

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.

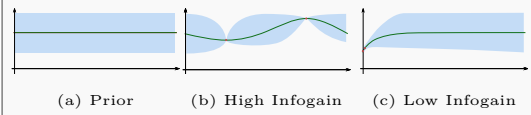
Definition 1.2 Utility Function **F:** Is a function that provide a ranking to judge uncertain situations.

1. Uncertainty Sampling for Regression

1.1. Maximizing the Information Gain

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

$$F(S) := H(f) - H(f|y_S) \stackrel{\text{eq. (4.17)}}{=} I(f; y_S) \quad (1.1)$$



Definition 1.3 Optimal Set of labels:

$$\{x_1, \dots, x_{|S|}\} = \arg \max_{S \subseteq D, |S| \leq T} F(S) \quad (1.2)$$

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:

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in D} F(S_t \cup \{x\}) \\ &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_x | f) \end{aligned} \quad (1.3)$$

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

$$x_t = \arg \max_{x \in D} \sigma_{t-1}^2(x) \quad (1.4)$$

this can then be maximized.

Let $A_t = \{x_1, \dots, x_t\}$ then it follows:

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

Algorithm 1.1 Greedy Uncertainty Sampling:

Given: $S_t := \{x_1, \dots, x_t\}$

1: **for** $t + 1 \dots T$ **do**

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in D} F(S_t \cup \{x\}) \\ &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_x | f) \end{aligned}$$

2: **end for**

Corollary 1.2 Diminishing Returns Property:

Mutal information satisfies modular submodularity (Property 4.9)

\Rightarrow adding a label/memasurement for some data point can only increase information:

$$\begin{aligned} F(A \cup \{x\}) - F(A) &\geq F(B \cup \{x\}) - F(B) \\ H(y_x | y_A) - H(y_x | f) &\geq H(y_x | y_B) - H(y_x | f) \\ \iff H(y_x | y_A) &\geq H(y_x | y_B) \end{aligned}$$

Note

For Gaussians processes the utility 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 Gaussian depends only on the covariance matrix and not the actual measurements.

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

$$F(S_T) \leq \underbrace{\left(1 - \frac{1}{e}\right)}_{\approx .63} \max_{S \subseteq D, |S| \leq T} F(S) \quad (1.6)$$

Note

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

1.2. Heteroscedastic Case

So far we considered homoscedastic noise [def. 38.22] but sometimes we may have heteroscedastic [def. 38.23] 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:**

$$x_t = \arg \max_{x \in D} \begin{matrix} \text{epistemic uncertainty} \\ \text{aleatoric uncertainty} \end{matrix} = \arg \max_{x \in D} \begin{matrix} \sigma_f^2(x) \\ \sigma_n^2(x) \end{matrix} \quad (1.7)$$

this can then be maximized.

Let $A_t = \{x_1, \dots, x_t\}$ then it follows:

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

2. Uncertainty Sampling for Classification

we now want to choose get/obtain labels for those samples that we are most unsure/uncertain about.
 \Rightarrow 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:

$$x_{t+1} = \arg \max_{x \in D} H(Y | x, x_{1:t}, y_{1:t}) = \arg \max_{x \in D} \quad (1.9)$$

$$= \arg \max_{x \in D} -\mathbf{p}(y | x, x_{1:t}, y_{1:t}) \quad (1.10)$$

$$= \arg \max_{x \in D} -\sum_y \mathbf{p}(y | x, x_{1:t}, y_{1:t}) \quad (1.11)$$

Notes

The posterior $\mathbf{p}(y | x, x_{1:t}, y_{1:t})$ 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. 38.22] but sometimes we may have heteroscedastic [def. 38.23] 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):

$$x_{t+1} = \arg \max_{\hat{x} \in D} I(\theta; \hat{y} | \hat{x}, x_{1:t}, y_{1:t}) \quad (1.12)$$

$$= \arg \max_{\hat{x} \in D} H(\hat{y} | \hat{x}, x_{1:t}, y_{1:t}) - \mathbb{E}_{\theta \sim \mathbf{p}(\cdot | x_{1:t}, y_{1:t})} [H(\hat{y} | \hat{x}, \hat{x}, y_{1:t})]$$

Explanation 1.1.

① $H(\hat{y} | \hat{x}, x_{1:t}, y_{1:t})$:

is the entropy of the predictive posterior distribution [def. 6.14], approximate using approximate inference section 9.

② $\mathbb{E}_{\theta \sim \mathbf{p}(\cdot | x_{1:t}, y_{1:t})} [H(\hat{y} | \hat{x}, \theta)]$:

is the conditional Entropy over the labels by drawing θ from the posterior distribution and averagin over them.

3. Examples

Example 1.1 Gaussian Information Gain:

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

4. Proofs

Proof 1.1: [def. 1.4]

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in D} F(S_t \cup \{x\}) = \arg \max_{x \in D} F(S_t \cup \{x\}) - F(S_t) \\ &= \arg \max_{x \in D} I(f; y_{S_t+x}) - I(f; y_{S_t}) \\ &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_x | f) \\ \text{eq. (4.17)} \quad &= \arg \max_{x \in D} H(y_{S_t+x}) - H(y_{S_t+x} | f) - H(y_{S_t}) + H(y_{S_t} | f) \\ &= \arg \max_{x \in D} H(y_{S_t}, x) - H(y_{S_t+x} | f) - H(y_{S_t}) + H(y_{S_t} | f) \\ \text{eq. (4.7)} \quad &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_{S_t+x} | f) + H(y_{S_t} | f) \\ &\quad - H(y_x | y_{S_t}) - H(y_x | f) \end{aligned}$$

Note ♣

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

Proof 1.2: [cor. 1.1]

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

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_x | f) \\ \text{eq. (4.24)} \quad &= \arg \max_{x \in D} \frac{1}{2} \ln(2\pi e) \sigma^2 x | S_t - \frac{1}{2} \ln(2\pi e) \sigma_n^2 \\ \text{eq. (23.62)} \quad &= \sigma_x^2 | S_t \end{aligned}$$

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

$$x_t = \arg \max_{x \in D} \sigma_{t-1}^2(x) \quad (1.13)$$

Proof 1.3: [cor. 1.4]

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

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in D} H(y_x | y_{S_t}) - H(y_x | f) \\ \text{eq. (4.24)} \quad &= \arg \max_{x \in D} \frac{1}{2} \ln(2\pi e) \sigma^2 x | S_t - \frac{1}{2} \ln(2\pi e) \sigma_n^2(x) \\ \text{eq. (23.62)} \quad &= \arg \max_{x \in D} \ln \frac{\sigma_f^2(x)}{\sigma_n^2(x)} \quad \text{eq. (23.74)} \quad = \arg \max_{x \in D} \frac{\sigma_f^2(x)}{\sigma_n^2(x)} \end{aligned}$$

Proof 1.4: [def. 1.6]

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

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:
 $f \in \mathcal{F} \quad f: D \mapsto \mathbb{R}$

that is expensive but from which we can draw noisy observations:

$$y_t = f(\mathbf{x}_t) + \epsilon \quad (2.2)$$

Goal

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

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

\Rightarrow 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 knowledge of f .

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

$$= T \max_{\mathbf{x} \in D} f(\mathbf{x}) - \sum_{t=1}^T f(\mathbf{x}_t) \quad (2.4)$$

r_t : instantaneous regret

Definition 2.4 (Time) Average Regret:

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

Definition 2.5

No/Sublinear Regret Algorithms:

$$\lim_{T \rightarrow \infty} \underbrace{\frac{R_T}{T}}_0 = 0 \quad \frac{R_T}{T} = o(1) \quad \forall \text{sequences } 1, \dots, T \quad (2.6)$$

Average regret

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 ϵ_t explore unif. at randomn:
 $a_{t+1} = \mathcal{U}(a_1, \dots, |\mathcal{A}|)$
- With probability $1 - \epsilon_t$ take action with highest known empirical mean payoff:

$$a_{t+1} = \arg \max_{a \in \mathcal{A}} \hat{\mu}_{a,T} \quad \hat{\mu}_{a,T} = \frac{1}{n_{a,T}} \sum_{s=1}^T \mathbb{1}_{\{a_s=a\}} v_{a,s} \quad (2.8)$$

4: 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]

$$\mathbf{x}_t = \arg \max_{\mathbf{x} \in D} \mu_{t-1}(\mathbf{x}) \quad (2.9)$$

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:

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

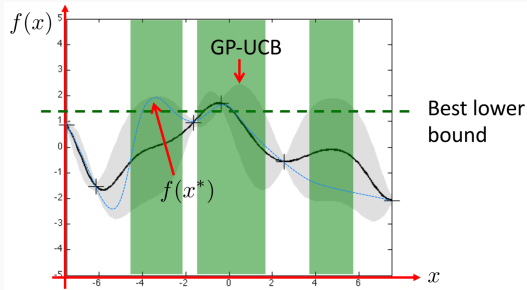
then it follows that:

$$f(\mathbf{x}^*) \geq \max \mu(x) - \beta\sigma(x) \quad (2.10)$$

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

$$f(\mathbf{x}^*) \geq \max \mu(x) - \beta\sigma(x) \geq \mu(\mathbf{x}') + \sigma(\mathbf{x}') \quad (2.11)$$

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:

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

Explanation 2.3 (Definition 2.7).

- $\beta_t \rightarrow \infty$ 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 β_t (needs to slowly decay with const: $\log t$):

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

with

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

Explanation 2.4 (γ_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:

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

Corollary 2.3 Squared Exponential Kernel:

For a squared exponential kernel [def. 11.13] it holds:

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

Corollary 2.4 Matern Kernel $\nu > 2$:

For a linear kernel [def. 11.14] it holds:

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

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:

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

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 θ of a given model $f \in \mathcal{F}$

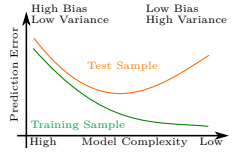
Definition 3.4 Model Assessment/Evaluation: Is the process of evaluating the performance of a model.

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 \Rightarrow big test error.



Definition 3.6 Labeled Data

$$\mathcal{D} := \{(\mathbf{x}_j, \mathbf{y}_j) \mid \mathbf{x}_j \in \mathcal{X}, \mathbf{y}_j \in \mathcal{Y}\} \quad \mathcal{D}/\mathcal{Z}$$

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 hyperparameters.

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:

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

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

$$\mathcal{R}(f) = \int \int f(\mathbf{x}, y) l(y, f(\mathbf{x})) dy d\mathbf{x}$$

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

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

Questions

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

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

1.2. Empirical Risk Minimization

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

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

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

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

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

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

Note: overfitting to the validation set

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

2. Cross Validation

Definition 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:



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:

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

⇒ 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 $X \sim \mathbf{p}$:

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

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

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

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

	log	units
≅ 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:

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

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 $X \sim \mathbf{p}$:

$$H(\mathbf{p}) = \mathbb{E}_X[I_X(x)] = \mathbb{E}_X\left[\log\frac{1}{\mathbf{p}_X(x)}\right] = -\mathbb{E}_X[\log \mathbf{p}_X(x)]$$

$$= -\sum_{i=1}^n \mathbf{p}(x_i) \log \mathbf{p}(x_i) \tag{4.2}$$

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

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

Notes

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

Property 4.1 Non negativity:

Entropy is always non-negative:
$$H(X) \geq 0 \quad \text{if } X \text{ is deterministic} \quad H(X) = 0 \tag{4.4}$$

1.2.1. Conditional Entropy

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

$$H(Y|X=x) = \mathbb{E}_{Y|X=x} \left[\log \frac{1}{\mathbf{p}_{Y|X}(Y|X=x)} \right]$$

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

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

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

$$= \mathbb{E}_{X,Y} \left[\log \frac{\mathbf{p}(x)}{\mathbf{p}(x,y)} \right]$$

Definition 4.6 Chain Rule for Entropy: proof 4.4

$$H(Y|X) = H(X,Y) - H(X)$$

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

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

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

Corollary 4.2 From eq. (4.17):

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

1.3. Cross Entropy

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

$$H(\mathbf{p}, \mathbf{q}) = \mathbb{E}_{x \sim \mathbf{p}} \left[\log \left(\frac{1}{\mathbf{q}(x)} \right) \right] \tag{4.10}$$

$$= -\mathbb{E}_{x \sim \mathbf{p}} [\log \mathbf{q}(x)] \tag{4.11}$$

$$= H(\mathbf{p}) + D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \tag{4.12}$$

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

1.4. Kullback-Leibler (KL) divergence

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

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

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

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

Explanation 4.3.

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

Intuition

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

Property 4.3 Non-Symmetric:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \neq D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p}) \quad \forall \mathbf{p}, \mathbf{q} \tag{4.14}$$

Property 4.4:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \geq 0 \tag{4.15}$$

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = 0 \iff \mathbf{p}(x) = \mathbf{q}(x) \forall x \in \mathcal{X} \tag{4.16}$$

Note

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

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

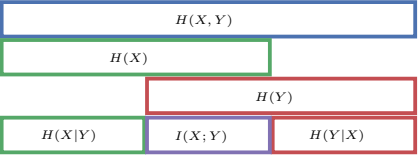
1.5. Mutual Information

Definition 4.9 Mutual Information/Information Gain: example 4.8 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.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.17}$$

$$= H(X) + H(Y) - H(X,Y)$$

$$= D_{\text{KL}}(\mathbf{p}_{X,Y} \parallel \mathbf{p}_X \mathbf{p}_Y)$$



Explanation 4.4 (Definition 4.9).

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

Property 4.5 Symmetry:

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

Property 4.6 Positiveness:

$$I(X;Y) \geq 0 \quad \text{if } X \perp Y \quad I(X;Y) = 0 \tag{4.19}$$

Property 4.7:

$$I(X;Y) \leq H(X) \quad I(X;Y) \leq H(Y) \tag{4.20}$$

Property 4.8 Self-Information:

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

Property 4.9 Montone Submodularity: Mutual information is monotone submodular^[def. 19.14]:

$$H(X,z) - H(x) \geq H(Y,z) - H(Y) \tag{4.21}$$

^[def. 4.6]
$$\iff H(z|X) \geq H(x|Y) \tag{4.22}$$

2. Proofs

Proof 4.1: ^[def. 4.7]

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

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

Proof 4.2: ^{[def. 4.15];}

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

$$\stackrel{??}{=} \min_h \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}} \left[\mathbb{E}_{y \sim p_{\mathcal{Y}|\mathcal{X}}} [(y - h(\mathbf{x}))^2 | \mathbf{x}] \right]$$

$$\stackrel{\heartsuit}{=} \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}} \left[\min_h(\mathbf{x}) \underbrace{\mathbb{E}_{y \sim p_{\mathcal{Y}|\mathcal{X}}} [(y - h(\mathbf{x}))^2 | \mathbf{x}]}_{\mathcal{R}_p(h, \mathbf{x})} \right] \stackrel{[def. 4.11]}{=}$$

Now lets minimize the conditional executed risk;

$$h^*(\mathbf{x}) = \arg \min_h \mathbb{E}_{y \sim p_{\mathcal{Y}|\mathcal{X}}} [(y - h(\mathbf{x}))^2 | \mathbf{x}] \quad (4.23)$$

$$0 \stackrel{!}{=} \frac{d}{dh^*} \mathcal{R}_p(h^*, \mathbf{x}) = \frac{d}{dh^*} \int (y - h^*)^2 p(y|x) dy$$

$$= \int \frac{d}{dh^*} (y - h^*)^2 p(y|x) dy = \int 2(y - h^*) p(y|x) dy$$

$$= -2h^* \underbrace{\int p(y|x) dy}_{=1} + 2 \underbrace{\int y p(y|x) dy}_{\mathbb{E}_Y[Y|X=x]}$$

Notes: \heartsuit

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

Note

$$\mathbb{E}[X] \mathbb{E}[Y|X] = \int_{\mathcal{X}} p_X(x) dx \int_{\mathcal{Y}} p(y|x) dy$$

$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_X(x) p(y|x) xy dx dy = \mathbb{E}[XY]$$

Proof 4.3: Definition 4.5

$$\mathbb{E}_{\mathcal{X}} [H(Y|X = x)] = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y|x) \log p(y|x)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x)} \right)$$

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

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

$$= - \sum_{x,y} p(x, y) \log p(x, y) + \sum_x p(x) \log \frac{1}{p(x)}$$

$$= H(X, Y) - H(X)$$

Proof 4.5: example 4.4

$$\text{KL}(p \parallel q) = \mathbb{E}_p [\log(p) - \log(q)]$$

$$= \mathbb{E}_p \left[\frac{1}{2} \log \frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2} (\mathbf{x} - \mu_p)^\top \Sigma_p^{-1} (\mathbf{x} - \mu_p) \right]$$

$$+ \frac{1}{2} (\mathbf{x} - \mu_q)^\top \Sigma_q^{-1} (\mathbf{x} - \mu_q)]$$

$$= \frac{1}{2} \mathbb{E}_p \left[\log \frac{|\Sigma_q|}{|\Sigma_p|} \right] - \frac{1}{2} \mathbb{E}_p \left[(\mathbf{x} - \mu_p)^\top \Sigma_p^{-1} (\mathbf{x} - \mu_p) \right]$$

$$+ \frac{1}{2} \mathbb{E}_p \left[(\mathbf{x} - \mu_q)^\top \Sigma_q^{-1} (\mathbf{x} - \mu_q) \right]$$

$$= \frac{1}{2} \log \frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2} \mathbb{E}_p \left[(\mathbf{x} - \mu_p)^\top \Sigma_p^{-1} (\mathbf{x} - \mu_p) \right]$$

$$+ \frac{1}{2} \mathbb{E}_p \left[(\mathbf{x} - \mu_q)^\top \Sigma_q^{-1} (\mathbf{x} - \mu_q) \right]$$

$$\frac{\mathbb{E}_p[a]}{\stackrel{(28.56)}{=} \mathbb{E}_p \left[\text{tr} \left\{ (\mathbf{x} - \mu_p)^\top \Sigma_p^{-1} (\mathbf{x} - \mu_p) \right\} \right]} = \mathbb{E}_p \left[\text{tr} \left\{ (\mathbf{x} - \mu_p) (\mathbf{x} - \mu_p)^\top \Sigma_p^{-1} \right\} \right]$$

$$\stackrel{(28.56)}{=} \mathbb{E}_p \left[\text{tr} \left\{ \Sigma_p \Sigma_p^{-1} \right\} \right]$$

$$\stackrel{(28.56)}{=} \mathbb{E}_p [\text{tr} \{I_d\}] = \mathbb{E}_p [d] = d$$

$$\frac{\mathbb{E}_p[b]}{\stackrel{(34.54)}{=} (\mu_p - \mu_q)^\top \Sigma_q^{-1} (\mu_p - \mu_q) + \text{tr} \left\{ \Sigma_q^{-1} \Sigma_p \right\}} =$$

3. Examples

Example 4.1 : Normal distribution has two population parameters: the mean μ and the variance σ^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:

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \ln |2\pi e \Sigma| \stackrel{\text{eq. (28.57)}}{=} \frac{1}{2} \ln ((2\pi e)^d |\Sigma|)$$

$$= \frac{d}{2} \ln (2\pi e)^d + \log |\Sigma| \quad (4.24)$$

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2) \Rightarrow \frac{1}{2} \ln |2\pi e| + \frac{1}{2} \sum_{i=1}^d \ln \sigma_i^2$$

Example 4.4 proof 4.5

KL Divergence of Gaussians:

Given two Gaussian distributions:

$$p = \mathcal{N}(\mu_p, \Sigma_p) \quad q = \mathcal{N}(\mu_q, \Sigma_q) \quad \text{it holds}$$

$$D_{\text{KL}}(p \parallel q) = \text{tr} \left(\Sigma_q^{-1} \Sigma_p \right) + (\mu_q - \mu_p)^\top \Sigma_q^{-1} (\mu_q - \mu_p) - d + \ln \left(\frac{|\Sigma_q|}{|\Sigma_p|} \right)$$

$$= \frac{2}{2}$$

Example 4.5 KL Divergence of Scalar Gaussians:

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

$$p = \mathcal{N}(\mu_p, \sigma_p^2)$$

$$D_{\text{KL}}(p \parallel q) = \frac{1}{2} \left(\frac{\sigma_p^2}{\sigma_q^2} (\mu_q - \mu_p)^2 \sigma_q^{-2} - 1 + \log \left(\frac{\sigma_q^2}{\sigma_p^2} \right) \right)$$

Example 4.6 KL Divergence of Diag. Gaussians:

$$\theta \sim q(\theta|\lambda) = \mathcal{N}(\mu_q, \text{diag}(\sigma_1^2, \dots, \sigma_d^2)) \quad \lambda = [\mu_{1:d} \quad \sigma_{1:d}]$$

$$p = \mathcal{N}(\mu_p, \text{diag}(\sigma_1^2, \dots, \sigma_d^2))$$

Example 4.7 KL Divergence of Gaussians:

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

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

Example 4.8 Gaussian Mutal Information:

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

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(\epsilon)$$

$$\stackrel{\text{eq. (4.24)}}{=} \frac{1}{2} \ln (2\pi e)^d |\Sigma + \sigma^2 I| - \frac{1}{2} \ln (2\pi e)^d |\sigma^2 I|$$

$$= \frac{1}{2} \ln \frac{(2\pi e)^d |\Sigma + \sigma^2 I|}{(2\pi e)^d |\sigma^2 I|}$$

$$= \frac{1}{2} \ln |I + \sigma^{-2} \Sigma|$$

Supervised Learning

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

Recall: goal of supervised learning

Given: training data:

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

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

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

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:

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

Note

The distribution $p_{\mathcal{X}, \mathcal{Y}}$ 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 $p_{\mathcal{X}, \mathcal{Y}}$.

$$R_p(h) = \mathbb{E}_{(\mathbf{x}, y) \sim p} [l(y; h(\mathbf{x}))] = \int_{\mathcal{D}} p(\mathbf{x}, y) l(y; h(\mathbf{x})) d\mathbf{x} dy$$

$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} p(\mathbf{x}, y) l(y, h(\mathbf{x})) d\mathbf{x} dy$$

$$\stackrel{??}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} l(y, h(\mathbf{x})) p(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dy \quad (4.25)$$

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 ($\mathbf{x} \sim p_{\mathcal{X}}(\mathbf{x})$), we can still calculate the expected risk given/conditioned on the known measurement \mathbf{x} :

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

Note: ^[def. 4.10] \longleftrightarrow ^[def. 4.11]

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

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

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

Problem

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

But: even though we do not know the distribution of $p_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}, y)$ 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 $p_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}, y)$:

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

Note

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

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

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

Objective

Given data generated *i.i.d.* from an distribution $p_{\mathcal{X}, \mathcal{Y}}(\mathbf{x}_i, y_i)$.

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

Considering: the least squares risk:

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

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

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

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 observe a \mathbf{x} multiple times we may still get different y values.
- Our model/prediction is unique, can only predict a specific y .
- Hence** even if our model fits exactly the data generating process $\mathbf{X} = \mathbf{x}$ we may still obtain different y 's because due to random/independent measurement noise/errors that the optimal bayes predictor still makes.

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 \Rightarrow 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 θ

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 statistic is a measurable function f that assigns a **single** value F to a sample of random variables:

$$\begin{aligned} \mathbf{X} &= \{X_1, \dots, X_n\} \\ f: \mathbb{R}^n &\mapsto \mathbb{R} \qquad F = f(X_1, \dots, X_n) \end{aligned}$$

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 probability 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/random variable used to estimate a true (population) parameter θ ^[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 θ .

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 \mathbf{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) \mathbf{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 β between *response-variables* \mathbf{x} and a *predictor-variable(s)* \mathbf{y} up to some error ϵ :
$$\mathbf{y} = f(\mathbf{x}, \beta) + \epsilon \quad (5.1)$$

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 β (but not necessarily the data):
$$\mathbf{y} = \beta^T \phi(\mathbf{x}) + \epsilon \quad (5.2)$$

Linearity
Linearity is w.r.t. the coefficients β_j .
Thus a model with transformed non-linear predictor^[def. 5.1] variables is still called *linear*.

Definition 5.5 Residual
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:
$$r_i := e_i = y_i - \hat{y}_i = y_i - \hat{\beta}^T \mathbf{x}_i \quad i = 1, \dots, n \quad (5.3)$$

Simple (linear) regression (SLR)

Definition 5.6
Simple Linear Regression: Is a *linear regression*^[def. 5.7] with only one explanatory variable^[def. 5.1]:
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n \quad (5.4)$$

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

$$= \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i = \beta^T \mathbf{x}_i + \epsilon_i \quad i = 1, \dots, n$$

$$\begin{bmatrix} \mathbf{x} \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{y} \end{bmatrix} \quad \mathbf{y} = \mathbf{X}\beta$$

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 than 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
Simple Linear Quadratic Regression: Is a *linear regression*^[def. 5.7] with two explanatory variables^[def. 5.1] written as:
$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i \quad i = 1, \dots, n \quad (5.6)$$

0.0.1. Existence

Corollary 5.1 Existence:
$$\exists \beta : \begin{bmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1p}\beta_p \\ x_{21}\beta_1 + x_{22}\beta_2 + \dots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + x_{n2}\beta_2 + \dots + x_{np}\beta_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (5.7)$$

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

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

Definition 5.10 Least Squares Regression $\text{lsq}(\mathbf{X}, \mathbf{y})$:
Minimizes the residual sum of squares:
$$\hat{\beta} \in \arg \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \arg \min_{\mathbf{u} \in \mathfrak{R}(\mathbf{X})} \|\mathbf{y} - \mathbf{u}\|_2^2 \quad (5.10)$$

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

Alternative Formulation
Sometimes people write eq. (5.10) as $\frac{1}{2} \arg \min_{\beta} \|\mathbf{r}\|_2^2$ which leads to the same solution^{eq. (23.63)}.

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. 38.22]:
$$\mathbb{V}[\epsilon_i] = \sigma^2 \quad \forall i \quad (5.11)$$

3. The errors are uncorrelated:
$$\text{Cov}[\epsilon_i, \epsilon_j] = 0 \quad \forall i \neq j \quad (5.12)$$

4. The errors are jointly normally distributed with mean 0 and constant variance σ^2 :
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i = 1, \dots, n \iff \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n) \quad (5.13)$$

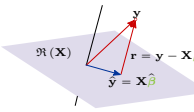
Definition 5.11
Simple Linear Regression Log-Likelihood:
Assume: a linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$
with Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$
With: $\mu = \mathbb{E}_{\epsilon}[\mathbf{y}] = \mathbb{E}_{\epsilon}[\mathbf{X}\beta + \epsilon] = \mathbf{X}\beta + 0$
 $\mathbb{V}_{\epsilon}[\mathbf{y}] = \mathbb{V}_{\epsilon}[\mathbf{X}\beta + \epsilon] = 0 + \mathbb{V}[\epsilon] = \mathbf{I}\sigma^2$
Thus: $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{I}\sigma^2)$ $\mathbf{Y}_i|\mathbf{X} \sim \mathcal{N}(x_i^T \beta, \sigma^2)$
with: $\theta = (\beta^T \sigma)^\top \in \mathbb{R}^{p+1}$
$$l_n(\mathbf{y}|\mathbf{X}, \theta) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 = -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2 \quad (5.14)$$

$$\theta^* \in \arg \max_{\theta \in \mathbb{R}^{p+1}} l_n(\mathbf{y}|\mathbf{X}, \theta) = \arg \min_{\theta \in \mathbb{R}^{p+1}} -l_n(\mathbf{y}|\mathbf{X}, \theta) \quad (5.15)$$

2.1. The Normal Equation

Definition 5.12
The Normal Equations:
Is the equation we need to solve in order to solve^{eq. (5.10)} or equivalently eq. (5.15) and is no longer an over determined system:
$$\mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{y} \quad \begin{matrix} \mathbf{X}^T \mathbf{X} \in \mathbb{R}^{p \times p} \\ \beta \in \mathbb{R}^p \\ \mathbf{X}^T \in \mathbb{R}^{p \times n} \\ \mathbf{y} \in \mathbb{R}^n \end{matrix} \quad (5.16)$$

Geometric Interpretation

Corollary 5.2
Geometric Interpretation:
We want to find $\arg \min_{\beta \in \mathbb{R}^n} \|\mathbf{X}\beta - \mathbf{y}\|_2^2$ which is equal to finding: $\arg \min \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$ $\hat{\mathbf{y}} \in \{\mathbf{X}\beta : \beta \in \mathbb{R}^n\} = \mathfrak{R}(\mathbf{X})$ but this minimum is equal to the orthogonal projection^[def. 28.22] of \mathbf{y} onto $\mathfrak{R}(\mathbf{X})$.


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

Proposition 5.2 Least Squares Solution:
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} := \mathbf{X}^\dagger \mathbf{y} \quad (5.17)$$

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/features/dimension of the problem.
Assume: $d \leq n$, that is we have an overdetermined system, more equations than unknowns.
1. Compute regular matrix (Matrix Product):
 $\mathbf{C} := \mathbf{X}^T \mathbf{X} \triangleq \mathcal{O}(n \cdot d^2)$.
2. Compute the r.h.s. vector (Matrix-Vector):
 $\mathbf{c} := \mathbf{X}^T \mathbf{y} \in \mathbb{R}^d \triangleq \mathcal{O}(nd)$.
3. Solve s.p.d. LSE via. **Cholesky decomposition:**
 $\mathbf{C}\mathbf{w} = \mathbf{c} \triangleq \mathcal{O}(d^3)$.
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 $\iff \mathbf{X}^T \mathbf{X}$ is invertible.
1. **Symmetric:** a transposed matrix times itself is symmetric $\Rightarrow \mathbf{C}$ is symmetric.
2. **Positive definite:**
$$\mathbf{w}^T \mathbf{C} \mathbf{w} = \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} = \|\mathbf{X} \mathbf{w}\|^2 > 0 \quad \forall \mathbf{w} \neq 0$$

has trivial kernel \nexists

QR Decomposition

2.1.3. Simple Linear Regression Solution

Definition 5.13
Linear Regression Solution:
$$\hat{\beta} = \underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\mathbf{X}^\dagger} \mathbf{y} \quad \text{with} \quad \begin{matrix} \Sigma^2 = (\mathbf{X}^T \mathbf{X})^{-1} \\ \mathbf{P} = \mathbf{X}^T \mathbf{y} \end{matrix} \quad (5.18)$$

 Σ^2 : Variance-Cov. M. \mathbf{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 $\mathbf{P}/\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T : \mathbf{y} \mapsto \hat{\mathbf{y}}$
Hat/Projection Matrix:
Is the matrix that projects the \mathbf{y} onto the $\hat{\mathbf{y}}$:
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} =: \mathbf{P}\mathbf{y} \quad (5.20)$$

Uniqueness
Theorem 5.1 : Let $\mathbf{A} \in \mathbb{R}^{p, p}$, $p \geq p$ then it holds that:
$$\mathbf{N}(\mathbf{A}) = \mathbf{N}(\mathbf{A}^T \mathbf{A}) \quad \mathfrak{R}(\mathbf{A}^T) = \mathfrak{R}(\mathbf{A}^T \mathbf{A}) \quad (5.21)$$

add rest from formulary
Theorem 5.2 Full-Rank Condition **F.R.C.:**
Equation 5.12 has a unique least squares solution given by:
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (5.22)$$

$$\iff \mathbf{N}(\mathbf{X}) = \{0\} \iff \text{rank}(\mathbf{X}) = p \quad p \geq p \quad (5.23)$$

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 $\theta \in \mathbb{R}^k$ 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. ϵ : noise/error term.

Lemma 5.1 : The conditional distribution D of Y given \mathbf{X} is equivalent to the unconditional distribution of the noise ϵ :

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

Example: Conditional linear Gaussian

Assume: a linear model $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

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

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

with: $\theta = (\mathbf{w}^\top \sigma)^\top \in \mathbb{R}^{n+1}$

Hence Y is distributed as a linear transformation of the \mathbf{X} variable plus some Gaussian noise ϵ : $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2) \Rightarrow$ 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:

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

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

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

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

$$\frac{\partial \ln(Y|\mathbf{X}, \theta)}{\partial \mathbf{w}} = \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{w}^\top \mathbf{x}_i) = \mathbf{0} \in \mathbb{R}^d$$

$$= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w} = \sum_{i=1}^n \mathbf{x}_i y_i$$

$$\frac{\partial \ln(Y|\mathbf{X}, \theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = 0$$

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

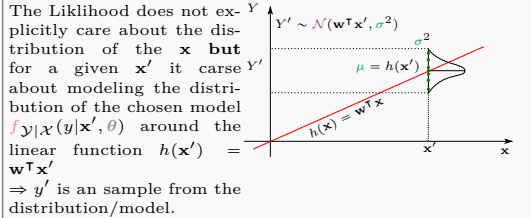
Note

- The mean μ of the normal distribution follows from:

$$\mathbb{E}[\mathbf{w}^\top \mathbf{x}_i + \epsilon_i] = \mathbb{E}[\mathbf{w}^\top \mathbf{x}_i] + \mathbb{E}[\epsilon_i] = \mathbf{w}^\top \mathbf{x}_i$$

const = 0
- The noise ϵ must have zero mean, otherwise it wouldn't be random anymore.
- The optimal function $h^*(\mathbf{x})$ determines the mean μ .
- We can also minimize:

$$\theta^* = \arg \max_{\theta} \hat{p}(Y|\mathbf{X}, \theta) = \arg \min_{\theta} -\hat{p}(Y|\mathbf{X}, \theta)$$



3.2. Conditional MLE $\hat{=}$ Least Squares

Assuming that the noise is i.i.d. Gaussian with constant variance σ , that is $\theta = (\mathbf{w}^\top \sigma)^\top$

and considering the negative log likelihood in order to minimize $\arg \max_{\alpha} \alpha = -\arg \min_{\alpha}$:

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

$$\arg \max_{\mathbf{w}} l_n(\mathbf{w}) \iff \arg \min_{\mathbf{w}} -l_n(\mathbf{w})$$

$$\arg \min_{\mathbf{w}} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \arg \min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \quad (5.25)$$

Thus Least squares regression equals Conditional MLE with a linear model + Gaussian noise.

Maximizing Likelihood \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 $\{y_1, \dots, y_n\} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\frac{\partial \ln(y|\mathbf{x}, \theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) \stackrel{!}{=} 0$$

$$\frac{\partial \ln(y|\mathbf{x}, \theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 \stackrel{!}{=} 0$$

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

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

Note

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

3.3. MLE for general conditional Gaussians

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

Given: data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ The MLE for general models h and i.i.d. Gaussian noise:

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

Is given by the least squares solution:

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

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

Other distributions

If we use other distributions instead of Gaussian noise, we obtain other loss functions e.g. L1-Norm for **Poisson Distribution**.

\Rightarrow if we know something about the distribution of the data we know which loss function we should chose.

Ridge Max Prior

Prior

Assume: prior $\mathbb{P}(\beta|\Sigma)$ on the model parameter β is gaussian as well and depends on the hyperparameter $(\stackrel{\text{def. 6.7}}{=} \Sigma \hat{=}$ co-variance matrix):

$$\begin{aligned} \beta &\sim \mathbf{p}^{\text{Ridge}}(\beta|\Sigma) = \mathcal{N}(\beta|0, \Sigma) \\ &\stackrel{[\text{def. 35.34}]}{=} (2\pi)^{-\frac{d+1}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \beta^\top \Sigma^{-1} \beta\right) \end{aligned}$$

$$l_n(\beta|\Sigma) = -\frac{1}{2} \ln \det(\Sigma)^{-1} - \frac{d+1}{2} \ln 2\pi - \frac{1}{2} \beta^\top \Sigma^{-1} \beta \quad (5.27)$$

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

Max Prior

$$\begin{aligned} \beta^* &\in \arg \max_{\beta \in \mathbb{R}^{d+1}} l_n(\beta|\Sigma) \\ &= \arg \max_{\beta \in \mathbb{R}^{d+1}} -\frac{1}{2} \ln \det(\Sigma)^{-1} - \frac{d+1}{2} \ln 2\pi - \frac{1}{2} \beta^\top \Sigma^{-1} \beta \\ 0 &\stackrel{!}{=} \frac{\partial}{\partial \beta^*} l_n(\beta^*|\Sigma) = -\frac{\partial}{\partial \beta^*} \beta^* \Sigma^{-1} \beta^* \stackrel{\text{eq. (5.36)}}{=} -2\Sigma^{-1} \beta^* \\ \beta^* &\in \arg \max_{\beta \in \mathbb{R}^{d+1}} \log p(\beta|\Sigma) = \arg \min_{\beta \in \mathbb{R}^{d+1}} -l_n(\beta|\Sigma) = 2\Sigma^{-1} \beta^* \end{aligned} \quad (5.28)$$

Log-MAP

$$\begin{aligned} \beta^* &\in \arg \max_{\beta \in \mathbb{R}^{d+1}} \mathbb{P}(\beta|\mathbf{X}, \mathbf{y}) \\ &\stackrel{\text{eq. (5.28)}}{=} \arg \min_{\beta \in \mathbb{R}^{d+1}} -\log \underbrace{\mathbb{P}(\beta|\Sigma)}_{\text{prior}} - \log \underbrace{\mathbb{P}(\mathbf{X}, \mathbf{y}|\beta)}_{\text{likelihood}} \\ &= \Sigma^{-1} \beta^* - \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y} + \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} \beta^* = 0 \\ &\iff (\Sigma^{-1} + \mathbf{X}^\top \mathbf{X} \sigma^{-2}) \beta^* = \sigma^{-2} \mathbf{X}^\top \mathbf{y} \\ &\quad (\sigma^2 \Sigma^{-1} + \mathbf{X}^\top \mathbf{X}) \hat{\beta} = \mathbf{X}^\top \mathbf{y} \\ \hat{\beta}^{\text{MAP}} &= (\sigma^2 \Sigma^{-1} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

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

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

which leads to:

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

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.

$$\|\mathbf{X}\beta - \mathbf{y}\|^2 + \|\Gamma\beta\|^2 \quad (5.31)$$

Γ : 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/Likelihood MAP inference

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

- $\|\mathbf{y} - \mathbf{X}\beta\|^2$ is forced to be small so that we find a weight vector β that matches the data as close as possible:

$$y_i = \beta_i^\top \mathbf{x}_i + \epsilon_i \quad \text{s.t.} \quad \sum_{i=1}^n \epsilon_i \text{ small}$$

In other words we want to fit the data well.

- $\beta^\top \Lambda \beta \stackrel{\text{ridge}}{=} \lambda \|\beta\|^2$ says chose a model with a small magnitude $\|\beta\|^2$.
Thus the smaller λ the bigger can the data faith fullness term be $\|\mathbf{y} - \mathbf{X}\beta\|^2$.

Note

The intercept β_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):

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

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

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

\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 $\lambda \xrightarrow{\sigma \uparrow} \infty$: $\|\beta\|^2$ 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 $\lambda \xrightarrow{\sigma \downarrow} 0$: $\|\mathbf{y} - \mathbf{X}\beta\|^2$ 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 $\Lambda^{-1} = \mathbf{1} \in \mathbb{R}^{d+1 \times d+1}$
- Λ is symmetric and diagonal.
- $(d+1)$ dimension as we included offset into β .

Heuristic Map Inference

A really large weight vector β will result in amplifying noise/larger variance/fluctuations \triangleq 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 $\mathbf{X}^T \mathbf{X}$ in eq. (5.30) is not invertible/regular/has not full rank. Then $(\mathbf{X}^T \mathbf{X} + \Lambda)$ will still be invertible/well posed. This was the original reason for L2/Ridge Regression.

MAP \triangleq Ridge

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

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

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

Ridge Regression MAP

Thus if we know our data β, σ we can chose λ statistically and do not need cross-validation.

Generalization

Regularized estimation can often be understood as MAP inference:

$$\arg \min_{\mathbf{w}} \sum_{i=1}^n l(\mathbf{w}^T \mathbf{x}_i; \mathbf{x}_i, y_i) + C(\mathbf{w}) =$$

$$= \arg \max_{\mathbf{w}} \prod_{i=1}^n \mathbb{P}(\mathbf{w}) \mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w}) = \arg \max_{\mathbf{w}} \mathbb{P}(\mathbf{w}|\text{data})$$

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

Priors

3.4. Laplace Prior \triangleq Lasso/L1-regularization

Intro

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

- bag of words with $d = \text{nb. of words} \gg \text{nb. of documents}$.
- Genome analysis $d = \text{nb. of genes} \gg \text{nb. of patients}$.

Problem: we have more unknowns/parameters than observations \Rightarrow no unique solution. **e.g.:** Trying to fit 1 data point with polynomial of degree 12.

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

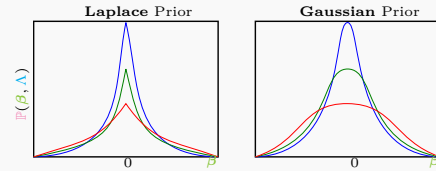
Idea: If most of the dimensions are irrelevant for the problem, then we can find a good (**sparse**) solution \triangleq **feature selection**/dimensionality reduction.

Given: Laplacian model prior $\beta \sim p(\beta|\Lambda)$:

$$\mathbb{P}_{\text{Lasso}}(\beta|\Lambda) \stackrel{\text{eq. (35.58)}}{=} \frac{\Lambda}{2} e^{-(\Lambda|\beta|)} = \prod_{j=1}^d \frac{\lambda_j}{2} e^{-\lambda_j |\beta_j|}$$

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

This leads to a L1 regularized model:



Thus: laplace priors gives sparseness, higher likelihood to get value at $\beta = 0$.

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

Laplacian MAP Prior Inference

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^d} \left\{ \|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda \|\beta\|_1 \right\}$$

$$= \arg \min_{\beta \in \mathbb{R}^d} \left\{ \|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda \sum_{i=1}^d |\beta_i| \right\} \quad (5.34)$$

$|\beta|_i$ does not change β_i while β_i^2 becomes very small for values $\in (0, 1)$ thus when minimizing the L2 error $\|\text{betac}\|^2 \rightarrow 0$ but not β_i while for L1 regularization will actually have to set β_i values to zero for large enough λ .

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. 23.12]) optimization problem eq. (5.34) is not differentiable at $\beta_i = 0$ and **thus** has no closed form solution as the L2 problem \Rightarrow quadratic programming.

3.5. Sparseness Priors/L0-regularization

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

\Rightarrow measure for the number of possible non-zero dimensions/-parameters in β .

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 \Rightarrow computationally difficult combinatorics.

Scalarization vs. Constrained Optimization

There are two equivalent ways of trading:

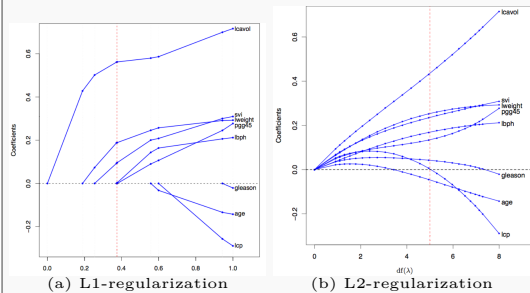
<ul style="list-style-type: none"> $g(\beta) = \ \mathbf{y} - \mathbf{X}\beta\ ^2$: the data term and $f(\beta)$: the Regularizer. <p>Scalarization</p> $\min_{\mathbf{w}} f(\beta) + Cg(\beta)$ <p>$g(\mathbf{w})$ $\begin{cases} C \text{ small} \\ C \text{ large} \end{cases}$</p>	<p>Constraint opt.</p> $\min_{\mathbf{w}} g(\beta) \text{ s.t. } f(\mathbf{w}) \leq B$ <p>$g(\mathbf{w})$ $\begin{cases} B \text{ small} \\ B \text{ large} \end{cases}$</p>
---	---

Note

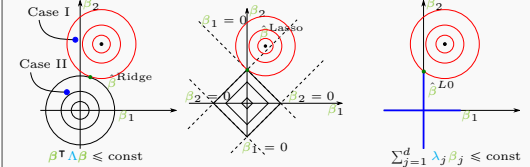
Scalarization and constrained optimization gives the same curves $\iff f, g$ are both convex functions.

