{P}(X_j | X_{j-1} = x_{j-1})$
5:  end for
```

#### 1.2. State Distributions

**Definition 13.8**
**Probability Distribution of the States** $\mathbf{q}_{n+1}$:

$$\begin{aligned} q_{n+1}\left(s_j\right) &= \mathbb{P}\left(X_{n+1} = s_j\right) \qquad \forall s_i \in S \\ &= \sum_{i=1}^{n} \mathbb{P}\left(X_n = s_i\right) \mathbb{P}\left(X_{n+1} = s_j | X_n = s_i\right) \\ &= \sum_{i=1}^{n} q_n\left(s_i\right) \mathrm{p}_{i,j}(t) \end{aligned} \tag{13.6}$$

$$\begin{aligned} \mathbf{q}_{n+1} &= \left[q_{n+1}\left(s_1\right) \quad \cdots \quad q_{n+1}\left(s_n\right)\right] \\ &= \mathbf{q}_n \mathrm{P}(t) \\ &= \left[q_n\left(s_1\right) \quad \cdots \quad q_n\left(s_n\right)\right] \begin{bmatrix} \mathrm{p}_{1,1} & \mathrm{p}_{1,2} & \cdots\cdots & \mathrm{p}_{1,n} \\ \mathrm{p}_{2,1} & \mathrm{p}_{2,2} & \cdots\cdots & \mathrm{p}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}_{n,1} & \mathrm{p}_{n,2} & \cdots\cdots & \mathrm{p}_{n,n} \end{bmatrix} (t) \end{aligned}$$

**Corollary 13.3** [proof 13.2]
**Time-homogeneous Markov Transition Probabilities:**

$$\mathbf{q}_{n+1} = \mathbf{q}_0 \mathrm{P}^{n+1} \tag{13.7}$$

**Definition 13.9 Stationary Distribution:**
A markov chain has a stationary distribution if it satisfies:

$$\lim_{N\to\infty} q_N\left(s_i\right) = \lim_{N\to\infty} \mathbb{P}\left(X_N = s_i\right) = \pi_i \qquad \forall s_i \in S$$

$$\lim_{N\to\infty} \mathbf{q}_N = \left[\pi_1 \quad \cdots \quad \pi_n\right] \iff \mathbf{q} = \mathbf{q}\mathrm{P}(N) \tag{13.8}$$

**Corollary 13.4 Existence of Stationary Distributions:**
A Markov Chain has a stationary distribution if and only if at least one state is *positive recurrent*!

add matrix version with eigenvector

add recurrent and transient states

#### 1.3. Properties of States

**Definition 13.10 Absorbing State/Sink:** Is a state $s_i$ that once entered cannot be left anymore:

$$\mathrm{p}_{ij}^{(n)}(t) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{13.9}$$

**Definition 13.11 Accessible State** $s_i \to s_j$:
A state $s_j$ is accessible from state $s_i$ iff:

$$\exists n : \quad \mathrm{p}_{ij}^{(n)}(t) > 0 \tag{13.10}$$

**Definition 13.12 Communicating States** $s_i \leftrightarrow s_j$:
Two states $s_j$ and $s_i$ are communicating iff:

$$\exists n_1 : \quad \mathrm{p}_{ij}^{(n_1)}(t) > 0 \quad \wedge \quad \exists n_2 : \quad \mathrm{p}_{ji}^{(n_2)}(t) > 0 \tag{13.11}$$

**Definition 13.13 Periodicity of States:** A state $s_i$ has period $k$ if any return to state $s_i$ must occur in multiples of $k$ time steps.
In other words $k$ is the *greatest common divisor* of the number of transitions by which state $s_i$ can be reached, starting from itself:

$$k = \gcd\{n > 0 : \mathrm{p}_{ii}^{(n)} = \mathbb{P}\left(X_n = s_i \mid X_0 = s_i\right) > 0\} \tag{13.12}$$

**Definition 13.14 Aperiodic State** $k = 1$:
Is a state $s_i$ with periodicity[def. 13.13] of one $\Leftrightarrow$ $k = 1$

**Corollary 13.5 :** A state $s_i$ is aperiodic if there exist two consecutive numbers $k$ and $k+1$ s.t. the chain can be in state $s_i$ at both time steps $k$ and $k+1$.

**Corollary 13.6 Absorbing State:** An absorbing state is an aperiodic state.

**Explanation 13.1** (Defintion 13.14). *Returns to state $s_i$ can occur at irregular times i.e. the state is not predictable.*
*In other words we cannot predict if the state will be revisited in multiples of $k$ times.*

#### 1.4. Characteristics of Markov Processes/Chains

**Definition 13.15**
**Time-homogeneous/Stationary Markov Chain:**
are markov chains[def. 13.3] where the transition probability is independent of time:

$$\mathrm{p}_{ji} = \mathbb{P}\left(X_t = s_j | X_{t-1} = s_i\right) = \mathbb{P}\left(X_{t-\tau} = s_j | X_{t-\tau} = s_i\right) \qquad \forall \tau \in \mathbb{N}_0 \tag{13.13}$$

**Corollary 13.7** $\mathrm{P}$
**Transition Matrices of Stationary MCs:**
Transition matrices of time-homogeneous markov chain are constant/time independent:

$$\mathrm{P}(t) = \mathrm{P} \tag{13.14}$$

**Definition 13.16 Aperiodic Makrov Chain:** Is a markov chain where all states are aperiodic:

$$\gcd\{n > 0 : \mathrm{p}_{ii}^{(n)} = \mathbb{P}\left(X_n = s_i \mid X_0 = s_i\right) > 0\} = 1 \qquad \forall i \in \{1, \dots, n\} \tag{13.15}$$

**Definition 13.17 Irreducable Markov Chain:** Is a Markov chain that has only *communicating states*[def. 13.12]:

$$s_j \leftrightarrow s_i \qquad \forall i, j \in \{1, \dots n\} \tag{13.16}$$

- $\implies$ no sinks[def. 13.10]
- $\implies$ every state can be reached from every other state

**Corollary 13.8 :** An *irreducable*[def. 13.17] markov chain is automatically *apperiodic*[def. 13.16] if it has at least one aperiodic state[def. 13.14] $\iff$ *ergodic*[def. 13.18].

**Corollary 13.9 :** A markov chain is *not-irreducable* if there exist two states with different periods.

**Definition 13.18** [example 13.1]
**Ergodic Markov Chain:** A finite markov chain is ergodic if there exist some number $N$ s.t. any state $s_j$ can be reached from any other state $s_i$ in any number of steps less or equal to a $N$.
$\Rightarrow$ a markov chains is ergodic if it is:

1. *Irreducable*[def. 13.17]
2. *Aperiodic*[def. 13.16]

**Corollary 13.10 Stationary Distribution:** An erdodic markov chain has a *unique* stationary distribution[def. 13.9] and converges to it starting from any initial state $q_0\left(s_i\right)$

#### 1.5. Types of Markov Chains

| | Observable | Unobservable |
|---|---|---|
| **Uncontrolled** | MC[def. 13.3] | HMM[def. 14.1] |
| **Controlled** | MDP[def. 15.1] | POMDP[def. 16.1] |

#### 1.6. Markov Chain Monte Carlo (MCMC)

### 2. Proofs

**Proof 13.1.** *[cor. 13.2]*

$$\mathbb{P}\left(X_0 = x_0, \dots, X_N = x_N\right) = \mathbb{P}\left(X_0 = x_0\right) \cdot \\ \cdot \mathbb{P}\left(X_1 = x_1 | X_0 = x_0\right) \cdot \mathbb{P}\left(X_2 = x_2 | X_1 = x_1, X_0 = x_0\right) \cdot \\ \cdots \mathbb{P}\left(X_N = x_N | X_{N-1} = x_{N-1}, \dots, X_0 = x_0\right)$$

*and then simply use the Markovian property*

**Proof 13.2.** *Corollary 13.3*

$$\mathbf{q}_{n+1} = \mathrm{P}\mathbf{q}_n = \left(\mathbf{q}_{n-1}\mathrm{P}\right)\mathrm{P} = \mathbf{q}_0 \mathrm{P}^{n+1}$$

### 3. Examples

**Example 13.1 Ergodic Markov Chain:**

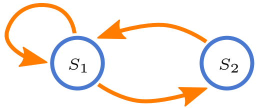
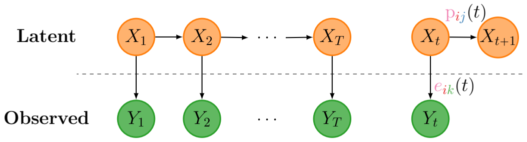


Figure 10: Ergodic for $N = 2$ (can reach $s_2$ at any $t \leqslant N$ after $N = 2$)

## Hidden Markov Model (HMM)

**Definition 14.1** $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathrm{P}, E)$
**Hidden Markov Model (HMM):**
Is a Markov Chain[def. 13.3] with hidden/latent states $S_j$ that are only partially observable by noisy/indirect observations[def. 14.2]: It is characterized by the 5-tuple of:

1. States[def. 13.1] $\mathcal{S} = \{s_1, \dots, s_n\}$
2. Actions[def. 15.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
3. Observations[def. 14.2] $\mathcal{O}/\mathcal{O}_{s_j} = \{o_1, \dots, o_m\}$
4. Transition Probabilities[def. 13.5] $\mathrm{p}\left(s_i, s_j\right)$
5. Emission/Output Probabilities[def. 14.3] $e_{ij}(t)$



**Definition 14.2 Observations** $\mathcal{O} = \{o_1, \dots, o_l\}$:
Are indirect or noisy observations that are related to the true states $s_j$.

**Definition 14.3**
**Emission/ Output Probabilities** $e_{ij}(t)$:
Given a state $X_t = s_i$ the output probability is the probability of the output random variable $Y_t$ to be in state $o_j$:

$$e_{ij}(t) = \mathbb{P}\left(Y_t = o_j | X_t = s_i\right) \qquad \begin{cases} \forall o_i \in \mathcal{O} \\ \forall s_j \in \mathcal{S} \end{cases} \tag{14.1}$$

# Markov Decision Processes (MDP)

**Definition 15.1** $(\mathcal{S}, \mathcal{A}, P_a, R_a)$
**Markov Decision Process (MDP):** A markov decision process is a *controlled* markov process/chain with an associated reward, where the transition can by steered by an actions. It is characterized by the 4-tuple of:

① States[def. 13.1] $\mathcal{S} = \{s_1, \dots, s_n\}$
② Actions[def. 15.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
③ Transition Probabilities[def. 15.3] $p_a(s_i, s_j)$
④ Rewards[def. 15.4] $r_a(s_i, s_j)$

**Definition 15.2**
**Actions** $\mathcal{A}_{s_i} = \{a_1, \dots, a_m\}$:
Is the set of possible actions from which we can choose at each state and may depend on the state $s_j$ itself.

**Definition 15.3 Transition Probability** $p_a(s_j, s_i)(t)$:
is the probability of a random variable $X_t$ in state $s_i$ to transition into state $s_j$ and depends also on the current action $a$:

$$p_a(s_j, s_i) = p(s_j|s_i, a) = \mathbb{P}(x_{t+1} = s_j|x_t = s_i, a_t = a) \quad \forall s_i, s_j \in \mathcal{S}, \forall a \in \mathcal{A} \tag{15.1}$$

**Definition 15.4 Reward** $r_a(s_i, s_j)$:
is a function or probability distribution that measures the immediate reward and may depend on a any subset of $(x_{t+1}, x_t, a)$:

$$(x_{t+1}, x_t, a) \mapsto R_{t+1} \in \mathcal{R} \subset \mathbb{R} \tag{15.2}$$

Markov decision processes require us to plan ahead. This is because the immediate reward[def. 15.4], that we obtain by greedily picking the best action may result in non-optimal local actions.

## 1. Policies and Values

**Definition 15.5**
**Optimizing Agent/ Decision Making Policy** $\pi(s_i)$:
Is a policy on how to choose an action $a \in \mathcal{A}$ based on a objective/value function[def. 15.8] and can be deterministic or randomized:

$$\pi : \mathcal{S} \mapsto \mathcal{A} \qquad \text{or} \qquad \pi : \mathcal{S} \mapsto \mathbb{P}(\mathcal{A}) \tag{15.3}$$

**Definition 15.6 Discounting Factor** $\gamma$:
Is a factor $\gamma \in [0, 1)$ that signifies that future rewards are less valuable then current rewards.

**Explanation 15.1** (Definition 15.6). *The reason for the discounting factor is that we may for example not even survive long enough to obtain future payoffs.*

**Definition 15.7 Expected Discounted Value** $J(\pi)$:
Is the *discounted* expected (reward) of the whole markov process:

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t))\right] \tag{15.4}$$

**Definition 15.8**
**Value Function** $V^\pi(x)$:
Is the *discounted* expected reward[def. 15.4] of the whole markov process given an inital state $X_0 = x$:

$$V^\pi(x) = J(\pi|X_0 = x) \tag{15.5}$$

$$= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \middle| X_0 = x\right] \tag{15.6}$$

$$\tag{15.7}$$

### 1.1. Calculating the value of $V^\pi$

**Definition 15.9** [proof 16.1]
**Value Iteration:**

$$\begin{aligned} V^\pi(x) &= J(\pi|X_0 = x) \qquad (15.8)\\ &= \mathbb{E}_{x'|x,\pi(x)}\left[r(x, \pi(x)) + \gamma V^\pi(x')\right] \\ &= r(x, \pi(x)) + \gamma\mathbb{E}_{x'|x,\pi(x)}\left[V^\pi(x')\right] \\ &= \boxed{r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, \pi(x)\right) V^\pi\left(x'\right)} \end{aligned}$$

We can now write this for all possible initial states as:

$$\mathbf{V}^\pi = \mathbf{r}^\pi + \gamma P^\pi \mathbf{V}^\pi \iff (\mathbf{I} - \gamma P^\pi)\mathbf{V}^\pi = \mathbf{r}^\pi \tag{15.9}$$

with:

$$\mathbf{V}^\pi = \begin{bmatrix} V^\pi(s_1) \\ \cdots \\ V^\pi(s_n) \end{bmatrix} \qquad \mathbf{r}^\pi = \begin{bmatrix} r^\pi(s_1, \pi(s_1)) \\ \cdots \\ r^\pi(s_n, \pi(s_n)) \end{bmatrix}$$

$$P^\pi = \begin{bmatrix} \mathbb{P}(s_1|s_1, \pi(s_1)) & \mathbb{P}(s_2|s_1, \pi(s_1)) & \cdots & \mathbb{P}(s_n|s_1, \pi(s_1)) \\ \mathbb{P}(s_1|s_2, \pi(s_2)) & \mathbb{P}(s_2|s_2, \pi(s_2)) & \cdots & \mathbb{P}(s_n|s_2, \pi(s_2)) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}(s_1|s_n, \pi(s_n)) & \mathbb{P}(s_2|s_n, \pi(s_n)) & \cdots & \mathbb{P}(s_n|s_n, \pi(s_n)) \end{bmatrix}$$

#### 1.1.1. Direct Mehtods

**Corollary 15.1 LU-decomposition** $\mathcal{O}(n^3)$:
The linear system from eq. (15.9):

$$(\mathbf{I} - \gamma P^\pi)\mathbf{V}^\pi = \mathbf{r}^\pi \tag{15.10}$$

can be solved *directly* using Gaussian elimination in polynomial time $\mathcal{O}(n^3)$.

**Note – invertebility**

If $\gamma < 1$ then $(\mathbf{I} - \gamma P^\pi)$ is full-rank/invertible as $\text{EVs}(P^\pi) \leqslant 1$.

#### 1.1.2. Fixed Point Iteration

**Corollary 15.2 Fixed-Point Iteration** $\mathcal{O}(n \cdot |\mathcal{S}|)$:
The linear system from eq. (15.9) can be solve using *fixed-point iteration*[def. 28.23] in at most $\mathcal{O}(n \cdot |\mathcal{S}|)$ (if every state $s_i$ is connected to every other state $s_j \in \mathcal{S}$)

**Algorithm 15.1 Fixed Point Iteration:**

```
    Input: Inital Guess: $V_0^\pi \overset{\text{i.e.}}{=} 0$
    for $t = 1, \dots, T$ do
        Use the fixed point method:
            $\mathbf{V}_t^\pi = \phi\mathbf{V}_t^\pi = \mathbf{r}^\pi + \gamma P^\pi \mathbf{V}_{t-1}^\pi$    (15.11)
    end for
```

**Corollary 15.3**
**Policy Iterration Contraction** [proof 16.2]:
Fixed point iteration of policy iteration is a contraction[def. 25.60] that leads to a fixed point $V^\pi$ with a rate depending on the discount factor $\gamma$.

$$\begin{aligned} \|V_t^\pi - V^\pi\| &= \|\phi V_{t-1}^\pi - \phi V^\pi\| \\ &\leqslant \gamma\|V_{t-1}^\pi - V^\pi\| = \gamma^t\|V_0^\pi - V^\pi\| \end{aligned} \tag{15.12}$$

**Explanation 15.2.**
- $\gamma \downarrow$: *the less we plan ahead/the smaller we choose $\gamma$ the shorter it takes to converge. But on the other hand we only care greedily about local optima and might miss global optima.*
- $\gamma \uparrow$: *the more we plan ahead/the larger we choose $\gamma$ the longer it takes to converge but we will explore all possibilities. But for to large $\gamma$ we will simply keep exploring without sticking to a optimal poin*

**Note contraction**

For a contraction:
- A unique fixed point exists
- We converge to the fixpoint

### 1.2. Choosing The Policy

**Question** how should we choose the $\pi$? **Idea** compute $J(\pi)$ for *every possible* policy:

$$\pi^* = \arg\max J(\pi) \tag{15.13}$$

**Problem** this is unfortunately infeasible as there exist $m^n = |\mathcal{A}|^{|\mathcal{S}|}$ policies that we need to calculate the value for.

**Note**

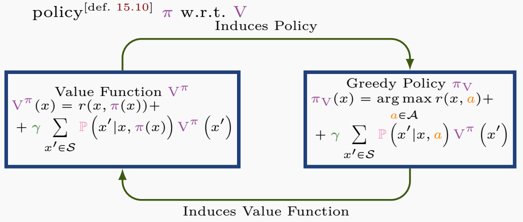
The problem is that $J/V^\pi$ depend on $\pi$ but if we do not know $\pi$ yet we cannot compute those.

#### 1.2.1. Greedy Policy

**Definition 15.10 Greedy Policy:**
**Assuming** we know $V^{\pi t-1}$ then we could choose a greedy policy:

$$\begin{aligned} a^* &= \pi_t(x) \qquad (15.14) \\ &:= \arg\max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^{\pi t-1}\left(x'\right) \end{aligned}$$

① Given a policy $\pi$ however we can calculate a value function $V^\pi$
② Given a value function $V$ we can induce a greedy policy[def. 15.10] $\pi$ w.r.t. $V$


Induces Policy

Value Function $V^\pi$
$V^\pi(x) = r(x, \pi(x)) +$
$+ \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, \pi(x)\right) V^\pi\left(x'\right)$

Greedy Policy $\pi_V$
$\pi_V(x) = \arg\max_{a \in \mathcal{A}} r(x, a) +$
$+ \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^\pi\left(x'\right)$

Induces Value Function

**Theorem 15.1 Optimality of Policies** **[Bellman]:**
A policy $\pi_V$ is optimal if and only if it is greedy w.r.t. its induced value function

**Definition 15.11 Non-linear Bellman Equation:** States that the optimal value is given by the action/policy that maximizes the value function eq. (15.8):

$$V^*(x) = \max_{a \in \mathcal{A}}\left[r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^*\left(x'\right)\right] \tag{15.15}$$

$$:= \max_{a \in \mathcal{A}} Q^*(x, a) \tag{15.16}$$

**Note**

This equation is non-linear due to the max in comparison to eq. (15.8).

#### 1.2.2. Policy Iteration

**Algorithm 15.2 Policy Iteration:**

```
    Initialize: Random Policy: $\pi$
    while Not converged $t = t + 1$ do
        Compute $V^{\pi t}(x)$
        $V^{\pi t}(x) = r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, \pi_t(x)\right) V^{\pi t}\left(x'\right)$
        Compute greedy policy $\pi_G$:
        $\pi_G(x) = \arg\max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^{\pi t}\left(x'\right)$
        Set $\pi_{t+1} \leftarrow \pi_G$
    end while
```

**Algorithm 15.2**

**Pros**
- Monotonically improves $V^{\pi t} \geqslant V^{\pi t-1}$
- is guaranteed to converge to an optimal policy/solution $\pi^*$ in polynomial *#iterations*: $\mathcal{O}\left(\frac{n^2 m}{1-\gamma}\right)$

**Cons**
- Complexity *per iteration* requires to evaluate the policy $V^\pi$ which requires us to solve a linear system.

#### 1.2.3. Value Iteration

**Definition 15.12 Value to Go** $V_t(x)$:
Is the maximal expected reward if we *start* in state $x$ and have $t$ time steps to go.

**Algorithm 15.3 Value Iteration** [proof 16.3]:

```
    Initialize: $V_0(x) = \max_{a \in \mathcal{A}} r(x, a)$
    for $t = 1, \dots, \infty$ do
        Compute:
    $Q_t(x, a) = r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V_{t-1}\left(x'\right)$    $\forall a \in \mathcal{A}$  $\forall x \in \mathcal{S}$
        for all $x \in \mathcal{S}$ let:
            $V_t(x) = \max_{a \in \mathcal{A}} Q_t(x, a)$
        if $\max_{x \in \mathcal{S}}|V_t(x) - V_{t-1}(x)| \leqslant \epsilon$ then
            break
        end if
    end for
    Choose greedy policy $\pi_{V_t}$ w.r.t. $V_t$
```

**Corollary 15.4** [proof 16.4]
**Value Iteration Contraction:**
Algorithm 15.3 is guaranteed to converge to a $\epsilon$ optimal policy:

$$\left\|\left(V_t - V^*\right)\right\|_\infty \leqslant \gamma^t \left\|\left(V_0 - V^*\right)\right\|_\infty \tag{15.17}$$

$$\implies t \approx \ln\frac{\gamma}{\epsilon}\left\|V_0 - V^*\right\|_\infty \quad \text{for} \quad \left\|\left(V_t - V^*\right)\right\|_\infty \leqslant \epsilon$$

**Algorithm 15.3**

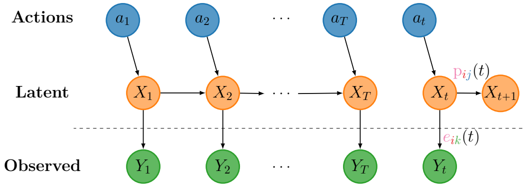
**Pros**
- Finds $\epsilon$-optimal solution in polynomial *#iterrations* $\mathcal{O}(\ln\frac{1}{\epsilon})$[cor. 15.4].
- Complexity *per iteration* requires us to solve a linear system $\mathcal{O}(m \cdot n \cdot s) = \mathcal{O}(|\mathcal{A}| \cdot |\mathcal{S}| \cdot s)$ where $s$ is the number of states we can reach.
  For small $s$ and small $m$ we are roughly linear w.r.t. the states $\mathcal{O}(n) = \mathcal{O}(|\mathcal{S}|)$

**Cons**
- Only $\epsilon$-optimal solution.

# Partially Observable MDP (POMDP)

**Definition 16.1** $(\mathcal{S}, \mathcal{A}, \mathcal{O}, P_a, E, R_a)$
**Partially Observable Markov Decision Process:**
A (POMDP) is a markov decision process[def. 15.1] with hidden markov states[def. 14.1]. It is characterized by the 6-tuple of:

① States[def. 13.1] $\mathcal{S} = \{s_1, \dots, s_n\}$
② Actions[def. 15.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
③ Observations[def. 14.2] $\mathcal{O}\,/\,\mathcal{O}_{s_j} = \{o_1, \dots, o_m\}$
④ Transition Probabilities[def. 15.3] $p_a(s_i, s_j)$
⑤ Emission/Output Probabilities[def. 14.3] $e_{ij}(t)$
⑥ Rewards[def. 15.4] $r_a(s_i, s_j)$



**Explanation 16.1.**
*Now our agent has only some indirect noisy observation of true state.*

# 1. POMDPs as MDPs

POMDPs can be converted into *belief state*?? MDPs[def. 15.1] by introducing a *belief state space* $\mathcal{B}$.

**Definition 16.2 History** $H_t$**:**
Is a sequence of actions, observations and rewards:

$$H_t = \{\{a_0, o_0, r_0\}, \ldots, \{a_0, o_0, r_0\}\}$$

**Definition 16.3 Belief State Space** $\mathcal{B}$**:** Is a $|\mathcal{S}| - 1$ dimensional simplex or ($|S|$-dimensional probability vector[def. 25.67]) whose elements $b$ are probabilities:

$$\mathcal{B} = \Delta(|\mathcal{S}|) = \left\{ b_t \in [0,1]^{|\mathcal{S}|} \;\middle|\; \sum_{x=1}^{n} b_t(x) = 1 \right\} \tag{16.1}$$

**Definition 16.4 Belief State** $b_t \in \mathcal{B}$**:** Is a probability distribution over the states $\mathcal{S}$ conditioned on the history $H_t$[def. 16.2].

## 1.1. Transition Model

**Definition 16.5** [proof 16.5]
**POMDP State/Posterior Update:**

$$\begin{aligned} b_{t+1}(s_i) =& \mathbb{P}(X_{t+1} = s_i | Y_{t+1} = o_k) \\ =& \frac{1}{Z} \underline{\mathbb{P}(Y_{t+1} = o_k | X_{t+1} = s_i, a_t)} \\ &\cdot \sum_{s_j \in Pa(s_i)} b_t(s_j) \mathbb{P}(X_{t+1} = s_i | X_t = s_j, a_t) \end{aligned} \tag{16.2}$$

**Definition 16.6 Stochastic Observation Model:**

$$\underline{\mathbb{P}(Y_{t+1} = o_k | b_t, a_t)} = \sum_{s_i \in \mathcal{S}} b_t(s_i) \mathbb{P}(Y_{t+1} = o_k | X_t = s_i, a_t) \tag{16.3}$$

## 1.2. Reward Function

**Definition 16.7 POMDP Reward Function:**

$$r(b_t, a_t) = \sum_{s_j \in \mathcal{S}} b_t(s_i) r(s_i, a_t) \tag{16.4}$$

**Note**

For finite horizon $T$, the set of reachable belief states is finite however exponential in $T$.

add defintion of simplex to math appendix which is basically this.

# 2. Proofs

## 2.1. Markov Decision Processes

**Proof 16.1.** *[def. 15.8]*

$$\begin{aligned} V^{\pi}(x) &= \mathbb{E}_{X_{1:\infty}} \left[ \sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &= \mathbb{E}_{\mathbf{X}} \left[ \gamma^0 r(X_0, \pi(X_0)) + \sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &\overset{\substack{\gamma^0=1 \\ X_0=x}}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}} \left[ \sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &\overset{\text{re-index}}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}} \left[ \sum_{t=0}^{\infty} \underline{\gamma^{t+1}} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x \right] \\ &= r(x, \pi(x)) + \underline{\gamma} \mathbb{E}_{\mathbf{X}} \left[ \sum_{t=0}^{\infty} \underline{\gamma^t} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x \right] \\ &= \quad r(x, \pi(x)) \\ &+ \gamma \mathbb{E}_{X_1} \left[ \mathbb{E}_{X_{2:\infty}} \left[ \sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \middle| X_1 = x' \right] \middle| X_0 = x \right] \\ &\overset{\text{law } 31.7}{=} r(x, \pi(x))) \\ &+ \gamma \sum_{x' \in S} \mathbb{P}(x' | x, \pi(x) \mathbb{E}_{X_{2:\infty}} \left[ \sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \middle| X_1 = x' \right] \\ &\overset{\text{eq. (13.13)}}{=} r(x, \pi(x))) \\ &+ \gamma \sum_{x' \in S} \mathbb{P}(x' | x, \pi(x) \underline{\mathbb{E}_{X_{2:\infty}} \left[ \sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \middle| X_0 = x' \right]} \\ &= \quad r(x, \pi(x)) + \gamma \sum_{x' \in S} \mathbb{P}\left(x' | x, \pi(x)\right) \underline{V^{\pi}\left(x'\right)} \end{aligned}$$

**Proof 16.2** ([cor. 15.3]). *Consider* $V, V' \in \mathbb{R}^n$ *and let* $\phi$*:*

$$\phi x := r^{\pi} + \gamma P^{\pi} x \qquad \Longrightarrow \qquad \phi V^{\pi} = V^{\pi}$$

*then it follows:*

$$\begin{aligned} \left\| \phi V - \phi V' \right\| &= \left\| \cancel{r^{\pi}} + \gamma P^{\pi} V - \cancel{r^{\pi}} - \gamma P^{\pi} V' \right\| \\ &= \left\| \gamma P^{\pi} \left( V - V' \right) \right\| \\ &\overset{\text{eq. (25.88)}}{\leqslant} \gamma \left\| P^{\pi} \right\| \cdot \left\| \left( V - V' \right) \right\| \\ &\overset{\text{i.e. } L_2}{\leqslant} \gamma \cdot 1 \cdot \left\| \left( V - V' \right) \right\|_2 \end{aligned}$$

**Proof 16.3.** *algorithm 15.3*

$$\begin{aligned} V_0(x) &= \max_{a \in \mathcal{A}} r(x, a) \\ V_1(x) &= \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x' | x, a\right) V_0\left(x'\right) \\ V_{t+1}(x) &= \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x' | x, a\right) V_t\left(x'\right) \end{aligned}$$

**Proof 16.4.** *[cor. 15.4] Let* $\phi : \mathbb{R}^n \mapsto \mathbb{R}^n$*, with:*

$$\left(\phi V^{*}\right)(x) = Q(x, a) = \max_{a} \left[ r(x, a) + \gamma \sum_{x'} \mathbb{P}\left(x' | x, a\right) \right]$$

*Bellman's theorem 15.1* $\qquad \phi V^{*} = V^{*}$
*and consider* $V, V' \in \mathbb{R}^n$

$$\begin{aligned} \left\| \phi V - \phi V' \right\|_{\infty} &= \max_{x} \left| (\phi V)(x) - (\phi V')(x) \right| \\ &= \max_{x} \left| \max_{a} Q(x, a) - \max_{a'} Q'(x, a') \right| \\ &\overset{\text{Property } 22.8}{\leqslant} \max_{x} \max_{a} \left| Q(x, a) - Q'(x, a) \right| \\ = \max_{x, a} | \cancel{r} &+ \gamma \sum_{x'} \mathbb{P}\left(x' | x, a\right) V(x') - \cancel{r} - \gamma \sum_{x'} \mathbb{P}\left(x' | x, a\right) V'(x') | \\ &= \gamma \max_{x, a} \left| \sum_{x'} \mathbb{P}\left(x' | x, a\right) \left( V(x') - V'(x') \right) \right| \\ &\overset{\text{eq. (25.88)}}{\leqslant} \gamma \max_{x, a} \overbrace{\left| \sum_{x'} \mathbb{P}\left(x' | x, a\right) \right|}^{\leqslant 1} \cdot \left| \left( V(x') - V'(x') \right) \right| \\ &\leqslant \gamma \cdot 1 \cdot \left\| \left( V(x') - V'(x') \right) \right\|_{\infty} \end{aligned}$$

**Note**

For the policy iteration the calculation was easier as the rewards canceled, however here we have the max.

## 2.2. MDPs

**Proof 16.5.** *Defintion 16.5 Directly by definition 7.5 and its corresponding proof 10.4 with additional action* $a_t$*:*

$$\begin{aligned} b_{t+1}(s_i) &= \mathbb{P}(X_{t+1} = s_i | y_{t+1}) \\ &= \frac{1}{Z} \mathbb{P}(y_{1:t+1} | s_i) \sum_{j=1} \overbrace{\underbrace{\mathbb{P}(X_t = s_j | y_{1:t})}_{b_t(s_j)} \mathbb{P}(s_i | s_j)}^{\mathbb{P}(X_{t+1} = s_i | y_{1:t})} \end{aligned}$$

# Reinforcement Learning

Now we are working with an *unknown* MDP[def. 15.1] meaning that:

1. we do no longer know the transition model[def. 15.3]
2. We do no longer know the reward function
3. We might not even know all the states

**However** we can observe them when taking steps.

**Note**

- Reinforcement learning is different than supervised learning as the data is no longer i.i.d. (data depends on previous action).
- Need to do exploration vs exploitation in order to learn policy and reward functions.

**Definition 17.1 Agent:**
Is the *learner/decision maker* of our *unknown* MDP.

**Definition 17.2 Environment:** Is the representation of the world in which our agents acts.

**Definition 17.3 On-Policy Learning:** At any given time the agent has full control which actions to pick.

**Definition 17.4 Off-Policy Learning:** The agent has to fix a policy in advance based on behavioral observations.

**Definition 17.5 Trajectory** $\tau$**:**
Is a set of consecutive 3-tuples of states, actions and rewards:

$$\tau = \{s_t, a_t, r_t\} \qquad t = 1, \dots, \tau \tag{17.1}$$

**Definition 17.6 Episodic Learning:** Is a setting where we generate multiple $K$-episodes of different trajectories $\left\{\tau^{(k)}\right\}_{i=1}^{K}$ from which the agent can learn.

**Explanation 17.1.** *For each episode the agent starts in a random state and follows a policy.*

## 1. Model Based Reinforcement Learning

**Proposition 17.1 Model Based RL:**
Try to learn the MDP [def. 15.1] by:

1. Estimating
   - the transition probabilities[def. 15.3] $\quad p_a(s_i, s_j)$
   - the reward function[def. 15.4] $\quad r(b_t, a_t)$
2. Optimizing the policy of the estimated MDP

### 1.1. Estimating Transitions and Rewards

**Formula 17.1 Estimating Transitions and Rewards:**
Given a data set $D = \{(\mathbf{x}_0, a_0, r_0, \mathbf{x}_1), (\mathbf{x}_1, a_1, r_1, \mathbf{x}_2), \dots\}$ we estimate the transitions and rewards using a categorical distribution[def. 32.23]:

$$N_{s_i|s_j,a} := \sum_{k=1}^{t} \delta_{\left(X_{k+1}=s_i|X_k=s_j,A_k=a\right)} \tag{17.2}$$

$$N_{s_j,a} := \sum_{k=1}^{t} \delta_{\left(X_k=s_j,A_k=a\right)} \tag{17.3}$$

$$p_a(s_i, s_j) \approx \frac{N_{s_i|s_j,a}}{N_{s_j,a}} \tag{17.4}$$

$$r(s_i, a) \approx \frac{1}{N_{s_i,a}} \sum_{k=1}^{t} \delta_{(X_k=s_i,A_k=a)} r(X_k, A_k) \tag{17.5}$$

### 1.2. Choosing the next step

How should we choose the action $a \in \mathcal{A}$ in order to balance exploration vs exploitation?

### 1.3. $\epsilon_t$ Greedy Learning

**Algorithm 17.1 Epsilon Greedy Learning:**

```
for $t = 1, \dots, T$ do
    Pick next action
        $a_t = \begin{cases} \arg\max_a Q_t(a) & \text{with probability } \epsilon_t \\ \text{random } a & \text{with probability } 1 - \epsilon_t \end{cases}$
end for
```

**Corollary 17.1 Necessary Condition for Convergence:**
If the sequence $\epsilon_t$ satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \epsilon_t < \infty, \qquad \sum_t \epsilon_t^2 < \infty \qquad (\text{i.e. } \epsilon_t = 1/t) \tag{17.6}$$

then algorithm 17.1 converges to an optimal policy with probability one.

add general definition of RM conditions and sequence

| **Pros** | **Cons** |
|---|---|
| • Simple | • Clearly sub optimal actions are not eliminated fast enough |

### 1.4. The $R_{\max}$ Algorithm

**Algorithm 17.2** [Brafman & Tennenholz '02]
**R-max Algorithm:**

```
    Initialize every state with:
        $\hat{r}(s_t, a) = R_{\max} \qquad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1 \qquad (17.7)$
    Set min. number $\Delta$ of observations for policy update
    Compute Policy $\pi_1$ of the MDP[def. 15.1] using $(\hat{p}, \hat{r})$:
        $\pi_t$
for $k = 1, \dots, K$ do
    Choose $a = \pi_t(x_t)$ and observe $(s, r)$
    Calculate:
        $N_{\mathbf{x}_t,a} += 1 \qquad r(x_t, a) += r(x_t, a) \qquad (17.8)$
        $N_{\mathbf{x}_{t+1}|\mathbf{x}_t,a} += 1 \qquad (17.9)$
    if $k == \Delta$ then
        Re-calculate (based on eqs. (17.4) and (17.5)):
        $\hat{r}(s_t, a) = R_{\max} \qquad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1$
            and update the policy $\pi_t = \pi_t(\hat{p}, \hat{r})$
    end if
end for
```

**Note**

Other ways of updating the policy at certains times exist.

**Problems**

**Cons**

- Memory: for all $a \in \mathcal{A}$, $\mathbf{x}_{t+1}, \mathbf{x}_t \in \mathcal{X}$ we need to store $\hat{p}_a(x_{t+1}|x_t, a)$ and $\hat{r}(s_t, a)$ which results in $|\mathcal{S}|^2|\mathcal{A}|$ (for dense MDP).
- Computation Time: We need to calculate the $\pi_t$ using policy (**??** 1.2.2) or value iteration (**??** 1.2.3) $|\mathcal{A}| \cdot |\mathcal{S}|$ whenever we update out policy.

#### 1.4.1. How many transitions do we need?

**Proposition 17.2** [proof 17.1]
**Number of Samples to bound Reward:**

$$\mathbb{P}\left(\hat{r}(s, a) - r(s, a) \leqslant \epsilon\right) \geqslant 1 - \delta \iff n \in \mathcal{O}\left(\frac{R_{\max^2}}{\epsilon^2} \log \frac{1}{\delta}\right) \tag{17.10}$$

**Theorem 17.1 :** Every $T$ timesteps, with high probability, $R_{\max}$ either:

- Obtain near optimal reward, or
- Visits at leas one unkown state-action pair

**Theorem 17.2 Performance of R-max:** With probability $\delta - 1$, $R_{\max}$ will reach an $\epsilon$-optimal policy in a number of steps that is polynomial in $|\mathcal{X}|, |\mathcal{A}|, T, 1/\epsilon$.

## 2. Model Free Reinforcement Learning

**Proposition 17.3 Model Free RL:**
Tries to estimate the value function[def. 15.8] directly in order to act greedily upon it.

- Policy Gradient Methods
- Actor Critic Methods

### 2.1. Temporal Difference Learning (TD)

**Assume** we fix a random intial policy $\pi$ and s.t. we have $\hat{V}_0^{\pi}(s_j)$.
**Goal**: want to calculate an unknown value function $V^{\pi}$.
If the reward and the next states are stochastic variables $(R, X)$ we can calculate the reward usingeq. (15.8):

$$\hat{V}^{\pi}(x_t) = \mathbb{E}_{X_{t+1},R}\left[R + \gamma \hat{V}^{\pi}(X')|X, a\right] \tag{17.11}$$

Now assume we observe a single example

$$(X_{t+1} = s_j, a, r, X_t = s_i)$$

then we can use monte carlos sampling[def. 33.6] with a single sample to approximate the expectation ineq. (17.11):

$$\hat{V}_{t+1}^{\pi}(s_i) = r + \gamma \hat{V}_t^{\pi}(s_j)$$

**Problem**: high variance of estimates ⇒ average with previous estimate.

**Definition 17.7 Temporal Difference (TD) Learning:**

$$\hat{V}(x_{t+1}) = (1 - \alpha_t)\hat{V}(x_t) + \alpha_t\left(r + \gamma \hat{V}(x_{t+1})\right) \tag{17.12}$$

**Corollary 17.2 Necessary Condition for Convergence:**
If the learning rate $\alpha_t$ satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \alpha_t < \infty, \qquad \sum_t \alpha_t^2 < \infty \qquad (\text{i.e. } \alpha_t = 1/t) \tag{17.13}$$

and all state-action pairs $(s_i, a_j)$ are chosen infinitely often, then we converge to the correct value function:

$$\mathbb{P}\left(\hat{V} \to \hat{V}^{\pi}\right) = 1 \tag{17.14}$$

### 2.2. Q-Learning

**Definition 17.8 Action Value/Q-Function:**

$$Q \tag{17.15}$$

#### 2.2.1. Policy Gradients

#### 2.2.2. Actor-Critic Methods

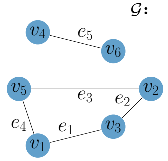
## 3. Proofs

**Proof 17.1.** *proposition 17.2 using hoeffdings bound*[def. 31.36] *with* $\delta$ *and* $b - a = R_{\max}$.

# Graph Theory

**Definition 18.1 Graph** $\mathcal{G}$:

A graph $\mathcal{G}$ is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a finite set of vertices $\mathcal{V}$[def. 18.4] and a multi set[def. 19.3] of edges $\mathcal{E}$[def. 18.10].



**Definition 18.2 Order** $n = |\mathcal{V}|$:
The order of a graph is the cardinality of its vertix set.

**Definition 18.3 Size** $m = |\mathcal{E}|$:
The size of a graph is the number of its edges.

**Corollary 18.1 $n$-Graph:** Is a graph $\mathcal{G}$[def. 18.1] of order $n$.

**Corollary 18.2 $(p, q)$-Graph:** Is a graph $\mathcal{G}$[def. 18.1] of order $p$ and size $q$.

## 1. Vertices

**Definition 18.4 Vertices/Nodes** $\mathcal{V}$:
Is a set of entities of a graph connected and related by edges in some way:

**Definition 18.5 Neighborhood $N(v)$:** The neighborhood of a vertix $v_i \in \mathcal{V}$ is the set of all adjacent vertices:

$$N(v_i) = \{v_k \in \mathcal{V} : \exists e_k = \{v_i, v_j\} \in \mathcal{E}, \forall v_j \in \mathcal{E}\} \qquad (18.1)$$

### 1.0.1. Adjacency Matrix

**Definition 18.6 (unweighted) Adjacency Matrix A:**
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its *adjacency matrix* is a square matrix $\mathbf{A} \in \mathbb{N}^{n,n}$ defined as:

$$\mathbf{A}_{i,j} := \begin{cases} 1 & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \qquad (18.2)$$

**Definition 18.7 weighted Adjacency Matrix A:**
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its *weighted adjacency matrix* is a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ defined as:

$$\mathbf{A}_{i,j} := \begin{cases} \theta_{ij} & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \qquad (18.3)$$

**Diagonal Elements**

For a graph without self-loops the diagonal elements of the adjacency are all zero.