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



The constraint formulation of the optimization problems can be plotted for two features β_1, β_2 as:



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

Likelihoods

3.6. Student's-t likelihood loss function

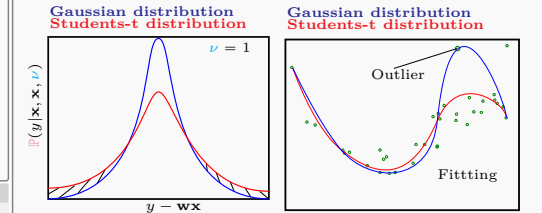
Students-t Distribution:

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

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



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

Students-t	$\mathbb{P}(y - \mathbf{w}^T \mathbf{x} > t) = \mathcal{O}(t^{-\alpha})$	$\alpha > 0$
	(Polynomial decay)	
Gaussian	$\mathbb{P}(y - \mathbf{w}^T \mathbf{x} > t) = \mathcal{O}(\exp^{-\alpha t})$	$\alpha > 0$
	(Exponential decay)	

\Rightarrow **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 $\text{Cov}[Y_i, Y_j|\mathbf{X}] = 0$. Hence we have i.i.d. samples with a corresponding conditional (log-)likelihood given by:

$$\mathcal{L}_n(\mathbf{y}|\mathbf{X}, \theta) \stackrel{\text{i.i.d.}}{=} \prod_{i=1}^n p(\mathbf{x}_i, y_i|\theta) = \prod_{i=1}^n \mathcal{N}(\beta^T \mathbf{x}_i, \sigma^2)$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2}\right)$$

$$= (\sigma^2 2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2\right)$$

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

Proof 5.2 Definition 5.13:

$$\beta^* \in \arg \min_{\beta \in \mathbb{R}^p} -\ln(\mathbf{y}|\mathbf{X}, \theta)$$

$$= \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2$$

$$= \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \arg \min_{\beta \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$\stackrel{\star}{\iff} (-2\mathbf{y}^T \mathbf{X} + 2\mathbf{X}^T \mathbf{X} \beta^*) \stackrel{!}{=} 0$$

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

Note: ★

$$\begin{aligned} & (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta + (\mathbf{X}\beta)^\top \mathbf{y} - (\mathbf{X}\beta)^\top (\mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top (\mathbf{X}\beta) \end{aligned}$$
$$\frac{\partial}{\partial \mathbf{x}} \mathbf{M}\mathbf{x} = \mathbf{M} \quad \text{and} \quad \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{M}\mathbf{x} = (\mathbf{M} + \mathbf{M}^\top)\mathbf{x} \quad (5.36)$$

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

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

Thus

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

combine proofs

Proof 5.3: ^[def. 5.12]

$$\begin{aligned} \text{lsq}(\mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta + (\mathbf{X}\beta)^\top \mathbf{y} - (\mathbf{X}\beta)^\top (\mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top (\mathbf{X}\beta) \end{aligned}$$
$$0 = \frac{\partial}{\partial \beta} \text{lsq}(\mathbf{X}, \mathbf{y}) = 2\mathbf{X}^\top \mathbf{X}\beta - 2\mathbf{X}^\top \mathbf{y} = 2\mathbf{X}^\top (\mathbf{X}\beta - \mathbf{y})$$

Note

$$\frac{\partial}{\partial \beta} \beta^\top \mathbf{X}^\top (\mathbf{X}\beta) \stackrel{\text{eq. (28.134)}}{=} (\mathbf{X}^\top \mathbf{X} + (\mathbf{X}^\top \mathbf{X})^\top)\beta = 2\mathbf{X}^\top \mathbf{X}\beta$$

Proof 5.4: Corollary 5.2

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

where $\mathbf{X} = \{\mathbf{x}_{:,1}, \dots, \mathbf{x}_{:,m}\}$ is the “basis” of the Range space:

$$(\mathbf{X}\beta - \mathbf{y})^\top \mathbf{x}_{:,j} = 0 \quad \forall j = 1, \dots, m$$

5. Examples

Example 5.1 Simple Linear Regression:

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

Example 5.2 Simple Linear Quadratic Regression:

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

Classification

6. Intro

Definition 5.18 Training Data \mathcal{D} :
 $\mathcal{D} := \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathcal{Y} := \{c_1, \dots, c_K\}\}$

Definition 5.19 Classifier c :
 Is a mapping that maps the features into classes:
 $c: \mathcal{X} \rightarrow \mathcal{Y}$ (5.38)

6.1. Types of Classification

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

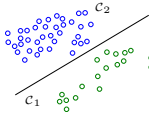
- Collectively/jointly exhaustiveness:
 $S = A \cup A^c$ (5.39)

- Mutual exclusivity:
 $s \in A \implies s \notin A^c \quad \forall s \in S$ (5.40)

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

Definition 5.21 Binary Classification:

Is a classification problem where the labels are binary:
 $\mathcal{Y} = \{c_1, c_2\} = \{-1, 1\}$ (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. 28.70] with one non-zero entry, where the index j of the non-zero entry indicates the class c_j :

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

s.t. $\mathbf{y}_i = \mathbf{e}_j \iff \mathbf{y}_i = \mathbf{c}_j$

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]:

$$y = c_k \implies y = \mathbf{e}_k \quad (5.42)$$

Definition 5.24 Soft Labels/Targets: Are observations $y \in \mathcal{Y}$ that are consider as noisy observations or probabilities \mathbf{p} . We can encode them using a probabilistic vector^[def. 28.71]:

$$y = [\mathbf{p}_1 \dots \mathbf{p}_K]^T \quad (5.43)$$

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

$\{-1, 1\}$

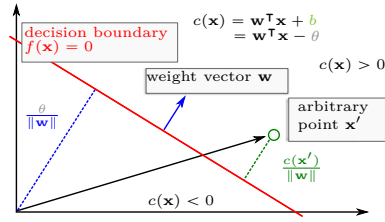
7.1. Linear Classification

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. 28.15]:

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

Explanation 5.2 (Definition 5.26).



- 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 θ vs. Bias b). The offset is called bias if it is considered as part of the classifier $\mathbf{w}^T \mathbf{x} + b$ 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:

$$\hat{\mathbf{w}}^T \mathbf{x} = \mathbf{w}^T \mathbf{x} y > 0 \quad \forall (\mathbf{x}, y) \in \mathcal{D} \quad (5.45)$$

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:

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

7.1.1. Normalization

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

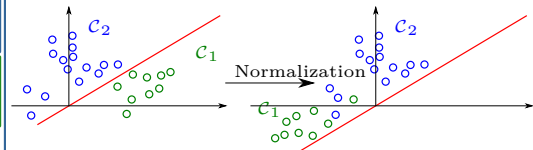
$$\mathbf{w} \leftarrow \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \quad \mathbf{x} \leftarrow \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \implies \mathbf{w}^T \mathbf{x} = (\mathbf{w}^T \ b) \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \mathbf{w}^T \mathbf{x} + b$$

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:

$$\tilde{\mathbf{x}} = \begin{cases} \mathbf{x} & \text{if } \mathbf{w}^T \mathbf{x} > \theta \\ -\mathbf{x} & \text{if } \mathbf{w}^T \mathbf{x} < \theta \end{cases} \quad (5.47)$$

Explanation 5.4 (proposition 5.4).

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



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 :

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

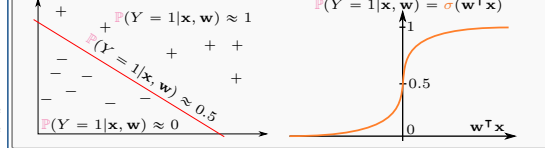
8. Logistic Regression

Bern($y; \sigma(\mathbf{w}^T \mathbf{x}, \sigma^2)$)

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. 35.22].

Problem: we need to convert/translate distance $\mathbf{w}^T \mathbf{x}$ into probability in order to use a bernoulli distribution.

Idea: use a sigmoidal function to convert distances $\mathbf{z} := \mathbf{w}^T \mathbf{x}$ into probabilities \Rightarrow Logistic Function^[def. 5.29].



8.1. Logistic Function

Definition 5.29 Sigmoid/Logistic Function:

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

Explanation 5.5 (Sigmoid/Logistic Function).

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

8.2. Logistic Regression

Definition 5.30 Logistic Regression:

models the likelihood of the output y as a Bernoulli Distribution^[def. 35.22] $y \sim \text{Bern}(\mathbf{p})$, where the probability \mathbf{p} is given by the Sigmoid function^[def. 5.29] of a linear regression:

$$p(y|\mathbf{x}, \mathbf{w}) = \text{Bern}(\sigma(\mathbf{w}^T \mathbf{x})) = \begin{cases} \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}} & \text{if } y = +1 \\ \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}}} & \text{if } y = -1 \end{cases} \stackrel{?? 5.5}{=} \frac{1}{1 + \exp(-y \cdot \mathbf{w}^T \mathbf{x})} = \sigma(-y \cdot \mathbf{w}^T \mathbf{x}) \quad (5.49)$$

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:

$$l_l(\mathbf{w}; \mathbf{x}, y) := \log(1 + \exp(-y \cdot \mathbf{w}^T \mathbf{x}))$$

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

Corollary 5.7 MLE for Logistic Regression:

$$l_n(\mathbf{w}) = \sum_{i=1}^n l_l = \sum_{i=1}^n \log(1 + \exp(-y_i \cdot \mathbf{w}^T \mathbf{x}_i)) \quad (5.51)$$

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 $\nabla_{\mathbf{w}} l_l(\mathbf{w})$:

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

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

Explanation 5.6.

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

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

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

Algorithm 5.1 Vanilla SGD for Logistic Regression:

Initialize: \mathbf{w}
 1: for $1, 2, \dots, T$ do
 2: Pick (\mathbf{x}, y) unif. at random from data \mathcal{D}
 3: $\hat{\mathbb{P}}(Y = -y|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y \cdot \mathbf{w}^T \mathbf{x})} = \sigma(y \cdot \mathbf{w}^T \mathbf{x})$
 \triangleright compute prob. of misclassif. with cur. model
 4: $\mathbf{w} = \mathbf{w} + \eta_t y \mathbf{x} \sigma(y \cdot \mathbf{w}^T \mathbf{x})$
 5: end for

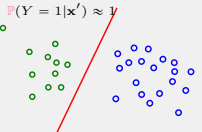
Making Predictions

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):

$$\mathbb{P}(y|\mathbf{x}, \hat{\mathbf{w}}) = \frac{1}{1 + \exp(-y \hat{\mathbf{w}}^T \mathbf{x})} \quad (5.53)$$

Drawback

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



8.2.2. Maximum a-Posteriori Estimates

adapt

8.3. Logistic regression and regularization

Adding Priors to Logistic Likelihood

- L2 (Gaussian prior):**

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

- L1 (Laplace prior):**

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

- Generalized:**

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

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

8.4. SGD for L2-regularized logistic regression

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

Thus: \mathbf{w} is pulled/shrunk 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:

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

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

Proof 5.6:

[def. 5.31]

$$\begin{aligned}
 l_n(\mathbf{w}) &= \arg \max_{\mathbf{w}} \mathbb{P}(y_{1:n} | \mathbf{x}_{1:n}, \mathbf{w}) = \arg \min_{\mathbf{w}} -\log \mathbb{P}(Y | \mathbf{X}, \mathbf{w}) \\
 &\stackrel{\text{i.i.d.}}{=} \arg \min_{\mathbf{w}} \sum_{i=1}^n -\log \mathbb{P}(y_i | \mathbf{x}_i, \mathbf{w}) \\
 \text{eq. (5.49)} \quad &= -\log \frac{1}{1 + \exp(-y_i \cdot \mathbf{w}^\top \mathbf{x}_i)} \\
 &= \log(1 + \exp(-y_i \cdot \mathbf{w}^\top \mathbf{x}_i)) =: l_l(\mathbf{w})
 \end{aligned}$$

Proof 5.7:

[def. 5.32]

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

Model Parameter Estimation

1. Maximum Likelihood Estimation

1.1. Likelihood Function

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

Consider: a probability density/mass function $f_{\mathcal{X}}(\mathbf{x}; \theta)$

Definition 6.1 Likelihood Function $\mathcal{L}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}_+$: 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 θ :

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

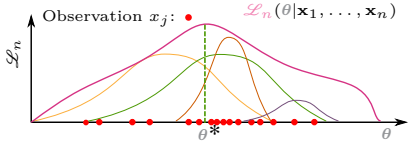


Figure 6: Possible Likelihood function in pink. Overlaid: possible candidate functions for Gaussian model explaining the observations.

Likelihood function is not a pdf

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

Corollary 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:

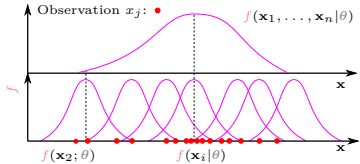


Figure 7: Bottom: probability distributions of the different data points \mathbf{x}_i given a fixed θ for a Gaussian distribution. Top: joint probability distribution of the i.i.d. data points $\{\mathbf{x}_i\}_{i=1}^n$ given a fixed θ

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

Notation

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

1.2. Maximum Likelihood Estimation (MLE)

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

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

Definition 6.2 Log Likelihood Function $l_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}$:

$$l_n(\theta|\mathbf{X}) = \log \mathcal{L}_n(\theta|\mathbf{X}) = \log f(\mathbf{X}|\theta) \quad (6.2)$$

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. (23.66):

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

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

$$\theta^* = \arg \max_{\theta \in \Theta} \mathcal{L}_n(\theta; \mathbf{x}) \quad \text{or} \quad \theta^* = \arg \max_{\theta \in \Theta} l_n(\theta; \mathbf{x}) \quad (6.3)$$

1.3. Maximization vs. Minimization

For optimization problems we minimize by convention. The logarithm is a concave function^[def. 23.25] \cap , thus if we calculate the extremal point we will obtain a maximum. If we want to calculate a minimum instead (i.e. in order to be compatible with some computer algorithm) we can convert the function into a convex function^{section 5} \cup by multiplying it by minus one and consider it as a loss function instead of a likelihood.

Definition 6.4 Negative Log-likelihood $-l_n(\theta|\mathbf{X})$:

$$\theta^* = \arg \max_{\theta \in \Theta} l_n(\theta|\mathbf{X}) = \arg \min_{\theta \in \Theta} -l_n(\theta|\mathbf{X}) \quad (6.4)$$

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

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

Note

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

Definition 6.5 Conditional (log) likelihood function:

Models the likelihood of a model with parameters θ given the data $\mathbf{Z} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

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

2. Maximum a posteriori estimation (MAP)

Idea

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

Thus the question arises if we can introduce a bit of bias into the MLE in turn of decreasing the variance?

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

2.1. Prior Distribution

Definition 6.6 Prior (Distribution) $\pi(\theta) = p(\theta)$:

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

$$\theta \sim \pi(\theta) = p(\theta) \quad (6.5)$$

Notes

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

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 $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$:

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

Definition 6.7 Hyperparameters $p_{\lambda}(\theta)$:

In most cases the prior distribution are parameterized that is the pdf $\pi(\theta|\lambda)$ depends on a set of parameters λ .

The parameters of the prior distribution, are called hyperparameters and are supplied due to believe/prior knowledge (and do not depend on the data) see example 6.1

2.2. Posterior Distribution

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

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

Definition 6.9

Posterior Distribution and Bayes Theorem:

Using Bayes theorem 34.3 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:

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

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Normalization}} \quad (6.9)$$

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

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 ?? θ^* that maximizes the posterior distribution.

2.2.2. Maximization

Definition 6.10

Maximum a-Posteriori Estimates (MAP):

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

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

Log-MAP estimator:

$$\theta^* = \arg \max_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \quad (6.12)$$

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

$$\stackrel{\text{eq. (23.63)}}{\propto} \arg \max_{\theta} \{p(\mathbf{y}|\theta, \mathbf{X}) \cdot p_{\lambda}(\theta)\}$$

Corollary 6.4 Negative Log MAP:

$$\theta^* = \arg \max_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \quad (6.13)$$

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

3. Proofs

Proof 6.1: 6.10:

$$\begin{aligned} p(\mathbf{X}, \mathbf{y}, \theta) &= \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \\ \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} &= \frac{p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)} \\ &= p(\mathbf{y}|\mathbf{X}, \theta)p(\theta|\mathbf{X})p(\mathbf{X}) \\ \stackrel{\text{eq. (6.6)}}{=} & p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)p(\mathbf{X}) \\ \Rightarrow \underline{p(\theta|\mathbf{X}, \mathbf{y})} &= \frac{p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)\cancel{p(\mathbf{X})}}{p(\mathbf{y}|\mathbf{X})\cancel{p(\mathbf{X})}} \end{aligned}$$

Note

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

4. Examples

Example 6.1 Hyperparameters Gaussian Prior:

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

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

Bayesian Inference/Modeling

Definition 6.11 Bayesian Inference: So far we only really looked at point estimators/estimates^[def. 37.8]. 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 \Rightarrow can deduce statistical properties of parameters from their distributions.

Definition 6.12 $p(w|y, X)/p(w|D)$
Posterior Probability Distribution:
 ① Specify the prior $p_\lambda(w)$
 ② Specify the likelihood $p(y|w, X)/p(D|w)$
 ③ Calculate the evidence $p(y|X)/p(D)$
 ④ Calculate the posterior distribution $P(w|y, X)/p(w|D)$

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

Definition 6.13 $p(y|X)/p(D)$
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:

$$p(y|X) = \int p(y|w, X) \cdot p_\lambda(w) dw = \int \text{Likelihood} \cdot \text{Prior} dw \quad (6.14)$$

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

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

$$p(f_*|x_*, X, y) = \int p(f_*|x_*, w) \cdot p(w|X, y) dw \quad (6.15)$$

it is calculated by weighting the likelihood^[def. 6.1] of the new observation x_* with the posterior of the observed data and averaging over all parameter values w .
 \Rightarrow obtain a distribution not depending on w .

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

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

Bayesian Filtering

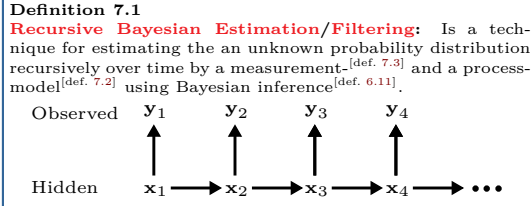


Figure 8: This problem corresponds to a *hidden Markov model* (HMM)^[def. 14.1]

$$x_t = (x_{t,1} \quad \dots \quad x_{t,n}) \quad y_t = (y_{t,1} \quad \dots \quad y_{t,m})$$

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

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

Corollary 7.1 Markov Property $x_t \perp\!\!\!\perp x_{1:t-2}|x_{t-1}$: The process models^[def. 7.2] is Markovian^[def. 38.14] i.e. the current state depends only on the previous state:

$$p(x_t|x_{1:t-1}) = p(x_t|x_{t-1}) \quad (7.1)$$

Definition 7.3 $y_t \sim p(y_t|x_t)$
Measurement/Sensor-Model/Likelihood: is a model h that maps observations/sensor measurements of our model y_t to the model state x_t

Corollary 7.2 $y_t \perp\!\!\!\perp y_{1:t-1}x_{1:t-1}|x_t$
Conditional Independent Measurements: The measurements y_t are conditionally independent of the previous observations $y_{1:t-1}$ given the current state x_t :

$$p(y_t|y_{1:t-1}, x_t) = p(y_t|x_t) \quad (7.2)$$

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:

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

Definition 7.4 Chapman-Kolmogorov eq. $p(x_t|y_{1:t-1})$
Prior Update/Prediction Step [proof 10.3]:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1}) dx_{t-1} \quad (7.3)$$

Prior Distribution:

$$p(x_0|y_{0-1}) = p(x_0) = p_0 \quad (7.4)$$

Definition 7.5 $p(x_t|y_{1:t})$
Posterior Distribution/Update Step [proof 10.4]:

$$p(x_t|y_{1:t}) = \frac{1}{Z_t} p(y_t|x_t) p(x_t|y_{1:t-1}) \quad (7.5)$$

Definition 7.6 Normalization [see proof 10.5]:

$$Z_t = p(y_t|y_{1:t-1}) = \int p(y_t|x_t)p(x_t|y_{1:t-1}) dx_t \quad (7.6)$$

Algorithm 7.1 Optimal Bayesian Filtering:
 1: **Input:** $p(x_0)$
 2: **while** Stopping Criterion not full-filled **do**
 3: **Prediction Step:**

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

 4: **Update Step:**

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

 with:

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

 5: **end while**

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

$$p(x_{1:t}, y_{1:t}) = p(x_1)p(y_1|x_1) \prod_{i=2}^t p(x_i|x_{i-1})p(y_i|x_i) \quad (7.7)$$

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

1. Kalman Filters

Definition 7.7 Kalman Filter Assumptions: Assumes a *linear*^[def. 23.15] process model^[def. 7.2], q with Gaussian model-noise v and a linear measurement model^[def. 7.3] h with Gaussian process-noise w .

Definition 7.8 Kalman Filter Model:
Process Model (7.8)

$$\begin{aligned} x^{(k)} &= A[k-1]x^{(k-1)} + u^{(k-1)} + v^{(k-1)} & \text{with} \\ x^{(0)} &\sim \mathcal{N}(x_0, P_0) & \text{and} & \quad v^{(k)} \sim \mathcal{N}(0, Q^{(k)}) \end{aligned} \quad (7.9)$$

Measurement Model

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

and define:

$$\hat{x}_p^{(k)} := E[x_p^{(k)}] \quad \text{and} \quad P_p^{(k)} := V[x_p^{(k)}] \quad (7.10)$$

$$\hat{x}_m^{(k)} := E[x_m^{(k)}] \quad \text{and} \quad P_m^{(k)} := V[x_m^{(k)}] \quad (7.11)$$

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

old Kalman algorithm (in shder Joseph Form I think)

Gaussian Processes (GP)

1. Gaussian Process Regression

add complexity! Only due to inverse

1.1. Gaussian Linear Regression

Given

- ① Linear Model with Gaussian Noise:

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^\top \mathbf{x} \\ \mathbf{y} &= f(\mathbf{x}) + \epsilon \end{aligned} \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}) \quad (8.1)$$

$$\Rightarrow \text{Gaussian Likelihood: } \mathbf{p}(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma_n^2 \mathbf{I})$$

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

Sought

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

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

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

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

Note

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

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

$$\mu_* = \frac{1}{\sigma_n^2} \mathbf{x}_*^\top \Sigma_w^{-1} \mathbf{X}^\top \mathbf{y} \quad \Sigma_* = \mathbf{x}_*^\top \Sigma_w^{-1} \mathbf{x}_*$$
 (8.2)

1.2. Kernelized Gaussian Linear Regression

Definition 8.3 Posterior Predictive Distribution:

$$\mathbf{p}(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \Sigma_*) \quad (8.3)$$

$$\mu_* \quad (8.4)$$

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

$$\mathbb{E}_{\mathbf{x} \sim \mathbf{p}}[g(\mathbf{x})] = \int g(\mathbf{x}) \mathbf{p}(\mathbf{x}) d\mathbf{x}$$

- Normalization constants:

$$\begin{aligned} \mathbf{p}(\theta|\mathbf{y}) &= \frac{1}{Z} \mathbf{p}(\theta, \mathbf{y}) & Z &= \int \mathbf{p}(\mathbf{y}|\theta) \mathbf{p}(\theta) d\theta \\ &= \frac{\mathbf{p}(\mathbf{y}|\theta) \mathbf{p}(\theta)}{Z} & &= \int \mathbf{p}(\theta) \prod_{i=1}^n \mathbf{p}(\mathbf{y}_i|\mathbf{x}_i, \theta) d\theta \end{aligned}$$

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 \mathbf{p} as well as possible.

1. Variational Inference

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

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

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

Definition 9.3 Variational Family of Distributions \mathcal{Q} : a set of probability distributions Q that is parameterized by the same *variational parameter* λ is called a variational family.

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 $\mathbf{p}(\theta|\mathcal{D})$ by a Gaussian probability distribution:

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

the distribution is given by:

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

$$\lambda_1 = \hat{\theta} = \arg \max_{\theta} \mathbf{p}(\theta|\mathbf{y})$$

with

$$\lambda_2 = \Sigma = H^{-1}(\hat{\theta}) = -\nabla \nabla_{\theta} \log \mathbf{p}(\hat{\theta}|\mathbf{y})$$

Corollary 9.1 : Taylor approximation of a function $\mathbf{p}(\theta|\mathbf{y}) \in \mathcal{C}^k$ around its mode $\hat{\theta}$ 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_λ is by minimizing the KL-divergence^[def. 4.8] between our approximate distribution q and our true posterior \mathbf{p} :

$$q_*^* \in \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\theta) \parallel \mathbf{p}(\theta|\mathbf{y})) = \arg \min_{\lambda \in \mathbb{R}^d} \text{KL}(q_\lambda(\theta) \parallel \mathbf{p}(\theta|\mathbf{y}))$$

Note

Usually we want to minimize $\text{KL}(\mathbf{p}(\theta|\mathbf{y}) \parallel q(\theta))$ but this is often infeasible s.t. we only minimize $\text{KL}(q(\theta) \parallel \mathbf{p}(\theta|\mathbf{y}))$

Definition 9.5 ELBO-Optimization Problem [proof 10.12]:

$$\begin{aligned} q_\lambda^* &\in \arg \min_{\{\lambda: q_\lambda \in \mathcal{Q}\}} \text{KL}(q_\lambda(\theta) \parallel \mathbf{p}(\theta|\mathbf{y})) \\ &= \arg \max_{\{\lambda: q_\lambda \in \mathcal{Q}\}} \mathbb{E}_{\theta \sim q_\lambda} [\log \mathbf{p}(\mathbf{y}, \theta)] + H(q_\lambda) \end{aligned} \quad (9.4)$$

$$= \arg \max_{\{\lambda: q_\lambda \in \mathcal{Q}\}} \mathbb{E}_{\theta \sim q_\lambda} [\log \mathbf{p}(\mathbf{y}|\theta)] - \text{KL}(q_\lambda(\theta) \parallel \mathbf{p}(\theta)) \quad (9.5)$$

$$:= \arg \max_{\{\lambda: q_\lambda \in \mathcal{Q}\}} \text{ELBO}(\lambda) \quad (9.6)$$

Attention: Sometimes people write simply \mathbf{p} for the posterior and $\cdot(\cdot)$ 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:

$$\text{ELBO}(q_\lambda) \leq \log \mathbf{p}(\mathbf{y}) \quad (9.7)$$

1.3.1. Maximizing The ELBO

Definition 9.7 Gradient of the ELBO Loss:

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

Problem

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

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

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:

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

Solutions:

- Score Gradients
- Reparameterization Trick: reparameterize a function s.t. it depends on another parameter and reformulate it s.t. it still returns the same value.

1.4. The Reparameterization Trick

Principle 9.1 proof 10.14
Reparameterization Trick: Let ϕ some base distribution from which we can sample and assume there exist an invertible function g s.t. $\theta = g(\epsilon, \lambda)$ then we can write θ in terms of a new distribution parameterized by $\epsilon \sim \phi(\epsilon)$:

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

we can then write by the law of the unconscious statistician law 34.6:

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

\Rightarrow the expectations does not longer depend on λ and we can pull in the gradient!

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

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:

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

Corollary 9.2 proof 10.3

Reparameterized ELBO for Gaussians:

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

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):

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

$$\theta \sim p_{\lambda}(\theta) = \mathbf{F} \quad \text{with} \quad \mathbf{F} = \mathbf{F}^L \circ \dots \circ \mathbf{F}^1$$

$$\mathbf{F}^l = \varphi \circ \mathbf{F}^l = \varphi(\mathbf{W}^l \mathbf{x} + \mathbf{b}^l)$$

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

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

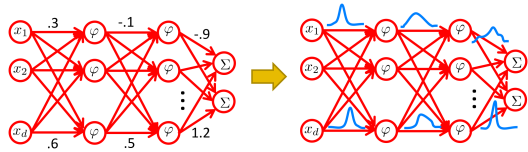


Figure 9

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

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

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. \Rightarrow need to approximate the posterior $p(\theta|\mathbf{y}, \mathbf{X})$ i.e. by fitting a Gaussian distribution to each weight of the posterior neural network.

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

$$\theta^* = \arg \max_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \stackrel{\text{eq. (6.13)}}{=} \arg \min_{\theta} \lambda \|\theta\|_2^2$$

$$- \sum_{i=1}^n \left(\frac{1}{2\sigma(\mathbf{x}_i; \theta)^2} \|y_i - \mu(\mathbf{x}_i; \theta)\|^2 + \frac{1}{2} \log \sigma(\mathbf{x}_i; \theta)^2 \right)$$

Explanation 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:

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

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.
 \Rightarrow 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:

$$p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{f}_{*}, \mathbf{x}_{*}, \mathbf{X}, \mathbf{y})}{p(\mathbf{x}_{*}, \mathbf{X}, \mathbf{y})}$$

$$= \frac{\int p(\mathbf{f}_{*}, \mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) d\mathbf{w}}{\int p(\mathbf{f}_{*}, \mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) d\mathbf{w}}$$

$$\stackrel{\text{eq. (34.19)}}{=} \frac{\int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) p(\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}}{\int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) p(\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}}$$

$$\stackrel{\text{eq. (34.19)}}{=} \frac{\int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}}{\int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}}$$

$$\stackrel{\text{eq. (34.19)}}{=} \int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}$$

$$\stackrel{\text{eq. (34.19)}}{=} \int p(\mathbf{f}_{*}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_{*}, \mathbf{X}, \mathbf{y}) d\mathbf{w}$$

Note ♣

- \mathbf{f}_{*} is independent of $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ given the fixed parameter \mathbf{w} .
- \mathbf{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:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}, \mathbf{w}|\mathbf{X}) d\mathbf{w} = \int p(\mathbf{y}|\mathbf{w}, \mathbf{X}) p(\mathbf{w}|\mathbf{X}) d\mathbf{w}$$

$$\stackrel{\text{eq. (6.6)}}{=} \int p(\mathbf{y}|\mathbf{w}, \mathbf{X}) p(\mathbf{w}) d\mathbf{w}$$

Proof 10.3: Definition 7.4:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) \stackrel{\text{eq. (34.19)}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t_1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1})$$

$$\stackrel{\text{independ.}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1})$$

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

Proof 10.4: Definition 7.5:

$$p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}_{1:t-1}) \stackrel{\text{eq. (34.23)}}{=} \begin{cases} p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1}) p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \\ p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \end{cases}$$

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

from which follows immediately eq. (7.5).

Proof 10.5: Definition 7.6:

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

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

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

Proof 10.6: [cor. 7.3]:

$$p(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) \stackrel{\text{eq. (34.19)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) p(\mathbf{x}_{1:t})$$

$$\stackrel{\text{law 34.2}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) p(\mathbf{x}_t|\mathbf{x}_{t-1:0}) \dots p(\mathbf{x}_2|\mathbf{x}_1) p(\mathbf{x}_1)$$

$$\stackrel{\text{eq. (7.1)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t}) \left(p(\mathbf{x}_1) \prod_{i=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1}) \right)$$

$$\stackrel{\text{law 34.2}}{=} \left(p(\mathbf{y}_1|\mathbf{x}_1) \dots p(\mathbf{y}_t|\mathbf{x}_t) \right) \left(p(\mathbf{x}_1) \prod_{i=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1}) \right)$$

$$\stackrel{[\text{cor. 7.2}]}{=} \left(p(\mathbf{y}_1|\mathbf{x}_1) \dots p(\mathbf{y}_t|\mathbf{x}_t) \right) \left(p(\mathbf{x}_1) \prod_{i=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1}) \right)$$

$$= p(\mathbf{y}_1|\mathbf{x}_1) p(\mathbf{x}_1) \prod_{i=1}^t p(\mathbf{y}_i|\mathbf{x}_i) p(\mathbf{x}_i|\mathbf{x}_{i-1})$$

Proof 10.7: [def. 8.1]

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w}) p(\mathbf{w})$$

$$\propto \exp \left(-\frac{1}{2} \frac{1}{\sigma_n^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right) \exp \left(-\frac{1}{2} \mathbf{w}^T \Sigma^{-1} \mathbf{w} \right)$$

$$\propto \exp \left\{ -\frac{1}{2} \frac{1}{\sigma_n^2} (\mathbf{y}^T \mathbf{y} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} + \sigma_n^2 \mathbf{w}^T \Sigma^{-1} \mathbf{w}) \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \frac{1}{\sigma_n^2} (\mathbf{y}^T \mathbf{y} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{w}^T (\mathbf{X}^T \mathbf{X} + \sigma_n^2 \Sigma^{-1}) \mathbf{w}) \right\}$$

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

$$p(\mathbf{w}|\mathcal{D}) \propto \exp \left(-\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \Sigma_w (\mathbf{w} - \bar{\mathbf{w}}) \right)$$

$$\propto \exp \left(-\frac{1}{2} \left(\mathbf{w}^T \Sigma_w \mathbf{w} - 2\mathbf{w}^T \Sigma_w \bar{\mathbf{w}} + \bar{\mathbf{w}}^T \Sigma_w \bar{\mathbf{w}} \right) \right)$$

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

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

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:

$\mathcal{L}_n(\theta) = \log(p(\theta|\mathbf{y})) = (\log \text{Prior} + \log \text{Likelihood})$
 we now approximate $\mathcal{L}_n(\theta)$ by a Taylor approximation around its maximum $\hat{\theta}$:

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

we can no derive the distribution:

$$p(\theta|\mathbf{y}) \approx \exp(\mathcal{L}_n(\theta)) = \exp(\log p(\theta|\mathbf{y}))$$

$$= p(\hat{\theta}) \exp \left(\frac{1}{2} \frac{\partial^2 \mathcal{L}_n}{\partial \theta^2} \Big|_{\hat{\theta}} \right)$$

$$= \sqrt{2\pi\sigma^2} p(\hat{\theta}) \mathcal{N}(\theta; \hat{\theta}, \sigma) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \mathcal{N}(\theta; \hat{\theta}, \sigma)$$

Notes

- the derivative of the maximum must be zero by definition

$$\frac{\partial \mathcal{L}_n}{\partial \theta} \Big|_{\hat{\theta}} = 0$$
- we approximate the normalization constant $\frac{1}{Z}$ by $\sqrt{2\pi\sigma^2} p(\hat{\theta})$.

Proof 10.10: [def. 9.4] 2D:

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

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

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\hat{\theta}) + (\theta - \hat{\theta})^T H(\hat{\theta}) (\theta - \hat{\theta})$$

$$= \mathcal{L}_n(\hat{\theta}) + \frac{1}{2} Q(\theta)$$

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

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

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

$$\mathcal{L}_n(\theta) \approx \mathcal{L}_n(\hat{\theta}) + (\theta - \hat{\theta})^T \nabla \mathcal{L}_n(\hat{\theta}) (\theta - \hat{\theta})$$

$$H(\theta) = \nabla \nabla^T \mathcal{L}_n(\theta) \quad \Sigma = H^{-1}(\hat{\theta})$$

$$p(\theta|\mathbf{y}) = \sqrt{(2\pi)^n \det(\Sigma)} p(\hat{\theta}) \mathcal{N}(\theta; \hat{\theta}, \Sigma)$$

$$\approx c \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \mathcal{N}(\theta; \hat{\theta}, \Sigma)$$

Proof 10.12: [def. 9.5]

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

Proof 10.13: [def. 9.6]

$$\begin{aligned}
\log p(y) &= \log \int p(y, \theta) d\theta = \log \int p(y|\theta) p(\theta) d\theta \\
&= \log \int p(y|\theta) \frac{p(\theta)}{q_\lambda(\theta)} q_\lambda(\theta) d\theta \\
&= \log \mathbb{E}_{\theta \sim q_\lambda} \left[p(y|\theta) \frac{p(\theta)}{q_\lambda(\theta)} \right] \\
&\stackrel{\text{eq. (34.55)}}{\geq} \mathbb{E}_{\theta \sim q_\lambda} \left[\log \left(p(y|\theta) \frac{p(\theta)}{q_\lambda(\theta)} \right) \right] \\
&= \mathbb{E}_{\theta \sim q_\lambda} \left[\log p(y|\theta) - \log \frac{p(\theta)}{q_\lambda(\theta)} \right] \\
&= \mathbb{E}_{\theta \sim q_\lambda} [\log p(y|\theta)] - \text{KL}(q_\lambda \parallel p(\cdot))
\end{aligned}$$

Proof 10.14: principle 9.1 Let:

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

then it follows immediately with ??:

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

Proof 10.15: [def. 10.3]

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

2. Examples

Example 10.1 Laplace Approximation
Logistic Regression Likelihood + Gaussian Prior:

Example 10.2 ELBO Bayesian Logistic Regression:
Suppose:

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

Then it follows for the terms of the ELBO:

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

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

Example 10.3 ELBO Gradient Gaussian: Suppose:

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

we can reparameterize using principle 9.1 by using:

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

from this it follows: (\mathbf{C} is the Cholesky factor of Σ)

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

from this it follows:

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

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

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

$$\text{Draw a mini batch } \begin{cases} \epsilon^{(1)}, \dots, \epsilon^{(m)} \\ j_1, \dots, j_m \sim \mathcal{U}(\{1, n\}) \end{cases}$$

$$= n \frac{1}{m} \sum_{j=1}^m \nabla_{\mathbf{C}, \mu} \log p(y_j | \mathbf{C}\epsilon + \mu, \mathbf{x}_j)$$

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

Example 10.4 BNN Likelihood Function Examples:

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

Kernels

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

Definition 11.1 Similarity Measure $\text{sim}(A, B)$: A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects.
No single definition of a similarity measure exists but often they are defined in terms of the inverse of distance metrics and they take on large values for similar objects and either zero or a negative value for very dissimilar objects.

Definition 11.2 Dissimilarity Measure $\text{dissim}(A, B)$: Is a measure of how dissimilar objects are, rather than how similar they are.
Thus it takes the largest values for objects that are really far apart from another.
Dissimilarities are often chosen as the squared norm of two difference vectors:
$$\|\mathbf{x} - \mathbf{y}\|^2 = \mathbf{x}^\top \mathbf{x} + \mathbf{y}^\top \mathbf{y} - 2\mathbf{x}^\top \mathbf{y} \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \quad (11.1)$$
$$\text{dissim}(\mathbf{x}, \mathbf{y}) = \text{sim}(\mathbf{x}, \mathbf{x}) + \text{sim}(\mathbf{y}, \mathbf{y}) - 2\text{sim}(\mathbf{x}, \mathbf{y})$$

Attention
It is better to rely on similarity measures instead of dissimilarity measures. Dissimilarities are often not adequate from a modeling point of view, because for objects that are really dissimilar/far from each other, we usually have the biggest problem to estimate their distance.
E.g. for a bag of words it is easy to determine similar words, but it is hard to estimate which words are most dissimilar. For normed vectors the only information of a dissimilarity defined as in eq. (11.1) becomes $2\mathbf{x}^\top \mathbf{y} = 2\text{dissim}(\mathbf{x}, \mathbf{y})$

Definition 11.3 Feature Map ϕ : 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:

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

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 there exists an inner product space^[def. 28.78] called **feature space** $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ and a map $\phi: \mathcal{X} \mapsto \mathcal{V}$ s.t.
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{V}} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad (11.3)$$

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

Example
Let $\mathbf{k}(\mathbf{x}, \mathbf{y}) := \mathbf{x}^\top \mathbf{A} \mathbf{y}$ thus the kernel measures the similarity between \mathbf{x} and \mathbf{y} by the inner product $\mathbf{x}^\top \mathbf{y}$ 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 between the self-similarities $\mathbf{k}(\mathbf{x}, \mathbf{x}) + \mathbf{k}(\mathbf{y}, \mathbf{y})$ and the cross-similarities $\mathbf{k}(\mathbf{x}, \mathbf{y})$:
$$\text{dissimilarity}(\mathbf{x}, \mathbf{y}) := \mathbf{k}(\mathbf{x}, \mathbf{x}) + \mathbf{k}(\mathbf{y}, \mathbf{y}) - 2\mathbf{k}(\mathbf{x}, \mathbf{y})$$

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 $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq \mathcal{X}$. Then the kernel matrix $\mathcal{K} \in \mathbb{R}^{n \times n}$ is defined by:
$$\mathcal{K} = \phi(\mathbf{X})\phi(\mathbf{X}^\top) = (\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n))(\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n))^\top$$
$$= \begin{pmatrix} \mathbf{k}(\mathbf{x}_1, \mathbf{x}_1) & \dots & \mathbf{k}(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(\mathbf{x}_n, \mathbf{x}_1) & \dots & \mathbf{k}(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_1) & \dots & \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_1) & \dots & \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_n) \end{pmatrix}$$
$$\mathcal{K}_{ij} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Corollary 11.3 $\mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$
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:
$$\mathcal{K} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top \quad (11.4)$$

\mathbf{V} : orthogonal matrix of eigenvectors $(\mathbf{v}_{t,i})_{i=1}^n$
 $\mathbf{\Lambda}$: diagonal matrix of eigenvalues λ_i
Assuming all eigenvalues λ_t are non-negative, we can calculate the mapping:
$$\phi: \mathbf{x}_i \mapsto \left(\sqrt{\lambda_t} \mathbf{v}_{t,i} \right)_{t=1}^n \in \mathbb{R}^n, \quad i = 1, \dots, n \quad (11.5)$$
which allows us to define the Kernel \mathcal{K} as:
$$\phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) = \sum_{t=1}^n \lambda_t \mathbf{v}_{t,i} \mathbf{v}_{t,j} = (\mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top)_{i,j} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \quad (11.6)$$

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:
 $\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}(\mathbf{y}, \mathbf{x}) = \phi(\mathbf{x})^\top \phi(\mathbf{y}) = \phi(\mathbf{y})^\top \phi(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$

Property 11.3 Non-negative Eigenvalues/p.s.d.s Form:
Let $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be an n -set of a *finite* input space \mathcal{V} . A kernel \mathbf{k} must induce 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 ?? 28.2.

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 $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, 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 ϕ s.t.

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

Theorem 11.2 General Mercers Theorem: Let Ω be a compact subset of \mathbb{R}^n . Suppose $\mathbf{k}: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ a gernal continuous symmetric function such that the integral operator:
$$T_{\mathbf{k}}: L_2(\mathbf{X}) \mapsto L_2(\mathbf{X}) \quad (T_{\mathbf{k}} f)(\cdot) = \int_{\Omega} \mathbf{k}(\cdot, \mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad (11.8)$$

is **positive**, that is it satisfies:
$$\int_{\Omega \times \Omega} \mathbf{k}(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} > 0 \quad \forall f \in L_2(\Omega)$$

Then we can expand $\mathbf{k}(\mathbf{x}, \mathbf{z})$ in a uniformly convergent series in terms of $T_{\mathbf{k}}$'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.
 \implies can replace the inner product by the kernel function.

Check if R or Rⁿ 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 $\phi(\mathbf{x})^\top \phi(\mathbf{y})$
explicitly but simply use the formula for the kernel:
$$\phi(\mathbf{x})^\top \phi(\mathbf{x}) = \mathbf{k}(\mathbf{x}, \mathbf{y}) \quad (11.9)$$
see examples 11.1 and 11.2

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

4. Types of Kernels

4.1. Stationary Kernels

Definition 11.7 Stationary Kernel: A stationary kernel is a kernel that only considers vector differences:
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}(\mathbf{x} - \mathbf{y}) \quad (11.10)$$
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:
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{k}(\|\mathbf{x} - \mathbf{y}\|_2) \quad (11.11)$$

Corollary 11.4 :
Isotropic \rightarrow Stationary

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

5.1. The Linear Kernel

Definition 11.9 Linear/String Kernel:
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y} \quad (11.12)$$

5.2. The Polynomial Kernel

Definition 11.10 Polynomial Kernel: represents all monomials^[def. 23.5] of degree up to m
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^\top \mathbf{y})^m \quad (11.13)$$

5.3. The Sigmoid Kernel

Definition 11.11 Sigmoid/tanh Kernel:
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \tanh \kappa \mathbf{x}^\top \mathbf{y} - b \quad (11.14)$$

5.4. The Exponential Kernel

Definition 11.12 Exponential Kernel: is an continuous kernel that is non-differential $\mathbf{k} \in \mathcal{C}^0$:
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_1}{\theta}\right) \quad (11.15)$$
 $\theta \in \mathbb{R}$: corresponds to a threshold.

5.5. The Gaussian Kernel

Definition 11.13 Gaussian/Squared Exp. Kernel/Radial Basis Functions (RBF):
Is an infinte dimensional smooth kernel $\mathbf{k} \in \mathcal{C}^\infty$ with some usefull properties
$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\theta^2}\right) \approx \begin{cases} 1 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ close} \\ 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ far away} \end{cases} \quad (11.16)$$

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

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

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

Note
If we chose h small, all data points not close to h will be 0/discarded \iff data points are considered as independent. Length of all vectors in **feature space** is one $\mathbf{k}(\mathbf{x}, \mathbf{x}) = e^0 = 1$.
Thus: Data points in input space are projected onto a high-(infintie)-dimensional sphere in feature space.
Classification: Cutting with hyperplanes 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 $\mathbf{k} \in \mathcal{C}^{[\nu]}$ by a positive parameter ν :

$$\mathbf{k}(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \|\mathbf{x} - \mathbf{y}\|_2}{\rho} \right)^\nu \mathcal{K}_\nu \left(\frac{\sqrt{2\nu} \|\mathbf{x} - \mathbf{y}\|_2}{\rho} \right) \quad (11.17)$$

$\nu, \rho \in \mathbb{R}_+$ ν : Smoothness ρ : Length scale

\mathcal{K}_ν modified Bessel function of the second kind

6. Kernel Engineering

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

6.1. Closure Properties/Composite Rules

Suppose we have two kernels:
 $\mathbf{k}_1: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \quad \mathbf{k}_2: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$
defined on the data space $\mathcal{X} \subseteq \mathbb{R}^d$. Then we may define using Composite Rules:
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_1(\mathbf{x}, \mathbf{x}') + \mathbf{k}_2(\mathbf{x}, \mathbf{x}') \quad (11.18)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_1(\mathbf{x}, \mathbf{x}') \cdot \mathbf{k}_2(\mathbf{x}, \mathbf{x}') \quad (11.19)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \alpha \mathbf{k}_1(\mathbf{x}, \mathbf{x}') \quad \alpha \in \mathbb{R}_+ \quad (11.20)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) f(\mathbf{x}') \quad (11.21)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) \quad (11.22)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = p(\mathbf{k}(\mathbf{x}, \mathbf{x}')) \quad (11.23)$$
$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \exp(\mathbf{k}(\mathbf{x}, \mathbf{x}')) \quad (11.24)$$

Where $f: \mathcal{X} \mapsto \mathbb{R}$ a real valued function
 $\phi: \mathcal{X} \mapsto \mathbb{R}^e$ the explicit mapping
 p a polynomial with pos. coefficients
 \mathbf{k}_3 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 $\Phi = [\phi(\mathbf{x}_1) \dots \phi(\mathbf{x}_n)]^\top \in \mathbb{R}^{d \times n}$.
Thus: $\mathcal{K} = \Phi^\top \Phi \in \mathbb{R}^{n \times n}$.
$$\mathbf{v}^\top \mathcal{K} \mathbf{v} = \mathbf{v}^\top \Phi^\top \Phi \mathbf{v} = (\Phi \mathbf{v})^\top \Phi \mathbf{v} = \|\Phi \mathbf{v}\|_2^2 \geq 0$$

Examples

Example 11.1 Calculating the Kernel by hand:

Let :

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

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

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

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

Operation Count:

- 2 · 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:

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

Operation Count:

- 2 multiplicaitons of $\mathbf{x}_{i1} \mathbf{x}_{j1}$ 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:

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

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??:
 $\mathbf{x} = [x_1 \quad x_n]^T \in \mathcal{S} \subseteq \mathbb{R}^n$ (12.1)

Definition 12.3 Input/Control Variables $\mathbf{u} \in \mathcal{A}$:
Are a variables \mathbf{u} that are directly related to the state space the propagation of to the state variables \mathbf{x} .

Definition 12.4 Output/Masurement Variables/State Observations: $\mathbf{y} \in \mathcal{O}$
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 Measurement/Output/Observation Model h :
Describes the mapping of the state \mathbf{x} onto the output \mathbf{y} .

Definition 12.7 (Discrete) State Space Model:
 $\mathbf{x}^{k+1} = f(t, \mathbf{x}^k, \mathbf{u}^k) \quad t = 1, \dots, K$ (12.2)
 $\mathbf{y}^k = h(t, \mathbf{x}^k, \mathbf{u}^k)$ (12.3)

Markov Models

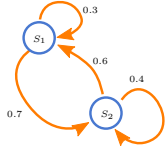
Definition 13.1 States $\mathcal{S} = \{s_1, \dots, s_n\}$:
A state s_i encodes all information of the current configuration of a system.

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. 34.7].
A (S, \mathcal{S}) -valued stochastic process $X = \{X_t : \Omega \rightarrow S\}_{t \in I}$ adapted to the filtration is said to possess the Markov property if:
$$\mathbb{P}(X_t \in A | \mathcal{F}_s) = \mathbb{P}(X_t \in A | X_s) \quad \forall A \in \mathcal{S} \quad s, t \in I \quad \text{s.t. } s < t$$
 (13.1)

1. Markov Chains

Definition 13.3 Markov Chain:
Is a sequence of random variables $\{X_i\}_{i \in \mathcal{T}}$ ^[def. 38.3] 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:
 $q_0(s_i) = \mathbb{P}(X_0 = s_i) \quad \forall s_i \in \mathcal{S}$
 $\iff \mathbf{q}_0 = [q_0(s_1) \quad \dots \quad q_0(s_n)]$ (13.2)

Definition 13.5 Transition Probability $\mathbb{P}_{ji}(t)$:
is the probability of a random variable X_t in state s_i to transition into state s_j :
 $\mathbb{P}_{ij}(t) = \mathbb{P}(X_{t+1} = s_j | X_t = s_i) \quad \forall s_i, s_j \in \mathcal{S}$ (13.3)

Definition 13.6 n^{th} Transition Probability $\mathbb{P}_{ji}^{(n)}(t)$:
denotes the probability of reaching state s_j from state s_i in n steps:
 $\mathbb{P}_{ij}^{(n)}(t) = \mathbb{P}(X_{t+n} = s_j | X_t = s_i) \quad \forall s_i, s_j \in \mathcal{S} \quad (13.4)$

Definition 13.7 Transition Matrix $\mathbb{P}(t)$:
The transition probabilities eq. (13.4) can be represented by a *row-stochastic matrix*?? $\mathbb{P}(t)$ where the i^{th} row represents the transition probabilities for the i^{th} state s_i i.e.

	To j
From i	$\begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$

Corollary 13.1 Row stochastic matrices and Graphs:
Row stochastic matrices?? represent graphs where the outgoing edges must sum to one:
 $\sum \delta^+(s_i) = 1$ (13.5)

1.1. Simulating Markov Chains

Corollary 13.2 proof 13.1
Realization of a Markov Chain:
 $\mathbb{P}(X_0 = x_0, \dots, X_N = x_N) = q_0(x_1) \prod_{n=1}^N \mathbb{P}_{n-1,n}(t)$

Algorithm 13.1 Forward Sampling:
Input: $\mathbf{q}(x_0)$ and \mathbb{P}
Output: $\mathbb{P}(X_{0:N})$
Sample $x_0 \sim \mathbb{P}(X_0)$
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} :
 $q_{n+1}(s_j) = \mathbb{P}(X_{n+1} = s_j) \quad \forall s_i \in \mathcal{S}$
$$= \sum_{i=1}^n \mathbb{P}(X_n = s_i) \mathbb{P}(X_{n+1} = s_j | X_n = s_i)$$

$$= \sum_{i=1}^n q_n(s_i) \mathbb{P}_{ij}(t)$$
 (13.6)
$$\mathbf{q}_{n+1} = [q_{n+1}(s_1) \quad \dots \quad q_{n+1}(s_n)]$$

$$= \mathbf{q}_n \mathbb{P}(t)$$

$$= [q_n(s_1) \quad \dots \quad q_n(s_n)] \begin{bmatrix} \mathbb{P}_{1,1} & \mathbb{P}_{1,2} & \dots & \mathbb{P}_{1,n} \\ \mathbb{P}_{2,1} & \mathbb{P}_{2,2} & \dots & \mathbb{P}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}_{n,1} & \mathbb{P}_{n,2} & \dots & \mathbb{P}_{n,n} \end{bmatrix} (t)$$

Corollary 13.3 [proof 13.2]
Time-homogeneous Markov Transition Probabilities:
 $\mathbf{q}_{n+1} = \mathbf{q}_0 \mathbb{P}^{n+1}$ (13.7)

Definition 13.9 Stationary Distribution:
A markov chain has a stationary distribution if it satisfies:
 $\lim_{N \rightarrow \infty} q_N(s_i) = \lim_{N \rightarrow \infty} \mathbb{P}(X_N = s_i) = \pi_i \quad \forall s_i \in \mathcal{S}$
 $\lim_{N \rightarrow \infty} \mathbf{q}_N = [\pi_1 \quad \dots \quad \pi_n] \iff \mathbf{q} = \mathbf{q} \mathbb{P}(N)$ (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*!

edit 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:
 $\mathbb{P}_{ij}^{(n)}(t) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$ (13.9)

Definition 13.11 Accessible State $s_i \rightarrow s_j$:
A state s_j is accessible from state s_i iff:
 $\exists n : \mathbb{P}_{ij}^{(n)}(t) > 0$ (13.10)

Definition 13.12 Communicating States $s_i \leftrightarrow s_j$:
Two states s_j and s_i are communicating iff:
 $\exists n_1 : \mathbb{P}_{ij}^{(n_1)}(t) > 0 \quad \wedge \quad \exists n_2 : \mathbb{P}_{ji}^{(n_2)}(t) > 0$ (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:
 $k = \gcd\{n > 0 : \mathbb{P}_{ii}^{(n)} = \mathbb{P}(X_n = s_i | X_0 = s_i) > 0\}$ (13.12)

Definition 13.14 Aperiodic State $k = 1$:
Is a state s_i with periodicity^[def. 13.13] of one $\iff k = 1$

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 (Definition 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:
 $\mathbb{P}_{ji} = \mathbb{P}(X_t = s_j | X_{t-1} = s_i) = \mathbb{P}(X_{t-\tau} = s_j | X_{t-\tau} = s_i) \quad \forall \tau \in \mathbb{N}_0$ (13.13)

Corollary 13.7 \mathbb{P}
Transition Matrices of Stationary MCs:
Transition matrices of time-homogeneous markov chain are constant/time independent:
 $\mathbb{P}(t) = \mathbb{P}$ (13.14)

Definition 13.16 Aperiodic Markov Chain: Is a markov chain where all states are aperiodic:
 $\gcd\{n > 0 : \mathbb{P}_{ii}^{(n)} = \mathbb{P}(X_n = s_i | X_0 = s_i) > 0\} = 1$
 $\forall i \in \{1, \dots, n\}$ (13.15)

Definition 13.17 Irreducible Markov Chain: Is a Markov chain that has only *communicating states*^[def. 13.12]:
 $s_j \leftrightarrow s_i \quad \forall i, j \in \{1, \dots, n\}$ (13.16)

- \implies no sinks^[def. 13.10]
- \implies every state can be reached from every other state

Corollary 13.8 : An *irreducible*^[def. 13.17] markov chain is automatically *aperiodic*^[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-irreducible* 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 Irreducible^[def. 13.17]
- 2 Aperiodic^[def. 13.16]

Corollary 13.10 Stationary Distribution: An ergodic markov chain has a *unique* stationary distribution^[def. 13.9] and converges to it starting from any initial state $q_0(s_i)$

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]
 $\mathbb{P}(X_0 = x_0, \dots, X_N = x_N) = \mathbb{P}(X_0 = x_0) \cdot \mathbb{P}(X_1 = x_1 | X_0 = x_0) \cdot \mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0) \cdot \dots \cdot \mathbb{P}(X_N = x_N | X_{N-1} = x_{N-1}, \dots, X_0 = x_0)$
and then simply use the Markovian property

Proof 13.2: Corollary 13.3
 $\mathbf{q}_{n+1} = \mathbb{P} \mathbf{q}_n = (\mathbf{q}_{n-1} \mathbb{P}) \mathbb{P} = \mathbf{q}_0 \mathbb{P}^{n+1}$

3. Examples

Example 13.1 Ergodic Markov Chain:

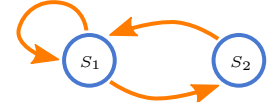
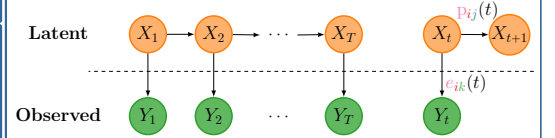


Figure 10: Ergodic for $N = 2$ (can reach s_2 at any $t \leq N$ after $N = 2$)

Hidden Markov Model (HMM)

Definition 14.1 $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{P}, \mathbb{E})$
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] $\mathbb{P}(s_i, s_j)$
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 :
 $e_{ij}(t) = \mathbb{P}(Y_t = o_j | X_t = s_i) \quad \begin{cases} \forall o_i \in \mathcal{O} \\ \forall s_j \in \mathcal{S} \end{cases}$ (14.1)

Markov Decision Processes (MDP)

Definition 15.1 $(S, \mathcal{A}, \mathbb{P}_a, R_a)$
Markov Decision Process (MDP): A markov decision process is a *controlled* markov process/chain with an associated reward, where the transition can be steered by an actions. It is characterized by the 4-tuple of:

- States^[def. 13.1] $S = \{s_1, \dots, s_n\}$
- Actions^[def. 15.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
- Transition Probabilities^[def. 15.3] $\mathbb{P}_a(s_i, s_j)$
- 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 $\mathbb{P}_a(s_j, s_i)(t)$:
 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 :

$$\mathbb{P}_a(s_j, s_i) = \mathbb{P}(s_j | s_i, a) = \mathbb{P}(x_{t+1} = s_j | x_t = s_i, a_t = a) \quad \forall s_i, s_j \in S, \forall a \in \mathcal{A} \quad (15.1)$$

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) :

$$(x_{t+1}, x_t, a) \mapsto R_{t+1} \in \mathcal{R} \subset \mathbb{R} \quad (15.2)$$

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:

$$\pi : S \mapsto \mathcal{A} \quad \text{or} \quad \pi : S \mapsto \mathbb{P}(\mathcal{A}) \quad (15.3)$$

Definition 15.6 Discounting Factor γ :
 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:

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \right] \quad (15.4)$$

Definition 15.8
Value Function $V^{\pi}(x)$:
 Is the *discounted* expected reward^[def. 15.4] of the whole markov process given an initial state $X_0 = x$:

$$V^{\pi}(x) = J(\pi) | X_0 = x = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \quad (15.5)$$

$$= \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \quad (15.6)$$

$$= \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \quad (15.7)$$

1.1. Calculating the value of V^{π}

Definition 15.9 [proof 16.1]
Value Iteration:

$$V^{\pi}(x) = J(\pi) | X_0 = x \quad (15.8)$$

$$= \mathbb{E}_{x'|x, \pi(x)} [r(x, \pi(x)) + \gamma V^{\pi}(x')] \quad (15.9)$$

$$= r(x, \pi(x)) + \gamma \mathbb{E}_{x'|x, \pi(x)} [V^{\pi}(x')] \quad (15.10)$$

$$= r(x, \pi(x)) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, \pi(x)) V^{\pi}(x') \quad (15.11)$$

We can now write this for all possible initial states as:

$$V^{\pi} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} V^{\pi} \iff (\mathbf{I} - \gamma \mathbf{P}^{\pi}) V^{\pi} = \mathbf{r}^{\pi} \quad (15.9)$$

with:

$$V^{\pi} = \begin{bmatrix} V^{\pi}(s_1) \\ \vdots \\ V^{\pi}(s_n) \end{bmatrix} \quad \mathbf{r}^{\pi} = \begin{bmatrix} r^{\pi}(s_1, \pi(s_1)) \\ \vdots \\ r^{\pi}(s_n, \pi(s_n)) \end{bmatrix}$$

$$\mathbf{P}^{\pi} = \begin{bmatrix} \mathbb{P}(s_1 | s_1, \pi(s_1)) & \mathbb{P}(s_2 | s_1, \pi(s_1)) & \dots & \mathbb{P}(s_n | s_1, \pi(s_1)) \\ \mathbb{P}(s_1 | s_2, \pi(s_2)) & \mathbb{P}(s_2 | s_2, \pi(s_2)) & \dots & \mathbb{P}(s_n | s_2, \pi(s_2)) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}(s_1 | s_n, \pi(s_n)) & \mathbb{P}(s_2 | s_n, \pi(s_n)) & \dots & \mathbb{P}(s_n | s_n, \pi(s_n)) \end{bmatrix}$$

1.1.1. Direct Methods

Corollary 15.1 LU-decomposition $\mathcal{O}(n^3)$:
 The linear system from eq. (15.9):

$$(\mathbf{I} - \gamma \mathbf{P}^{\pi}) V^{\pi} = \mathbf{r}^{\pi} \quad (15.10)$$
 can be solved *directly* using Gaussian elimination in polynomial time $\mathcal{O}(n^3)$.

Note – invertibility

If $\gamma < 1$ then $(\mathbf{I} - \gamma \mathbf{P}^{\pi})$ is full-rank/invertible as EVs(\mathbf{P}^{π}) ≤ 1 .

1.1.2. Fixed Point Iteration

Corollary 15.2 Fixed-Point Iteration $\mathcal{O}(n \cdot |S|)$:
 The linear system from eq. (15.9) can be solve using *fixed-point iteration*^[def. 31.30] in at most $\mathcal{O}(n \cdot |S|)$ (if every state s_i is connected to every other state $s_j \in S$)

Algorithm 15.1 Fixed Point Iteration:
Input: Initial Guess: $V_0^{\pi} \stackrel{\text{i.e.}}{=} 0$
 1: **for** $t = 1, \dots, T$ **do**
 2: Use the fixed point method:

$$V_t^{\pi} = \phi V_t^{\pi} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} V_{t-1}^{\pi} \quad (15.11)$$

 3: **end for**

Corollary 15.3
Policy Iteration Contraction [proof 16.2]:
 Fixed point iteration of policy iteration is a contraction^[def. 28.63] that leads to a fixed point V^{π} with a rate depending on the discount factor γ .

$$\|V_t^{\pi} - V^{\pi}\| = \|\phi V_{t-1}^{\pi} - \phi V^{\pi}\| \leq \gamma \|V_{t-1}^{\pi} - V^{\pi}\| = \gamma^t \|V_0^{\pi} - V^{\pi}\| \quad (15.12)$$

Explanation 15.2.

- $\gamma \downarrow$: the less we plan ahead/the smaller we choose γ 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 γ the longer it takes to converge but we will explore all possibilities. But for to large γ we will simply keep exploring without sticking to a optimal point

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 π ? **Idea** compute $J(\pi)$ for every possible policy:

$$\pi^* = \arg \max_{\pi} J(\pi) \quad (15.13)$$

Problem this is unfortunately infeasible as there exist $m^n = |\mathcal{A}|^{|S|}$ policies that we need to calculate the value for.

Note

The problem is that J/V^{π} depend on π but if we do not know π yet we cannot compute those.

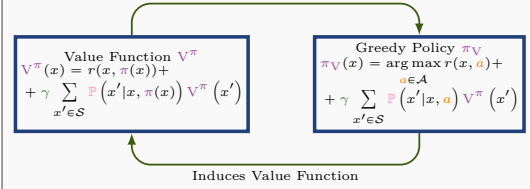
1.2.1. Greedy Policy

Definition 15.10 Greedy Policy:
Assuming we know $V^{\pi^{t-1}}$ then we could choose a greedy policy:

$$a^* = \pi_t(x) \quad (15.14)$$

$$:= \arg \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, a) V^{\pi^{t-1}}(x')$$

- Given a policy π however we can calculate a value function V^{π}
- Given a value function V we can induce a greedy policy^[def. 15.10] π w.r.t. V Induces Policy



Theorem 15.1 Optimality of Policies [Bellman]:
 A policy π_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):

$$V^*(x) = \max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, a) V^*(x') \right] \quad (15.15)$$

$$:= \max_{a \in \mathcal{A}} Q^*(x, a) \quad (15.16)$$

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: π
 1: **while** Not converged $t = t + 1$ **do**
 2: Compute $V^{\pi^t}(x)$

$$V^{\pi^t}(x) = r(x, \pi(x)) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, \pi_t(x)) V^{\pi^t}(x')$$

 3: Compute greedy policy π_G :

$$\pi_G(x) = \arg \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, a) V^{\pi^t}(x')$$

 4: Set $\pi_{t+1} \leftarrow \pi_G$
 5: **end while**

Algorithm 15.2

Pros

- Monotonically improves $V^{\pi^t} \geq V^{\pi^{t-1}}$
- is guaranteed to converge to an optimal policy/solution π^* in polynomial #iterations: $\mathcal{O}\left(\frac{n^2 m}{1-\gamma}\right)$

Cons

- Complexity *per iteration* requires to evaluate the policy V^{π} 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: $V_0(x) = \max_{a \in \mathcal{A}} r(x, a)$
 1: **for** $t = 1, \dots, \infty$ **do**
 2: Compute:

$$Q_t(x, a) = r(x, a) + \gamma \sum_{x' \in S} \mathbb{P}(x' | x, a) V_{t-1}(x') \quad \forall a \in \mathcal{A} \quad \forall x \in S$$

 3: for all $x \in S$ let:

$$V_t(x) = \max_{a \in \mathcal{A}} Q_t(x, a)$$

 4: **if** $\max_{x \in S} |V_t(x) - V_{t-1}(x)| \leq \epsilon$ **then**
 break
 5: **end if**
 6: **end for**
 7: Choose greedy policy π_{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 ϵ optimal policy:

$$\|V_t - V^*\|_{\infty} \leq \gamma^t \|V_0 - V^*\|_{\infty} \quad (15.17)$$

$$\implies t \approx \ln \frac{\gamma}{\epsilon} \|V_0 - V^*\|_{\infty} \quad \text{for} \quad \|V_t - V^*\|_{\infty} \leq \epsilon$$

Algorithm 15.3

Pros

- Finds ϵ -optimal solution in polynomial #iterations $\mathcal{O}(\ln \frac{1}{\epsilon})$ ^[cor. 15.4].
- Complexity *per iteration* requires us to solve a linear system $\mathcal{O}(m \cdot n \cdot s) = \mathcal{O}(|\mathcal{A}| \cdot |S| \cdot s)$ 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 $\mathcal{O}(n) = \mathcal{O}(|S|)$

Cons

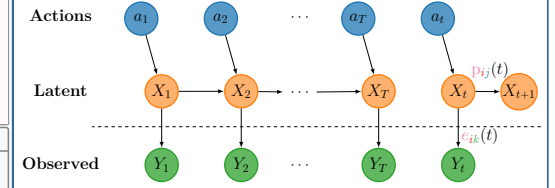
- Only ϵ -optimal solution.

Partially Observable MDP (POMDP)

Definition 16.1 $(S, \mathcal{A}, \mathcal{O}, \mathbb{P}_a, E, R_a)$

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:

- States^[def. 13.1] $S = \{s_1, \dots, s_n\}$
- Actions^[def. 15.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
- Observations^[def. 14.2] $\mathcal{O}/\mathcal{O}_{s_j} = \{o_1, \dots, o_m\}$
- Transition Probabilities^[def. 15.3] $\mathbb{P}_a(s_i, s_j)$
- Emission/Output Probabilities^[def. 14.3] $e_{ij}(t)$
- 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:
 $H_t = \{\{a_0, o_0, r_0\}, \dots, \{a_{t-1}, o_{t-1}, r_{t-1}\}\}$

Definition 16.3 Belief State Space \mathcal{B} : Is a $|\mathcal{S}| - 1$ dimensional simplex or $(|\mathcal{S}|)$ -dimensional probability vector^[def. 28.71] whose elements b are probabilities:

$$\mathcal{B} = \Delta(|\mathcal{S}|) = \left\{ b_t \in [0, 1]^{|\mathcal{S}|} \mid \sum_{x=1}^n b_t(x) = 1 \right\} \quad (16.1)$$

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:

$$\begin{aligned} b_{t+1}(s_i) &= \mathbb{P}(X_{t+1} = s_i | Y_{t+1} = o_k) \\ &= \frac{1}{Z} \mathbb{P}(Y_{t+1} = o_k | X_{t+1} = s_i, a_t) \\ &\quad \cdot \sum_{s_j} b_t(s_j) \mathbb{P}(X_{t+1} = s_i | X_t = s_j, a_t) \quad (16.2) \\ &= \sum_{s_j \in \mathcal{S}} P a(s_i) \end{aligned}$$

Definition 16.6 Stochastic Observation Model:

$$\mathbb{P}(Y_{t+1} = o_k | b_t, a_t) = \sum_{s_i \in \mathcal{S}} b_t(s_i) \mathbb{P}(Y_{t+1} = o_k | X_t = s_i, a_t) \quad (16.3)$$

1.2. Reward Function

Definition 16.7 POMDP Reward Function:

$$r(b_t, a_t) = \sum_{s_j \in \mathcal{S}} b_t(s_j) r(s_j, a_t) \quad (16.4)$$

Note

For finite horizon T , the set of reachable belief states is finite however exponential in T .

add definition of simplex to math appendix which is basically this.

2. Proofs

2.1. Markov Decision Processes

Proof 16.1: [def. 15.8]

$$\begin{aligned} V^\pi(x) &= \mathbb{E}_{X_{1:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &= \mathbb{E}_{\mathbf{X}} \left[\gamma^0 r(X_0, \pi(X_0)) + \sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &\stackrel{\gamma^0=1}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}} \left[\sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x \right] \\ &\stackrel{\text{re-index}}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}} \left[\sum_{t=0}^{\infty} \gamma^{t+1} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x \right] \\ &= r(x, \pi(x)) + \gamma \mathbb{E}_{\mathbf{X}} \left[\sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x \right] \\ &= r(x, \pi(x)) + \gamma \mathbb{E}_{X_1} \left[\mathbb{E}_{X_{2:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \mid X_1 = x' \right] \mid X_0 = x \right] \\ &\stackrel{\text{law}}{=} r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}(x' | x, \pi(x)) \mathbb{E}_{X_{2:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \mid X_1 = x' \right] \\ &\stackrel{\text{eq. (13.13)}}{=} r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}(x' | x, \pi(x)) \mathbb{E}_{X_{2:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x' \right] \\ &= r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}(x' | x, \pi(x)) \underline{V^\pi(x')} \end{aligned}$$

Proof 16.2 [cor. 15.3]: Consider $V, V' \in \mathbb{R}^n$ and let ϕ :
 $\phi x := r^\pi + \gamma P^\pi x \implies \phi V^\pi = V^\pi$

then it follows:

$$\begin{aligned} \|\phi V - \phi V'\| &= \left\| \cancel{r^\pi} + \gamma P^\pi V - \cancel{r^\pi} - \gamma P^\pi V' \right\| \\ &= \left\| \gamma P^\pi (V - V') \right\| \\ &\stackrel{\text{eq. (28.91)}}{\leq} \gamma \|P^\pi\| \cdot \|V - V'\| \\ &\stackrel{\text{i.e. } L_2}{\leq} \gamma \cdot 1 \cdot \|V - V'\|_2 \end{aligned}$$

Proof 16.3: algorithm 15.3

$$\begin{aligned} V_0(x) &= \max_{a \in \mathcal{A}} r(x, a) \\ V_1(x) &= \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}(x' | x, a) V_0(x') \\ V_{t+1}(x) &= \max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}(x' | x, a) V_t(x') \end{aligned}$$

Proof 16.4: [cor. 15.4] Let $\phi : \mathbb{R}^n \mapsto \mathbb{R}^n$, with:

$$(\phi V^*) (x) = Q(x, a) = \max_a \left[r(x, a) + \gamma \sum_{x'} \mathbb{P}(x' | x, a) \right]$$

Bellman's theorem 15.1

and consider $V, V' \in \mathbb{R}^n$

$$\begin{aligned} \|\phi V - \phi V'\|_\infty &= \max_x \left| (\phi V)(x) - (\phi V')(x) \right| \\ &= \max_x \left| \max_a Q(x, a) - \max_{a'} Q'(x, a') \right| \\ &\stackrel{\text{Property 23.9}}{\leq} \max_x \max_a \left| Q(x, a) - Q'(x, a) \right| \\ &= \max_{x, a} \left| \cancel{r} + \gamma \sum_{x'} \mathbb{P}(x' | x, a) V(x') - \cancel{r} - \gamma \sum_{x'} \mathbb{P}(x' | x, a) V'(x') \right| \\ &= \gamma \max_{x, a} \left| \sum_{x'} \mathbb{P}(x' | x, a) (V(x') - V'(x')) \right| \\ &\stackrel{\leq 1}{\leq} \gamma \max_{x, a} \left| \sum_{x'} \mathbb{P}(x' | x, a) \right| \cdot \left| (V(x') - V'(x')) \right| \\ &\stackrel{\text{eq. (28.91)}}{\leq} \gamma \max_{x, a} \left| \sum_{x'} \mathbb{P}(x' | x, a) \right| \cdot \|V(x') - V'(x')\|_\infty \\ &\leq \gamma \cdot 1 \cdot \|V(x') - V'(x')\|_\infty \end{aligned}$$

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 :

$$\begin{aligned} b_{t+1}(s_i) &= \mathbb{P}(X_{t+1} = s_i | y_{t+1}) \\ &= \frac{1}{Z} \mathbb{P}(y_{1:t+1} | s_i) \sum_{j=1}^n \underbrace{\mathbb{P}(X_t = s_j | y_{1:t})}_{b_t(s_j)} \mathbb{P}(s_i | s_j) \end{aligned}$$

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 τ :
Is a set of consecutive 3-tuples of states, actions and rewards:
 $\tau = \{s_t, a_t, r_t\} \quad t = 1, \dots, \tau$ (17.1)

Definition 17.6 Episodic Learning: Is a setting where we generate multiple K -episodes of different trajectories $\{\tau^{(k)}\}_{i=1}^K$ 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] $p_a(s_i, s_j)$
 - the reward function^[def. 15.4] $r(b_t, a_t)$
- 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. 35.23]:

$$N_{s_i|s_j,a} := \sum_{k=1}^t \delta(X_{k+1}=s_i|X_k=s_j, A_k=a) \quad (17.2)$$

$$N_{s_j,a} := \sum_{k=1}^t \delta(X_k=s_j, A_k=a) \quad (17.3)$$

$$p_a(s_i, s_j) \approx \frac{N_{s_i|s_j,a}}{N_{s_j,a}} \quad (17.4)$$

$$r(s_i, a) \approx \frac{1}{N_{s_i,a}} \sum_{k=1}^t \delta(X_k=s_i, A_k=a) r(X_k, A_k) \quad (17.5)$$

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. ϵ_t Greedy Learning

Algorithm 17.1 Epsilon Greedy Learning:

```
1: for  $t = 1, \dots, T$  do
2:   Pick next action
      $a_t = \begin{cases} \arg \max_a Q_t(a) & \text{with probability } \epsilon_t \\ \text{random } a & \text{with probability } 1 - \epsilon_t \end{cases}$ 
3: end for
```

Corollary 17.1 Necessary Condition for Convergence:

If the sequence ϵ_t satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \epsilon_t < \infty, \quad \sum_t \epsilon_t^2 < \infty \quad (\text{i.e. } \epsilon_t = 1/t) \quad (17.6)$$

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:

$$\hat{r}(s_t, a) = R_{\max} \quad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1 \quad (17.7)$$

Set min. number Δ of observations for policy update

Compute Policy π_1 of the MDP^[def. 15.1] using (\hat{p}, \hat{r}) :
 π_t

```
1: for  $k = 1, \dots, K$  do
2:   Choose  $a = \pi_t(x_t)$  and observe  $(s, r)$ 
3:   Calculate:
        $N_{\mathbf{x}_t, a} + = 1 \quad r(x_t, a) + = r(x_t, a) \quad (17.8)$ 
        $N_{\mathbf{x}_{t+1}|\mathbf{x}_t, a} + = 1 \quad (17.9)$ 
```

```
4:   if  $k == \Delta$  then
5:     Re-calculate (based on eqs. (17.4) and (17.5)):
        $\hat{r}(s_t, a) = R_{\max} \quad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1$ 
       and update the policy  $\pi_t = \pi_t(\hat{p}, \hat{r})$ 
```

```
6:   end if
```

```
7: end for
```

Note

Other ways of updating the policy at certain 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 π_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:

$$\mathbb{P}(\hat{r}(s, a) - r(s, a) \leq \epsilon) \geq 1 - \delta \iff n \in \mathcal{O}\left(\frac{R_{\max}^2}{\epsilon^2} \log \frac{1}{\delta}\right) \quad (17.10)$$

Theorem 17.1 : Every T timesteps, with high probability, R_{\max} either:

- Obtain near optimal reward, or
- Visits at least one unknown state-action pair

Theorem 17.2 Performance of R-max: With probability $\delta - 1$, R_{\max} will reach an ϵ -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 initial policy π and s.t. we have $\hat{V}_0^{\pi}(s_j)$.

Goal: want to calculate an unknown value function V^{π} .

If the reward and the next states are stochastic variables (R, X) we can calculate the reward using eq. (15.8):

$$\hat{V}^{\pi}(x_t) = \mathbb{E}_{X_{t+1}, R} [R + \gamma \hat{V}^{\pi}(X') | X, a] \quad (17.11)$$

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. 36.6] 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 \Rightarrow average with previous estimate.

Definition 17.7 Temporal Difference (TD) Learning:

$$\hat{V}(x_{t+1}) = (1 - \alpha_t) \hat{V}(x_t) + \alpha_t (r + \gamma \hat{V}(x_{t+1})) \quad (17.12)$$

Corollary 17.2 Necessary Condition for Convergence:
If the learning rate α_t satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \alpha_t < \infty, \quad \sum_t \alpha_t^2 < \infty \quad (\text{i.e. } \alpha_t = 1/t) \quad (17.13)$$

and all state-action pairs (s_i, a_j) are chosen infinitely often, then we converge to the correct value function:

$$\mathbb{P}(\hat{V} \rightarrow \hat{V}^{\pi}) = 1 \quad (17.14)$$

2.2. Q-Learning

Definition 17.8 Action Value/Q-Function:

$$Q \quad (17.15)$$

2.2.1. Policy Gradients

2.2.2. Actor-Critic Methods

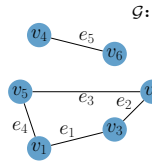
3. Proofs

Proof 17.1: proposition 17.2 using hoeffdings bound^[def. 34.38] with δ and $b - a = R_{\max}$.

Graph Theory

Definition 18.1 Graph

A graph \mathcal{G} is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a finite set of vertices \mathcal{V} ^[def. 40.4] and a multi set^[def. 19.3] of edges \mathcal{E} ^[def. 40.10].



Definition 18.2 Order

$n = |\mathcal{V}|$: The order of a graph is the cardinality of its vertex 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. 40.1] of order n .

Corollary 18.2 (p, q) -Graph: Is a graph \mathcal{G} ^[def. 40.1] 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 vertex $v_i \in \mathcal{V}$ is the set of all adjacent vertices:

$$N(v_i) = \{v_k \in \mathcal{V} : \exists e_k = \{v_i, v_j\} \in \mathcal{E}, \forall v_j \in \mathcal{E}\} \quad (18.1)$$

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:

$$\mathbf{A}_{i,j} := \begin{cases} 1 & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (18.2)$$

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:

$$\mathbf{A}_{i,j} := \begin{cases} \theta_{ij} & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (18.3)$$

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 Vertex

δ : The degree of a vertex v is the cardinality of the neighborhood^[def. 40.5] – the number of adjacent vertices:

$$\deg(v_i) = \delta(v) = |N(v)| = \sum_{j=1}^{j \leq i} \mathbf{A}_{i,j} \quad (18.4)$$

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:

$$\mathbf{D}_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (18.5)$$

2. Edges

Definition 18.10 Edges

$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$: Represent some relation between edges^[def. 40.4] and are represented by two-element subset sets of the vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ and } v_j \text{ connected} \quad (18.6)$$

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:

$$U \subseteq \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \quad (18.7)$$

4. Components

Definition 18.12 Component: A connected component of a graph \mathcal{G} is a *connected*^[def. 40.20] subgraph^[def. 40.11] 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:

$$W = \{v_k, v_{k+1}\}_{k \in \mathcal{K}} \quad (18.8)$$

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 vertices 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. 40.15] of a graph \mathcal{G} where the last visited vertex is the one from which we started.

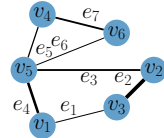
6. Different Kinds of Graphs

7. Weighted Graph

Definition 18.17 Weighted Graph:

Is a graph \mathcal{G} where edges are associated with a weight:

$$\exists \theta_i := \text{weight}(e_i) \quad \forall e_i \in \mathcal{E}$$



8. Spanning Graphs

Definition 18.18 Spanning Graph:

Is a subgraph^[def. 40.11] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for which it holds:

$$U = \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \quad (18.9)$$



8.1. Minimum Spanning Graph

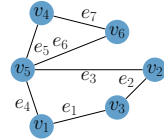
Definition 18.19 Minimum Spanning Graph: Is a spanning graph^[def. 40.18] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with minimal weights/distance of the edges.

9. Connected Graphs

Definition 18.20 (Weakly) Connected Graph:

Is a graph \mathcal{G} ^[def. 40.1] where there exists a path between any two vertices:

$$\exists P(v_i, \dots, v_j) \quad \forall v_i, v_j \in \mathcal{V} \quad (18.10)$$



Corollary 18.3 Strongly Connected Graph: A directed Graph^[def. 40.22] is called strongly connected if every nodes is *reachable* from every other node.

Corollary 18.4 Components of Connected Graphs: A connected Graph^[def. 40.20] consist of one component $c(\mathcal{G}) = 1$.

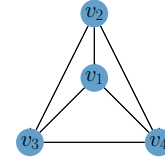
9.1. Fully Connected/Complete

Definition 18.21 Fully Connected/Complete Graph:

Is a connected graph \mathcal{G} ^[def. 40.20] where each node is connected to every other node.

$$\exists e \forall \{v_i, v_j\} \quad \forall v_i, v_j \in \mathcal{V} \quad (18.11)$$

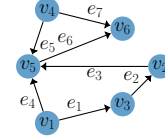
$$|\mathcal{V}| = \frac{1}{2}|\mathcal{V}|(|\mathcal{V}| - 1) \quad (18.12)$$



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. 40.23].



Definition 18.23 Directed Edges/Arcs: Represent some *directional* relationship between edges^[def. 40.4] and are represented by *ordered* two-element subset sets of vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ goes to } v_j \quad (18.13)$$

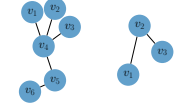
9.3. Trees And Forests

9.3.1. Acyclic Graphs

Definition 18.24 Acyclic Graphs: Are graphs^[def. 40.1] where no cycles^[def. 40.16] exist.

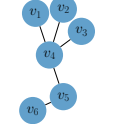
Definition 18.25 Forests:

Are acyclic graphs^[def. 40.24]:



Definition 18.26 Trees:

Are acyclic graphs^[def. 40.24] that are connected^[def. 40.20].



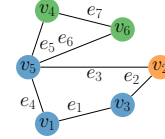
10. Graph Layering

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. 40.4] into subsets

$$\{\mathcal{V}_1, \dots, \mathcal{V}_L\} \subseteq \mathcal{V}$$

$$\text{s.t.} \quad \mathcal{V} = \mathcal{V}_1 \cup \dots \cup \mathcal{V}_L \quad (18.14)$$



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:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \quad l_{ij} := \begin{cases} -1 & \text{if } i \neq j \text{ and } e_{ij} \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } e_{ij} \notin \mathcal{E} \\ \deg(v_i) & \text{if } i = j \end{cases} \quad (18.15)$$

Corollary 18.5 title:

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 $\{\}/\emptyset$: 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] 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 $2^{|S|}$.

1. Closure

Definition 19.7 Closure: A set is *closed* under an operation Ω if performance of that operations onto members of the set always produces a member of that set.

2. Open vs. Closed Sets

Definition 19.8 Open Sets:

- Euclidean Spaces**: A subset $U \in \mathbb{R}$ is open, if for every $x \in U$ it exists $\epsilon(x) \mathbb{R}_+$ s.t. a point $y \in \mathbb{R}$ belongs to U if:
$$\|x - y\|_2 < \epsilon(x) \tag{19.1}$$
- Metric Spaces**^[def. 28.65]: a Subset U of a metric space (M, d) is open if:
$$\exists \epsilon > 0 : \text{ if } d(x, y) < \epsilon \quad \forall y \in M, \forall x \in U \implies y \in U \tag{19.2}$$
- Topological Spaces**^[def. 30.2]: Let (X, τ) be a topological space. A set A is said to be open if it is contained in τ .

[add image](#)

Definition 19.9 Closed Set: Is the complement of an open set^[def. 19.8].

Definition 19.10 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.

3. Number Sets

3.1. The Real Numbers \mathbb{R}

3.1.1. Intervals

Definition 19.11 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 :
$$[a, b] = \{x \in \mathbb{R} | a \leq x \leq b\} \tag{19.3}$$

Definition 19.12 Open Interval (a, b) : The open interval of a and b is the set of all real numbers that are within a and b :
$$(a, b) = \{x \in \mathbb{R} | a < x \leq b\} \tag{19.4}$$

3.2. The Rational Numbers \mathbb{Q}

Example 19.1 Power Set/Cardinality of $S = \{x, y, z\}$: The subsets of S are:
 $\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}$ and hence the power set of S is $\mathcal{P}(S) = \{\{\emptyset\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

4. Set Functions

4.1. Submodular Set Functions

Definition 19.13 Submodular Set Functions: A submodular function $f : 2^\Omega \mapsto \mathbb{R}$ is a function that satisfies:
$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \quad \forall A \subseteq B \subset \Omega \quad \{x\} \in \Omega \setminus B \tag{19.5}$$

Explanation 19.1 (Definition 19.13). *Adding 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 19.14 Montone Submodular Function: A *monotone* submodular function is a submodular function^[def. 19.13] that satisfies:
$$f(A) \leq f(B) \quad \forall A \subseteq B \subseteq \Omega \tag{19.6}$$

Explanation 19.2 (Definition 19.14). *Adding more elements to a set will always increase the information/value gain.*

4.2. Complex Numbers

Definition 19.15 Complex Conjugate \bar{z} : The complex conjugate of a complex number $z = x + iy$ is defined as:
$$\bar{z} = x - iy \tag{19.7}$$

[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 :
$$\bar{x} = x \implies x \in \mathbb{R} \tag{19.8}$$

Formula 19.1 Euler's Formula:
$$e^{\pm ix} = \cos x \pm i \sin x \tag{19.9}$$

Formula 19.2 Euler's Identity:
$$e^{\pm i} = -1 \tag{19.10}$$

Note

$$e^n = 1 \Leftrightarrow n = i2\pi k, \quad k \in \mathbb{N} \tag{19.11}$$

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.
$$f : A \mapsto S \quad A \in \mathbb{N} \tag{20.1}$$

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. $(2, 4, 6, 8, 10, 12, \dots)$.
$$t_n = t_0 + nd \quad d : \text{difference between two terms} \tag{20.2}$$

1.2. Geometric Sequence

Definition 20.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \dots)$.
$$t_n = t_0 \cdot r^n \quad r : \text{ratio between two terms} \tag{20.3}$$

Property 20.1 Sum of Geometric Sequence:
$$\sum_{k=1}^n ar^{k-1} = \frac{a(1 - r^n)}{1 - r} \tag{20.4}$$

2. Converging Sequences

2.1. Pointwise Convergence

Definition 20.6 $\lim_{n \rightarrow \infty} f_n = f$ **pointwise** **Pointwise Convergence**[?]: Let (f_n) be a sequence of functions with the same domain^[def. 23.8] and codomain^[def. 23.9]. The sequence is said to convergence pointwise to its *pointwise limit function* f if it satisfies:
$$\lim_{n \rightarrow \infty} |f_n(x) - f(x)| = 0 \quad \forall x \in \text{dom}(f_i) \tag{20.5}$$

2.2. Uniform Convergence

Definition 20.7 $\lim_{n \rightarrow \infty} f_n = f$ **uniform**/ $f_n \xrightarrow{\mathcal{U}} f$ **Uniform Convergence**[?]: Let (g_n) be a sequence of functions with the same domain^[def. 23.8] and codomain^[def. 23.9]. The sequence is said to convergence uniformly to its *pointwise limit function* f if it satisfies:
$$\exists \epsilon > 0 : \exists n \geq 1 \quad \sup_{x \in \text{dom}(f_i)} |g_n(x) - f(x)| < \epsilon \quad \forall x \in \text{dom}(f_i) \tag{20.6}$$

Note

Uniform convergence is characterized by the uniform norm??, and is stronger than pointwise convergence.

Toplogy

Definition 21.1 Topological Space[?] (X, τ) : Is an ordered pair (X, τ) , where X is a set and τ is a topology^[def. 30.1] on X .

Definition 21.2 Topological Space[?] (X, τ) : Is an ordered pair (X, τ) , where X is a set and τ is a topology^[def. 30.1] on X .

1. Weak Topologies

Definition 21.3 Weak Topology $\mathcal{C}(\mathcal{K}; \mathbb{R})$: Is the coresets topology s.t all cont. linear functionals w.r.t. to the strong topology are continuous. Neighbourhood Basis:
$$\{f | |l_1| < \epsilon_1, \dots, |l_n| < \epsilon_n, \forall \epsilon_i, \forall n, \forall \text{lin. functions } f\} \tag{21.1}$$

Note

The weak closure:

- is usually larger as the uniform closure, as for the weak closure there are many more convergence sequences
- is easier to calculate than the uniform closure

2. Compact Space

Corollary 21.1 Euclidean Space: In the euclidean case, a set $X \in \mathbb{R}$ is compact iff:

- it is closed^[def. 19.9]
- bounded

3. Closure

Definition 21.4 Closure of a Set[?] $\text{cl}_{X, \tau}(S) / \bar{S}$: The closure of a subset S of a topological space^[def. 30.2] (X, τ) is defined equivalantly by:

- Is the union of S and its boundary ∂S .
- is the set S together with its limit points.

Note

If the topological space X, τ is clear from context, then the closure of a set S is often written simply as \bar{S} .

Corollary 21.2 Uniform Closure $\|\cdot\|_\infty$: The uniform closure of a set of functions A is the *space of all functions that can be approximated* by a sequence (f_n) of uniformly-converging functions from A . ^[def. 20.7] functions

Corollary 21.3 Weak Closure:

<div>Logic</div> <div>1. Boolean Algebra</div> <div>1.1. Basic Operations</div>	
Definition 22.1 Conjunction/AND	\wedge :
Definition 22.2 Disjunction/OR	\vee :
Definition 22.3 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. <div> $x \wedge y = xy = \min(x, y)$ $x \vee y = x + y = \max(x, y)$ $\neg x = 1 - x$ $x \oplus y = (x + y) \cdot (\neg x + \neg y) = x \cdot \neg y + \neg x \cdot y$ </div>	
Note: non-linearity of XOR	
$(x + y) \cdot (\neg x + \neg y) = -x^2 - y^2 - 2xy + 2x + 2y$	
1.2. Boolean Identities	
Property 22.1 Idempotence:	
$x \wedge x \equiv x$ and $x \vee x \equiv x$	(22.1)
Property 22.2 Identity Laws:	
$x \wedge \text{true} \equiv x$ and $x \vee \text{false} \equiv x$	(22.2)
Property 22.3 Zero Law's:	
$x \wedge \text{false} \equiv \text{false}$ and $x \vee \text{true} \equiv \text{true}$	(22.3)
Property 22.4 Double Negation:	
$\neg \neg x \equiv x$	(22.4)
Property 22.5 Complementation:	
$x \wedge \neg x \equiv \text{false}$ and $x \vee \neg x \equiv \text{true}$	(22.5)
Property 22.6 Commutativity:	
$x \vee y \equiv y \vee x$ and $x \wedge y \equiv y \wedge x$	(22.6)
Property 22.7 Associativity:	
$(x \vee y) \vee z \equiv x \vee (y \vee z)$	(22.7)
$(x \wedge y) \wedge z \equiv x \wedge (y \wedge z)$	(22.8)
Property 22.8 Distributivity:	
$x \vee (y \wedge z) \equiv (x \vee y) \wedge (x \vee z)$	(22.9)
$x \wedge (y \vee z) \equiv (x \wedge y) \vee (x \wedge z)$	(22.10)
Property 22.9 De Morgan's Laws:	
$\neg(x \vee z) \equiv (\neg x \wedge \neg y)$	(22.11)
$\neg(x \wedge z) \equiv (\neg x \vee \neg y)$	(22.12)
Note	
The algebra axioms come in pairs that can be obtained by interchanging \wedge and \vee .	
1.3. Normal Forms	
Definition 22.4 Literal	[example 22.1]: Literals are atomic formulas or their negations
Definition 22.5 Negation Normal Form (NNF): A formula F is in negation normal form is the negation operator \neg is only applied to literals ^[def. 22.4] and the only other operators are \wedge and \vee .	
Definition 22.6 Conjunctive Normal Form (CNF): An boolean algebraic expression F is in CNF if it is a <i>conjunction</i> of <i>clauses</i> , where each clause is a disjunction of <i>literals</i> ^[def. 22.4] $L_{i,j}$:	
$F_{\text{CNF}} = \bigwedge_{i=1}^n \left(\bigvee_{j=1}^{m_i} L_{i,j} \right)$	

Definition 22.7 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. 22.4] $L_{i,j}$:

$$F_{\text{DNF}} = \bigvee_{i=1}^n \left(\bigwedge_{j=1}^{m_i} L_{i,j} \right)$$
(22.14)

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 22.1:

- Using *De Morgan's laws*Property 22.9 and double negationProperty 22.4 transform F into *Negation Normal Form*^[def. 22.5]:

$\neg \neg x$	by	x
$\neg(x \wedge y)$	by	$(\neg x \vee \neg y)$
$\neg(x \vee y)$	by	$(\neg x \wedge \neg y)$
$\neg \text{true}$	by	false
$\neg \text{false}$	by	true

- Using distributive lawsProperty 22.8 substitute all:

$x \wedge (y \vee z)$	by	$(x \wedge y) \vee (x \wedge z)$
$(y \vee z) \wedge x$	by	$(y \wedge x) \vee (z \wedge x)$
$x \wedge \text{true}$	by	true
$\text{true} \wedge x$	by	true

- Using the identityProperty 22.2 and zero laws Property 22.3 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 22.2]

To obtain a DNF formula from a truth table we need to have a *conjunctive*^[def. 22.3] for each row where F is true.

2. Examples

Example 22.1 Literals:

Boolean literals: $x, \neg y, s$

Not boolean literals: $\neg \neg x, (x \wedge y)$

Example 22.2 DNF from truth tables:

	x	y	z	F
	0	0	0	1
Need a conjunction of:	0	0	1	0
• $(\neg x \wedge \neg y \wedge \neg z)$	0	1	0	0
• $(\neg x \wedge y \wedge z)$	0	1	1	1
• $(x \wedge \neg y \wedge \neg z)$	1	0	0	1
• $(x \wedge y \wedge z)$	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

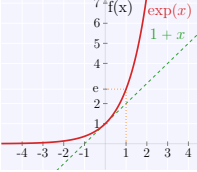
1. Functional Analysis

1.1. Elementary Functions

1.1.1. Exponential Numbers

Definition 23.1 Exponential Function

$$\exp(x) = e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n \quad (23.1)$$

$\exp : \mathbb{K} \mapsto \mathbb{K}:$


Definition 23.2 Exponential/Euler Number

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 2.7182 \quad (23.2)$$

Properties Defining the Exponential Function

Property 23.1:
$$\exp(x + y) = \exp(x) \cdot \exp(y) \quad (23.3)$$

Property 23.2:
$$\exp(x) \leq 1 + x \quad (23.4)$$

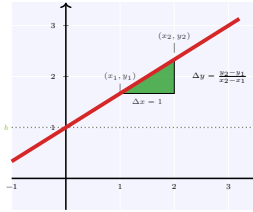
1.1.2. Affine Linear Functions

Definition 23.3 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 $t_b(x) = x + b$:

$$M = \{f : \mathbb{R} \mapsto \mathbb{R} \mid f(x) = (s_a \circ t_b)(x) = ax + b, \quad a, b \in \mathbb{R}\} \quad (23.5)$$



Formula 23.1

Linear Function from Point and slope

Given a point (x_1, y_1) and a slope a we can derive:

$$f(x) = a \cdot (x - x_0) + y_0 = ax + (y_1 - ax_0) \quad (23.6)$$

Formula 23.2 Linear Function from two Points:

$$f(x) = a \cdot (x - x_p) + y_p = ax + (y_p - ax_p) \quad (23.7)$$
$$a = \frac{y_1 - y_0}{x_1 - x_0} \quad p = \{0 \text{ or } 1\}$$

1.1.3. Polynomials

Definition 23.4 Polynomial:

A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:

$$\mathcal{P}_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1} + a_nx^n \quad (23.8)$$

Corollary 23.1 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 23.5 Monomial:

Is a polynomial with only one term.