### 1.0.2. Degree Matrix

**Definition 18.8 Degree of a Vertix** $\delta$:
The degree of a vertix $v$ is the cardinality of the neighborhood[def. 18.5] – the number of adjacent vertices:

$$\deg(v_i) = \delta(v) = |N(v)| = \sum_{j=1}^{j<i} \mathbf{A}_{ij} \qquad (18.4)$$

**Definition 18.9 Degree Matrix D:**
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its degree matrix is a diagonal matrix $\mathbf{D} \in \mathbb{N}^{n,n}$ defined as:

$$\mathbf{D}_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \qquad (18.5)$$

## 2. Edges

**Definition 18.10 Edges** $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$:
Represent some relation between edges[def. 18.4] and are represented by two-element subset sets of the vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \quad \Longleftrightarrow \quad v_i \text{ and } v_j \text{ connected} \qquad (18.6)$$

**Proposition 18.1 Number of Edges:** A graph $\mathcal{G}$ with $n = |\mathcal{V}|$ has between $\left[0, \frac{1}{2}n(n-1)\right]$ edges.

## 3. Subgraph

**Definition 18.11 Subgraph** $\mathcal{H} \subseteq \mathcal{G}$:
A graph $\mathcal{H} = (U, F)$ is a *subgraph* of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ iff:

$$U \subseteq \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \qquad (18.7)$$

## 4. Components

**Definition 18.12 Component:** A connected component of a graph $\mathcal{G}$ is a *connected*[def. 18.20] subgraph[def. 18.11] of $\mathcal{G}$ that is *maximal by inclusion* – there exist no larger connected containing subgraphs.
The number of components of a graph $\mathcal{G}$ is defined as $c(\mathcal{G})$.

## 5. Walks, Paths and cycles

**Definition 18.13 Walk:** A walk of a graph $\mathcal{G}$ as a sequence of vertices with corresponding edges:

$$W = \{v_k, v_{k+1}\}_k^K \in \mathcal{E} \qquad (18.8)$$

**Definition 18.14 Length of a Walk $K$:** Is the number of edges of that Walk.

**Definition 18.15 Path $P$:** Is a walk of a graph $\mathcal{G}$ where all visited vertics are distinct (no-repetitions).

**Attention:** Some use the terms walk for paths and simple paths for paths.

**Definition 18.16 Cycle:** Is a path[def. 18.15] of a graph $\mathcal{G}$ where the last visited vertix is the one from which we started.

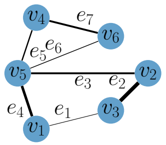
## 6. Different Kinds of Graphs

## 7. Weighted Graph

**Definition 18.17 Weighted Graph:**

Is a graph $\mathcal{G}$ where edges are associated with a weight:

$$\exists \theta_i := \text{weight}(e_i) \qquad \forall e_i \mathcal{E}$$

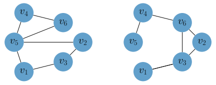


## 8. Spanning Graphs

**Definition 18.18 Spanning Graph:**

Is a subgraph[def. 18.11] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for which it holds:

$$U = \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \qquad (18.9)$$



### 8.1. Minimum Spanning Graph

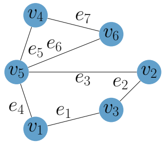
**Definition 18.19 Minimum Spanning Graph:** Is a spanning graph[def. 18.18] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with minimal weights/distance of the edges.

## 9. Connected Graphs

**Definition 18.20 (Weakly) Connected Graph:**

Is a graph $\mathcal{G}$[def. 18.1] where there exists a path between any two vertices:

$$\exists P(v_i, \ldots, v_j) \qquad \forall v_i, v_j \in \mathcal{V} \qquad (18.10)$$



**Corollary 18.3 Strongly Connected Graph:** A directed Graph[def. 18.22] is called strongly connected if every nodes is *reachable* from every other node.

**Corollary 18.4 Components of Connected Graphs:** A connected Graph[def. 18.20] consist of one component $c(\mathcal{G}) = 1$.

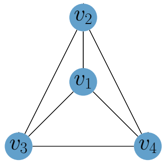
### 9.1. Fully Connected/Complete

**Definition 18.21 Fully Connected/Complete Graph:**

Is a connected graph $\mathcal{G}$[def. 18.20] where each node is connected to every other node.

$$\exists e \forall \{v_i, v_j\} \qquad \forall v_i, v_j \in \mathcal{V} \qquad (18.11)$$

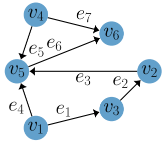
$$|\mathcal{V}| = \frac{1}{2}|\mathcal{V}|(|\mathcal{V}| - 1) \qquad (18.12)$$



### 9.2. Directed Graphs

**Definition 18.22 Directed Graph/Digraph (DG):**

A directed graph $\mathcal{G}$ is a graph where edges are direct arcs[def. 18.23].



**Definition 18.23 Directed Edges/Arcs:** Represent some *directional* relationship between edges[def. 18.4] and are represented by *ordered* two-element subset sets of vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \quad \Longleftrightarrow \quad v_i \text{ goes to } v_j \qquad (18.13)$$

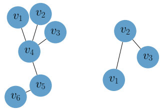
### 9.3. Trees And Forests

### 9.3.1. Acyclic Graphs

**Definition 18.24 Acyclic Graphs:** Are graphs[def. 18.1] where no cycles[def. 18.16] exist.

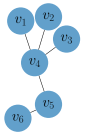
**Definition 18.25 Forests:**

Are acyclic graphs[def. 18.24]:



**Definition 18.26 Trees:**

Are acyclic graphs[def. 18.24] that are connected[def. 18.20].



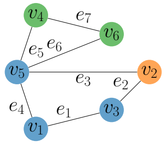
## 10. Graph Layering

**Definition 18.27 Graph Layering:**

Given a graph $\mathcal{G}$ a layering of the graph is a partition of its node set $\mathcal{V}$[def. 18.4] into subsets

$$\{\mathcal{V}_1, \ldots, \mathcal{V}_L\} \subseteq \mathcal{V}$$

$$\text{s.t.} \quad \mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_L \qquad (18.14)$$



## 11. Bisection Algorithms

### 11.1. Local Approaches

### 11.2. Global Approaches

### 11.2.1. Spectral Decomposition

**Definition 18.28 Graph Laplacian (Matrix)** $\mathbf{L}(\mathcal{G})$:
Given a graph with $n$ vertices and $m$ edges has a graph laplacian matrix defined as:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \qquad l_{ij} := \begin{cases} -1 & \text{if } i \neq j \text{ and } e_{ij} \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } e_{ij} \notin \mathcal{E} \\ \deg(v_i) & \text{if } i = j \end{cases} \qquad (18.15)$$

**Corollary 18.5 title:**

### 11.2.2. Inertial Bisection

# Math Appendix

## Set Theory

**Definition 19.1 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 19.2 Empty Set** $\{\}/\varnothing$**:**
is the unique set having no elements/cardinality[def. 19.5] zero.

**Definition 19.3 Multiset/Bag:** Is a set-like object in which multiplicity[def. 19.4] matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

**Definition 19.4 Multiplicity:** The multiplicity $n_a$ of a member $a$ of a multiset[def. 19.3] $\mathcal{S}$ is the number of times it appears in that set.

**Definition 19.5 Cardinality** $|S|$**:** Is the number of elements that are contained in a set.

**Definition 19.6 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 19.7 Closure:** A set is *closed* under an operation $\Omega$ if performance of that operations onto members of the set always produces a member of that set.

**Definition 19.8 Bounded Set:** A set $S \subset \mathbb{R}^n$is *bounded* if there exists a constant $K$ s.t. the absolute value of every component of every element of $S$ is less or equal to K.

### 1. Number Sets

#### 1.1. The Real Numbers $\mathbb{R}$

#### 1.1.1. Intervals

**Definition 19.9 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 $b$:

$$[a, b] = \{x \in \mathbb{R} \,|\, a \leqslant x \leqslant b\} \tag{19.1}$$

**Definition 19.10 Open Interval** $(a, b)$**:**
The open interval of $a$ and $b$ is the set of all real numbers that are within $a$ and $b$:

$$(a, b) = \{x \in \mathbb{R} \,|\, a < x \leqslant <\} \tag{19.2}$$

#### 1.2. The Rational Numbers $\mathbb{Q}$

**Example 19.1 Power Set/Cardinality of** $S = \{x, y, z\}$**:**
The subsets of S are:
$\{\varnothing\}$, $\{x\}$, $\{y\}$, $\{z\}$, $\{x, y\}$, $\{x, z\}$, $\{y, z\}$, $\{x, y, z\}$ and hence the power set of $S$ is $\mathcal{P}(S) = \{\{\varnothing\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

### 2. Set Functions

#### 2.1. Submoduluar Set Functions

**Definition 19.11 Submodular Set Functions:** A submodular function $f : 2^\Omega \mapsto \mathbb{R}$ is a function that satisifies:

$$f(A \cup \{x\}) - f(A) \geqslant f(B \cup \{X\}) - F(B) \qquad \begin{aligned} &\forall A \subseteq B \subset \Omega \\ &\{x\} \in \Omega \setminus B \end{aligned} \tag{19.3}$$

**Explanation 19.1** (Definition 19.11)**.** *Addaing an element $x$ to the the smaller subset $A$ yields at least as much information/value gain as adding it to the larger subset $B$.*

**Definition 19.12 Montone Submodular Function:**
A *monotone* submodular function is a submodular function[def. 19.11] that satisifies:

$$f(A) \leqslant f(B) \qquad \forall A \subseteq B \subseteq \Omega \tag{19.4}$$

**Explanation 19.2** (Definition 19.12)**.** *Adding more elements to a set will always increase the information/value gain.*

#### 2.2. Complex Numbers

**Definition 19.13 Complex Conjugate** $\bar{z}$**:**
The complex conjugate of a complex number $z = x + iy$ is defined as:

$$\bar{z} = x - iy \tag{19.5}$$

add kind of wikipedia image

**Corollary 19.1 Complex Conjugate Of a Real Number:**
The complex conjugate of a real number $x \in \mathbb{R}$ is $x$:

$$\bar{x} = x \qquad \Longrightarrow \qquad x \in \mathbb{R} \tag{19.6}$$

**Formula 19.1 Euler's Formula:**

$$e^{\pm ix} = \cos x \pm i \sin x \tag{19.7}$$

**Formula 19.2 Euler's Identity:**

$$e^{\pm i} = -1 \tag{19.8}$$

**Note**

$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\, 2\pi k, \qquad k \in \mathbb{N} \tag{19.9}$$

## Sequences&Series

**Definition 20.1 Index Set:** Is a set[def. 19.1] $A$, 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.

$$f : A \mapsto S \qquad A \in \mathbb{N} \tag{20.1}$$

**Definition 20.2 Sequence** $(a_n)_{n\in A}$**:**
A sequence is an by an *index set $A$ enumerated* multiset[def. 19.3] (repetitions are allowed) of objects in which *order does matter*.

**Definition 20.3 Series:** is an infinite ordered set of terms combined together by addition.

### 1. Types of Sequences

#### 1.1. Arithmetic Sequence

**Definition 20.4 Arithmetic Sequence:** Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \ldots)$.

$$t_n = t_0 + nd \qquad d \text{ :difference between two terms} \tag{20.2}$$

#### 1.2. Geometric Sequence

**Definition 20.5 Geometric Sequence:** Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \ldots)$.

$$t_n = t_0 \cdot r^n \qquad r \text{ :ratio between two terms} \tag{20.3}$$

# Logic

## 1. Boolean Algebra

### 1.1. Basic Operations

**Definition 21.1 Conjunction/AND** $\wedge$:

**Definition 21.2 Disjunction/OR** $\vee$:

**Definition 21.3 Negation/NOT** $\neg$:

#### 1.1.1. Expression as Integer

If the truth values $\{0,1\}$ are interpreted as integers then the basic operations can be represent with basic arithmetic operations.

$$\begin{aligned} x \wedge y &= xy = \min(x,y) \\ x \vee y &= x + y = \max(x,y) \\ \neg x &= 1 - x \\ x \oplus y &= (x+y)\cdot(\neg x + \neg y) = x \cdot \neg y + \neg x \cdot y \end{aligned}$$

**Note: non-linearity of XOR**

$$(x+y)\cdot(\neg x + \neg y) = -x^2 - y^2 - 2xy + 2x + 2y$$

### 1.2. Boolean Identities

**Property 21.1 Idempotence:**

$$x \wedge x \equiv x \qquad \text{and} \qquad x \vee x \equiv x \tag{21.1}$$

**Property 21.2 Identity Laws:**

$$x \wedge \text{true} \equiv x \qquad \text{and} \qquad x \vee \text{false} \equiv x \tag{21.2}$$

**Property 21.3 Zero Law's:**

$$x \wedge \text{false} \equiv \text{false} \qquad \text{and} \qquad x \vee \text{true} \equiv \text{true} \tag{21.3}$$

**Property 21.4 Double Negation:**

$$\neg\neg x \equiv x \tag{21.4}$$

**Property 21.5 Complementation:**

$$x \wedge \neg x \equiv \text{false} \qquad \text{and} \qquad x \vee \neg x \equiv \text{true} \tag{21.5}$$

**Property 21.6 Commutativity:**

$$x \vee y \equiv y \vee x \qquad \text{and} \qquad x \wedge y \equiv y \wedge x \tag{21.6}$$

**Property 21.7 Associativity:**

$$(x \vee y) \vee z \equiv x \vee (y \vee z) \tag{21.7}$$

$$(x \wedge y) \vee z \equiv x \vee (y \wedge z) \tag{21.8}$$

**Property 21.8 Distributivity:**

$$x \vee (y \wedge z) \equiv (x \vee y) \wedge (x \vee z) \tag{21.9}$$

$$x \wedge (y \vee z) \equiv (x \wedge y) \vee (x \wedge z) \tag{21.10}$$

**Property 21.9 De Morgan's Laws:**

$$\neg(x \vee z) \equiv (\neg x \wedge \neg y) \tag{21.11}$$

$$\neg(x \wedge z) \equiv (\neg x \vee \neg y) \tag{21.12}$$

**Note**

The algebra axioms come in pairs that can be obtained by interchanging $\wedge$ and $\vee$.

### 1.3. Normal Forms

**Definition 21.4 Literal** [example 21.1]:
Literals are atomic formulas or their negations

**Definition 21.5 Negation Normal Form (NNF):** A formula $F$ is in negation normal form is the negation operator $\neg$ is only applied to literals[def. 21.4] and the only other operators are $\wedge$ and $\vee$.

**Definition 21.6 Conjunctive Normal Form (CNF):** An boolean algebraic expression $F$ is in CNF if it is a *conjunction* of *clauses*, where each clause is a disjunction of *literals*[def. 21.4] $L_{i,j}$:

$$F_{\text{CNF}} = \bigwedge_{i=1}^{n} \left( \bigvee_{j=1}^{m_i} L_{i,j} \right) \tag{21.13}$$

**Definition 21.7 Disjunctive Normal Form (DNF):** An boolean algebraic expression $F$ is in DNF if it is a *disjunction* of *clauses*, where each clause is a conjunction of *literals*[def. 21.4] $L_{i,j}$:

$$F_{\text{DNF}} = \bigvee_{i=1}^{n} \left( \bigwedge_{j=1}^{m_i} L_{i,j} \right) \tag{21.14}$$

**Note**

- true is a CNF with no clause and a single literal.
- false is a CNF with a single clause and no literals

#### 1.3.1. Transformation to CNF and DNF

**DNF**

**Algorithm 21.1:**

① Using *De Morgan's laws*Property 21.9 and double negationProperty 21.4 transform $F$ into *Negation Normal Form*[def. 21.5]:

| | | |
|---|---|---|
| $\neg\neg x$ | by | $x$ |
| $\neg(x \wedge y)$ | by | $(\neg x \vee \neg y)$ |
| $\neg(x \vee y)$ | by | $(\neg x \wedge \neg y)$ |
| $\neg$true | by | false |
| $\neg$false | by | true |

② Using distributive lawsProperty 21.8 substitute all:

| | | |
|---|---|---|
| $x \wedge (y \vee z)$ | by | $(x \wedge y) \vee (x \wedge z)$ |
| $(y \vee z) \wedge x$ | by | $(y \wedge x) \vee (z \wedge x)$ |
| $x \wedge$ true | by | true |
| true $\wedge\ x$ | by | true |

③ Using the identityProperty 21.2 and zero laws Property 21.3 remove true from any cause and delete all clauses containing false.

**Note**

For the CNF form simply use duality for step 2 and 3 i.e. swap $\wedge$ and $\vee$ and true and false.

**Using Truth Tables** [example 21.2]

To obtain a DNF formula from a truth table we need to have a *conjunctive*[def. 21.3] for each row where $F$ is true.

## 2. Examples

**Example 21.1 Literals:**
Boolean literals: $x, \neg y, s$
Not boolean literals: $\neg\neg x$, $(x \wedge y)$

**Example 21.2 DNF from truth tables:**

Need a conjunction of:

- $(\neg x \wedge \neg y \wedge \neg z)$
- $(\neg x \wedge y \wedge z)$
- $(x \wedge \neg y \wedge \neg z)$
- $(x \wedge y \wedge z)$

| x | y | z | F |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 |

$$(\neg x \wedge \neg y \wedge \neg z) \wedge (\neg x \wedge y \wedge z) \wedge (x \wedge \neg y \wedge \neg z) \wedge (x \wedge y \wedge z)$$

# Calculus and Analysis

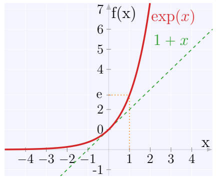
## 1. Functional Analysis

### 1.1. Elementary Functions

#### 1.1.1. Exponential Numbers

**Definition 22.1 Exponential Function** $\exp : \mathbb{K} \mapsto \mathbb{K}$:

$$\exp(x) = e^{x} = \sum_{n=0}^{\infty} \frac{x}{n!} = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n} \qquad (22.1)$$



**Definition 22.2 Exponential/Euler Number** e:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n} = 2.7182 \qquad (22.2)$$

**Properties Defining the Expeontial Function**

**Property 22.1:**

$$\exp(x+y) = \exp(x) + \exp(y) \qquad (22.3)$$

**Property 22.2:**

$$\exp(x) \leqslant 1 + x \qquad (22.4)$$

#### 1.1.2. Affine Linear Functions

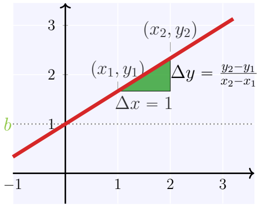
**Definition 22.3 Affine Linear Function** $f(x) = ax + b$**:** An affine linear function are functions that can be defined by a scaling $s_a(x) = ax$ plus a translation $t_b(x) = x + b$:

$$M = \{f : \mathbb{R} \mapsto \mathbb{R} | f(x) = (s_a \circ t_b)(x) = ax + b, \quad a, b \in \mathbb{R}\} \qquad (22.5)$$

$$f(x) = ax + b$$

$$f(0) = b$$

$$f'(x) = a$$



**Formula 22.1** [proof 22.1]
**Linear Function from Point and slope** $f(x_0) = y_1$**:**
Given a point $(x_1, y_1)$ and a slope $a$ we can derive:

$$f(x) = a \cdot (x - x_0) + y_0 = ax + (y_1 - ax_0) \qquad (22.6)$$

**Formula 22.2 Linear Function from two Points:**

$$f(x) = a \cdot (x - x_p) + y_p = ax + (y_p - ax_p) \qquad (22.7)$$

$$a = \frac{y_1 - y_0}{x_1 - x_0} \qquad p = \{ \text{ or } 2\}1$$

#### 1.1.3. Polynomials

**Definition 22.4 Polynomial:** A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:

$$\mathcal{P}_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1}x^{n-1} + a_n x^n \qquad (22.8)$$

**Corollary 22.1 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 22.5 Monomial:** Is a polynomial with only one term.

### 1.2. Functional Compositions

**Definition 22.6 Functional Compositions** $f \circ g$**:** Let $f : A \mapsto B$ and $g : D \mapsto C$ be to mappings s.t. $\operatorname{codom}(f) \subseteq D$ then we can define a composition function $(f \circ g) A \mapsto D$ as:

$$h(\mathbf{x}) = (g \circ f)(\mathbf{x}) = g(f(\mathbf{x})) \qquad \text{with} \qquad \mathbf{x} \in A \qquad (22.9)$$

**Corollary 22.2 Nested Functional Composition:**

$$F_{k:1}(\mathbf{x}) = (F_k \circ \cdots \circ F_1)(\mathbf{x}) = F_k\Big(F_{k-1} \circ \cdots \circ (F_1(\mathbf{x}))\Big) \qquad (22.10)$$

## 2. Proofs

**Proof 22.1** (formula 22.1).

$$f(x_0) = y_0 = ax_0 + b \qquad \Rightarrow \qquad b = y_0 - ax_0$$

**Definition 22.7 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 22.8 Discriminant:** $\delta = b^2 - 4ac$

**Definition 22.9 Solution to** [def. 22.7]**:**

$$x_{\pm} = \frac{-b \pm \sqrt{\delta}}{2a} \qquad \textbf{or} \qquad x_{\pm} = \frac{1}{2}\left(-p \pm \sqrt{p^2 - 4q}\right)$$

**Theorem 22.1**
**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:

$$F(X) = \int_a^x f(t)\,\mathrm{d}t \qquad (22.11)$$

Then it follows:

$$F'(x) = f(x) \qquad \forall x \in (a, b) \qquad (22.12)$$

**Theorem 22.2**
**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 $[a, b]$:

$$\int_a^b f(t)\,\mathrm{d}t = F(b) - F(a) \qquad \Longleftrightarrow \qquad \int_a^x \frac{\partial}{\partial x} F(t)\,\mathrm{d}t = F(x) \qquad (22.13)$$

**Definition 22.10 Domain of a function** $\operatorname{dom}(\cdot)$**:**
**Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the set of all possible input values $\mathcal{X}$ is called the domain of $f$ – $\operatorname{dom}(f)$.

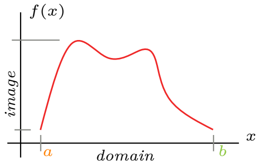
**Definition 22.11**
**Codomain/target set of a function** $\operatorname{codom}(\cdot)$**:**
**Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the codaomain of that function is the set $\mathcal{Y}$ into which all of the output of the function is **constrained** to fall.

**Definition 22.12 Image (Range) of a function:** $f[\cdot]$

**Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the image of that function is the set to which the function can actually map:

$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \qquad (22.14)$$

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



**Misnomer Range:** The term Range is ambiguous s.t. certain books refer to it as codomain and other as image.

**Definition 22.13 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 $f$:

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \qquad (22.15)$$

**Example 22.1 :**
**Given** $f : \mathbb{R} \to \mathbb{R}$
defined by $f : x \mapsto x^2 \iff f(x) = x^2$
$\operatorname{dom}(f) = \mathbb{R}$, $\operatorname{codom}(f) = \mathbb{R}$ **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:

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \qquad (22.16)$$

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

$$x \leqslant y \iff f(x) \leqslant f(y) \qquad \forall x, y \in \operatorname{dom}(f) \qquad (22.17)$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \qquad \forall x, y \in \operatorname{dom}(f) \qquad (22.18)$$

**Definition 22.15 (strictly) Decreasing Functions:**
A function $f$ is called monotonically decreasing/decreasing or non-increasing if:

$$x \geqslant y \iff f(x) \geqslant f(y) \qquad \forall x, y \in \operatorname{dom}(f) \qquad (22.19)$$

And *strictly* decreasing if:

$$x > y \iff f(x) > f(y) \qquad \forall x, y \in \operatorname{dom}(f) \qquad (22.20)$$

**Definition 22.16 Monotonic Function:** A function $f$ is called monotonic iff either $f$ is increasing or decreasing.

**Definition 22.17 Linear Function:**
A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

$$L(\mathbf{x} + \mathbf{y}) = L(\mathbf{x}) + L(\mathbf{y})$$

$$L(\alpha\mathbf{x}) = \alpha L(\mathbf{x}) \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

**Corollary 22.3 Linearity of Differentiation:** The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(af(x) + bg(x)\right) = a\frac{\mathrm{d}}{\mathrm{d}x}f(x) + b\frac{\mathrm{d}}{\mathrm{d}x}g(x) \qquad a, b \in \mathbb{R} \qquad (22.21)$$

**Definition 22.18 Quadratic Function:**
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\intercal}\mathbf{A}\mathbf{x} + \mathbf{b}^{\intercal}\mathbf{x} + c \qquad (22.22)$$

## 3. Smoothness

**Definition 22.19 Smoothness of a Function** $\mathcal{C}^k$**:**
**Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the function is said to be of class $k$ if it is differentiable up to order $k$ **and** continuous, on its entire domain:

$$f \in \mathcal{C}^k(\mathcal{X}) \iff \exists f', f'', \ldots, f^{(k)} \text{ continuous} \qquad (22.23)$$

**Note**

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

#### 3.0.1. Continuous Functions

**Definition 22.20 Continuous Function** $\mathcal{C}^0$**:** Functions that do not have any jumps or peaks.

#### 3.0.2. Piece wise Continuous Functions

**Definition 22.21 Piecewise Linear Functions** $\mathcal{C}^0_{\mathbf{pw}}$**:**

#### 3.0.3. Continously Differentiable Function

**Corollary 22.4 Continuously Differentiable Function** $\mathcal{C}^1$**:** Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f : \mathcal{X} \to \mathcal{Y}$ of the class must satisfy:

$$f \in \mathcal{C}^1(\mathcal{X}) \iff f' \text{ continuous} \qquad (22.24)$$

#### 3.0.4. Smooth Functions

**Corollary 22.5 Smooth Function** $\mathcal{C}^{\infty}$**:** Is a function $f : \mathcal{X} \to \mathcal{Y}$ that has derivatives infinitely many times differerntiable.

$$f \in \mathcal{C}^{\infty}(\mathcal{X}) \iff f', f'', \ldots, f^{(\infty)} \qquad (22.25)$$

### 3.1. Lipschitz Continuous Functions

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

**Definition 22.22 Lipschitz Continuity:** A Lipschitz continuous function is a function $f$ whose rate of change is bound by a Lipschitz Constant $L$:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leqslant L\|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}, \qquad L > 0 \tag{22.26}$$

**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 ⇒ tells us something about robustness.

#### 3.1.1. Lipschitz Continuous Gradient

**Definition 22.23 Lipschitz Continuous Gradient:** A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has $L$-*Lipschitz continuous gradient* if it satisfies:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leqslant L\|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \mathtt{dom}(f), \qquad L > 0 \tag{22.27}$$

if $f \in \mathcal{C}^2$, this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \preccurlyeq L\mathbf{I} \qquad \forall \mathbf{x} \in \mathtt{dom}(f), \qquad L > 0 \tag{22.28}$$

**Lemma 22.1 Descent Lemma** [Poorfs 22.5,??]: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (22.27) over its domain, then it holds that:

$$\left|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^{\mathsf{T}}(\mathbf{x} - \mathbf{y})\right| \leqslant \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{22.29}$$

**Note**

If $f$ is twice differentiable then the largest eigenvalue of the Hessian (Definition 23.8) of $f$ is uniformly upper bounded by $L$

### 3.2. L-Smooth Functions

**Definition 22.24 $L$-Smoothness:** A $L$-smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

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

with $\qquad \forall \mathbf{x}, \mathbf{y} \in \mathtt{dom}(f), \qquad L > 0 \quad (22.30)$

If $f$ is a twice differentiable this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \preccurlyeq L\mathbf{I} \qquad L > 0 \tag{22.31}$$

**Theorem 22.3** [proof 22.6]
**L-Smoothness of convex functions:**
A *convex* and L-Smooth function ([def. 22.24]) has a *Lipschitz continuous gradient*eq. (22.27) thus it holds that:

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})^{\mathsf{T}}(\mathbf{x} - \mathbf{y}) \leqslant \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{22.32}$$

**Note**

$L$-smoothnes is a weaker condition than $L$-Lipschitz continuous gradients

## 4. Convexity

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

**Definition 22.25 Convex Functions:**
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:

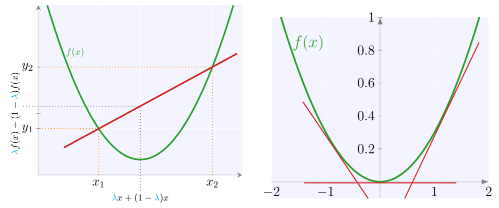
$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) \tag{22.33}$$
$$\forall \lambda \in [0,1] \qquad \forall x, y \in \mathtt{dom}(f)$$

If $f$ is a differentiable function this is equivalent to:

$$f(x) \geqslant f(y) + \nabla f(y)^{\mathsf{T}}(x - y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{22.34}$$

If $f$ is a twice differentiable function this is equivalent to:

$$\nabla^2 f(x) \succcurlyeq 0 \qquad \forall x, y \in \mathtt{dom}(f) \tag{22.35}$$



(a) (b)

**Definition 22.26 Concave Functions:**
A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if it satisfies:

$$f(\lambda x + (1-\lambda)y) \geqslant \lambda f(x) + (1-\lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f)\\ \forall \lambda \in [0,1]\end{array} \tag{22.36}$$

**Corollary 22.6 Convexity → global minimima:** Convexity implies that all local minima (if they exist) are global minima.

**Definition 22.27 Stricly Convex Functions:**
A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \qquad \begin{array}{l}\forall x, y \in \mathtt{dom}(f)\\ \forall \lambda \in [0,1]\end{array}$$

If $f$ is a differentiable function this is equivalent to:

$$f(x) > f(y) + \nabla f(y)^{\mathsf{T}}(x - y) \qquad \forall x, y \in \mathtt{dom}(f) \tag{22.37}$$

If $f$ is a twice differentiable function this is equivalent to:

$$\nabla^2 f(x) > 0 \qquad \forall x, y \in \mathtt{dom}(f) \tag{22.38}$$

**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. (22.37) implies that $f(x)$ is above the tangent $f(x) + \nabla f(x)^{\mathsf{T}}(y - x)$ for all $x, y \in \mathtt{dom}(f)$
- ?? implies that $f(x)$ is flat or curved upwards

**Corollary 22.7 Strict Convexity → Uniqueness:**
Strict convexity implies a unique minimizer ⟺ at most one global minimum.

**Corollary 22.8 :** A twice differentiable function of one variable $f : \mathbb{R} \to \mathbb{R}$ is convex on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval $\mathcal{X}$:

$$f''(x) \geqslant 0 \qquad \forall x \in \mathcal{X} \tag{22.39}$$

**Definition 22.28 $\mu$-Strong Convexity:**
Let $\mathcal{X}$ be a Banach space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. A function $f : \mathcal{X} \to \mathbb{R}$ is called strongly convex iff the following equation holds:

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

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

$$f(y) \geqslant f(x) + \nabla f(x)^{\mathsf{T}}(y - x) + \frac{\mu}{2}\|y - x\|_2^2 \tag{22.40}$$

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

$$\nabla^2 f(x) \succcurlyeq \mu\mathbf{I} \qquad \forall x, y \in \mathcal{X} \qquad \mu > 0 \tag{22.41}$$

**Corollary 22.9 Strong Convexity implies Strict Convexity:**

https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex

**Property 22.3:**

$$f(\mathbf{y}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})^{\mathsf{T}}(\mathbf{x} - \mathbf{y}) + \frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \tag{22.42}$$

**Intuition**

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

**Size of $\mu$**

The parameter $\mu$ specifies how strongly the bounding quadratic function/approximation is.

**Proof 22.2.** *eq.* (22.41) *analogously to* **Proof** *eq.* (22.31)

**Note**

If $f$ is twice differentiable then the smallest eigenvalue of the Hessian ([def. 23.8]) of $f$ is uniformly lower bounded by $\mu$
**Hence** strong convexity can be considered as the analogous to smoothness

**Example 22.2 Quadratic Function:** A quadratic function eq. (22.22) is convex if:

$$\nabla^2_{\mathbf{x}}\text{eq. (22.22)} = \mathbf{A} \succcurlyeq 0 \tag{22.43}$$

**Corollary 22.10 :**
Strong convexity ⇒ Strict convexity ⇒ Convexity

### 4.1. Properties that preserve convexity

**Property 22.4 Non-negative weighted Sums:** Let $f$ be a convex function then $g(x)$ is convex as well:

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

**Property 22.5 Composition of Affine Mappings:** Let $f$ be a convex function then $g(x)$ is convex as well:

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

**Property 22.6 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.
⇒**Geometrically**: its graph remains unchanged after reflection about the y-axis.

$$f(-x) = f(x) \tag{22.44}$$

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

$$f(-x) = -f(x) \tag{22.45}$$

**Theorem 22.4 Rules:**
**Let** $f$ be even and $f$ odd respectively.

$g =: f \cdot f$ is even $\qquad g =: f \cdot f$ is even
$g =: f \cdot f$ is odd $\qquad$ the same holds for division

**Examples**

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

$$\begin{aligned} &x\text{-Shift:} && f(x - c) \Rightarrow \text{shift to the right} \\ & && f(x + c) \Rightarrow \text{shift to the left} && (22.46) \\ &y\text{-Shift:} && f(x) \pm c \Rightarrow \text{shift up/down} && (22.47) \end{aligned}$$

**Proof 22.3. eq. (22.46)** $f(x_n - c)$ *we take the $x$-value at $x_n$ but take the $y$-value at $x_o := x_n - c$*
⇒ *we shift the function to $x_n$.*

**Euler's formula**

$$e^{\pm ix} = \cos x \pm i \sin x \tag{22.48}$$

**Euler's Identity**

$$e^{\pm i} = -1 \tag{22.49}$$

**Note**

$$e^{n} = 1 \Leftrightarrow n = i\,2\pi k, \qquad k \in \mathbb{N} \tag{22.50}$$

**Corollary 22.11 Every norm is a convex function:** By using definition [def. 22.25] and the triangular inequality it follows (with the exception of the L0-norm):

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

### 4.2. Taylor Expansion

**Definition 22.29 Taylor Expansion:**

$$T_n(x) = \sum_{i=0}^{n} \frac{1}{n!} f^{(i)}(x_0) \cdot (x - x_0)^{(i)} \tag{22.51}$$
$$= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \mathcal{O}(x^3) \tag{22.52}$$

**Definition 22.30 Incremental Taylor:**
**Goal**: evaluate $T_n(x)$ (eq. (22.52)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^{n} \frac{h^i}{n!} f^{(i)}(x_0) i^{-1} \tag{22.53}$$
$$= f(x_0) \pm hf'(x_0) + \frac{h^2}{2}f''(x_0) \pm f'''(x_0)(h)^3 + \mathcal{O}(h^4)$$

**Note**

If we chose $\Delta x$ small enough it is sufficient to look only at the first two terms.

**Definition 22.31 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

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathsf{T}} H(\mathbf{x} - \mathbf{x}_0) \tag{22.54}$$

**Definition 22.32 Argmax:** The argmax of a function defined on a set $D$ is given by:

$$\arg\max_{x \in D} f(x) = \{x | f(x) \geqslant f(y), \forall y \in D\} \tag{22.55}$$

**Definition 22.33 Argmin:** The argmin of a function defined on a set $D$ is given by:

$$\arg\min_{x \in D} f(x) = \{x | f(x) \leqslant f(y), \forall y \in D\} \tag{22.56}$$

**Corollary 22.12 Relationship** $\arg\min \leftrightarrow \arg\max$**:**

$$\arg\min_{x \in D} f(x) = \arg\max_{x \in D} -f(x) \tag{22.57}$$

**Property 22.7 Argmax Identities:**

1. **Shifting**:
$$\forall \lambda \text{ const} \qquad \arg\max f(x) = \arg\max f(x) + \lambda \tag{22.58}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \qquad \arg\max f(x) = \arg\max \lambda f(x) \tag{22.59}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \qquad \arg\max f(x) = \arg\min \lambda f(x) \tag{22.60}$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f)$$
$$\arg\max f(x) = \arg\min \frac{1}{f(x)} \tag{22.61}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions[def. 22.14] $g$ it holds that:
$$\arg\max g(f(x)) = \arg\max f(x) \tag{22.62}$$

**Definition 22.34 Max:** The maximum of a function $f$ defined on the set $D$ is given by:

$$\max_{x \in D} f(x) = f(x^*) \qquad \text{with} \qquad \forall x^* \in \arg\max_{x \in D} f(x) \tag{22.63}$$

**Definition 22.35 Min:** The minimum of a function $f$ defined on the set $D$ is given by:

$$\min_{x\in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\min_{x\in D} f(x) \qquad (22.64)$$

**Corollary 22.13 Relationship** min ↔ max**:**

$$\min_{x\in D} f(x) = -\max_{x\in D} -f(x) \qquad (22.65)$$

**Property 22.8 Max Identities:**

1. **Shifting**:
$$\forall \lambda \text{ const} \qquad \max\{f(x)+\lambda\} = \lambda + \max f(x) \qquad (22.66)$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \qquad \max \lambda f(x) = \lambda \max f(x) \qquad (22.67)$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \qquad \max \lambda f(x) = \lambda \min f(x) \qquad (22.68)$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \texttt{dom}(f) \qquad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \qquad (22.69)$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions[def. 22.14] $g$ it holds that:
$$\max g(f(x)) = g(\max f(x)) \qquad (22.70)$$

**Definition 22.36 Supremum:** The supremum of a function defined on a set $D$ is given by:

$$\sup_{x\in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y|y\geqslant f(x),\forall x\in D} y \qquad (22.71)$$

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

**Definition 22.37 Infinmum:** The infinmum of a function defined on a set $D$ is given by:

$$\inf_{x\in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y|y\leqslant f(x),\forall x\in D} y \qquad (22.72)$$

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

**Corollary 22.14 Relationship** sup ↔ inf**:**

$$\in_{x\in D} f(x) = -\sup_{x\in D} -f(x) \qquad (22.73)$$

**Note**

The supremum/infinmum 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 $\iff$ infinmum/supremum.

**Definition 22.38 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.

$$y(t) = f(x(t), t) \xrightarrow[\forall \tau]{\text{time-invariance}} y(t-\tau) = f(x(t-\tau), t) \qquad (22.74)$$

**Definition 22.39 Inverse Function $g = f^{-1}$:**
A function $g$ is the inverse function of the function $f : A \subset \mathbb{R} \to B \subset \mathbb{R}$ if

$$f(g(x)) = x \qquad \forall x \in \texttt{dom}(g) \qquad (22.75)$$

and

$$g(f(u)) = u \qquad \forall u \in \texttt{dom}(f) \qquad (22.76)$$

**Property 22.9**
**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 22.5 The Existence of an Inverse Function:**
A function has an inverse function if and only if it is one-to-one.

**Corollary 22.15 Inverse functions and strict monotonicity:** If a function $f$ is strictly monotonic [def. 22.16] on its entire domain, then it is one-to-one and therefore has an inverse function.