Cubic Polynomials

Definition 23.6 Cubic Polynomials:

Are polynomials of degree^[cor. 23.1] 3 and have four coefficients:

$$f(x) = a_3x^3 + a_2x^2 + a_1x + a_0 \quad (23.9)$$

1.2. Functional Compositions

Definition 23.7 Functional Compositions

Let $f : A \mapsto B$ and $g : D \mapsto C$ be two mappings s.t. $\text{codom}(f) \subseteq D$ then we can define a composition function $(f \circ g)A \mapsto C$ as:

$$h(x) = (g \circ f)(x) = g(f(x)) \quad \text{with } x \in A \quad (23.10)$$

Corollary 23.2 Nested Functional Composition:

$$F_{k;1}(x) = (F_k \circ \dots \circ F_1)(x) = F_k(F_{k-1} \circ \dots \circ (F_1(x))) \quad (23.11)$$

2. Proofs

Proof 23.1 formula 23.1:

$$f(x_0) = y_0 = ax_0 + b \quad \Rightarrow \quad b = y_0 - ax_0$$

Theorem 23.1

First Fundamental Theorem of Calculus:

Let f be a continuous real-valued function defined on a closed interval $[a, b]$.

Let F be the function defined $\forall x \in [a, b]$ by:

$$F(x) = \int_a^x f(t) dt \quad (23.12)$$

Then it follows:

$$F'(x) = f(x) \quad \forall x \in (a, b) \quad (23.13)$$

Theorem 23.2

Second Fundamental Theorem of Calculus:

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

$$\int_a^b f(t) dt = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t) dt = F(x) \quad (23.14)$$

Definition 23.8 Domain of a function

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

Definition 23.9

Codomain/target set of a function

Given a function $f : X \rightarrow Y$, the codomain of that function is the set Y into which all of the output of the function is constrained to fall.

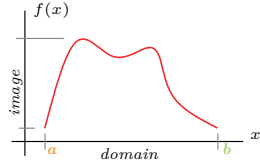
Definition 23.10 Image (Range) of a function: $f[\cdot]$

Given a function $f : X \rightarrow Y$, the image of that function is the set to which the function can actually map:

$$\{y \in Y \mid y = f(x), \quad \forall x \in X\} := f[X] \quad (23.15)$$

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

The image is thus a subset of a function's codomain.



Misnomer Range:

The term Range is ambiguous s.t. certain books refer to it as codomain and other as image.

Definition 23.11 Inverse Image/Preimage $f^{-1}(\cdot)$:

Let $f : X \mapsto Y$ be a function, and A a subset set of its codomain Y .

Then the preimage of A under f is the set of all elements of the domain X , that map to elements in A under f :

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \quad (23.16)$$

Example 23.1 :

Given

$$f : \mathbb{R} \rightarrow \mathbb{R}$$

defined by

$$f : x \mapsto x^2 \iff f(x) = x^2$$

$\text{dom}(f) = \mathbb{R}, \text{codom}(f) = \mathbb{R}$ but its image is $f[\mathbb{R}] = \mathbb{R}_+$.

Image (Range) of a subset

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

$$f[A] = \{y \in Y \mid y = f(x), \quad \forall x \in A\} \quad (23.17)$$

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 23.12 (strictly) Increasing Functions:

A function f is called monotonically increasing/increasing/non-decreasing if:

$$x \leq y \iff f(x) \leq f(y) \quad \forall x, y \in \text{dom}(f) \quad (23.18)$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \quad \forall x, y \in \text{dom}(f) \quad (23.19)$$

Definition 23.13 (strictly) Decreasing Functions:

A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geq y \iff f(x) \geq f(y) \quad \forall x, y \in \text{dom}(f) \quad (23.20)$$

And **strictly** decreasing if:

$$x > y \iff f(x) > f(y) \quad \forall x, y \in \text{dom}(f) \quad (23.21)$$

Definition 23.14 Monotonic Function:

A function f is called monotonic iff either f is **increasing** or **decreasing**.

Definition 23.15 Linear Function:

A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

$$L(x + y) = L(x) + L(y)$$
$$L(\alpha x) = \alpha L(x) \quad \forall x, y \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

Corollary 23.3 Linearity of Differentiation:

The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{d}{dx} (af(x) + bg(x)) = a \frac{d}{dx} f(x) + b \frac{d}{dx} g(x) \quad a, b \in \mathbb{R} \quad (23.22)$$

Definition 23.16 Quadratic Function:

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

$$f(x) = \frac{1}{2} x^T A x + b^T x + c \quad (23.23)$$

3. Norms

3.1. Infinity/Supremum Norm

Definition 23.17 Infinity/Supremum Norm:

$$\|f\|_{\infty} := \sup_{x \in \text{dom}(f)} |f(x)| \quad (23.24)$$

Note

In order to make this a proper norm one usually considers *bounded functions* s.t.:

$$\|f\|_{\infty} \leq M < \infty$$

Corollary 23.4 Infinity Norm induced Metric:

The infinity norm naturally induces a metric^[def. 28.64]:

$$d := (f, g) := \|f - g\|_{\infty} \quad (23.25)$$

4. Smoothness

Definition 23.18 Smoothness of a Function C^k :

Given a function $f : X \rightarrow Y$, the function is said to be of class k if it is differentiable up to order k **and** continuous, on its entire domain:

$$f \in C^k(X) \iff \exists f', f'', \dots, f^{(k)} \text{ continuous} \quad (23.26)$$

Note

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

4.0.1. Continuous Functions

Definition 23.19 Continuous Function C^0 :

Functions that do not have any jumps or peaks.

4.0.2. Piece wise Continuous Functions

Definition 23.20 Piecewise Linear Functions C^0_{pw} :

4.0.3. Continuously Differentiable Function

Corollary 23.5 Continuously Differentiable Function C^1 :

Is the class of functions that consists of all differentiable functions whose derivative is continuous.

Hence a function $f : X \rightarrow Y$ of the class must satisfy:

$$f \in C^1(X) \iff f' \text{ continuous} \quad (23.27)$$

4.0.4. Smooth Functions

Corollary 23.6 Smooth Function C^∞ : Is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that has derivatives infinitely many times differerntiable.

$$f \in C^\infty(\mathcal{X}) \iff f', f'', \dots, f^{(\infty)} \quad (23.28)$$

4.1. Lipschitz Continuous Functions

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

Definition 23.21 Lipschitz Continuity:

A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Constant L :

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y}, \quad L > 0 \quad (23.29)$$

Note

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

4.1.1. Lipschitz Continuous Gradient

Definition 23.22 Lipschitz Continuous Gradient:

A continuously differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has L -Lipschitz continuous gradient if it satisfies:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f), \quad L > 0 \quad (23.30)$$

if $f \in C^2$, this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leq L \mathbf{I} \quad \forall \mathbf{x} \in \text{dom}(f), \quad L > 0 \quad (23.31)$$

Lemma 23.1 Descent Lemma [Poorfs 23.5,??]:

If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has Lipschitz continuous gradient eq. (23.30) over its domain, then it holds that:

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})| \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (23.32)$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian (Definition 24.8) of f is uniformly upper bounded by L

4.2. L-Smooth Functions

Definition 23.23 L-Smoothness:

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

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

with $\forall \mathbf{x}, \mathbf{y} \in \text{dom}(f), \quad L > 0 \quad (23.33)$

If f is a twice differentiable this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leq L \mathbf{I} \quad L > 0 \quad (23.34)$$

Theorem 23.3 [proof 23.6]

L-Smoothness of convex functions:

A convex and L -Smooth function (def. 23.23) has a Lipschitz continuous gradienteq. (23.30) thus it holds that:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (23.35)$$

Note

L -smoothnes is a weaker condition than L -Lipschitz continuous gradients

5. Convexity and Concavity

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

Definition 23.24 Convex Functions:

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if it satisfies:

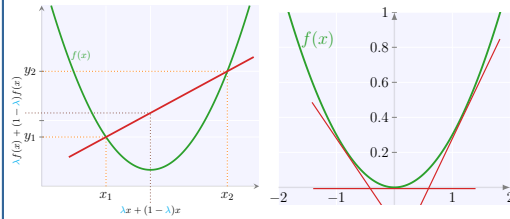
$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \lambda \in [0, 1] \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (23.36)$$

If f is a differentiable function this is equivalent to:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (23.37)$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (23.38)$$



Definition 23.25 Concave Functions:

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is concave if it satisfies:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \geq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad \forall \lambda \in [0, 1] \quad (23.39)$$

Corollary 23.7 Convexity \rightarrow global minimima: Convexity implies that all local minima (if they exist) are global minima.

5.1. Properties

Property 23.3 Monotonicity of the Derivative:

$$\begin{array}{ll} \text{convex} & f'(a) < f'(b) \\ \text{If } f : \mathbb{R} \mapsto \mathbb{R} \text{ is} & \\ \text{concave} & f'(a) > f'(b) \end{array} \quad a < b, \quad a, b \in \mathbb{R} \quad (23.40)$$

5.1.1. Properties that preserve convexity

Property 23.4 Non-negative weighted Sums: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

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

Property 23.5 Composition of Affine Mappings: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

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

Property 23.6 Pointwise Maxima: Let f be a convex function then $g(\mathbf{x})$ is convex as well:

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

5.2. Strict Convexity/Concavity

Definition 23.26 Stricly Convex Functions:

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad \forall \lambda \in [0, 1]$$

If f is a differentiable function this is equivalent to:

$$f(\mathbf{x}) > f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (23.41)$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(\mathbf{x}) > 0 \quad \forall \mathbf{x}, \mathbf{y} \in \text{dom}(f) \quad (23.42)$$

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. (23.41) implies that $f(\mathbf{x})$ is above the tangent $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$
- ?? implies that $f(\mathbf{x})$ is flat or curved upwards

Corollary 23.8 Strict Convexity \rightarrow Uniqueness:

Strict convexity implies a unique minimizer \iff at most one global minimum.

Corollary 23.9 : A twice differentiable function of one variable $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval \mathcal{X} :

$$f''(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \mathcal{X} \quad (23.43)$$

5.3. Strong Convexity/Concavity

Definition 23.27 μ -Strong Convexity:

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

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

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad (23.44)$$

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

$$\nabla^2 f(\mathbf{x}) \geq \mu \mathbf{I} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad \mu > 0 \quad (23.45)$$

Corollary 23.10

Strong Convexity implies Strict Convexity:

<https://math.stackexchange.com/question/2090991/proof-for-strongly-convex-function-is-strictly-convex>

Property 23.7:

$$f(\mathbf{y}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{1}{2\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \quad (23.46)$$

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 μ

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

Proof 23.2: eq. (23.45) analogously to **Proof** eq. (23.34)

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian (def. 24.8) of f is uniformly lower bounded by μ

Hence strong convexity can be considered as the analogous to smoothness

Example 23.2 Quadratic Function: A quadratic function eq. (23.23) is convex if:

$$\nabla_{\mathbf{x}}^2 \text{eq. (23.23)} = \mathbf{A} \geq 0 \quad (23.47)$$

Corollary 23.11 :

Strong convexity \Rightarrow Strict convexity \Rightarrow Convexity

Functions

Even Functions: have rotational symmetry with respect to the origin.

\Rightarrow **Geometrically:** its graph remains unchanged after reflection about the y -axis.

$$f(-x) = f(x) \quad (23.48)$$

Odd Functions: are symmetric w.r.t. to the y -axis.

\Rightarrow **Geometrically:** its graph remains unchanged after rotation of 180 degrees about the origin.

$$f(-x) = -f(x) \quad (23.49)$$

Theorem 23.4 Rules:

Let f be even and f odd respectively.

$$\begin{array}{ll} g =: f \cdot f \text{ is even} & g =: f \cdot f \text{ is even} \\ g =: f \cdot f \text{ is odd} & \text{the same holds for division} \end{array}$$

Examples

Even: $\cos x, |x|, c, x^2, x^4, \dots \exp(-x^2/2)$.

Odd: $\sin x, \tan x, x, x^3, x^5, \dots$

$$\begin{array}{ll} \mathbf{x}\text{-Shift:} & f(\mathbf{x} - \mathbf{c}) \Rightarrow \text{shift to the right} \\ & f(\mathbf{x} + \mathbf{c}) \Rightarrow \text{shift to the left} \end{array} \quad (23.50)$$

$$\mathbf{y}\text{-Shift:} \quad f(\mathbf{x}) \pm \mathbf{c} \Rightarrow \text{shift up/down} \quad (23.51)$$

Proof 23.3: eq. (23.50) $f(\mathbf{x}_n - \mathbf{c})$ we take the x -value at \mathbf{x}_n but take the y -value at $\mathbf{x}_0 := \mathbf{x}_n - \mathbf{c}$
 \Rightarrow we shift the function to \mathbf{x}_n .

Euler's formula

$$e^{\pm i x} = \cos x \pm i \sin x \quad (23.52)$$

Euler's Identity

$$e^{\pm i} = -1 \quad (23.53)$$

Note

$$e^n = 1 \Leftrightarrow n = i 2\pi k, \quad k \in \mathbb{N} \quad (23.54)$$

Corollary 23.12 Every norm is a convex function: By using definition (def. 23.24) and the triangular inequality it follows (with the exception of the L_0 -norm):

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

5.4. Taylor Expansion

Definition 23.28 Taylor Expansion:

$$T_n(\mathbf{x}) = \sum_{i=0}^n \frac{1}{i!} f^{(i)}(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)^{(i)} \quad (23.55)$$

$$= f(\mathbf{x}_0) + f'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} f''(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^2 + \mathcal{O}(x^3) \quad (23.56)$$

Definition 23.29 Incremental Taylor:

Goal: evaluate $T_n(\mathbf{x})$ (eq. (23.56)) at the point $\mathbf{x}_0 + \Delta \mathbf{x}$ in order to propagate the function $f(\mathbf{x})$ by $h = \Delta \mathbf{x}$:

$$T_n(\mathbf{x}_0 \pm h) = \sum_{i=0}^n \frac{h^i}{i!} f^{(i)}(\mathbf{x}_0) i^{-1} \quad (23.57)$$

$$= f(\mathbf{x}_0) \pm h f'(\mathbf{x}_0) + \frac{h^2}{2} f''(\mathbf{x}_0) \pm f'''(\mathbf{x}_0)(h)^3 + \mathcal{O}(h^4)$$

Note

If we chose $\Delta \mathbf{x}$ small enough it is sufficient to look only at the first two terms.

Definition 23.30 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\mathbf{x} \in X$, $f : X \mapsto \mathbb{R}$ and $f \in C^2$ then it holds that

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top H(\mathbf{x} - \mathbf{x}_0) \quad (23.58)$$

Definition 23.31 Argmax: The argmax of a function defined on a set D is given by:

$$\arg \max_{x \in D} f(x) = \{x | f(x) \geq f(y), \forall y \in D\} \quad (23.59)$$

Definition 23.32 Argmin: The argmin of a function defined on a set D is given by: $\arg \min_{x \in D} f(x) = \{x f(x) \leq f(y), \forall y \in D\} \quad (23.60)$
Corollary 23.13 Relationship $\arg \min \leftrightarrow \arg \max$: $\arg \min_{x \in D} f(x) = \arg \max_{x \in D} -f(x) \quad (23.61)$
Property 23.8 Argmax Identities: <ol style="list-style-type: none"> Shifting: $\forall \lambda \text{ const} \quad \arg \max f(x) = \arg \max f(x) + \lambda \quad (23.62)$ Positive Scaling: $\forall \lambda > 0 \text{ const} \quad \arg \max f(x) = \arg \max \lambda f(x) \quad (23.63)$ Negative Scaling: $\forall \lambda < 0 \text{ const} \quad \arg \max f(x) = \arg \min \lambda f(x) \quad (23.64)$ Positive Functions: $\forall \arg \max f(x) > 0, \forall x \in \text{dom}(f)$ $\arg \max f(x) = \arg \min \frac{1}{f(x)} \quad (23.65)$ Strictly Monotonic Functions: for all strictly monotonic increasing functions^[def. 23.12] g it holds that: $\arg \max g(f(x)) = \arg \max f(x) \quad (23.66)$
Definition 23.33 Max: The maximum of a function f defined on the set D is given by: $\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \max f(x) \quad (23.67)$
Definition 23.34 Min: The minimum of a function f defined on the set D is given by: $\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg \min f(x) \quad (23.68)$
Corollary 23.14 Relationship $\min \leftrightarrow \max$: $\min_{x \in D} f(x) = -\max_{x \in D} -f(x) \quad (23.69)$
Property 23.9 Max Identities: <ol style="list-style-type: none"> Shifting: $\forall \lambda \text{ const} \quad \max \{f(x) + \lambda\} = \lambda + \max f(x) \quad (23.70)$ Positive Scaling: $\forall \lambda > 0 \text{ const} \quad \max \lambda f(x) = \lambda \max f(x) \quad (23.71)$ Negative Scaling: $\forall \lambda < 0 \text{ const} \quad \max \lambda f(x) = \lambda \min f(x) \quad (23.72)$ Positive Functions: $\forall \arg \max f(x) > 0, \forall x \in \text{dom}(f) \quad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \quad (23.73)$ Strictly Monotonic Functions: for all strictly monotonic increasing functions^[def. 23.12] g it holds that: $\max g(f(x)) = g(\max f(x)) \quad (23.74)$
Definition 23.35 Supremum: The supremum of a function defined on a set D is given by: $\sup_{x \in D} f(x) = \{y y \geq f(x), \forall x \in D\} = \min_{y y \geq f(x), \forall x \in D} y \quad (23.75)$ <p>and is the smallest value y that is equal or greater $f(x)$ for any $x \iff$ smallest upper bound.</p>
Definition 23.36 Infimum: The infimum of a function defined on a set D is given by: $\inf_{x \in D} f(x) = \{y y \leq f(x), \forall x \in D\} = \max_{y y \leq f(x), \forall x \in D} y \quad (23.76)$ <p>and is the biggest value y that is equal or smaller $f(x)$ for any $x \iff$ largest lower bound.</p>
Corollary 23.15 Relationship $\sup \leftrightarrow \inf$: $\sup_{x \in D} f(x) = -\inf_{x \in D} -f(x) \quad (23.77)$

Note The supremum/infimum is necessary to handle unbound function that seem to converge and for which the max/min does not exist as the argmax/argmin may be empty. E.g. consider $-e^x/e^x$ for which the max/min converges toward 0 but will never reached s.t. we can always choose a bigger $x \Rightarrow$ there exists no argmax/argmin \Rightarrow need to bound the functions from above/below \iff infimum/supremum.
Definition 23.37 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. $y(t) = f(x(t), t) \xrightarrow[\nabla \tau]{\text{time-invariance}} y(t - \tau) = f(x(t - \tau), t) \quad (23.78)$
Definition 23.38 Inverse Function $g = f^{-1}$: A function g is the inverse function of the function $f : A \subset \mathbb{R} \rightarrow B \subset \mathbb{R}$ if $f(g(x)) = x \quad \forall x \in \text{dom}(g) \quad (23.79)$ and $g(f(u)) = u \quad \forall u \in \text{dom}(f) \quad (23.80)$
Property 23.10 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 23.5 The Existence of an Inverse Function: A function has an inverse function if and only if it is one-to-one.
Corollary 23.16 Inverse functions and strict monotonicity: If a function f is strictly monotonic ^[def. 23.14] on its entire domain, then it is one-to-one and therefore has an inverse function.
6. Special Functions
6.1. The Gamma Function
Definition 23.39 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (??) to the real and complex numbers (with a positive real part): $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad \Re(z) > 0 \quad (23.81)$ $\Gamma(n) \xleftrightarrow[n \in \mathbb{N}]{\iff} \Gamma(n) = (n-1)!$
7. Proofs
Proof 23.4: lemma 23.1 for C^1 functions: Let $g(t) \equiv f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$ from the FToC (theorem 23.2) we know that: $\int_0^1 g'(t) dt = g(1) - g(0) = f(\mathbf{x}) - f(\mathbf{y})$
It then follows from the reverse: $ f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) $ $\stackrel{\text{Chain. R.}}{\stackrel{\text{FToC}}{=}} \left \int_0^1 \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \right $ $= \left \int_0^1 (\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt \right $ $= \left \int_0^1 (\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) dt \right $ $\stackrel{\text{C.S.}}{\leq} \left \int_0^1 \ \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y})\ \cdot \ \mathbf{x} - \mathbf{y}\ dt \right $ $\stackrel{\text{eq. (23.30)}}{=} \left \int_0^1 L \ \mathbf{y} + t(\mathbf{x} - \mathbf{y}) - \mathbf{y}\ \cdot \ \mathbf{x} - \mathbf{y}\ dt \right $ $= \left L \ \mathbf{x} - \mathbf{y}\ ^2 \int_0^1 t dt \right = \frac{L}{2} \ \mathbf{x} - \mathbf{y}\ _2^2$

Proof 23.5: ?? for C^2 functions: $f(\mathbf{y}) \stackrel{\text{Taylor}}{=} f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(z) (\mathbf{y} - \mathbf{x})$ <p>Now we plug in $\nabla^2 f(\mathbf{x})$ and recover eq. (23.33):</p> $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top L (\mathbf{y} - \mathbf{x})$
Proof 23.6: theorem 23.3: With the definition of convexity for a differentiable function (eq. (23.41)) it follows $f(x) - f(y) + \nabla f(y)^\top (x - y) \geq 0$ $\Rightarrow f(x) - f(y) + \nabla f(y)^\top (x - y) $ <p style="text-align: center;">if eq. $\stackrel{(23.41)}{=} f(x) - f(y) + \nabla f(y)^\top (x - y)$</p> <p>with lemma 23.1 and ^[def. 23.23] it follows theorem 23.3</p>

Differential Calculus

1. Mean Value Theorem

Theorem 24.1 Mean Value Theorem: Let $f : [a, b] \rightarrow \mathbb{R}$ be continuous function, differentiable on the open interval (a, b) , with $a < b$. Then there exist some $c \in (a, b)$ s.t.

$$f'(c) = \frac{f(b) - f(a)}{b - a} = \frac{1}{b - a} \int_a^b f(x) dx \quad (24.1)$$

2. The Product Rule

Rule 24.1 (Product /Leibniz Rule).

Let u, v be two differentiable functions $u, v \in \mathcal{C}^1$ then it holds that:

$$\frac{d(u(x)v(x))}{dx} = (uv)' = u'v + v'u \quad (24.2)$$

3. The Chain Rule

Formula 24.1 Generalized Chain Rule:

Let $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^k$ and $\mathbf{G} : \mathbb{R}^k \rightarrow \mathbb{R}^m$ be to general maps then it holds:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} (\mathbf{G} \circ \mathbf{F}) &= \left(\frac{\partial \mathbf{G}}{\partial \mathbf{F}} \circ \mathbf{F} \right) \cdot \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \\ \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n} \quad \mathbb{R}^n \rightarrow (\mathbb{R}^{m \times k} \cdot \mathbb{R}^{k \times n}) \end{aligned} \quad \begin{aligned} \partial \mathbf{F} : \mathbb{R}^n &\rightarrow \mathbb{R}^{k \times n} \\ \partial \mathbf{G} : \mathbb{R}^k &\rightarrow \mathbb{R}^{m \times k} \end{aligned} \quad (24.3)$$

4. Directional Derivative

5. Partial Differentiation

Definition 24.1 Partial Derivative:

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real valued function, its partial derivative $\partial_i f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as the directional derivative?? along the coordinate axis of one of its variables:

$$\begin{aligned} \partial_i f(\mathbf{x}) &= \frac{\partial f}{\partial x_i} = D_{x_i} f = \lim_{h \rightarrow 0} \frac{f(\mathbf{x}, x_i \leftarrow x_i + h) - f(\mathbf{x})}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h} \end{aligned} \quad (24.4)$$

5.1. The Gradient

5.1.1. The Nabla Operator

Definition 24.2 Nabla Operator/Del ∇ : 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:

$$\nabla = \sum_{i=1}^n \frac{\partial}{\partial x_i} \hat{\mathbf{e}}_i = \begin{bmatrix} \frac{\partial}{\partial x_1}(\mathbf{x}) \\ \frac{\partial}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n}(\mathbf{x}) \end{bmatrix} \quad (24.5)$$

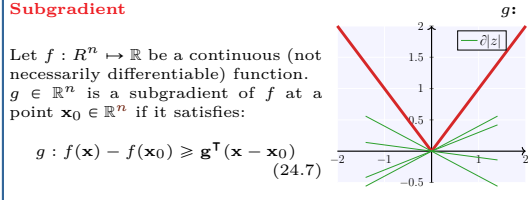
Definition 24.3 Gradient:

Given a *scalar valued* function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ its gradient $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined as vector \mathbb{R}^n of the partial derivatives^[def. 24.1] w.r.t. all coordinate axes:

$$\text{grad } f(\mathbf{x}) := \nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix} = \left(\frac{\partial f}{\partial \mathbf{x}} \right)^T \quad (24.6)$$

5.1.2. The Subderivative

Definition 24.4 Subgradient



Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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:

$$g : f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad (24.7)$$

Definition 24.5 Subderivative

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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. 24.4] g : $\partial f(\mathbf{x}_0) \{ g : f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathbb{R}^n \}$ (24.8)

Heuristic

We can guess the sub derivative at a point by looking at all the slopes that are smaller then the graph.

5.2. The Jacobian

Definition 24.6 Jacobian/Jacobi Matrix

Given a *vector valued* function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ its derivative $\mathbf{J}_{\mathbf{f}} : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$ with components $\partial_{ij} \mathbf{f} = \partial_i f_j : \mathbb{R}^n \rightarrow \mathbb{R}$ is a vector valued function defined as:

$$\begin{aligned} \mathbf{J}(\mathbf{f}(\mathbf{x})) &= \mathbf{J}_{\mathbf{f}}(\mathbf{x}) = \mathbf{Df} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}) = \frac{\partial (f_1, \dots, f_m)}{\partial (x_1, \dots, x_n)}(\mathbf{x}) \quad (24.9) \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} \\ \dots &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix} \end{aligned}$$

Explanation 24.1. Rows of the Jacobian are transposed gradients^[def. 24.3] of the component functions f_1, \dots, f_m .

Corollary 24.1 :

6. Second Order Derivatives

Definition 24.7 Second Order Derivative $\frac{\partial^2}{\partial x_i \partial x_j}$:

Theorem 24.2

Symmetry of second derivatives/Schwartz's Theorem: Given a continuous and twice differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ then its second order partial derivatives commute:

$$\frac{\partial}{\partial x_i} \frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\partial f}{\partial x_i}$$

6.1. The Hessian

Definition 24.8 Hessian Matrix:

Given a function $f : \mathbb{R} \rightarrow \mathbb{R}^n$ its Hessian $\in \mathbb{R}^{n \times n}$ is defined as:

$$\mathbf{H}(\mathbf{f})(\mathbf{x}) = \mathbf{H}_{\mathbf{f}}(\mathbf{x}) = \mathbf{J}(\nabla f(\mathbf{x}))^T \quad (24.10)$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix}$$

and it corresponds to the Jacobian of the Gradient. Due to the differentiability and theorem 24.2 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 24.2 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:

$$\mathbf{d}^T \mathbf{H} \mathbf{d} \iff \mathbf{d}^T \sum_{i=1}^n \lambda_i \mathbf{v}_i \iff \text{if } \mathbf{d} = \mathbf{v}_j \quad \mathbf{d}^T \lambda_j \mathbf{v}_j$$

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

7. Extrema

Definition 24.9 Critical/Stationary Point: Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, that is differentiable at a point \mathbf{x}_0 then it is called a **critical point** if the functions derivative vanishes at that point:

$$f'(\mathbf{x}_0) = 0 \iff \nabla_{\mathbf{x}} f(\mathbf{x}_0) = 0$$

Corollary 24.3 Second Derivative Test $f : \mathbb{R} \rightarrow \mathbb{R}$:

Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is twice differentiable at a stationary point x ^[def. 24.9] then it follows that:

- $f''(x) > 0 \iff \begin{matrix} f'(x + \epsilon) > 0 & \text{slope points uphill} \\ f'(x - \epsilon) < 0 & \text{slope points downhill} \end{matrix}$
 $f(x)$ is a local minimum
- $f''(x) < 0 \iff \begin{matrix} f'(x + \epsilon) > 0 & \text{slope points downhill} \\ f'(x - \epsilon) < 0 & \text{slope points uphill} \end{matrix}$
 $f(x)$ is a local maximum

$\epsilon > 0$ sufficiently small enough

Corollary 24.4 Second Derivative Test $f : \mathbb{R}^n \rightarrow \mathbb{R}$:

Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at a stationary point \mathbf{x} ^[def. 24.9] then it follows that:

- If \mathbf{H} is **p.d** $\iff \forall \lambda_i > 0 \in \mathbf{H} \rightarrow f(\mathbf{x})$ is a local min.
- If \mathbf{H} is **n.d** $\iff \forall \lambda_i < 0 \in \mathbf{H} \rightarrow f(\mathbf{x})$ 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.

8. Proofs

Proof 24.1: Definition 24.4 $f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^T (\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathbb{R}^n$ corresponds to a line (see formula 23.1) at the point \mathbf{x}_0 with slope \mathbf{g}^T . Thus we search for all lines with smaller slope then function graph.

9. Examples

Example 24.1 Subderivatives Absolute Value Function

$|x| : f : \mathbb{R} \rightarrow \mathbb{R}$ with $f(x) = |x|$ at the point $x = 0$ it holds:

$$f(x) - f(0) \geq gx \iff \text{the interval } [-1; 1]$$

For $x \neq 0$ the subgradient is equal to the gradient. Thus it follows for the subderivatives/differentials:

$$\partial |x| = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Integral Calculus

Theorem 25.1 Important Integral Properties:

Addition $\int_a^b f(x) \, dx = \int_a^c f(x) \, dx + \int_c^b f(x) \, dx$ (25.1)

Reflection $\int_a^b f(x) \, dx = - \int_b^a f(x) \, dx$ (25.2)

Translation $\int_a^b f(x) \, dx \stackrel{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x \mp c) \, dx$ (25.3)

f Odd $\int_{-a}^a f(x) \, dx = 0$ (25.4)

f Even $\int_{-a}^a f(x) \, dx = 2 \int_0^a f(x) \, dx$ (25.5)

Proof 25.1: eqs. (25.4) and (25.5)

$$\begin{aligned} I &:= \int_{-a}^a f(x) \, dx = \int_{-a}^0 f(x) \, dx + \int_0^a f(x) \, dx \\ &\stackrel{t=-x}{dt=-dx} \underset{a}{=} \int_a^0 f(-x) \, dx + \int_0^a f(x) \, dx \\ &= \int_0^a f(-x) + f(x) \, dx = \begin{cases} 0 & \text{if } f \text{ odd} \\ 2I & \text{if } f \text{ even} \end{cases} \end{aligned}$$

Definition 25.1 Integration by Parts:

$$\int_a^b u \, dv = uv \Big|_a^b - \int_a^b v \, du \tag{25.6}$$

1. Integral Theorems

1.1. Greens Identities

Theorem 25.2 Greens First Identity:

Let $\bar{\Omega} = \Omega \cup \partial\Omega$, for all vector fields $\mathbf{j} \in (C^1_{\text{pw}}(\bar{\Omega}))^d$ and scalar functions $v \in C^1_{\text{pw}}(\bar{\Omega})$ it holds:

$$\int_{\Omega} \mathbf{j}^\top \text{grad } v \, d\mathbf{x} = - \int_{\Omega} \text{div } \mathbf{j} v \, d\mathbf{x} + \int_{\partial\Omega} \mathbf{j}^\top \mathbf{n} v \, dS \tag{25.7}$$

add multidimensional product rule and gauss theorem from NPDE

Differential Equations

Definition 25.2 [??]
Differential Operator:
A differential operator \mathcal{L} is a mapping of a suitable function space onto another function space, involving only values of the function argument and its derivatives in the same point:
 $\mathcal{L} : C^n(\Omega) \mapsto C^k(\Omega), k < n$
Note: \mathcal{L} is a differential operator of order $k - n$.

Definition 25.3 Linear Differential Operator:
Is a differential operator \mathcal{L} that satisfies:
 $\mathcal{L}(\alpha u + \beta v) = \alpha \mathcal{L}(u) + \beta \mathcal{L}(v) \quad \forall \alpha, \beta \in \mathbb{R} \quad (25.8)$

Ordinary Differential Quations
Partial Differential Equations (PDE)s

Definition 27.1 Partial Differential Equation:
Let $\mathbf{u} = \mathbf{u}(x_1, \dots, x_n) : \mathbb{R}^k \mapsto \mathbb{R}^l$ be an unknown function depending on $\mathbf{x} = (x_1 \dots x_k)$ and let f be a known function.
The known function \mathcal{F} , depending on differentials of the non-known function \mathbf{u} is called a Partial Differential equation:
 $\mathcal{F}\left(\mathbf{u}, \frac{\partial \mathbf{u}}{\partial x_1}, \dots, \frac{\partial \mathbf{u}^n}{\partial x_i^j}, \dots, \frac{\partial \mathbf{u}^n}{\partial x_j^j}, f\right) = \mathcal{F}(\mathbf{u}, D\mathbf{u}, \dots, D^n \mathbf{u}, f) = 0$
or $\mathcal{L}(\mathbf{u}) = f$ in $\Omega \quad (27.1)$

Corollary 27.1 Dependent Variables:
 $\mathbf{u} : \mathbb{R}^k \mapsto \mathbb{R}^l \quad (27.2)$

Corollary 27.2 Independent Variables:
 $\mathbf{x} = (x_1 \dots x_k) \quad (27.3)$

Definition 27.2 Order n :
Is the highest partial derivative that appears in a PDE.

1. Algebraic Types

1.1. Linearity

Definition 27.3 [??]
Linear PDEs:
A linear PDE naturally defines a linear operator [def. 25.3]. A linear PDE must be linear regarding the unknown function \mathbf{u} . In other words all dependent variables \mathbf{u} and their corresponding derivatives depend only on the independent variables x_1, x_2, \dots, x_m :
 $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y + c(x, y)\mathbf{u} = d(x, y) \quad (27.4)$

Definition 27.4 [??]
Semilinear PDEs:
Are PDEs whose coefficients of the highest order n -terms are functions depending only on the independent variables but not onto the dependent variables \mathbf{u} or their derivatives.
Thus the PDE is **linear** regarding to the highest order terms:
 $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (27.5)$

Definition 27.5 [??]
Quasilinear PDEs:
Are PDEs whose coefficients of the highest order (n) terms are functions only depending on the independent variables and on the dependent variables \mathbf{u} and their derivatives up to an order $m < n$, that is smaller than the highest order terms n :
 $a(x, y, \mathbf{u})\mathbf{u}_x + b(x, y, \mathbf{u})\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (27.6)$

Definition 27.6 [??]
Fully Non-linear PDEs:
Are PDEs where all terms of the highest order n are non-linear:
 $a(x, y, \mathbf{u}, \mathbf{u}')\mathbf{u}_x + b(x, y, \mathbf{u}, \mathbf{u}')\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (27.7)$
Note: $\neg(\text{Quasilinear} \Leftrightarrow \text{Fully Nonlinear})$

1.2. Homogeneity

Definition 27.7 Homogeneous $\mathcal{L}(\mathbf{u}) = 0$:
All terms depend on \mathbf{u} or on derivatives of \mathbf{u} .

Definition 27.8 Non-Homogeneous $\mathcal{L}(\mathbf{u}) = f$:
Their exists non-zero terms f that do not depend on \mathbf{u} or on derivatives of \mathbf{u} .

1.3. Constant Coefficients

Definition 27.9 PDEs with Constant Coefficients:
Is a PDE whose coefficients a, b, c, \dots are constants i.e. independent variables.

1.4. 2nd-Order Linear PDEs in two variables

Definition 27.10
2nd-Order Linear PDEs in two Variables:
 $\mathcal{L}(\mathbf{u}) = a\mathbf{u}_{xx} + 2b\mathbf{u}_{xy} + c\mathbf{u}_{yy} + d\mathbf{u}_x + e\mathbf{u}_y + f\mathbf{u} = g \quad (27.8)$
where a, b, \dots, g are functions depending on x and y .

Definition 27.11 Principal Part: Is the operator \mathcal{L}_0 , that consists of the second-(=highest) order parts of \mathcal{L} :
 $\mathcal{L}_2(\mathbf{u}) := a\mathbf{u}_{xx} + 2b\mathbf{u}_{xy} + c\mathbf{u}_{yy}$

Definition 27.12 PDEs Discriminante: Is defined by:
 $\delta(\mathcal{L}) := -\det\left(\begin{pmatrix} a & b \\ b & c \end{pmatrix}\right) = b^2 - ac \quad (27.9)$

Explanation 27.1. It turns out that many fundamental properties of the solution of eq. (27.8) are determined by its principal part, or rather by the sign of the discriminant $\delta(\mathcal{L})$.

Definition 27.13 [??]
Parabolic PDEs: Let [def. 27.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called hyperbolic if:
 $\delta(\mathcal{L}) = b^2 - ac = 0 \quad (27.10)$

Definition 27.14 [??]
Hyperbolic PDEs: Let [def. 27.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called hyperbolic if:
 $\delta(\mathcal{L}) = b^2 - ac > 0 \quad (27.11)$

Definition 27.15 [??]
Parabolic PDEs: Let [def. 27.10] be a PDE defined on $\Omega \subset \mathbb{R}^2$, then the PDE is called elliptic if:
 $\delta(\mathcal{L}) = b^2 - ac < 0 \quad (27.12)$

Explanation 27.2.
The reason for this categorization are normal quadratic equations in two variables:
 $Ax^2 + By^2 + Cxy + Dx + Ey + f = 0$
If $B^2 - 4AC = 0 \Rightarrow$ the equation is a parabola.
If $B^2 - 4AC > 0 \Rightarrow$ the equation is a hyperbola.
If $B^2 - 4AC < 0 \Rightarrow$ the equation is an ellipse.

2. Method Of Characteristics

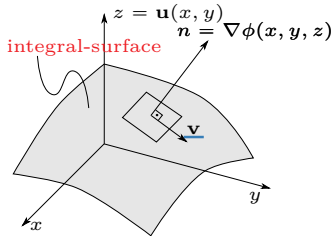
Is a method that makes use of geometrical aspects in order to solve 1st-order PDEs with two variables by constructing integral surfaces and can be used to solve PDEs of the type:
Linear: $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y) \quad (27.13)$
Semilin.: $a(x, y)\mathbf{u}_x + b(x, y)\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (27.14)$
Quasilin.: $a(x, y, \mathbf{u})\mathbf{u}_x + b(x, y, \mathbf{u})\mathbf{u}_y = c(x, y, \mathbf{u}) \quad (27.15)$

Formula 27.1 Method of Characteristics:
 $x := x(r; s) \quad y := y(r; s) \quad z := u(r; s)$
Parameter.: $\lambda(r; s) := x(r; s)\mathbf{e}_x + y(r; s)\mathbf{e}_y + z(r; s)\mathbf{e}_z$
 $\frac{\partial \lambda}{\partial r}(r; s) = (a, b, c)$
 $\mathbf{v} := v(x(r; s), y(r; s), z(r; s))$
E.g.
 $\frac{\partial x}{\partial r}(r; s) = \dot{x} = a(\lambda_s(r))$
 $\frac{\partial y}{\partial r}(r; s) = \dot{y} = b(\lambda_s(r))$
 $\frac{\partial z}{\partial r}(r; s) = \dot{z} = c(\lambda_s(r))$
Compact:
 $\dot{x} = a(x, y, u) \quad \dot{y} = b(x, y, u) \quad \dot{u} = c(x, y, u)$
I.C.: $x(0; s) = x_0(s) \quad y_0(0; s) = y_0(s) \quad u_0(0; s) = u_0(s)$

Definition 27.16 Integral Surface ϕ :
An function $\phi : \mathbb{R}^3 \mapsto \mathbb{R}$ is a an integral surface of a vector field $\mathbf{V} : \mathbb{R}^3 \mapsto \mathbb{R}^3$ if ϕ is a surface that has in every point a tangent plane containing a vector $\mathbf{v} = (a \ b \ c)$ of \mathbf{V} .

Corollary 27.3 PDEs and Integral Surfaces:
The solution of a PDE $\mathbf{u}(x, y)$ can be thought of as an integral surface:
 $z = u(x, y) \quad$ or implicitly $\phi(x, y, z) = u(x, y) - z \quad (27.16)$

Explanation 27.3 ([proof ??])
Integral Surface and PDEs).
The solution $\mathbf{u}(x, y)$ of eq. (27.13) can be sought of as an surface $z = \mathbf{u}(x, y)$ in \mathbb{R}^3 or in implicit form $\phi(x, y, z) := \mathbf{u}(x, y) - z$.



Let: $\mathbf{n}(x, y) := \text{grad } \phi = \begin{pmatrix} \phi_x \\ \phi_y \\ \phi_z \end{pmatrix} = \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix} \quad$ and

Let $\mathbf{V} := \begin{pmatrix} a(x, y) \\ b(x, y) \\ c(x, y) \end{pmatrix}$ be a vector field $\mathbb{R}^3 \mapsto \mathbb{R}^3$ and

$\mathbf{n}(x, y) := \text{grad } \phi = \begin{pmatrix} \phi_x \\ \phi_y \\ \phi_z \end{pmatrix} = \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix}$

Idea: we can rewrite eq. (27.13) as:

$\left\langle \begin{pmatrix} a & b & c \end{pmatrix}^\top, \nabla \phi(x, y, z) \right\rangle = \left\langle \begin{pmatrix} a(x, y) \\ b(x, y) \\ c(x, y) \end{pmatrix}, \begin{pmatrix} \mathbf{u}_x \\ \mathbf{u}_y \\ -1 \end{pmatrix} \right\rangle = 0$

Geometric Interpretation:
 \mathbf{v} is orthogonal to the normal \mathbf{n} for all points $(x, y, \mathbf{u}(x, y))$.
Hence every vector $\mathbf{v} = (a \ b \ c)^\top$ lies in the tangent plane containing ϕ .
Consequently in order to find a surface ϕ (and thus also a solution \mathbf{u}), we need to search for ϕ s.t. the vector \mathbf{v} lies in the tangent plane for every possible point of ϕ .

Idea

$\gamma(r)$ We first simplify the task and start by constructing/finding integral curves λ and then we construct the integral surface ϕ out of this curves.

3. Linear Equations

Definition 27.17
Characteristic/Integral Curve $\lambda_s(r) = \lambda(r; s)$:
Given a vector field \mathbf{V} an integral curve $\lambda(r)$ of that vector field, is a curve parameterized by parameter r :

$\lambda(r) := x(r)\mathbf{e}_x + y(r)\mathbf{e}_y + z(r)\mathbf{e}_z = \begin{pmatrix} x(r) \\ y(r) \\ z(r) \end{pmatrix} \quad (27.17)$

s.t. at each point r of the curve a vector \mathbf{v} of the vector field:

$\mathbf{v} = \begin{pmatrix} a(x(r), y(r)) \\ b(x(r), y(r)) \\ c(x(r), y(r)) \end{pmatrix} \in \mathbf{V} \quad (27.18)$

is tangent to the curve:

$\frac{d\lambda(r)}{dr} = \mathbf{V}(\lambda(r)) = \begin{pmatrix} a(x(r), y(r)) \\ b(x(r), y(r)) \\ c(x(r), y(r)) \end{pmatrix} = \begin{pmatrix} a(\lambda(r)) \\ b(\lambda(r)) \\ c(\lambda(r)) \end{pmatrix} \quad (27.19)$

Definition 27.18 Characteristic Equations:
The set of ordinary differential equations of a PDE arising from Equation (27.19) are called characteristic equations:

$\frac{dx(r)}{dr} = \dot{x} = \underline{a(\lambda(r))} = a(r) \quad (27.20)$

$\frac{dy(r)}{dr} = \dot{y} = \underline{b(\lambda(r))} = b(r) \quad (27.21)$

$\frac{dz(r)}{dr} = \dot{z} = \underline{c(\lambda(r))} = c(r) \quad (27.22)$

Problem: in order to get a unique solution we need to specify initial conditions.

Idea: If a characteristic has an arbitrary point in common with the integral surface ϕ then the whole characteristic λ will lie in the integral surface.

Proof 27.1: Let: $\phi(\lambda(r)) = u(x(r), y(r)) - z(r)$
 $\Rightarrow \frac{d\phi}{dr} = u_x \frac{dx}{dr} + u_y \frac{dy}{dr} - 1 \stackrel{27.22}{=} \begin{pmatrix} u_x \\ u_y \\ -1 \end{pmatrix} \cdot \begin{pmatrix} a(\lambda(r), y(r)) \\ b(\lambda(r), y(r)) \\ c(\lambda(r), y(r)) \end{pmatrix} = \begin{pmatrix} u_x \\ u_y \\ -1 \end{pmatrix} \cdot \dot{\lambda}(r) = 0$

Thus: $\phi(\lambda(r_0)) = 0 \Leftrightarrow \phi(\lambda(r)) = 0, \forall r$

Definition 27.19
Characteristic (Curve) $\lambda_s(r) = \lambda(r; s)$:
is an integral curve of the vector field \mathbf{V} that is uniquely determined by a parameter s .

Consequence: For every characteristic s we need to specify one initial point on the integral surface in order to have all the characteristics lie within the integralsurface.

Idea: we define another curve $\Gamma(s)$ on the integralsurface that transvers all the characteristic curves $\lambda_s(r)$ transversal (=angle between $\Gamma(s)$ and $\lambda_s(r)$ is never zero $\Leftrightarrow \Gamma(s) \nparallel \lambda_s(r)$).

Definition 27.20 Inital Condition: $s \mapsto \Gamma(s), \Gamma : \mathbb{R} \mapsto \mathbb{R}^3$
 $\lambda_s(r) = \begin{pmatrix} x_s(r) \\ y_s(r) \\ z_s(r) \end{pmatrix}, \quad \Gamma(s) = \begin{pmatrix} x_0(s) \\ y_0(s) \\ z_0(s) \end{pmatrix} \quad \lambda_s(0) \stackrel{!}{=} \Gamma(s)$
 $\Rightarrow \underline{x_s(0) = x_0(s)} \quad \underline{y_s(0) = y_0(s)} \quad \underline{z_s(0) = z_0(s)}$

Definition 27.21
Projected Characteristic Curves $\gamma(\tau)$:
Are curves in the plane of the independent variables of our PDE, along which u is constant or satisfies certain conditions. If u is constant along $g(\tau)$ then the initial data is simply propagated along those characteristic curves:

$\frac{d}{d\gamma} u(\gamma(\tau), \tau) = 0 \Leftrightarrow u(\gamma(\tau), \tau) = u_0(\gamma(\tau)) \quad (27.23)$

Hint: If the PDE is linear, then the two first characteristics do not depend on u and can be solved directly, u will then be constant along those characteristics:

$$\begin{matrix} a(x,y)u_x + b(x,y)u_y = c(x,y) \\ \frac{dx}{dr} = a & \frac{dy}{dr} = b & \frac{du}{dr} = c \end{matrix} \text{ implies } \frac{dy}{dx} = \frac{b(x,y)}{a(x,y)}$$

Hint: If we divide the PDE by a we have to solve a PDE less, beacause the first ODE will allways be:

$$\dot{x} = 1 \Rightarrow x = r \Rightarrow x_s(r) = x_0(s)$$

4. Quasilinear Equations

Solving Quasilinear Equations

$$\begin{matrix} a(x,y,u)u_x + b(x,y,u)u_y = c(x,y,u) \\ u|_{\Gamma}(r,s) = \phi(s) \\ \frac{dx}{dr} = a(x,y,u) & \frac{dy}{dr} = b(x,y,u) & \frac{du}{dr} = c(x,y,u) \\ x_s(0) = x_0(s) & y_s(0) = y_0(s) & z_s(0) = \phi(s) \end{matrix}$$

Results

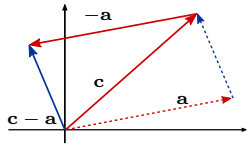
Now the projected characteristic curves may depend on u as well as on x,y. **Thus** the first two characteristics are no longer decoupled form the third one.

1. We may get projected characteristic curves crossing themselves.
2. u is no longer constant along the projected characteristic curves, rather the PDE reduces to an ODE satisfying certain conditions along this curves.

Linear Algebra

1. Vectors

Definition 28.1 Vector Subtraction:



$$\mathbf{b} = \mathbf{c} - \mathbf{a} \quad (28.1)$$

2. Linear Systems of Equations

2.1. Gaussian Elimination

2.1.1. Rank

Definition 28.2 Matrix Rank

The ranks of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as the dimension^[def. 28.13] of the vector space spanned^[def. 28.9] by its row or column vectors:

$$\begin{aligned} \text{rank}(\mathbf{A}) &= \dim(\{\mathbf{a}_{:,1}, \dots, \mathbf{a}_{:,n}\}) \\ &= \dim(\{\mathbf{a}_{1,:}, \dots, \mathbf{a}_{m,:}\}) \\ &\stackrel{\text{def. 28.50}}{=} \dim(\mathfrak{R}(\mathbf{A})) \end{aligned} \quad (28.2)$$

Corollary 28.1 :

- 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:

$$\text{rank}(\mathbf{A}) \leq \min\{n, m\} \quad (28.3)$$

Property 28.1 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:

$$\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\} \quad (28.4)$$

Rank-1 Matrix

Definition 28.3 Rank-1 Matrix:

Is a matrix of rank one. A tensor product of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ results in a rank one matrix:

$$\mathbf{uv}^T = \mathbf{A} \in \mathbb{R}^{n,n} \quad (28.5)$$

Definition 28.4 Rank-1 Modification/Update:

Adding a rank-1 matrix to another matrix is called rank-1 modification:

$$\mathbf{X} = \mathbf{X} + \mathbf{uv}^T \quad (28.6)$$

3. Sparse Linear Systems

Definition 28.5 Sparse Matrix

$\mathbf{A} \in \mathbb{K}^{m,n}$, $m, n \in \mathbb{N}_{>0}$: A matrix \mathbf{A} is sparse if:

$$\text{nnz}(\mathbf{A}) \ll mn \quad \mathbf{A} \in \mathbb{K}^{m,n}, m, n \in \mathbb{N}_{>0} \quad (28.7)$$

$$\text{nnz} := \#\{(i, j) \in \{1, \dots, m\} \times \{1, \dots, n\} : a_{i,j} \neq 0\}$$

4. Vector Spaces

4.1. Vector Space

Definition 28.6 Vector Space: TODO

4.2. Vector Subspace

Definition 28.7 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:

$$\mathbf{u}, \mathbf{v} \in U \implies \mathbf{u} + \mathbf{v} \in U \quad (28.8)$$

$$\mathbf{u} \in U \implies \lambda \mathbf{u} \in U \quad \forall \lambda \in \mathbb{K} \quad (28.9)$$

Definition 28.8 Linear combination:

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:

$$\mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{v}_i = \alpha_1 \mathbf{v}_1 + \dots + \alpha_n \mathbf{v}_n \quad \alpha_i \in \mathbb{K} \quad (28.10)$$

Definition 28.9

Span/Linear Hull

Is the set of all possible linear combinations^[def. 28.8] of finite set $X = \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathcal{V}$ of a \mathbb{K} vector space \mathcal{V} :

$$\langle X \rangle = \text{span}(X) = \left\{ \mathbf{v} \mid \sum_{i=1}^n \alpha_i \mathbf{v}_i, \forall \alpha_i \in \mathbb{K} \right\} \quad (28.11)$$

Definition 28.10 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. 28.9] \mathcal{V} :

$$\text{span}(\mathbf{v}_1, \dots, \mathbf{v}_m) = \mathcal{V} \quad (28.12)$$

Explanation 28.1 (Definition 28.10).

The generating set of vector space (or set of vectors) $\mathcal{V} \stackrel{i.e.}{=} \mathbb{R}^n$ 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 $\text{span}(X)$.

Definition 28.11 Linear Independence: A set of vector $\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \in \mathcal{V}$ is called linear independent if the satisfy:

$$\mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{v}_i = \mathbf{0} \iff \alpha_1 = \dots = \alpha_n = 0 \quad (28.13)$$

Corollary 28.2 : 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:

$$\langle X \rangle \subsetneq \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \quad (28.14)$$

4.3. Basis

Definition 28.12 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:

$$\langle \mathfrak{B} \rangle = \mathcal{V} \quad \text{and} \quad \mathfrak{B} \text{ is a linear independent generating set} \quad (28.15)$$

Corollary 28.3 : The unit vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ build a standard basis of the \mathbb{R}^n .

Corollary 28.4 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. 28.8] of \mathfrak{B} by a unique set of coefficients α_i :

$$\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{b}_i \quad \begin{matrix} \alpha_1, \dots, \alpha_n \in \mathbb{K} \\ \mathbf{b}_1, \dots, \mathbf{b}_n \in \mathfrak{B} \end{matrix} \quad (28.16)$$

4.3.1. Dimensionality

Definition 28.13 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} :

$$\dim(\mathcal{V}) := |\mathfrak{B}| = n \in \mathbb{N}_0 \quad (28.17)$$

Corollary 28.5 : 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 $\dim(\mathcal{V}) = \infty$.

4.4. Affine Subspaces

Definition 28.14 Affine Subspaces: Given a \mathbb{K} -vector space \mathcal{V} of dimension $\dim(\mathcal{V}) \geq 2$ a sub vector space^[def. 28.7] U of \mathcal{V} defined as:

$$\mathcal{W} := \mathbf{v} + U = \{\mathbf{v} + \mathbf{x} \mid \mathbf{x} \in U\} \quad \mathbf{v} \in \mathcal{V} \quad (28.18)$$

Corollary 28.6 Direction: The sub vector spaces U are called directions of \mathcal{V} and it holds:

$$\dim(\mathcal{W}) := \dim(U) \quad (28.19)$$

4.4.1. Hyperplanes

Definition 28.15 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. 28.16]:

$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} - d = 0\} \quad (28.20)$$

Corollary 28.7 Half spaces: A hyperplane $\mathcal{H} \in \mathbb{R}^{d-1}$ separates its d -dimensional ambient space into two half spaces:

$$\mathcal{H}^+ = \{x \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} + b > 0\} \quad (28.21)$$

$$\mathcal{H}^- = \{x \in \mathbb{R}^d \mid \hat{\mathbf{n}}^T \mathbf{x} + b < 0\} = \mathbb{R}^d - \mathcal{H}^+ \quad (28.22)$$

Notes

Hyperplanes in \mathbb{R}^2 are lines and hyperplanes in \mathbb{R}^3 are planes.

Hess Normal Form

Definition 28.16 Hess Normal Form: Is an equation to describe hyperplanes^[def. 28.15] in \mathbb{R}^d :

$$\mathbf{r}^T \hat{\mathbf{n}} - d = 0 \iff \hat{\mathbf{n}}^T (\mathbf{r} - \mathbf{r}_0) \quad \mathbf{r}_0 := \mathbf{r}^T d \geq 0 \quad (28.23)$$

where all points described by the vector $\mathbf{r} \in \mathbb{R}^d$, that satisfy this equations lie on the hyperplane.

Note

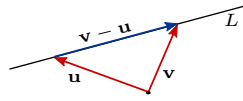
The direction of the unit normal vector is usually chosen s.t. $\mathbf{r}^T \hat{\mathbf{n}} \geq 0$.

4.4.2. Lines

Definition 28.17 Lines: Lines are a set^[def. 19.1] of the form: $L = \mathbf{u} + \mathbb{K} \mathbf{v} = \{\mathbf{u} + \lambda \mathbf{v} \mid \lambda \in \mathbb{K}\} \quad \mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{v} \neq \mathbf{0} \quad (28.24)$

Two Point Formula

Definition 28.18 Two Point Formula:



$$L = \mathbf{u} + \mathbb{K} \mathbf{v} \quad (28.25)$$

4.4.3. Planes

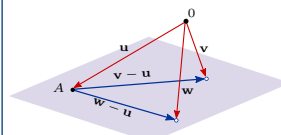
Definition 28.19 Planes: Planes are sets defined as:

$$E = \mathbf{u} + \mathbb{K} \mathbf{v} + \mathbb{K} \mathbf{w} = \{\mathbf{u} + \lambda \mathbf{v} + \mu \mathbf{w} \mid \lambda, \mu \in \mathbb{K}\} \quad (28.26)$$

$$\mathbf{u}, \mathbf{w} \in \mathcal{V} \quad \text{s.t. } \mathbf{v}, \mathbf{u} \neq \mathbf{0} \quad \text{and} \quad \mathbf{v}, \mathbf{w} \text{ lin. indep.}$$

Parameterform

Definition 28.20 Two Point Formula:

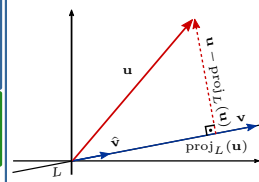


$$E = \mathbf{u} + \mathbb{K}(\mathbf{v} - \mathbf{u}) + \mathbb{K}(\mathbf{w} - \mathbf{u}) \quad (28.27)$$

4.4.4. Minimal Distance of Vector Subspaces

Projections in 2D

Definition 28.21 2D Vector Projection [Proof 28.17, 28.18]:



$$\begin{aligned} \mathbf{u}_v &= \text{proj}_L(\mathbf{u}) \\ &= u_v \hat{\mathbf{v}} = (\mathbf{u}^T \hat{\mathbf{v}}) \hat{\mathbf{v}} \\ &= \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v} = \frac{\mathbf{u}^T \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \mathbf{v} \end{aligned} \quad (28.28)$$

Corollary 28.8

2D Projection Matrix \mathbf{P} : Is the matrix that satisfies:

$$\mathbf{P} \mathbf{u} = \text{proj}_L(\mathbf{u}) \quad \mathbf{P} = \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} = \frac{\mathbf{v} \mathbf{v}^T}{\|\mathbf{v}\|^2} \quad (28.29)$$

Proof 28.1: [Corollary 28.8]

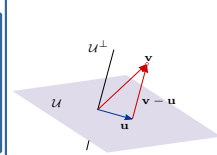
$$\frac{1}{\mathbf{v}^T \mathbf{v}} \mathbf{u}^T \mathbf{v} \mathbf{v} = \frac{1}{\mathbf{v}^T \mathbf{v}} \mathbf{v} (\mathbf{v}^T \mathbf{u}) = \frac{1}{\mathbf{v}^T \mathbf{v}} (\mathbf{v} \mathbf{v}^T) \mathbf{u}$$

General Projections

Definition 28.22

General Vector Projection:

Is the orthogonal projection \mathbf{u} of a vector \mathbf{v} onto a sub-vector space \mathcal{U}



$$\mathbf{u} = \sum_{i=1}^n \alpha_i \mathbf{b}_i \quad (28.30)$$

$$\mathbf{A} \mathbf{A}^T \alpha_i = \mathbf{A}^T \mathbf{v} \quad \mathbf{A} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix}$$

where $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a basis of the vector subspace \mathcal{U} .

Theorem 28.1 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. 28.67] projection

$$p: \begin{cases} \mathcal{V} \rightarrow \mathcal{U} \\ \mathbf{v} \mapsto \mathbf{u} \end{cases} \quad (28.31)$$

the vector $\mathbf{u}' := \mathbf{v} - \mathbf{u}$ representing the distance between \mathbf{u} and \mathbf{v} and is minimal:

$$\|\mathbf{u}'\| = \|\mathbf{v} - \mathbf{u}\| \leq \|\mathbf{v} - \mathbf{w}\| \quad \forall \mathbf{w} \in \mathcal{U} \quad \mathbf{u}' \in \mathcal{U}^\perp \quad (28.32)$$

4.5. Affine Subspaces

4.6. Planes

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

5. Matrices

Special Kind of Matrices

5.1. Symmetric Matrices

Definition 28.23 Symmetric Matrices: A matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is called *symmetric* if it satisfies:
$$\mathbf{A} = \mathbf{A}^T \tag{28.33}$$

Property 28.2 [proof ??]
Eigenvalues of real symmetric Matrices: The eigenvalues of a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are real:
$$\text{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geq 0}\}_{i=1}^n \tag{28.34}$$

Property 28.3 [proof ??]
Orthogonal Eigenvector basis: Eigenvectors of real symmetric matrices with distinct eigenvalues are orthogonal.

Corollary 28.9
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{R}^{n,n}$ is a real *symmetric*[def. 28.23] matrix then its eigenvectors are *orthogonal* and its eigen-decomposition[def. 28.86] is given by:
$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T \tag{28.35}$$

5.2. Orthogonal Matrices

Definition 28.24 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. 28.68] basis:
$$\langle \mathbf{q}_{\cdot i}, \mathbf{q}_{\cdot j} \rangle = \delta_{ij} \quad \text{and} \quad \langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_{ij} \tag{28.36}$$

This is exactly true if the inverse of \mathbf{Q} equals its transpose:
$$\mathbf{Q}^{-1} = \mathbf{Q}^T \iff \mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n \tag{28.37}$$

Attention: *Orthogonal* matrices are sometimes also called *orthonormal matrices*.

5.3. Hermitian Matrices

Definition 28.25 Conjugate Transpose $\mathbf{A}^H / \mathbf{A}^*$
Hermitian Conjugate/Adjoint Matrix:
The conjugate transpose of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is defined as:
$$\mathbf{A}^H := (\mathbf{A}^T)^* = \overline{\mathbf{A}^T} \iff \mathbf{a}_{i,j}^H = \overline{\mathbf{a}_{j,i}} \quad \begin{matrix} 1 \leq i \leq n \\ 1 \leq j \leq m \end{matrix} \tag{28.38}$$

Definition 28.26
Hermitian/Self-Adjoint Matrices $\mathbf{A} = \mathbf{A}^H$:
A hermitian matrix is complex square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ who is equal to its own *conjugate transpose*[def. 28.25]:
$$\mathbf{A} = \mathbf{A}^H = \overline{\mathbf{A}^T} \iff \mathbf{a}_{i,j} = \overline{\mathbf{a}_{j,i}} \quad i \in \{1, \dots, n\} \tag{28.39}$$

Corollary 28.10 : [def. 28.25] implies that \mathbf{A} must be a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Corollary 28.11 Real Hermitian Matrices: From [cor. 19.1] it follows:
$$\mathbf{A} \in \mathbb{R}^{n \times n} \text{ hermitian} \implies \mathbf{A} \text{ real symmetric} \tag{28.40}$$

Property 28.4 [proof 28.15]
Eigenvalues of Hermitan Matrices: The eigenvalues of a hermitian matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ are real:
$$\text{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geq 0}\}_{i=1}^n \tag{28.41}$$

Property 28.5 [proof 28.16]
Orthogonal Eigenvector basis: Eigenvectors of hermitian matrices with distinct eigenvalues are orthogonal.

Corollary 28.12
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{C}^{n,n}$ is a hermitian matrix[def. 28.26] then its eigendecomposition[def. 28.86] is given by:
$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^H \tag{28.42}$$

5.4. Unitary Matrices

Definition 28.27 Unitary Matrix $\mathbf{U}\mathbf{U}^H$:
is a complex square matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ whose inverse[def. 28.41] is equal to its *conjugate transpose*[def. 28.25]:
$$\mathbf{U}^H \mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I} \tag{28.43}$$

Corollary 28.13 Real Unitary Matrix: A real matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ that is unitary is an *orthogonal matrix*[def. 28.24].

Property 28.6
Preservation of Euclidean Norm [proof 28.14]:
Orthogonal and unitary matrices $\mathbf{Q} \in \mathbb{K}^{n,n}$ do not affect the 2-norm:
$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathbb{K}^n \tag{28.44}$$

5.5. Similar Matrices

Definition 28.28 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.:
$$\exists \mathbf{S} : \mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{28.45}$$

Corollary 28.14
Similarity Transformation/Conjugation:
The mapping:
$$\mathbf{A} \mapsto \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{28.46}$$

is called *similarity transformation*

Corollary 28.15
Eigenvalues of Similar Matrices [proof 28.13]:
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. (28.46) \mathbf{B} has the same eigenvalues with transformed eigenvectors:
$$\{\{\lambda_i, \mathbf{u}_i\}\}_{i=1}^n \quad \mathbf{u}_i := \mathbf{S}^{-1}\mathbf{v}_i \tag{28.47}$$

5.6. Skew Symmetric Matrices

Definition 28.29
Key Symmetric/Antisymmetric Matrices:
$$\mathbf{A}^T = -\mathbf{A} \tag{28.48}$$

5.7. Triangular Matrix

Definition 28.30 Triangular Matrix: An upper (lower) triangular matrix, is a matrix whose element's below (above) the main diagonal are all zero:

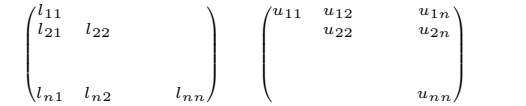


Figure 11: Lower Tri. Mat. Figure 12: Upper Tri. Mat.

5.7.1. Unitriangular Matrix

Definition 28.31 Unitriangular Matrix: An upper (lower) unitriangular matrix, is a upper (lower) triangular matrix[def. 28.30] whose diagonal elements are all ones.

5.7.2. Strictly Triangular Matrix

Definition 28.32 Strictly Triangular Matrix: An upper (lower) strictly triangular matrix, is a upper (lower) triangular matrix[def. 28.30] whose diagonal elements are all zero.

5.8. Block Partitioned Matrices

Definition 28.33 Block Partitioned Matrix:
A matrix $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ can be partitioned into a *block partitioned matrix*:
$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \tag{28.49}$$

Definition 28.34 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*:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \quad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l}, \mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^k, \mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^l \tag{28.50}$$

5.8.1. Schur Complement

Definition 28.35 Schur Complement: Given a block partitioned matrix[def. 28.33] $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ its Schur complements are given by:
$$\mathbf{S}_A = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \quad \mathbf{S}_D = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \tag{28.51}$$

5.8.2. Inverse of Block Partitioned Matrix

Definition 28.36 proof 28.3
Inverse of a Block Partitioned Matrix:
Given a block partitioned matrix[def. 28.33] $\mathbf{M} \in \mathbb{R}^{k+l, k+l}$ its inverse \mathbf{M}^{-1} can be partitioned as well:
$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \mathbf{M}^{-1} = \begin{bmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{C}} & \tilde{\mathbf{D}} \end{bmatrix} \tag{28.52}$$

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}_A^{-1}\mathbf{C}\mathbf{A}^{-1} & \tilde{\mathbf{C}} &= -\mathbf{S}_A^{-1}\mathbf{C}\mathbf{A}^{-1} \\ \tilde{\mathbf{B}} &= -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}_A^{-1} & \tilde{\mathbf{D}} &= \mathbf{S}_A^{-1} \end{aligned}$$

where $\mathbf{S}_A = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is the Schur complement of \mathbf{A} .

5.9. Properties of Matrices

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

Definition 28.37 Square Root:

5.9.2. Trace

Definition 28.38 Trace: The trace of an $\mathbf{A} \in \mathbb{R}^{n \times n}$ matrix is defined as:
$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \dots + a_{nn} \tag{28.53}$$

Property 28.7 Trace of a Scalar:
$$\text{tr}(\mathbb{R}) = \mathbb{R} \tag{28.54}$$

Property 28.8 Trace of Transpose:
$$\text{tr}(\mathbf{A}^T) = \text{tr}(\mathbf{A}) \tag{28.55}$$

Property 28.9 Trace of multiple Matrices:
$$\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CBA}) \tag{28.56}$$

6. Matrices and Determinants

6.1. Determinants

6.1.1. Laplace/Cofactor Expansion

Definition 28.39 Minor:

Definition 28.40 Cofactors:

Properties

Property 28.10 Determinant times Scalar $\det(\alpha\mathbf{A})$:
Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ it holds:
$$\det(\alpha \cdot \mathbf{A}) = \alpha^n \mathbf{A} \tag{28.57}$$

6.2. Inverse of Matrices

Definition 28.41 Inverse Matrix \mathbf{A}^{-1} :

6.2.1. Invertability

Definition 28.42
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:

- $\det(\mathbf{A}) = 0$
- $\mathbf{A}\mathbf{x} = \mathbf{b}$ has either
 - no solution \mathbf{x}
 - infinitely many solutions \mathbf{x}
- $\dim(\mathbf{A}) < n$
- $\nexists \mathbf{B} : \mathbf{B} = \mathbf{A}^{-1}$

Transformations And Mapping

7. Linear & Affine Mappings/Transformations

7.1. Linear Mapping

Definition 28.43
Linear Mapping: A linear mapping, function or transformation is a map $l : V \mapsto W$ between two \mathbb{K} -vector spaces^[def. 28.6] V and W if it satisfies:

$$l(\mathbf{x} + \mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y}) \quad (\text{Additivity}) \quad (28.58)$$
$$l(\alpha \mathbf{x}) = \alpha l(\mathbf{x}) \quad \forall \alpha \in \mathbb{K} \quad (\text{Homogeneity}) \quad (28.59)$$
$$\forall \mathbf{x}, \mathbf{y} \in V$$

Proposition 28.1^[proof 28.8]
Equivalent Formulations: Definition 28.43 is equivalent to:

$$l(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha l(\mathbf{x}) + \beta l(\mathbf{y}) \quad \forall \alpha, \beta \in \mathbb{K} \quad \forall \mathbf{x}, \mathbf{y} \in V \quad (28.60)$$

Corollary 28.16 Superposition Principle:
Definition 28.43 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 28.17^[proof 28.10]
A linear mapping $\iff \mathbf{A}\mathbf{x}$:
For every matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ the map:

$$l_{\mathbf{A}} : \begin{cases} \mathbb{K}^n & \rightarrow \mathbb{K}^m \\ \mathbf{x} & \mapsto \mathbf{A}\mathbf{x} \end{cases} \quad (28.61)$$

is a linear map and every linear map l can be represented by a matrix vector product:

$$l \text{ is linear} \iff \exists \mathbf{A} \in \mathbb{K}^{n \times m} : f(\mathbf{x}) = \mathbf{A}\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{K}^m \quad (28.62)$$

Principle 28.1^[proof 28.9]
Principle of linear continuation: A linear mapping $l : \mathcal{V} \mapsto \mathcal{W}$ is determined by the image of the basis \mathcal{B} of \mathcal{V} :

$$l(\mathbf{v}) = \sum_{i=1}^n \beta_i l(\mathbf{b}_i) \quad \mathcal{B}(\mathcal{V}) = \{\mathbf{b}_1, \dots, \mathbf{b}_n\} \quad (28.63)$$

Property 28.11^[proof 28.11]
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. 28.43].

Definition 28.44 Level Sets:

7.2. Affine Mapping

Definition 28.45 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:

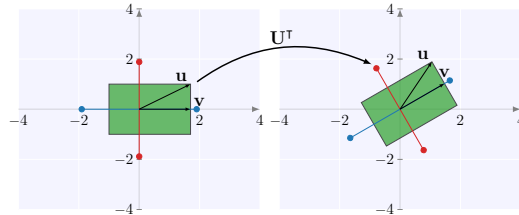
$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (28.64)$$

is called an affine transformation of \mathbf{x} .

7.3. Orthogonal Transformations

Definition 28.46 Orthogonal Transformation:
A linear transformation $T : \mathcal{V} \mapsto \mathcal{V}$ of an inner product space^[def. 28.78] is an orthogonal transformation if preserves the inner product:

$$T(\mathbf{u}) \cdot T(\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{V} \quad (28.65)$$



Corollary 28.18 Orthogonal Matrix Transformation:
An orthogonal matrix^[def. 28.24] \mathbf{Q} provides an orthogonal transformation:

$$(\mathbf{Q}\mathbf{u})^T (\mathbf{Q}\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} \quad (28.66)$$

Explanation 28.2 (Improper Rotations).
Orthogonal transformations in two or three dimensional euclidean space^[def. 28.46] represent improper rotations:

- Stiff Rotations
- Reflections
- Reflections+Rotations

Corollary 28.19 Preservation of Orthogonality: Orthogonal transformation preserves orthogonality.

Corollary 28.20^[proof 28.6]
Preservation of Norm:
An orthogonal transformation $\mathbf{Q} : \mathcal{V} \mapsto \mathcal{V}$ preserves the length/norm:

$$\|\mathbf{u}\|_{\mathcal{V}} = \|\mathbf{Q}\mathbf{u}\|_{\mathcal{V}} \quad (28.67)$$

Corollary 28.21 Preservation of Angle:
An orthogonal transformation T preserves the angle^[def. 28.66] of its vectors:

$$\angle(\mathbf{u}, \mathbf{v}) = \angle(T(\mathbf{u}), T(\mathbf{v})) \quad (28.68)$$

7.4. Kernel & Image
7.4.1. Kernel

Definition 28.47 Kernel/Null Space $\mathcal{N}/\varphi^{-1}(\{0\})$:
Let φ be a linear mapping^[def. 28.43] between two \mathbb{K} -vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The kernel of φ is defined as:

$$\mathcal{N}(\varphi) := \varphi^{-1}(\{0\}) = \{\mathbf{v} \in \mathcal{V} \mid \varphi(\mathbf{v}) = \mathbf{0}\} \subseteq \mathcal{V} \quad (28.69)$$

Definition 28.48 Right Null Space $\mathcal{N}(\mathbf{A})$:
If $\varphi = \mathbf{A} \in \mathbb{K}^{m \times n}$ then the eq. (28.69) is equal to:

$$\mathcal{N}(\mathbf{A}) = \varphi_{\mathbf{A}}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^n \mid \mathbf{A}\mathbf{v} = \mathbf{0}\} \in \mathbb{K}^m \quad (28.70)$$

Definition 28.49 Left Null Space $\mathcal{N}(\mathbf{A}^T)$:
If $\varphi = \mathbf{A} \in \mathbb{K}^{m \times n}$ then the left null space is defined as:

$$\mathcal{N}(\mathbf{A}^T) = \varphi_{\mathbf{A}^T}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^m \mid \mathbf{A}^T \mathbf{v} = \mathbf{0}\} \in \mathbb{K}^n \quad (28.71)$$

Note
The term left null space stems from the fact that:

$$(\mathbf{A}^T \mathbf{x})^T = \mathbf{0} \quad \text{is equal to} \quad \mathbf{x}^T \mathbf{A} = \mathbf{0}$$

7.4.2. Image

Definition 28.50 Image/Range \mathcal{R}/φ :
Let φ be a linear mapping^[def. 28.43] between two \mathbb{K} -vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The image of φ is defined as:

$$\mathcal{R}(\varphi) := \varphi(\mathcal{V}) = \{\varphi(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}\} \subseteq \mathcal{W} \quad (28.72)$$

Definition 28.51 Column Space $\mathcal{A}\mathbf{x}$:
If $\varphi = \mathbf{A} = (\mathbf{c}_1 \quad \dots \quad \mathbf{c}_n) \in \mathbb{K}^{m \times n}$ then eq. (28.72) is equal to:

$$\begin{aligned} \mathcal{R}(\mathbf{A}) &= \varphi_{\mathbf{A}}(\mathbb{K}^n) = \{\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{K}^n\} = \left\langle \mathbf{c}_1 \quad \dots \quad \mathbf{c}_n \right\rangle \\ &= \left\{ \mathbf{v} \mid \sum_{i=1}^n \alpha_i \mathbf{c}_i, \forall \alpha_i \in \mathbb{K} \right\} \end{aligned} \quad (28.73)$$

Definition 28.52 Row Space $\mathbf{A}^T \mathbf{x}$:
If $\varphi = \mathbf{A} = (\mathbf{r}_1^T \quad \dots \quad \mathbf{r}_m^T) \in \mathbb{K}^{m \times n}$ then the column space is defined as:

$$\begin{aligned} \mathcal{R}(\mathbf{A}^T) &= \varphi_{\mathbf{A}}(\mathbb{K}^m) = \{\mathbf{A}^T \mathbf{x} \mid \mathbf{x} \in \mathbb{K}^m\} = \left\langle \mathbf{r}_1 \quad \dots \quad \mathbf{r}_m \right\rangle \\ &= \left\{ \mathbf{v} \mid \sum_{i=1}^m \alpha_i \mathbf{r}_i, \forall \alpha_i \in \mathbb{K} \right\} \end{aligned} \quad (28.74)$$

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

8. Eigenvalues and Vectors

Definition 28.53 Eigenvalues: Given a square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ the eigenvalues

Definition 28.54 Spectrum: The spectrum of a square matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ is the set of its eigenvalues^[def. 28.53]:

$$\text{spectrum}(\mathbf{A}) = \lambda(\mathbf{A}) = \{\lambda_1, \dots, \lambda_n\} \quad (28.78)$$

Formula 28.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix \mathbf{A} its eigenvalues can be calculated by:

$$\{\lambda_1, \lambda_2\} \in \frac{\text{tr}(\mathbf{A}) \pm \sqrt{\text{tr}(\mathbf{A})^2 - 4 \det(\mathbf{A})}}{2} \quad (28.79)$$

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

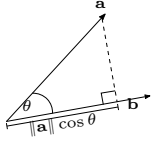
9. Vector Algebra

9.1. Dot/Standard Scalar Product

Definition 28.55 Scalar Projection

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} :

$$a_b = \|\mathbf{a}\| \cos \theta_{a,b} = \mathbf{a} \cdot \tilde{\mathbf{b}} \quad (28.80)$$



Definition 28.56 Standard Scalar/Dot Product:

Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ the standard scalar product is defined as:

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= \mathbf{u}^T \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^n u_i v_i = u_1 v_1 + \dots + u_n v_n \\ &= \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta = u_v \tilde{\mathbf{v}} = v_u \tilde{\mathbf{u}} \quad \theta \in [0, \pi] \end{aligned} \quad (28.81)$$

Explanation 28.3 (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 28.12 Orthogonal Direction

For $\theta \in [-\pi, \pi/2]$ rad $\cos \theta = 0$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = 0 \iff \mathbf{u} \perp \mathbf{v} \quad (28.82)$$

Note: Perpendicular

Perpendicular corresponds to orthogonality of two lines.

Property 28.13 Maximizing Direction:

For $\theta = 0$ rad $\cos \theta = 1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \quad (28.83)$$

Property 28.14 Minimizing Direction:

For $\theta = \pi$ rad $\cos \theta = -1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = -\|\mathbf{u}\| \|\mathbf{v}\| \quad (28.84)$$

Definition 28.57 Vector Projection:

General Projection via normal equation into inner product stuff i.e. with projection theorem

9.2. Cross Product

9.3. Outer Product

Definition 28.58 Outer Product: $\mathbf{u}\mathbf{v}^T = \mathbf{u} \otimes \mathbf{v}$: Given two vectors $\mathbf{u} \in \mathbb{K}^m$, $\mathbf{v} \in \mathbb{K}^n$ their outer product is defined as:

$$\begin{aligned} \mathbf{u} \otimes \mathbf{v} &= \mathbf{u} \mathbf{v}^H = \begin{bmatrix} u_1 & & u_m \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \\ &= \begin{bmatrix} u_1 \odot \bar{v}_1 & & u_m \odot \bar{v}_n \end{bmatrix} = \begin{bmatrix} u_1 \bar{v}_1 & u_1 \bar{v}_2 & u_1 \bar{v}_n \\ u_2 \bar{v}_1 & u_2 \bar{v}_2 & u_2 \bar{v}_n \\ u_m \bar{v}_1 & u_m \bar{v}_2 & u_m \bar{v}_n \end{bmatrix} \end{aligned} \quad (28.85)$$

Proposition 28.2

Rank of Outer Product: The outer product of two vectors is of rank one:

$$\text{rank}(\mathbf{u} \otimes \mathbf{v}) = 1 \quad (28.86)$$

9.4. Vector Norms

Definition 28.59 Norm $\|\cdot\|_{\mathcal{V}}$:

Let \mathcal{V} be a vector space over a field F , a norm on \mathcal{V} is a map: $\|\cdot\|_{\mathcal{V}} : \mathcal{V} \mapsto \mathbb{R}_+$ (28.87)

that satisfies:

$$\|\mathbf{x}\|_{\mathcal{V}} = 0 \iff \mathbf{x} = 0 \quad (\text{Definitness}) \quad (28.88)$$

$$\|\alpha \mathbf{x}\|_{\mathcal{V}} = |\alpha| \|\mathbf{x}\|_{\mathcal{V}} \quad (\text{Homogeneity}) \quad (28.89)$$

$$\|\mathbf{x} + \mathbf{y}\|_{\mathcal{V}} \leq \|\mathbf{x}\|_{\mathcal{V}} + \|\mathbf{y}\|_{\mathcal{V}} \quad (\text{Triangular Inequality}) \quad (28.90)$$

$$\alpha \in \mathbb{K} \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$$

Explanation 28.4 (Definition 28.59).

A norm is a measures of the size of its argument.

Corollary 28.24 Normed vector space: Is a vector space \mathcal{V} over a field F , on which a norm $\|\cdot\|_{\mathcal{V}}$ can be defined.

9.4.1. Cauchy Schwartz

Definition 28.60

Cauchy Schwartz Inequality: $|\mathbf{u}^T \mathbf{v}| \leq \|\mathbf{u}\| \|\mathbf{v}\|$ (28.91)

9.4.2. Triangular Inequality

Definition 28.61 **Triangular Inequality:** States that the length of the sum of two vectors is lower or equal to the sum of their individual lengths:

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \quad (28.92)$$

Corollary 28.25 Reverse Triangular Inequality:

$$\begin{aligned} -\|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}} &\leq \|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}} \leq \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}} \\ \text{resp. } \|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}} &\leq \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}} \end{aligned}$$

9.5. Distances

Definition 28.62

Distance Function/Measure $d : S \times S \mapsto \mathbb{R}_+$: Let S be a set, a distance functions is a mapping d that satisfies:

$$d(x, x) = 0 \quad (\text{Zero Identity Distance}) \quad (28.93)$$

$$d(x, y) = d(y, x) \quad (\text{Symmetry}) \quad (28.94)$$

$$d(x, z) \leq d(x, y) + d(y, z) \quad (\text{Triangular Identity}) \quad (28.95)$$

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

Explanation 28.5 (Definition 28.62).

Is measuring the distance between two things.

9.5.1. Contraction

Definition 28.63 Contraction: Given a metric space (M, d) is a mapping $f : M \mapsto M$ that satisfies:

$$d(f(x), f(y)) \leq \lambda d(x, y) \quad \lambda \in [0, 1] \quad (28.96)$$

Add metric spaces

9.6. Metrics

Definition 28.64 Metric

Is a distance measure^[def. 28.62] that additionally satisfies the **identity of indiscernibles**:

$$d(x, y) = 0 \iff x = y \quad \forall x, y \in S$$

Corollary 28.26 Metric \rightarrow Norm: Every norm $\|\cdot\|_{\mathcal{V}}$ on a vector space \mathcal{V} over a field F induces a metric by:

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

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

$$1. \text{ Homogeneity/Scaling: } d(\alpha x, \alpha y)_{\mathcal{V}} = |\alpha| d(x, y)_{\mathcal{V}}$$

$$2. \text{ Translational Invariance: } d(x + \alpha, y + \alpha) = d(x, y)$$

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 \mathbf{a} is similar to \mathbf{b} and \mathbf{b} is similar to \mathbf{c} it does not imply that \mathbf{a} is similar to \mathbf{c} .

Note

(bilinear form $\xrightarrow{\text{induces}}$)

inner product $\xrightarrow{\text{induces}}$ norm $\xrightarrow{\text{induces}}$ metric.

9.6.1. Metric Space

Definition 28.65 Metric Space (M, d) : A metric space is a pair (M, d) of a set M and a metric^[def. 28.64] d defined on M :

$$d : M \times M \mapsto \mathbb{R}_+ \quad (28.97)$$

10. Angles

Definition 28.66 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. 28.78] \mathcal{V} . The angle $\alpha \in [0, \pi]$ between \mathbf{u}, \mathbf{v} is defined by:

$$\angle(\mathbf{u}, \mathbf{v}) := \alpha \quad \cos \alpha = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \quad \mathbf{u}, \mathbf{v} \in \mathcal{V} \quad \alpha \in [0, \pi] \quad (28.98)$$

11. Orthogonality

Definition 28.67 Orthogonal Vectors: Let \mathcal{V} be an inner-product space^[def. 28.78]. A set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n\} \in \mathcal{V}$ is called *orthogonal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \quad \forall i \neq j \quad (28.99)$$

11.1. Orthonormality

Definition 28.68 Orthonormal Vectors: Let \mathcal{V} be an inner-product space^[def. 28.78]. A set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_n, \dots\} \in \mathcal{V}$ is called *orthonormal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \forall i, j \quad (28.100)$$

12. Special Kind of Vectors

12.1. Binary/Boolean Vectors

Definition 28.69

Binary/Boolean Vectors/Bit Maps \mathbb{B}^n : Are vectors that contain only zero or one values:

$$\mathbb{B}^n = \{0, 1\}^n \quad (28.101)$$

Definition 28.70

R-Sparse Boolean Vectors

Are boolean vectors that contain exact r one values:

$$\mathbb{B}_r^n = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^T \mathbf{x} = \sum_{i=1}^n x_i = r \right\} \quad \mathbb{B}_r^n; \quad (28.102)$$

12.2. Probabilistic Vectors

Definition 28.71 Probabilistic Vectors: Are vectors that represent probabilities and satisfy:

$$\left\{ \mathbf{x} \in [0, 1]^n : \sum_{i=1}^n x_i = 1 \right\} \quad (28.103)$$

13. Vector Spaces and Measures

13.1. Bilinear Forms

13.2. Quadratic Forms

13.2.1. Min/Max Value

Corollary 28.27 ^[proof 28.20]

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} :

$$\mathbf{v}_1 \in \arg \min_{\mathbf{x}^T \mathbf{x} = 1} \mathbf{x}^T \mathbf{A} \mathbf{x} \quad \mathbf{v}_1 \in \arg \max_{\mathbf{x}^T \mathbf{x} = 1} \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (28.104)$$

Note

$$(\mathbf{Q}^T \tilde{\mathbf{n}})^T \mathbf{Q}^T \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^T \mathbf{Q} \mathbf{Q}^T \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^T \tilde{\mathbf{n}} = 1$$

13.2.2. Skew Symmetric Matirx

Corollary 28.28

Quadratic Form of Skew Symmetric matrix: The quadratic form of a skew symmetric matrix^[def. 28.29] vanishes:

$$\alpha = \mathbf{x}^T \mathbf{A}_{\text{skew}} \mathbf{x} = (\mathbf{x}^T \mathbf{A}_{\text{skew}}^T \mathbf{x})^T = (\mathbf{x}^T \mathbf{A}_{\text{skew}} \mathbf{x})^T = -\alpha \quad (28.105)$$

Which can only hold iff $\alpha = 0$.

13.3. Inner Product – Generalization of the dot product

Definition 28.72 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:

$$\begin{aligned} a(\alpha u + \beta v, w) &= \alpha a(u, w) + \beta a(v, w) \\ a(u, \alpha v + \beta w) &= \alpha a(u, v) + \beta a(u, w) \end{aligned}$$

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

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

Definition 28.73 Symmetric bilinear form: A bilinear form a on \mathcal{V} is symmetric if and only if:

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

Definition 28.74 Positive (semi) definite bilinear form:

A symmetric bilinear form a on a vector space \mathcal{V} over a field F is **positive definite** if and only if:

$$a(u, u) > 0 \quad \forall u \in \mathcal{V} \setminus \{0\} \quad (28.106)$$

$$\text{And positive semidefinite} \iff \geq \quad (28.107)$$

Corollary 28.29 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**:

$$a(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}' = (\mathbf{A} \mathbf{x}')^T \mathbf{x}$$

Definition 28.75 Positive (semi) definite Matrix $>$:

A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive definite** if and only if:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \iff \mathbf{A} > 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\} \quad (28.108)$$

$$\text{And positive semidefinite} \iff \geq \quad (28.109)$$

Corollary 28.30

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*.

$$\forall \lambda_i \in \text{eigen}(\mathbf{A}) > 0 \quad (28.110)$$

$$\text{And positive semidefinite} \iff \geq \quad (28.111)$$

Note

Positive definite matrices are often assumed to be symmetric but that is not necessarily true.

Proof 28.2: ?? 28.2 (for real matrices): Let \mathbf{v} be an eigenvector of \mathbf{A} then it follows:

$$\begin{aligned} & \stackrel{?? 28.2}{0} < \mathbf{v}^T \mathbf{A} \mathbf{v} = \mathbf{v}^T \lambda \mathbf{v} = \|\mathbf{v}\| \lambda \end{aligned}$$

Corollary 28.31 Positive Definiteness and Determinant:

The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

Definition 28.76 Negative (semi) definite Matrix <:
A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **negative definite** if and only if:
 $\mathbf{x}^\top \mathbf{A} \mathbf{x} < 0 \iff \mathbf{A} < 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ (28.112)
And **negative semidefinite** $\iff \leq$ (28.113)

Theorem 28.3 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:

- $\mathbf{A} > 0 \iff \det(\mathbf{A}^{(k)}) > 0 \quad k = 1, \dots, n$ (28.114)
- $\mathbf{A} < 0 \iff (-1)^k \det(\mathbf{A}^{(k)}) > 0 \quad k = 1, \dots, n$ (28.115)
- \mathbf{A} is indefinite if the first $\det(\mathbf{A}^{(k)})$ 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 $\det(\mathbf{A}^{(k)})$ that breaks both patterns is 0.

14. Inner Products

Definition 28.77 Inner Product: Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on \mathcal{V} is a map:
 $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \mapsto F \subseteq \mathbb{K} \quad K = \mathbb{R} \text{ or } \mathbb{C}$ (28.116)
that satisfies:

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

Definition 28.78 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 vector space over a field F together with an **inner product** $\langle \cdot, \cdot \rangle_{\mathcal{V}}$.

Corollary 28.32 Inner product \rightarrow S.p.d. Bilinear Form:
Let \mathcal{V} be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on \mathcal{V} is a positive definite symmetric bilinear form on \mathcal{V} .

Example: scalar prodct

Let $a(u, v) = u^\top \mathbf{I} v$ 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 definition $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, whereas bilinear forms must not.

Corollary 28.33 Inner product induced norm
 $\langle \cdot, \cdot \rangle_{\mathcal{V}} \rightarrow \|\cdot\|_{\mathcal{V}}$: Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ induces a norm of the form:
 $\|\mathbf{x}\|_{\mathcal{V}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \quad \mathbf{x} \in \mathcal{V}$
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 28.34 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{V} \times \mathcal{V} \mapsto F$ induces an energy norm:
 $\|\mathbf{x}\|_a := (a(\mathbf{x}, \mathbf{x}))^{\frac{1}{2}} = \sqrt{a(\mathbf{x}, \mathbf{x})} \quad \mathbf{x} \in \mathcal{V}$

15. Matrix Algebra

16. Matrix Norms

16.1. Operator Norm

Definition 28.79 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:
 $\|\mathbf{A}\|_{\mu, \nu} := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_{\mu}}{\|\mathbf{x}\|_{\nu}} = \sup_{\|\mathbf{x}\|_{\nu}=1} \|\mathbf{A}\mathbf{x}\|_{\mu} \quad \|\cdot\|_{\mu} : \mathbb{K}^m \mapsto \mathbb{R}$ (28.117)

Explanation 28.6 (Definition 28.79). *Is a measure for the largest factor by which a matrix \mathbf{A} can stretch a vector $\mathbf{x} \in \mathbb{R}^n$.*

16.2. Induced Norms

Corollary 28.35 Induced Norms: Let $\|\cdot\|_p : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ defined as:

$$\|\mathbf{A}\|_p := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{y}\|_p=1} \|\mathbf{A}\mathbf{y}\|_p \quad (28.118)$$

Explanation 28.7 ([Corollary 28.35]).
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 $\{\mathbf{y} \mid \|\mathbf{y}\| = 1\}$ is compact, thus if we consider finite matrices the supremum is attained and we may replace it by the max.