## 5. Special Functions

### 5.1. The Gamma Function

**Definition 22.40 The gamma function $\Gamma(\alpha)$:** Is extension of the factorial function (**??**) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\,\mathrm{d}x \qquad \Re(z) > 0 \qquad (22.77)$$

$$\Gamma(n) \quad \overset{n\in\mathbb{N}}{\iff} \quad \Gamma(n) = (n-1)!$$

## 6. Proofs

**Proof 22.4.** *lemma 22.1 for $\mathcal{C}^1$ functions:*
*Let $g(t) \equiv f(\mathbf{y} + t(\mathbf{x}-\mathbf{y}))$ from the FToC (theorem 22.2) we know that:*

$$\int_0^1 g'(t)\,\mathrm{d}t = g(1) - g(0) = f(\mathbf{x}) - f(\mathbf{y})$$

*It then follows from the reverse:*

$$\begin{aligned}
&|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})| \\
\overset{\substack{Chain.\ R\\ FToC}}{=} &\left|\int_0^1 \nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t - \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})\right| \\
= &\left|\int_0^1 (\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t\right| \\
= &\left|\int_0^1 (\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t\right| \\
\overset{C.S.}{\leqslant} &\left|\int_0^1 \|\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y})\| \cdot \|\mathbf{x}-\mathbf{y}\|\,\mathrm{d}t\right| \\
\overset{eq.\ (22.27)}{=} &\left|\int_0^1 L\|\mathbf{y}+t(\mathbf{x}-\mathbf{y}) - \mathbf{y}\| \cdot \|\mathbf{x}-\mathbf{y}\|\,\mathrm{d}t\right| \\
= &\left|L\|\mathbf{x}-\mathbf{y}\|^2 \int_0^1 t\,\mathrm{d}t\right| = \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2
\end{aligned}$$

**Proof 22.5.** **??** *for $\mathcal{C}^2$ functions:*

$$f(\mathbf{y}) \overset{Taylor}{=} f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal\nabla^2 f(z)(\mathbf{y}-\mathbf{x})$$

*Now we plug in $\nabla^2 f(\mathbf{x})$ and recover eq. (22.30):*

$$f(\mathbf{y}) \leqslant f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal L(\mathbf{y}-\mathbf{x})$$

**Proof 22.6.** *theorem 22.3:*
*With the definition of convexity for a differentiable function (eq. (22.37)) it follows*

$$\begin{aligned}
&f(x) - f(y) + \nabla f(y)^\intercal(x-y) \geqslant 0 \\
\Rightarrow &|f(x) - f(y) + \nabla f(y)^\intercal(x-y)| \\
\overset{if\ eq.\ (22.37)}{=} &f(x) - f(y) + \nabla f(y)^\intercal(x-y)
\end{aligned}$$

*with lemma 22.1 and [def. 22.24] it follows theorem 22.3*

# Differential Calculus

## 1. The Chain Rule

**Formula 23.1 Generalized Chain Rule:**
Let $\mathbf{F}:\mathbb{R}^n\mapsto\mathbb{R}^k$ and $\mathbf{G}:\mathbb{R}^k\mapsto\mathbb{R}^m$ be to general maps then it holds:

$$\underbrace{\partial\left(\mathbf{G}\circ\mathbf{F}\right)}_{\mathbb{R}^n\mapsto\mathbb{R}^{m\times n}} = \underbrace{\left(\partial\mathbf{G}\circ\mathbf{F}\right)\cdot\partial\mathbf{F}}_{\mathbb{R}^n\mapsto\left(\mathbb{R}^{m\times k}\cdot\mathbb{R}^{k\times n}\right)} \qquad \begin{aligned}\partial F &: \mathbb{R}^n\mapsto\mathbb{R}^{k\times n}\\ \partial G &: \mathbb{R}^k\mapsto\mathbb{R}^{m\times k}\end{aligned} \tag{23.1}$$

## 2. Directional Derivative

## 3. Partial Differentiation

**Definition 23.1 Paritial Derivative:**
Let $f:\mathbb{R}^n\mapsto\mathbb{R}$ be a real valued function, its partial derivative $\partial_i f:\mathbb{R}^n\mapsto\mathbb{R}$ is defined as the directional derivative**??** along the coordinate axis of one of its variables:

$$\partial_i f(\mathbf{x}) = \frac{\partial f}{\partial x_i} = D_{x_i}f = \lim_{h\to 0}\frac{f(\mathbf{x}, x_i\leftarrow x_i+h)-f(\mathbf{x})}{h}$$
$$= \lim_{h\to 0}\frac{f(x_1,\dots,x_i+h,\dots,x_n)-f(x_1,\dots,x_i,\dots,x_n)}{h} \tag{23.2}$$

### 3.1. The Gradient

### 3.1.1. The Nabla Operator

**Definition 23.2 Nabla Operator/Del** $\nabla$:
Given a cartesian coordinate system $\mathbb{R}^n$ with coordinates $x_1,\dots,x_n$ and associated unit vectors $\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_n$ its *del* operator is defined as:

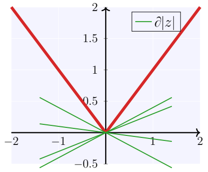
$$\nabla = \sum_{i=1}^{n}\frac{\partial}{\partial x_i}\tilde{\mathbf{e}}_i = \begin{bmatrix}\frac{\partial}{\partial x_1}(\mathbf{x})\\ \frac{\partial}{\partial x_2}(\mathbf{x})\\ \vdots\\ \frac{\partial}{\partial x_n}(\mathbf{x})\end{bmatrix} \tag{23.3}$$

**Definition 23.3 Gradient:**
Given a *scalar valued* function $f:\mathbb{R}^n\mapsto\mathbb{R}$ its gradient $\nabla f:\mathbb{R}^n\mapsto\mathbb{R}^n$ is defined as vector $\mathbb{R}^n$ of the partial derivatives[def. 23.1] w.r.t. all coordinate axes:

$$\operatorname{grad} f(\mathbf{x}) := \nabla f(\mathbf{x}) = \begin{bmatrix}\frac{\partial f}{\partial x_1}(\mathbf{x})\\ \frac{\partial f}{\partial x_2}(\mathbf{x})\\ \vdots\\ \frac{\partial f}{\partial x_n}(\mathbf{x})\end{bmatrix} = \left(\frac{\partial f}{\partial\mathbf{x}}\right)^{\mathsf{T}} \tag{23.4}$$

### 3.1.2. The Subderivative

**Definition 23.4** [proof 23.1]
**Subgradient** $g$:



Let $f:R^n\mapsto\mathbb{R}$ be a continuous (not necessarily differentiable) function. $g\in\mathbb{R}^n$ is a subgradient of $f$ at a point $\mathbf{x}_0\in\mathbb{R}^n$ if it satisfies:

$$g: f(\mathbf{x})-f(\mathbf{x}_0)\geqslant\mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0) \tag{23.5}$$

**Definition 23.5** [example 23.1]
**Subderivative** $\partial f(\mathbf{x}_0)$:
Let $f:R^n\mapsto\mathbb{R}$ be a continuous (not necessarily differentiable) function. The subdifferential of $f$ at a point $\mathbf{x}_0\in\mathbb{R}^n$ is defined as the set of all possible subgradients[def. 23.4] $g$:

$$\partial f(\mathbf{x}_0)\left\{g : f(\mathbf{x})-f(\mathbf{x}_0)\geqslant\mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0)\quad\forall\mathbf{x}\in\mathbb{R}^n\right\} \tag{23.6}$$

**Heuristic**

We can guess the sub derivative at a point by looking at all the slopes that are smaller then the graph.

### 3.2. The Jacobian

**Definition 23.6**
**Jacobian/Jacobi Matrix** $\mathbf{Df},\mathbf{J_f}$:
Given a *vector valued* function

$$\mathbf{f}:\mathbb{R}^n\mapsto\mathbb{R}^m \qquad \text{its derivative} \qquad \mathbf{J_f}:\mathbb{R}^n\mapsto\mathbb{R}^{m\times n}$$

with components $\partial_{ij}\mathbf{f}=\partial_i f_j:\mathbb{R}^n\mapsto\mathbb{R}$ is a vector valued function defined as:

$$\mathbf{J}(\mathbf{f}(\mathbf{x})) = \mathbf{J_f}(\mathbf{x}) = \mathbf{Df} = \frac{\partial\mathbf{f}}{\partial\mathbf{x}}(\mathbf{x}) = \frac{\partial(f_1,\dots,f_m)}{\partial(x_1,\dots,x_n)}(\mathbf{x}) \tag{23.7}$$
$$= \begin{bmatrix}\frac{\partial\mathbf{f}}{\partial x_1}\cdots\cdots\frac{\partial\mathbf{f}}{\partial x_n}\end{bmatrix} = \begin{bmatrix}\nabla^{\mathsf{T}}\mathbf{f}_1\\ \vdots\\ \nabla^{\mathsf{T}}f_m\end{bmatrix}$$
$$= \begin{bmatrix}\frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x})\cdots\cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x})\\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \ddots & \frac{\partial f_2}{\partial x_n}(\mathbf{x})\\ \vdots & & \vdots\\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \frac{\partial f_m}{\partial x_n}(\mathbf{x})\end{bmatrix}$$

**Explanation 23.1.** *Rows of the Jaccobian are transposed gradients*[def. 23.3] *of the component functions $f_1,\dots,f_m$.*

**Corollary 23.1 :**

## 4. Second Order Derivatives

**Definition 23.7 Second Order Derivative** $\frac{\partial^2}{\partial x_i\partial x_j}$:

**Theorem 23.1**
**Symmetry of second derivatives/Schwartz's Theorem:**
Given a continuous and twice differentiable function $f:\mathbb{R}^n\mapsto\mathbb{R}$ then its second order partial derivatives commute:

$$\frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}$$

### 4.1. The Hessian

**Definition 23.8 Hessian Matrix:**
Given a function $f:\mathbb{R}\mapsto\mathbb{R}^n$ its Hessian$\in\mathbb{R}^{n\times n}$ is defined as:

$$\mathbf{H}(\mathbf{f})(\mathbf{x}) = \mathbf{H}_f(\mathbf{x}) = \mathbf{J}(\nabla\mathbf{f}(\mathbf{x}))^T \tag{23.8}$$
$$= \begin{bmatrix}\frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1\partial x_2}(\mathbf{x})\cdots\cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}(\mathbf{x})\\ \frac{\partial^2 f}{\partial x_2\partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2\partial x_n}(\mathbf{x})\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_n\partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n\partial x_2}(\mathbf{x})\cdots\cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x})\end{bmatrix}$$

and it corresponds to the Jacobian of the Gradient.
Due to the differentiability and theorem 23.1 it follows that the Hessian is (if it exists):

- Symmetric
- Real

**Corollary 23.2 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 $\{(\lambda_1,\mathbf{v}_1),\dots,\lambda_n,\mathbf{v}_n\}$.
Not let $\mathbf{d}$ be a directional unit vector then the second derivative in that direction is given by:

$$\mathbf{d}^{\mathsf{T}}\mathbf{H}\mathbf{d} \iff \mathbf{d}^{\mathsf{T}}\sum_{i=1}^{n}\lambda_i\mathbf{v}_i \overset{\text{if }\mathbf{d}=\mathbf{v}_j}{\iff} \mathbf{d}^{\mathsf{T}}\lambda_j\mathbf{v}_j$$

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

## 5. Extrema

**Definition 23.9 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:

$$f'(\mathbf{x}_0)=0 \qquad\iff\qquad \nabla_{\mathbf{x}}f(\mathbf{x}_0)=0$$

**Corollary 23.3 Second Derivative Test** $f:\mathbb{R}\mapsto\mathbb{R}$**:**
Suppose $f:\mathbb{R}\mapsto\mathbb{R}$ is twice differentiable at a stationary point $x$ [def. 23.9] then it follows that:

- $f''(x)>0 \iff \begin{array}{ll} f'(x+\epsilon)>0 & \text{slope points uphill}\\ f'(x-\epsilon)<0 & \text{slope points downhill}\\ f(x)\text{ is a local minimum} & \end{array}$
- $f''(x)<0 \iff \begin{array}{ll} f'(x+\epsilon)>0 & \text{slope points downhill}\\ f'(x-\epsilon)<0 & \text{slope points uphill}\\ f(x)\text{ is a local maximum} & \end{array}$

$\epsilon>0$ sufficiently small enough

**Corollary 23.4 Second Derivative Test** $f:\mathbb{R}^n\mapsto\mathbb{R}$**:**
Suppose $f:\mathbb{R}^n\mapsto\mathbb{R}$ is twice differentiable at a stationary point $\mathbf{x}$ [def. 23.9] then it follows that:

- If $\mathbf{H}$ is p.d $\iff \forall\lambda_i>0\in\mathbf{H} \quad\to\quad f(\mathbf{x})$ is a local min.
- If $\mathbf{H}$ is n.d $\iff \forall\lambda_i<0\in\mathbf{H} \quad\to\quad f(\mathbf{x})$ is a local max.
- If $\exists\lambda_i>0\in\mathbf{H}$ and $\exists\lambda_i<0\in\mathbf{H}$ then $\mathbf{x}$ is a local maximum in one cross section of $f$ but a local minimum in another
- If $\exists\lambda_i=0\in\mathbf{H}$ 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 $\mathbf{H}$ is positive definite for a minima $\mathbf{x}^*$ of a *quadratic* function $f$ then this point must be a global minimum of that function.

## 6. Proofs

**Proof 23.1.** *Definition 23.4* $f(\mathbf{x})\geqslant f(\mathbf{x}_0)+\mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0)\quad\forall\mathbf{x}\in\mathbb{R}^n$ *corresponds to a line (see formula 22.1) at the point* $\mathbf{x}_0$ *with slope* $\mathbf{g}^{\mathsf{T}}$.
*Thus we search for all lines with smaller slope then function graph.*

## 7. Examples

**Example 23.1 Subderivatives Absolute Value Function** $|x|$**:** $f:\mathbb{R}\mapsto\mathbb{R}$ with $f(x)=|x|$ at the point $x=0$ it holds:

$$f(x)-f(0)\geqslant gx \qquad\Longrightarrow\qquad \text{the interval }[-1;1]$$

For $x\neq 0$ the subgradient is equal to the gradient. Thus it follows for the subderivatives/differentials:

$$\partial|x| = \begin{cases}-1 & \text{if } x<0\\ [-1,1] & \text{if } x=0\\ 1 & \text{if } x>0\end{cases}$$

**Theorem 24.1 Important Integral Properties:**

**Addition**
$$\int\limits_a^b f(x)\,\mathrm{d}x = \int\limits_a^c f(x)\,\mathrm{d}x + \int\limits_c^b f(x)\,\mathrm{d}x \tag{24.1}$$

**Reflection**
$$\int\limits_a^b f(x)\,\mathrm{d}x = -\int\limits_b^a f(x)\,\mathrm{d}x \tag{24.2}$$

**Translation**
$$\int\limits_a^b f(x)\,\mathrm{d}x \overset{u:=x\pm c}{=} \int\limits_{a\pm c}^{b\pm c} f(x \mp c)\,\mathrm{d}x \tag{24.3}$$

$f$ **Odd**
$$\int\limits_{-a}^{a} f(x)\,\mathrm{d}x = 0 \tag{24.4}$$

$f$ **Even**
$$\int\limits_{-a}^{a} f(x)\,\mathrm{d}x = 2\int\limits_0^a f(x)\,\mathrm{d}x \tag{24.5}$$

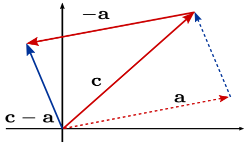
**Proof 24.1. eqs. (24.4) and (24.5)**

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

# Linear Algebra

## 1. Vectors

**Definition 25.1 Vector Substraction:**



$$\mathbf{b} = \mathbf{c} - \mathbf{a} \tag{25.1}$$

## 2. Linear Systems of Equations

### 2.1. Gaussian Elimination

### 2.1.1. Rank

**Definition 25.2 Matrix Rank** $\mathfrak{rank}$:
The ranks of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as the the dimension[def. 25.10] of the vector space spaned[def. 25.6] by its row or column vectors:

$$\begin{aligned}\mathfrak{rank}(\mathbf{A}) &= \dim\left(\{\mathbf{a}_{:,1},\dots,\mathbf{a}_{:,n}\}\right)\\ &= \dim\left(\{\mathbf{a}_{1,:},\dots,\mathbf{a}_{m,:}\}\right)\\ &\overset{\text{def. } 25.47}{=} \dim(\mathfrak{R}(\mathbf{A}))\end{aligned} \tag{25.2}$$

**Corollary 25.1 :**

- The column-and row-ranks of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ are equal.
- The rank of a non-symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is limited by the smaller dimension:
$$\mathfrak{rank}(\mathbf{A}) \leqslant \min\{n, m\} \tag{25.3}$$

**Property 25.1 Rank of Matrix Product:** Let $\mathbf{A} \in \mathbb{R}^{m,n}$ and $\mathbf{B} \in \mathbb{R}^{n,p}$ then the rank of the matrix product is limited:

$$\mathfrak{rank}(\mathbf{AB}) \leqslant \min\{\mathfrak{rank}(\mathbf{A}), \mathfrak{rank}(\mathbf{B})\} \tag{25.4}$$

## 3. Vector Spaces

### 3.1. Vector Space

**Definition 25.3 Vector Space:** TODO

### 3.2. Vector Subspace

**Definition 25.4 Vector Subspaces:**
A non-empty subset $U$ of a $\mathbb{K}$-vector space $\mathcal{V}$ is called a subspace of $\mathcal{V}$ if it satisfies:

$$\mathbf{u},\mathbf{v} \in U \quad\Longrightarrow\quad \mathbf{u}+\mathbf{v} \in U \tag{25.5}$$

$$\mathbf{u} \in U \quad\Longrightarrow\quad \lambda\mathbf{u} \in U \qquad \forall\lambda \in \mathbb{K} \tag{25.6}$$

**Definition 25.5 Linearcombination:**
Let $X = \{\mathbf{v}_1,\dots,\mathbf{v}_n\} \subset \mathcal{V}$ be a non-empty and finite subset of vectors of an $\mathbb{K}$-vector space $\mathcal{V}$. A *linear combination* of $X$ is a combination of the vectors defined as:

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i = \alpha_1\mathbf{v}_1 + \ldots + \alpha_n\mathbf{v}_n \qquad \alpha_i \in \mathbb{K} \tag{25.7}$$

**Definition 25.6**
**Span/Linear Hull** $\langle X\rangle$:
Is the set of all possible linear combinations[def. 25.5] of finite set $X = \{\mathbf{v}_1,\dots,\mathbf{v}_n\} \subset \mathcal{V}$ of a $\mathbb{K}$ vector space $\mathcal{V}$:

$$\langle X\rangle = \operatorname{span}(X) = \left\{\mathbf{v}\middle|\, \sum_{i=1}^{n} \alpha_i\mathbf{v}_i, \forall\alpha_i \in \mathbb{K}\right\} \tag{25.8}$$

**Definition 25.7 Generating Set:** A *generating set* of vectors $X = \{\mathbf{v}_1,\dots,\mathbf{v}_m\} \in \mathcal{V}$ of a vector spaces $\mathcal{V}$ is a set of vectors that *span*[def. 25.6] $\mathcal{V}$:

$$\operatorname{span}(\mathbf{v}_1\dots,\mathbf{v}_m) = \mathcal{V} \tag{25.9}$$

**Explanation 25.1** (Definition 25.7).

*The generating set of vector space (or set of vectors)* $\mathcal{V} \overset{i.e.}{=} \mathbb{R}^n$ *is a subset* $X = \{\mathbf{v}_1,\dots,\mathbf{v}_m\} \subset \mathcal{V}$ *s.t. every element of* $\mathcal{V}$ *can be produced by* $\operatorname{span}(X)$.

**Definition 25.8 Linear Independence:** A set of vector $\{\mathbf{v}_1,\dots,\mathbf{v}_n\} \in \mathcal{V}$ is called linear independent if the satisfy:

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i\mathbf{v}_i = \mathbf{0} \quad\Longleftrightarrow\quad \alpha_1 = \ldots = \alpha_n = 0 \tag{25.10}$$

**Corollary 25.2 :** A set of vector $\{\mathbf{x}_1,\dots,\mathbf{x}_n\} \in \mathcal{V}$ is called linear independent, if for every subset $X = \mathbf{x}_1,\dots,\mathbf{x}_m \subseteq \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ it holds that:

$$\langle X\rangle \subsetneq \{\mathbf{x}_1,\dots,\mathbf{x}_n\} \tag{25.11}$$

### 3.3. Basis

**Definition 25.9 Basis** $\mathfrak{B}$:
A subset $\mathfrak{B} = \{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ of a $\mathbb{K}$-vector space $\mathcal{V}$ is called a basis of $\mathcal{V}$ if:

$$\langle\mathfrak{B}\rangle = \mathcal{V} \quad \text{and} \quad \mathfrak{B} \text{ is a linear independent generating set} \tag{25.12}$$

**Corollary 25.3 :** The unit vectors $\mathbf{e}_1,\dots,\mathbf{e}_n$ build a standard basis of the $\mathbb{R}^n$.

**Corollary 25.4 Basis Representation:**
Let $\mathfrak{B}$ be a basis of a $\mathbb{K}$-vector space $\mathcal{V}$, then it holds that every vector $\mathbf{v} \in \mathcal{V}$ can be represented as a linear combination[def. 25.5] of $\mathfrak{B}$ by a unique set of coefficients $\alpha_i$:

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i\mathbf{b}_i \qquad \begin{matrix}\alpha_1,\dots,\alpha_n \in \mathbb{K}\\ \mathbf{b}_1,\dots,\mathbf{b}_n \in \mathfrak{B}\end{matrix} \tag{25.13}$$

### 3.3.1. Dimensionality

**Definition 25.10 Dimension of a vector space** $\dim(\mathcal{V})$:
Let $\mathcal{V}$ be a vector space. The dimension of $\mathcal{V}$ is defined as the number of necessary basis vectors $\mathfrak{B} = \{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ in order to span $\mathcal{V}$:

$$\dim(\mathcal{V}) := |\mathfrak{B}| = n \in \mathbb{N}_0 \tag{25.14}$$

**Corollary 25.5 :** $n$-linearly independent vectors of a $\mathbb{K}$-vector space $\mathcal{V}$ with finite dimension $n$ constitute a basis.

**Note**

If $\mathcal{V}$ is infinite $\dim(\mathcal{V}) = \infty$.

### 3.4. Affine Subspaces

**Definition 25.11 Affine Subspaces:** Given a $\mathbb{K}$-vector space $\mathcal{V}$ of dimension $\dim(\mathcal{V}) \geqslant 2$ a sub vector space[def. 25.4] $U$ of $\mathcal{V}$ defined as:

$$\mathcal{W} := \mathbf{v} + U = \{\mathbf{v}+\mathbf{x} | \mathbf{x} \in U\} \qquad \mathbf{v} \in \mathcal{V} \tag{25.15}$$

**Corollary 25.6 Direction:** The sub vector spaces $U$ are called *directions* of $\mathcal{V}$ and it holds:

$$\dim(\mathcal{W}) := \dim(U) \tag{25.16}$$

### 3.4.1. Hyperplanes

**Definition 25.12 Hyperplane** $\mathcal{H}$:
A hyperplane is a $d-1$ dimensional subspace of an $d$-dimensional ambient space that can be specified by the hess normal form[def. 25.13]:

$$\mathcal{H} = \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \hat{\mathbf{n}}^\intercal\mathbf{x} - d = 0\right\} \tag{25.17}$$

**Corollary 25.7 Half spaces:** A hyperplane $\mathcal{H} \in \mathbb{R}^{d-1}$ separates its $d$-dimensional ambient space into two half spaces:

$$\mathcal{H}^+ = \left\{x \in \mathbb{R}^d \,\middle|\, \tilde{\mathbf{n}}^\intercal\mathbf{x} + b > 0\right\} \tag{25.18}$$

$$\mathcal{H}^- = \left\{x \in \mathbb{R}^d \,\middle|\, \tilde{\mathbf{n}}^\intercal\mathbf{x} + b < 0\right\} = \mathbb{R}^d - \mathcal{H}^+ \tag{25.19}$$

**Notes**

Hyperplanes in $\mathbb{R}^2$ are lines and hyperplanes in $R^3$ are lines.

**Hess Normal Form**

**Definition 25.13 Hess Normal Form:**
Is an equation to describe hyperplanes[def. 25.12] in $\mathbb{R}^d$:

$$\mathbf{r}^\intercal\tilde{\mathbf{n}} - d = 0 \tag{25.20}$$

where all points described by the vector $\mathbf{r} \in \mathbb{R}^d$, *that satisfy* this equations lie on the hyperplane.

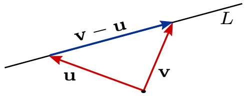
### 3.4.2. Lines

**Definition 25.14 Lines:** Lines are a set[def. 19.1] of the form:

$$L = \mathbf{u} + \mathbb{K}\,\mathbf{v} = \{\mathbf{u} + \lambda\mathbf{v} | \lambda \in \mathbb{K}\} \qquad \mathbf{u},\mathbf{v} \in \mathcal{V}, \mathbf{v} \neq 0 \tag{25.21}$$

**Two Point Formula**

**Definition 25.15 Two Point Formula:**



$$L = \mathbf{u} + \mathbb{K}\,\mathbf{v} \tag{25.22}$$

### 3.4.3. Planes

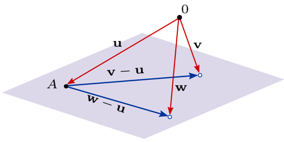
**Definition 25.16 Planes:** Planes are sets defined as:

$$E = \mathbf{u} + \mathbb{K}\,\mathbf{v} + \mathbb{K}\,\mathbf{w} = \{\mathbf{u} + \lambda\mathbf{v} + \mu\mathbf{w} | \lambda, \mu \in \mathbb{K}\} \tag{25.23}$$

$$\mathbf{u}, ,\mathbf{w} \in \mathcal{V} \qquad \text{s.t. } \mathbf{v},\mathbf{u} \neq \mathbf{0} \qquad \text{and} \qquad \mathbf{v},\mathbf{w} \text{ lin. indep.}$$

**Parameterform**

**Definition 25.17 Two Point Formula:**

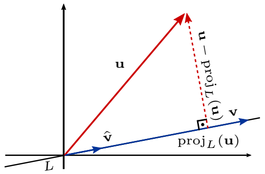


$$\begin{aligned}E = \mathbf{u} &+ \mathbb{K}(\mathbf{v}-\mathbf{u})\\ &+ \mathbb{K}(\mathbf{w}-\mathbf{u})\end{aligned} \tag{25.24}$$

### 3.4.4. Minimal Distance of Vector Subspaces

**Projections in 2D**

**Definition 25.18 2D Vector Projection [Proof 25.17,25.18]:**



$\mathbf{u}$
$\mathbf{u}-\text{proj}_L(\mathbf{u})$
$\mathbf{v}$
$\hat{\mathbf{v}}$
$\text{proj}_L(\mathbf{u})$
$L$

$$\begin{aligned}\mathbf{u}_\mathbf{v} &= \text{proj}_L(\mathbf{u})\\ &= u_v\tilde{\mathbf{v}} = \left(\mathbf{u}^\intercal\tilde{\mathbf{v}}\right)\tilde{\mathbf{v}}\\ &= \frac{\mathbf{u}^\intercal\mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} = \frac{\mathbf{u}^\intercal\mathbf{v}}{\mathbf{v}^\intercal\mathbf{v}}\mathbf{v}\end{aligned} \tag{25.25}$$

**Corollary 25.8** [proof 25.8]
**2D Projection Matrix** **P:** Is the matrix that satisfies:

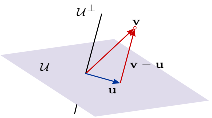
$$\mathbf{Pu} = \text{proj}_L(\mathbf{u}) \qquad \mathbf{P} = \frac{\mathbf{v}\mathbf{v}^\intercal}{\mathbf{v}^T\mathbf{v}} = \frac{\mathbf{v}\mathbf{v}^\intercal}{\|\mathbf{v}\|^2} \tag{25.26}$$

**Proof 25.1.** [Corollary 25.8]

$$\frac{1}{\mathbf{v}^\intercal\mathbf{v}}\mathbf{u}^\intercal\mathbf{v}\mathbf{v} = \frac{1}{\mathbf{v}^\intercal\mathbf{v}}\mathbf{v}\left(\mathbf{v}^\intercal\mathbf{u}\right) = \frac{1}{\mathbf{v}^\intercal\mathbf{v}}\left(\mathbf{v}\mathbf{v}^\intercal\right)\mathbf{u}$$

**General Projections**

**Definition 25.19** [proof 25.19]
**General Vector Projection:**
Is the orthogonal projection $\mathbf{u}$ of a vector $\mathbf{v}$ onto a sub-vector space $\mathcal{U}$



$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i\mathbf{b}_i \tag{25.27}$$

$$\mathbf{A}\mathbf{A}^\intercal\alpha_i = \mathbf{A}^\intercal\mathbf{v} \qquad \mathbf{A} = \begin{pmatrix}\mathbf{b}_1\\ \vdots\\ \vdots\\ \mathbf{b}_n\end{pmatrix}$$

where $\mathfrak{B} = \{\mathbf{b}_1,\dots,\mathbf{b}_n\}$ is a basis of the vector subspace $\mathcal{U}$.

**Theorem 25.1 Projection Theorem:** Let $\mathcal{U}$ a sub vector space of a finite euclidean vector space $\mathcal{V}$. Then there exists for every vector $\mathbf{v} \in \mathcal{V}$ a vector $\mathbf{u} \in \mathcal{U}$ obtained by an *orthogonal*[def. 25.63] projection

$$p : \begin{cases}\mathcal{V} \to \mathcal{U}\\ \mathbf{v} \mapsto \mathbf{u}\end{cases} \tag{25.28}$$

the vector $u' := \mathbf{v} - \mathbf{u}$ representing the distance between $\mathbf{u}$ and $\mathbf{v}$ and is minimal:

$$\|\mathbf{u}'\| = \|\mathbf{v}-\mathbf{u}\| \leqslant \|\mathbf{v}-\mathbf{w}\| \qquad \forall\mathbf{w} \in \mathcal{U} \qquad \mathbf{u}' \in \mathcal{U}^\perp \tag{25.29}$$

### 3.5. Affine Subspaces

### 3.6. Planes

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

# 4. Matrices

## Special Kind of Matrices

### 4.1. Symmetric Matrices

**Definition 25.20 Symmetric Matrices:** A matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is called *symmetric* if it satisfies:

$$\mathbf{A} = \mathbf{A}^\intercal \tag{25.30}$$

**Property 25.2** [proof ??]
**Eigenvalues of real symmetric Matrices:** The eigenvalues of a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are real:

$$\operatorname{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geqslant 0}\}_{i=1}^{n} \tag{25.31}$$

**Property 25.3** [proof ??]
**Orthogonal Eigenvector basis:** Eigenvectors of real symmetric matrices with distinct eigenvalues are orthogonal.

**Corollary 25.9**
**Eigendecomposition Symmetric Matrices:** If $\mathbf{A} \in \mathbb{R}^{n,n}$ is a real *symmetric*[def. 25.20] matrix then its eigenvectors are *orthogonal* and its eigen-decomposition[def. 25.82] is given by:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^\intercal \tag{25.32}$$

### 4.2. Orthogonal Matrices

**Definition 25.21 Orthogonal Matrix:** A real valued square matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is said to be orthogonal if its row vectors (and respectively its column vectors) build an orthonormal[def. 25.64] basis:

$$\langle \mathbf{q}_{:i}, \mathbf{q}_{:j} \rangle = \delta_{ij} \qquad \text{and} \qquad \langle \mathbf{q}_{i:}, \mathbf{q}_{j:} \rangle = \delta_{ij} \tag{25.33}$$

This is exactly true if the inverse of $\mathbf{Q}$ equals its transpose:

$$\mathbf{Q}^{-1} = \mathbf{Q}^\intercal \qquad \iff \qquad \mathbf{Q}\mathbf{Q}^\intercal = \mathbf{Q}^\intercal\mathbf{Q} = \mathbf{I}_n \tag{25.34}$$

**Attention:** *Orthogonal* matrices are sometimes also called *orthonormal matrices.*

### 4.3. Hermitian Matrices

**Definition 25.22 Conjugate Transpose Hermitian Conjugate/Adjoint Matrix:** $\mathbf{A}^{\mathsf{H}}/\mathbf{A}^*$

The conjugate transpose of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is defined as:

$$\mathbf{A}^{\mathsf{H}} := (\overline{\mathbf{A}}^\intercal) = \overline{\mathbf{A}^\intercal} \qquad \iff \qquad \mathbf{a}^{\mathsf{H}}_{i,j} = \bar{\mathbf{a}}_{j,i} \qquad \begin{matrix} 1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m \end{matrix} \tag{25.35}$$

**Definition 25.23**
**Hermitian/Self-Adjoint Matrices** $\mathbf{A} = \mathbf{A}^{\mathsf{H}}$:
A hermitian matrix is complex square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ who is equal to its own *conjugate transpose*[def. 25.22]:

$$\mathbf{A} = \mathbf{A}^{\mathsf{H}} = \overline{\mathbf{A}^\intercal} \qquad \iff \qquad \mathbf{a}_{i,j} = \bar{\mathbf{a}}_{j,i} \qquad i \in \{1, \dots, n\} \tag{25.36}$$

**Corollary 25.10 :** [def. 25.22] implies that $\mathbf{A}$ must be a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

**Corollary 25.11 Real Hermitian Matrices:** From [cor. 19.1] it follows:

$$\mathbf{A} \in \mathbb{R}^{n \times n} \;\; \textit{hermitian} \qquad \Longrightarrow \qquad \mathbf{A} \;\; \textit{real symmetric}^{[\text{def. } 25.20]} \tag{25.37}$$

**Property 25.4** [proof 25.15]
**Eigenvalues of Hermitan Matrices:** The eigenvalues of a hermitian matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ are real:

$$\operatorname{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geqslant 0}\}_{i=1}^{n} \tag{25.38}$$

**Property 25.5** [proof 25.16]
**Orthogonal Eigenvector basis:** Eigenvectors of hermitian matrices with distinct eigenvalues are orthogonal.

**Corollary 25.12**
**Eigendecomposition Symmetric Matrices:** If $\mathbf{A} \in \mathbb{C}^{n,n}$ is a hermitian matrix[def. 25.23] then its eigendecomposition[def. 25.82] is given by:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathsf{H}} \tag{25.39}$$

### 4.4. Unitary Matrices

**Definition 25.24 Unitary Matrix** $\mathbf{U}\mathbf{U}^{\mathsf{H}}$:
is a complex square matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ whose inverse[def. 25.38] is equal to its *conjugate transpose*[def. 25.22]:

$$\mathbf{U}^{\mathsf{H}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathsf{H}} = \mathbf{I} \tag{25.40}$$

**Corollary 25.13 Real Unitary Matrix:** A real matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ that is unitary is an *orthogonal matrix*[def. 25.21].

**Property 25.6**
**Preservation of Euclidean Norm** [proof 25.14]:
Orthogonal and unitary matrices $\mathbf{Q} \in \mathbb{K}^{n,n}$ do not affect the 2-norm:

$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \qquad \forall \mathbf{x} \in \mathbb{K}^n \tag{25.41}$$

### 4.5. Similar Matrices

**Definition 25.25 Similar Matrices:** Two square matrices $\mathbf{A} \in \mathbb{K}^{n \times n}$ and $\mathbf{B} \in \mathbb{K}^{n \times n}$ are called *similar* if there exists a invertible matrix $\mathbf{S} \in \mathbb{K}^{n \times n}$ s.t.:

$$\exists \mathbf{S}: \qquad \mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{25.42}$$

**Corollary 25.14**
**Similarity Transformation/Conjugation:**
The mapping:

$$\mathbf{A} \mapsto \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{25.43}$$

is called *similarity transformation*

**Corollary 25.15**
**Eigenvalues of Similar Matrices** [proof 25.13]:

If $\mathbf{A} \in \mathbb{K}^{n \times n}$ has the eigenvalue-eigenvector pairs $\{\{\lambda_i, \mathbf{v}_i\}\}_{i=1}^{n}$ then its *conjugate*eq. (25.43) $\mathbf{B}$ has the same eigenvalues with transformed eigenvectors:

$$\{\{\lambda_i, \mathbf{u}_i\}\}_{i=1}^{n} \qquad \mathbf{u}_i := \mathbf{S}^{-1}\mathbf{v}_i \tag{25.44}$$

### 4.6. Skew Symmetric Matrices

**Definition 25.26**
**Skey Symmetric/Antisymmetric Matrices:**

$$\mathbf{A}^\intercal = -\mathbf{A} \tag{25.45}$$

### 4.7. Triangular Matrix

**Definition 25.27 Triangular Matrix:** An upper (lower) triangular matrix, is a matrix whose element's below (above) the main diagonal are all zero:

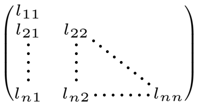


Figure 12: Lower Tri. Mat.

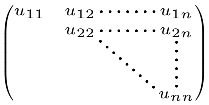


Figure 13: Upper Tri. Mat.

#### 4.7.1. Unitriangular Matrix

**Definition 25.28 Unitriangular Matrix:** An upper (lower) unitriangular matrix, is a upper (lower) triangular matrix[def. 25.27] whose diagonal elements are all ones.

#### 4.7.2. Strictly Triangular Matrix

**Definition 25.29 Strictly Triangular Matrix:** An upper (lower) strictly triangular matrix, is a upper (lower) triangular matrix[def. 25.27] whose diagonal elements are all zero.

### 4.8. Block Partitioned Matrices

**Definition 25.30 Block Partitioned Matrix:**
A matrix $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ can be partitioned into a *block partitioned matrix*:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \qquad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \tag{25.46}$$

**Definition 25.31 Block Partitioned Linear System:**
A linear system $\mathbf{M}\mathbf{x} = \mathbf{b}$ with $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^{k+l}$ can be partitioned into a *block partitioned system*:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \qquad \begin{matrix} \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \\ \mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^{k}, \mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^{l} \end{matrix} \tag{25.47}$$

#### 4.8.1. Schur Complement

**Definition 25.32 Schur Complement:** Given a block partitioned matrix[def. 25.30] $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ its Schur complements are given by:

$$\mathbf{S}_A = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \qquad \mathbf{S}_D = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \tag{25.48}$$

#### 4.8.2. Inverse of Block Partitioned Matrix

**Definition 25.33** proof 25.3
**Inverse of a Block Partitioned Matrix:**
Given a block partitioned matrix[def. 25.30] $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ its inverse $\mathbf{M}^{-1}$ can be partitioned as well:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \qquad \mathbf{M}^{-1} = \begin{bmatrix} \widetilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{bmatrix} \tag{25.49}$$

$$\widetilde{\mathbf{A}} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}_{\mathbf{A}}^{-1}\mathbf{C}\mathbf{A}^{-1} \qquad \tilde{\mathbf{C}} = -\mathbf{S}_{\mathbf{A}}^{-1}\mathbf{C}\mathbf{A}^{-1}$$

$$\tilde{\mathbf{B}} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}_{\mathbf{A}}^{-1} \qquad \widetilde{\mathbf{D}} = \mathbf{S}_{\mathbf{A}}^{-1}$$

where $\mathbf{S}_{\mathbf{A}} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is the Schur complement of $\mathbf{A}$.

### 4.9. Properties of Matrices

#### 4.9.1. Square Root of p.s.d. Matrices

**Definition 25.34 Square Root:**

#### 4.9.2. Trace

**Definition 25.35 Trace:** The trace of an $\mathbf{A} \in \mathbb{R}^{n \times n}$ matrix is defined as:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \dots + a_{nn} \tag{25.50}$$

**Property 25.7 Trace of a Scalar:**

$$\operatorname{tr}(\mathbb{R}) = \mathbb{R} \tag{25.51}$$

**Property 25.8 Trace of Transpose:**

$$\operatorname{tr}(\mathbf{A}^\intercal) = \operatorname{tr}(\mathbf{A}) \tag{25.52}$$

**Property 25.9 Trace of multiple Matrices:**

$$\operatorname{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \operatorname{tr}(\mathbf{B}\mathbf{C}\mathbf{A}) = \operatorname{tr}(\mathbf{C}\mathbf{B}\mathbf{A}) \tag{25.53}$$

# 5. Matrices and Determinants

## 5.1. Determinants

### 5.1.1. Laplace/Cofactor Expansion

**Definition 25.36 Minor:**

**Definition 25.37 Cofactors:**

**Properties**

**Property 25.10 Determinant times Scalar** $\det(\alpha\mathbf{A})$:
Given a matirx $\mathbf{A} \in \mathbb{R}^{n \times n}$ it holds:

$$\det(\alpha \cdot \mathbf{A}) = \alpha^n \mathbf{A} \tag{25.54}$$

## 5.2. Inverese of Matrices

**Definition 25.38 Inverse Matrix** $\mathbf{A}^{-1}$:

### 5.2.1. Invertability

**Definition 25.39**
**Singular/Non-Invertible Matrix** $\det(\mathbf{A}) = 0$:
A square matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is singular or non-invertible if it satisfies the following and equal conditions:

- $\det(\mathbf{A}) = 0$
- $\dim(\mathbf{A}) < n$
- $\nexists \mathbf{B} : \mathbf{B} = \mathbf{A}^{-1}$
- $\mathbf{A}\mathbf{x} = \mathbf{b}$ has either
  - no solution $\mathbf{x}$
  - infinitely many solutions $\mathbf{x}$

# Transformations And Mapping

## 6. Linear & Affine Mappings/Transformations

### 6.1. Linear Mapping

**Definition 25.40**
**Linear Mapping:** A linear mapping, function or transformation is a map $l : V \mapsto W$ between two $\mathbb{K}$-vector spaces[def. 25.3] $V$ and $W$ if it satisfies:

$$l(\mathbf{x}+\mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y}) \qquad \text{(Additivity)} \tag{25.55}$$
$$l(\alpha\mathbf{x}) = \alpha l(\mathbf{x}) \quad \forall \alpha \in \mathbb{K} \qquad \text{(Homogenitivity)} \tag{25.56}$$
$$\forall \mathbf{x}, \mathbf{y} \in V$$

**Proposition 25.1** [proof 25.8]
**Equivalent Formulations:** Definition 25.40 is equivalent to:

$$l(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha l(\mathbf{x}) + \beta l(\mathbf{y}) \qquad \begin{matrix}\forall \alpha, \beta \in \mathbb{K}\\ \forall \mathbf{x}, \mathbf{y} \in V\end{matrix} \tag{25.57}$$

**Corollary 25.16 Superposition Principle:**
Definition 25.40 is also known as the superposition principle: *"the net response caused by two or more signals is the sum of the responses that would have been caused by each signal individually."*

**Corollary 25.17** [proof 25.10]
**A linear mapping $\iff$ Ax:**
For every matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ the map:

$$l_{\mathbf{A}} : \begin{cases} \mathbb{K}^n & \rightarrow \quad \mathbb{K}^m \\ \mathbf{x} & \mapsto \quad \mathbf{A}\mathbf{x} \end{cases} \tag{25.58}$$

is a *linear map* and every linear map $l$ can be represented by a matrix vector product:

$$l \text{ is linear} \iff \exists \mathbf{A} \in \mathbb{K}^{n\times m} : f(x) = \mathbf{A}\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{K}^m \tag{25.59}$$

**Principle 25.1** [proof 25.9]
**Principle of linear continuation:** A linear mapping $l : \mathcal{V} \mapsto \mathcal{W}$ is determined by the image of the basis $\mathfrak{B}$ of $\mathcal{V}$:

$$l(\mathbf{v}) = \sum_{i=1}^{n} \beta_i l(b_i) \qquad \mathfrak{B}(\mathcal{V}) = \{b_1, \ldots, b_n\} \tag{25.60}$$

**Property 25.11** [proof 25.11]
**Compositions of linear mappings are linear $f \circ g$:** Let $g, f$ be linear functions mapping from $\mathcal{V}$ to $\mathcal{W}$ (i.e. matching) then it holds that $f \circ g$ is a linear[def. 25.40].