16.3. Induced Norms

16.3.1. 1-Norm

Definition 28.80 Column Sum Norm $\|\mathbf{A}\|_1$:
 $\|\mathbf{A}\|_1 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$ (28.119)

16.3.2. ∞ -Norm

Definition 28.81 Row Sum Norm $\|\mathbf{A}\|_{\infty}$:
 $\|\mathbf{A}\|_{\infty} = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_{\infty}}{\|\mathbf{x}\|_{\infty}} = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$ (28.120)

16.3.3. Spectral Norm

Spectral Radius & Singular Value

Definition 28.82 Spectral Radius $\rho(\mathbf{A})$:
The spectral radius is defined as the largest eigenvalue of a matrix:
 $\rho(\mathbf{A}) = \max \{|\lambda| \mid \lambda \in \text{eigenval}(\mathbf{A})\}$ (28.121)

Definition 28.83 Singular Value σ_i :
Given a matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ its n real and positive singular values are defined as:
 $\sigma(\mathbf{A}) := \left\{ \left\{ \sqrt{\lambda_i} \right\}_{i=1}^n \mid \lambda_i \in \text{eigenval}(\mathbf{A}^\top \mathbf{A}) \right\}$ (28.122)

Spectral Norm

Definition 28.84 L2/Spectral Norm $\|\mathbf{A}\|_2$:
 $\|\mathbf{A}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x}}$ (28.123)
 $= \max_{\|\mathbf{x}\|_2=1} \sqrt{\rho(\mathbf{A}^\top \mathbf{A})} =: \sigma_{\max}(\mathbf{A})$ (28.124)

16.4. Energy Norm

16.5. Forbenius Norm

Definition 28.85 Forbenius Norm $\|\mathbf{A}\|_F$:
The *Forbenius norm* $\|\cdot\|_F : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ is defined as:
 $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}^2|} = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^\text{H})}$ (28.125)

16.6. Distance

17. Decompositions

17.1. Eigen/Spectral decomposition

Definition 28.86 $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$, [proof 28.25]
Eigendecomposition/ Spectral Decomposition :
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a *diagonalizable* square matrix and define by $\mathbf{X} = [\mathbf{x}_1 \quad \dots \quad \mathbf{x}_n] \in \mathbb{R}^{n \times n}$ a non-singular matrix whose column vectors are the eigenvectors of \mathbf{A} with associated eigenvalue matrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$. Then \mathbf{A} can be represented as:
 $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ (28.126)

Proposition 28.3 Diagonalization: If non of \mathbf{A} eigenvalues are zero it can be diagonalized:
 $\mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \mathbf{\Lambda}$ (28.127)

Proposition 28.4 Existence:
 $\exists \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \iff \mathbf{A}$ diagonalizable (28.128)

17.2. QR-Decompositions

17.3. Singular Value Decomposition

Definition 28.87
Singular Value Decomposition (SVD) $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\text{H}$:
For any matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ there exist unitary matrices^[def. 28.27]
 $\mathbf{U} \in \mathbb{K}^{m,m} \quad \mathbf{V} \in \mathbb{K}^{n,n}$
and a (generalized) digonal matrix:
 $\mathbf{\Sigma} \in \mathbb{R}^{m,n} \quad p := \min\{m, n\}$
 $\mathbf{\Sigma} = \text{gendia}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m,n}$
such that:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\text{H} \quad (28.129)$$

$$= \left(\begin{array}{c|c|c|c} \text{u}_1 & \text{u}_r & \text{u}_{r+1} & \text{u}_m \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline \text{u}_1 & \text{u}_r & \text{u}_{r+1} & \text{u}_m \\ \hline \end{array} \right) \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r & \\ & & & 0 & \\ & & & & \ddots & \\ & & & & & & 0 \end{array} \right) \left(\begin{array}{c} \text{v}_1 \\ \text{v}_2 \\ \text{v}_3 \\ \text{v}_4 \\ \text{v}_5 \\ \text{v}_6 \end{array} \right)$$

(Image full, economical, alternative representation, range and kernel
https://math.ubt.uni-bonn.de/~math/lehre/ss19/lineal/288_SVD_Separ.pdf

17.3.1. Eigenvalues

Proposition 28.5 [proof 28.23]:
The eigenvalues of a matrix $\mathbf{A}^\top \mathbf{A}$ are positive.

Proposition 28.6 [proof 28.24]
Similarity Transformation: The unitary matrix \mathbf{V} provides a *similarity transformation*^[cor. 28.14] of $\mathbf{A}^\top \mathbf{A}$ into a diagonal matrix $\mathbf{\Sigma}^\top \mathbf{\Sigma}$:
 $\mathbf{\Sigma}^\top \mathbf{\Sigma} \mapsto \mathbf{V}^\text{H} \mathbf{A}^\top \mathbf{A} \mathbf{V}$ (28.130)

Corollary 28.36 eigenval($\mathbf{A}^\top \mathbf{A}$) = eigenval($\mathbf{\Sigma}^\top \mathbf{\Sigma}$):
From proposition 28.6 and ^[cor. 28.15] it follows that:
 $\text{eigenval}(\mathbf{A}^\top \mathbf{A}) = \text{eigenval}(\mathbf{\Sigma}^\top \mathbf{\Sigma})$ (28.131)
 $\implies \|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^\top \mathbf{A})} = \sqrt{\lambda_{\max}} = \sigma_{\max}$

Note

λ and *singularvalue* corresponds to the eigenvalues/singular-values of $\mathbf{A}^\top \mathbf{A}$ and not \mathbf{A}

17.3.2. Best Lower Rank Approximation...

Theorem 28.4 Eckart Young Theorem: Given a matrix $\mathbf{X} \in \mathbb{K}^{m,n}$ the *reduced* SVD \mathbf{X} defined as:

$$\mathbf{X}_k := \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\text{H} \quad \mathbf{U}_k := [\mathbf{u}_{:,1} \quad \dots \quad \mathbf{u}_{:,k}] \in \mathbb{K}^{m,k} \\ \mathbf{\Sigma}_k = \text{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k,k} \\ k \leq \min\{m, n\} \\ \mathbf{V}_k = [\mathbf{v}_{:,1} \quad \dots \quad \mathbf{v}_{:,k}] \in \mathbb{K}^{n,k}$$

provides the best lower k rank approximation of \mathbf{X} :
 $\min_{\mathbf{Y} \in \mathbb{K}^{n,m} : \text{rank}(\mathbf{Y}) \leq k} \|\mathbf{X} - \mathbf{Y}\|_F = \|\mathbf{X} - \mathbf{X}_k\|_F$ (28.132)

18. Matric Calculus

18.1. Derivatives

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{b}) = \mathbf{b} \\ \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A} \mathbf{x} = \mathbf{A} \quad (28.133)$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x} \quad (28.134)$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{A} \mathbf{x}) = \mathbf{A}^\top \mathbf{b} \quad \frac{\partial}{\partial \mathbf{X}} (\mathbf{c}^\top \mathbf{X} \mathbf{b}) = \mathbf{c} \mathbf{b}^\top \quad \frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|_2} \\ \frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x} \quad \frac{\partial}{\partial \mathbf{X}} (\|\mathbf{X}\|_F^2) = 2\mathbf{X} \\ \frac{\partial}{\partial \mathbf{x}} \|\mathbf{x}\|_1 = \frac{\mathbf{x}}{|\mathbf{x}|} \\ \frac{\partial}{\partial \mathbf{x}} (\|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2^2) = 2(\mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b}) \quad \frac{\partial}{\partial \mathbf{X}} (|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1} \\ \frac{\partial}{\partial \mathbf{x}} (\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1} \frac{\partial \mathbf{Y}}{\partial \mathbf{x}} \mathbf{Y}^{-1}$$

19. Proofs

Proof 28.3: ^[def. 28.36]
 $\mathbf{M}\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{I}_{k,k} & \mathbf{0}_{k,l} \\ \mathbf{0}_{l,k} & \mathbf{I}_{l,l} \end{bmatrix}$ (28.135)

19.1. Vector Algebra

Proof 28.4 Definition 28.56:
eq. (29.19)
(1): $\|a - b\| \stackrel{\text{eq. (29.19)}}{=} \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \cos \theta$
(2): $\|a - b\| = (a - b)(a - b) = \|a\|^2 + \|b\|^2 - 2(ab)$
 $\|a - b\| = \|a - b\| \implies ab = \|a\|\|b\| \cos \theta$

Proof 28.5 Proposition 28.2: 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

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u} \mathbf{v}^\text{H} = \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 \\ \vdots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix}$$

19.2. Mappings

Proof 28.6: Corollary 28.20
 $\|\mathbf{Q}\mathbf{x}\|^2 = (\mathbf{Q}\mathbf{x})^\top \mathbf{Q}\mathbf{x} = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q}\mathbf{x} = \mathbf{x}^\top \mathbf{x} = \|\mathbf{x}\|^2$

Proof 28.7: Corollary 28.21 Follows immediately from definition 28.66 in combination with eqs. (28.65) and (28.67).

Proof 28.8: Proposition 28.1:
 $\implies l(\alpha \mathbf{x} + \beta \mathbf{y}) \stackrel{28.58}{=} l(\alpha \mathbf{x}) + l(\beta \mathbf{y}) \stackrel{28.59}{=} \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$
 $\leftarrow \begin{array}{l} l(\alpha \mathbf{x} + \mathbf{0}) = \alpha l(\mathbf{x}) \\ l(\mathbf{1x} + \mathbf{1y}) = l(\mathbf{x}) + l(\mathbf{y}) \end{array}$

Proof 28.9 principle 28.1:
Every vector $\mathbf{v} \in \mathcal{V}$ can be represented by a basis eq. (28.16) of \mathcal{V} . With *homogeneity*eq. (28.59) and *additivity*eq. (28.58) it follows for the image of all $\mathbf{v} \in \mathcal{V}$:
 $l(\mathbf{v}) = l(\alpha_1 b_1 + \dots + \alpha_n b_n) = l\alpha_1 (b_1) + \dots + l(\alpha_n b_n)$ (28.136)
 \implies the image of the basis of \mathcal{V} determines the linear mapping.

Proof 28.10 Proof [Corollary 28.17]:
 $\implies l_{\mathbf{A}}(\alpha \mathbf{x} + \mathbf{y}) = \mathbf{A}(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{A} \mathbf{y} = \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$
 \longleftarrow Let \mathfrak{B} be a standard normal basis of \mathcal{V} with eq. (28.136):

$$l(\mathbf{x}) = \sum_{i=1}^n x_i l(\mathbf{e}_i) = \sum_{i=1}^n x_i \mathbf{A}_{:,i} = \mathbf{A} \mathbf{x} \quad \mathbf{A}_{:,i} := l(\mathbf{e}_i) \in \mathbb{R}^n$$

Proof 28.11 Proof Property 28.11:
 $(g \circ f)(\alpha \mathbf{x}) = g(f(\alpha \mathbf{x})) = g(\alpha f(\mathbf{x})) = \alpha (g \circ f)(\mathbf{x})$
 $(g \circ f)(\mathbf{x} + \mathbf{y}) = g(f(\mathbf{x} + \mathbf{y})) = g(f(\mathbf{x}) + f(\mathbf{y}))$
 $= (g \circ f)(\mathbf{x}) + (g \circ f)(\mathbf{y})$
or even simpler as every linear form can be represented by a matrix product:
 $f(y) = \mathbf{A} \mathbf{y} \quad g(z) = \mathbf{B} \mathbf{z} \quad \implies \quad (f \circ g)(\mathbf{x}) = \mathbf{A} \mathbf{B} \mathbf{x} := \mathbf{C} \mathbf{x}$

Proof 28.12: [Corollary 28.22] Let $\mathbf{y} \in \mathcal{N}(\mathbf{A})$ ($\mathbf{z} \in \mathcal{N}(\mathbf{A}^\top)$) then it follows:
 $\mathcal{N}(\mathbf{A}) \perp \mathfrak{R}(\mathbf{A}^\top) \quad (\mathbf{A}^\top \mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top \mathbf{A} \mathbf{y} = \mathbf{x}^\top \mathbf{0} = 0$
 $\mathcal{N}(\mathbf{A}^\top) \perp \mathfrak{R}(\mathbf{A}) \quad (\mathbf{A} \mathbf{x})^\top \mathbf{z} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{z} = \mathbf{x}^\top \mathbf{0} = 0$

19.3. Special Matrices

Proof 28.13 [Corollary 28.15]: 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 28.14 Property 28.6:
 $\|\mathbf{Q} \mathbf{x}\|_2^2 = (\mathbf{Q} \mathbf{x})^\top \mathbf{Q} \mathbf{x} = \mathbf{x}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{x} = \|\mathbf{x}\|_2^2$

Proof 28.15: Property 28.4
Let $\mathbf{A} \in \mathbb{K}^{n \times n}$ be a hermitian matrix^[def. 28.26] and let $\lambda \in \mathbb{K}$ be an eigenvalue of \mathbf{A} with corresponding eigenvector $\mathbf{v} \in \mathbb{K}^n$:
 $\lambda(\bar{\mathbf{v}}^\top \mathbf{v}) = \bar{\mathbf{v}}^\top \lambda \mathbf{v} = \bar{\mathbf{v}}^\top \mathbf{A} \mathbf{v} = \overline{(\bar{\mathbf{v}}^\top \mathbf{A} \mathbf{v})} = \bar{\mathbf{A}} \mathbf{v}^\top \mathbf{v} = \bar{\lambda}(\bar{\mathbf{v}}^\top \mathbf{v})$
 $\lambda(\bar{\mathbf{v}}^\top \mathbf{v}) = \bar{\lambda}(\bar{\mathbf{v}}^\top \mathbf{v})$
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 (19.8))

Proof 28.16: ??
old

19.4. Vector Spaces

Proof 28.17 Definition 28.21: We know that $\text{proj}_L(\mathbf{u})$ must be a vector times a certain magnitude:
 $\text{proj}_L(\mathbf{u}) = \alpha \tilde{\mathbf{v}} \quad \alpha \in \mathbb{K} \quad (28.137)$
the magnitude follows from the scalar projection^[def. 28.55] in the direction of \mathbf{v} which concludes the derivation.

Proof 28.18 Definition 28.21 (via orthogonality): We know that $\mathbf{u} - \text{proj}_L(\mathbf{u})$ must be orthogonal^[def. 28.67] to \mathbf{v}
 $(\mathbf{u} - \text{proj}_L(\mathbf{u}))^\top \mathbf{v} = (\mathbf{u} - \alpha \mathbf{v})^\top \mathbf{v} = 0 \implies \alpha = \frac{\mathbf{u}^\top \mathbf{v}}{\mathbf{v}^\top \mathbf{v}}$

Proof 28.19: Definition 28.22 Let $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ a basis of \mathcal{U} s.t. by ^[cor. 28.4]:

$$\mathbf{u} = \sum_{i=1}^n \alpha_i \mathbf{b}_i$$

the coefficients $\{\alpha_i\}_{i=1}^n$ need to be determined. We know that:
 $\mathbf{v} - \mathbf{u} \perp \mathbf{b}_1, \dots, \mathbf{v} - \mathbf{u} \perp \mathbf{b}_n$

$$\implies \left(\mathbf{v} - \sum_{i=1}^n \alpha_i \mathbf{b}_i \right) \cdot \mathbf{b}_j = 0 \quad j = 1, \dots, n$$

this linear system of equations can be rewritten as:

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_n \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_n \end{pmatrix} \mathbf{v}$$

...
. . .

Proof 28.20: Corollary 28.27
Let $\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$ be the eigendecomposition^[cor. 28.12] of \mathbf{A} then it follows:

$$\begin{aligned} \min_{\tilde{\mathbf{n}}^\top \tilde{\mathbf{n}}=1} \tilde{\mathbf{n}}^\top \mathbf{A} \tilde{\mathbf{n}} &= \min_{\|\tilde{\mathbf{n}}\|=1} \tilde{\mathbf{n}}^\top (\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top) \tilde{\mathbf{n}} \\ &= \min_{\|\tilde{\mathbf{n}}\|=1} (\mathbf{Q}^\top \tilde{\mathbf{n}})^\top \mathbf{\Lambda} (\mathbf{Q}^\top \tilde{\mathbf{n}}) \\ &= \min_{\mathbf{x}^\top \mathbf{x}=1} \mathbf{x}^\top \mathbf{\Lambda} \mathbf{x} \quad \mathbf{x} := \mathbf{Q}^\top \tilde{\mathbf{n}} \\ &= \min_{\mathbf{x}^\top \mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2 \mathbf{\Lambda}_{ii} = \min_{\mathbf{x}^\top \mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2 \lambda_i \end{aligned}$$

Thus in order to obtain the minimum value we need to choose the eigenvector that leads to the smallest eigenvalue.

19.5. Norms

Proof 28.21: ?? 28.21
 $|\mathbf{u} \cdot \mathbf{v}| \stackrel{\text{eq. (28.81)}}{=} \|\mathbf{u}\| \|\mathbf{v}\| |\cos \theta| \leq \|\mathbf{u}\| \|\mathbf{v}\|$

Proof 28.22: Definition 28.61
 $\|\mathbf{u} + \mathbf{v}\|^2 = (\mathbf{u} + \mathbf{v})(\mathbf{u} + \mathbf{v}) = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\mathbf{u} \cdot \mathbf{v})$
from cauchy schwartz we know:
 $\mathbf{u} \cdot \mathbf{v} \leq |\mathbf{u} \cdot \mathbf{v}| \stackrel{\text{eq. (28.91)}}{\leq} \|\mathbf{u}\| \|\mathbf{v}\|$
 $\|\mathbf{u} + \mathbf{v}\|^2 \leq \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\|\mathbf{u}\| \|\mathbf{v}\|) = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2$

19.6. Decompositions
19.6.1. Symmetric - Antisemitic

Definition 28.88 Symmetric - Antisymmetric Decomposition: Any matrix $\mathbf{A} \in \mathbb{K}^{n \times n}$ can be decomposed into the sum of a symmetric matrix^[def. 28.23] \mathbf{A}^{sym} and a skew-symmetric matrix^{??} \mathbf{A}^{skes} :

$$\mathbf{A} = \mathbf{A}^{\text{sym}} + \mathbf{A}^{\text{skew}} \quad \begin{aligned} \mathbf{A}^{\text{sym}} &= \frac{1}{2} \left(\mathbf{A} + \mathbf{A}^{\text{H}} \right) \\ \mathbf{A}^{\text{skew}} &= \frac{1}{2} \left(\mathbf{A} - \mathbf{A}^{\text{H}} \right) \end{aligned} \quad (28.138)$$

19.6.2. SVD

Proof 28.23 [Corollary 28.5]: $\mathbf{B} := \mathbf{A}^\top \mathbf{A}$ corresponds to a symmetric positive definite form^[def. 28.75]:
 $\mathbf{x}^\top \mathbf{B} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} = \|\mathbf{A} \mathbf{x}\|_2^2 > 0$
thus Proposition 28.6 follows immediately form [Corollary 28.2].

Proof 28.24 Proposition 28.6:

$$\mathbf{A}^\top \mathbf{A} \stackrel{\text{SVD}}{=} \left(\mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\text{H}} \right)^{\text{H}} \mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \mathbf{V} \mathbf{\Sigma}^{\text{H}} \underbrace{\mathbf{U}^{\text{H}} \mathbf{U}}_{\mathbf{I}_m} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \mathbf{V} \mathbf{\Sigma}^{\text{H}} \mathbf{\Sigma} \mathbf{V}^{\text{H}}$$

$$\implies \mathbf{V}^{\text{H}} \mathbf{A}^\top \mathbf{A} \mathbf{V} = \mathbf{\Sigma}^\top \mathbf{\Sigma}$$

19.6.3. Eigendecomposition

Proof 28.25 Definition 28.86:

$$\mathbf{A} \mathbf{X} = \left[\lambda_1 \mathbf{x}_1 \quad \lambda_n \mathbf{x}_n \right] = \mathbf{X} \mathbf{\Lambda}$$

.....

Geometry

Corollary 29.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The **affine transformation** of $\phi : \hat{\Omega} \rightarrow \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d - c}{b - a} (x - a) + c \tag{29.1}$$

Proof 29.1: [cor. 29.1] By [def. 28.45] we want a function $f : [a, b] \rightarrow [c, d]$ that satisfies:

$$f(a) = c \qquad \text{and} \qquad f(b) = d$$

additionally $f(x)$ has to be a linear function ([def. 23.15]), that is the output scales the same way as the input scales.

Thus it follows:

$$\frac{d - c}{b - a} = \frac{f(x) - f(a)}{x - a} \qquad \Longleftrightarrow \qquad f(x) = \frac{d - c}{b - a} (x - a) + c$$

Trigonometry

0.1. Trigonometric Functions

0.1.1. Sine

Definition 29.1 Sine:

$$\sin \alpha = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{a}{c} \tag{29.2}$$

0.1.2. Cosine

Definition 29.2 Cosine:

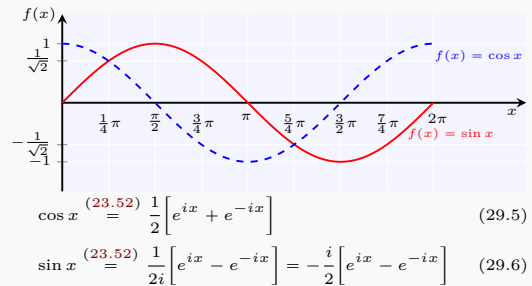
$$\cos \alpha = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{b}{c} \tag{29.3}$$

0.1.3. Tangens

Definition 29.3 Tangens:

$$\tan \alpha = \frac{\text{opposite}}{\text{adjacent}} = \frac{a}{b} = \frac{a/c}{b/c} = \frac{\sin \alpha}{\cos \alpha} \tag{29.4}$$

0.1.4. Trigonometric Functions and the Unit Circle
Sine and Cosine



Note

Using theorem 29.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos \alpha \quad \text{and} \quad \sin(\alpha \pm \pi) = -\sin \alpha \tag{29.7}$$

0.1.5. Sinh

Definition 29.4 Sinh:

$$\sinh x \stackrel{(eq. (23.52))}{=} \frac{1}{2} \left[e^x - e^{-x} \right] = -i \sin(ix) \tag{29.8}$$

Property 29.1: $\sinh x = 0$ has a unique root at $x = 0$.

0.1.6. Cosh

Definition 29.5 Cosh:

$$\cosh x \stackrel{(23.52)}{=} \frac{1}{2} \left[e^x + e^{-x} \right] = \cos(ix) \tag{29.9}$$

$$\tag{29.10}$$

Property 29.2: $\cosh x$ is strictly positive.

Proof 29.2:

$$e^x = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{29.11}$$

0.2. Addition Theorems

Theorem 29.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin \alpha \cos \beta \pm \cos \alpha \sin \beta \tag{29.12}$$

$$\cos(\alpha \pm \beta) = \cos \alpha \cos \beta \mp \sin \alpha \sin \beta \tag{29.13}$$

0.3. Werner Formulas

Werner Formulas

$$\sin \alpha \cos \beta = \frac{1}{2} \left[\sin(\alpha + \beta) + \sin(\alpha - \beta) \right] \tag{29.14}$$

$$\sin \alpha \sin \beta = \frac{1}{2} \left[\cos(\alpha - \beta) - \cos(\alpha + \beta) \right] \tag{29.15}$$

$$\cos \alpha \cos \beta = \frac{1}{2} \left[\cos(\alpha + \beta) + \cos(\alpha - \beta) \right] \tag{29.16}$$

Note

Using theorem 29.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos \alpha \quad \text{and} \quad \sin(\alpha \pm \pi) = -\sin \alpha \tag{29.17}$$

0.4. Law of Cosines

Law 29.1 Law of Cosines

[proof 29.3]:

relates the three side of a *general* triangle to each other.

$$a^2 = b^2 + c^2 - 2bc \cos \theta_{b,c} \tag{29.18}$$

Law 29.2 Law of Cosines for Vectors

[proof 29.4]:

relates the length of vectors to each other.

$$\|\mathbf{a}\|^2 = \|\mathbf{c} - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2 - 2\|\mathbf{b}\|\|\mathbf{c}\|\cos \theta_{\mathbf{b},\mathbf{c}} \tag{29.19}$$

Law 29.3 Pythagorean theorem: special case of ?? for right triangle:

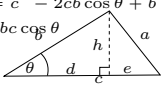
$$a^2 = b^2 + c^2 \tag{29.20}$$

1. Proofs

Proof 29.3: Law 29.1 From the definition of the sine and cosine we know that:

$$\sin \theta = \frac{h}{b} \Rightarrow \underline{h} \qquad \text{and} \qquad \cos \theta = \frac{d}{b} \Rightarrow \underline{d}$$

$$\begin{aligned} \frac{e}{a^2} &= c - \underline{d} = c - b \cos \theta \\ a^2 &= \frac{e}{c^2} + \frac{h}{b^2} = c^2 - 2cb \cos \theta + b^2 \cos^2 \theta + b^2 \sin^2 \theta \\ &= c^2 + b^2 - 2bc \cos \theta \end{aligned}$$



Proof 29.4: Law 29.2 Notice that $\mathbf{c} = \mathbf{a} + \mathbf{b} \Rightarrow \mathbf{a} = \mathbf{c} - \mathbf{b}$ and we can either use ?? 29.3 or notice that:

$$\begin{aligned} \|\mathbf{c} - \mathbf{b}\|^2 &= (\mathbf{c} - \mathbf{b}) \cdot (\mathbf{c} - \mathbf{b}) \\ &= \mathbf{c} \cdot \mathbf{c} - 2\mathbf{c} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} \\ &= \|\mathbf{c}\|^2 + \|\mathbf{b}\|^2 - 2(\|\mathbf{c}\|\|\mathbf{b}\|\cos \theta) \end{aligned}$$

Topology

Definition 30.1 Topology of set

τ :

Let X be a set. A collection τ of open?? subsets of X is called *topology* of X if it satisfies:

- $\emptyset \in \tau$ and $X \in \tau$
- Any finite or infinite union of subsets of τ is contained in τ :

$$\{U_i : i \in \mathbf{I}\} \subseteq \tau \qquad \Longrightarrow \qquad \cup_{i \in \mathbf{I}} U_i \in \tau \tag{30.1}$$

- The intersection of a finite number of elements of τ also belongs to τ :

$$\{U_i\}_{i=1}^n \in \tau \qquad \Longrightarrow \qquad U_1 \cap \dots \cap U_n \in \tau \tag{30.2}$$

Definition 30.2 Topological Space[?]

(X, τ) :

Is an ordered pair (X, τ) , where X is a set and τ is a topology[def. 30.1] on X .

Numerical Methods

Machine Arithmetic's

Machine/Floating Point Numbers

Definition 31.1 (IEEE)
Institute of Electrical and Electronics Engineers:
Is a engineering associations that defines a standard on how computers should treat machine numbers in order to have certain guarantees.

Definition 31.2 Machine/Floating Point Numbers \mathbb{M} :
Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{M} \subset \mathbb{R}$

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

Corollary 31.1 Closure:
Machine numbers \mathbb{F} are not *closed*^[def. 19.7] under basic arithmetic operations:

$$\mathbb{F} \Omega \mathbb{F} \mapsto \not\mathbb{F} \quad \Omega = \{+, -, *, /\} \quad (31.1)$$

Note

Corollary 31.1 provides a problem as the computer can only represent floating point number \mathbb{F} .

Definition 31.3 Overflow: Result is bigger then the biggest representable floating point number.

Definition 31.4 Underflow: Result is smaller then the smaller representable floating point number i.e. to close to zero.

1.1.2. The Rounding Unit

Definition 31.5 **Rounding Function/Unit** rd/\sim :
Let $x \in \mathbb{K}$ be a number real or complex number. The rounding function approximates x by the nearest machine number $\tilde{x} \in \mathbb{F}$:

$$\text{rd} : \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg \min_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases} \quad (31.2)$$

Notes

- If this is ambiguous (there are two possibilities), then it takes the larger one:
- Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

Definition 31.6 Floating Point Operation $\tilde{\Omega}$:
Is a basic arithmetic operation between two floating point numbers $x \in \mathbb{F}$ rounded back to the nearest floating point number:

$$\mathbb{F} \tilde{\Omega} \mathbb{F} \mapsto \mathbb{F} \quad \tilde{\Omega} := \text{rd} \circ \Omega \quad \Omega = \{+, -, *, /\} \quad (31.3)$$

Definition 31.7 Absolute Error: Let $\tilde{x} \in \mathbb{K}$ be an approximation of $x \in \mathbb{K}$ then the absolute error is defined by:

$$\epsilon_{\text{abs}} := |x - \tilde{x}| \quad (31.4)$$

Definition 31.8 Relative Error: Let $\tilde{x} \in \mathbb{K}$ be an approximation of $x \in \mathbb{K}$ then the relative error is defined by:

$$\epsilon_{\text{abs}} := \frac{|x - \tilde{x}|}{|x|} \quad (31.5)$$

Note

We are interested in the relative error as it controls the number of *correct/significant* digits l of the approximation \tilde{x} of $x \in \mathbb{K}$:

$$\epsilon_{\text{abs}} := \frac{|x - \tilde{x}|}{|x|} \leq 10^{-l} \quad l \in \mathbb{N}_{>0} \quad (31.6)$$

1.1.3. The Machine Epsilon

Definition 31.9 **The Machine Epsilon:** **EPS**
The machine epsilon EPS is the largest possible *relative* rounding error^[def. 31.8]:

$$\text{EPS} := \max_{x \in I \setminus 0} \frac{|\text{rd}(x) - x|}{|x|} \quad I := [\min|\mathbb{M}|, \max|\mathbb{M}|] \in \mathbb{K} \quad (31.7)$$

Corollary 31.2 Relative Error of Flop:
The *relative* error^[def. 31.8] of any floating point operation^[def. 31.6] is bounded by the machine epsilon^[def. 31.9]:

$$\begin{aligned} \text{EPS}_{\text{rel}}(\tilde{\Omega}(x, y)) &:= \frac{|\tilde{\Omega}(x, y) - \Omega(x, y)|}{|\Omega(x, y)|} \\ &= \frac{|(\text{rd} - \mathbf{I})\Omega(x, y)|}{|\Omega(x, y)|} \leq \text{EPS} \end{aligned} \quad (31.8)$$

Corollary 31.3 EPS for Machine Number: For machine numbers EPS can be computed by:

$$\text{EPS} = \frac{1}{2} B^{1-m} \quad (31.9)$$

Type	EPS
double	$2.2 \cdot 10^{-16}$
float	$1.1 \cdot 10^{-23}$
FP16	$9.76 \cdot 10^{-4}$

Axiom of Round off Analysis

Axiom 31.1 Axiom of Round off Analysis:
Let $x, y \in \mathbb{F}$ be (normalized) floats and assume that $x\tilde{\Omega}y \in \mathbb{F}$ (i.e. no over/underflow). Then it holds that:

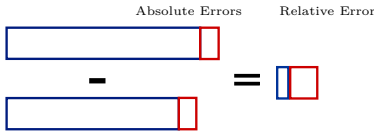
$$\begin{aligned} x\tilde{\Omega}y &= (x\Omega y)(1 + \delta) \quad \Omega = \{+, -, *, /\} \\ \tilde{f}(x) &= f(x)(1 + \delta) \quad f \in \{\exp, \sin, \cos, \log, \dots\} \end{aligned} \quad (31.10)$$

with $|\delta| < \text{EPS}$

Explanation 31.1 (axiom 31.1). *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 $x\Omega y$.*

1.1.4. Cancellation

Definition 31.10 Cancellation:
Is the extreme amplification of *relative* errors^[def. 31.8] when subtracting numbers of almost equal size.



Roundoff Errors

2.0.1. Tricks

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. 31.11] 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 31.11 Log sum Exponential:

$$\text{LogSumExp}(x_1, \dots, x_n) := \log \left(\sum_{i=1}^n e^{x_i} \right) \quad (31.11)$$

Formula 31.1 [proof 31.3]

Log-Sum-Exp Trick:

$$\log \left(\sum_{i=1}^n e^{x_i} \right) = a + \log \sum_{i=1}^n e^{x_i - a} \quad a := \max_{i \in \{1, \dots, n\}} x_i \quad (31.12)$$

Explanation 31.2 (formula 31.1). *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.

Definition 31.12 Partition Π :

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

Asymptotic Complexity

3.1. O-Notation

3.1.1. Small $o(\cdot)$ Notation

Definition 31.13 Little o Notation:

$$f(n) = o(g(n)) \iff \lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0 \quad (31.13)$$

3.1.2. Big $\mathcal{O}(\cdot)$ Notation

3.2. Basic Operations

4. Rate Of Convergence

Definition 31.14 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:

$$\lim_{k \mapsto \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^k - \mathbf{x}^* \right\|} = \rho \tag{31.14}$$
$$\iff \lim_{k \mapsto \infty} \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \rho \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \quad \forall k \in \mathbb{N}_0$$

Definition 31.15 Linear/Exponential Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *linearly* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho \in (0, 1) \qquad \qquad \qquad \forall k \in \mathbb{N}_0 \tag{31.15}$$

Definition 31.16 Superlinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *superlinear* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho = 1 \tag{31.16}$$

Definition 31.17 Sublinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *sublinear* to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ if it satisfies:

$$\rho = 0 \quad \iff \quad \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| = o \left(\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \right) \tag{31.17}$$

Definition 31.18 Logarithmic Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *logarithmically* to \mathbf{x}^* if it converges *sublinear*^[def. 31.17] and additoinally satisfies

$$\rho = 0 \quad \iff \quad \left\| \mathbf{x}^{k+2} - \mathbf{x}^{k+1} \right\| = o \left(\left\| \mathbf{x}^{k+1} - \mathbf{x}^k \right\| \right) \tag{31.18}$$

add explanation 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:
- $$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \underbrace{(1 - \alpha)}_{:= \rho} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|$$
2. and that $(1 - \alpha) = \exp(-\alpha)$ from which follows that:
- eq. (31.19)
$$\iff \left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq e^{-\alpha} \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|$$

Definition 31.19 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 $p \in \{2, \dots\}$ to \mathbf{x}^* if it satisfies:

$$\lim_{k \mapsto \infty} \frac{\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\|}{\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\|^p} = C \qquad C < 1 \tag{31.19}$$

Does this even exist/check if this is true

Definition 31.20 Exponential Convergence: A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate ρ to \mathbf{x}^* if in the asymptotic limit $k \rightarrow \infty$ it satisfies:

$$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \leq \rho^k \left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \qquad \rho < 1 \tag{31.20}$$

$$\left\| \mathbf{x}^{k+1} - \mathbf{x}^* \right\| \in o \left(\left\| \mathbf{x}^{(k)} - \mathbf{x}^* \right\| \right) \tag{31.21}$$

5. Basic Operations

Operation	#mul/div	#add/sub	asyp. comp
Dot Prod.	n	$n - 1$	$\mathcal{O}(n)$
Tensor Prod.	nm	0	$\mathcal{O}(nm)$
Matrix Prod.	mnk	$mk(n - 1)$	$\mathcal{O}(nmk)$

6.1. Direct Methods

6.1.1. Gaussian Elimination

Definition 31.22 Row Echelon Matrix: Is a rectangular matrix where:

- All non-zero rows are above any zero rows.
- Each pivot of a row has a larger column index than the pivot of the row above.
- All entries below a pivot are zero.

Note

In case of square matrix this is a unit diagonal matrix.

$$\begin{aligned} & \begin{matrix} a_{11}x_1 & + & a_{12}x_2 & + & & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & & + & a_{2n}x_n & = & b_2 \end{matrix} \\ \mathbf{Ax} = \mathbf{b} & \iff \\ & \begin{matrix} a_{n1}x_1 & + & a_{n2}x_2 & + & & + & a_{nn}x_n & = & b_n \end{matrix} \end{aligned}$$

and consists of two steps:

- ① Forward Elimination $\mathcal{O}(n^3)$ – transforming \mathbf{A} into an upper diagonal form $[\mathbf{U}|\mathbf{b}']$:
$$\begin{aligned} & \begin{matrix} a_{11}x_1 + & a_{12}x_2 & + & & + & a_{1n}x_n & = & b_1 \\ & \overset{(1)}{a_{22}}x_2 & + & \overset{(1)}{a_{23}}x_3 & + & & + & \overset{(1)}{a_{2n}}x_n & = & b_2 \\ & & \overset{(2)}{a_{33}}x_3 & + & & + & \overset{(2)}{a_{3n}}x_n & = & b_3 \end{matrix} \end{aligned}$$
- ② Back Substitution Elimination $\mathcal{O}(n^2)$ – calculating the unknown's \mathbf{x} from \mathbf{U} :
$$a_{(n-1)n}x_n = b_n$$

Is in principle the same as Gauss elimination but reduce the matrix into row-reduced echelon form^[def. 31.22].

Algorithm 31.1 Forward Elimination:
Transforms $\mathbf{Ax} = \mathbf{b}$ into row-echelon form^[def. 31.22].

Given:

- 1: **for** $k = 1, \dots, n - 1$ **do**
- 2: **pivot** $\leftarrow \mathbf{A}(k, k)$
- 3: **for** $i = k + 1, \dots, n$ **do**
- 4: $l_{ik} \leftarrow \frac{\mathbf{A}(i, k)}{\text{pivot}}$
- 5: **for** $j = k + 1, \dots, n$ **do**
- 6: $a_{ij}^{(k)} = \mathbf{A}(i, j) - l_{ik} \mathbf{A}(k, j)$
- 7: **end for**
- 8: $b_i^{(k)} = \mathbf{b}(i) - l_{ik} \mathbf{b}(k)$
- 9: **end for**
- 10: **end for**

Corollary 31.5 Complexity:

$$\sum_{i=1}^{n-1} (n-1)(2(n-i)+3) = n(n-1) \left(\frac{2}{3}n + \frac{7}{6} \right) = \mathcal{O} \left(\frac{2}{3}n^3 \right) \quad (31.22)$$

Algorithm 31.2 Backward Substitution:

Given U:

- 1: $x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$
- 2: **for** $i = n - 1, n - 2, \dots, 1$ **do**
- 3:
$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^n a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}}$$
- 4: **end for**

6.1.2. LU-Decomposition

Corollary 31.7 [proof ??]
LU decomposition Complexity:
 $\frac{2}{3}n^3 + \frac{1}{3}n^2$

6.1.4. Symmetric Positive Definite Matrices

Cholesky Decomposition

Corollary 31.8 [proof 31.5]
Cholesky decomposition Complexity:
 $\frac{1}{3}n^3 + \frac{1}{3}n^2$

7. Non-linear Systems of Equations

7.1. Iterative Methods

Definition 31.26

General Non-linear System of Equations (NLSE) F :
Is a system of non-linear equations F (that do **not** satisfy linearity??):
$$F:\subseteq \mathbb{R}^n \mapsto \mathbb{R}^n \quad \text{seek to find} \quad \mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{0} \quad (31.25)$$

Definition 31.27 Stationary m -point Iteration ϕ_F :
Let $n, m \in \mathbb{R}$ and let $U \subseteq (R^n)^m = \mathbb{R}^n \times \dots \times \mathbb{R}^n$ be a set.
A function $\phi : U \mapsto \mathbb{R}^n$, is called (m -point) iteration function
if it produces an iterative sequence $\left(\mathbf{x}^{(k)}\right)_k$ of approximate
solutions to eq. (31.25), using the m most recent iterates:
$$\mathbf{x}^{(k)} = \phi_F\left(\mathbf{x}^{(k-1)}, \dots, \mathbf{x}^{(k-m)}\right) \quad (31.26)$$

Inital Guess $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(m-1)}$

Note

Stationary as ϕ does no explicitly depend on k .

Definition 31.28 Fixed Point \mathbf{x}^* :
Is a point \mathbf{x}^* for which the sequence does not change any-
more:
$$\mathbf{x}^{(k-1)} = \mathbf{x}^*$$

$$\mathbf{x}^* = \phi_F\left(\mathbf{x}^{(k-1)}, \dots, \mathbf{x}^{(k-m)}\right) \quad \text{with}$$

$$\mathbf{x}^{(k-m)} = \mathbf{x}^* \quad (31.27)$$

7.1.1. Convergence

Question

Does the sequence $\left(\mathbf{x}^{(k)}\right)_k$ converge to a limit:
$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}^* \quad (31.28)$$

7.1.2. Consistency

Definition 31.29 Consistent m -point Iterative Method:
A stationary m -point method^[def. 31.27] is *consistent* with a non-
linear system of equations^[def. 31.26] F iff:
$$F\left(\mathbf{x}^*\right) \iff \phi_F\left(\mathbf{x}^*, \dots, \mathbf{x}^*\right) = \mathbf{x}^* \quad (31.29)$$

7.1.3. Speed of Convergence

add cvg, consistency, speed of cvrg...

7.2. Fixed Point Iterations $m = 1$

Definition 31.30 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 $F(\mathbf{x}) = 0$:
$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \quad \text{Inital Guess: } \mathbf{x}^{(0)} \quad (31.30)$$

Corollary 31.9 Consistency: If ϕ_F is *continuous* and
 $\mathbf{x}^* = \lim_{k \rightarrow \infty} x^{(k)}$ then \mathbf{x}^* is a fixed point^[def. 31.28] of ϕ .

Algorithm 31.3 Fixed Point Iteration:

Input: Inital Guess: $\mathbf{x}^{(0)}$
1: Rewrite $F(\mathbf{x}) = 0$ into a form of $\mathbf{x} = \phi_F(\mathbf{x})$
 \triangleright There exist many ways
2: **for** $k = 1, \dots, T$ **do**
3: Use the fixed point method:

$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \quad (31.31)$$

4: **end for**

add examples and rest

8. Numerical Quadrature

Definition 31.31 Order of a Quadrature Rule:
The **order** of a quadrature rule $\mathcal{Q}_n : C^0([a, b]) \rightarrow \mathbb{R}$ is defined as:
$$\text{order}(\mathcal{Q}_n) := \max \left\{ n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \int_a^b p(t) dt \quad \forall p \in \mathcal{P}_n \right\} + 1 \tag{31.32}$$

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.

8.1. Composite Quadrature

Definition 31.32 Composite Quadrature:
Given a mesh $\mathcal{M} = \{a = x_0 < x_1 < \dots < x_m = b\}$ apply a Q.R. \mathcal{Q}_n to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \dots, m \triangleq \text{p.w.}$ Quadrature:
$$\int_a^b f(t) dt = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t) dt = \sum_{j=1}^m \mathcal{Q}_n(f|_{I_j}) \tag{31.33}$$

Lemma 31.1 Error of Composite quadrature Rules:
Given a function $f \in C^k([a, b])$ with integration domain:
$$\sum_{i=1}^m h_i = |b - a| \quad \text{for } \mathcal{M} = \{x_j\}_{j=1}^m$$

Let: $h_{\mathcal{M}} = \max_j |x_j, x_{j-1}|$ be the **mesh-width**
Assume an equal number of quadrature nodes for each interval $I_j = [x_{j-1}, x_j]$ of the mesh \mathcal{M} i.e. $n_j = n$.
Then the error of a quadrature rule $\mathcal{Q}_n(f)$ of order q is given by:
$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k, q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k, q\}}\right) \quad \text{for } n \rightarrow \infty \\ &\stackrel{[\text{cor. 23.6}]}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^q\right) \quad \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{31.34}$$

Definition 31.33 Complexity W : Is the number of function evaluations \triangleq number of quadrature points.
$$W(\mathcal{Q}(f)_n) = \# \text{f-eval} \triangleq n \tag{31.35}$$

Lemma 31.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :
$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \iff \epsilon_n(f) = cn^{-q} \quad c \in \mathbb{R}_+$$

the error complexity is **algebraic** (??) and is given by:
$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{31.36}$$

Proof 31.1: lemma 31.2: **Assume:** we want to reduce the error by a factor of ϵ_n by increasing the number of quadrature points $n_{\text{new}} = a \cdot n_{\text{old}}$.
Question: what is the additional effort ($\#$ f-eval) needed in order to achieve this reduction in error?
$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \implies n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{31.37}$$

8.1.1. Simpson Integration

Definition 31.34 Simpson Integration:

Filtering Algorithms

10. Signals

Definition 31.35 Time Discrete Signal: Is a bounded sequence^[def. 20.2] $(x_j)_{j \in \mathbb{Z}} \in l^\infty(\mathbb{Z})$.

Definition 31.36 Sampling:

Corollary 31.10 Finite Time Discrete Signal:

11. Channels/Filters

Definition 31.37 Channel/Filter: *F*
Is a mapping of signals to signals $F ::$
$$F : l^\infty(\mathbb{Z}) \mapsto l^\infty(\mathbb{Z}) \tag{31.38}$$

Property 31.1 Finite Channel/Filter: A filter $F : l^\infty(\mathbb{Z}) \mapsto l^\infty(\mathbb{Z})$

Property 31.2 Causal Channel/Filter:

Explanation 31.3. *The response cannot start before the signal has been feed into the filter.*

Definition 31.38 Time Shift Operator: *S_m*

Property 31.3 Time-invariant Channel/Filter:

Explanation 31.4. *The response of the filter should not depend at which time we pass the signal to the filter.*

Property 31.4 Linear Channel/Filter:

Definition 31.39 Linear Time-invariant Finite Input Response Filter LT-FIR:

11.1. Impulse Responses

Definition 31.40 Impulse:

Definition 31.41 Impulse Response *h:*

Corollary 31.11 [proof 31.2]
Signal in terms of Impulse Responses: We can write any arbitrary discrete signal as weighted sum of time shifted impulses:
$$F(x_j) = \tag{31.39}$$
$$(F(x_j))_j = \tag{31.40}$$

Proof 31.2 ^[cor. 31.11]:

11.2. Discrete Convolution

Definition 31.42 LT-FIR formula:

Proofs

Proof 31.3 Log Sum Trickformula 31.1:
$$\begin{aligned} \text{LSE} &= \log \left(\sum_{i=1}^n e^{x_i} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} e^a \right) \\ &= \log \left(e^a \sum_{i=1}^n e^{x_i - a} \right) = \log \left(\sum_{i=1}^n e^{x_i - a} \right) + \log(e^a) \\ &= \log \left(\sum_{i=1}^n e^{x_i - a} \right) + a \end{aligned}$$

Proof 31.4 LU-Complexity^[cor. 31.7]:
For eliminating the first column we need to eliminate $n - 1$ rows by n additions and n multiplications which equals $(n - 1)2n$. For the second column we need for $n - 2$ rows $n - 1$ additions and $n - 1$ multiplications which equals $(n - 2)2(n - 1)$ thus to eliminate all n columns we have:
$$\sum_{i=1}^n (n - i + 1) \cdot 2(n - i)$$
using the index $l = n - i + 1$ we can write this as:
$$\begin{aligned} \sum_{i=1}^n (n - i + 1) \cdot 2(n - i) &= 2 \sum_{i=1}^n (j + 1) \cdot (j) = 2 \sum_{l=0}^n j^2 + 1 \\ &= 2 \left(\frac{1}{3} n^3 - \frac{1}{3} n \right) \end{aligned}$$

add rules for sums of n and n^2

Proof 31.5 Cholesky Complexity^[cor. 31.8]: **U** and **L** “are the same” as we have a s.p.d. matrix s.t. we can simply half the forward elimination complexity of the LU-decomposition^[cor. 31.7]:
$$\frac{1}{2} \frac{2}{3} n^3 + \frac{1}{3} n^2 \tag{31.41}$$

Optimization

Definition 32.1 First Order Method: A first-order method is an algorithm that chooses the k -th iterate in $\mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_{k-1})\} \quad \forall k = 1, 2, \dots \quad (32.1)$

Note

Gradient descent is a first order method

1. Linear Optimization

1.1. Polyhedra

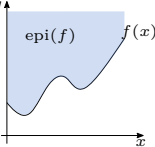
Definition 32.2 Polyhedron: Is a set $P \in \mathbb{R}^n$ that can be described by the *finite* intersection of m closed *half spaces*??:

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_j \mathbf{x} \leq b_j, j = 1, \dots, m\}$$
$$\mathbf{A} \in \mathbb{R}^{m \times n} \quad \mathbf{b} \in \mathbb{R}^m \quad (32.2)$$

1.1.1. Polyhedral Function

Definition 32.3 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 graph:

$$\text{epi}(f) := \{(\mathbf{x}, y) \in \mathbb{R}^n \mid y \geq f(\mathbf{x})\} \subseteq \mathbb{R}^{n+1} \quad (32.3)$$


Definition 32.4 Polyhedral Function: A function f is *polyhedral* if its epigraph $\text{epi}(f)$ ^[def. 32.3] is a polyhedral set ^[def. 32.2]:

$$f \text{ is polyhedral} \iff \text{epi}(f) \text{ is polyhedral} \quad (32.4)$$

2. Lagrangian Optimization Theory

Add: derivation of lagrange function

Definition 32.5 (Primal) Constraint Optimization:

Given an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:

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

Definition 32.6 Lagrange Function:

$$\mathcal{L}(\alpha, \beta, \mathbf{w}) := f(\mathbf{w}) + \alpha \mathbf{g}(\mathbf{w}) + \beta \mathbf{h}(\mathbf{w}) \quad (32.5)$$

Extremal Conditions

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

For the inequality constraints $g(\mathbf{x}) \leq 0$ we distinguish two situations:

Case I : $g(\mathbf{x}^*) < 0$ switch const. off
Case II : $g(\mathbf{x}^*) \geq 0$ optimize using active eq. constr.

$$\frac{\partial}{\partial \alpha} \mathcal{L}(\mathbf{x}) = g(\mathbf{x}) \stackrel{!}{=} 0 \quad \text{Constraint satisfaction}$$

Definition 32.7 Lagrangian Dual Problem: Is given by:

$$\begin{aligned} \text{Find} \quad \max_{\alpha, \beta} \theta(\alpha, \beta) &= \inf_{\mathbf{w} \in \Omega} \mathcal{L}(\mathbf{w}, \alpha, \beta) \\ \text{s.t.} \quad \alpha_i &\geq 0 \quad 1 \leq i \leq k \end{aligned}$$

Solution Strategy

- Find the extremal point \mathbf{w}^* of $\mathcal{L}(\mathbf{w}, \alpha, \beta)$:
$$\left. \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^*} \stackrel{!}{=} 0 \quad (32.6)$$
- Insert \mathbf{w}^* into \mathcal{L} and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left. \frac{\partial \theta}{\partial \beta} \right|_{\beta=\beta^*} \stackrel{!}{=} 0 \quad (32.7)$$
- 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 (α^*, β^*) .

Theorem 32.1 Upper Bound Dual Cost: Let $\mathbf{w} \in \Omega$ be a feasible solution of the primal problem ^[def. 32.5] and (α, β) a *feasible solution* of the respective dual problem ^[def. 32.7].

Then it holds that:

$$f(\mathbf{w}) \geq \theta(\alpha, \beta) \quad (32.8)$$

Proof 32.1:

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

Corollary 32.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup \{\theta(\alpha, \beta) : \alpha \geq 0\} \leq \inf \{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leq 0, \mathbf{h}(\mathbf{w}) = 0\} \quad (32.9)$$

Theorem 32.2 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:

$$f(\mathbf{w}^*) = \theta(\alpha^*, \beta^*) \quad (32.10)$$

Definition 32.8 Convex Optimization: Given: a *convex*

$$\begin{aligned} \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{s.t.} \quad \mathbf{x} \in S \end{aligned} \quad (32.11)$$

Often S is specified using linear inequalities:

$$\text{e.g.} \quad S = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{Ax} \leq \mathbf{b}\}$$

Theorem 32.3 Strong Duality: Given an convex optimization problem:

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

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

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

Theorem 32.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

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

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

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

$$\frac{\partial \mathcal{L}(\mathbf{w}, \alpha, \beta)}{\partial \mathbf{w}} \stackrel{!}{=} 0 \quad \frac{\partial \mathcal{L}(\mathbf{w}^*, \alpha, \beta)}{\partial \beta} \stackrel{!}{=} 0 \quad (32.12)$$

under the conditions that:

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

Consequence

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

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

Combinatorics

Permutations

Definition 33.1 **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 sequence^[def. 20.2] – **order does matter**.

Definition 33.2 **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:

$$P_n(\mathcal{S}) = n! = \prod_{i=0}^{n-1} (n - \textcolor{red}{i}) = n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 1 \tag{33.1}$$

Explanation 33.1. If we have i.e. three distinct elements $\{\bullet, \textcolor{blue}{\bullet}, \textcolor{green}{\bullet}\}$ For the first element \bullet that we arrange we have three possible choices where to put it. However this reduces the number of possible choices for the second element $\textcolor{green}{\bullet}$ to only two. Consequently for the last element $\textcolor{blue}{\bullet}$ we have no choice left.



Definition 33.3 **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 $j \in \{1, \dots, k\}$ of the multiset \mathcal{S} . The permutation of \mathcal{S} is given by:

$$P_{n_1, \dots, n_k}(\mathcal{S}) = \frac{n!}{n_1! \cdot \dots \cdot n_k!} \quad \text{s.t.} \quad \sum_{j=1}^k n_j \leq n \quad k < n \tag{33.2}$$

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) \Rightarrow less possibilities to arrange the elements uniquely.

Picking things from a bag

1. Combinations

Definition 33.4 **k -Combination:** A k -combination of a set \mathcal{S} of *distinct* elements of size $n = \mathcal{S}$ is a subset \mathcal{S}_k (**order does not matter**) of $k = |\mathcal{S}_k|$, *chosen* from \mathcal{S} .

Note

Thus unlike in a permutation we just care about what we pick and not how it ends up beeing arranged.

Definition 33.5 **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 n choose k :

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n - k)!} \tag{33.3}$$

2. Variation

Definition 33.6 **Variation:** A k -variation of a set \mathcal{S} of size $n = \mathcal{S}$ is

1. a selection/combination^[def. 33.4] 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. 33.2] of that subset \mathcal{S}_k (with or without repetition) into a sequence^[def. 20.2]

Definition 33.7 **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:

$$V_k^n(\mathcal{S}) = \binom{n}{k} k! = \frac{n!}{(n - k)!} \tag{33.4}$$

Note

Sometimes also denotes as P_k^n .

Definition 33.8 **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:

$$\bar{V}_k^n(\mathcal{S}) = n^k \tag{33.5}$$

Stochastics

Definition 33.9 Stochastics:	Is a collective term for the areas of <i>probability theory</i> and <i>statistics</i> .
Definition 33.10 Statistics:	Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.
Definition 33.11 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 33.12 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 quora/hh	
Note: Stochastics vs. Stochastic	
Stochastics is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is a <i>adjective</i> , describing that a certain phenomena is governed by uncertainty i.e. a process.	
Probability Theory	
Definition 34.1 Probability Space	$W = \{\Omega, \mathcal{F}, \mathbb{P}\}$: Is the unique triple $\{\Omega, \mathcal{F}, \mathbb{P}\}$, where Ω is its sample space, \mathcal{F} its σ -algebra of events, and \mathbb{P} its probability measure.
Definition 34.2 Sample Space Ω :	[example 34.1] Is the set of all possible outcomes (elementary events) [cor. 34.5] of an experiment.
Definition 34.3 Event	[example 34.2] A : An “event” is a subset of the sample space Ω and is a property which can be observed to hold or not to hold <i>after</i> the experiment is done. Mathematically speaking not every subset of Ω is an event and has an associated probability. Only those subsets of Ω that are part of the corresponding σ -algebra \mathcal{F} are events and have their assigned probability.
Corollary 34.1 : If the outcome ω of an experiment is in the subset A , then the event A is said to “have occurred”.	
Corollary 34.2 Complement Set	A^C : is the contrary event of A .
Corollary 34.3 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 34.4 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 34.5 The Elementary Event	ω : Is a “singleton”, i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω .
Corollary 34.6 The Sure Event	Ω : Is equal to the sample space as it contains all possible elementary events.
Corollary 34.7 The Impossible Event	\emptyset : The impossible event i.e. nothing is happening is denoted by the empty set.
Definition 34.4 The Family of All Events	$\mathcal{A}/2^\Omega$: The set of all subset of the sample space Ω called family of all events is given by the power set of the sample space $\mathcal{A} = 2^\Omega$ (for finite sample spaces).

Definition 34.5 Probability	$\mathbb{P}(A)$: Is a number associated with every A , that measures the likelihood of the event to be realized “a priori”. The bigger the number the more likely the event will happen. 1. $0 \leq \mathbb{P}(A) \leq 1$ 2. $\mathbb{P}(\Omega) = 1$ 3. If $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
Note	We can think of the probability of an event A as the limit of the "frequency" of repeated experiments: $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \frac{\delta_n(A)}{n} \quad \text{where} \quad \delta(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$
1. Sigma Algebras	
Definition 34.6 Sigma Algebra	[Proof 34.3] σ : A set \mathcal{F} of subsets of Ω is called a σ -algebra on Ω if the following properties apply <ul style="list-style-type: none">$\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$If $A \in \mathcal{F}$ then $\Omega \setminus A = A^C \in \mathcal{F}$: The complementary subset of A is also in Ω.For all $A_i \in \mathcal{F} : \bigcup_{i=1}^\infty A_i \in \mathcal{F}$
Explanation 34.1 ([def. 34.6]). <i>The σ-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.</i>	
Corollary 34.8 \mathcal{F}_{\min}: $\mathcal{F} = \{\emptyset, \Omega\}$ is the simplest σ -algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.	
Corollary 34.9 \mathcal{F}_{\max}: $\mathcal{F} = 2^\Omega$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.	
Definition 34.7 Measurable Space	(Ω, \mathcal{F}) : Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.
Corollary 34.10 \mathcal{F}-measurable Event	$A_i \in \mathcal{F}$: The measurable events A_i of \mathcal{F} are called <i>\mathcal{F}-measurable</i> or <i>measurable sets</i> .
Definition 34.8 Sigma Algebra generated by a subset of Ω	[Example 34.4] $\sigma(C)$: Let C be a class of subsets of Ω . The σ -algebra generated by C , denoted by $\sigma(C)$, is the <i>smallest</i> sigma algebra \mathcal{F} that included all elements of C .
Definition 34.9 Borel σ-algebra	[Example 34.5] $\mathcal{B}(\mathbb{R})$: The Borel σ -algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ -algebra containing all open intervals in \mathbb{R} . The sets in contained in $\mathcal{B}(\mathbb{R})$ are called Borel sets. The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward. For all real numbers $a, b \in \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ contains various sets.
Why do we need Borel Sets	
So far we only looked at atomic events ω , with the help of sigma algebras we are now able to measure continuous events s.a. $[0, 1]$.	
Definition 34.10 Borel Set:	
Corollary 34.11 Generating Borel σ-Algebra [Proof 34.1]: The Borel σ -algebra of \mathbb{R} is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ (\mathbb{Q} =rationals).	

Definition 34.11 (\mathbb{P})-trivial Sigma Algebra:	is a σ -algebra \mathcal{F} for which each event has a probability of zero or one: $\mathbb{P}(A) \in \{0, 1\} \quad \forall A \in \mathcal{F} \quad (34.1)$
Interpretation	A trivial sigma algebra means that all events are almost surely constant and that there exist no non-trivial information. An example of a trivial sigma algebra is $\mathcal{F}_{\min} = \{\Omega, \emptyset\}$.
2. Measures	
Definition 34.12 Measure	μ : A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map: $\mu : \mathcal{F} \mapsto [0, \infty] \quad (34.2)$
for which holds: <ul style="list-style-type: none">$\mu(\emptyset) = 0$countable additivity [def. 34.13]	
Definition 34.13 Countable/σ-Additive Function:	Given a function μ defined on a σ -algebra \mathcal{F} . The function μ is said to be countable additive if for every countable sequence of pairwise disjoint elements $(F_i)_{i \geq 1}$ of \mathcal{F} it holds that: $\mu\left(\bigcup_{i=1}^\infty F_i\right) = \sum_{i=1}^\infty \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \emptyset \quad \forall j \neq k \quad (34.3)$
Corollary 34.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds: $F \cap G = \emptyset \implies \mu(F \cup G) = \mu(F) + \mu(G) \quad (34.4)$	
Explanation 34.2. <i>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.</i>	
Definition 34.14 Equivalent Measures	[Example 34.6] $\mu \sim \nu$: Let μ and ν be two measures defined on a measurable space [def. 34.7] (Ω, \mathcal{F}) . The two measures are said to be equivalent if it holds that: $\mu(A) > 0 \iff \nu(A) > 0 \quad \forall A \subseteq \mathcal{F} \quad (34.5)$
this is equivalent to μ and ν having equivalent null sets: $\mathcal{N}_\mu = \mathcal{N}_\nu \quad \mathcal{N}_\mu = \{A \in \mathcal{A} \mu(A) = 0\} \quad \mathcal{N}_\nu = \{A \in \mathcal{A} \nu(A) = 0\} \quad (34.6)$	
Definition 34.15 Measure Space	$(\mathcal{F}, \Omega, \mu)$: The triplet of sample space, sigma algebra and a measure is called a measure space.
2.1. Borel Measures	
Definition 34.16 Borel Measure:	A Borel Measure is any measure [def. 34.12] μ defined on the Borel σ -algebra [def. 34.9] $\mathcal{B}(\mathbb{R})$.
2.1.1. The Lebesgue Measure	
Definition 34.17 Lebesgue Measure on \mathcal{B}	λ : Is the Borel measure [def. 34.16] defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns for every half-open interval $(a, b]$ interval its length: $\lambda((a, b]) := b - a \quad (34.7)$

Corollary 34.13 Lebesgue Measure of Atomites:	
<ul style="list-style-type: none">The Lebesgue measure of a set containing only one point must be zero: $\lambda(\{a\}) = 0 \quad (34.8)$The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2, \dots, a_n\}$ must be zero: $\lambda(A) + \sum_{i=1}^n \lambda(\{a_i\}) = 0 \quad (34.9)$The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2, \dots\}$ can be either zero, positive and finite or infinite.	
3. Probability/Kolomogorov's Axioms 1931	
One problem we are still having is the range of μ , by standardizing the measure we obtain a well defined measure of events.	
Axiom 34.1 Non-negativity:	The probability of an event is a non-negative real number: If $A \in \mathcal{F}$ then $\mathbb{P}(A) \geq 0 \quad (34.10)$
Axiom 34.2 Unitaarity:	The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one: The certain event $\mathbb{P}(\Omega) = 1 \quad (34.11)$
Axiom 34.3 σ-additivity:	If $A_1, A_2, A_3, \dots \in \mathcal{F}$ are mutually disjoint, then: $\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu(A_i) \quad (34.12)$
Corollary 34.14 : As a consequence of this it follows: $\mathbb{P}(\emptyset) = 0 \quad (34.13)$	
Corollary 34.15 Complementary Probability: $\mathbb{P}(A^C) = 1 - \mathbb{P}(A) \quad \text{with} \quad A^C = \Omega - A \quad (34.14)$	
Definition 34.18 Probability Measure	\mathbb{P} : a probability measure is function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ defined on a σ -algebra \mathcal{F} of a sample space Ω that satisfies the probability axioms.
4. Conditional Probability	
Definition 34.19 Conditional Probability:	Let A, B be events, with $\mathbb{P}(B) \neq 0$. Then the conditional probability of the event A given B is defined as: $\mathbb{P}(A B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \mathbb{P}(B) \neq 0 \quad (34.15)$
5. Independent Events	
Theorem 34.1 Independent Events: Let A, B be two events. A and B are said to be independent iff: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \quad \mathbb{P}(A B) = \mathbb{P}(A), \quad \mathbb{P}(B) > 0$ $\mathbb{P}(B A) = \mathbb{P}(B), \quad \mathbb{P}(A) > 0 \quad (34.16)$	
Note The requirement of no impossible events follows from [def. 34.19]	
Corollary 34.16 Pairwise Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>pairwise independent</i> if every pair of events is independent: $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \mathbb{P}(A_j) \quad i \neq j, \quad \forall i, j \in \mathcal{A} \quad (34.17)$	
Corollary 34.17 Mutal Independent Evenest: A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is <i>mutal independent</i> if every event A_j is independent of any intersection of the other events: $\mathbb{P}\left(\bigcap_{i=i}^k B_i\right) = \prod_{i=1}^k \mathbb{P}(B_i) \quad \forall \{B_i\}_{i=1}^k \subseteq \{A_i\}_{i=1}^n \quad k \leq n, \quad \{A_i\}_{i=1}^n \in \mathcal{A} \quad (34.18)$	

6. Product Rule

Law 34.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \quad (34.19)$$

Law 34.2

Generalized Product Rule/Chain Rule: is the generalization of the product rule?? to n events $\{A_i\}_{i=1}^n$

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^n E_i\right) &= \prod_{k=1}^n \mathbb{P}\left(E_k \mid \bigcap_{i=1}^{k-1} E_i\right) = \\ &= \mathbb{P}(E_n | E_{n-1} \cap \dots \cap E_1) \cdot \mathbb{P}(E_{n-1} | E_{n-2} \cap \dots \cap E_1) \cdots \\ &\quad \cdots \mathbb{P}(E_3 | E_2 \cap E_1) \mathbb{P}(E_2 | E_1) \mathbb{P}(E_1) \end{aligned} \quad (34.20)$$

7. Law of Total Probability

Definition 34.20 Complete Event Field: A complete event field $\{A_i : i \in I \subseteq \mathbb{N}\}$ is a countable or finite partition of Ω that is the partitions $\{A_i : i \in I \subseteq \mathbb{N}\}$ are a *disjoint union* of the sample space:

$$\bigcup_{i \in I} A_i = \Omega \quad A_i \cap A_j = \emptyset \quad i \neq j, \forall i, j \in I \quad (34.21)$$

Theorem 34.2

Law of Total Probability/Partition Equation: Let $\{A_i : i \in I\}$ be a complete event field^[def. 34.20] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \quad (34.22)$$

8. Bayes Theorem

Law 34.3 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \mathbb{P}(B) > 0 \quad (34.23)$$

follows directly from eq. (34.19).

Theorem 34.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field^[def. 34.20] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \quad (34.24)$$

proof ?? 34.2

Distributions on \mathbb{R}

9.1. Distribution Function

Definition 34.21 Distribution Function of \mathbb{P} F : The *distribution function* F induced by a probability measure \mathbb{P} on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((-\infty, x]) \quad (34.25)$$

Theorem 34.4 : A function F is the distribution function of a (unique) probability on $(\mathbb{R}, \mathcal{B})$ iff:

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

Corollary 34.18 : 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 $\mathbb{P} = \mathbb{Q}$.

9.2. 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:

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

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

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

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

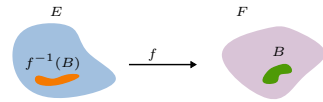
Probability for an event in Ω

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

Probability for an event in E

Definition 34.22 \mathcal{E} -measurable function: Let (E, \mathcal{E}) and (F, \mathcal{F}) be two measurable spaces. A function $f : E \mapsto F$ is called measurable (relative to \mathcal{E} and \mathcal{F}) if

$$\forall B \in \mathcal{F} : \quad f^{-1}(B) = \{\omega \in E : f(\omega) \in B\} \in \mathcal{E} \quad (34.26)$$



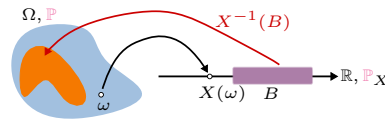
Interpretation

The pre-image^[def. 23.11] 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 34.23 Random Variable: A real-valued random variable (vector) X , defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ is an \mathcal{E} -measurable function mapping, if it maps its sample space Ω into a target space (F, \mathcal{F}) :

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \quad (34.27)$$

Since X is \mathcal{E} -measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 34.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ -algebra \mathcal{B} of \mathbb{R} .

Corollary 34.20 Random Variables of Borel Sets: Given that we work with Borel σ -algebras then the definition of a random variable is equivalent to (due to ^[cor. 34.11]):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \quad (34.28)$$

Definition 34.24

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 34.21 Indicator Functions

An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \quad (34.29)$$

We know that a probability measure \mathbb{P} on \mathbb{R} is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 34.22 : 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

$$\begin{aligned} \{X \leq a\} &= \{\omega : X(\omega) \leq a\} = X^{-1}((-\infty, a]) \in \mathcal{E}, \forall a \in \mathbb{R} \\ \text{or} \quad \{X < a\} &\in \mathcal{E}. \end{aligned}$$

Explanation 34.3 (^[cor. 34.22]). A random variable is a function that is measurable if and only if its distribution function is defined.

9.3. The Law of Random Variables

Definition 34.25 Law/Distribution of X $\mathcal{L}(X)$:

Let X be a r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in (E, \mathcal{E}) , then the *distribution*/law of X is defined as:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \quad (34.30)$$

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \quad \forall B \in \mathcal{E}$$

Note

- Sometimes \mathbb{P}^X is also called the *image* of \mathbb{P} by X
- The law can also be written as:

$$\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)) = (\mathbb{P} \circ X^{-1})(B)$$

Theorem 34.5 : The law/distribution of X is a probability measure \mathbb{P} on (E, \mathcal{E}) .

Definition 34.26

(Cumulative) Distribution Function F_X :

Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leq x) \quad (34.31)$$

Corollary 34.23 : The distribution of \mathbb{P}^X of a real valued r.v. is entirely characterized by its cumulative distribution function F_X ^[def. 34.33].

Property 34.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \quad (34.32)$$

Property 34.2: Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad (34.33)$$

9.4. Probability Density Function

Definition 34.27 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 34.28 Probability Density Function: Let X be a r.v. with associated cdf F_X . If F_X is continuously integrable for all $x \in \mathbb{R}$ then X has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy \quad (34.34)$$

or alternatively:

$$f_X(x) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x \leq X \leq x + \epsilon)}{\epsilon} \quad (34.35)$$

Corollary 34.24 $\mathbb{P}(X = b) = 0, \quad \forall b \in \mathbb{R}$:

$$\mathbb{P}(X = b) = \lim_{a \rightarrow b} \mathbb{P}(a < X \leq b) = \lim_{a \rightarrow b} \int_a^b f(x) dx = 0 \quad (34.36)$$

Corollary 34.25 : From ^[cor. 34.24] it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Corollary 34.26 :

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (34.37)$$

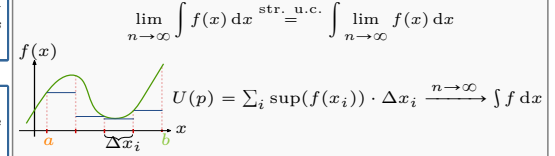
Notes

- 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*.

9.5. 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

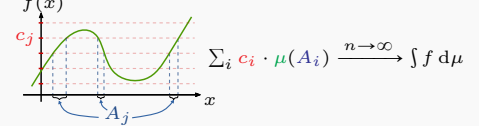


Idea

Partition domain by function values of equal size i.e. values that lie within the same sets/have the same value A_j build up the partitions w.r.t. to the variable x .

Problem: we do not know how big those sets/partitions on the x -axis will be.

Solution: we can use the measure μ of our measure space $(\Omega, \mathcal{A}, \mu)$ in order to obtain the size of our sets $A_j \Rightarrow$ we do not have to care anymore about discontinuities, as we can measure the size of our sets using our measure.



Definition 34.29 Lebesgue Integral:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n c_i \mu(A_i) = \int_{\Omega} f d \mu \quad f(x) \approx c_i \quad \forall x \in A_i \quad (34.38)$$

Definition 34.30

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:

$$X = \sum_{i=1}^n a_i \mathbb{1}_{A_i} \quad a_i \in \mathbb{R} \quad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \quad (34.39)$$

9.6. 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 34.1.

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 34.31 Independent Random Variables:

Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leq x | Y \leq y) = \mathbb{P}(X \leq x) \quad \forall x, y \in \mathbb{R} \quad (34.40)$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(X \leq x, Y \leq y) \\ &= F_X(x) F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \quad (34.41)$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B) \quad \forall A, B \in \mathcal{B} \quad (34.42)$$

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 34.32
Independent Identically Distributed:

10. Product Rule

Law 34.4 Product Rule: Let X, Y be two random variables then their jo

Law 34.5
Generalized Product Rule/Chain Rule:

11. Change Of Variables Formula

Formula 34.1
(Scalar Discret) Change of Variables: Let X be a discret rv $X \in \mathcal{X}$ with pmf \mathbf{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. 23.14) 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:
$$\mathbf{p}_Y(y) = \sum_{x_i \in \mathcal{X}_y} \mathbf{p}_X(x_i) = \sum_{x \in \mathcal{Y}: g(x)=y} \mathbf{p}_X(x) \quad (34.43)$$

see proof ?? 34.3

Formula 34.2
(Scalar Continuous) Change of Variables:
Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic(def. 23.14) function.
Define a new r.v. Y as
$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \quad (34.44)$$

then the pdf of Y is given by:
$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X(x) \left| \frac{d}{dy} (g^{-1}(y)) \right| \quad (34.45)$$

$$= f_X(x) \frac{1}{\left| \frac{dy}{dx} \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx}(g^{-1}(y)) \right|} \quad (34.46)$$

Formula 34.3
(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. 23.14) function $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$
Define a new r.v. Y as
$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \quad (34.47)$$

and let $h(x) := g(x)^{-1}$ then the pdf of Y is given by:
$$\begin{aligned} f_Y(y) &= f_X(x_1, \dots, x_n) \cdot |J| \\ &= f_X(h_1(y), \dots, h_n(y)) \cdot |J| \\ &= f_X(y) |\det D_x h(x)| \Big|_{x=y} \\ &= f_X(g^{-1}(y)) \left| \det \left(\frac{\partial g}{\partial x} \right) \right|^{-1} \end{aligned} \quad (34.48)$$

where $J = \det Dh$ is the Jacobian(def. 24.6).
See also proof ?? 34.6 and example 34.8

Note
A monotonic function is required in order to satisfy inevitability.

Probability Distributions on \mathbb{R}^n

13. Joint Distribution

Definition 34.33
Joint (Cumulative) Distribution Function $F_{\mathbf{X}}$:
Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector in \mathbb{R}^n , then its cumulative distribution function is defined as:
$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}^X((-\infty, \mathbf{x}]) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) \\ &= \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \end{aligned} \quad (34.49)$$

Definition 34.34 Joint Probability Distribution:
Let $\mathbf{X} = (X_1, \dots, X_n)$ 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* $f_{\mathbf{X}}$ defined by:
$$F_X(x) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_{\mathbf{X}}(y_1, \dots, y_n) dy_1 \dots dy_n \quad (34.50)$$

or alternatively:
$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(x_1 \leq X_1 \leq x_1 + \epsilon, \dots, x_n \leq X_n \leq x_n + \epsilon)}{\epsilon} \quad (34.51)$$

13.1. Marginal Distribution

Definition 34.35 Marginal Distribution:

14. The Expectation

Definition 34.36 Expectation:
$$\mathbb{E}[X] = \int_{\Omega} X(\omega) \mathbb{P}(d\omega) = \int_{\Omega} X d\mathbb{P} \quad (34.52)$$

Corollary 34.27 Expectation of simple r.v.:
If X is a simple(def. 34.30) r.v. its *expectation* is given by:
$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \quad (34.53)$$

14.1. Properties
14.1.1. Linear Operators

14.1.2. Quadratic Form
Definition 34.37 proof 34.7
Expectation of a Quadratic Form:
Let $\epsilon \in \mathbb{R}^n$ be a random vector with $\mathbb{E}[\epsilon] = \mu$ and $\mathbb{V}[\epsilon] = \Sigma$:
$$\mathbb{E}[\epsilon^T A \epsilon] = \text{tr}(A \Sigma) + \mu^T A \mu \quad (34.54)$$

14.2. The Jensen Inequality
Theorem 34.6 Jensen Inequality: Let X be a random variable and g some function, then it holds:
$$\begin{aligned} g(\mathbb{E}[X]) &\leq \mathbb{E}[g(X)] & \text{if } g \text{ is convex} & \text{[def. 23.24]} \\ g(\mathbb{E}[X]) &\geq \mathbb{E}[g(X)] & \text{if } g \text{ is concave} & \text{[def. 23.25]} \end{aligned} \quad (34.55)$$

14.3. Law of the Unconscious Statistician

Law 34.6 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 :
$$\mathbb{E}_Y[y] = \mathbb{E}_X[g(x)] \quad (34.56)$$

Consequence
Hence if we \mathbf{p}_X we do not have to first calculate \mathbf{p}_Y in order to calculate $\mathbb{E}_Y[y]$.

14.4. Properties

14.5. Law of Iterated Expectation (LIE)

Law 34.7 [proof 34.8]
Law of Iterated Expectation (LIE):
$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]] \quad (34.57)$$

14.6. Hoeffdings Bound

Definition 34.38 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:
$$\mathbb{P}(|\mu_{\mathbf{X}} - \mathbb{E}[X]| \geq \epsilon) \leq 2 \exp \left(\frac{-2n^2 \epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \stackrel{[0,1]}{=} 2e^{-2n\epsilon^2} \quad (34.58)$$

Explanation 34.4. The difference of the expectation from the empirical average to be bigger than ϵ is upper bound in probability.

15. Moment Generating Function (MGF)

Definition 34.39 Moment of Random Variable: The i -th moment of a random variable X is defined as (if it exists):
$$m_i := \mathbb{E}[X^i] \quad (34.59)$$

Definition 34.40 ψ_X
Moment Generating Function (MGF):
$$\psi_X(t) = \mathbb{E}[e^{tX}] \quad t \in \mathbb{R} \quad (34.60)$$

Corollary 34.28 Sum of MGF: The moment generating function of a sum of n independent variables $(X_j)_{1 \leq j \leq n}$ is the product of the moment generating functions of the components:
$$\psi_{S_n}(t) = \psi_{X_1}(t) \dots \psi_{X_n}(t) \quad S_n := X_1 + \dots + X_n \quad (34.61)$$

Corollary 34.29 : The i -th moment of a random variable is the i -th derivative of its associated moment generating function evaluated zero:
$$\mathbb{E}[X^i] = \psi_X^{(i)}(0) \quad (34.62)$$

16. 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 34.41 $\hat{\mu}$
Fourier Transformed Probability Measure:
$$\hat{\mu} = \int e^{i\langle u, x \rangle} \mu(dx) \quad (34.63)$$

Corollary 34.30 : As $e^{i\langle u, x \rangle}$ can be rewritten using formulae. (19.9) and (19.10) it follows:
$$\hat{\mu} = \int \cos(\langle u, x \rangle) \mu(dx) + i \int \sin(\langle u, x \rangle) \mu(dx) \quad (34.64)$$

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 34.42 Characteristic Function φ_X : Let \mathbf{X} be an \mathbb{R}^n -valued random variable. Its characteristic function φ_X is defined on \mathbb{R}^n as:
$$\begin{aligned} \varphi_{\mathbf{X}}(u) &= \int e^{i\langle u, \mathbf{x} \rangle} \mathbb{P}^X(d\mathbf{x}) = \widehat{\mathbb{P}^X}(u) \\ &= \mathbb{E}[e^{i\langle u, \mathbf{x} \rangle}] \end{aligned} \quad (34.65) \quad (34.66)$$

Corollary 34.31 : The characteristic function φ_X of a distribution always exists as it is equal to the Fourier transform of the probability measure, which always exists.

Note
This is an advantage over the moment generating function.

Theorem 34.7 : Let μ be a probability measure on \mathbb{R}^n . Then $\hat{\mu}$ is a bounded continuous function with $\hat{\mu}(0) = 1$.
add proof

Theorem 34.8 Uniqueness Theorem: The Fourier Transform $\hat{\mu}$ of a probability measure μ on \mathbb{R}^n characterizes μ . That is, if two probability measures on \mathbb{R}^n admit the same Fourier transform, they are equal.
add proof

Corollary 34.32 : Let $\mathbf{X} = (X_1, \dots, X_n)$ be an \mathbb{R}^n -valued random variable. Then the real valued r.v.'s $(X_j)_{1 \leq j \leq n}$ are independent if and only if:

$$\varphi_X(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \quad (34.67)$$

Proofs

Proof 34.1: [cor. 34.11]: Let \mathcal{C} denote all open intervals. Since every open set in \mathbb{R} is the countable union of open intervals[def. 19.12], it holds that $\sigma(\mathcal{C})$ is the Borel σ -algebra of \mathbb{R} .
Let \mathcal{D} denote all intervals of the form $(-\infty, a]$, $a \in \mathbb{Q}$.
Let $a, b \in \mathcal{C}$, and let

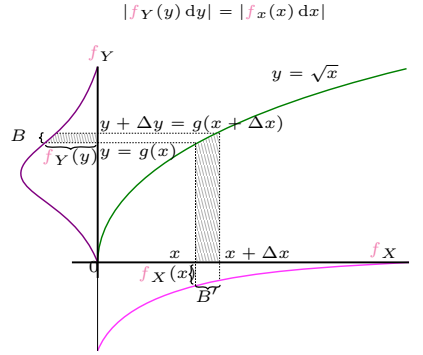
- $(a_n)_{n>1}$ be a sequence of rationals decreasing to a and
- $(b_n)_{n>1}$ be a sequence of rationals increasing strictly to b

 $(a, b) = \cup_{n=1}^{\infty} (a_n, b_n] = \cup_{n=1}^{\infty} ((-\infty, b_n] \cap (-\infty, a_n]^C)$
Thus $\mathcal{C} \subset \sigma(\mathcal{D})$, whence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$ but as each element of \mathcal{D} is a closed subset, $\sigma(\mathcal{D})$ must also be contained in the Borel sets \mathcal{B} with
$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{D}) \subset \mathcal{B}$$

Proof 34.2: theorem 34.3 Plug eq. (34.22) into the denominator and eq. (24.2) into the nominator and then use [def. 34.19]:
$$\frac{\mathbb{P}(B|A_j) \mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i) \mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

Proof 34.3: ??:
$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = \mathbf{p}_Y(y)$$

Proof 34.4: ?? (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:



Proof 34.5: ?? from CDF:
$$\mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \begin{cases} \mathbb{P}(X \leq g^{-1}(y)) & \text{if } g \text{ is increas.} \\ \mathbb{P}(X \geq g^{-1}(y)) & \text{if } g \text{ is decreas.} \end{cases}$$

If g is monotonically increasing:
$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$

If g is monotonically decreasing:
$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{d}{dy} g^{-1}(y)$$

Proof 34.6: ??: 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:

$y = g(x) \Rightarrow \mathbb{P}(y) = \mathbb{P}(g(x))$ (for disc. rv.)

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t) dt \xrightarrow{\Delta x \rightarrow 0} |\Delta x \cdot f_X(x)|$$

For y we use Taylor (??)

$$g(x + \Delta x) \stackrel{\text{eq. (23.56)}}{=} g(x) + \frac{dg}{dx} \Delta y \quad \text{for } \Delta x \rightarrow 0$$
$$= y + \Delta y \quad \text{with } \Delta y := \frac{dg}{dx} \cdot \Delta x \quad (34.68)$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\mathbb{P}(X \in B') = \int_y^{y+\Delta y} f_Y(t) dt \xrightarrow{\Delta y \rightarrow 0} |\Delta y \cdot f_Y(y)|$$
$$= \left| \frac{dg}{dx}(x) \Delta x \cdot f_Y(y) \right|$$

Now we simply need to related the surface of the two pdfs:

$B = [x, x + \Delta x]$ same surfaces \propto $[y, y + \Delta y] = B'$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$
$$\xLeftrightarrow{\Delta y \rightarrow 0} |f_Y(y) \cdot \Delta y| = \left| f_Y(y) \cdot \frac{dg}{dx}(x) \Delta x \right| = |f_X(x) \cdot \Delta x|$$
$$f_Y(y) \cdot \left| \frac{dg}{dx}(x) \right| |\Delta x| = f_X(x) \cdot |\Delta x|$$
$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left| \frac{dg}{dx}(x) \right|} = \frac{f_X(g^{-1}(y))}{\left| \frac{dg}{dx} g^{-1}(y) \right|}$$

Proof 34.7: ^[def. 34.37]

$$\mathbb{E}[\epsilon^T A \epsilon] \stackrel{\text{eq. (28.54)}}{=} \mathbb{E}[\text{tr}(\epsilon^T A \epsilon)]$$
$$\stackrel{\text{eq. (28.56)}}{=} \mathbb{E}[\text{tr}(A \epsilon \epsilon^T)]$$
$$= \text{tr}(\mathbb{E}[A \epsilon \epsilon^T])$$
$$= \text{tr}(A \mathbb{E}[\epsilon \epsilon^T])$$
$$= \text{tr}(A (\Sigma + \mu \mu^T))$$
$$= \text{tr}(A \Sigma) + \text{tr}(A \mu \mu^T)$$
$$\stackrel{\text{eq. (28.54)}}{=} \text{tr}(A \Sigma) + A \mu \mu^T$$

Proof 34.8: law 34.7

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}_X(x) = \sum_x x \cdot \sum_y \mathbb{P}_{X,Y}(x, y)$$
$$= \sum_x x \cdot \sum_y \mathbb{P}_{X|Y}(x|y) \cdot \mathbb{P}_Y(y)$$
$$= \sum_y \mathbb{P}_Y(y) \cdot \sum_x x \cdot \mathbb{P}_{X|Y}(x|y)$$
$$= \sum_y \mathbb{P}_Y(y) \cdot \mathbb{E}[X|Y] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$$

Examples

- Example 34.1 :**
- Toss of a coin (with head and tail): $\Omega = \{H, T\}$.
 - Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$
 - A cubic die: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
 - The positive integers: $\Omega = \{1, 2, 3, \dots\}$
 - The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$
- Example 34.2 :**
- 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 34.3 :** If the sample space is a die toss $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, the sample space may be that we are only told whether an even or odd number has been rolled:
- $$\mathcal{F} = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

Example 34.4 : If we are only interested in the subset $A \in \Omega$ of our experiment, then we can look at the corresponding generating σ -algebra $\sigma(A) = \{\emptyset, A, A^C, \Omega\}$.

- Example 34.5 :**
- open half-lines: $(-\infty, a)$ and (a, ∞) ,
 - union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
 - closed interval: $[a, b] = \overline{(-\infty, a) \cup (b, \infty)}$,
 - 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 $(a, b] = (-\infty, b] \cap (a, \infty)$,
 - every set containing only one real number: $\{a\} = \bigcap_{n=1}^{\infty} (a - \frac{1}{n}, a + \frac{1}{n})$,
 - every set containing finitely many real numbers: $\{a_1, \dots, a_n\} = \bigcup_{k=1}^n \{a_k\}$.

Example 34.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\}$$
$$\mathbb{P}(\{1, 2, 3\}) = \{2/3, 1/6, 1/6\}$$
$$\tilde{\mathbb{P}}(\{1, 2, 3\}) = \{1/3, 1/3, 1/3\}$$

Example 34.7 :

add example fat book p.1280

add example prob th book 4

Example 34.8 ??: Let $X, Y \stackrel{\text{ind.}}{\sim} \mathcal{N}(0, 1)$.

Question: proof that:

$$U = X + Y \quad V = X - 1$$

are indepdent and normally distributed:

$$h(u, v) = \begin{cases} h_1(u, v) = \frac{u+v}{2} \\ h_2(u, v) = \frac{u-v}{2} \end{cases} \quad J = \det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$
$$f_{U,V} = f_{X,Y}(\underline{x}, \underline{y}) \cdot \frac{1}{2}$$
$$\stackrel{\text{indp.}}{=} f_X(\underline{x}) \cdot f_Y(\underline{y})$$
$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$
$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\left\{ \left(\frac{u+v}{2} \right)^2 + \left(\frac{u-v}{2} \right)^2 \right\} / 2}$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{v^2}{4}}$$

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:

$$\mathbb{P}_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapt

Definition 35.1 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

$$\mathbb{P}(\omega) = 1 \iff \omega \text{ happens a.s.} \quad (35.1)$$

Definition 35.2 Probability Mass Function (PMF):

Definition 35.3 Discrete Random Variable (DVR): The set of possible values \bar{x} of \mathcal{X} is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \dots, 8\} \quad \mathcal{X} = \mathbb{N} \quad (35.2)$$

Definition 35.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \rightarrow [0, \infty)$ that satisfies:

Non-negativity: $f(x) \geq 0, \quad \forall x \in \mathbb{R}^n \quad (35.3)$

Normalization: $\int_{-\infty}^{\infty} f(x) dx \stackrel{!}{=} 1 \quad (35.4)$

Must be integrable (35.5)

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 ∞ . 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:

$$\mathbb{P}(X = x) = 0 \quad \text{a.s.}$$

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 35.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf (^[def. 35.4]) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx \quad (35.6)$$

Property 35.1 Zero Probability: If X is a continuous rrv (^[def. 35.5]), then:

$$\mathbb{P}(X = a) = 0 \quad \forall a \in \mathbb{R} \quad (35.7)$$

Property 35.2 Open vs. Closed Intervals: For any real numbers a and b , with $a < b$ it holds:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) \quad (35.8)$$

\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 35.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Proof 35.1: Property 35.1:

$$\mathbb{P}(X = a) = \lim_{\Delta x \rightarrow 0} \mathbb{P}(X \in [a, a + \Delta x])$$
$$= \lim_{\Delta x \rightarrow 0} \int_a^{a+\Delta x} f_X(x) dx = 0$$

Proof 35.2: Property 35.2:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b)$$
$$= \mathbb{P}(a < X < b) = \int_a^b f_X(x) dx$$

Definition 35.6 Support of a probability density function: The support of the density of a pdf $f_X(\cdot)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\text{supp}(\cdot) f_X := \{x \in \mathcal{X} | f_X(x) > 0\} \quad (35.9)$$

Note: this is not a rigorous definition.

Theorem 35.1 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 35.2 Identically Distributed: From theorem 35.1 it follows that to RV X and Y that have exactly the same pdf follow the same distribution. We say X and Y are **identically distributed**.

0.1. Cumulative Distribution Fuction

Definition 35.7 Cumulative distribution function (CDF): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The (cumulative) distribution function of a real-valued random variable X is the function given by:

$$F_X(x) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}$$

Property 35.3: Monotonically Increasing

$$x \leq y \iff F_X(x) \leq F_X(y) \quad \forall x, y \in \mathbb{R} \quad (35.10)$$

Upper Limit

$$\lim_{x \rightarrow \infty} F_X(x) = 1 \quad (35.11)$$

Lower Limit

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad (35.12)$$

Definition 35.8 CDF of a discret rv X: Let X be discret rv with pdf \mathbb{P}_X , then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{t=-\infty}^x \mathbb{P}_X(t)$$

Definition 35.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X , then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \iff \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 35.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X , then it holds that:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) \quad (35.13)$$

Proof 35.3: ^[def. 35.9]:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^x f_X(t) dt$$

Proof 35.4: lemma 35.1:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$$

or by the fundamental theorem of calculus (theorem 23.2):

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t) dt = \int_a^b \frac{\partial F_X(t)}{\partial t} dt = [F_X(t)]_a^b$$

Theorem 35.2 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 35.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x} \in \mathcal{X}} \bar{x} \mathbb{P}_x(\bar{x}) \quad (35.14)$$

Definition 35.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{x} \in \mathcal{X}} \bar{x} f_x(\bar{x}) d\bar{x} \quad (35.15)$$

Law 35.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \quad (35.16)$$

Property 35.4 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:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \quad (35.17)$$

Thus \mathbb{E} is a **linear** operator (^[def. 23.15]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 35.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \quad (35.18)$$

Property 35.5 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:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}] \mathbf{B} \quad (35.19)$$

Proof 35.5: eq. (35.24):

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{p}_{X,Y}(x, y)xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} \mathbf{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathbf{p}_Y(y)y = \mathbb{E}[X] \mathbb{E}[Y]\end{aligned}$$

Definition 35.12 Autocorrelation/Crosscorrelation $\gamma(t_1, t_2)$: Describes the covariance (def. 35.16) between the two values of a stochastic process $(\mathbf{X}_t)_{t \in T}$ at different time points t_1 and t_2 .
 $\gamma(t_1, t_2) = \text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}] = \mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})]$ (35.20)

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \text{Cov}[\mathbf{X}_t, \mathbf{X}_t] \stackrel{\text{eq. (35.35)}}{=} \mathbf{V}[\mathbf{X}_t] \quad (35.21)$$

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

more to prob theory maybe

Definition 35.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{x} \mathbf{p}_x(\mathbf{x}) \quad (35.22)$$

Definition 35.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\mathbf{x} \in \mathcal{X}} \mathbf{x} f_x(\mathbf{x}) d\mathbf{x} \quad (35.23)$$

Law 35.2 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \quad (35.24)$$

Property 35.6 Translation and scaling: If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, c \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \quad (35.25)$$

Thus \mathbb{E} is a linear operator [def. 23.15].