**Definition 25.41 Level Sets:**

### 6.2. Affine Mapping

**Definition 25.42 Affine Transformation/Map:**
Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ then:

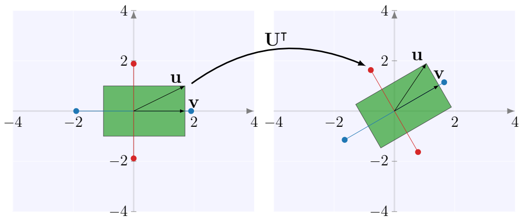
$$\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{25.61}$$

is called an affine transformation of $\mathbf{x}$.

### 6.3. Orthogonal Transformations

**Definition 25.43 Orthogonal Transformation:**
A linear transformation $T : \mathcal{V} \mapsto \mathcal{V}$ of an inner product space[def. 25.74] is an orthogonal transformation if preserves the inner product:

$$T(\mathbf{u}) \cdot T(\mathbf{v})\mathbf{u} \cdot \mathbf{v} \qquad \forall \mathbf{u}, \mathbf{v} \in \mathcal{V} \tag{25.62}$$



**Corollary 25.18 Orthogonal Matrix Transformation:**
An orthogonal matrix[def. 25.21] $\mathbf{Q}$ provides an orthogonal transformation:

$$(\mathbf{Q}\mathbf{u})^\intercal(\mathbf{Q}\mathbf{v}) = \mathbf{u}\mathbf{v} \tag{25.63}$$

**Explanation 25.2** (Improper Rotations).
*Orthogonal transformations in two or three dimensional euclidean space[def. 25.43] represent improper rotations:*

- *Stiff Rotations*
- *Reflections*
- *Reflections+Rotations*

**Corollary 25.19 Preservation of Orthogonality:** Orthogonal transformation preserver orthogonality.

**Corollary 25.20** [proof 25.6]
**Preservation of Norm:**
An orthogonal transformation $\mathbf{Q} : \mathcal{V} \mapsto \mathcal{V}$ preservers the *length/norm*:

$$\|\mathbf{u}\|_{\mathcal{V}} = \|\mathbf{Q}\mathbf{u}\|_{\mathcal{V}} \tag{25.64}$$

**Corollary 25.21 Preservation of Angle:**
An orthogonal transformation $T$ preservers the *angle*[def. 25.62] of its vectors:

$$\angle(\mathbf{u}, \mathbf{v}) = \angle(T(\mathbf{u}), T(\mathbf{v})) \tag{25.65}$$

### 6.4. Kernel & Image

### 6.4.1. Kernel

**Definition 25.44 Kernel/Null Space** $\mathtt{N}/\varphi^{-1}(\{0\})$**:**
Let $\varphi$ be a linear mapping[def. 25.40] between two a $\mathbb{K}$-vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The *kernel* of $\varphi$ is defined as:

$$\mathtt{N}(\varphi) := \varphi^{-1}(\{\mathbf{0}\}) = \{\mathbf{v} \in \mathcal{V} \mid \varphi(\mathbf{v}) = \mathbf{0}\} \subseteq \mathcal{V} \tag{25.66}$$

**Definition 25.45 Right Null Space** $\mathtt{N}(\mathbf{A})$**:**
If $\varphi = \mathbf{A} =\in \mathbb{K}^{m\times n}$ then the eq. (25.66) is equal to:

$$\mathtt{N}(\mathbf{A}) = \varphi_{\mathbf{A}}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^n \mid \mathbf{A}\mathbf{v} = \mathbf{0}\} \in \mathbb{K}^m \tag{25.67}$$

**Definition 25.46 Left Null Space** $\mathtt{N}(\mathbf{A}^\intercal)$**:**
If $\varphi = \mathbf{A} =\in \mathbb{K}^{m\times n}$ then the *left* null space is defined as:

$$\mathtt{N}(\mathbf{A}^\intercal) = \varphi_{\mathbf{A}^\intercal}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^m \mid \mathbf{A}^\intercal\mathbf{v} = \mathbf{0}\} \in \mathbb{K}^n \tag{25.68}$$

**Note**
The term *left* null space stems from the fact that:

$$(\mathbf{A}^\intercal\mathbf{x})^\intercal = \mathbf{0} \qquad \text{is equal to} \qquad \mathbf{x}^\intercal\mathbf{A} = \mathbf{0}$$

### 6.4.2. Image

**Definition 25.47 Image/Range** $\mathfrak{R}/\varphi$**:**
Let $\varphi$ be a linear mapping[def. 25.40] between two a $\mathbb{K}$-vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The *imgae* of $\varphi$ is defined as:

$$\mathfrak{R}(\varphi) := \varphi(\mathcal{V}) = \{\varphi(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}\} \subseteq \mathcal{W} \tag{25.69}$$

**Definition 25.48 Column Space** $\mathbf{A}\mathbf{x}$**:**
If $\varphi = \mathbf{A} = (\mathbf{c}_1 \cdots\cdots \mathbf{c}_n) \in \mathbb{K}^{m\times n}$ then eq. (25.69) is equal to:

$$\begin{aligned}\mathfrak{R}(\mathbf{A}) = \varphi_{\mathbf{A}}(\mathbb{K}^n) &= \left\{\mathbf{A}\mathbf{x}\middle|\forall \mathbf{x} \in \mathbb{K}^n\right\} = \left\langle(\mathbf{c}_1 \cdots\cdots \mathbf{c}_n)\right\rangle \\ &= \left\{\mathbf{v}\middle| \sum_{i=1}^{n} \alpha_i\mathbf{c}_i, \forall \alpha_i \in \mathbb{K}\right\}\end{aligned} \tag{25.70}$$

**Definition 25.49 Row Space** $\mathbf{A}^\intercal\mathbf{x}$**:**
If $\varphi = \mathbf{A} = (\mathbf{r}_1^\intercal \cdots\cdots \mathbf{r}_m^\intercal) \in \mathbb{K}^{m\times n}$ then the column space is defined as:

$$\begin{aligned}\mathfrak{R}(\mathbf{A}^\intercal) = \varphi_{\mathbf{A}}(\mathbb{K}^m) &= \left\{\mathbf{A}^\intercal\mathbf{x}\middle|\forall \mathbf{x} \in \mathbb{K}^m\right\} = \left\langle(\mathbf{r}_1 \cdots\cdots \mathbf{r}_m)\right\rangle \\ &= \left\{\mathbf{v}\middle| \sum_{i=1}^{m} \alpha_i\mathbf{r}_i, \forall \alpha_i \in \mathbb{K}\right\}\end{aligned} \tag{25.71}$$

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

**Corollary 25.22 Orthogonality** [proof 25.12]**:**
The *right* (*left*) null space[def. 25.44] is *orthogonal*[def. 25.63] to the *row*[def. 25.49] (*column*[def. 25.48]) space:

$$\mathtt{N}(\mathbf{A}) \perp \mathfrak{R}(\mathbf{A}^\intercal) \qquad \text{and} \qquad \mathtt{N}(\mathbf{A}^\intercal) \perp \mathfrak{R}(\mathbf{A}) \tag{25.72}$$

### 6.4.3. Rank Nullity Theorem

**Theorem 25.2 Rank-Nullity theorem:**
Let $\mathcal{V}$ be a finite vector space and let $\varphi$ be a linear mapping $\varphi : \mathcal{V} \mapsto \mathcal{W}$ then it holds:

$$\dim(\mathcal{V}) = \dim \underbrace{\left(\varphi^{-1}(\{\mathbf{0}\})\right)}_{\text{Kernel}} + \dim \underbrace{(\varphi(\mathcal{V}))}_{\text{Image}} \tag{25.73}$$

**Corollary 25.23 Representation as Standardbases:**
For every linear mapping $\varphi : \mathbb{K}^n \mapsto \mathbb{K}^m$ there exists a matrix $\mathbf{A}$ that represents this mapping:

$$\varphi = \varphi_{\mathbf{A}} = (\varphi(\mathbf{e}_1) \cdots\cdots \varphi(\mathbf{e}_n)) \in \mathbb{K}^{m\times n} \tag{25.74}$$

where

## 7. Eigenvalues and Vectors

**Definition 25.50 Eigenvalues:** Given a square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ the eigenvalues

finish

**Definition 25.51 Spectrum:** The spectrum of a square matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ is the set of its eigenvalues[def. 25.50]:

$$\text{spectrum}(\mathbf{A}) = \lambda(\mathbf{A}) = \{\lambda_1, \ldots, \lambda_n\} \tag{25.75}$$

**Formula 25.1 Eigenvalues of a 2x2 matrix:** Given a 2x2-matrix $\mathbf{A}$ its eigenvalues can be calculated by:

$$\{\lambda_1, \lambda_2\} \in \frac{\mathtt{tr}(\mathbf{A}) \pm \sqrt{\mathtt{tr}(\mathbf{A})^2 - 4\det(\mathbf{A})}}{2} \tag{25.76}$$
$$\text{with} \qquad \mathtt{tr}(\mathbf{A}) = a + d \qquad \det(\mathbf{A}) = ad - bc$$

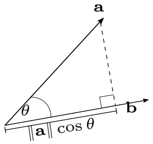
# 8. Vector Algebra

## 8.1. Dot/Standard Scalar Product

**Definition 25.52 Scalar Projection** $a_b$:

The scalar projection of a vector $\mathbf{a}$ onto a vector $\mathbf{b}$ is the *scalar* magnitude of the shadow/projection of the vector $\mathbf{a}$ onto $\mathbf{b}$:

$$a_b = \|\mathbf{a}\| \cos\theta_{a,b} = \mathbf{a}\tilde{\mathbf{b}} \tag{25.77}$$



**Definition 25.53** [proof 25.4]
**Standard Scalar/Dot Product:**
Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ the standard scalar product is defined as:

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v} = \mathbf{u}^\mathsf{T}\mathbf{v} = \langle \mathbf{u}, \mathbf{v}\rangle = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + \cdots + u_n v_n \\ = \|a\|\|b\| \cos\theta = u_v \hat{\mathbf{v}} = v_u \hat{\mathbf{u}} \quad \theta \in [0, \pi]\end{aligned} \tag{25.78}$$

**Explanation 25.3** (Geometric Interpretation).
*It is the magnitude of one vector times the magnitude of the shadow/scalar projection of the other vector.*
*Thus the dot product tells you:*
1. *How much are two vectors pointing into the same direction*
2. *With what magnitude*

**Property 25.12 Orthogonal Direction** $\perp$:
For $\theta \in [-\pi, \pi/2]$ rad $\cos\theta = 0$ and it follows:

$$\mathbf{u}\cdot\mathbf{v} = 0 \iff \mathbf{u} \perp \mathbf{v} \tag{25.79}$$

**Note: Perpendicular**

Perpendicular corresponds to orthogonality of two lines.

**Property 25.13 Maximizing Direction:**
For $\theta = 0$ rad $\cos\theta = 1$ and it follows:

$$\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \tag{25.80}$$

**Property 25.14 Minimizing Direction:**
For $\theta = \pi$rad $\cos\theta = -1$ and it follows:

$$\mathbf{u}\cdot\mathbf{v} = -\|\mathbf{u}\|\|\mathbf{v}\| \tag{25.81}$$

**Definition 25.54 Vector Projecion:**

General Projection via normal equation into inner product stuff i.e. with projection theorem

## 8.2. Cross Product

## 8.3. Outer Product

**Definition 25.55 Outer Product** $\mathbf{u}\mathbf{v}^\mathsf{T} = \mathbf{u} \otimes \mathbf{v}$:
Given two vectors $\mathbf{u} \in \mathbb{K}^m$, $\mathbf{v} \in \mathbb{K}^n$ their outer product is defined as:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^\mathsf{H} = [\mathbf{u}_1 \cdots\cdots \mathbf{u}_m] \begin{bmatrix} \bar{\mathbf{v}}_1 \\ \vdots \\ \bar{\mathbf{v}}_n \end{bmatrix} \tag{25.82}$$

$$= \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 \\ \vdots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1\bar{\mathbf{v}}_1 & \mathbf{u}_1\bar{\mathbf{v}}_2 \cdots\cdots & \mathbf{u}_1\bar{\mathbf{v}}_n \\ \mathbf{u}_2\bar{\mathbf{v}}_1 & \mathbf{u}_2\bar{\mathbf{v}}_2 \cdots\cdots & \mathbf{u}_2\bar{\mathbf{v}}_n \\ \vdots & \vdots \quad \ddots & \vdots \\ \mathbf{u}_m\bar{\mathbf{v}}_1 & \mathbf{u}_m\bar{\mathbf{v}}_2 \cdots\cdots & \mathbf{u}_m\bar{\mathbf{v}}_n \end{bmatrix}$$

**Proposition 25.2** [proof 25.5]
**Rank of Outer Product:** The outer product of two vectors is of rank one:

$$\mathfrak{rank}(\mathbf{u} \otimes \mathbf{v}) = 1 \tag{25.83}$$

## 8.4. Vector Norms

**Definition 25.56 Norm** $\|\cdot\|_\mathcal{V}$:
Let $\mathcal{V}$ be a vector space over a field $F$, a norm on $\mathcal{V}$ is a map:

$$\|\cdot\|_\mathcal{V} : \mathcal{V} \mapsto \mathbb{R}_+ \tag{25.84}$$

that satisfies:

$$\|\mathbf{x}\|_\mathcal{V} = 0 \iff \mathbf{x} = 0 \qquad \text{(Definitness)} \tag{25.85}$$

$$\|\alpha\mathbf{x}\|_\mathcal{V} = |\alpha|\|\mathbf{x}\|_\mathcal{V} \qquad \text{(Homogenity)} \tag{25.86}$$

$$\|\mathbf{x}+\mathbf{y}\|_\mathcal{V} \leqslant \|\mathbf{x}\|_\mathcal{V} + \|\mathbf{y}\|_\mathcal{V} \qquad \text{(Triangular Inequality)} \tag{25.87}$$

$$\alpha \in \mathbb{K} \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$$

**Explanation 25.4** (Definition 25.56).
*A norm is a measures of the size of its argument.*

**Corollary 25.24 Normed vector space:** Is a vector space $\mathcal{V}$ over a field $F$, on which a norm $\|\cdot\|_\mathcal{V}$ can be defined.

### 8.4.1. Cauchy Schwartz

**Definition 25.57** [proof 25.21]
**Cauchy Schwartz Inequality:**

$$|\mathbf{u}^\mathsf{T}\mathbf{v}| \leqslant \|\mathbf{u}\|\|\mathbf{v}\| \tag{25.88}$$

### 8.4.2. Triangular Inequality

**Definition 25.58** [proof 25.22]
**Triangular Inequality:** States that the length of the sum of two vectors is lower or equal to the sum of their individual lengths:

$$\|\mathbf{u}+\mathbf{v}\| \leqslant \|\mathbf{u}\| + \|\mathbf{v}\| \tag{25.89}$$

**Corollary 25.25 Reverse Triangular Inequality:**

$$-\|\mathbf{x}-\mathbf{y}\|_\mathcal{V} \leqslant \|\mathbf{x}\|_\mathcal{V} - \|\mathbf{y}\|_\mathcal{V} \leqslant \|\mathbf{x}-\mathbf{y}\|_\mathcal{V}$$

resp. $$\big|\|\mathbf{x}\|_\mathcal{V} - \|\mathbf{y}\|_\mathcal{V}\big| \leqslant \|\mathbf{x}-\mathbf{y}\|_\mathcal{V}$$

## 8.5. Distances

**Definition 25.59 Distance Function/Measure:** Let $S$ be a set a distance functions is a mapping $d$ defines as:

$$d : S \times S \mapsto \mathbb{R}_+$$

that satisfies:

$$d(x, x) = 0 \qquad \text{(Zero Identity Distance)} \tag{25.90}$$

$$d(x, y) = d(y, x) \qquad \text{(Symmetry)} \tag{25.91}$$

$$d(x, z) \leqslant d(x, y) + d(y, z) \qquad \text{(Triangular Identiy)} \tag{25.92}$$

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

**Explanation 25.5** (Definition 25.59).
*Is measuring the distance between two things.*

### 8.5.1. Contraction

**Definition 25.60 Contraction:** Given a metric space $(M, d)$ is a mapping $f : M \mapsto M$ that satisfies:

$$d(f(x), f(y)) \leqslant \lambda d(x, y) \qquad \lambda \in [0, 1) \tag{25.93}$$

add metric spaces

## 8.6. Metrics

**Definition 25.61 Metric:**
Is a distance measure that additonally satisfies: $\forall x, y \in S$
$d(x, y) = 0 \iff x = y$ (identity of indiscernibles)

**Corollary 25.26 Metric→Norm:** Every norm $\|\cdot\|_\mathcal{V}$ on a vector space $\mathcal{V}$ over a field $F$ induces a metric by:

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

metric induced by norms additionally satisfy: $\forall x, y \in \mathcal{V}$, $\alpha \in F \subseteq \mathbb{K}$ $K = \mathbb{R}$ or $\mathbb{C}$
1. Homogenity/Scaling: $d(\alpha x, \alpha y)_\mathcal{V} = |\alpha| d(x, y)_\mathcal{V}$
2. Translational 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:

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

**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**

(bilinear form$\xrightarrow{\text{induces}}$)
inner product$\xrightarrow{\text{induces}}$norm$\xrightarrow{\text{induces}}$metric.

# 9. Angles

**Definition 25.62 Angle between Vectors** $\angle(\mathbf{u}, \mathbf{v})$: Let $\mathbf{u}, \mathbf{v} \in \mathbb{K}^n$ be two vectors of an inner product space[def. 25.74] $\mathcal{V}$. The angle $\alpha \in [0, \pi]$ between $\mathbf{u}, \mathbf{v}$ is defined by:

$$\angle(\mathbf{u}, \mathbf{v}) := \alpha \qquad \cos\alpha = \frac{\mathbf{u}^\mathsf{T}\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \qquad \begin{array}{l}\mathbf{u}, \mathbf{v} \in \mathcal{V} \\ \alpha \in [0, \pi]\end{array} \tag{25.94}$$

# 10. Orthogonality

**Definition 25.63 Orthogonal Vectors:** Let $\mathcal{V}$ be an inner-product space[def. 25.74]. A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \in \mathcal{V}$ is called *orthogonal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \qquad \forall i \neq \tag{25.95}$$

## 10.1. Orthonormality

**Definition 25.64 Orthonormal Vectors:** Let $\mathcal{V}$ be an inner-product space[def. 25.74]. A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n, \ldots\} \in \mathcal{V}$ is called *orthonormal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \forall i, j \tag{25.96}$$

# 11. Special Kind of Vectors

## 11.1. Binary/Boolean Vectors

**Definition 25.65**
**Binary/Boolean Vectors/Bit Maps** $\mathbb{B}^n$: Are vectors that contain only zero or one values:

$$\mathbb{B}^n = \{0, 1\}^n \tag{25.97}$$

**Definition 25.66**
**R-Sparse Boolean Vectors** $\mathbb{B}^n_r$:
Are boolean vectors that contain exact $r$ one values:

$$\mathbb{B}^n_r = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^\mathsf{T}\mathbf{x} = \sum_{i=1}^{n} \mathbf{x} = r \right\} \tag{25.98}$$

## 11.2. Probablistic Vectors

**Definition 25.67 Probabilistic Vectors:** Are vectors that represent probabilities and satisfy:

$$\left\{ \mathbf{x} \in [0, 1]^n : \sum_{i=1}^{n} x_i = 1 \right\} \tag{25.99}$$

# 12. Vector Spaces and Measures

## 12.1. Bilinear Forms

## 12.2. Quadratic Forms

### 12.2.1. Min/Max Value

**Corollary 25.27** [proof 25.20]
**Extreme Value:** The minimum/maximum of a quadratic form?? with a quadratic matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is given by the eigenvector corresponding to the smallest/largest eigenvector of $\mathbf{A}$:

$$\mathbf{v}_1 \in \operatorname*{arg\,min}_{\mathbf{x}^\mathsf{T}\mathbf{x}=1} \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \qquad \mathbf{v}_1 \in \operatorname*{arg\,max}_{\mathbf{x}^\mathsf{T}\mathbf{x}=1} \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \tag{25.100}$$

**Note**

$$(\mathbf{Q}^\mathsf{T}\tilde{\mathbf{n}})^\mathsf{T}\mathbf{Q}^\mathsf{T}\tilde{\mathbf{n}} = \tilde{\mathbf{n}}^\mathsf{T}\mathbf{Q}\mathbf{Q}^\mathsf{T}\tilde{\mathbf{n}} = \tilde{\mathbf{n}}^\mathsf{T}\tilde{\mathbf{n}} = 1$$

### 12.2.2. Skew Symmetric Matirx

**Corollary 25.28**
**Quadratic Form of Skew Symmetric matrix:** The quadratic form of a skew symmetric matrix[def. 25.26] vanishes:

$$\alpha = \mathbf{x}^\mathsf{T}\mathbf{A}_{\text{skew}}\mathbf{x} = \left(\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}_{\text{skew}}\mathbf{x}\right)^\mathsf{T} = \left(\mathbf{x}^\mathsf{T}\mathbf{A}_{\text{skew}}\mathbf{x}\right)^\mathsf{T} = -\alpha \tag{25.101}$$

Which can only hold iff $\alpha = 0$.

## 12.3. Inner Product – Generalization of the dot product

**Definition 25.68 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:

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

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

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

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

**Definition 25.69 Symmetric bilinear form:** A bilinear form $a$ on $\mathcal{V}$ is symmetric if and only if:

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

**Definition 25.70 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:

$$a(u, u) > 0 \qquad \forall u \in \mathcal{V} \backslash \{0\} \tag{25.102}$$

And positive semidefinte $\iff \geqslant$ (25.103)

**Corollary 25.29 Matrix induced Bilinear Form:**
For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any *symmetric* matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ induces a bilinear form:

$$a(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}' = (\mathbf{A}\mathbf{x}')\mathbf{x},$$

**Definition 25.71 Positive (semi) definite Matrix** $\succ$:
A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is positive defintie if and only if:

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0 \iff \mathbf{A} \succ 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \backslash \{0\} \tag{25.104}$$

And **positive semidefinte** $\iff \succcurlyeq$ (25.105)

**Corollary 25.30** [proof 25.2]
**Eigenvalues of positive (semi) definite matrix:**
A positive definite matrix is a matrix where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive.*

$$\forall \lambda_i \in \text{eigenv}(\mathbf{A}) > 0 \tag{25.106}$$

And positive semidefinte $\iff \geqslant$ (25.107)

**Note**

Positive definite matrices are often assumed to be symmetric but that is not necessarily true.

**Proof 25.2.** ?? *25.2 (for real matrices):*
*Let* $\mathbf{v}$ *be an eigenvector of* $\mathbf{A}$ *then it follows:*

$$0 \overset{??\ 25.2}{<} \mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v} = \mathbf{v}^\mathsf{T}\lambda\mathbf{v} = \|\mathbf{v}\|\lambda$$

**Corollary 25.31 Positive Definiteness and Determinant:** The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

**Definition 25.72 Negative (semi) definite Matrix $\prec$:**
A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is negative defintie if and only if:

$$\mathbf{x}^\intercal \mathbf{A}\mathbf{x} < 0 \quad \iff \quad \mathbf{A} \prec 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\} \tag{25.108}$$

$$\text{And } \textbf{negative semidefinte} \iff \preccurlyeq \tag{25.109}$$

**Theorem 25.3 Sylvester's criterion:** Let $\mathbf{A}$ be *symmetric/Hermitian* matrix and denote by $\mathbf{A}^{(k)}$ the $k \times k$ upper left sub-matrix of $\mathbf{A}$.
Then it holds that:

- $$\mathbf{A} \succ 0 \quad \iff \quad \det\left(\mathbf{A}^k\right) > 0 \qquad k = 1, \dots, n \tag{25.110}$$
- $$\mathbf{A} \prec 0 \quad \iff \quad (-1)^k \det\left(\mathbf{A}^k\right) > 0 \qquad k = 1, \dots, n \tag{25.111}$$
- $\mathbf{A}$ is indefinite if the first $\det\left(\mathbf{A}^k\right)$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive ($\mathbf{A}$ can be anything of the previous three) if the first $\det\left(\mathbf{A}^k\right)$ that breaks both patterns is 0.

## 13. Inner Products

**Definition 25.73 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:

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \mapsto F \subseteq \mathbb{K} \qquad K = \mathbb{R} \text{ or } \mathbb{C} \tag{25.112}$$

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

1. (Conjugate) Stmmetry: $\quad \langle x, y \rangle = \overline{\langle x, y \rangle}$.
2. Linearity in the first argument:
$$\langle \alpha x + \beta y, z \rangle = \alpha\langle x, z \rangle + \beta\langle y, z \rangle$$
3. Positve-definiteness:
$$\langle x, x \rangle \geqslant 0 : x = 0 \iff \langle x, x \rangle = 0$$

**Definition 25.74 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 25.32 Inner product↦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** $a(u, v) = u^\intercal \mathbf{I} v$ 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 $\langle \mathbf{x}, \mathbf{x} \rangle \geqslant 0$, whereas bilinear forms must not.

**Corollary 25.33 Inner product induced norm $\langle \cdot, \cdot \rangle_{\mathcal{V}} \to \|\cdot\|_{\mathcal{V}}$:** Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ induces a norm of the form:

$$\|\mathbf{x}\|_{\mathcal{V}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \qquad \mathbf{x} \in \mathcal{V}$$

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

**Corollary 25.34 Energy Norm:** A *s.p.d.* bilinear form $a : \mathcal{V} \times \mathcal{V} \mapsto F$ induces an energy norm:

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

## 14. Matrix Algebra

## 15. Matrix Norms

### 15.1. Operator Norm

**Definition 25.75 Operator/Induced Norm:**
Let $\|\cdot\|_\mu : \mathbb{K}^m \mapsto \mathbb{R}$ and $\|\cdot\|_\nu : \mathbb{K}^n \mapsto \mathbb{R}$ be vector norms.
The operator norm is defined as:

$$\|\mathbf{A}\|_{\mu,\nu} := \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \mathbf{x}\neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_\mu}{\|\mathbf{x}\|_\nu} = \sup_{\|\mathbf{x}\|_\nu = 1} \|\mathbf{A}\mathbf{x}\|_\mu \qquad \|\cdot\|_\mu : \mathbb{K}^m \mapsto \mathbb{R} \tag{25.113}$$

**Explanation 25.6** (Definition 25.75). *Is a measure for the largest factor by which a matrix* $\mathbf{A}$ *can stretch a vector* $\mathbf{x} \in \mathbb{R}^n$.

### 15.2. Induced Norms

**Corollary 25.35 Induced Norms:** Let $\|\cdot\|_p : \mathbb{K}^{m\times n} \mapsto \mathbb{R}$ defined as:

$$\|\mathbf{A}\|_p := \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \mathbf{x}\neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{y}\|_p = 1} \|\mathbf{A}\mathbf{y}\|_p \tag{25.114}$$

**Explanation 25.7** ([Corollary 25.35]).
*Induced norms are matrix norms induced by vector norms as we:*
- *Only work with vectors* $\mathbf{A}\mathbf{x}$
- *And use the normal p-vector norms* $\|\cdot\|_p$

**Note supremum**

The set of vectors $\{\mathbf{y} | \|\mathbf{y}\| = 1\}$ is compact, thus if we consider finite matrices the supremum is attained and we may replace it by the max.

### 15.3. Induced Norms

#### 15.3.1. 1-Norm

**Definition 25.76 Column Sum Norm** $\|\mathbf{A}\|_1$**:**

$$\|\mathbf{A}\|_1 = \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \mathbf{x}\neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1\leqslant j\leqslant n} \sum_{i=1}^m |a_{ij}| \tag{25.115}$$

#### 15.3.2. ∞-Norm

**Definition 25.77 Row Sum Norm** $\|\mathbf{A}\|_\infty$**:**

$$\|\mathbf{A}\|_\infty = \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \mathbf{x}\neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1\leqslant i\leqslant m} \sum_{j=1}^n |a_{ij}| \tag{25.116}$$

#### 15.3.3. Spectral Norm L2-Norm

**Spectral Radius & Singular Value**

**Definition 25.78 Spectral Radius** $\rho(\mathbf{A})$**:**
The spectral radius is defined as the largest eigenvalue of a matrix:

$$\rho(\mathbf{A}) = \max\{\lambda | \lambda \in \operatorname{eigenval}(\mathbf{A})\} \tag{25.117}$$

**Definition 25.79 Singular Value** $\sigma_i$**:**
Given a matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ its $n$ real and positive singular values are defined as:

$$\sigma(\mathbf{A}) := \left\{\left\{\sqrt{\lambda_i}\right\}_{i=1}^n \middle| \lambda_i \in \operatorname{eigenval}\left(\mathbf{A}^\intercal \mathbf{A}\right)\right\} \tag{25.118}$$

**Spectral Norm**

**Definition 25.80 L2/Spectral Norm** $\|\mathbf{A}\|_2$**:**

$$\|\mathbf{A}\|_2 = \sup_{\substack{\mathbf{x}\in\mathbb{K}^n \\ \|\mathbf{x}\|_2 = 1}} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \sqrt{\mathbf{x}^\intercal \mathbf{A}^\intercal \mathbf{A}\mathbf{x}} \tag{25.119}$$

$$= \max_{\|\mathbf{x}\|_2 = 1} \sqrt{\rho(\mathbf{A}^\intercal \mathbf{A}))} =: \sigma_{\max}(\mathbf{A}) \tag{25.120}$$

### 15.4. Energy Norm

### 15.5. Forbenius Norm

**Definition 25.81 Forbenius Norm** $\|\mathbf{A}\|_F$**:**
The *Forbenius norm* $\|\cdot\|_F : \mathbb{K}^{m\times n} \mapsto \mathbb{R}$ is defined as:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}^2|} = \sqrt{\operatorname{tr}\left(\mathbf{A}\mathbf{A}^\mathsf{H}\right)} \tag{25.121}$$

### 15.6. Distance

## 16. Decompositions

### 16.1. Eigen/Spectral decomposition

**Definition 25.82** $\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}$,[proof 25.25]
**Eigendecomposition/ Spectral Decomposition :**
Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a *diagonalizable* square matrix and define by $\mathbf{X} = [\mathbf{x}_1 \cdots\cdots \mathbf{x}_n] \in \mathbb{R}^{n\times n}$ a non-singular matrix whose column vectors are the eigenvectors of $\mathbf{A}$ with associated eigenvalue matrix $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Then $\mathbf{A}$ can be represented as:

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \tag{25.122}$$

**Proposition 25.3 Diagonalization:** If non of $\mathbf{A}$ eigenvalues are zero it can be diagonalized:

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \boldsymbol{\Lambda} \tag{25.123}$$

**Proposition 25.4 Existence:**

$$\exists \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \iff \mathbf{A} \text{ diagonalizable} \tag{25.124}$$

### 16.2. QR-Decompositions

### 16.3. Singular Value Decomposition

**Definition 25.83**
**Singular Value Decomposition (SVD)** $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{H}$**:**
For any matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ there exist unitary matrices[def. 25.24]

$$\mathbf{U} \in \mathbb{K}^{m,m} \qquad \mathbf{V} \in \mathbb{K}^{n,n}$$

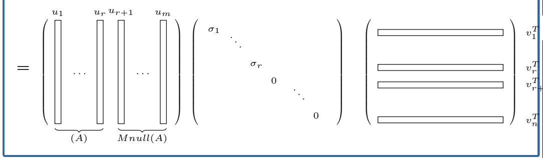
and a (generalized) digonal matrix:

$$p := \min\{m, n\}$$

$$\boldsymbol{\Sigma} \in \mathbb{R}^{m,n} \qquad \boldsymbol{\Sigma} = \operatorname{gendiag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m,n}$$

such that:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{H} \tag{25.125}$$



Images full, economical, alternative representation, range and kernel
https://math.mit.edu/classes/18.095/2016IAP/lec2/SVD_Notes.pdf

#### 16.3.1. Eigenvalues

**Proposition 25.5** [proof 25.23]**:**
The eigenvalues of a matrix $\mathbf{A}^\intercal \mathbf{A}$ are positive.

**Proposition 25.6** [proof 25.24]
**Similarity Transformation:** The unitary matrix $\mathbf{V}$ provides a *similarity transformation*[cor. 25.14] of $\mathbf{A}^\intercal \mathbf{A}$ into a diagonal matrix $\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma} \mapsto \mathbf{V}^\mathsf{H} \mathbf{A}^\intercal \mathbf{A}\mathbf{V} \tag{25.126}$$

**Corollary 25.36 eigenval($\mathbf{A}^\intercal \mathbf{A}$) = eigenval($\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma}$):**
From proposition 25.6 and [cor. 25.15] it follows that:

$$\operatorname{eigenval}(\mathbf{A}^\intercal \mathbf{A}) = \operatorname{eigenval}(\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma}) \tag{25.127}$$

$$\Longrightarrow \quad \|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^\intercal \mathbf{A})} = \sqrt{\lambda_{\max}} = \sigma_{\max}$$

**Note**

$\lambda$ and *singularvalue* corresponds to the eigenvalues/singular-values of $\mathbf{A}^\intercal \mathbf{A}$ and not $\mathbf{A}$

#### 16.3.2. Best Lower Rank Approximation

**Theorem 25.4 Eckart Yound Theorem:** Given a matrix $\mathbf{X} \in \mathbb{K}^{m,n}$ the *reduced* SVD $\mathbf{X}$ defined as:

$$\mathbf{U}_k := [\mathbf{u}_{:,1} \cdots\cdots \mathbf{u}_{:,k}] \in \mathbb{K}^{m,k}$$

$$\mathbf{X}_k := \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\mathsf{H} \qquad \boldsymbol{\Sigma}_k = \operatorname{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k,k}$$

$$k \leqslant \min\{m, n\}$$

$$\mathbf{V}_k = [\mathbf{v}_{:,1} \cdots\cdots \mathbf{v}_{:,k}] \in \mathbb{K}^{n,k}$$

provides the best lower $k$ rank approximation of $\mathbf{X}$:

$$\min_{\mathbf{Y}\in\mathbb{K}^{n,m}:\mathfrak{rank}(\mathbf{Y})\leqslant k} \|\mathbf{X} - \mathbf{Y}\|_F = \|\mathbf{X} - \mathbf{X}_k\|_F \tag{25.128}$$

## 17. Matric Calculus

### 17.1. Derivatives

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{b}) = \mathbf{b}$$
$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A} \tag{25.129}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\intercal \mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^\intercal)\mathbf{x} \tag{25.130}$$

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

## 18. Proofs

**Proof 25.3.** *[def. 25.33]*

$$\mathbf{M}\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{I}_{k,k} & \mathbf{0}_{k,l} \\ \mathbf{0}_{l,k} & \mathbf{I}_{l,l} \end{bmatrix} \tag{25.131}$$

### 18.1. Vector Algebra

**Proof 25.4** (Definition 25.53).

$$\textbf{(1)}: \underline{\|a - b\|} \overset{eq.\ (26.19)}{=} \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \cos\theta$$
$$\textbf{(2)}: \underline{\|a - b\|} = (a - b)(a - b) = \|a\|^2 + \|b\|^2 - 2(ab)$$
$$\underline{\|a - b\|} = \underline{\|a - b\|} \quad \Rightarrow \quad ab = \|a\|\|b\| \cos\theta$$

**Proof 25.5** (Proposition 25.2). *The outer product of* $\mathbf{u}$ *with* $\mathbf{v}$ *corresponds to a scalar multiplication of* $\mathbf{v}$ *with elements* $u_i$ *thus the rank must be that of* $\mathbf{v}$, *which is a vector and hence of rank*

*1*
$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^\mathsf{H} = \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 \\ \vdots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix}$$

### 18.2. Mappings

**Proof 25.6.** *Corollary 25.20*

$$\|\mathbf{Q}\mathbf{x}\|^2 = (\mathbf{Q}\mathbf{x})^\intercal \mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal \mathbf{Q}^\intercal \mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal \mathbf{x} = \|\mathbf{x}\|^2$$

**Proof 25.7.** *Corollary 25.21 Follows immediately from definition 25.62 in combination with eqs.* (25.62) *and* (25.64).

**Proof 25.8.** *Proposition 25.1:*

$$\Longrightarrow \quad l(\alpha\mathbf{x} + \beta\mathbf{y}) \overset{25.55}{=} l(\alpha\mathbf{x}) + l(\beta\mathbf{y}) \overset{25.56}{=} \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$$
$$\Longleftarrow \quad l(\alpha\mathbf{x} + \mathbf{0}) = \alpha l(\mathbf{x})$$
$$l(1\mathbf{x} + 1\mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y})$$

**Proof 25.9** (principle 25.1).
*Every vector* $\mathbf{v} \in \mathcal{V}$ *can be represented by a basis eq.* (25.13) *of* $\mathcal{V}$. *With homogentityeq.* (25.56) *and additivityeq.* (25.55) *it follows for the image of all* $\mathbf{v} \in \mathcal{V}$:

$$l(\mathbf{v}) = l(\alpha_1 b_1 + \cdots + \alpha_n b_n) = l\alpha_1(b_1) + \cdots + l(\alpha_n) b_n \tag{25.132}$$

$\Rightarrow$ *the image of the basis of* $\mathcal{V}$ *determines the linear mapping.*

**Proof 25.10** (Proof [Corollary 25.17]).

$$\Longrightarrow \quad l_{\mathbf{A}}(\alpha\mathbf{x}+\mathbf{y}) = \mathbf{A}\left(\alpha\mathbf{x}+\beta\mathbf{y}\right) = \alpha\mathbf{A}\mathbf{x}+\beta\mathbf{A}\mathbf{y} = \alpha l(\mathbf{x})+\beta l(\mathbf{y})$$

$\Longleftarrow$ *Let $\mathfrak{B}$ be a standard normal basis of $\mathcal{V}$ with eq. (25.132):*

$$l(\mathbf{x}) = \sum_{i=1}^{n} x_i l\left(\mathbf{e}_i\right) = \sum_{i=1}^{n} x_i \mathbf{A}_{:,i} = \mathbf{A}\mathbf{x} \qquad \mathbf{A}_{:,i} := \mathbf{l}(\mathbf{e_i}) \in \mathbb{R}^n$$

**Proof 25.11** (Proof Property 25.11).

$$\begin{aligned}(g\circ f)(\alpha\mathbf{x}) &= g(f(\alpha\mathbf{x})) = g(\alpha f(\mathbf{x})) = \alpha(g\circ f)(\mathbf{x})\\ (g\circ f)(\mathbf{x}+\mathbf{y}) &= g(f(\mathbf{x}+\mathbf{y})) = g(f(\mathbf{x})+f(\mathbf{y}))\\ &= (g\circ f)(\mathbf{x}) + (g\circ f)(\mathbf{y})\end{aligned}$$

*or even simpler as every linear form can be represented by a matrix product:*

$$f(y) = \mathbf{A}\mathbf{y} \quad g(z) = \mathbf{B}\mathbf{z} \quad \Rightarrow \quad (f\circ g)(\mathbf{x}) = \mathbf{A}\mathbf{B}\mathbf{x} := \mathbf{C}\mathbf{x}$$

**Proof 25.12.** [Corollary 25.22] *Let $\mathbf{y} \in N(\mathbf{A})$ ($\mathbf{z} \in N(\mathbf{A}^\intercal)$) then it follows:*

$$\begin{aligned} N(\mathbf{A}) \perp \mathfrak{R}(\mathbf{A}^\intercal) \qquad & (\mathbf{A}^\intercal\mathbf{x})^\intercal\mathbf{y} = \mathbf{x}^\intercal\mathbf{A}\mathbf{y} = \mathbf{x}^\intercal\mathbf{0} = 0\\ N(\mathbf{A}^\intercal) \perp \mathfrak{R}(\mathbf{A}) \qquad & (\mathbf{A}\mathbf{x})^\intercal\mathbf{z} = \mathbf{x}^\intercal\mathbf{A}^\intercal\mathbf{z} = \mathbf{x}^\intercal\mathbf{0} = 0\end{aligned}$$

## 18.3. Special Matrices

**Proof 25.13** ([Corollary 25.15]). *Let $\mathbf{u} = \mathbf{S}^{-1}\mathbf{v}$ then it follows:*

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S}\mathbf{u} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}\mathbf{v} = \lambda\mathbf{S}^{-1}\mathbf{v} = \lambda\mathbf{u}$$

**Proof 25.14** (Property 25.6).

$$\|\mathbf{Q}\mathbf{x}\|_2^2 = (\mathbf{Q}\mathbf{x})^\intercal\mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal\mathbf{Q}^\intercal\mathbf{Q}\mathbf{x} = \|\mathbf{x}\|_2^2$$

**Proof 25.15.** *Property 25.4*
*Let $\mathbf{A} \in \mathbb{K}^{n\times n}$ be a hermitian matrix*[def. 25.23] *and let $\lambda \in \mathbb{K}$ be an eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\mathbf{v} \in \mathbb{K}^n$:*

$$\lambda(\bar{\mathbf{v}}^\intercal\mathbf{v}) = \bar{\mathbf{v}}^\intercal\lambda\mathbf{v} = \bar{\mathbf{v}}^\intercal\mathbf{A}\mathbf{v} = \overline{(\bar{\mathbf{v}}^\intercal\mathbf{A}\mathbf{v})} = \overline{\mathbf{A}\mathbf{v}}^\intercal\mathbf{v} = \bar{\lambda}\left(\bar{\mathbf{v}}^\intercal\mathbf{v}\right)$$

$$\lambda\left(\bar{\mathbf{v}}^\intercal\mathbf{v}\right) = \bar{\lambda}\left(\bar{\mathbf{v}}^\intercal\mathbf{v}\right)$$

1. $\bar{\mathbf{v}}\mathbf{v} = \sum_{i=1}^n |v_i|^2 > 0$ *as* $\mathbf{v} \neq \mathbf{0}$
2. $\lambda = \bar{\lambda}$ *which can only hold for* $\lambda \in \mathbb{R}$ *(Equation (19.6))*

**Proof 25.16.** ??

add

## 18.4. Vector Spaces

**Proof 25.17** (Definition 25.18). *We know that $proj_L(\mathbf{u})$ must be a vector times a certain magnitude:*

$$proj_L(\mathbf{u}) = \alpha\tilde{\mathbf{v}} \qquad \alpha \in \mathbb{K} \tag{25.133}$$

*the magnitude follows from the scalar projection*[def. 25.52] *in the direction of $\mathbf{v}$ which concludes the derivation.*

**Proof 25.18** (Definition 25.18 (via orthogonality)). *We know that $\mathbf{u} - proj_L(\mathbf{u})$ must be orthogonal*[def. 25.63] *to $\mathbf{v}$*

$$(\mathbf{u} - proj_L(\mathbf{u}))^\intercal\mathbf{v} = (\mathbf{u}-\alpha\mathbf{v})^\intercal\mathbf{v} = 0 \Rightarrow \qquad \alpha = \frac{\mathbf{u}^\intercal\mathbf{v}}{\mathbf{v}^\intercal\mathbf{v}}$$

**Proof 25.19.** *Definition 25.19 Let $\mathfrak{B} = \{\mathbf{b}_1,\dots,\mathbf{b}_n\}$ a basis of $\mathcal{U}$ s.t. by* [cor. 25.4]*:*

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i\mathbf{b}_i$$

*the coefficients $\{\alpha_i\}_{i=1}^n$ need to be determined. We know that:*

$$\mathbf{v}-\mathbf{u} \perp \mathbf{b}_1,\dots,\mathbf{v}-\mathbf{u}\perp\mathbf{b}_n$$

$$\Longrightarrow \qquad \left(\mathbf{v} - \sum_{i=1}^{n}\alpha_i\mathbf{b}_i\right)\cdot\mathbf{b}_j = 0 \qquad j = 1,\dots,n$$

*this linear system of equations can be rewritten as:*

$$(\mathbf{b}_1\cdots\cdots\mathbf{b}_n)\begin{pmatrix}\mathbf{b}_1\\ \vdots\\ \mathbf{b}_n\end{pmatrix}\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix} = \begin{pmatrix}\mathbf{b}_1\\ \vdots\\ \mathbf{b}_n\end{pmatrix}\mathbf{v}$$

**Proof 25.20.** *Corollary 25.27*
*Let $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\intercal$ be the eigendecomposition*[cor. 25.12] *of $\mathbf{A}$ then it follows:*

$$\begin{aligned}\min_{\tilde{\mathbf{n}}^\intercal\tilde{\mathbf{n}}=1} \tilde{\mathbf{n}}^\intercal\mathbf{A}\tilde{\mathbf{n}} &= \min_{\|\tilde{\mathbf{n}}\|=1} \tilde{\mathbf{n}}^\intercal(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\intercal)\tilde{\mathbf{n}}\\ &= \min_{\|\tilde{\mathbf{n}}\|=1} (\mathbf{Q}^\intercal\tilde{\mathbf{n}})^\intercal\mathbf{\Lambda}(\mathbf{Q}^T\tilde{\mathbf{n}})\\ &= \min_{\mathbf{x}=1}\mathbf{x}^\intercal\mathbf{\Lambda}\mathbf{x} \qquad \mathbf{x} := \mathbf{Q}^\intercal\tilde{\mathbf{n}}\\ &= \min_{\mathbf{x}=1}\sum_{i=1}^{n}\mathbf{x}_i^2\Lambda_{ii} = \min_{\mathbf{x}=1}\sum_{i=1}^{n}\mathbf{x}_i^2\lambda_i\end{aligned}$$

*Thus in order to obtain the minimum value we need to choose the eigenvector that leads to the smallest eigenvalue.*

## 18.5. Norms

**Proof 25.21.** ?? *25.21*

$$|\mathbf{u}\cdot\mathbf{v}| \overset{eq.\ (25.78)}{=} \|\mathbf{u}\|\|\mathbf{v}\||\cos\theta| \leqslant \|\mathbf{u}\|\|\mathbf{v}\|$$

**Proof 25.22.** *Definition 25.58*

$$\|\mathbf{u}+\mathbf{v}\|^2 = (\mathbf{u}+\mathbf{v})(\mathbf{u}+\mathbf{v}) = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\mathbf{u}\cdot\mathbf{v})$$

*from cauchy schwartz we know:*

$$\mathbf{u}\cdot\mathbf{v} \leqslant |\mathbf{u}\cdot\mathbf{v}| \overset{eq.\ (25.88)}{\leqslant} \|\mathbf{u}\|\|\mathbf{v}\|$$

$$\|\mathbf{u}+\mathbf{v}\|^2 \leqslant \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\|\mathbf{u}\|\|\mathbf{v}\|) = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2$$

## 18.6. Decompositions

### 18.6.1. Symmetric - Antisemitic

**Definition 25.84 Symmetric - Antisymmetric Decomposition:** Any matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ can be decomposed into the sum of a *symmetric matrix*[def. 25.20] $\mathbf{A}^{\text{sym}}$ and a *skew-symmetric matrix*?? $\mathbf{A}^{\text{skes}}$:

$$\mathbf{A} = \mathbf{A}^{\text{sym}} + \mathbf{A}^{\text{skew}} \qquad \begin{aligned}\mathbf{A}^{\text{sym}} &= \frac{1}{2}\left(\mathbf{A}+\mathbf{A}^{\mathsf{H}}\right)\\ \mathbf{A}^{\text{skew}} &= \frac{1}{2}\left(\mathbf{A}-\mathbf{A}^{\mathsf{H}}\right)\end{aligned} \tag{25.134}$$

### 18.6.2. SVD

**Proof 25.23** ([Corollary 25.5]). $\mathbf{B} := \mathbf{A}^\intercal\mathbf{A}$ *corresponds to a symmetric positive definite form*[def. 25.71]*:*

$$\mathbf{x}^\intercal\mathbf{B}\mathbf{x} = \mathbf{x}^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 > 0$$

*thusProposition 25.6 follows immediately form* [Corollary 25.2].

**Proof 25.24** (Proposition 25.6).

$$\mathbf{A}^\intercal\mathbf{A} \overset{SVD}{=} \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}\right)^{\mathsf{H}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}} = \mathbf{V}\mathbf{\Sigma}^{\mathsf{H}}\underbrace{\mathbf{U}^{\mathsf{H}}\mathbf{U}}_{\mathbf{I}_m}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}} = \mathbf{V}\mathbf{\Sigma}^{\mathsf{H}}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$$

$$\Longrightarrow \qquad \mathbf{V}^{\mathsf{H}}\mathbf{A}^\intercal\mathbf{A}\mathbf{V} = \mathbf{\Sigma}^\intercal\mathbf{\Sigma}$$

### 18.6.3. Eigendecomposition

**Proof 25.25** (Definition 25.82).

$$\mathbf{A}\mathbf{X} = \left[\lambda_1\mathbf{x}_1\cdots\cdots\lambda_n\mathbf{x}_n\right] = \mathbf{X}\mathbf{\Lambda}$$

# Geometry

**Corollary 26.1 Affine Transformation in 1D: Given:** numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The affine transformation of $\phi : \hat{\Omega} \to \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{26.1}$$

**Proof 26.1.** [cor. 26.1] *By* [def. 25.42] *we want a function* $f : [a, b] \to [c, d]$ *that satisfies:*

$$f(a) = c \qquad \textbf{and} \qquad f(b) = d$$

*additionally* $f(x)$ *has to be a linear function* ([def. 22.17])*, that is the output scales the same way as the input scales.*
**Thus** *it follows:*

$$\frac{d-c}{b-a} = \frac{f(x)-f(a)}{x-a} \qquad \iff \qquad f(x) = \frac{d-c}{b-a}(x-a) + c$$

## Trigonometry

### 0.1. Trigonometric Functions

### 0.1.1. Sine

**Definition 26.1 Sine:**

$$\sin\alpha = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{a}{c} \tag{26.2}$$

### 0.1.2. Cosine

**Definition 26.2 Cosine:**

$$\cos\alpha\alpha = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{b}{c} \tag{26.3}$$

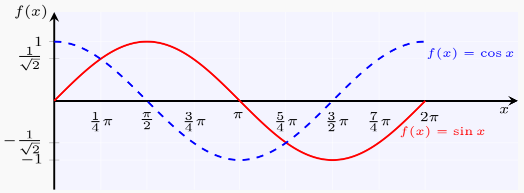
### 0.1.3. Tangens

**Definition 26.3 Tangens:**

$$\cos\alpha\alpha = \frac{\text{opposite}}{\text{adjacent}} = \frac{a}{b} = \frac{a/c}{b/c} = \frac{\sin\alpha}{\cos\alpha} \tag{26.4}$$

### 0.1.4. Trigonometric Functions and the Unit Circle

**Sine and Cosine**



$$\cos x \overset{(22.48)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{26.5}$$

$$\sin x \overset{(22.48)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{26.6}$$

**Note**

Using theorem 26.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{26.7}$$

### 0.1.5. Sinh

**Definition 26.4 Sinh:**

$$\sinh x \overset{(eq.\ (22.48))}{=} \frac{1}{2}\left[e^{x} - e^{-x}\right] = -\mathrm{i}\sin(\mathrm{i}\,x) \tag{26.8}$$

**Property 26.1:** $\sinh x = 0$ has a unique root at $x = 0$.

### 0.1.6. Cosh

**Definition 26.5 Cosh:**

$$\cosh x \overset{(22.48)}{=} \frac{1}{2}\left[e^{x} + e^{-x}\right] = \cos(\mathrm{i}\,x) \tag{26.9}$$

$$\tag{26.10}$$

**Property 26.2:** $\cosh x$ is strictly positive.

**Proof 26.2.**

$$e^{x} = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{26.11}$$

### 0.2. Addition Theorems

**Theorem 26.1 Addition Theorems:**

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{26.12}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{26.13}$$

### 0.3. Werner Formulas

**Werner Formulas**

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right] \tag{26.14}$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \tag{26.15}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right] \tag{26.16}$$

**Note**

Using theorem 26.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{26.17}$$

### 0.4. Law of Cosines

**Law 26.1 Law of Cosines** [proof 26.3]**:** relates the three side of a *general* triangle to each other.

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{b,c} \tag{26.18}$$

**Law 26.2 Law of Cosines for Vectors** [proof 26.4]**:** relates the length of vectors to each other.

$$\|\mathbf{a}\|^2 = \|\mathbf{c} - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2 - 2\|\mathbf{b}\|\|\mathbf{c}\|\cos\theta_{\mathbf{b},\mathbf{c}} \tag{26.19}$$

**Law 26.3 Pythagorean theorem:** special case of **??** for right triangle:

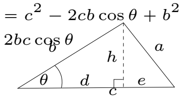
$$a^2 = b^2 + c^2 \tag{26.20}$$

## 1. Proofs

**Proof 26.3.** *Law 26.1 From the defintion of the sine and cosine we know that:*

$$\sin\theta = \frac{h}{b} \Rightarrow \underline{h} \qquad \textit{and} \qquad \cos\theta = \frac{d}{b} \Rightarrow \underline{d}$$

$$\begin{aligned} \underline{e} &= c - \underline{d} = c - b\cos\theta \\ a^2 &= \underline{e}^2 + \underline{h}^2 = c^2 - 2cb\cos\theta + b^2\cos^2\theta + b^2\sin^2\theta \\ &= c^2 + b^2 - 2bc\cos\theta \end{aligned}$$



**Proof 26.4.** *Law 26.2 Notice that* $\mathbf{c} = \mathbf{a} + \mathbf{b} \Rightarrow \mathbf{a} = \mathbf{c} - \mathbf{b}$ *and we can either use* **??** *26.3 or notice that:*

$$\begin{aligned} \|\mathbf{c} - \mathbf{b}\|^2 &= (\mathbf{c} - \mathbf{b}) \cdot (\mathbf{c} - \mathbf{b}) \\ &= \mathbf{c} \cdot \mathbf{c} - 2\mathbf{c} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} \\ &= \|\mathbf{c}\|^2 + \|\mathbf{b}\|^2 - 2\left(\|\mathbf{c}\|\|\mathbf{b}\|\cos\theta\right) \end{aligned}$$

# Topology

# Numerical Methods

## 1. Machine Arithmetic's

### 1.1. Machine Numbers

**Definition 28.1 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 28.2 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 $x\widetilde{\Omega}y = \mathbf{fl}(x\Omega y)$

**Corollary 28.1 Closure:** Machine numbers $\mathbb{F}$ are not *closed*[def. 19.7] under basic arithmetic operations:

$$\mathbb{F}\,\Omega\,\mathbb{F} \mapsto \not\mathbb{F} \qquad \Omega = \{+, -, *, /\} \tag{28.1}$$

**Note**

Corollary 28.1 provides a problem as the computer can only represent floating point number $\mathbb{F}$.

**Definition 28.3 Floating Point Operation** $\widetilde{\Omega}$**:** Is a basic arithmetic operation that obtains a number $x \in \mathbb{F}$ by applying a function rd:

$$\mathbb{F}\,\widetilde{\Omega}\,\mathbb{F} \mapsto \mathbb{F} \qquad \begin{aligned}\widetilde{\Omega} &:= \mathrm{rd} \circ \Omega \\ \Omega &= \{+, -, *, /\}\end{aligned} \tag{28.2}$$

**Definition 28.4 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:

$$\mathrm{rd} : \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg \min\limits_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases} \tag{28.3}$$

**Consequence**

Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

**Axiom 28.1 Axiom of Round off Analysis:** Let $x, y \in \mathbb{F}$ be (normalized) floats and assume that $x\widetilde{\Omega}y \in \mathbb{F}$ (i.e. no over/underflow). Then it holds that:

$$\begin{aligned} x\widetilde{\Omega}y &= (x\Omega y)(1+\delta) & \Omega &= \{+, -, *, /\} \\ \widetilde{f}(x) &= f(x)(1+\delta) & f &\in \{\exp, \sin, \cos, \log, \ldots\} \end{aligned} \tag{28.4}$$

with $|\delta| < \mathrm{EPS}$

**Explanation 28.1** (axiom 28.1)**.** *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 $x\Omega y$.*

Finish

**Definition 28.5 Overflow:** Result is bigger then the biggest representable floating point number.

**Definition 28.6 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. 28.7] 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 28.7 Log sum Exponential:**

$$\mathrm{LogSumExp}(x_1, \ldots, x_n) := \log\left(\sum_{i=1}^{n} \mathrm{e}^{x_i}\right) \tag{28.5}$$

**Formula 28.1 Log-Sum-Exp Trick:**

$$\log\left(\sum_{i=1}^{n} \mathrm{e}^{x_i}\right) = a + \log \sum_{i=1}^{n} \mathrm{e}^{x_i - a} \qquad a := \max_{i \in \{1, \ldots, n\}} x_i \tag{28.6}$$

**Explanation 28.2** (formula 28.1)**.** *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 28.1.**

$$\begin{aligned} LSE = \log\Big(\sum_{i=1}^{n} e^{x_i}\Big) &= \log\Big(\sum_{i=1}^{n} e^{x_i - a} e^{a}\Big) \\ &= \log\Big(e^{a} \sum_{i=1}^{n} e^{x_i - a}\Big) = \log\Big(\sum_{i=1}^{n} e^{x_i - a}\Big) + \log(e^{a}) \\ &= \log\Big(\sum_{i=1}^{n} e^{x_i - a}\Big) + a \end{aligned}$$

**Definition 28.8 Partition** $\Pi$**:** Given an interval $[0, T]$ a sequence of values $0 < t_0 < \cdots < t_n < T$ is called a partition $\Pi(t_0, \ldots, t_n)$ of this interval.

## 2. Convergence

### 2.1. O-Notation

#### 2.1.1. Small $o(\cdot)$ Notation

**Definition 28.9 Little $o$ Notation:**

$$f(n) = o(g(n)) \iff \lim_{n \to \infty} \frac{f(n)}{g(n)} = 0 \tag{28.7}$$

#### 2.1.2. Big $\mathcal{O}(\cdot)$ Notation

### 2.2. Rate Of Convergence

**Definition 28.10 Rate of Convergence:** Is a way to measure the rate of convergence of a sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ to a value to $\mathbf{x}^*$. Let $\rho \in [0, 1]$ be the *rate of convergence* and define:

$$\lim_{k \mapsto \infty} \frac{\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{k} - \mathbf{x}^*\right\|} = \rho \tag{28.8}$$

$$\iff \lim_{k \mapsto \infty} \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \rho \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \qquad \forall k \in \mathbb{N}_0$$

**Definition 28.11 Linear/Exponential Convergence:** A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *linearly* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho \in (0, 1) \qquad \forall k \in \mathbb{N}_0 \tag{28.9}$$

**Definition 28.12 Superlinear Convergence:** A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *superlinear* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 1 \tag{28.10}$$

**Definition 28.13 Sublinear Convergence:** A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *sublinear* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 0 \iff \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| = o\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|\right) \tag{28.11}$$

**Definition 28.14 Logarithmic Convergence:** A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *logarithmically* to $\mathbf{x}^*$ if it converges *sublinear*[def. 28.13] and additoinally satisfies

$$\rho = 0 \iff \left\|\mathbf{x}^{k+2} - \mathbf{x}^{k+1}\right\| = o\left(\left\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\right\|\right) \tag{28.12}$$

add explaintion why

**Exponetial Convergence**

Linear convergence is sometimes called exponential convergence. This is due to the fact that:

1. We often have expressions of the form:
$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \underbrace{(1 - \alpha)}_{:=\rho} \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$$
2. and that $(1 - \alpha) = \exp(-\alpha)$ from which follows that:
$$\text{eq. (28.13)} \iff \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \mathrm{e}^{-\alpha} \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$$

**Definition 28.15 Convergence of order $p$:** In order to distinguish *superlinear convergence* we define the order of convergence.

A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \ldots\}$ to $\mathbf{x}^*$ if it satisfies:

$$\lim_{k \mapsto \infty} \frac{\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^{p}} = C \qquad C < 1 \tag{28.13}$$

Does this even exist/check if this is ture

**Definition 28.16 Exponential Convergence:** A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate $\rho$ to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ it satisfies:

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \rho^{k} \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \qquad \rho < 1 \tag{28.14}$$

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \in o\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|\right) \tag{28.15}$$

## 3. Linear Systems of Equations

### 3.1. Direct Methods

#### 3.1.1. LU-Decomposition

**Definition 28.17 LU Decomposition:**

#### 3.1.2. Symmetric Positive Definite Matrices

For linear systems with s.p.d.[def. 25.71] matrices $\mathbf{A}$ the LU-decomposition[def. 28.17] simplifies to the Cholesky Decomposition[def. 28.18].

**Cholesky Decomposition**

**Definition 28.18 Cholesky Decomposition:** Let $\mathbf{A}$ be a s.p.d.[def. 25.71] then it can be factorized into:

$$\mathbf{A} = \mathbf{G}\mathbf{G}^{\intercal} \qquad \text{with} \qquad \mathbf{G} := \mathbf{L}\mathbf{D}^{1/2} \tag{28.16}$$

### 3.2. Iterative Methods

## 4. Iterative Methods for Non-linear Systems

**Definition 28.19**
**General Non-linear System of Equations (NLSE)** $F$**:** Is a system of non-linear equations $F$ (that do **not** satisfy linearity**??**):

$$F :\subseteq \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \text{seek to find} \qquad \mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{0} \tag{28.17}$$

add linearity under linear algebra

**Definition 28.20 Stationary $m$-point Iteration** $\phi_F$**:** Let $n, m \in \mathbb{R}$ and let $U \subseteq (R^n)^m = \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ be a set. The function $\phi : U \mapsto \mathbb{R}^n$, called ($m$-point) iteration function is an iterative algorithm that produces an iterative sequence $\left(\mathbf{x}^{(k)}\right)_k$ of approximate solutions to eq. (28.17), using the $m$ most recent iterates:

$$\mathbf{x}^{(k)} = \phi_F\left(\mathbf{x}^{(k-1)}, \ldots, \mathbf{x}^{(k-m)}\right) \tag{28.18}$$

Inital Guess $\qquad \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m-1)}$

**Note**

*Stationary* as $\phi$ does no explicitly depend on $k$.

**Definition 28.21 Fixed Point** $\mathbf{x}^*$**:** Is a point $\mathbf{x}^*$ for which the sequence does not change anymore:

$$\mathbf{x}^* = \phi_F\left(\mathbf{x}^{(k-1)}, \ldots, \mathbf{x}^{(k-m)}\right) \quad \text{with} \quad \begin{aligned} \mathbf{x}^{(k-1)} &= \mathbf{x}^* \\ &\vdots \\ \mathbf{x}^{(k-m)} &= \mathbf{x}^* \end{aligned} \tag{28.19}$$

#### 4.0.1. Convergence

**Question**

Does the sequence $\left(\mathbf{x}^{(k)}\right)_k$ converge to a limit:

$$\lim_{k \to \infty} \mathbf{x}^{(k)} = \mathbf{x}^* \tag{28.20}$$

#### 4.0.2. Consistency

**Definition 28.22 Consistent $m$-point Iterative Method:** A *stationary* $m$-point method[def. 28.20] is *consistent* with a non-lineary system of equations[def. 28.19] $F$ iff:

$$F\left(\mathbf{x}^*\right) \iff \phi_F\left(\mathbf{x}^*, \ldots, \mathbf{x}^*\right) = \mathbf{x}^* \tag{28.21}$$

#### 4.0.3. Speed of Convergence

add cvg, consitency, speed of cvng,...

### 4.1. Fixed Point Iterations $m = 1$

**Definition 28.23 Fixed Point Iteration:** Is a 1-point method $\phi_F : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ that seeks a fixed point $\mathbf{x}^*$ to solve $F(\mathbf{x}) = 0$:

$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \qquad \text{Inital Guess: } \mathbf{x}^{(0)} \tag{28.22}$$

**Corollary 28.2 Consistency:** If $\phi_F$ is *continuous* and $\mathbf{x}^* = \lim_{k \to \infty} x^{(k)}$ then $\mathbf{x}^*$ is a fixed point[def. 28.21] of $\phi$.

**Algorithm 28.1 Fixed Point Iteration:**

```
Input: Inital Guess: $\mathbf{x}^{(0)}$
Rewrite $F(\mathbf{x}) = 0$ into a form of $\mathbf{x} = \phi_F(\mathbf{x})$
                                        ▷ There exist many ways
for $k = 1, \ldots, T$ do
    Use the fixed point method:
        $\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right)$        (28.23)
end for
```

add examples and rest

## 5. Numerical Quadrature

**Definition 28.24 Order of a Quadrature Rule:** The order of a quadrature rule $\mathcal{Q}_n : \mathcal{C}^0([a, b]) \to \mathbb{R}$ is defined as:

$$\mathrm{order}(\mathcal{Q}_n) := \max\left\{n \in \mathbb{N}_0 : \mathcal{Q}_n(p) =\in_a^b p(t)\,\mathrm{d}t \quad \forall p \in \mathcal{P}_n\right\} + 1 \tag{28.24}$$

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

**Note**

Is a quality measure for quadrature rules.

### 5.1. Composite Quadrature

**Definition 28.25 Composite Quadrature:** **Given** a mesh $\mathcal{M} = \{a = x_0 < x_1 < \ldots < x_m = b\}$ apply a Q.R. $\mathcal{Q}_n$ to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \ldots, m \triangleq$ p.w. Quadrature:

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{j=1}^{m} \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t = \sum_{j=1}^{m} \mathcal{Q}_n(f_{I_j}) \tag{28.25}$$

**Lemma 28.1 Error of Composite quadrature Rules:**
**Given** a function $f \in \mathcal{C}^{k}([a, b])$ with integration domain:

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

**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:

$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k,q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k,q\}}\right) && \text{for } n \to \infty \\ &\overset{\text{[cor. 22.5]}}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{q}\right) && \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{28.26}$$

**Definition 28.26 Complexity $W$:** Is the number of function evaluations ≙ number of quadrature points.

$$W(\mathcal{Q}(f)_n) = \#\text{f-eval} \triangleq n \tag{28.27}$$

**Lemma 28.2 Error-Complexity $W(\epsilon_n(f))$:** Relates the complexity to the quadrature error.
**Assuming** and quadrature error of the form :

$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \qquad \Longleftrightarrow \qquad \epsilon_n(f) = cn^{-q} \qquad c \in \mathbb{R}_+$$

the error complexity is algebraic (**??**) and is given by:

$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{28.28}$$

**Proof 28.2.** *lemma 28.2:* **Assume***: we want to reduce the error by a factor of $\epsilon_n$ by increasing the number of quadrature points $n_{new} = a \cdot n_{old}$.*
**Question***: what is the additional effort (#f-eval) needed in order to achieve this reduction in error?*

$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \qquad \Rightarrow \qquad n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{28.29}$$

### 5.1.1. Simpson Integration

**Definition 28.27 Simpson Integration:**

# Optimization

**Definition 29.1 Fist Order Method:** A first-order method is an algorithm that chooses the $k$-th iterate in

$$\mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \dots \nabla f(\mathbf{x}_{k-1})\} \qquad \forall k = 1, 2, \dots \tag{29.1}$$

**Note**

Gradient descent is a first order method

## 1. Linear Optimization

### 1.1. Polyhedra

**Definition 29.2 Polyhedron:** Is a set $P \in \mathbb{R}^n$ that can be described by the *finite* intersection of $m$ closed *half spaces*??:

$$P = \{\mathbf{x} \in \mathbb{R}^n \,|\mathbf{A}\mathbf{x} \leqslant \mathbf{b}\} = \{\mathbf{x} \in \mathbb{R}^n \,|\mathbf{a}_j\mathbf{x} \leqslant b_j, j = 1, \dots, m\}$$

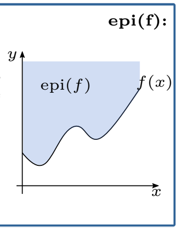
$$\mathbf{A} \in \mathbb{R}^{m \times n} \qquad \mathbf{b} \in \mathbb{R}^m \tag{29.2}$$

#### 1.1.1. Polyhedral Function

**Definition 29.3 Epigraph/Subgraph** **epi(f):**

The epigraph of a function $f \in \mathbb{R}^n \mapsto \mathbb{R}$ is defined as the set of point that lie above its gaph:

$$\text{epi}(f) := \{(\mathbf{x}, y) \in \mathbb{R}^n \,|y \geqslant f(\mathbf{x})\} \subseteq \mathbb{R}^{n+1} \tag{29.3}$$



**Definition 29.4 Polyhedral Function:** A function $f$ is *polyhedral* if its epigraph epi$(f)$[def. 29.3] is a polyhedral set[def. 29.2]:

$$f \text{ is polyhedral} \iff \text{epi}(f) \text{ is polyhedral} \tag{29.4}$$

## 2. Lagrangian Optimization Theory

**Add:** derivation of lagragne function

**Definition 29.5 (Primal) Constraint Optimization:** **Given** an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:

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

**Definition 29.6 Lagrange Function:**

$$\mathscr{L}(\alpha, \beta, \mathbf{w}) := f(\mathbf{w}) + \alpha\mathbf{g}(\mathbf{w}) + \beta\mathbf{h}(\mathbf{w}) \tag{29.5}$$

**Extremal Conditions**

$\nabla \mathscr{L}(\mathbf{x}) \overset{!}{=} 0$ Extremal point $\mathbf{x}^*$

$\frac{\partial}{\partial\beta}\mathscr{L}(\mathbf{x}) = h(\mathbf{x}) \overset{!}{=} 0$ Constraint satifisfaction

For the inequality constraints $g(\mathbf{x}) \leqslant 0$ we distinguish two situations:

Case I : $g(\mathbf{x}^*) < 0$ switch const. off

Case II : $g(\mathbf{x}^*) \geqslant 0$ optimze using active eq. constr.

$\frac{\partial}{\partial\alpha}\mathscr{L}(\mathbf{x}) = g(\mathbf{x}) \overset{!}{=} 0$ Constraint satifisfaction

**Definition 29.7 Lagrangian Dual Problem:** Is given by:

$$\begin{aligned} \text{Find} \quad &\max \theta(\alpha, \beta) = \inf_{\mathbf{w}\in\Omega} \mathscr{L}(\mathbf{w}, \alpha, \beta) \\ \textbf{s.t.} \quad &\alpha_i \geqslant 0 \qquad 1 \leqslant i \leqslant k \end{aligned}$$

**Solution Strategy**

1. Find the extremal point $\mathbf{w}^*$ of $\mathscr{L}(\mathbf{w}, \alpha, \beta)$:
$$\left.\frac{\partial \mathscr{L}}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^*} \overset{!}{=} 0 \tag{29.6}$$
2. Insert $\mathbf{w}^*$ into $\mathscr{L}$ and find the extremal point $\beta^*$ of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left.\frac{\partial \theta}{\partial \beta}\right|_{\beta=\beta^*} \overset{!}{=} 0 \tag{29.7}$$
3. Calculate the solution $\mathbf{w}^*(\beta^*)$ of the constraint minimization problem.

**Value of the Problem**

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

**Theorem 29.1 Upper Bound Dual Cost:** Let $\mathbf{w} \in \Omega$ be a feasible solution of the primal problem [def. 29.5] and $(\alpha, \beta)$ a feasible solution of the respective dual problem [def. 29.7]. Then it holds that:

$$f(\mathbf{w}) \geqslant \theta(\alpha, \beta) \tag{29.8}$$

**Proof 29.1.**

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

**Corollary 29.1 Duality Gap Corollary:** The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup\{\theta(\alpha, \beta) : \alpha \geqslant 0\} \leqslant \inf\{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leqslant 0, \mathbf{h}(\mathbf{w}) = 0\} \tag{29.9}$$

**Theorem 29.2 Optimality:** The triple $(\mathbf{w}^*, \alpha^*, \beta^*)$ 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:

$$f(\mathbf{w}^*) = \theta(\alpha^*, \beta^*) \tag{29.10}$$

**Definition 29.8 Convex Optimization:** **Given**: a convex function f and a convex set S solve:

$$\begin{aligned} &\min f(\mathbf{x}) \\ \text{s.t.} \quad &\mathbf{x} \in S \end{aligned} \tag{29.11}$$

Often S is specified using linear inequalities:

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

**Theorem 29.3 Strong Duality:** Given an convex optimization problem:

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

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

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

**Theorem 29.4 Kuhn-Tucker Conditions:** Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

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

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 $\alpha^*, \beta^*$ s.t.:

$$\frac{\partial \mathscr{L}(\mathbf{w}, \alpha, \beta)}{\partial \mathbf{w}} \overset{!}{=} 0 \qquad \frac{\partial \mathscr{L}(\mathbf{w}^*, \alpha, \beta)}{\partial \beta} \overset{!}{=} 0 \tag{29.12}$$

under the condtions that:

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

**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.

# Combinatorics

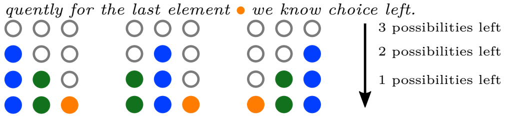
## 1. Permutations

**Definition 30.1 Permutation:** A $n$-Permutation is the (re)*arrangement* of $n$ elements of a set[def. 19.1] $\mathcal{S}$ of size $n = |\mathcal{S}|$ into a sequences[def. 20.2].

**Definition 30.2 Number of Permutations of a Set $n!$:**
Let $\mathcal{S}$ be a set[def. 19.1] $n = |\mathcal{S}|$ *distinct* objects. The number of permutations of $\mathcal{S}$ is given by:

$$P_n(\mathcal{S}) = n! = \prod_{i=0}^{n-1}(n-i) = n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot 1 \tag{30.1}$$

**Explanation 30.1.** *If we have i.e. three distinct elements {●, ●, ●} For the first element ● that we arrange we have three possible choices where to put it. However this reduces the number of possible choices for the second element ● to only two. Consequently for the last element ● we know choice left.*



**Definition 30.3**
**Number of Permutations of a Multiset:**
Let $\mathcal{S}$ be a multi set[def. 19.3] with $n = |\mathcal{S}|$ total and $k$ *distinct* objects. Let $n_j$ be the multiplicity[def. 19.4] of the member $j \in \{1,\ldots,k\}$ of the multiset $\mathcal{S}$. The permutation of $\mathcal{S}$ is given by:

$$P_{n_1,\ldots,n_k}(\mathcal{S}) = \frac{n!}{n_1!\cdot\ldots\cdot n_k} \qquad \text{s.t.} \quad \sum_{j=1}^{k} n_j \leqslant n \qquad k < n \tag{30.2}$$

**Note**

We need to divide by the permutations as sequence/order does not change if we exchange objects of the same kind (e.g. red ball by red ball) ⇒ less possibilities to arrange the elements uniquely.

## 2. Combinations

**Definition 30.4 $k$-Combination:**
A $k$-combination of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is a subset $\mathcal{S}_k$ (order does not matter) of $k = |\mathcal{S}_k|$ *distinct* elements, *chosen* from $\mathcal{S}$.

**Definition 30.5 Number of $k$-Combinations $C_{n,k}$:**
The number of $k$-combinations of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is given by:

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{30.3}$$

## 3. Variation

**Definition 30.6 Variation:**
A $k$-variation of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is

1. a selection/combination[def. 30.4] of a subset $\mathcal{S}_k$ (order does not matter) of $k$-*distinct* elements $k = |\mathcal{S}_k|$, *chosen* from $\mathcal{S}$
2. and an $k$ arrangement/permutation[def. 30.2] of that subset $\mathcal{S}_k$ (with or without repetition) into a sequence[def. 20.2]

**Definition 30.7**
**Number of Variations without repetitions $V_k^n$:**
Let $\mathcal{S}$ be a set[def. 19.1] $n = |\mathcal{S}|$ *distinct* objects from which we choose $k$ elements. The number of variations of size $k = |\mathcal{S}_k|$ of the set $\mathcal{S}$ *without repetitions* is given by:

$$V_k^n(\mathcal{S}) = \binom{n}{k}k! = \frac{n!}{(n-k)!} \tag{30.4}$$

**Note**

Sometimes also denotes as $P_k^n$.

**Definition 30.8**
**Number of Variations with repetitions $\bar{V}_k^n$:**
Let $\mathcal{S}$ be a set[def. 19.1] $n = |\mathcal{S}|$ *distinct* objects from which we choose $k$ elements. The number of variations of size $k = |\mathcal{S}_k|$ of the set $\mathcal{S}$ from which we *choose and always return* is given by:

$$\bar{V}_k^n(\mathcal{S}) = n^k \tag{30.5}$$

»»»> d641cc7fe8996718a0ca20ca61519f18af078387

# Stochastics

**Definition 30.9 Stochastics:** Is a collective term for the areas of *probability theory* and *statistics*.

**Definition 30.10 Statistics:** Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.

**Definition 30.11 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 30.12 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.

Improve those definitions: maybe ask on quorahho

**Note: Stochastics vs. Stochastic**

Stochasticss is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is a *adjective*, describing that a certain phenomena is governed by uncertainty i.e. a process.

# Probability Theory

**Definition 31.1 Probability Space** $W = \{\Omega, \mathcal{F}, \mathbb{P}\}$**:**
Is the unique triple $\{\Omega, \mathcal{F}, \mathbb{P}\}$, where $\Omega$ is its sample space, $\mathcal{F}$ its $\sigma$-algebra of events, and $\mathbb{P}$ its probability measure.

**Definition 31.2** [example 31.1]
**Sample Space** $\Omega$**:**
Is the set of all possible outcomes (elementary events [cor. 31.5]) of an experiment.

**Definition 31.3** [example 31.2]
**Event** $A$**:**
An "event" is a subset of the sample space $\Omega$ and is a property which can be observed to hold or not to hold *after* the experiment is done.
Mathematically speaking not every subset of $\Omega$ is an event and has an associated probability.
Only those subsets of $\Omega$ that are part of the corresponding $\sigma$-algebra $\mathcal{F}$ are events and have their assigned probability.

**Corollary 31.1 :** If the outcome $\omega$ of an experiment is in the subset $A$, then the event $A$ is said to "have occured".

**Corollary 31.2 Complement Set** $A^{\mathrm{C}}$**:**
is the contrary event of $A$.

**Corollary 31.3 The Union Set** $A \cup B$**:**
Let $A$, $B$ be two events. The event "$A$ or $B$" is interpreted as the union of both.

**Corollary 31.4 The Intersection Set** $A \cap B$**:**
Let $A$, $B$ be two events. The event "$A$ and $B$" is interpreted as the intersection of both.

**Corollary 31.5 The Elementary Event** $\omega$**:**
Is a "singleton", i.e. a subset $\{\omega\}$ containing a single outcome $\omega$ of $\Omega$.

**Corollary 31.6 The Sure Event** $\Omega$**:**
Is equal to the sample space as it contains all possible elementary events.

**Corollary 31.7 The Impossible Event** $\varnothing$**:**
The impossible event i.e. nothing is happening is denoted by the empty set.

**Definition 31.4 The Family of All Events** $\mathcal{A}/2^{\Omega}$**:**
The set of all subset of the sample space $\Omega$ called family of all events is given by the power set of the sample space $\mathcal{A} = 2^{\Omega}$ (for finite sample spaces).

**Definition 31.5 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. $0 \leqslant \mathbb{P}(A) \leqslant 1$
2. $\mathbb{P}(\Omega) = 1$
3. If $A \cap B = \varnothing$ 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:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{\delta_n(A)}{n} \qquad \text{where} \qquad \delta(A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$$

## 0.1. Sigma Algebras

**Definition 31.6** [Proof 31.3]
**Sigma Algebra** $\sigma$**:**
A set $\mathcal{F}$ of subsets of $\Omega$ is called a $\sigma$-algebra on $\Omega$ if the following properties apply

- $\Omega \in \mathcal{F}$ and $\varnothing \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $\Omega \backslash A = A^{\mathrm{C}} \in \mathcal{F}$:
  The complementary subset of A is also in $\Omega$.
- For all $A_i \in \mathcal{F} : \bigcup_{i=1} A_i \in \mathcal{F}$

**Explanation 31.1** ([def. 31.6]). *The $\sigma$-algebra determines what events we can measure, it represents all of the possible events of the experiment that we can detect.*
*Thus the sigma algebra is a mathematical construct that tells us how much information we obtain once we conduct some experiment.*

**Corollary 31.8** $\mathcal{F}_{\min}$**:** $\mathcal{F} = \{\varnothing, \Omega\}$ is the simplest $\sigma$-algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.

**Corollary 31.9** $\mathcal{F}_{\max}$**:** $\mathcal{F} = 2^{\Omega}$ consists of all subsets of $\Omega$ and thus corresponds to full information i.e. we know if and which event happened.

**Definition 31.7 Measurable Space** $\{\Omega, \underline{\mathcal{F}}\}$**:**
Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.

**Corollary 31.10 $\mathcal{F}$-measurable Event** $A_i \in \mathcal{F}$**:**
The measurable events $A_i$ of $\mathcal{F}$ are called *$\mathcal{F}$-measurable* or *measurable sets*.

**Definition 31.8** [Example 31.4]
**Sigma Algebra generated by a subset of $\Omega$** $\sigma(\mathcal{C})$**:**
Let $\mathcal{C}$ be a class of subsets of $\Omega$. The $\sigma$-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the *smallest* sigma algebra $\mathcal{F}$ that included all elements of $\mathcal{C}$.

**Definition 31.9** [Example 31.5]
**Borel $\sigma$-algebra** $\mathcal{B}(\mathbb{R})$**:**
The Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R})$ is the smallest $\sigma$-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.

**Why do we need Borel Sets**

So far we only looked at atomic events $\omega$, with the help of sigma algebras we are now able to measure continuous events s.a. $[0, 1]$.

**Corollary 31.11 Generating Borel $\sigma$-Algebra**[Proof 31.1]**:**
The Borel $\sigma$-algebra of $\mathbb{R}$ is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ ($\mathbb{Q}$ =rationals).

**Definition 31.10 ($\mathbb{P}$)-trivial Sigma Algebra:**
is a $\sigma$-algebra $\mathcal{F}$ for which each event has a probability of zero or one:

$$\mathbb{P}(A) \in \{0, 1\} \qquad \forall A \in \mathcal{F} \tag{31.1}$$

**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 $\mathcal{F}_{\min} = \{\Omega, \varnothing\}$.

## 0.2. Measures

**Definition 31.11 Measure** $\mu$**:**
A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map:

$$\mu : \mathcal{F} \mapsto [0, \infty] \tag{31.2}$$

for which holds:

- $\mu(\varnothing) = 0$
- countable additivity [def. 31.12]

**Definition 31.12 Countable/$\sigma$-Additive Function:**
Given a function $\mu$ defined on a $\sigma$-algebra $\mathcal{F}$.
The function $\mu$ is said to be countable additive if for every countable sequence of pairwise disjoint elements $(F_i)_{i\geqslant 1}$ of $\mathcal{F}$ it holds that:

$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \varnothing \quad \forall j \neq k \tag{31.3}$$

**Corollary 31.12 Additive Function:** A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds:

$$F \cap G = \varnothing \quad \Longrightarrow \quad \mu(F \cup G) = \mu(F) + \mu(G) \tag{31.4}$$

**Explanation 31.2.** *If we take two events that cannot occur simultaneously, then the probability that at least one of the events occurs is just the sum of the measures (probabilities) of the original events.*

**Definition 31.13** [Example 31.6]
**Equivalent Measures** $\mu \sim \nu$**:**
Let $\mu$ and $\nu$ be two measures defined on a measurable space[def. 31.7] $(\Omega, \mathcal{F})$. The two measures are said to be equivalent if it holds that:

$$\mu(A) > 0 \iff \nu(A) > 0 \qquad \forall A \subseteq \mathcal{F} \tag{31.5}$$

this is equivalent to $\mu$ and $\nu$ having equivalent null sets:

$$\mathcal{N}_\mu = \mathcal{N}_\nu \qquad \begin{aligned} \mathcal{N}_\mu &= \{A \in \mathcal{A} | \mu(A) = 0\} \\ \mathcal{N}_\nu &= \{A \in \mathcal{A} | \nu(A) = 0\} \end{aligned} \tag{31.6}$$

**Definition 31.14 Measure Space** $\{\mathcal{F}, \Omega, \underline{\mu}\}$**:**
The triplet of sample space, sigma algebra and a measure is called a measure space.

**Definition 31.15 Lebesgue Measure on $\mathcal{B}$** $\lambda$**:**
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:

$$\lambda([a, b]) := b - a \tag{31.7}$$

**Corollary 31.13 Lebesgue Measure of Atomitcs:**

- The Lebesgue measure of a set containing only one point must be zero:
  $$\lambda(\{a\}) = 0 \tag{31.8}$$
- The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2 \ldots, a_n\}$ must be zero:
  $$\lambda(A) + \sum_{i=1}^{n} \lambda(\{a_i\}) = 0 \tag{31.9}$$
- The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2 \ldots, \}$ 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 $\mu$, by standardizing the measure we obtain a well defined measure of events.

**Axiom 31.1 Non-negativity:** The probability of an event is a non-negative real number:

$$\text{If } A \in \mathcal{F} \qquad \text{then} \qquad \mathbb{P}(A) \geqslant 0 \tag{31.10}$$

**Axiom 31.2 Unitairity:** The probability that at least one of the elementary events in the entire sample space $\Omega$ will occur is equal to one:

$$\text{The certain event} \qquad \mathbb{P}(\Omega) = 1 \tag{31.11}$$

**Axiom 31.3 $\sigma$-additivity:** If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are mutually disjoint, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{31.12}$$

**Corollary 31.14 :** As a consequence of this it follows:

$$\mathbb{P}(\varnothing) = 0 \tag{31.13}$$

**Corollary 31.15 Complementary Probability:**

$$\mathbb{P}(A^{\mathrm{C}}) = 1 - \mathbb{P}(A) \qquad \text{with} \qquad A^{\mathrm{C}} = \Omega - A \tag{31.14}$$

**Definition 31.16 Probability Measure** $\mathbb{P}$**:**
a probability measure is function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ defined on a $\sigma$-algebra $\mathcal{F}$ of a sample space $\Omega$ that satisfies the probability axioms.

## 1. Conditional Probability

**Definition 31.17 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:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) \neq 0 \tag{31.15}$$

## 2. Independent Events

**Theorem 31.1**
**Independent Events:** Let $A$, $B$ be two events. $A$ and $B$ are said to be independent iffy:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \qquad \begin{aligned} \mathbb{P}(A|B) &= \mathbb{P}(A), \quad \mathbb{P}(B) > 0 \\ \mathbb{P}(B|A) &= \mathbb{P}(B), \quad \mathbb{P}(A) > 0 \end{aligned} \tag{31.16}$$

**Note**

The requirement of no impossible events follows from [def. 31.17]

**Corollary 31.16 Pairwise Independent Evenest:**
A finite set of events $\{A_i\}_{i=1}^{n} \in \mathcal{A}$ is *pairwise independent* if every pair of events is independent:

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cap \mathbb{P}(A_j) \quad i \neq j, \quad \forall i, j \in \mathcal{A} \tag{31.17}$$

**Corollary 31.17 Mutal Independent Evenest:**
A finite set of events $\{A_i\}_{i=1}^{n} \in \mathcal{A}$ is *mutal independent* if every event $A_j$ is independent of any intersection of the other events:

$$\mathbb{P}\left(\bigcap_{i=i}^{k} B_i\right) = \prod_{i=1}^{k} \mathbb{P}(B_i) \qquad \begin{aligned} &\forall \{B_i\}_{i=1}^{k} \subseteq \{A_i\}_{i=1}^{n} \\ &k \leqslant n, \qquad \{A_i\}_{i=1}^{n} \in \mathcal{A} \end{aligned} \tag{31.18}$$

## 3. Product Rule

**Law 31.1 Product Rule:** Let $A$, $B$ be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{31.19}$$

**Law 31.2**
**Generalized Product Rule/Chain Rule:** is the generalization of the product rule?? to $n$ events $\{A_i\}_{i=1}^{n}$

$$\mathbb{P}\left(\bigcap_{i=i}^{k} E_i\right) = \prod_{k=1}^{n} \mathbb{P}\left(E_k \middle| \bigcap_{i=i}^{k-1} E_i\right) = \tag{31.20}$$

$$= \mathbb{P}(E_n|E_{n-1} \cap \ldots \cap E_1) \cdot \mathbb{P}(E_{n-1}|E_{n-2} \cap \ldots \cap E_1) \cdots$$
$$\cdots \mathbb{P}(E_3|E_2 \cap E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)$$

## 4. Law of Total Probability

**Definition 31.18 Complete Event Field:** A complete event field $\{A_i : i \in I \subseteq \mathbb{N}\}$ is a countable or finite partition of $\Omega$ that is the partitions $\{A_i : i \in I \subseteq \mathbb{N}\}$ are a *disjoint union* of the sample space:

$$\bigcup_{i\in I} A_i = \Omega \qquad A_i \cap A_j = \varnothing \qquad i \neq j, \forall i, j \in I \tag{31.21}$$

**Theorem 31.2**
**Law of Total Probability/Partition Equation:**
Let $\{A_i : i \in I\}$ be a complete event field[def. 31.18] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{31.22}$$

## 5. Bayes Theorem

**Law 31.3 Bayes Rule:** Let $A, B$ be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) > 0 \tag{31.23}$$

follows directly fromeq. (31.19).

**Theorem 31.3 Bayes Theorem:** Let $\{A_i : i \in I\}$ be a complete event field[def. 31.18] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i\in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{31.24}$$

proof **??** 31.2

# Distributions on $\mathbb{R}$

### 6.1. Distribution Function

**Definition 31.19 Distribution Function of $\mathbb{P}$** $F$**:** The *distribution function* $F$ induced by a a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((\infty, x]) \tag{31.25}$$

**Theorem 31.4 :** 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
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$

**Corollary 31.18 :** 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 function/map that determines a quantity of interest based on the outcome $\omega \in \Omega$ of a random experiment. Thus $X$ is not really a variable in the classical sense but a variable with respect to the outcome of an experiment. Its value is determined in two steps:

① The outcome of an experiment is a random quantity $\omega \in \Omega$

② The outcome $\omega$ 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 $\Omega$ into another space $\mathcal{E}$, usually $\mathcal{E} = \mathbb{R}$ or $\mathcal{E} = \mathbb{R}^n$:

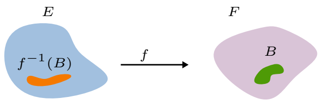
$$X : \Omega \mapsto \mathcal{E} \qquad \omega \mapsto X(\omega)$$

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 $\Omega$:

$$\underbrace{\mathbb{P}_X(E)}_{\text{Probability for an event in } E} = \mathbb{P}(\{\omega : X(\omega) \in E\}) = \mathbb{P}(X \in E) = \mathbb{P}\overbrace{\left(X^{-1}(E)\right)}^{\text{Probability for an event in } \Omega}$$

**Definition 31.20 $\mathcal{E}$-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

$$\forall B \in \mathcal{F} : \qquad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \tag{31.26}$$



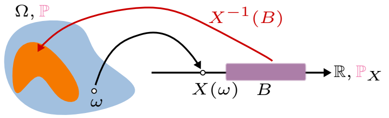
**Interpretation**

The pre-image[def. 22.13] 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 31.21 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 $\Omega$ into a target space $(F, \mathcal{F})$:

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \tag{31.27}$$

Since $X$ is $\mathcal{E}$-measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



**Corollary 31.19 :** Usually $F = \mathbb{R}$, which usually amounts to using the Borel $\sigma$-algebra $\mathcal{B}$ of $\mathbb{R}$.

**Corollary 31.20 Random Variables of Borel Sets:** Given that we work with Borel $\sigma$-algebras then the definition of a random variable is equivalent to (due to [cor. 31.11]):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \tag{31.28}$$

**Definition 31.22**
**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 31.21 Indicator Functions** $I_A(\omega)$**:** An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{31.29}$$

We know that a probability measure $\mathbb{P}$ on $\mathbb{R}$ is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

**Corollary 31.22 :** Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let $(E, \mathcal{E})$ be an arbitrary measurable space. Let $X$ be a real value function on $E$.
Then it holds that $X$ is measurable *if and only if*

$$\{X \leqslant a\} = \{\omega : X(\omega) \leqslant a\} = X^{-1}((-\infty, a]) \in \mathcal{E}, \forall a \in \mathbb{R}$$

or $\{X<a\} \in \mathcal{E}$.

**Explanation 31.3** ([cor. 31.22])**.** *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 31.23 Law/Distribution of X** $\mathscr{L}(X)$**:** Let $X$ be a r.v. on $\{\Omega, \mathcal{F}, \mathbb{P}\}$, with values in $\underline{(E, \mathcal{E})}$, then the *distribution/law* of $X$ is defined as:

$$\begin{aligned} &\mathbb{P} : \mathcal{B} \mapsto [0, 1] \\ &\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \qquad \forall b \in \mathcal{E} \end{aligned} \tag{31.30}$$

**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 31.5 :** The law/distribution of $X$ is a probability measure $\mathbb{P}$ on $\underline{(E, \mathcal{E})}$.

**Definition 31.24**
**(Cumulative) Distribution Function** $F_X$**:**
Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leqslant x) \tag{31.31}$$

**Corollary 31.23 :** The distribution of $\mathbb{P}^X$ of a real valued r.v. is entirely characterized by its cumulative distribution function $F_X$[def. 31.31].

**Property 31.1:**

$$\mathbb{P}(X > x) = 1 - F_X(x) \tag{31.32}$$

**Property 31.2:** Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leqslant B) = F_X(b) - F_X(a) \tag{31.33}$$

### 6.4. Probability Density Function

**Definition 31.25 Continuous Random Variable:** Is a r.v. for which a probability density function $f_X$ exists.

**Definition 31.26 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:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \tag{31.34}$$

or alternatively:

$$f_X(x) = \lim_{\epsilon\to 0} \frac{\mathbb{P}(x \leqslant X \leqslant x + \epsilon)}{\epsilon} \tag{31.35}$$

**Corollary 31.24** $\mathbb{P}(X = b) = 0, \qquad \forall b \in \mathbb{R}$**:**

$$\mathbb{P}(X = b) = \lim_{a\to b} \mathbb{P}(a < X \leqslant b) = \lim_{a\to b} \int_a^b f(x) = 0 \tag{31.36}$$

**Corollary 31.25 :** From [cor. 31.24] it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leqslant X < b) \\ &= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X <\leqslant b) \end{aligned}$$

**Corollary 31.26 :**

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{31.37}$$

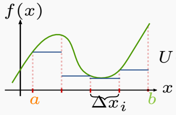
**Notes**

- 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**

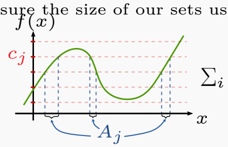
- Difficult to extend to higher dimensions – general domains of definitions $f : \Omega \mapsto \mathbb{R}$
- Depends on continuity
- Integration of limit processes require strong uniform convergence in order to integrate limit processes
  $$\lim_{n\to\infty} \int f(x)\,\mathrm{d}x \overset{\text{str. u.c.}}{=} \int \lim_{n\to\infty} f(x)\,\mathrm{d}x$$



$$U(p) = \textstyle\sum_i \sup(f(x_i)) \cdot \Delta x_i \xrightarrow{n\to\infty} \int f\,\mathrm{d}x$$

**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 $\mu$ 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.



$$\textstyle\sum_i c_i \cdot \mu(A_i) \xrightarrow{n\to\infty} \int f\,\mathrm{d}\mu$$

**Definition 31.27 Lebesgue Integral:**

$$\lim_{n\to\infty} \sum_{i=1}^{n} c_i \mu(A_i) = \int_\Omega f\,\mathrm{d}\mu \qquad \begin{array}{l} f(x) \approx c_i \\ \forall x \in A_i \end{array} \tag{31.38}$$

**Definition 31.28**
**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:

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \qquad a_i \in \mathbb{R} \qquad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \tag{31.39}$$

### 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 31.1.
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 31.29 Independent Random Variables:** Two real valued random variables $X$ and $Y$ are said to be independent iff:

$$\mathbb{P}(X \leqslant x|Y \leqslant y) = \mathbb{P}(X \leqslant x) \qquad \forall x, y \in \mathbb{R} \tag{31.40}$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leqslant x\} \cap \{Y \leqslant y\}) = \mathbb{P}(X \leqslant x, Y \leqslant y) \\ &= F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \tag{31.41}$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad \forall A, B \in \mathcal{B} \tag{31.42}$$

**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 31.30**
**Independent Identically Distributed:**

## 7. Product Rule

**Law 31.4 Product Rule:** Let $X, Y$ be two random variables then their jo

**Law 31.5**
**Generalized Product Rule/Chain Rule:**

## 8. Change Of Variables Formula

**Formula 31.1**
**(Scalar Discret) Change of Variables: Let** $X$ be a discret rv $X \in \mathcal{X}$ with pmf $p_X$ 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. 22.16]) 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:

$$p_Y(y) = \sum_{x_i\in\mathcal{X}_y} p_X(x_i) = \sum_{x\in\mathcal{Y}:g(x)=y} p_X(x) \tag{31.43}$$

see proof **??** 31.3

**Formula 31.2**
**(Scalar Continuous) Change of Variables:**
Let $X \sim f_X$ be a continuous r.v. and let $g$ be an arbitrary strictly monotonic[def. 22.16] function.
Define a new r.v. $Y$ as

$$\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\} \tag{31.44}$$

then the pdf of $Y$ is given by:

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\right| \tag{31.45}$$

$$= f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|} \tag{31.46}$$

**Formula 31.3**
**(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. 22.16] function

$$g : \mathbb{R}^n \mapsto \mathbb{R}^m$$

Define a new r.v. $Y$ as

$$\mathcal{Y} = \{\mathbf{y} | \mathbf{y} = g(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}\} \tag{31.47}$$

and let $h(\mathbf{x}) := g(\mathbf{x})^{-1}$ then the pdf of $Y$ is given by:

$$\begin{aligned} f_Y(\mathbf{y}) &= f_X(x_1, \dots, x_n) \cdot |J| \\ &= f_X(h_1(\mathbf{y}), \dots, h_n(\mathbf{y})) \cdot |J| \\ &= f_X(\mathbf{y})|\det D_{\mathbf{x}} h(\mathbf{x})|\Big|_{\mathbf{x}=\mathbf{y}} \\ &= f_X(g^{-1}(\mathbf{y}))\left|\det\left(\frac{\partial g}{\partial \mathbf{x}}\right)\right|^{-1} \end{aligned} \tag{31.48}$$

where $J = \det Dh$ is the Jaccobian[def. 23.6].
See also proof **??** 31.6 and example 31.8

**Note**

A monotonic function is required in order to satisfy inevitability.

# Probability Distributions on $\mathbb{R}^n$

## 10. Joint Distribution

**Definition 31.31**
**Joint (Cumulative) Distribution Function** $F_{\mathbf{X}}$**:**
Let $\mathbf{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$, then its *cumulative distribution function* is defined as:

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}^X((-\infty, \mathbf{x}]) = \mathbb{P}(\mathbf{X} \leqslant \mathbf{x}) \\ &= \mathbb{P}(X_1 \leqslant x_1, \dots X_n \leqslant x_n) \end{aligned} \tag{31.49}$$

**Definition 31.32 Joint Probability Distribution:**
Let $\mathbf{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$ with associated cdf $F_{\mathbf{X}}$. If $F_{\mathbf{X}}$ is continuously integrable for all $\mathbf{x} \in \mathbb{R}$ then $\mathbf{X}$ has a *probability density* $f_X$ defined by:

$$F_X(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\mathbf{X}}(y_1, \dots, y_n)\,\mathrm{d}y_1\,\mathrm{d}y_n \tag{31.50}$$

or alternatively:

$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + \epsilon, \dots, x_n \leqslant X_n \leqslant x_n + \epsilon)}{\epsilon} \tag{31.51}$$

### 10.1. Marginal Distribution

**Definition 31.33 Marginal Distribution:**

## 11. The Expectation

**Definition 31.34 Expectation:**

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int_\Omega X\,\mathrm{d}\mathbb{P} \tag{31.52}$$

**Corollary 31.27 Expectation of simple r.v.:**
If $X$ is a simple[def. 31.28] r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \tag{31.53}$$

### 11.1. Properties

#### 11.1.1. Linear Operators

add

#### 11.1.2. Quadratic Form

**Definition 31.35** proof 31.7
**Expectation of a Quadratic Form:**
Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be a random vector with $\mathbb{E}[\boldsymbol{\epsilon}] = \mu$ and $\mathbb{V}[\boldsymbol{\epsilon}] = \Sigma$:

$$\mathbb{E}\left[\boldsymbol{\epsilon}^\intercal \mathbf{A} \boldsymbol{\epsilon}\right] = \operatorname{tr}(\mathbf{A}\Sigma) + \boldsymbol{\mu}^\intercal \mathbf{A} \boldsymbol{\mu} \tag{31.54}$$

### 11.2. The Jensen Inequality

**Theorem 31.6 Jensen Inequality:** Let $X$ be a random variable and $g$ some function, then it holds:

$$\begin{matrix} g(\mathbb{E}[X]) \geqslant \mathbb{E}[g(X)] & \multirow{2}{*}{if} & g \text{ is convex}^{[\text{def. } 22.25]} \\ g(\mathbb{E}[X]) \leqslant \mathbb{E}[g(X)] & & g \text{ is concave}^{[\text{def. } 22.26]} \end{matrix} \tag{31.55}$$

### 11.3. Law of the Unconscious Statistician

**Law 31.6 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 $X$:

$$\mathbb{E}_Y[y] = \mathbb{E}_X[g(x)] \tag{31.56}$$

**Consequence**

Hence if we $\mathrm{p}_X$ we do not have to first calculate $\mathrm{p}_Y$ in order to calculate $\mathbb{E}_Y[y]$.

### 11.4. Properties

### 11.5. Law of Iterated Expectation (LIE)

**Law 31.7** [proof 31.8]
**Law of Iterated Expectation (LIE):**

$$\mathbb{E}[X] = \mathbb{E}_Y \mathbb{E}[X|Y] \tag{31.57}$$

### 11.6. Hoeffdings Bound

**Definition 31.36 Hoeffdings Bound:**
Let $\mathbf{X} = \{X_i\}_{i=1}^n$ be i.i.d. random variables strictly bounded by the interval $[a, b]$ then it holds:

$$\mathbb{P}(|\mu_{\mathbf{X}} - \mathbb{E}[X]| \geqslant \epsilon) \leqslant 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \overset{[0,1]}{=} 2\mathrm{e}^{-2n\epsilon^2} \tag{31.58}$$

**Explanation 31.4.** *The difference of the expectation from the empirical average to be bigger than $\epsilon$ is upper bound in probability.*

## 12. Moment Generating Function (MGF)

**Definition 31.37 Moment of Random Variable:** The $i$-th moment of a random variable $X$ is defined as (if it exists):

$$m_i := \mathbb{E}\left[X^i\right] \tag{31.59}$$

**Definition 31.38** $\psi_X$
**Moment Generating Function (MGF):**

$$\psi_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \qquad t \in \mathbb{R} \tag{31.60}$$

**Corollary 31.28 Sum of MGF:** The moment generating function of a sum of $n$ independent variables $(X_j)_{1\leqslant j\leqslant n}$ is the product of the moment generating functions of the components:

$$\psi_{S_n}(t) = \psi_{X_1}(t)\cdots\psi_{X_n}(t) \qquad S_n := X_1 + \dots X_n \tag{31.61}$$

**Corollary 31.29 :** The $i$-th moment of a random variable is the $i$-th derivative of its associated moment generating function evaluated zero:

$$\mathbb{E}\left[X^i\right] = \psi_X^{(i)}(0) \tag{31.62}$$

## 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 31.39** $\hat{\mu}$
**Fourier Transformed Probability Measure:**

$$\hat{\mu} = \int \mathrm{e}^{i\langle u,x\rangle}\mu(\mathrm{d}x) \tag{31.63}$$

**Corollary 31.30 :** As $e^{i\langle u,x\rangle}$ can be rewritten using formulaeqs. (19.7) and (19.8) it follows:

$$\hat{\mu} = \int \cos(\langle u, x\rangle)\,\mu(\mathrm{d}x) + i\int \sin(\langle u, x\rangle)\,\mu(\mathrm{d}x) \tag{31.64}$$

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 31.40 Characteristic Function** $\varphi_X$**:** Let $\mathbf{X}$ be an $\mathbb{R}^n$-valued random variable. Its characteristic function $\varphi_{\mathbf{X}}$ is defined on $\mathbb{R}^n$ as:

$$\varphi_{\mathbf{X}}(u) = \int \mathrm{e}^{i\langle \mathbf{u},\mathbf{x}\rangle}\mathbb{P}^X(\mathrm{d}\mathbf{x}) = \widehat{\mathbb{P}^X}(\mathbf{u}) \tag{31.65}$$

$$= \mathbb{E}\left[\mathrm{e}^{i\langle \mathbf{u},\mathbf{x}\rangle}\right] \tag{31.66}$$

**Corollary 31.31 :** The characteristic function $\varphi_X$ 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 31.7 :** Let $\mu$ be a probability measure on $\mathbb{R}^n$. Then $\hat{\mu}$ is a bounded continuous function with $\hat{\mu}(0) = 1$.

add proof

**Theorem 31.8 Uniqueness Theorem:** The Fourier Transform $\hat{\mu}$ of a probability measure $\mu$ on $\mathbb{R}^n$ *characterizes* $\mu$. That is, if two probability measures on $\mathbb{R}^n$ admit the same Fourier transform, they are equal.

add proof

**Corollary 31.32 :** Let $\mathbf{X} = (X_1, \dots, X_n)$ be an $\mathbb{R}^n$-valued random variable. Then the real valued r.v.'s $(X_j)_{1\leqslant j\leqslant n}$ are independent if and only if:

$$\varphi_X(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \tag{31.67}$$

### Proofs

**Proof 31.1.** *[cor. 31.11]: Let $\mathcal{C}$ denote all open intervals. Since every open set in $\mathbb{R}$ is the countable union of open intervals[def. 19.10], it holds that $\sigma(\mathcal{C})$ is the Borel $\sigma$-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 $b$*

$$(a, b) = \cup_{n=1}^\infty (a_n, b_n] = \cup_{n=1}^\infty \left((-\infty, b_n] \cap (-\infty, a_n]^{\mathrm{C}}\right)$$

*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*

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma((D) \subset \mathcal{B}$$

**Proof 31.2.** *theorem 31.3 Plug eq. (31.22) into the denominator and* **??** *into the nominator and then use [def. 31.17]:*

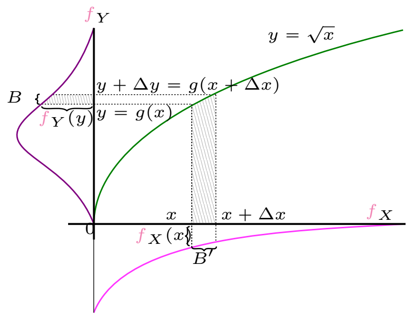
$$\frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i\in I}\mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

**Proof 31.3.** **??***:*

$$Y = g(X) \qquad \Longleftrightarrow \qquad \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = \mathrm{p}_Y(y)$$

**Proof 31.4.** **??** *(non-formal): The probability contained in a differential area must be invariant under a change of variables that is:*

$$|f_Y(y)\,\mathrm{d}y| = |f_x(x)\,\mathrm{d}x|$$



**Proof 31.5.** **??** *from CDF:*

$$\mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \begin{cases} \mathbb{P}(X \leqslant g^{-1}(y)) & \textit{if } g \textit{ is increas.} \\ \mathbb{P}(X \geqslant g^{-1}(y)) & \textit{if } g \textit{ is decreas.} \end{cases}$$

*If $g$ is monotonically increasing:*

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X(g^{-1}(y)) = f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$

*If $g$ is monotonically decreasing:*

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$

**Proof 31.6.** **??***:* **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:*

$$y = g(x) \qquad \Rightarrow \qquad \mathbb{P}(y) = \mathbb{P}(g(x)) \textit{ (for disc. rv.)}$$

*Now lets consider the probability for the continuous r.v.s:*

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t)\,\mathrm{d}t \xrightarrow{\Delta x \to 0} |\Delta x \cdot f_x(x)|$$

*For $y$ we use Taylor (**??**)*

$$\begin{aligned} g(x + \Delta x) &\overset{eq.\ (22.52)}{=} g(x) + \frac{\mathrm{d}g}{\mathrm{d}x}\Delta y && \textit{for } \Delta x \to 0 \\ &= y + \Delta y && \textit{with } \Delta y := \frac{\mathrm{d}g}{\mathrm{d}x} \cdot \Delta x \end{aligned} \tag{31.68}$$

**Thus** *for $\mathbb{P}(Y \in B')$ it follows:*

$$\begin{aligned} \mathbb{P}(X \in B') &= \int_y^{y+\Delta y} f_Y(t)\,\mathrm{d}t \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)| \\ &= \left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x \cdot f_Y(y)\right| \end{aligned}$$

*Now we simply need to related the surface of the two pdfs:*

$$B = [x, x + \Delta x] \overset{\textit{same surfaces}}{\propto} [y, y + \Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\overset{\Delta y \to 0}{\Longleftrightarrow} |f_Y(y) \cdot \Delta y| = \left|f_Y(y) \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x\right| = |f_X(x) \cdot \Delta x|$$

$$f_Y(y)\left|\cdot\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right||\Delta x| = f_X(x) \cdot |\Delta x|$$

$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}g^{-1}(y)\right|}$$

**Proof 31.7.** *[def. 31.35]*

$$\begin{aligned}
\mathbb{E}\left[\epsilon^{\intercal}\mathbf{A}\epsilon\right] &\overset{eq.\ (25.51)}{=} \mathbb{E}\left[tr(\epsilon^{\intercal}\mathbf{A}\epsilon)\right] \\
&\overset{eq.\ (25.53)}{=} \mathbb{E}\left[tr(\mathbf{A}\epsilon\epsilon^{\intercal})\right] \\
&= tr\left(\mathbb{E}\left[\mathbf{A}\epsilon\epsilon^{\intercal}\right]\right) \\
&= tr\left(\mathbf{A}\mathbb{E}\left[\epsilon\epsilon^{\intercal}\right]\right) \\
&= tr\left(\mathbf{A}\left(\Sigma + \mu\mu^{\intercal}\right)\right) \\
&= tr\left(\mathbf{A}\Sigma\right) + tr\left(\mathbf{A}\mu\mu^{\intercal}\right) \\
&\overset{eq.\ (25.51)}{=} tr\left(\mathbf{A}\Sigma\right) + \mathbf{A}\mu\mu^{\intercal}
\end{aligned}$$

**Proof 31.8.** *law 31.7*

$$\begin{aligned}
\mathbb{E}[X] = \sum_x x \cdot \mathrm{p}_X(x) &= \sum_x x \cdot \sum_y \mathrm{p}_{X,Y}(x,y) \\
&= \sum_x x \cdot \sum_y \mathrm{p}_{X|Y}(x|y) \cdot \mathrm{p}_Y(y) \\
&= \sum_y \mathrm{p}_Y(y) \cdot \sum_x x \cdot \mathrm{p}_{X|Y}(x|y) \\
&= \sum_y \mathrm{p}_Y(y) \cdot \mathbb{E}[X|Y] = \mathbb{E}_Y\left[\mathbb{E}[X|Y]\right]
\end{aligned}$$

### Examples

**Example 31.1 :**

- 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: $\Omega = \{1, 2, 3, \ldots\}$
- The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$

**Example 31.2 :**

- 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 31.3 :** 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:

$$\mathcal{F} = \{\varnothing, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

**Example 31.4 :** If we are only interested in the subset $A \in \Omega$ of our experiment, then we can look at the corresponding generating $\sigma$-algebra $\sigma(A) = \left\{\varnothing, A, A^{\mathrm{C}}, \Omega\right\}$.

**Example 31.5 :**

- open half-lines: $(-\infty, a)$ and $(a, \infty)$,
- union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
- closed interval: $[a, b] = \overline{(-\infty, \cup a) \cup (b, \infty)}$,
- 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: $\{a_1, \ldots, a_n\} = \bigcup_{k=1}^{n} a_k$.

**Example 31.6 Equivalent (Probability) Measures:**

$$\Omega = \{1, 2, 3\} \qquad \begin{aligned} \mathbb{P}(\{1, 2, 3\}) &= \{2/3, 1/6, 1/6\} \\ \tilde{\mathbb{P}}(\{1, 2, 3\}) &= \{1/3, 1/3, 1/3\} \end{aligned}$$

**Example 31.7 :**

add example fat book p.1286

add example prob th book 4

**Example 31.8 ??:** Let $X, Y \overset{\text{ind.}}{\sim} \mathcal{N}(0, 1)$.
**Question**: proof that:

$$U = X + Y \qquad\qquad V = X - 1$$

are indepdent and normally distributed:

$$h(u, v) = \begin{cases} h_1(u, v) = \underline{\frac{u+v}{2}} \\ h_2(u, v) = \underline{\frac{u-v}{2}} \end{cases} \qquad J = \det\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$

$$\begin{aligned}
f_{U,V} &= f_{X,Y}(\underline{x}, \underline{y}) \cdot \frac{1}{2} \\
&\overset{\text{indp.}}{=} f_X(\underline{x}) \cdot f_X(\underline{y}) \\
&= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\underline{x}^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\underline{y}^2}{2}} \\
&= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left\{\left(\underline{\frac{u+v}{2}}\right)^2 + \left(\underline{\frac{u-v}{2}}\right)^2 / 2\right\}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{2}} \mathrm{e}^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} \mathrm{e}^{-\frac{v^2}{4}}
\end{aligned}$$

Thus $U, V$ are independent r.v. distributed as $\mathcal{N}(0, 2)$.

# 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:

$$\mathrm{p}_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapet

**Definition 32.1 Almost Surely $\mathbb{P}$-(a.s.):**
**Let** $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \qquad \iff \qquad \omega \text{ happens a.s.} \tag{32.1}$$

**Definition 32.2 Probability Mass Function (PMF):**

**Definition 32.3 Discrete Random Variable (DVR):** The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \ldots, 8\} \qquad\qquad \mathcal{X} = \mathbb{N} \tag{32.2}$$

**Definition 32.4 Probability Density Function (PDF):** Is real function $f : \mathbb{R}^n \to [0, \infty)$ that satisfies:

**Non-negativity:**

$$f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n \tag{32.3}$$

**Normalization:**

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \overset{!}{=} 1 \tag{32.4}$$

**Must be integrable** (32.5)

**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 $\infty$. 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:

$$\mathbb{P}(X = x) = 0 \qquad\qquad \textbf{a.s.}$$

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 32.5 Continuous Random Variable (CRV):** A real random variable (rrv) $X$ is said to be (absolutely) continuous if there exists a pdf ([def. 32.4]) $f_X$ s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,\mathrm{d}x \tag{32.6}$$

**Property 32.1 Zero Probability:** If $X$ is a continuous rrv ([def. 32.5]), then:

$$\mathbb{P}(X = a) = 0 \qquad\qquad \forall a \in \mathbb{R} \tag{32.7}$$

**Property 32.2 Open vs. Closed Intervals:** For any real numbers $a$ and $b$, with $a < b$ it holds:

$$\begin{aligned}
\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a \leqslant X < b) &= \mathbb{P}(a < X \leqslant b) \\
&= \mathbb{P}(a < X < b)
\end{aligned} \tag{32.8}$$

$\iff$ 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 32.1 :** In particular for any real numbers $a$ and $b$ with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_x(x)\,\mathrm{d}x$$

**Proof 32.1.** *Property 32.1:*

$$\begin{aligned}
\mathbb{P}(X = a) &= \lim_{\Delta x \to 0} \mathbb{P}(X \in [a, a + \Delta x]) \\
&= \lim_{\Delta x \to 0} \int_a^{a + \Delta x} f_X(x)\,\mathrm{d}x = 0
\end{aligned}$$

**Proof 32.2.** *Property 32.2:*

$$\begin{aligned}
\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a \leqslant X < b) &= \mathbb{P}(a < X \leqslant b) \\
&= \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x
\end{aligned}$$

**Definition 32.6 Support of a probability density function:** The support of the density of a pdf $f_X(.)$ is the set of values of the random variable $X$ s.t. its pdf is non-zero:

$$\mathsf{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{32.9}$$

**Note**: this is not a rigorous definition.

**Theorem 32.1 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 32.2 Identically Distributed:** From theorem 32.1 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 Fucntion

**Definition 32.7 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:

$$F_X(x) = \mathbb{P}(X \leqslant x) \qquad\qquad \forall x \in \mathbb{R}$$

**Property 32.3:**

**Monotonically Increasing**

$$x \leqslant y \iff F_X(x) \leqslant F_X(y) \quad \forall x, y \in \mathbb{R} \tag{32.10}$$

**Upper Limit**

$$\lim_{x \to \infty} F_X(x) = 1 \tag{32.11}$$

**Lower Limit**

$$\lim_{x \to -\infty} F_X(x) = 0 \tag{32.12}$$

**Definition 32.8 CDF of a discret rv X: Let** $X$ be discret rv with pdf $\mathrm{p}_X$, then the CDF of $X$ is given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{t=-\infty}^{x} \mathrm{p}_X(t)$$

**Definition 32.9 CDF of a continuous rv X: Let** $X$ be continuous rv with pdf $f_X$, then the CDF of $X$ is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \qquad \iff \qquad \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

**Lemma 32.1 Probability Interval:** Let $X$ be a continuous rrv with pdf $f_X$ and cumulative distribution function $F_X$, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = F_X(b) - F_X(a) \tag{32.13}$$

**Proof 32.3.** *[def. 32.9]:*

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

**Proof 32.4.** *lemma 32.1:*

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$

*or by the fundamental theorem of calculus (theorem 22.2):*

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,\mathrm{d}t = \int_a^b \frac{\partial F_X(t)}{\partial t}\,\mathrm{d}t = [F_X(t)]|_a^b$$

**Theorem 32.2 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 32.10 Expectation (disc. case):**

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\mathbf{x}} \in \mathcal{X}} \bar{\mathbf{x}}\mathrm{p}_x(\bar{\mathbf{x}}) \tag{32.14}$$

**Definition 32.11 Expectation (cont. case):**

$$\mathbb{E}_x[x] := \int_{\bar{\mathbf{x}} \in \mathcal{X}} \bar{\mathbf{x}} f_x(\bar{\mathbf{x}})\,\mathrm{d}\bar{\mathbf{x}} \tag{32.15}$$

**Law 32.1 Expectation of independent variables:**

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{32.16}$$

**Property 32.4 Translation and scaling:** If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \tag{32.17}$$

**Thus** $\mathbb{E}$ is a linear operator ([def. 22.17]).

**Note: Expectation of the expectation**

The expectation of a r.v. $X$ is a constant hence with Property 32.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{32.18}$$

**Property 32.5 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:

$$\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[(\mathbf{XB})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B} \tag{32.19}$$

**Proof 32.5.** *eq. (32.24):*

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{p}_{X,Y}(x, y)xy \\
&\overset{??}{=} \sum_{x \in \mathcal{X}} \mathrm{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathrm{p}_Y(y)y = \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

**Definition 32.12 Autocorrelation/Crosscorelation $\gamma(t_1, t_2)$:** Describes the covariance ([def. 32.16]) between the two values of a stochastic process $(\mathbf{X}_t)_{t \in T}$ at different time points $t_1$ and $t_2$.

$$\gamma(t_1, t_2) = \mathrm{Cov}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right] = \mathbb{E}\left[\left(\mathbf{X}_{t_1} - \mu_{t_1}\right)\left(\mathbf{X}_{t_2} - \mu_{t_2}\right)\right] \tag{32.20}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \mathrm{Cov}\left[\mathbf{X}_t, \mathbf{X}_t\right] \overset{eq.\ (32.35)}{=} = \mathbb{V}\left[\mathbf{X}_t\right] \tag{32.21}$$

**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 $\mathbf{x}(t)$ 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 autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

# 2. Key Figures

## 2.1. The Expectation

move to prob theory maybe

**Definition 32.13 Expectation (disc. case):**

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\mathbf{x}}\in\mathcal{X}} \bar{\mathbf{x}}\,\mathrm{p}_x(\bar{\mathbf{x}}) \tag{32.22}$$

**Definition 32.14 Expectation (cont. case):**

$$\mathbb{E}_x[x] := \int_{\bar{\mathbf{x}}\in\mathcal{X}} \bar{\mathbf{x}}\,f_x(\bar{\mathbf{x}})\,\mathrm{d}\bar{\mathbf{x}} \tag{32.23}$$

**Law 32.2 Expectation of independent variables:**

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{32.24}$$

**Property 32.6 Translation and scaling:** If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \tag{32.25}$$

**Thus** $\mathbb{E}$ is a linear operator[def. 22.17].

**Property 32.7**
**Affine Transformation of the Expectation:**
If $\mathbf{X} \in \mathbb{R}^n$ is a randomn vector, $\mathbf{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mu + \mathbf{b} \tag{32.26}$$

**Note: Expectation of the expectation**

The expectation of a r.v. $X$ is a constant hence with Property 32.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{32.27}$$

**Property 32.8 Matrix×Expectation:** If $\mathbf{X} \in \mathbb{R}^n$ is a randomn vector and $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\,\mathbf{B} \tag{32.28}$$

**Proof 32.6.** *eq. (32.24):*

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathrm{p}_{X,Y}(x,y)xy\\ &\overset{\mathbf{??}}{=} \sum_{x\in\mathcal{X}} \mathrm{p}_X(x)x \sum_{y\in\mathcal{Y}} \mathrm{p}_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

## 2.2. The Variance

**Definition 32.15 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:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \overset{\text{see } \mathbf{??}\ 32.7}{=} \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{32.29}$$

### 2.2.1. Properties

**Property 32.9 Variance of a Constant:** If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic ⇒ we have no uncertainty ⇒ no variance:

$$\mathbb{V}[a] = 0 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{32.30}$$

see shift and scaling for proof **??** 32.8

**Property 32.10 Shifting and Scaling:**

$$\mathbb{V}[a + bX] = a^2\sigma^2 \qquad \textbf{with} \qquad a \in \mathbb{R} \tag{32.31}$$

see **??** 32.8

**Property 32.11** [proof 32.9]
**Affine Transformation of the Variance:**
If $\mathbf{X} \in \mathbb{R}^n$ is a randomn vector, $\mathbf{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{V}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal \tag{32.32}$$

**Definition 32.16 Covariance:** The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned} \tag{32.33}$$

see **??** 32.10

**Definition 32.17 Covariance Matrix:** The variance of a $k$-dimensional random vector $\mathbf{X} = \begin{pmatrix}X_1 & \dots & X_k\end{pmatrix}$ is given by a p.s.d. eq. (25.105) 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:

$$\begin{aligned}\mathbb{V}[\mathbf{X}] :=&\Sigma(\mathbf{X}) := \mathrm{Cov}[\mathbf{X},\mathbf{X}] :=\\ =&\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\intercal\right]\\ =&\mathbb{E}\left[\mathbf{X}\mathbf{X}^\intercal\right] - \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}]^\intercal \in [-\infty,\infty]\end{aligned} \tag{32.34}$$

$$= \begin{bmatrix}\mathbb{V}[X_1] & \cdots & \mathrm{Cov}[X_1, X_k]\\ \vdots & \ddots & \\ \mathrm{Cov}[X_k, X_1] & \cdots & \mathbb{V}[X_k]\end{bmatrix}$$

$$= \begin{bmatrix}\mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)]\\ \vdots & \ddots & \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)]\end{bmatrix}$$

**Note: Covariance and Variance**

The variance is a special case of the covariance in which two variables are identical:

$$\mathrm{Cov}[X,X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{32.35}$$

add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

**Property 32.12 Translation and Scaling:**

$$\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X,Y) \tag{32.36}$$

**Property 32.13**
**Affine Transformation of the Covariance:**
If $\mathbf{X} \in \mathbb{R}^n$ is a randomn vector, $\mathbf{A} \in \mathbb{R}^{m\times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathrm{Cov}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal = \mathbf{A}\Sigma(\mathbf{X})\mathbf{A}^\intercal \tag{32.37}$$

**Definition 32.18 Correlation Coefficient:** Is the standardized version of the covariance:

$$\begin{aligned}\mathrm{Corr}[\mathbf{X}] :&= \frac{\mathrm{Cov}[\mathbf{X}]}{\sigma_{X_1}\cdots\sigma_{X_k}} \in [-1,1]\\ &= \begin{cases}+1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R}\\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R}\end{cases}\end{aligned} \tag{32.38}$$

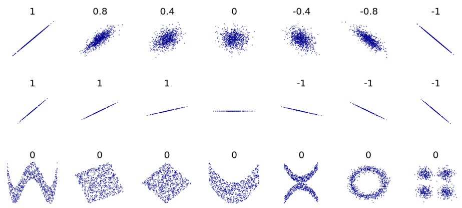


Figure 14: Several sets of $(x, y)$ points, with their correlation coefficient

**Law 32.3 Translation and Scaling:**

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \tag{32.39}$$

**Note**

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 14), **but** not the slope of that relationship (middle row fig. 14) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 14 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$ implies that there does not exist a **linear** relationship between the random variables X and Y.

**Difference Covariance&Correlation**

1. Variance is affected by scaling and covariance not **??** and law 32.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

**Law 32.4 Covariance of independent RVs:** The covariance/correlation of two independent variable's (**??**) is zero:

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\\ &\overset{\text{eq. (32.24)}}{=} = \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0\end{aligned}$$

**Zero covariance/correlation⇏ independence**

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0 \nRightarrow \mathrm{p}_{X,Y}(x,y) = \mathrm{p}_X(x)\mathrm{p}_Y(y)$$

**For example**: let $X \sim \mathcal{U}([-1,1])$ and let $Y = X^2$.

1. Clearly $X$ and $Y$ are dependent
2. **But** the covariance/correlation between $X$ and $Y$ is non-zero:

$$\begin{aligned}\mathrm{Cov}(X{,}Y) &= \mathrm{Cov}(X, X^2) = \mathbb{E}\left[X\cdot X^2\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right]\\ &= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \overset{\text{eq. (32.63)}}{\underset{\text{eq. (32.52)}}{=}} 0 - 0\cdot\mathbb{E}\left[X^2\right]\end{aligned}$$

⇒ the relationship between Y and X must be non-linear.

**Definition 32.19 Quantile:** Are specific values $q_\alpha$ in the range[def. 22.12] of a random variable $X$ that are defined as the value for which the cumulative probability is less then $q_\alpha$ with probability $\alpha \in (0,1)$:

$$q_\alpha : \mathbb{P}(X \leqslant x) = F_X(q_\alpha) = \alpha \quad \xrightarrow{F\text{ invert.}} \quad q_\alpha = F_X^{-1}(\alpha) \tag{32.40}$$

add figure

# 3. Proofs

**Proof 32.7.** *eq.* (32.29)

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right]\\ &\overset{\textit{Property } 32.6}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

**Proof 32.8.** *Property 32.10*

$$\begin{aligned}\mathbb{V}[a + bX] &= \mathbb{E}\left[(a + bX - \mathbb{E}[a + bX])^2\right]\\ &= \mathbb{E}\left[\left(\cancel{a} + bX - \cancel{a} - b\mathbb{E}[X]\right)^2\right]\\ &= \mathbb{E}\left[(bX - b\mathbb{E}[X])^2\right]\\ &= \mathbb{E}\left[b^2(X - \mathbb{E}[X])^2\right]\\ &= b^2\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = b^2\sigma^2\end{aligned}$$

**Proof 32.9.** *Property 32.11*

$$\begin{aligned}\mathbb{V}(\mathbf{A}\mathbf{X} + b) &= \mathbb{E}\left[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{X}\mathbf{A}])^2\right] + 0 =\\ =&\mathbb{E}\left[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^\intercal\right]\\ =&\mathbb{E}\left[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{A}(\mathbf{X} - (\mathbb{E}[\mathbf{X}]))^\intercal\right]\\ =&\mathbb{E}\left[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - (\mathbb{E}[\mathbf{X}])^\intercal\mathbf{A}^\intercal\right]\\ =&\mathbf{A}\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - (\mathbb{E}[\mathbf{X}])^\intercal\right]\mathbf{A}^\intercal = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal\end{aligned}$$

**Proof 32.10.** *eq.* (32.33)

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]\,Y + \mathbb{E}[X]\,\mathbb{E}[Y]]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

## Discrete Distributions

**Definition 32.20 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 $\mathrm{Bern}(\mathrm{p})$

**Definition 32.21 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 32.22 Bernoulli Distribution $X \sim \mathrm{Bern}(\mathrm{p})$:**
$X$ is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter $\mathrm{p}$ that signifies the success probability:

$$\begin{aligned}\mathrm{p}(x;\mathrm{p}) &= \begin{cases}\mathrm{p} & \text{for } x=1\\ 1-\mathrm{p} & \text{for } x=0\end{cases} \iff \begin{cases}\mathbb{P}(X=1)=\mathrm{p}\\ \mathbb{P}(X=0)=1-\mathrm{p}\end{cases}\\ &= \mathrm{p}^x\cdot(1-\mathrm{p})^{1-x} \quad \text{for } x\in\{0,1\}\end{aligned}$$

$$\mathbb{E}[X]=\mathrm{p} \qquad (32.41) \qquad \mathbb{V}[X]=\mathrm{p}(1-\mathrm{p}) \qquad (32.42)$$

### 4.2. Multinoulli/Categorical Distribution $\mathrm{Cat}(n,\mathrm{p})$

**Definition 32.23**
**Multinulli/Categorical Distribution $X\sim\mathrm{Cat}(\mathbf{p})$:**
Is the generalization of the Bernoulli distribution[def. 32.22] to a sample space[def. 31.2] of $k$ individual items $\{c_1,\dots,c_c\}$ with probabilities $\mathbf{p}=\{\mathrm{p}_1,\dots,\mathrm{p}_k\}$:

$$p(x=c_i|\mathbf{p})=\mathrm{p}_i \qquad\iff\qquad p(x|\mathbf{p})=\prod_i^k \mathrm{p}_i^{\delta[x=c_i]}$$

$$\sum_{j=1}^k \mathrm{p}_j=1 \qquad \mathrm{p}_j\in[0,1] \qquad \forall j=1,\dots,k \qquad (32.43)$$

$$\mathbb{E}[X]=\mathbf{p} \qquad \mathbb{V}[X]_{i,j}=\Sigma_{i,j}=\begin{cases}\mathrm{p}_i(1-\mathrm{p}_i) & \text{if } i=j\\ -\mathrm{p}_i\mathrm{p}_j & \text{if } i\neq j\end{cases}$$

**Corollary 32.3**
**One-hot encoded Categorical Distribution:**
If we encode the $k$ categories by a *sparse vectors*[def. 25.66] with norm one:

$$\mathbb{B}_r^n=\left\{\mathbf{x}\in\{0,1\}^n:\mathbf{x}^\intercal\mathbf{x}=\sum_{i=1}^n\mathbf{x}=1\right\}$$

s.t. $\qquad \mathbf{x}_j=\mathbf{e}_j \qquad\iff\qquad \mathbf{x}=c_j$

then we can rewrite eq. (32.43) as:

$$\mathrm{p}(\mathbf{x}|\mathbf{p})=\prod_i^k\mathbf{x}_i\cdot\mathrm{p}_i \qquad \sum_{j=1}^k\mathrm{p}_j=1 \qquad (32.44)$$

### 4.3. Binomial Distribution $\mathcal{B}(n,\mathrm{p})$

**Definition 32.24 Binomial Distribution:**
Models the probability of exactly $X$ success given a fixed number *$n$-Bernoulli experiments*[def. 32.21], where the probability of success of a single experiment is given by $\mathrm{p}$:

$$\mathrm{p}(x)=\binom{n}{x}\mathrm{p}^x(1-\mathrm{p})^{n-x}$$

$n$ :nb. of repetitions
$x$ :nb. of successes
$\mathrm{p}$ :probability of success

$$\mathbb{E}[X]=n\mathrm{p} \qquad (32.45) \qquad \mathbb{V}[X]=n\mathrm{p}(1-\mathrm{p}) \qquad (32.46)$$

**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:

$$\underbrace{(\mathrm{p}\cdots\mathrm{p})}_{\text{x-times}}\cdot\underbrace{(q\cdots q)}_{n-x\text{-times}}=\mathrm{p}^x q^{n-x}$$

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.

$$P(n;n_1,n_2)=\frac{n!}{x!(n-x)!}=\binom{n}{x}$$

### 4.4. Geometric Distribution $\mathrm{Geom}(\mathrm{p})$

**Definition 32.25 Geometric Distribution $\mathrm{Geom}(\mathrm{p})$:**
Models the probability of the number $X$ of Bernoulli trials[def. 32.21] *until the first success*

$$\mathrm{p}(x)=\mathrm{p}(1-\mathrm{p})^{x-1}$$

$x$ :nb. of repetitions *until first success*
$\mathrm{p}$ :success probability *of single Bernoulli experiment*

$$F(x)=\sum_{i=1}^x\mathrm{p}(1-\mathrm{p})^{i-1}\overset{??}{=}1-(1-\mathrm{p})^x$$

$$\mathbb{E}[X]=\frac{1}{\mathrm{p}} \qquad (32.47) \qquad \mathbb{V}[X]=\frac{1-\mathrm{p}}{\mathrm{p}^2} \qquad (32.48)$$

**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 $\mathrm{p}(x)$:
$$x\geqslant\frac{\mathrm{p}(x)}{1-\mathrm{p}}$$
- $\log(1-\mathrm{p})\approx-\mathrm{p}$ for small $\mathrm{p}$

### 4.5. Poisson Distribution $\mathrm{Pois}(\lambda)$

**Definition 32.26 Poisson Distribution:** Is an extension of the binomial distribution, where the realization $x$ of the random variable $X$ may attain values in $\mathbb{Z}_{\geqslant0}$.
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.

$$\mathrm{p}(x)=\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad \begin{matrix}\lambda>0\\ x\in\mathbb{Z}_{\geqslant0}\end{matrix} \qquad (32.49)$$

**Event Rate $\lambda$:** describes the average number of events in a single interval.

$$\mathbb{E}[X]=\lambda \qquad (32.50) \qquad \mathbb{V}[X]=\lambda \qquad (32.51)$$

1

## Continuous Distributions

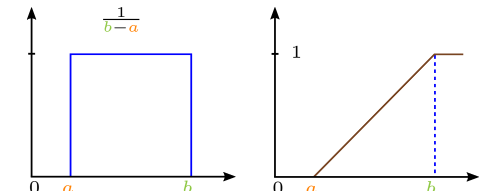
### 5.1. Uniform Distribution $\mathcal{U}(a,b)$

**Definition 32.27 Uniform Distribution $\mathcal{U}(a,b)$:**
Is probability distribution, where all intervals of the **same** length on the distribution's support[def. 32.6] $\mathbf{supp}(\mathcal{U}[a,b])=[a,b]$ are equally probable/likely.

$$\mathrm{f}(x)=\frac{1}{b-a}\mathbb{1}_{x\in[a;b)}=\begin{cases}\frac{1}{b-a}=\text{const} & \text{if} \quad a\leqslant x\leqslant b\\ 0 & \qquad\text{else}\end{cases} \qquad (32.52)$$

$$F(x)=\begin{cases}0 & & x<a\\ \frac{x-a}{b-a} & \text{if} & a\leqslant x\leqslant b\\ 1 & & x>b\end{cases} \qquad (32.53)$$

$$\mathbb{E}[X]=\frac{a+b}{2} \qquad \mathbb{V}(X)=\frac{(b-a)^2}{12} \qquad (32.54)$$



(a) $f_X(x)$ (b) $F_X(x)$

### 5.2. Exponential Distribution $\exp(\lambda)$

**Definition 32.28 Exponential Distribution $X\sim\exp(\lambda)$:**
Is the continuous analogue to the geometric distribution [def. 32.25].
It describes the probability $\mathrm{f}(x;\lambda)$ 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 $x$.

$$\mathrm{f}(x;\lambda)=\begin{cases}\lambda\mathrm{e}^{-\lambda x} & \text{if} \quad x\geqslant0\\ 0 & \qquad x<0\end{cases} \qquad (32.55)$$

$$F(x;\lambda)=\begin{cases}1-\mathrm{e}^{-\lambda x} & \text{if} \quad x\geqslant0\\ 0 & \qquad x<0\end{cases} \qquad (32.56)$$

$$\mathbb{E}[X]=\frac{1}{\lambda} \qquad \mathbb{V}(X)=\frac{1}{\lambda^2} \qquad (32.57)$$

### 5.3. Laplace Distribution

**Definition 32.29 Laplace Distribution:**

Laplace Distibution $\qquad f(\mathbf{x};\mu,\sigma)=\dfrac{1}{2\sigma}\exp\left(-\dfrac{|\mathbf{x}-\mu|}{\sigma}\right) \qquad (32.58)$

### 5.4. The Normal Distribution $\mathcal{N}(\mu,\sigma)$

**Definition 32.30 Normal Distribution $\mathbf{X}\sim\mathcal{N}(\mu,\sigma^2)$:**
Is a symmetric distribution where the population parameters $\mu$, $\sigma^2$ are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X]=\mu \qquad \mathbb{V}(X)=\sigma^2 \qquad (32.59)$$

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad (32.60)$$

$$F(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^x\exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\}\mathrm{d}u \qquad (32.61)$$

$$x\in\mathbb{R} \qquad \text{or} \qquad -\infty<x<\infty$$

$$\varphi_X(u)=\exp\left\{iu\mu-\frac{u^2\sigma^2}{2}\right\} \qquad (32.62)$$

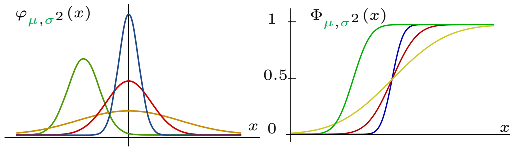


Figure 16: $\mu=0$, $\sigma^2=0.2$ $\quad\mu=0$, $\sigma^2=1.0$ $\quad\mu=0$, $\sigma^2=5.0$ $\quad\mu=-2$, $\sigma^2=0.5$

**Property 32.14:** $\mathbb{P}_X(\mu-\sigma\leqslant x\leqslant\mu+\sigma)=0.66$

**Property 32.15:** $\mathbb{P}_X(\mu-2\sigma\leqslant x\leqslant\mu+2\sigma)=0.95$

### 5.5. The Standard Normal distribution $\mathcal{N}(0,1)$

**Historic Problem:** the cumulative distribution eq. (32.61) 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 32.31**
**Standard Normal Distribution $\mathbf{X}\sim\mathcal{N}(0,1)$:**

$$\mathbb{E}[X]=0 \qquad \mathbb{V}(X)=1 \qquad (32.63)$$

$$f(x;0,1)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2} \qquad (32.64)$$

$$F(x;0,1)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^x\mathrm{e}^{-\frac{1}{2}u^2}\mathrm{d}u \qquad (32.65)$$

$$x\in\mathbb{R} \qquad \text{or} \qquad -\infty<x<\infty$$

$$\psi_X(u)=\mathrm{e}^{\frac{u^2}{2}} \qquad \varphi_X(u)=\mathrm{e}^{-\frac{u^2}{2}} \qquad (32.66)$$

**Corollary 32.4**
**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 $\alpha$-quantile[def. 32.19] is then denoted by:

$$z_\alpha=\Phi^{-1}(\alpha) \qquad \alpha\in(0,1) \qquad (32.67)$$

#### 5.5.1. Calculating Probabilities

**Property 32.16 Symmetry:** Let $z>0$

$$\begin{aligned}\mathbb{P}(Z\leqslant z) &= \Phi(z) && (32.68)\\ \mathbb{P}(Z\leqslant -z) &= \Phi(-z)=1-\Phi(z) && (32.69)\\ \mathbb{P}(-a\leqslant Z\leqslant b) &= \Phi(b)-\Phi(-a)=\Phi(b)-(1-\Phi(a))\\ &\overset{a=b=z}{=} 2\Phi(z)-1 && (32.70)\end{aligned}$$

### 5.5.2. Linear Transformations of Normal Dist.

**Proposition 32.1 Linear Transformation** [proof 32.11]: 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:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \quad \Longleftrightarrow \quad f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \tag{32.71}$$

**Proposition 32.2 Standardization** [proof 32.12]: 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:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0, 1) \tag{32.72}$$

**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 32.3** [proof 32.13]
**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:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{32.73}$$

## 6. The Multivariate Normal distribution

**Definition 32.32 Multivariate Normal/Gaussian**: An $\mathbb{R}^n$-valued random variable $\mathbf{X} = (X_1 \quad \dots, X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma : \quad \mathscr{L}\left(\sum_{i=1}^{n} \alpha_i X_j\right) = \mathcal{N}(\mu, \sigma^2) \qquad \forall \alpha_i \in \mathbb{R} \tag{32.74}$$

(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 32.33**
**Multivariate Normal distribution $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$**:
A $k$-dimensional random vector

$$\mathbf{X} = (X_1 \quad \dots \quad X_n)^\intercal \quad \text{with} \quad \boldsymbol{\mu} = (\mathbb{E}[\mathbf{x}_1] \quad \dots \quad \mathbb{E}[\mathbf{x}_k])^\intercal$$

**and** $k \times k$ **p.s.d**.covariance matrix:

$$\boldsymbol{\Sigma} := \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\intercal] = [\mathrm{Cov}[\mathbf{x}_i, \mathbf{x}_j], 1 \leqslant i, j \leqslant k]$$

follows a $k$-dim multivariate normal/Gaussian distribution if its law[def. 31.23] satisfies:

$$\begin{aligned} f_{\mathbf{X}}(X_1, \dots, X_k) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) && (32.75) \\ &= \underbrace{\frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}}}_{\text{Normalisation}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right) \end{aligned}$$

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{i\mathbf{u}^\intercal\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}\Sigma\mathbf{u}\right\} \tag{32.76}$$

### 6.1. Joint Gaussian Distributions

**Definition 32.34 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 $\mathbf{Z} = \begin{bmatrix} X & Y \end{bmatrix}^\intercal$ is again a GRV.

**Property 32.17** proof 32.15
**Joint Independent Gaussian Random Variables:** Let $X_1, \dots, X_n$ be $\mathbb{R}$-valued *independent* random variables with laws $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Then the law of $\mathbf{X} = (X_1 \quad \dots \quad X_n)$ is a (multivariate) Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \cdots\cdots 0 \\ 0 & \sigma_2^2 \cdots\cdots 0 \\ \vdots & \vdots \quad \ddots \quad \vdots \\ 0 & 0 \cdots\cdots \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{32.77}$$

**Corollary 32.5 Quadratic Form:**
If $\mathbf{x}$ and $\mathbf{y}$ are both independent GRVs

$$\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x) \qquad \mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$$

then they are jointly Gaussian[def. 32.34] given by:

$$\begin{aligned} &\mathrm{p}(\mathbf{x}, \mathbf{y}) = \mathrm{p}(x)\mathrm{p}(y) && (32.78) \\ &\propto \exp\left(-\frac{1}{2}\left\{(\mathbf{x} - \boldsymbol{\mu}_x)^\intercal \boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) + (\mathbf{y} - \boldsymbol{\mu}_y)^\intercal \boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right\}\right) \\ &= \exp\left(-\frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_x)^\intercal \quad (\mathbf{y} - \boldsymbol{\mu}_y)^\intercal]\begin{bmatrix} \boldsymbol{\Sigma}_x^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_y^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{x} - \boldsymbol{\mu}_x \\ \mathbf{y} - \boldsymbol{\mu}_y \end{bmatrix}\right) \\ &\mathrel{\widehat{=}} \exp -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^\intercal \boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z) \end{aligned}$$

**Property 32.18**
**Marginal Distribution of Multivariate Gaussian:** Let $\mathbf{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a 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:

$$A = \{i_1, \dots, i_k\} \qquad i_j \in V \tag{32.79}$$

is given by:

$$\mathbf{X} = \left(X_{i_1} \quad \dots \quad X_{i_k}\right)^\intercal \sim \mathcal{N}(\mu_A, \Sigma_{AA}) \tag{32.80}$$

$$\Sigma = \begin{bmatrix} \sigma_{i_1,i_1}^2 & \cdots\cdots & \sigma_{i_1,i_k}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{i_k,i_1}^2 & \cdots\cdots & \sigma_{i_k,i_k}^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

### 6.2. Conditional Gaussian Distributions

**Property 32.19 Conditional Gaussian Distribution:** Let $\mathbf{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a 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 $V$:

$$A = \{i_1, \dots, i_k\} \qquad B = \{j_1, \dots, j_m\} \qquad i_l, j_{l'} \in V$$

then the conditional distribution of the random vector $\mathbf{X}_A$, conditioned on $\mathbf{X}_B$ given by $\mathrm{p}(\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B)$ is:

$$\mathbf{X}_A = \left(X_{i_1} \quad \dots \quad X_{i_k}\right)^\intercal \sim \mathcal{N}\left(\mu_{A|B}, \Sigma_{A|B}\right) \tag{32.81}$$

$$\boxed{\begin{aligned} \mu_{A|B} &= \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{x}_B - \mu_B) \\ \Sigma_{A|B} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \end{aligned}}$$

**Note**

Can be proofed using the matrix inversion lemma but is a very tedious computation.

maybe add sometime

**Corollary 32.6**
**Conditional Distribution of Joint Gaussian's:** Let $\mathbf{X}$ and $\mathbf{Y}$ be jointly Gaussian random vectors:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\intercal & \\ \mathbf{B} & \end{bmatrix}\right) \tag{32.82}$$

then the *marginal* distribution of $\mathbf{x}$ conditioned on $\mathbf{y}$ can be written as: $X \sim \mathcal{N}\left(\mu_{X|Y}, \Sigma_{X|Y}\right)$

$$\boxed{\begin{aligned} \boldsymbol{\mu}_{X|Y} &= \boldsymbol{\mu}_X + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \boldsymbol{\mu}_Y) \\ \Sigma_{X|Y} &= \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\intercal \end{aligned}} \tag{32.83}$$

add proofs

### 6.3. Transformations

**Property 32.20 Multiples of Gaussian's $\mathbf{A}X$:** Let $\mathbf{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a an $\mathbb{R}^n$ valued Gaussian and let $\mathbf{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$Y = \mathbf{A}X \in \mathbb{R} \qquad Y \sim \mathcal{N}\left(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^\intercal\right) \tag{32.84}$$

**Property 32.21 Affine Transformation of GRVs: Let** $\mathbf{y} \in \mathbb{R}^n$ be GRV, $\mathbf{A} \in \mathbb{R}^{d \times n}$, $b \in \mathbb{R}^d$ and let $\mathbf{x}$ be defined by the affine transformation[def. 25.42]:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + b \qquad \mathbf{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$$

**Then** $\mathbf{x}$ is a GRV (see **??** 32.14).

**Property 32.22 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:

$$\mathbf{z} = \mathbf{A}_x\mathbf{x} + \mathbf{A}_y\mathbf{y} \qquad \mathbf{A}_x \in \mathbb{R}^{d \times n}, \mathbf{A}_x \in \mathbb{R}^{d \times m}$$

**Then** $\mathbf{z}$ is GRV (see **??** 32.16).

**Definition 32.35 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 32.36 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:

$$\mathrm{f}(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\beta x} & \text{if} \quad x > 0 \\ 0 & \phantom{\text{if}} \quad x \leqslant 0 \end{cases} \tag{32.85}$$

$$\Gamma(\alpha) \overset{\text{eq. (22.77)}}{=} \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{32.86}$$

**with** $\alpha, \beta \in \mathbb{R}_{>0}$

## 7. Student's t-distribution

**Definition 32.37 Student' t-distribution:**

add

### 7.1. Delta Distribution

**Definition 32.38 The delta function $\delta(\mathbf{x})$:**
The delta/dirac function $\delta(\mathbf{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\mathbf{x})f(\mathbf{x})\,\mathrm{d}\mathbf{x} = f(0)$$

for any integrable function $f$ on $\mathbb{R}$.
**Or** alternativly by:

$$\begin{aligned} \delta(x - x_0) &= \lim_{\sigma \to 0} \mathcal{N}(x|x_0, \sigma) && (32.87) \\ &\approx \infty\mathbb{1}_{\{x = x_0\}} && (32.88) \end{aligned}$$

**Property 32.23 Properties of $\delta$:**

- **Normalization**: The delta function integrates to 1:
$$\int_{\mathbb{R}} \delta(x)\,\mathrm{d}x = \int_{\mathbb{R}} \delta(x) \cdot c_1(x)\,\mathrm{d}x = c_1(0) = 1$$
where $c_1(x) = 1$ is the constant function of value 1.
- **Shifting**:
$$\int_{\mathbb{R}} \delta(x - x_0)f(x)\,\mathrm{d}x = f(x_0) \tag{32.89}$$
- **Symmetry**: $\int_{\mathbb{R}} \delta(-x)f(x)\,\mathrm{d}x = f(0)$
- **Scaling**: $\int_{\mathbb{R}} \delta(\alpha x)f(x)\,\mathrm{d}x = \frac{1}{|\alpha|}f(0)$

**Note**

- In mathematical terms $\delta$ is not a function but a **gernalized function**.
- We may regard $\delta(x - x_0)$ as a density with all its probability mass centered at the signle point $x_0$.
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normaldistribution eq. (32.87) would be a non-differentiable/discret form of the dirac measure.

**Definition 32.39 Heaviside Step Function:**

$$H(x) := \frac{\mathrm{d}}{\mathrm{d}x}\max\{x, 0\} \qquad x \in \mathbb{R}_{\neq 0} \tag{32.90}$$

or alternatively:

$$H(x) := \int_{-\infty}^{x} \delta(s)\,\mathrm{d}s \tag{32.91}$$

## Proofs

**Proof 32.11.** *proposition 32.1: Let $X$ be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:*

$$\begin{aligned} F_Y(y) &\overset{y>0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \leqslant \frac{y-a}{b}\right) \\ &= F_X\left(\frac{y-a}{b}\right) \\ F_Y(y) &\overset{y<0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \geqslant \frac{y-a}{b}\right) \\ &= 1 - F_X\left(\frac{y-a}{b}\right) \end{aligned}$$

*Differentiating both expressions w.r.t. $y$ leads to:*

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \begin{cases} \frac{1}{b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \\ \frac{1}{-b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \end{cases} = \frac{1}{|b|} f_X(x)\left(\frac{y-a}{b}\right)$$

*eq. (32.71)).*
*in order to prove that $Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right)$ we simply plug $f_X$ in the previous expression:*

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma|b|}\exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma|b|}\exp\left\{-\frac{1}{2}\left(\frac{y - (a + b\mu)}{\sigma|b|}\right)^2\right\} \end{aligned}$$

**Proof 32.12.** *proposition 32.2: Let $X$ be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:*

$$Z := \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = aX + b \quad \textit{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

$$\overset{\text{eq. (32.71)}}{\sim} \mathcal{N}\left(a\mu + b, a^2\sigma^2\right) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$$

**Proof 32.13.** *proposition 32.3: Let $X$ be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:*

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leqslant x) \overset{\substack{-\mu \\ \div\sigma}}{=} \mathbb{P}\left(\frac{X-\mu}{\sigma} \leqslant \frac{x-\mu}{\sigma}\right) = \mathbb{P}\left(Z \leqslant \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

**Proof 32.14.** *Property 32.21 scalar case*
**Let** $y \sim \mathrm{p}(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ *and define* $\mathbf{x} = ay + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}$
**Using** *the Change of variables formula it follows:*

$$\mathrm{p}_x(\bar{x}) \overset{\text{eq. (31.46)}}{=} \frac{\mathrm{p}_y(\bar{y})}{|\frac{\mathrm{d}x}{\mathrm{d}y}|} \qquad \left[\quad |\frac{\mathrm{d}x}{\mathrm{d}y}| = a \quad\right]$$

$$\begin{aligned} &\overset{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}}\exp\left(-\frac{1}{2\sigma^2}\Big(\overbrace{\frac{\bar{x}-b}{a}}^{\bar{y}(\bar{x})} - \mu\Big)^2\right) \\ &= \frac{1}{\sqrt{2\pi a^2\mu^2}}\exp\left(-\frac{1}{2\sigma^2 a^2}\Big(\bar{x}\underbrace{-b - a\mu}_{\mu_x}\Big)^2\right) \end{aligned}$$

**Hence** $x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

**Note**

We can also verify that we have calculated the right mean and variance by:

$$\begin{aligned} \mathbb{E}[x] &= \mathbb{E}[ay + b] = a\mathbb{E}[y] + b = a\mu + b \\ \mathbb{V}[x] &= \mathbb{V}[ay + b] = a^2\mathbb{V}[y] = a^2\sigma^2 \end{aligned}$$

**Proof 32.15.** ??

$$
\begin{aligned}
\mathrm{p}_{\mathbf{X}}(\mathbf{u}) &= \prod_{i}^{n} \mathrm{p}_{X_i}(u_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \\
\varphi_{\mathbf{X}}(\mathbf{u}) &= \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1 u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_n u_n^2\right\} \\
&= \exp\left\{ i\sum_{i}^{n} u_n\mu_n - \frac{1}{2}\underset{\cdots\cdots}{\sum_{i}^{n}\sigma_n u_n^2} \right\} = \exp\left\{ i\mathbf{u}^{\intercal}\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}\underset{\cdots}{\Sigma\mathbf{u}} \right\}
\end{aligned}
$$

**Proof 32.16.** *Property 32.22*

*From Property 32.21 it follows immediately that* $\mathbf{z}$ *is GRV* $\mathbf{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ *with:*

$$
\mathbf{z} = \mathbf{A}\xi \qquad \textbf{with} \qquad \mathbf{A} = \begin{bmatrix}\mathbf{A}_x & \mathbf{A}_y\end{bmatrix} \textit{ and } \xi = \begin{pmatrix}\mathbf{x} & \mathbf{y}\end{pmatrix}
$$

*Knowing that* $\mathbf{z}$ *is a GRV it is sufficient to calculate* $\mu_z$ *and* $\Sigma_z$ *in order to characterize its distribution:*

$$
\begin{aligned}
\mathbb{E}[\mathbf{z}] &= \mathbb{E}\left[\mathbf{A}_x x + \mathbf{A}_y y\right] = \mathbf{A}_x\mu_x + \mathbf{A}_y\mu_y \\
\mathbb{V}[\mathbf{z}] &= \mathbb{V}[\mathbf{A}\xi] \overset{??}{=} \mathbf{A}\mathbb{V}[\xi]\mathbf{A}^{\intercal} \\
&= \begin{bmatrix}\mathbf{A}_x & \mathbf{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\mathbf{A}_x & \mathbf{A}_y\end{bmatrix}^{\intercal} \\
&= \begin{bmatrix}\mathbf{A}_x & \mathbf{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\mathbf{A}_x^{\intercal} \\ \mathbf{A}_y^{\intercal}\end{bmatrix} \\
&= \mathbf{A}_x\mathbb{V}[x]\mathbf{A}_x^{\intercal} + \mathbf{A}_y\mathbb{V}[y]\mathbf{A}_y^{\intercal} \\
&+ \underbrace{\mathbf{A}_y\mathrm{Cov}[y,x]\mathbf{A}_x^{\intercal}}_{=0 \textit{by independence}} + \underbrace{\mathbf{A}_x\mathrm{Cov}[x,y]\mathbf{A}_y^{\intercal}}_{=0 \textit{by independence}} \\
&= \mathbf{A}_x\Sigma_x\mathbf{A}_x^{\intercal} + \mathbf{A}_y\Sigma_y\mathbf{A}_y^{\intercal}
\end{aligned}
$$

**Note**

Can also be proofed by using the normal definition of [def. 32.15] and tedious computations.

**Proof 32.17.** *Equation (32.43) If* $\mathbf{x} = c_i$ *i.e. the outcome* $c_i$ *has occurred then it follows:*

$$
\prod_{j}^{k} \mathrm{p}_i^{\delta[x=c_i]} = \mathrm{p}_1^0 \cdots \mathrm{p}_i^1 \cdots \mathrm{p}_k^0 = 1 \cdots \mathrm{p}_i \cdots 1 = p(\mathbf{x} = c_i|\mathbf{p})
$$

# Sampling Methods

## 1. Sampling Random Numbers

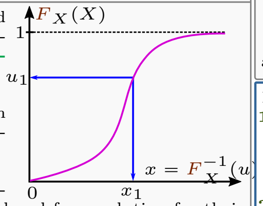
Most math libraries have uniform **random number generator** (**RNG**) i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a, b]$ (eq. (32.52)).
Furthermore repeated calls to these RNG are independent, that is:

$$\begin{aligned} p_{U_1,U_2}(u_1, u_2) &\overset{??}{=} p_{U_1}(u_1) \cdot p_{U_2}(u_2) \\ &= \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

**Question**: using samples $\{u_1, \ldots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discreet or continuous PDFs?

## 2. Inverse-transform Technique

**Idea**



Can make use of section 1 and the fact that CDF are increasing functions ([def. 22.14]). **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.

### 2.1. Continuous Case

**Definition 33.1 One Continuous Variable: Given**: a desired continuous pdf $f_X$ and uniformly distributed rn $\{u_1, u_2, \ldots\}$:

1. Integrate the desired pdf $f_X$ in order to obtain the desired cdf $F_X$:
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \qquad (33.1)$$
2. Set $F_X(X) \overset{!}{=} U$ 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:
$$U = F_X(X) = F_X\Big(\underbrace{F_X^{-1}(U)}_{X}\Big) \qquad (33.2)$$
4. Plug in the uniformly distributed rn:
$$x_i = F_X^{-1}(u_i) \qquad \textbf{s.t.} \qquad x_i \sim f_X \qquad (33.3)$$

**Definition 33.2 Multiple Continuous Variable:**
**Given**: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \qquad (33.4)$$
2. Use [def. 33.3] to first get a rv for $y$ of $Y \sim f_Y(y)$.
3. Then with this fixed $y$ use [def. 33.3] again to get a value for $x$ of $X \sim f_{X|Y}(x|y)$.

**Proof 33.1.** *[def. 33.3]:*
**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. 22.14]).*
*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 $x = \underline{F_X^{-1}(u)}$.*
**Now** *let $a$ arbitrary:*

$$F_X(a) = \mathbb{P}(\underline{x} \leqslant a) = \mathbb{P}(F_X^{-1}(U) \leqslant a)$$

*Since $F_X$ is strictly increasing:*

$$\mathbb{P}\left(F_X^{-1}(U) \leqslant a\right) = \mathbb{P}(U \leqslant F_X(a)) \overset{\text{eq. (32.52)}}{=} \int_0^{F_X(a)} 1\,\mathrm{d}t = F_X(a)$$

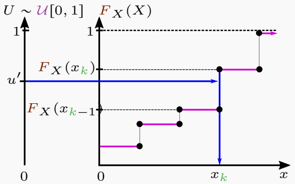
**Note**

Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing ([def. 22.14]) we may always define an auxiliary function by its infinimum:

$$\hat{F}_X^{-1} := \inf\{x | F_X(X) \geqslant 0\} \qquad u \in [0, 1] \qquad (33.5)$$

### 2.2. Discret Case

**Idea**



**Given**: a desired discret pmf $p_X$ s.t. $\mathbb{P}(X = x_i) = p_X(x_i)$ and uniformly distributed rn $\{u_1, u_2, \ldots\}$.
**Goal**: given a uniformly distributed rn $u$ determine $k$ s.t.:

$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff F_X(x_{k-1}) < u \leqslant F_X(x_k) \qquad (33.6)$$

and return $x_k$.

**Definition 33.3 One Discret Variable:**

1. Compute the CDF of $p_X$ ([def. 32.8])
$$F_X(x) = \sum_{t=-\infty}^{x} p_X(t) \qquad (33.7)$$
2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find $k^i$ ($\triangleq$ inversion) s.t.:
$$F_X\left(x_{k(i)-1}\right) < u_i \leqslant F_X\left(x_{k(i)}\right) \qquad \forall u_i \qquad (33.8)$$

**Proof 33.2.** **??**: *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:*

$$F_X(x_{k-1}) < u \leqslant F_X(x_k) \qquad (33.9)$$

**Now** *observe that:*

$$\begin{aligned} u \leqslant F_X(x_k) &= F_X(x_{k-1}) + p_X(x_k) \\ \Rightarrow F_X(x_{k-1}) < u &\leqslant F_X(x_{k-1}) + p_X(x_k) \end{aligned}$$

*The probability of $U$ being in $(F_X(x_{k-1}), F_X(x_k)]$ is:*

$$\begin{aligned} \mathbb{P}(U \in [F_X(x_{k-1}), F_X(x_k)]) &= \int_{F_X(x_{k-1})}^{F_X(x_k)} p_U(t)\,\mathrm{d}t \\ = \int_{F_X(x_{k-1})}^{F_X(x_k)} 1\,\mathrm{d}t &= \int_{F_X(x_{k-1})}^{F_X(x_{k-1})+p_X(x_k)} 1\,\mathrm{d}t = p_X(x_k) \end{aligned}$$

**Hence** *the random variable $x_k \in \mathcal{X}$ has the pdf $p_X$.*

**Definition 33.4**
**Multiple Continuous Variables (Option 1):**
**Given**: a pdf of multiple rvs $p_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $p_{X,Y}$:
$$p_{X,Y} = p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y) \qquad (33.10)$$
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 33.5**
**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**
$$\begin{aligned} &p_Z(1) = p_{X,Y}(1, 1), p_Z(2) = p_{X,Y}(1, 2), \ldots \\ &\ldots, p_Z(N_x \cdot N_y) = p_{X,Y}(N_x, N_y) \end{aligned}$$

Then simply apply **??** to the auxillary pdf $p_Z$

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \qquad (33.11)$$
2. Use [def. 33.3] to first get a rv for $y$ of $Y \sim f_Y(y)$.
3. Then with this fixed $y$ use [def. 33.3] again to get a value for $x$ of $X \sim f_{X|Y}(x|y)$.

nice examples see comment in code text

## 3. Monte Carlo Methods

### 3.1. Monte Carlo (MC) Integration

Integration methods s.a. Simpson integration[def. 28.27] suffer heavily from the curse of dimensionality.
An n-order[def. 28.24] quadrature scheme $\mathcal{Q}_n$ in 1-dimension is usually of order $n/d$ in d-dimensions.
**Idea** estimate an integral stochastically by drawing sample from some distribution.

**Definition 33.6 Monte Carlo Integration:**

$$3 + 4 \qquad (33.12)$$

### 3.2. Rejection Sampling

### 3.3. Importance Sampling

# Descriptive Statistics

## 1. Populations and Distributions

**Definition 34.1 Population** $\{x_i\}_{i=1}^N$**:**
is the entire set of entities from which we can draw sample.

**Definition 34.2**
**Families of Probability Distributions** $p_\theta$**:**
Are probability distributions that vary only by a set of hyper parameters $\theta$[def. 34.1].

**Definition 34.3** [example 34.3]
**Population/Statistical Parameter** $\theta$**:**
Are the parameters defining families of probability distributions[def. 34.2]

**Explanation 34.1** (Definition 34.1). *Such hyper parameters are often characterized by populations following a certain family of distributions with the help of a stastistc. Hence they are called population or statistical parameters.*

### 1.1. Characteristics of Populations

**Definition 34.4 Population Mean:** Given a population $\{x_i\}_{i=1}^N$ of size $N$ its variance is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^N x_i \tag{34.1}$$

**Definition 34.5 Population Variance:** Given a population $\{x_i\}_{i=1}^N$ of size $N$ its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2 \tag{34.2}$$

**Note**

The population variance and mean are equally to the mean derived from the true distribution of the population.

## 2. Sample Statistics

**Definition 34.6 (Sample) Statistic:** A statistc is a measuarble function $T$ that assigns a **single** value $t$ to a sample of random variables or population:

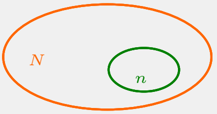
$$t : \mathbb{R}^n \mapsto \mathbb{R} \qquad t = T(X_1, \dots, X_n)$$

E.g. $T$ could be the mean, variance,...

**Definition 34.7 Degrees of freedom of a Statistic:** Is the number of values in the final calculation of a statistic that are free to vary.

**Note**

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



## 3. Point and Interval Estimation

Assume a population $X$ with a given sample $\{x_i\}_{i=1}^n$ follows some family of distributions:

$$X \sim p_X(;\theta) \tag{34.3}$$

how can we estimate the correct value of the parameter $\theta$ or some function of that parameter $\tau(\theta)$?

### 3.1. Point Estimates

**Definition 34.8 (Point) Estimator** $\hat{\theta}$**:**
Is a statistic[def. 34.6] that tries estimates an unknown parameter $\theta$ of an underlying family of distributions[def. 34.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution:

$$\hat{\theta} = t(\mathbf{x}_1, \dots, \mathbf{x}_n) \tag{34.4}$$

**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 $\theta$.
The most prevalent forms of interval estimation are:

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

#### 3.1.1. Empirical Mean

**Definition 34.9 Sample/Empirical Mean** $\bar{x}$**:**
The sample mean is an estimate/statistic of the population mean[def. 34.4] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n x_i \tag{34.5}$$

**Corollary 34.1** [proof 34.1]
**Unbiased Sample Mean:**
The sample mean estimator is unbiased:

$$\mathbb{E}[\hat{\mu}_X] = \mu \tag{34.6}$$

**Corollary 34.2** [Proof 34.2]
**Variance of the Sample Mean:**
The variance of the sample mean estimator is given by:

$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n}\sigma_X^2 \tag{34.7}$$

#### 3.1.2. Empirical Variance

**Definition 34.10 Biased Sample Variance:**
The sample variance is an estimate/statistic of the population variance[def. 34.5] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2 \tag{34.8}$$

**Definition 34.11** [proof 34.3]
**(Unbiased) Sample Variance:**
The unbiased form of the sample variance[def. 34.10] is given by:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)^2 \tag{34.9}$$

**Definition 34.12 Bessel's Correction:** The factor

$$\frac{n}{n-1} \tag{34.10}$$

is called Bessel's correction. Multiplying the uncorrected population variance eq. (34.8) by this term yields an unbiased estimated of the variance.

**Attention:**

- The Bessel correction holds for the variance but not for the standard deviation.
- Usually only the unbiased variance is used and sometimes also denoted by $s_n^2$

### 3.2. Interval Estimates

**Definition 34.13 Interval Estimator** $\hat{\theta}$**:**
Is an estimator that tries to bound an unknown parameter $\theta$ of an underlying family of distributions[def. 34.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution.
Let $\theta \in \Theta$ and define two point statistics[def. 34.6] $g$ and $h$ then an interval estimate is defined as:

$$\mathbb{P}(L_n < \theta < U_n) = \gamma \qquad \begin{matrix} \forall\theta \in \Theta & L_n = g(\mathbf{x}_1, \dots, \mathbf{x}_n) \\ \gamma \in [0,1] & U_n = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \end{matrix} \tag{34.11}$$

# Statistical Tests

## 4. Parametric Hypothesis Testing

**Definition 34.14 Parametric Hypothesis Testing:**
Hypothesis testing is a statistical procedure in which a hypothesis is tested based on sampled data $X_1, \dots, X_n$.

### 4.1. Null Hypothesis

**Definition 34.15 Null Hypothesis** $H_0$**:**
A null hypothesis $H_0$ is an *assumption* on a population[def. 34.1] parameter[def. 34.3] $\theta$:

$$H_0 : \theta = \theta_0 \tag{34.12}$$

**Note**

Often, a null hypothesis cannot be verified, but can only be falsified.

**Definition 34.16 Alternative Hypothesis** $H_A/H_1$**:**
The alternative hypothesis $H_1$ is an *assumption* on a population[def. 34.1] parameter[def. 34.3] $\theta$ that is opposite to the null hypothesis.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{34.13}$$

### 4.2. Test Statistic

The decision on the hypothesis test is based on a sample from the population $X(n) = \{X_1, \dots, X_n\}$ however the decision is usually not based on single sample but a sample statistic[def. 34.6] as this is easier to use.

**Definition 34.17** [example 34.4]
**Test Statistic/Testing Parameter** $T$**:**
Is a sample statistic[def. 34.6] used for hypothesis tests in order to give evidence for or against a hypothesis:

$$t_n = T(D_n) = T(\{X_1, \dots, X_n\}) \tag{34.14}$$

### 4.3. Sampling Distribution

**Definition 34.18** $T_{\theta_0}(t)$
**Null Distribution/Sampling Distribution under** $H_0$**:**
Let $D_n = \{X_1, \dots X_n\}$ be a random sample from the true population $p_{\text{pop}}$ and let $T(D_n)$ be a test statistic of that sample.
The probability distribution of the test statistic under the assumption that the null hypothesis is true is called *sampling distribution*:

$$t \sim T_{\theta_0} = T(t|H_0 \text{ true}) \qquad X_i \sim p_{\text{pop}} \tag{34.15}$$

### 4.4. The Critical Region

Given a sample $D_n = \{X_1, \dots, X_n\}$ of the true population $p_{\text{pop}}$ how should we decide whether the null hypothesis should be rejected or not?
**Idea:** let $\mathcal{T}$ be the be the set of all possible values that the sample statistic $T$ can map to. Now lets split $\mathcal{T}$ in two disjunct sets $\mathcal{T}_0$ and $\mathcal{T}_1$:

$$\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \qquad \mathcal{T}_0 \cap \mathcal{T}_1 = \varnothing$$

- if $t_n = T(X_n) \in \mathcal{T}_0$ we accept the null hypothesis $H_0$
- if $t_n = T(X_n) \in \mathcal{T}_1$ we reject the null hypothesis for $H_1$

**Definition 34.19 Critical/Rejection Region** $\mathcal{T}_1$**:** Is the set of all values of the test statistic[def. 34.17] $t_n$ that causes us to reject the Null Hypothesis in favor for the alternative hypothesis $H_A$:

$$K = \mathcal{T}_1 = \{\mathcal{T} : H_0 \text{ rejected}\} \tag{34.16}$$

**Definition 34.20 Acceptance Region** $\mathcal{T}_0$**:** Is the region where we accept the null hypothesis $H_0$.

$$\mathcal{T}_0 = \{\mathcal{T} : H_0 \text{ accepted}\} \tag{34.17}$$

**Definition 34.21 Critical Value** c**:**
Is the value of *the critical region* $c \in \mathcal{T}_1$ which is closest to the *region of acceptance*[def. 34.20]:

### 4.5. Type I&II Errors

**Definition 34.22**
**False Positive** **Type I Error:**
Is the rejection of the null hypothesis $H_0$, even-tough it is true

$$\begin{aligned} &\text{Test rejects } H_0|H_0 \text{ true} \\ \iff\; &t_n \in \mathcal{T}_1|H_0 \text{ true} \end{aligned} \tag{34.18}$$

**Definition 34.23**
**False Negative** **Type II Error:**
Is the acceptance of a null hypothesis $H_0$, even-tough its false:

$$\begin{aligned} &\text{Test accepts } H_0|H_A \text{ true} \\ \iff\; &t_n \in \mathcal{T}_0|H_A \text{ true} \end{aligned} \tag{34.19}$$

**Types of Errors**

| Decision | $H_0$ **true** | $H_0$ **false** |
|---|---|---|
| **Accept** | 🟩 **TN** | 🟥 **Type II (FN)** |
| **Reject** | 🟥 **Type I (FP)** | 🟩 **TP** |

### 4.6. Statistical Significance & Power

**Question:** how should we choose the split $\{\mathcal{T}_0, \mathcal{T}_1\}$?
The bigger we choose $\Theta_1$ (and thus the smaller $\Theta_0$) the more likely it is to accept the alternative.
**Idea:** take the position of the adversary and choose $\Theta_1$ so small that $\theta \in \Theta_1$ has only a small *probability* of occurring.

**Definition 34.24** [example 34.5]
**(Statistical) Significance** $\alpha$**:**
A study's defined significance level $\alpha$ denotes the probability to incur a *Type I Error*[def. 34.22]:

$$\mathbb{P}(t_n \in \mathcal{T}_1|H_0 \text{ true}) = \mathbb{P}(\text{test rejects } H_0|H_0 \text{ true}) \leqslant \alpha \tag{34.20}$$

**Definition 34.25 Probability Type II Error** $\beta$**:**
A test probability to for a *false negative*[def. 34.23] is defined as:

$$\beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_0|H_1 \text{ true}) = \mathbb{P}(\text{test accepts } H_0|H_1 \text{ true}) \tag{34.21}$$

**Definition 34.26 (Statistical) Power** $1-\beta$**:**
A study's power $1-\beta$ denotes a tests probability for a *true positive*:

$$\begin{aligned} 1-\beta(t_n) &= \mathbb{P}(t_n \in \mathcal{T}_1|H_1 \text{ true}) && (34.22) \\ &= \mathbb{P}(\text{test rejects } H_0|H_1 \text{ true}) && (34.23) \end{aligned}$$

**Corollary 34.3 Types of Split:**
The Critical region is chosen s.t. we incur a Type I Error with probability less than $\alpha$, which corresponds to the type of the test[def. 34.16]:

$$\begin{aligned} &\mathbb{P}(c_2 \leqslant X \leqslant c_1) \leqslant \alpha && \text{two-sided} \\ \text{or} \quad &\mathbb{P}(c_2 \leqslant X) \leqslant \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leqslant c_1) \leqslant \frac{\alpha}{2} \\ &\mathbb{P}(c_2 \leqslant X) \leqslant \alpha && \text{one-sided} \\ &\mathbb{P}(X \leqslant c_1) \leqslant \alpha && \text{one-sided} \end{aligned}$$

| Truth<br>Decission | 🟩 $H_0$ **true** | 🟥 $H_0$ **false** |
|---|---|---|
| $H_0$ **accept** | $1-\alpha$ | $1-\beta$ |
| $H_0$ **rejected** | $\alpha$ | $\beta$ |

### 4.7. P-Value

**Definition 34.27 P-Value** $p$**:**
Given a test statistic $t_n = T(X_1, \dots, X_n)$ the p-value $p \in [0,1]$ is the smallest value s.t. we reject the null hypothesis:

$$p := \inf_\alpha \{\alpha | t_n \in \mathcal{T}_1\} \qquad t_n = T(X_1, \dots, X_n) \tag{34.24}$$

**Explanation 34.2.**

- *The smaller the p-value the less likely is an observed statistic $t_n$ and thus the higher is the evidence against a null hypothesis.*
- *A null hypothesis has to be rejected if the p-value is bigger than the chosen significance niveau $\alpha$.*

## 5. Conducting Hypothesis Tests

1. Select an appropriate test statistic[def. 34.17] $T$.
2. Define the null hypothesis $H_0$ and the alternative hypothesis $H_1$ for $T$.
3. Find the sampling distribution[def. 34.18] $T_{\theta_0}(t)$ for $T$, given $H_0$ true.
4. Chose the significance level $\alpha$
5. Evaluate the test statistic $t_n = T(X_1, \dots, X_n)$ for the sampled data.
6. Determine the p-value $p$.
7. Make a decision (accept or reject $H_0$)

### 5.1. Tests for Normally Distributed Data

Let us consider an i.i.d. sample of observations $\{x_i\}_{i=1}^n$, of a normally distributed population $X_{\text{pop}} \sim \mathcal{N}(\mu, \sigma^2)$.
From eqs. (34.6) and (34.7) it follows that the *mean of the sample* is distributed as:

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

thus the mean of the sample $\overline{X}_n$ should equal the mean $\mu$ of the population. We now want to test the null hypothesis:

$$H_0 : \mu = \mu_0 \qquad \iff \qquad \overline{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n) \tag{34.25}$$

This is obviously only likely if the realization $\bar{x}_n$ is close to $\mu_0$.

#### 5.1.1. Z-Test $\sigma$ known

**Definition 34.28 Z-Test:**
For a realization of $Z$ with $\{x_i\}_{i=1}^n$ and mean $\bar{x}_n$:

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

we *reject the null hypothesis* $H_0 : \mu = \mu_0$ for the alternative $H_A$ for significance niveau[def. 34.24] $\alpha$ if:

$$\begin{aligned}
|z| \geqslant z_{1-\frac{\alpha}{2}} \quad &\iff z \leqslant z_{\frac{\alpha}{2}} \vee z \geqslant z_{1-\frac{\alpha}{2}} \\
&\iff z \in \mathcal{T}_1 = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, \infty\right) \\
z \geqslant z_{1-\alpha} \quad &\iff z \in \mathcal{T}_1 = \left[z_{1-\alpha}, \infty\right) \\
z \leqslant z_{\alpha} = -z_{1-\alpha} \quad &\iff z \in \mathcal{T}_1 = (-\infty, -z_{\alpha}] = (\infty, -z_{1-\alpha}]
\end{aligned} \tag{34.26}$$

**Notes**

- Recall from [def. 32.19] and [cor. 32.4] that:
  $z_{\alpha} \overset{\text{i.e. } \alpha=0.05}{=} z_{0.05} = \Phi^{-1}(\alpha) \iff \mathbb{P}(Z \leqslant z_{0.05}) = 0.05$
- $|z| \geqslant z_{1-\frac{\alpha}{2}}$ which stands for:
  $$\mathbb{P}(Z \leqslant z_{0.05}) + \mathbb{P}(Z \geqslant z_{0.95}) = \mathbb{P}(Z \leqslant -z_{1-0.05}) + \mathbb{P}(Z \geqslant z_{0.95}) \\ = \mathbb{P}(|Z| \geqslant z_{0.95})$$
  can be rewritten as:
  $$z \geqslant z_{1-\frac{\alpha}{2}} \quad \vee \quad -z \geqslant z_{1-\frac{\alpha}{2}} \iff z \leqslant -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$$
- One usually goes over to the standard normal distribution proposition 32.2 and thus test how far one is away from zero mean ⇒ Z-test.
- We thus inquire a Type I error with probability $\alpha$ and should be small i.e. 1%.

#### 5.1.2. t-Test $\sigma$ unknown

In reality we usually do not know the true $\sigma$ of the whole data set and thus calculate it over our sample. This however increases uncertainty and thus our sample does no longer follow a normal distribution but a **t-distribution** wiht $n-1$ degrees of freedom:

$$T \sim t_{n-1} \tag{34.27}$$

**Definition 34.29 t-Test:**
For a realization of $T$ with $\{x_i\}_{i=1}^n$ and mean $\bar{x}_n$:

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

*we reject the null hypothesis* $H_0 : \mu = \mu_0$ for the alternative $H_A$ if:

$$\begin{aligned}
&|t| \geqslant t_{n-1,1-\frac{\alpha}{2}} \\
&\qquad \iff t \in \mathcal{T}_1 = \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\
&t \geqslant t_{n-1,1-\alpha} \\
&\qquad \iff t \in \mathcal{T}_1 = \left[t_{n-1,1-\alpha}, \infty\right) \\
&t \leqslant t_{n-1,\alpha} = -t_{n-1,1-\alpha} \\
&\qquad \iff t \in \mathcal{T}_1 = (-\infty, -t_{n-1,\alpha}] = (\infty, -t_{n-1,1-\alpha}]
\end{aligned}$$

**Notes**

- The t-distribution has fatter tails as the normal distribution ⇒ rare event become more likely
- For $n \to \infty$ the t-distribution goes over into the normal distribution
- The t-distribution gains a degree of foredoom for each sample and loses one for each parameter we are interested in ⇒ $n$-samples and we are interested in one parameter $\mu$.

### 5.2. Confidence Intervals

Now we are interested in the opposite of the critical region[def. 34.19] namely the region of plaussible values.

**Definition 34.30 Confidence Interval $I$:**
Let $D_n = \{X_1, \dots, X_n\}$ be a *sample* of observations and $T_n$ a sample statistic of that sample. The confidence interval is defined as:

$$I(D_n) = \{\theta_0 : T_n(D_n) \in \mathcal{T}_0\} = \{\theta_0 : H_0 \text{ is not rejected}\} \tag{34.28}$$

**Corollary 34.4 :** The confidence interval captures the unkown parameter $\theta$ with probability $1 - \alpha$:

$$\mathbb{P}_{\theta}(\theta \in I(D_n)) = \mathbb{P}(T_n(D_n) \in \mathcal{T}_0) = 1 - \alpha \tag{34.29}$$

add page 91 confidence intervals z-test and t-test

## 6. 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?

## 7. Examples

**Example 34.1 ??: Let** $x$ be uniformly distributed on $[0, 1]$ ([def. 32.27]) with pmf $\mathrm{p}_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{p}_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y\mathrm{p}_y(y) \Rightarrow x = \int_{-\infty}^{y} \mathrm{p}_y(t)\,\mathrm{d}t = F_Y(x)$$

**Example 34.2 ??: Let**

add https://www.youtube.com/watch?v=WUUb7VIRzgg

**Example 34.3 Family of Distributions:** The family of normal distribution $\mathcal{N}$ has two parameters $\left\{\mu, \sigma^2\right\}$

**Example 34.4 Test Statistic:** Lets assume the test statistic follows a normal distribution:

$$T \sim \mathcal{N}(\mu; 1)$$

however we are unsure about the population parameter[def. 34.3] $\theta = \mu$ but assume its equal to $\theta_0$ thus the null-and alternative hypothesis are:

$$H_0 : \mu = \mu_0 \qquad\qquad H_1 : \mu \neq \mu_0$$

**Example 34.5 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 what the manufacture claims is true, so we want to test:

$$H_0 : \mathrm{p} = \mathrm{p}_0 = 0.1 \qquad \text{vs.} \qquad H_A : \mathrm{p} > 0.1$$

We model the number of number of defective goods using the binomial distribution[def. 32.24]

$$\begin{aligned} X \sim&\mathcal{B}(n, \mathrm{p}) \\ \sim&T(n, \mathrm{p}) \end{aligned}, n = 20 \qquad \mathbb{P}(X \geqslant x) = \sum_{k=x}^{n} \binom{n}{k} \mathrm{p}^k (1-\mathrm{p})^{n-k}$$

from this we find:

$$\mathbb{P}_{\mathrm{p}_0}(X \geqslant 4) = 1 - \mathbb{P}_{\mathrm{p}_0}(X \leqslant 3) = 0.13$$
$$\mathbb{P}_{\mathrm{p}_0}(X \geqslant 5) = 1 - \mathbb{P}_{\mathrm{p}_0}(X \leqslant 4) = 0.04 \leqslant \alpha$$

thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
⇒ throw away null hypothesis for the 5% niveau in favor to the alternative.
⇒ the 5% significance niveau is given by $K = \{5, 6, \ldots, 20\}$

**Note**

If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X \leqslant x-1)$

## 8. Proofs

**Proof 34.1.** *[cor. 34.1]:*

$$\mathbb{E}\left[\hat{\mu}_X\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}[\underbrace{\mu + \cdots + \mu}_{1,\ldots,n}]$$

**Proof 34.2.** *[cor. 34.2]:*

$$\mathbb{V}\left[\hat{\mu}_X\right] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \overset{\textit{Property } 32.10}{=} \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right]$$
$$\frac{1}{n^2} n\mathbb{V}[X] = \frac{1}{n}\sigma^2$$

**Proof 34.3.** *definition 34.11:*

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_X^2\right] =&\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
=&\frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
=&\frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
=&\frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\cdot n\bar{x} + n\bar{x}^2\right] \\
=&\frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
=&\frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
=&\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
=&\frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\
=&\frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \left(\sigma^2 + n\mu^2\right)\right] \\
=&\frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}$$

# Stochastic Calculus

## Stochastic Processes

**Definition 35.1**
**Random/Stochastic Process** $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$:
An ($\mathbb{R}^d$-valued) stochastic process is a collection of ($\mathbb{R}^d$ *−valued*) random variables $X_t$ 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 $\{t_1, t_2, \ldots\}$. Therefore, the random process $X$ can be written as a function:

$$X : \mathcal{T} \subseteq \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}^d \quad \iff \quad (t, \omega) \mapsto X(t, \omega) \qquad (35.1)$$

**Definition 35.2 Sample path/Trajector/Realization:** Is the *stochastic/noise signal* $r(\cdot, \omega)$ on the index set[def. 20.1] $\mathcal{T}$, that we obtain be sampling $\omega$ from $\Omega$.

**Notation**

Even though the r.v. $X$ is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$

**Corollary 35.1** $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\} > 0$
**Strictly Positive Stochastic Processes:** A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called strictly positive if it satisfies:

$$X_t > 0 \qquad \mathbb{P}\text{-a.s.} \qquad \forall t \in \mathcal{T} \qquad (35.2)$$

**Definition 35.3**
**Random/Stochastic Chain** $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$:
is a collection of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$[def. 31.1]. The random variables are ordered by an associated index set[def. 20.1] $\mathcal{T}$ and take values in the same mathematical *discrete state space*[def. 35.5] S, which must be measurable w.r.t. some $\sigma$-algebra[def. 31.6] $\Sigma$.
Therefore for a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space $(S, \Sigma)$, the random *chain* $X$ is a collection of $S$-valued random variables that can be written as:

$$X : \mathcal{T} \times \Omega \mapsto S \quad \iff \quad (t, \omega) \mapsto X(t, \omega) \qquad (35.3)$$

**Definition 35.4 Index/Parameter Set** $\mathcal{T}$:
Usually represents time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \ldots\}$.

**Definition 35.5 State Space** $S$:
Is the range/possible values of the random variables of a stochastic process[def. 35.1] and must be measurable[def. 31.7] w.r.t. some $\sigma$-algebra $\Sigma$.

**Sample-vs. State Space**

Sample space[def. 31.2] hints that we are working with probabilities i.e. probability measures will be defined on our sample space.
State space is used in dynamics, it implies that there is a time progression, and that our system will be in different states as time progresses.

**Definition 35.6 Sample path/Trajector/Realization:** Is the *stochastic/noise signal* $r(\cdot, \omega)$ on the index set $\mathcal{T}$, that we obtain be sampling $\omega$ from $\Omega$.

**Notation**

Even though the r.v. $X$ is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$

### 1.1. Filtrations

**Definition 35.7 Filtration** $\mathbb{F} = \{\mathcal{F}_t\}_{t \geqslant 0}$:
A collection $\{\mathcal{F}_t\}_{t \geqslant 0}$ of sub *$\sigma$-algebras*[def. 31.6] $\{\mathcal{F}_t\}_{t \geqslant 0} \in \mathcal{F}$ is called filtration if it is *increasing*:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \qquad \forall s \leqslant t \qquad (35.4)$$

**Explanation 35.1** (Definition 35.7). *A filtration describes the flow of information i.e. with time we learn more information.*

**Definition 35.8**
**Filtered Probability Space** $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geqslant 0}, \mathbb{P})$:
A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t \geqslant 0}$ is called a *filtered probability* space.

**Definition 35.9 Adapted Process:** A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called adapted *to a* filtration $\mathbb{F}$ if:

$$X_t \text{ is } \mathcal{F}_t\text{-measurable} \qquad \forall t \qquad (35.5)$$

That is the value of $X_t$ is observable at time $t$

**Definition 35.10 Predictable Process:** A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called predictable *w.r.t. a* filtration $\mathbb{F}$ if:

$$X_t \text{ is } \mathcal{F}_{t-1}\text{-measurable} \qquad \forall t \qquad (35.6)$$

That is 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 $r = (r_k)_{k=1,\ldots,T}$ should be predictable.

**Corollary 35.2 :** The amount of information of an adapted random process is increasing see example 35.1.

## 2. Martingales

**Definition 35.11 Martingales:** A stochastic process $X(t)$ is a martingale on a *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geqslant 0}, \mathbb{P})$ if the following conditions hold:

1. Given $s \leqslant t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value:
$$\forall s \leqslant t \qquad \mathbb{E}\left[X(t) | \mathcal{F}_s\right] = X(s) \quad \text{a.s.} \qquad (35.7)$$
2. The expectation is finite:
$$\mathbb{E}\left[|X(t)|\right] < \infty \quad \forall t \geqslant 0 \qquad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geqslant 0} \text{ adapted} \qquad (35.8)$$

**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 35.12 Auto Covariance** $\gamma(t_2 - t_1)$:
Describes the covariance[def. 32.16] between two values of a stochastic process $(\mathbf{X}_t)_{t \in \mathcal{T}}$ at different time points $t_1$ and $t_2$.

$$\gamma(t_1, t_2) = \mathrm{Cov}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right] = \mathbb{E}\left[\left(\mathbf{X}_{t_1} - \mu_{t_1}\right)\left(\mathbf{X}_{t_2} - \mu_{t_2}\right)\right] \qquad (35.9)$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \mathrm{Cov}\left[\mathbf{X}_t, \mathbf{X}_t\right] \overset{\text{eq. (32.35)}}{=} \mathbb{V}\left[\mathbf{X}_t\right] \qquad (35.10)$$

**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 35.13 Auto Correlation** $\rho(t_2 - t_1)$:
Is the scaled version of the auto-covariance[def. 35.12]:

$$\rho(t_2 - t_1) = \mathrm{Corr}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right] \qquad (35.11)$$
$$= \frac{\mathrm{Cov}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}\left[\left(\mathbf{X}_{t_1} - \mu_{t_1}\right)\left(\mathbf{X}_{t_2} - \mu_{t_2}\right)\right]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}}$$

## 3. Different kinds of Processes

### 3.1. Markov Process

**Definition 35.14 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:

$$\mathbb{P}\left(X(t_{n+1}) \in B | X(t_1), \ldots, X(t_n)\right) = \mathbb{P}\left(X(t_{n+1}) \in B | X(t_n)\right)$$

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 $s$:

$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \quad 0 \leqslant s < t \qquad (35.12)$$

**Interpretation**

In order to predict the future only the current/last value counts.

**Corollary 35.3 Transition Density:** The transition probability of a continuous distribution $\mathrm{p}$ can be calculated via:

$$\mathbb{P}(s, x, t, B) = \int_B \mathrm{p}(s, x, t, y)\, \mathrm{d}y \qquad (35.13)$$

### 3.2. Gaussian Process

**Definition 35.15 Gaussian Process:** Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:

$$X(t) \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right) \quad \forall t \in T \qquad (35.14)$$

### 3.3. Diffusions

**Definition 35.16 Diffusion:** Is a Markov Process[def. 35.14] for which it holds that:

$$\mu(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}\left[X(t + \Delta t) - X(t) | X(t)\right] \qquad (35.15)$$
$$\sigma^2(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}\left[\left(X(t + \Delta t) - X(t)\right)^2 | X(t)\right] \qquad (35.16)$$

See **??**/eq. (35.16) for simple proof of eq. (35.15)/**??**.

- $\mu(t, X(t))$ is called **drift**
- $\sigma^2(t, X(t))$ is called **diffusion coefficient**

**Interpretation**

There exist not discontinuities for the trajectories.

### 3.4. Brownian Motion/Wienner Process

**Definition 35.17**
***d*-dim standard Brownian Motion/Wienner Process:**
Is an $\mathbb{R}^d$ valued *stochastic process*[def. 35.1] $(W_t)_{t \in \mathcal{T}}$ starting at $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies:

1. **Normal Independent Increments**: the increments are *normally distributed independent random variables*:
$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}\left(0, (t_i - t_{i-1})\mathbb{1}_{d \times d}\right) \quad \forall i \in \{1, \ldots, T\} \qquad (35.17)$$
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
$$\lim_{t \to 0} \frac{\mathbb{P}\left(|W(t + \Delta t) - W(t)| \geqslant \delta\right)}{\Delta t} = 0 \qquad \forall \delta > 0 \qquad (35.18)$$
4. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \qquad (35.19)$$

**Notation**

- In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wienner process.
- **However** in some sources the Wienner process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

**Corollary 35.4** $W_t \sim \mathcal{N}(0, \sigma)$:
The random variable $W_t$ follows the $\mathcal{N}(0, \sigma)$ law

$$\mathbb{E}\left[W(t)\right] = \mu = 0 \qquad (35.20)$$
$$\mathbb{V}\left[W(t)\right] = \mathbb{E}\left[W^2(t)\right] = \sigma^2 = t \qquad (35.21)$$

See **??** 35.5

#### 3.4.1. Properties of the Wienner Process

**Property 35.1 Non-Differentiable Trajectories:**
The sample paths of a Brownian motion are not differentiable:

$$\frac{\mathrm{d}W(t)}{t} = \lim_{t \to 0} \mathbb{E}\left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t}\right)^2\right]$$
$$= \lim_{t \to 0} \frac{\mathbb{E}\left[W(t + \Delta t) - W(t)\right]}{\Delta t} = \lim_{t \to 0} \frac{\sigma^2}{\Delta t} = \infty$$

$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
$\xrightarrow{\text{solution}}$ Ito Calculus see section 36.

**Property 35.2 Auto covariance Function:**
The auto-covariance[def. 35.12] for a Wienner process

$$\mathbb{E}\left[(W(t) - \mu t)(W(t') - \mu t')\right] = \min(t, t') \qquad (35.22)$$

**Property 35.3:** A standard Brownian motion is a

**Quadratic Variation**

**Definition 35.18 Total Variation:** The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$LV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)| \qquad (35.23)$$
$$\mathcal{S} = \left\{\Pi\{x_0, \ldots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{[\text{def. } 28.8]} \text{ of } [a, b]\right\}$$

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 35.19**
**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:

$$QV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)|^2 \qquad (35.24)$$
$$\mathcal{S} = \left\{\Pi\{x_0, \ldots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{[\text{def. } 28.8]} \text{ of } [a, b]\right\}$$

**Corollary 35.5 Bounded (quadratic) Variation:**
The (quadratic) variation[def. 35.18] of a function is bounded if it is finite:

$$\exists M \in \mathbb{R}_+ : \quad LV_{[a,b]}(f) \leqslant M \quad \left(QV_{[a,b]}(f) \leqslant M\right) \quad \forall \Pi \in \mathcal{S} \qquad (35.25)$$

**Theorem 35.1 Variation of Wienner Process:** Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:

$$\mathbb{P}\left(\omega : LV(W(\omega)) < \infty\right) = 0 \qquad (35.26)$$

**Theorem 35.2**
**Quadratic Variation of standard Brownian Motion:**
The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:

$$\lim_{N \to \infty} \sum_{k=1}^{N} \left[W\left(k\frac{T}{N}\right) - W\left((k - 1)\frac{T}{N}\right)\right]^2 = T$$
$$\text{with probability 1} \qquad (35.27)$$

See **??**

**Corollary 35.6 :** theorem 35.2 can also be written as:

$$(\mathrm{d}W(t))^2 = \mathrm{d}t \qquad (35.28)$$

### 3.4.2. Lévy's Characterization of BM

**Theorem 35.3**
**$d$-dim standard BM/Wienner Process by Paul Lévy:**
An $\mathbb{R}^d$ valued *adapted stochastic process*[def's. 35.1, 35.7] $(W_t)_{t\in\mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{R}_+}$, that satisfies:

① **Start**

$$W(0) := W_0 = 0 \qquad a.s. \tag{35.29}$$

② **Continuous Martingale**: $W_t$ is an a.s. *continuous* martingale[def. 35.11] w.r.t. the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ under $\mathbb{P}$.

③ **Quadratic Variation**:

$$W_t^2 - t \text{ is also an martingale} \iff QV(W_t) = t \tag{35.30}$$

is a standard Brownian motion[def. 35.24]. Proof see **??** 35.8

## Further Stochastic Processes

### 3.4.3. White Noise

undestand script and add

**Definition 35.20 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):
$$\mathbb{E}\left[\epsilon*[k]\right] = 0 \qquad \forall k \in T_{\text{discret}} \tag{35.31}$$
- Zero autocorrelation[def. 35.13] $\gamma$ i.e. the signals of different times are in no-way correlated:
$$\begin{aligned}\gamma(\epsilon*[k], \epsilon*[k+n]) &= \mathbb{E}\left[\epsilon*[k]\epsilon*[k+n]^{\intercal}\right]\\ &= \mathbb{V}\left[\epsilon*[k]\right]\delta_{\text{discret}}[n]\\ &\qquad \forall k, n \in T_{\text{discret}}\end{aligned} \tag{35.32}$$

**With** $\qquad \delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else}\end{cases}$

See proofs

**Definition 35.21 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):
$$\mathbb{E}\left[\epsilon*(t)\right] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{35.33}$$
- Zero autocorrelation[def. 35.13] $\gamma$ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon*(t), \epsilon*(t+\tau)) = \mathbb{E}\left[\epsilon*(t)\epsilon*(t+\tau)^{\intercal}\right] \tag{35.34}$$
$$\overset{\text{eq. (32.88)}}{=} \mathbb{V}\left[\epsilon*(t)\right]\delta(t-\tau) = \begin{cases}\mathbb{V}\left[\epsilon*(t)\right] & \text{if } \tau = 0\\ 0 & \text{else}\end{cases}$$
$$\forall t, \tau \in T_{\text{continuous}} \tag{35.35}$$

**Definition 35.22 Homoscedastic Noise:** Has constant variability for all observations/time-steps:

$$\mathbb{V}\left[\epsilon_{i,t}\right] = \sigma^2 \qquad \begin{aligned}&\forall t = 1, \ldots, T\\ &\forall i = 1, \ldots, N\end{aligned} \tag{35.36}$$

**Definition 35.23 Heteroscedastic Noise:** Is noise whose variability may vary with each observation/time-step:

$$\mathbb{V}\left[\epsilon_{i,t}\right] = \sigma(i,t)^2 \qquad \begin{aligned}&\forall t = 1, \ldots, T\\ &\forall i = 1, \ldots, N\end{aligned} \tag{35.37}$$

### 3.4.4. Generalized Brownian Motion

**Definition 35.24 Brownian Motion:**
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 35.17], and define:

$$X_t = \mu t + \sigma W_t \qquad t \in \mathbb{R}_+ \qquad \begin{aligned}&\mu \in \mathbb{R} &&: \text{drift parameter}\\ &\sigma \in \mathbb{R}_+ &&: \text{scale parameter}\end{aligned} \tag{35.38}$$

then $\{X_t\}_{t\in\mathbb{R}_+}$ is normally distributed with mean $\mu t$ and variance $t\sigma^2$ $X_t \sim \mathcal{N}\left(\mu t, \sigma^2 t\right)$.

**Theorem 35.4 Normally Distributed Increments:**
If $W(T)$ is a Brownian motion, then $W(t) - W(0)$ is a normal random variable with mean $\mu t$ and variance $\sigma^2 t$, where $\mu, \sigma \in \mathbb{R}$. From this it follows that $W(t)$ is distributed as:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}}\exp\left\{-\frac{(x-\mu t)^2}{2\sigma^2 t}\right\} \tag{35.39}$$

**Corollary 35.7 :** More generally we may define the process:

$$t \mapsto f(t) + \sigma W_t \tag{35.40}$$

which corresponds to a noisy version of $f$.

**Corollary 35.8**
**Brownian Motion as a Solution of an SDE:** A stochastic process $X_t$ follows a BM with drift $\mu$ and scale $\sigma$ if it satisfies the following SDE:

$$\mathrm{d}X(t) = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) \tag{35.41}$$
$$X(0) = 0 \tag{35.42}$$

### 3.4.5. 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 35.25 Geometric Brownian Motion:**
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 35.17] the exponential transform:

$$X(t) = \exp\left(W(t)\right) = \exp\left(\mu t + \sigma W(t)\right) \qquad t \in \mathbb{R}_+ \tag{35.43}$$

is called geometric Brownian motion

**Corollary 35.9 Log-normal Returns:** For a geometric BM we obtain log-normal returns:

$$\ln\left(\frac{S_t}{S_0}\right) = \mu t + \sigma W(t) \iff \mu t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \tag{35.44}$$

meaning that the mean and the variance of the process (stock) *log-returns* grow over time linearly.

**Corollary 35.10**
**Geometric BM as a Solution of an SDE:**
A stochastic process $X_t$ follows a geometric BM with drift $\mu$ and scale $\sigma$ if it satisfies the following SDE:

$$\begin{aligned}\mathrm{d}X(t) &= X(t)\left(\mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)\right)\\ &= \mu X(t)\,\mathrm{d}t + \sigma X(t)\,\mathrm{d}W(t)\end{aligned} \tag{35.45}$$
$$X(0) = 0 \tag{35.46}$$

### 3.4.6. Locally Brownian Motion

**Definition 35.26 Locally Brownian Motion:**
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 35.17] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = \mu\left(X(t), t\right)\mathrm{d}t + \sigma\left(X(t), t\right)\mathrm{d}W(t) \tag{35.47}$$

**Note**

A local Brownian motion is an generalization of a geometric Brownian motion.

### 3.4.7. Ornstein-Uhlenbeck Process

**Definition 35.27 Ornstein-Uhlenbeck Process:**
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 35.17] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = -aX(t)\,\mathrm{d}t + b\sigma\,\mathrm{d}W(t) \qquad a > 0 \tag{35.48}$$

### 3.5. Poisson Processes

**Definition 35.28 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 35.29 Poisson Process:** A Poisson Process with *rate* $\lambda \in \mathbb{R}_{\geqslant 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$, having a discrete *state space* $N = \{0, 1, 2, \ldots\}$ and satisfies:

1. $X_0 = 0$
2. The increments follow a Poisson distribution[def. 32.26]:
$$\mathbb{P}\left((X_t - X_s) = k\right) = \frac{\lambda(t-s)}{k!}\mathrm{e}^{-\lambda(t-s)} \qquad \begin{aligned}&0 \leqslant s < t < \infty\\ &\forall k \in \mathbb{N}\end{aligned}$$
3. No correlation of (non-overlapping) increments:
$$\forall t_0 < t_1 < \cdots < t_n : \text{the increments are independent}$$
$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}} \tag{35.49}$$

**Interpretation**

A Poisson Process is a *continuous-time* process with *discrete, positive* realizations in $\in \mathbb{N}_{\geqslant 0}$

**Corollary 35.11 Probability of events:** Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}\left(X_{(t+\Delta t)} - X_t \neq 0\right) = \lambda\Delta t + o(\Delta t^2) \qquad t \text{ small i.e. } t \to 0 \tag{35.50}$$

1. Thus the probability of an event happening during $\Delta t$ is proportional to time period and the rate $\lambda$
2. The probability of two or more events to happen *during* $\Delta t$ is of order $o(\Delta t^2)$ and thus extremely small (as *Deltat* is small).

**Definition 35.30 Differential of a Poisson Process:** The differential of a Poisson Process is defined as:

$$\mathrm{d}X_t = \lim_{\Delta t \to \mathrm{d}t}\left(X_{(t+\Delta t)} - X_t\right) \tag{35.51}$$

**Property 35.4 Probability of Events for differential:** With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\mathbb{P}\left(\mathrm{d}X_t = 0\right) = 1 - \lambda \tag{35.52}$$
$$\mathbb{P}\left(|\mathrm{d}X_t| = 1\right) = \lambda \tag{35.53}$$

## Proofs

**Proof 35.1.** *eq. (35.15):*
*Let by $\delta$ 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:*

$$\begin{aligned}\mathbb{E}\left[x(n)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1) \pm \delta\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1)\right]\\ &\overset{\text{induction}}{=} \mathbb{E}\left[x_{n-1}\right] = \ldots\mathbb{E}\left[x(0)\right] = 0\end{aligned}$$

*Thus in expectation the particles goes nowhere.*

**Proof 35.2.** *eq. (35.16):*
*Let by $\delta$ 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:*

$$\begin{aligned}\mathbb{E}\left[x(n)^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1) \pm \delta\right]^2\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2\right]\\ &\overset{\text{ind.}}{=} \mathbb{E}\left[x_{n-1}^2\right] + \delta^2 = \mathbb{E}\left[x_{n-2}^2\right] + 2\delta^2 = \ldots\\ &= \mathbb{E}\left[x(0)\right] + n\delta^2 = n\delta^2\end{aligned}$$

*as* $n = \frac{\textit{time}}{\textit{step-size}} = \frac{t}{\Delta x}$ *it follows:*

$$\sigma^2 = \mathbb{E}\left[x^2(n)\right] - \mathbb{E}\left[x(n)\right]^2 = \mathbb{E}\left[x^2(n)\right] = \frac{\delta^2}{\Delta x}t \tag{35.54}$$

*Thus in expectation the particles goes nowhere.*

**Proof 35.3.** *eq. (35.34):*

$$\begin{aligned}\gamma(\epsilon*[k], \epsilon*[k+n]) &= \mathrm{Cov}\left[\epsilon*[k], \epsilon*[k+1]\right]\\ &= \mathbb{E}\left[\left(\epsilon*[k] - \mathbb{E}\left[\epsilon*[k]\right]\right)\left(\epsilon*[k+n] - \mathbb{E}\left[\epsilon*[k+n]\right]\right)^{\intercal}\right]\\ &\overset{\text{eq. (35.31)}}{=} \mathbb{E}\left[\left(\epsilon*[k]\right)\left(\epsilon*[k+n]\right)\right]\end{aligned}$$

**Proof 35.4.** *[cor. 35.4]:*
*Since $B_t - B_s$ is the increment over the interval $[s,t]$, it is the same in distribution as the increeent over the interval $[s-s, t-s] = [0, t-s]$*

*Thus* $\qquad B_t - B_s \sim B_{t-s} - B_0$

*but as $B_0$ is a.s. zero by definition eq. (35.19) it follows:*

$$B_t - B_s \sim B_{t-s} \qquad B_{t-s} \sim \mathcal{N}(0, t-s)$$

**Proof 35.5.** *[cor. 35.4]:*

$$W(t) = W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t)$$
$$\Rightarrow \qquad \mathbb{E}\left[X\right] = 0 \qquad \mathbb{V}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = t$$

**Proof 35.6.** *theorem 35.2:*

$$\begin{aligned}&\sum_{k=0}^{N-1}\left[W\left(t_k\right) - W\left(t_{k-1}\right)\right]^2 && t_k = k\frac{T}{N}\\ &= \sum_{k=0}^{N-1} X_k^2 && X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right)\\ &= \sum_{k=0}^{N-1} Y_k = n\left(\frac{1}{n}\sum_{k=0}^{N-1} Y_k\right) && \mathbb{E}\left[Y_k\right] = \frac{T}{N}\\ &\overset{S.L.L.N}{=} n\frac{T}{n} = T\end{aligned}$$

**Proof 35.7.** *theorem 35.3 ②:*

1. *first we need to show eq. (35.7):* $\qquad \mathbb{E}\left[W_t|\mathcal{F}_s\right] = W_s$
*Due to the fact that $W_t$ is $\mathcal{F}_t$ measurable i.e. $W_t \in \mathcal{F}_t$ we know that:*
$$\mathbb{E}\left[W_t|\mathcal{F}_t\right] = W_t \tag{35.55}$$
$$\begin{aligned}\mathbb{E}\left[W_t|\mathcal{F}_s\right] &= \mathbb{E}\left[W_t - W_s + W_s|\mathcal{F}\right]\\ &= \mathbb{E}\left[W_t - W_s|\mathcal{F}_s\right] + \mathbb{E}\left[W_s|\mathcal{F}_s\right]\\ &\overset{\text{eq. (35.55)}}{=} \mathbb{E}\left[W_t - W_s\right] + W_s\\ &\overset{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s\end{aligned}$$
2. *second we need to show eq. (35.8):* $\qquad \mathbb{E}\left[|X(t)|\right] < \infty$
$$\mathbb{E}\left[|W(t)|\right]^2 \overset{??}{\leqslant} \mathbb{E}\left[|W(t)|^2\right] = \mathbb{E}\left[W^2(t)\right] = t \leqslant \infty$$

**Proof 35.8.** *theorem 35.3 ③: $W_t^2 - t$ is a martingale?*
*Using the binomial formula we can write and adding $W_s - W_s$:*

$$W_t^2 = (W_t - W_s)^2 + 2W_s\left(W_t - W_s\right) + W_s^2$$

*using the expectation:*

$$\begin{aligned}\mathbb{E}\left[W_t^2|\mathcal{F}_s\right] &= \mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}_s\right] + \mathbb{E}\left[2W_s\left(W_t - W_s\right)|\mathcal{F}_s\right]\\ &\quad + \mathbb{E}\left[W_s^2|\mathcal{F}_s\right]\\ &\overset{\text{eq. (35.55)}}{=} \mathbb{E}\left[(W_t - W_s)^2\right] + 2W_s\mathbb{E}\left[(W_t - W_s)\right] + W_s^2\\ &\overset{\text{eq. (35.21)}}{=} \mathbb{V}\left[W_t - W_s\right] + 0 + W_s^2\\ &\quad t - s + W_s^2\end{aligned}$$

*from this it follows that:*

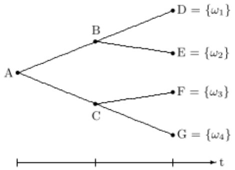
$$\mathbb{E}\left[W_t^2 - t|\mathcal{F}_s\right] = W_s^2 - s \tag{35.56}$$

*understand why* $\mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}\right] = \mathbb{E}\left[(W_t - W_s)^2\right]$

### Examples

**Example 35.1 :**

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 $\omega$ 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.



$$\mathcal{F} = \begin{cases} \{\varnothing, \Omega\} & t \in [0, T/2) \\ \{\varnothing, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^{\Omega} & t = T \end{cases} \tag{35.57}$$

Thus, $\mathcal{F}_0$ represents initial information whereas $\mathcal{F}_\infty$ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geqslant 0}, \mathbb{P})$.

## Ito Calculus