Property 35.7

Affine Transformation of the Expectation:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mu + b \quad (35.26)$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 35.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \quad (35.27)$$

Property 35.8 Matrix×Expectation: If $\mathbf{X} \in \mathbb{R}^n$ is a random vector and $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\mathbf{B} \quad (35.28)$$

Proof 35.6: eq. (35.24):

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbf{p}_{X,Y}(x, y)xy \\ &\stackrel{??}{=} \sum_{x \in \mathcal{X}} \mathbf{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathbf{p}_Y(y)y = \mathbb{E}[X] \mathbb{E}[Y]\end{aligned}$$

2.2. The Variance

Definition 35.15 Variance $\mathbf{V}[X]$: 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 expected value $\mathbb{E}[X]$ and is defined by:

$$\mathbf{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] \stackrel{\text{see ?? 35.7}}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad (35.29)$$

2.2.1. Properties

Property 35.9 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic \Rightarrow we have no uncertainty \Rightarrow no variance:

$$\mathbf{V}[a] = 0 \quad \text{with} \quad a \in \mathbb{R} \quad (35.30)$$

see shift and scaling for proof ?? 35.8

Property 35.10 Shifting and Scaling:

$$\mathbf{V}[a + bX] = a^2 \sigma^2 \quad \text{with} \quad a \in \mathbb{R} \quad (35.31)$$

see ?? 35.8

Property 35.11

[proof 35.9]

Affine Transformation of the Variance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\mathbf{V}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbf{V}[\mathbf{X}]\mathbf{A}^T \quad (35.32)$$

Definition 35.16 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]\end{aligned} \quad (35.33)$$

see ?? 35.10

Definition 35.17 Covariance Matrix: The variance of a k -dimensional random vector $\mathbf{X} = (X_1 \dots X_k)$ is given by a p.s.d. eq. (28.109) 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:

$$\mathbf{V}[\mathbf{X}] := \Sigma(\mathbf{X}) := \text{Cov}[\mathbf{X}, \mathbf{X}] := \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \quad (35.34)$$

$$\begin{aligned}&= \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T \in [-\infty, \infty] \\ &= \begin{bmatrix} \mathbf{V}[X_1] & \text{Cov}[X_1, X_2] & \dots & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \text{Cov}[X_k, X_1] & \mathbf{V}[X_k] & \dots & \dots & \dots \\ \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)] & \dots & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)] & \dots & \dots & \dots \end{bmatrix}\end{aligned}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\text{Cov}[X, X] = \mathbf{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \quad (35.35)$$

add <http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/>

Property 35.12 Translation and Scaling:

$$\text{Cov}(a + bX, c + dY) = bd \text{Cov}(X, Y) \quad (35.36)$$

Property 35.13

Affine Transformation of the Covariance:

If $\mathbf{X} \in \mathbb{R}^n$ is a random vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^m$ then it holds:

$$\text{Cov}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbf{V}[\mathbf{X}]\mathbf{A}^T = \mathbf{A}\Sigma(\mathbf{X})\mathbf{A}^T \quad (35.37)$$

Definition 35.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}\text{Corr}[\mathbf{X}] &:= \frac{\text{Cov}[\mathbf{X}]}{\sigma_{X_1} \dots \sigma_{X_k}} \in [-1, 1] \\ &= \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases}\end{aligned} \quad (35.38)$$

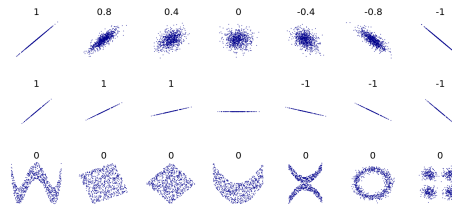


Figure 13: Several sets of (x, y) points, with their correlation coefficient

Law 35.3 Translation and Scaling:

$$\text{Corr}(a + bX, c + dY) = \text{sign}(b)\text{sign}(d)\text{Cov}(X, Y) \quad (35.39)$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 13), but not the slope of that relationship (middle row fig. 13) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 13 has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
- Zero covariance/correlation $\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0$ implies that there does not exist a linear relationship between the random variables X and Y .

Difference Covariance&Correlation

1. Variance is affected by scaling and covariance not ?? and law 35.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 35.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (??) is zero:

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] \\ &\stackrel{\text{eq. (35.24)}}{=} \mathbb{E}[X] \mathbb{E}[Y] - \mathbb{E}[X] \mathbb{E}[Y] = 0\end{aligned}$$

Zero covariance/correlation \nRightarrow independence

$\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0 \nRightarrow \mathbf{p}_{X,Y}(x, y) = \mathbf{p}_X(x)\mathbf{p}_Y(y)$
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:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}(X, X^2) = \mathbb{E}[X \cdot X^2] - \mathbb{E}[X] \mathbb{E}[X^2] \\ &= \mathbb{E}[X^3] - \mathbb{E}[X] \mathbb{E}[X^2] \stackrel{\text{eq. (35.63)}}{=} 0 - 0 \cdot \mathbb{E}[X^2] \\ &\stackrel{\text{eq. (35.52)}}{=} 0\end{aligned}$$

 \Rightarrow the relationship between Y and X must be non-linear.

Definition 35.19 Quantile: Are specific values q_α in the range [def. 23.10] of a random variable X that are defined as the value for which the cumulative probability is less then q_α with probability $\alpha \in (0, 1)$:

$$q_\alpha : \mathbb{P}(X \leq x) = F_X(q_\alpha) = \alpha \quad \xrightarrow{F \text{ invert.}} \quad q_\alpha = F_X^{-1}(\alpha) \quad (35.40)$$

add figure

3. Proofs

Proof 35.7: eq. (35.29)

$$\begin{aligned}\mathbf{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &\stackrel{\text{Property 35.6}}{=} \mathbb{E}[X^2] - 2\mathbb{E}[X] \mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2\end{aligned}$$

Proof 35.8: Property 35.10

$$\begin{aligned}\mathbf{V}[a + bX] &= \mathbb{E}[(a + bX - \mathbb{E}[a + bX])^2] \\ &= \mathbb{E}[(a + bX - a - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[(bX - b\mathbb{E}[X])^2] \\ &= \mathbb{E}[b^2(X - \mathbb{E}[X])^2] \\ &= b^2 \mathbb{E}[(X - \mathbb{E}[X])^2] = b^2 \sigma^2\end{aligned}$$

Proof 35.9: Property 35.11

$$\begin{aligned}\mathbf{V}(\mathbf{A}\mathbf{X} + b) &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^2] + 0 = \\ &= \mathbb{E}[(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])(\mathbf{A}\mathbf{X} - \mathbb{E}[\mathbf{A}\mathbf{X}])^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{A}^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{A}^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] \mathbf{A}^T = \mathbf{A}\mathbf{V}[\mathbf{X}]\mathbf{A}^T\end{aligned}$$

Proof 35.10: eq. (35.33)

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Discrete Distributions

Definition 35.20 Multivariate Distribution: the variate refers to the number of input variables i.e. a m-variate distribution has m-input variables whereas a uni-variate distribution has only one.

Dimensional vs. Multivariate

The dimension refers to the number of dimensions we need to embed the function. If the variables of a function are independent than the dimension is the same as the number of inputs but the number of input variables can also be less.

4.1. Bernoulli Distribution

Bern(p)

Definition 35.21 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 35.22 Bernoulli Distribution $X \sim \text{Bern}(\mathbf{p})$: X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter \mathbf{p} that signifies the success probability:

$$\mathbf{p}(x; \mathbf{p}) = \begin{cases} \mathbf{p} & \text{for } x = 1 \\ 1 - \mathbf{p} & \text{for } x = 0 \end{cases} \iff \begin{cases} \mathbb{P}(X = 1) = \mathbf{p} \\ \mathbb{P}(X = 0) = 1 - \mathbf{p} \end{cases} \\ = \mathbf{p}^x \cdot (1 - \mathbf{p})^{1-x} \quad \text{for } x \in \{0, 1\}$$

$$\mathbb{E}[X] = \mathbf{p} \quad (35.41) \quad \mathbb{V}[X] = \mathbf{p}(1 - \mathbf{p}) \quad (35.42)$$

4.2. Multinoulli/Categorical Distribution

Cat(n, p)

Definition 35.23 Multinoulli/Categorical Distribution $X \sim \text{Cat}(\mathbf{p})$: Is the generalization of the Bernoulli distribution^[def. 35.22] to a sample space^[def. 34.2] of k individual items $\{c_1, \dots, c_k\}$ with probabilities $\mathbf{p} = \{\mathbf{p}_1, \dots, \mathbf{p}_k\}$:

$$p(x = c_i | \mathbf{p}) = \mathbf{p}_i \iff p(x | \mathbf{p}) = \prod_i \mathbf{p}_i^{\delta[x=c_i]} \\ \sum_{j=1}^k \mathbf{p}_j = 1 \quad \mathbf{p}_j \in [0, 1] \quad \forall j = 1, \dots, k \quad (35.43) \\ \mathbb{E}[X] = \mathbf{p} \quad \mathbb{V}[X]_{i,j} = \Sigma_{i,j} = \begin{cases} \mathbf{p}_i(1 - \mathbf{p}_i) & \text{if } i = j \\ -\mathbf{p}_i\mathbf{p}_j & \text{if } i \neq j \end{cases}$$

Corollary 35.3

One-hot encoded Categorical Distribution: If we encode the k categories by a *sparse vectors*^[def. 28.70] with norm one:

$$\mathbb{B}_r^n = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^\top \mathbf{x} = \sum_{i=1}^n \mathbf{x}_i = 1 \right\} \\ \text{s.t.} \quad \mathbf{x}_j = \mathbf{e}_j \iff \mathbf{x} = \mathbf{c}_j \\ \text{then we can rewrite eq. (35.43) as:} \\ \mathbf{p}(\mathbf{x} | \mathbf{p}) = \prod_i \mathbf{x}_i \cdot \mathbf{p}_i \quad \sum_{j=1}^k \mathbf{p}_j = 1 \quad (35.44)$$

4.3. Binomial Distribution

B(n, p)

Definition 35.24 Binomial Coefficient:

The binomial coefficient occurs inside the binomial distribution^[def. 35.25] and signifies the different combinations/order that x out of n successes can happen see also ^[def. 33.5].

Definition 35.25 Binomial Distribution [proof 35.1]: Models the probability of exactly x success given a fixed number of n -Bernoulli experiments^[def. 35.21], where the probability of success, of a single experiment is given by \mathbf{p} :

$$\mathbf{p}(x) = \binom{n}{x} \mathbf{p}^x (1 - \mathbf{p})^{n-x} \quad \begin{matrix} n: \text{nb. of repetitions} \\ x: \text{nb. of successes} \\ \mathbf{p}: \text{probability of success} \end{matrix} \\ \mathbb{E}[X] = n\mathbf{p} \quad (35.45) \quad \mathbb{V}[X] = n\mathbf{p}(1 - \mathbf{p}) \quad (35.46)$$

Explanation 35.1. Lets consider a box of n balls consisting of black and white balls. If we want to know the probability of drawing first exactly x white and then exactly $n - x$ black balls we can simply calculate:

$$\underbrace{(\mathbf{p} \cdots \mathbf{p})}_{x\text{-times}} \cdot \underbrace{(q \cdots q)}_{n-x\text{-times}} = \mathbf{p}^x q^{n-x}$$

But if we do not care about the order we need to increase the probability by the different ways we can achieve this result by adding the binomial coefficient.

4.4. Geometric Distribution

Geom(p)

Definition 35.26 Geometric Distribution Geom(p): Models the probability of the number X of Bernoulli trials^[def. 35.21] until the first success

$$\mathbf{p}(x) = \mathbf{p}(1 - \mathbf{p})^{x-1} \quad \begin{matrix} x: \text{nb. of repetitions until first success} \\ \mathbf{p}: \text{success probability of single Bernoulli experiment} \end{matrix}$$

$$F(x) = \sum_{i=1}^x \mathbf{p}(1 - \mathbf{p})^{i-1} \text{ eq. (20.4)} = 1 - (1 - \mathbf{p})^x$$

$$\mathbb{E}[X] = \frac{1}{\mathbf{p}} \quad (35.47) \quad \mathbb{V}[X] = \frac{1 - \mathbf{p}}{\mathbf{p}^2} \quad (35.48)$$

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 $\mathbf{p}(x)$:

$$x \geq \frac{\mathbf{p}(x)}{1 - \mathbf{p}}$$

- $\log(1 - \mathbf{p}) \approx -\mathbf{p}$ for small \mathbf{p}

4.5. Poisson Distribution

Pois(λ)

Definition 35.27 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geq 0}$. It expresses the probability of a given number of events X occurring in a fixed interval if those events occur independently of the time since the last event.

$$\mathbf{p}(x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad \lambda > 0 \quad x \in \mathbb{Z}_{\geq 0} \quad (35.49)$$

Event Rate λ: describes the average number of events in a single interval.

$$\mathbb{E}[X] = \lambda \quad (35.50) \quad \mathbb{V}[X] = \lambda \quad (35.51)$$

Continuous Distributions

5.1. Uniform Distribution

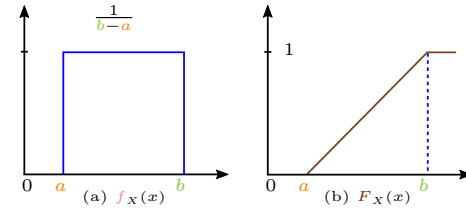
U(a, b)

Definition 35.28 Uniform Distribution $\mathcal{U}(a, b)$: Is probability distribution, where all intervals of the same length on the distribution's support^[def. 35.6] $\text{supp}(\mathcal{U}[a, b]) = [a, b]$ are equally probable/likely.

$$f(x) = \frac{1}{b - a} \mathbb{1}_{x \in [a; b]} = \begin{cases} \frac{1}{b - a} = \text{const} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases} \quad (35.52)$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x - a}{b - a} & \text{if } a \leq x \leq b \\ 1 & x > b \end{cases} \quad (35.53)$$

$$\mathbb{E}[X] = \frac{a + b}{2} \quad \mathbb{V}(X) = \frac{(b - a)^2}{12} \quad (35.54)$$



5.2. Exponential Distribution

exp(λ)

Definition 35.29 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution^[def. 35.26].

It describes the probability $f(x; \lambda)$ 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 .

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (35.55)$$

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (35.56)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \mathbb{V}(X) = \frac{1}{\lambda^2} \quad (35.57)$$

5.3. Laplace Distribution

Definition 35.30 Laplace Distribution:

$$\text{Laplace Distribution} \quad f(\mathbf{x}; \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|\mathbf{x} - \mu|}{\sigma}\right) \quad (35.58)$$

5.4. The Normal Distribution

N(μ, σ)

Definition 35.31 Normal Distribution $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$: Is a symmetric distribution where the population parameters μ, σ^2 are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X] = \mu \quad \mathbb{V}(X) = \sigma^2 \quad (35.59)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\} \quad (35.60)$$

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right\} du \quad (35.61)$$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \quad (35.62)$$

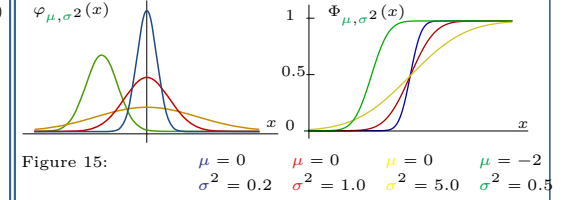


Figure 15: $\mu = 0, \sigma^2 = 0.2$ (green), $\mu = 0, \sigma^2 = 1.0$ (red), $\mu = 0, \sigma^2 = 5.0$ (yellow), $\mu = -2, \sigma^2 = 0.5$ (blue)

Property 35.14: $\mathbb{P}_X(\mu - \sigma \leq x \leq \mu + \sigma) = 0.66$

Property 35.15: $\mathbb{P}_X(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.95$

5.5. The Standard Normal distribution

N(0, 1)

Historic Problem: the cumulative distribution eq. (35.61) 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 35.32

Standard Normal Distribution $\mathbf{X} \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}[X] = 0 \quad \mathbb{V}(X) = 1 \quad (35.63)$$

$$f(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (35.64)$$

$$F(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}u^2} du \quad (35.65)$$

$$x \in \mathbb{R} \quad \text{or} \quad -\infty < x < \infty \\ \psi_X(u) = e^{-\frac{u^2}{2}} \quad \varphi_X(u) = e^{-\frac{u^2}{2}} \quad (35.66)$$

Corollary 35.4

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 α -quantile^[def. 35.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \quad \alpha \in (0, 1) \quad (35.67)$$

5.5.1. Calculating Probabilities

Property 35.16 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leq z) = \Phi(z) \quad (35.68)$$

$$\mathbb{P}(Z \leq -z) = \Phi(-z) = 1 - \Phi(z) \quad (35.69)$$

$$\mathbb{P}(-a \leq Z \leq b) = \Phi(b) - \Phi(-a) = \Phi(b) - (1 - \Phi(a)) \\ \stackrel{a=b=z}{=} 2\Phi(z) - 1 \quad (35.70)$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 35.1 Linear Transformation [proof 35.12]: Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the linear transformed r.v. Y given by the *affine transformation* $Y = a + bX$ with $a \in \mathbb{R}, b \in \mathbb{R}_+$ follows:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \quad (35.71)$$

Proposition 35.2 Standardization [proof 35.13]: 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:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0, 1) \quad (35.72)$$

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 35.3 Standardization of the CDF: Let $F_X(X)$ be the cumulative distribution function of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function $\Phi_Z(z)$ of the standardized random normal variable $Z \sim \mathcal{N}(0, 1)$ is related to $F_X(X)$ by:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad (35.73)$$

6. The Multivariate Normal distribution

Definition 35.33 Multivariate Normal/Gaussian: An \mathbb{R}^n -valued random variable $\mathbf{X} = (X_1 \dots X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma : \mathcal{L}\left(\sum_{i=1}^n \alpha_i X_i\right) = \mathcal{N}(\mu, \sigma^2) \quad \forall \alpha_i \in \mathbb{R} \quad (35.74)$$

(possible degenerated $\mathcal{N}(0, 0)$ for $\forall \alpha_i = 0$)

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 35.34 Multivariate Normal distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$: A k -dimensional random vector $\mathbf{X} = (X_1 \dots X_n)^T$ with $\mu = (\mathbb{E}[x_1] \dots \mathbb{E}[x_k])^T$ and $k \times k$ **p.s.d.** covariance matrix: $\Sigma := \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = [\text{Cov}[x_i, x_j], 1 \leq i, j \leq k]$ follows a k -dim multivariate normal/Gaussian distribution if its law^(def. 34.25) satisfies:

$$f_{\mathbf{X}}(X_1, \dots, X_k) = \mathcal{N}(\mu, \Sigma) \quad (35.75)$$

$$= \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)\right)$$

Normalisation

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{i\mathbf{u}^T \mu - \frac{1}{2}\mathbf{u} \Sigma \mathbf{u}\right\} \quad (35.76)$$

6.1. Joint Gaussian Distributions

Definition 35.35 Jointly Gaussian Random Variables: Two random variables X, Y both scalars or vectors, are said to be **jointly Gaussian** if the joint vector random variable $\mathbf{Z} = [X \ Y]^T$ is again a GRV.

Property 35.17 proof 35.16
Joint Independent Gaussian Random Variables: Let X_1, \dots, X_n be \mathbb{R} -valued *independent* random variables with laws $\mathcal{N}(\mu_i, \sigma_i^2)$. Then the law of $\mathbf{X} = (X_1 \dots X_n)$ is a (multivariate) Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \quad (35.77)$$

Corollary 35.5 Quadratic Form: If \mathbf{x} and \mathbf{y} are both independent GRVs $\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$ $\mathbf{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$ then they are jointly Gaussian^[def. 35.35] given by:

$$\mathbf{p}(\mathbf{x}, \mathbf{y}) = \mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{y}) \quad (35.78)$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\mathbf{x} - \mu_x)^T \Sigma_x^{-1}(\mathbf{x} - \mu_x) + (\mathbf{y} - \mu_y)^T \Sigma_y^{-1}(\mathbf{y} - \mu_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}\left[(\mathbf{x} - \mu_x)^T \quad (\mathbf{y} - \mu_y)^T\right] \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_y^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mu_x \\ \mathbf{y} - \mu_y \end{bmatrix}\right)$$

$$\cong \exp -\frac{1}{2}(\mathbf{z} - \mu_z)^T \Sigma_z^{-1}(\mathbf{z} - \mu_z)$$

Property 35.18 Marginal Distribution of Multivariate Gaussian: Let $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ be a \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. The k -variate marginal distribution of the Gaussian indexed by a subset of the variables:

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad \mathbf{i}_j \in V \quad (35.79)$$

is given by:

$$\mathbf{X} = (X_{i_1} \dots X_{i_k})^T \sim \mathcal{N}(\mu_A, \Sigma_{AA}) \quad (35.80)$$

$$\Sigma = \begin{bmatrix} \sigma_{i_1, i_1}^2 & \sigma_{i_1, i_k}^2 \\ \vdots & \vdots \\ \sigma_{i_k, i_1}^2 & \sigma_{i_k, i_k}^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

6.2. Conditional Gaussian Distributions

Property 35.19 Conditional Gaussian Distribution: Let $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ be a \mathbb{R}^n valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. Suppose we take two disjoint subsets of V :

$$\mathbf{A} = \{i_1, \dots, i_k\} \quad \mathbf{B} = \{j_1, \dots, j_m\} \quad \mathbf{i}_l, \mathbf{j}_{l'} \in V$$

then the conditional distribution of the random vector \mathbf{X}_A , conditioned on \mathbf{X}_B given by $\mathbf{p}(\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B)$ is:

$$\mathbf{X}_A = (X_{i_1} \dots X_{i_k})^T \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B}) \quad (35.81)$$

$$\mu_{A|B} = \mu_A + \Sigma_{AB} \Sigma_{BB}^{-1}(\mathbf{x}_B - \mu_B)$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}$$

Note
Can be proofed using the matrix inversion lemma but is a very tedious computation.

Corollary 35.6 Conditional Distribution of Joint Gaussian's: Let \mathbf{X} and \mathbf{Y} be jointly Gaussian random vectors:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix}\right) \quad (35.82)$$

then the *marginal* distribution of \mathbf{x} conditioned on \mathbf{y} can be written as:

$$X \sim \mathcal{N}(\mu_{X|Y}, \Sigma_{X|Y})$$

$$\begin{bmatrix} \mu_{X|Y} \\ \Sigma_{X|Y} \end{bmatrix} = \begin{bmatrix} \mu_X + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \mu_Y) \\ \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T \end{bmatrix} \quad (35.83)$$

6.3. Transformations

Property 35.20 Multiples of Gaussian's **A** \mathbf{X} : Let $\mathbf{X} = (X_1 \dots X_n)^T \sim \mathcal{N}(\mu, \Sigma)$ be a \mathbb{R}^n valued Gaussian and let $\mathbf{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \in \mathbb{R}^d \quad \mathbf{Y} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T) \quad (35.84)$$

Property 35.21 Affine Transformation of GRVs: Let $\mathbf{y} \in \mathbb{R}^n$ be GRV, $\mathbf{A} \in \mathbb{R}^{d \times n}$, $\mathbf{b} \in \mathbb{R}^d$ and let \mathbf{x} be defined by the **affine transformation**^[def. 28.45]:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{b} \quad \mathbf{A} \in \mathbb{R}^{d \times n}, \mathbf{b} \in \mathbb{R}^d$$

Then \mathbf{x} is a GRV (see ?? 35.15).

Property 35.22 Linear Combination of jointly GRVs: Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ two jointly GRVs, and let \mathbf{z} be defined as:

$$\mathbf{z} = \mathbf{A}_x \mathbf{x} + \mathbf{A}_y \mathbf{y} \quad \mathbf{A}_x \in \mathbb{R}^{d \times n}, \mathbf{A}_y \in \mathbb{R}^{d \times m}$$

Then \mathbf{z} is GRV (see ?? 35.17).

Definition 35.36 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 35.37 Gamma Distribution $X \sim \Gamma(x, \alpha, \beta)$: Is a widely used distribution that is related to the exponential distribution, Erlang distribution, and chi-squared distribution as well as Normal distribution:

$$f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \quad (35.85)$$

$$\Gamma(\alpha) \stackrel{\text{eq. (23.81)}}{=} \int_0^\infty t^{\alpha-1} e^{-t} dt \quad (35.86)$$

with $\alpha, \beta \in \mathbb{R}_{>0}$

6.5. Chi-Square Distribution χ_k^2

Definition 35.38 Student' t-distribution:

6.7. Delta Distribution

Definition 35.39 The delta function $\delta(\mathbf{x})$: The delta/dirac function $\delta(\mathbf{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$$

for any integrable function f on \mathbb{R} .
Or alternatively by:

$$\delta(x - x_0) = \lim_{\sigma \rightarrow 0} \mathcal{N}(x|x_0, \sigma) \quad (35.87)$$

$$\approx \propto \mathbb{1}_{\{x=x_0\}} \quad (35.88)$$

Property 35.23 Properties of δ :

- Normalization:** The delta function integrates to 1: $\int_{\mathbb{R}} \delta(x) dx = \int_{\mathbb{R}} \delta(x) \cdot c_1(x) dx = c_1(0) = 1$ where $c_1(x) = 1$ is the constant function of value 1.
- Shifting:** $\int_{\mathbb{R}} \delta(x - x_0) f(x) dx = f(x_0) \quad (35.89)$

- Symmetry:** $\int_{\mathbb{R}} \delta(-x) f(x) dx = f(0)$
- Scaling:** $\int_{\mathbb{R}} \delta(\alpha x) f(x) dx = \frac{1}{|\alpha|} f(0)$

Note

- In mathematical terms δ is not a function but a **generalized 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. (35.87) would be a non-differentiable/discret form of the dirac measure.

Definition 35.40 Heaviside Step Function:

$$H(x) := \frac{d}{dx} \max\{x, 0\} \quad x \in \mathbb{R}_{\neq 0} \quad (35.90)$$

or alternatively:

$$H(x) := \int_{-\infty}^x \delta(s) ds \quad (35.91)$$

Proofs

Proof 35.11 Definition 35.25: Consider a sequence of n random $\{X_i\}_{i=1}^n$ Bernoulli experiments^[def. 35.22] with success probability p .

Define the r.v. Y_n to be the sum of the n Bernoulli variables:

$$Y_n = \sum_{i=1}^n X_i \quad n \in \mathbb{N}$$

i.e. the total number of successes. Now lets calculate the probability density function f_n of Y_n . First let $(x_1 \dots x_n) \in \{0, 1\}^n$ and let $y = \sum_{i=1}^n x_i$ a bit sting of zeros and ones, with one occuring y times.

$$\mathbb{P}((X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)) = (\underbrace{p \dots p}_y) \cdot (\underbrace{q \dots q}_{n-y\text{-times}}) = p^y (1-p)^{n-y}$$

However we need to take into account that there exists further realization $\mathbf{X} = \mathbf{x}$, that correspond to different orders of the elements in our two classes $\{0, 1\}$ which leads to $\frac{n!}{y!(n-y)!} = \binom{n}{y}$:

$$f_n(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y \in \{0, 1, \dots, n\}$$

Proof 35.12: proposition 35.1: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_Y(y) \stackrel{y \geq 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \leq \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

$$F_Y(y) \stackrel{y < 0}{=} \mathbb{P}_Y(Y \leq y) = \mathbb{P}(a + bX \leq y) = \mathbb{P}_X\left(X \geq \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right)$$

Differentiating both expressions w.r.t. y leads to:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{b} \frac{dF_X\left(\frac{y-a}{b}\right)}{dy} = \frac{1}{|b|} f_X(x) \left(\frac{y-a}{b}\right) & \text{if } y \geq 0 \\ \frac{1}{-b} \frac{dF_X\left(\frac{y-a}{b}\right)}{dy} = \frac{1}{|b|} f_X(x) \left(\frac{y-a}{b}\right) & \text{if } y < 0 \end{cases}$$

eq. (35.71)).

in order to prove that $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$ we simply plug f_X in the previous expression:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma}|b|} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\} = \frac{1}{\sqrt{2\pi\sigma}|b|} \exp\left\{-\frac{1}{2}\left(\frac{y - (a + b\mu)}{\sigma|b|}\right)^2\right\}$$

Proof 35.13: proposition 35.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{1}{\sigma} X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

eq. (35.71) $\mathcal{N}(a\mu + b, a^2\sigma^2) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0, 1)$

Proof 35.14: proposition 35.3: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leq x) \stackrel{-\mu}{=} \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Proof 35.15: Property 35.21 scalar case

Let $y \sim p(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\mathbf{x} = ay + b$ $a \in \mathbb{R}_+, b \in \mathbb{R}$

Using the Change of variables formula it follows:

$$p_x(\bar{x}) \stackrel{\text{eq. (34.46)}}{=} \frac{p_y(\bar{y})}{\left|\frac{dx}{dy}\right|} \left[\begin{array}{c} \bar{y}(\bar{x}) \\ \left|\frac{dx}{dy}\right| = a \end{array} \right]$$

$$\begin{aligned} \bar{y} &= \frac{\bar{x}-b}{a} = \frac{1}{a} \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\underbrace{\bar{x}-b}_a - \mu\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi a^2 \mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2} \left(\underbrace{\bar{x}-b-a\mu}_{\mu_x}\right)^2\right) \end{aligned}$$

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 35.16: ??

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{u}) &= \prod_{i=1}^n p_{X_i}(u_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right) \\ \varphi_{\mathbf{X}}(\mathbf{u}) &= \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_nu_n^2\right\} \\ &= \exp\left\{i\sum_n u_n\mu_n - \frac{1}{2}\sum_n \sigma_nu_n^2\right\} = \exp\left\{i\mathbf{u}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}\right\} \end{aligned}$$

Proof 35.17: Property 35.22

From Property 35.21 it follows immediately that \mathbf{z} is GRV

$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ with:

$\mathbf{z} = \mathbf{A}\boldsymbol{\xi}$ with $\mathbf{A} = \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}$ and $\boldsymbol{\xi} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$

Knowing that \mathbf{z} is a GRV it is sufficient to calculate $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ in order to characterize its distribution:

$$\begin{aligned} \mathbb{E}[\mathbf{z}] &= \mathbb{E}[\mathbf{A}_x x + \mathbf{A}_y y] = \mathbf{A}_x \mu_x + \mathbf{A}_y \mu_y \\ \mathbb{V}[\mathbf{z}] &= \mathbb{V}[\mathbf{A}\boldsymbol{\xi}] \stackrel{??}{=} \mathbf{A} \mathbb{V}[\boldsymbol{\xi}] \mathbf{A}^\top \\ &= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[x] & \text{Cov}[x, y] \\ \text{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix}^\top \\ &= \begin{bmatrix} \mathbf{A}_x & \mathbf{A}_y \end{bmatrix} \begin{bmatrix} \mathbb{V}[x] & \text{Cov}[x, y] \\ \text{Cov}[y, x] & \mathbb{V}[y] \end{bmatrix} \begin{bmatrix} \mathbf{A}_x^\top \\ \mathbf{A}_y^\top \end{bmatrix} \\ &= \mathbf{A}_x \mathbb{V}[x] \mathbf{A}_x^\top + \mathbf{A}_y \mathbb{V}[y] \mathbf{A}_y^\top \\ &\quad + \underbrace{\mathbf{A}_y \text{Cov}[y, x] \mathbf{A}_x^\top}_{=0 \text{by independence}} + \underbrace{\mathbf{A}_x \text{Cov}[x, y] \mathbf{A}_y^\top}_{=0 \text{by independence}} \\ &= \mathbf{A}_x \boldsymbol{\Sigma}_x \mathbf{A}_x^\top + \mathbf{A}_y \boldsymbol{\Sigma}_y \mathbf{A}_y^\top \end{aligned}$$

Note

Can also be proofed by using the normal definition of [def. 35.15] and tedious computations.

Proof 35.18: Equation (35.43) If $\mathbf{x} = c_i$ i.e. the outcome c_i has occurred then it follows:

$$\prod_j^k p_i^{\delta_{[x=c_i]}} = p_1^0 \cdots p_i^1 \cdots p_k^0 = 1 \cdots p_i \cdots 1 = p(\mathbf{x} = c_i | \mathbf{p})$$

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. (35.52)). Furthermore repeated calls to these RNG are independent, that is:

$$\begin{aligned} \mathbb{P}_{U_1, U_2}(u_1, u_2) &\stackrel{??}{=} \mathbb{P}_{U_1}(u_1) \cdot \mathbb{P}_{U_2}(u_2) \\ &= \begin{cases} 1 & \text{if } u_1, u_2 \in [a, b] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Question: using samples $\{u_1, \dots, u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discrete or continuous PDFs?

2. Inverse-transform Technique

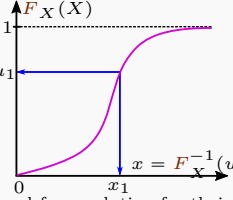
Idea

Can make use of section 1 and the fact that CDF are increasing functions ([def. 23.12]). **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 36.1 One Continuous Variable: **Given:** a desired continuous pdf f_X and uniformly distributed rn $\{u_1, u_2, \dots\}$:

- Integrate the desired pdf f_X in order to obtain the desired cdf F_X :

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad (36.1)$$

- Set $F_X(X) \stackrel{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0, 1]$.

- Invert this equation/find the inverse $F_X^{-1}(U)$ i.e. solve:

$$U = F_X(X) = F_X\left(\underbrace{F_X^{-1}(U)}_X\right) \quad (36.2)$$

- Plug in the uniformly distributed rn:

$$x_i = F_X^{-1}(u_i) \quad \text{s.t.} \quad x_i \sim f_X \quad (36.3)$$

Definition 36.2 Multiple Continuous Variable:

Given: a pdf of multiple rvs $f_{X,Y}$:

- Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (36.4)$$

- Use [def. 36.3] to first get a rv for y of $Y \sim f_Y(y)$.

- Then with this fixed y use [def. 36.3] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof 36.1: [def. 36.3]:

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. 23.12]).

Then for any $u \in [0, 1]$ there must exist a **unique** x s.t. $F_X(x) = u$.

Thus F_X must be invertible and we may write $x = F_X^{-1}(u)$.

Now let a arbitrary:

$$F_X(a) = \mathbb{P}(x \leq a) = \mathbb{P}(F_X^{-1}(U) \leq a)$$

Since F_X is strictly increasing:

$$\begin{aligned} \mathbb{P}\left(F_X^{-1}(U) \leq a\right) &= \mathbb{P}(U \leq F_X(a)) \\ &\stackrel{\text{eq. (35.52)}}{=} \int_0^{F_X(a)} 1 dt = F_X(a) \end{aligned}$$

Note

Strictly speaking we may not assume that a CDF is **strictly** increasing but we as all CDFs are weakly increasing ([def. 23.12]) we may always define an auxiliary function by its infimum:

$$\hat{F}_X^{-1} := \inf \{x | F_X(x) \geq 0\} \quad u \in [0, 1] \quad (36.5)$$

2.2. Discret Case

Idea

Given: a desired $U \sim \mathcal{U}[0, 1]$ and uniformly distributed rn $\{u_1, u_2, \dots\}$. **Goal:** given a uniformly distributed rn u determine k s.t.:

$$\sum_{i=1}^{k-1} \mathbb{P}(X = x_i) < U \leq \sum_{i=1}^k \mathbb{P}(X = x_i) \iff F_X(x_{k-1}) < u \leq F_X(x_k) \quad (36.6)$$

and return x_k .

Definition 36.3 One Discret Variable:

- Compute the CDF of p_X ([def. 35.8])

$$F_X(x) = \sum_{t=-\infty}^x p_X(t) \quad (36.7)$$

- Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i ($\hat{=}$ inversion) s.t.:

$$F_X(x_{k(i)-1}) < u_i \leq F_X(x_{k(i)}) \quad \forall u_i \quad (36.8)$$

Proof 36.2: ??: 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:

$$F_X(x_{k-1}) < u \leq F_X(x_k) \quad (36.9)$$

Now observe that:

$$\begin{aligned} u &\leq F_X(x_k) = F_X(x_{k-1}) + p_X(x_k) \\ \Rightarrow F_X(x_{k-1}) &< u \leq F_X(x_{k-1}) + p_X(x_k) \end{aligned}$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\begin{aligned} \mathbb{P}(U \in [F_X(x_{k-1}), F_X(x_k)]) &= \int_{F_X(x_{k-1})}^{F_X(x_k)} p_U(t) dt \\ &= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1 dt = F_X(x_k) - F_X(x_{k-1}) = p_X(x_k) \end{aligned}$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X .

Definition 36.4

Multiple Continuous Variables (Option 1):

Given: a pdf of multiple rvs $p_{X,Y}$:

- Use the product rule (??) in order to decompose $p_{X,Y}$:

$$p_{X,Y} = p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y) \quad (36.10)$$

- Use ?? to first get a rv for y of $Y \sim p_Y(y)$.

- Then with this fixed y use ?? again to get a value for x of $X \sim p_{X|Y}(x|y)$.

Definition 36.5

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

$$\begin{aligned} p_Z(1) &= p_{X,Y}(1, 1), p_Z(2) = p_{X,Y}(1, 2), \dots \\ \dots, p_Z(N_x \cdot N_y) &= p_{X,Y}(N_x, N_y) \end{aligned}$$

Then simply apply ?? to the auxillary pdf p_Z

- Use the product rule (??) in order to decompose $f_{X,Y}$:

$$f_{X,Y} = f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad (36.11)$$

- Use [def. 36.3] to first get a rv for y of $Y \sim f_Y(y)$.

- Then with this fixed y use [def. 36.3] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Also examples see comment in code text

3. Monte Carlo Methods

3.1. Monte Carlo (MC) Integration

Integration methods s.a. Simpson integration ([def. 31.34]) suffer heavily from the curse of dimensionality. An n-order ([def. 31.31]) 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 36.6 Monte Carlo Integration:

$$3 + 4 \quad (36.12)$$

3.2. Rejection Sampling

3.3. Importance Sampling

Descriptive Statistics

1. Populations and Distributions

Definition 37.1 Population $\{x_i\}_{i=1}^N$:
Is the entire set of entities from which we can draw sample.

Definition 37.2 Families of Probability Distributions p_θ :
Are probability distributions that vary only by a set of hyper parameters θ [def. 37.1].

Definition 37.3 Population/Statistical Parameter θ :
Are the parameters defining families of probability distributions[def. 37.2]

Explanation 37.1 (Definition 37.1). *Such hyper parameters are often characterized by populations following a certain family of distributions with the help of a statistic. Hence they are called population or statistical parameters.*

1.1. Characteristics of Populations

Definition 37.4 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (37.1)$$

Definition 37.5 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (37.2)$$

Note
The population variance and mean are equally to the mean derived from the true distribution of the population.

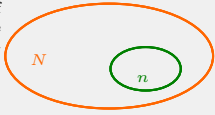
2. Sample Statistics

Definition 37.6 (Sample) Statistic: A statistic is a measurable function T that assigns a **single** value t to a sample of random variables or population:
$$t: \mathbb{R}^n \mapsto \mathbb{R} \quad t = T(X_1, \dots, X_n)$$

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

Definition 37.7 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:
$$X \sim p_X(\cdot; \theta) \quad (37.3)$$

how can we estimate the correct value of the parameter θ or some function of that parameter $\tau(\theta)$?

3.1. Point Estimates

Definition 37.8 (Point) Estimator $\hat{\theta}$:
Is a statistic[def. 37.6] that tries estimates an unknown parameter θ of an underlying family of distributions[def. 37.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution:
$$\hat{\theta} = t(\mathbf{x}_1, \dots, \mathbf{x}_n) \quad (37.4)$$

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 θ .
The most prevalent forms of interval estimation are:
• Confidence intervals (frequentist method).
• Credible intervals (Bayesian method).

3.1.1. Empirical Mean

Definition 37.9 Sample/Empirical Mean \bar{x} :
The sample mean is an estimate/statistic of the population mean[def. 37.4] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i \quad (37.5)$$

Corollary 37.1 [proof 37.1]
Unbiased Sample Mean:
The sample mean estimator is unbiased:
$$\mathbb{E}[\hat{\mu}_X] = \mu \quad (37.6)$$

Corollary 37.2 [Proof 37.2]
Variance of the Sample Mean:
The variance of the sample mean estimator is given by:
$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n} \sigma_X^2 \quad (37.7)$$

3.1.2. Empirical Variance

Definition 37.10 Biased Sample Variance:
The sample variance is an estimate/statistic of the population variance[def. 37.5] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (37.8)$$

Definition 37.11 [proof 37.3]
(Unbiased) Sample Variance:
The unbiased form of the sample variance[def. 37.10] is given by:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2 \quad (37.9)$$

Definition 37.12 Bessel's Correction: The factor $\frac{n}{n-1}$ (37.10)
is called Bessel's correction. Multiplying the uncorrected population variance eq. (37.8) 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 37.13 Interval Estimator $\hat{\theta}$:
Is an estimator that tries to bound an unknown parameter θ of an underlying family of distributions[def. 37.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution.
Let $\theta \in \Theta$ and define two point statistics[def. 37.6] g and h then an interval estimate is defined as:
$$\mathbb{P}(L_n < \theta < U_n) = \gamma \quad \forall \theta \in \Theta \quad \begin{matrix} L_n = g(\mathbf{x}_1, \dots, \mathbf{x}_n) \\ U_n = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \end{matrix} \quad \gamma \in [0, 1] \quad (37.11)$$

Statistical Tests

4. Parametric Hypothesis Testing

Definition 37.14 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 37.15 Null Hypothesis H_0 :
A null hypothesis H_0 is an *assumption* on a population parameter[def. 37.3] θ :

$$H_0: \theta = \theta_0 \quad (37.12)$$

Note
Often, a null hypothesis cannot be verified, but can only be falsified.

Definition 37.16 Alternative Hypothesis H_A/H_1 :
The alternative hypothesis H_1 is an *assumption* on a population parameter[def. 37.3] θ that is opposite to the null hypothesis.

$$H_A: \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \quad (37.13)$$

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. 37.6] as this is easier to use.

Definition 37.17 Test Statistic/Testing Parameter T :
Is a sample statistic[def. 37.6] used for hypothesis tests in order to give evidence for or against a hypothesis:
$$t_n = T(D_n) = T(\{X_1, \dots, X_n\}) \quad (37.14)$$

4.3. Sampling Distribution

Definition 37.18 Null Distribution/Sampling Distribution under $T_{\theta_0}(t)$ H_0 :
Let $D_n = \{X_1, \dots, X_n\}$ be a random sample from the true population p_{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*:
$$t \sim T_{\theta_0} = T(t|H_0 \text{ true}) \quad X_i \sim p_{\text{pop}} \quad (37.15)$$

4.4. The Critical Region

Given a sample $D_n = \{X_1, \dots, X_n\}$ of the true population p_{pop} how should we decide whether the null hypothesis should be rejected or not?
Idea: let \mathcal{T} be the set of all possible values that the sample statistic T can map to. Now lets split \mathcal{T} in two disjoint sets \mathcal{T}_0 and \mathcal{T}_1 :
$$\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \quad \mathcal{T}_0 \cap \mathcal{T}_1 = \emptyset$$

• 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 37.19 Critical/Rejection Region \mathcal{T}_1 :
Is the set of all values of the test statistic[def. 37.17] t_n that causes us to reject the Null Hypothesis in favor for the alternative hypothesis H_A :
$$K = \mathcal{T}_1 = \{T: H_0 \text{ rejected}\} \quad (37.16)$$

Definition 37.20 Acceptance Region \mathcal{T}_0 :
Is the region where we accept the null hypothesis H_0 .
$$\mathcal{T}_0 = \{T: H_0 \text{ accepted}\} \quad (37.17)$$

Definition 37.21 Critical Value c :
Is the value of the *critical region* $c \in \mathcal{T}_1$ which is closest to the *region of acceptance*[def. 37.20]:

4.5. Type I&II Errors

Definition 37.22 False Positive **Type I Error:**
Is the rejection of the null hypothesis H_0 , even-tough it is true
$$\text{Test rejects } H_0 | H_0 \text{ true} \iff t_n \in \mathcal{T}_1 | H_0 \text{ true} \quad (37.18)$$

Definition 37.23 False Negative **Type II Error:**
Is the acceptance of a null hypothesis H_0 , even-tough its false:
$$\text{Test accepts } H_0 | H_A \text{ true} \iff t_n \in \mathcal{T}_0 | H_A \text{ true} \quad (37.19)$$

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 Θ_1 (and thus the smaller Θ_0) the more likely it is to accept the alternative.
Idea: take the position of the adversary and choose Θ_1 so small that $\theta \in \Theta_1$ has only a small *probability* of occurring.

Definition 37.24 (Statistical) Significance α :
A study's defined significance level α denotes the probability to incur a *Type I Error*[def. 37.22]:
$$\mathbb{P}(t_n \in \mathcal{T}_1 | H_0 \text{ true}) = \mathbb{P}(\text{test rejects } H_0 | H_0 \text{ true}) \leq \alpha \quad (37.20)$$

Definition 37.25 Probability Type II Error β :
A test probability to for a *false negative*[def. 37.23] is defined as:
$$\beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_0 | H_1 \text{ true}) = \mathbb{P}(\text{test accepts } H_0 | H_1 \text{ true}) \quad (37.21)$$

Definition 37.26 (Statistical) Power $1 - \beta$:
A study's power $1 - \beta$ denotes a tests probability for a *true positive*:
$$1 - \beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_1 | H_1 \text{ true}) = \mathbb{P}(\text{test rejects } H_0 | H_1 \text{ true}) \quad (37.22)$$

Corollary 37.3 Types of Split:
The Critical region is chosen s.t. we incur a Type I Error with probability less than α , which corresponds to the type of the test[def. 37.16]:

$$\mathbb{P}(c_2 \leq X \leq c_1) \leq \alpha \quad \text{two-sided}$$

or
$$\mathbb{P}(c_2 \leq X) \leq \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leq c_1) \leq \frac{\alpha}{2}$$

$$\mathbb{P}(c_2 \leq X) \leq \alpha \quad \text{one-sided}$$

$$\mathbb{P}(X \leq c_1) \leq \alpha \quad \text{one-sided}$$

	Truth	
	H_0 true	H_0 false
Decission		
H_0 accept	$1 - \alpha$	$1 - \beta$
H_0 rejected	α	β

4.7. P-Value

Definition 37.27 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:
$$p := \inf \{\alpha | t_n \in \mathcal{T}_1\} \quad t_n = T(X_1, \dots, X_n) \quad (37.24)$$

Explanation 37.2.
• 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 α .

5. Conducting Hypothesis Tests

- ① Select an appropriate test statistic^[def. 37.17] T .
- ② Define the null hypothesis H_0 and the alternative hypothesis H_1 for T .
- ③ Find the sampling distribution^[def. 37.18] $T_{\theta_0}(t)$ for T , given H_0 true.
- ④ Chose the significance level α
- 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. (37.6) and (37.7) it follows that the *mean of the sample* is distributed as:

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

thus the mean of the sample \bar{X}_n should equal the mean μ of the population. We now want to test the null hypothesis:

$$H_0 : \mu = \mu_0 \iff \bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n) \quad (37.25)$$

This is obviously only likely if the realization \bar{x}_n is close to μ_0 .

5.1.1. Z-Test σ known

Definition 37.28 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. 37.24] α if:
$$|z| \geq z_{1-\frac{\alpha}{2}} \iff z \leq z_{\frac{\alpha}{2}} \vee z \geq z_{1-\frac{\alpha}{2}}$$
$$\iff z \in \mathcal{T}_1 = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, \infty\right)$$
$$z \geq z_{1-\alpha} \iff z \in \mathcal{T}_1 = [z_{1-\alpha}, \infty)$$
$$z \leq z_{\alpha} = -z_{1-\alpha} \iff z \in \mathcal{T}_1 = (-\infty, -z_{\alpha}] = (\infty, -z_{1-\alpha}] \quad (37.26)$$

Notes

- Recall from ^[def. 35.19] and ^[cor. 35.4] that:
$$z_{\alpha} \stackrel{\text{i.e. } \alpha=0.05}{=} z_{0.05} = \Phi^{-1}(\alpha) \iff \mathbb{P}(Z \leq z_{0.05}) = 0.05$$
- $|z| \geq z_{1-\frac{\alpha}{2}}$ which stands for:
$$\mathbb{P}(Z \leq z_{0.05}) + \mathbb{P}(Z \geq z_{0.95}) = \mathbb{P}(Z \leq -z_{1-0.05}) + \mathbb{P}(Z \geq z_{0.95}) = \mathbb{P}(|Z| \geq z_{0.95})$$
can be rewritten as:
$$z \geq z_{1-\frac{\alpha}{2}} \vee -z \geq z_{1-\frac{\alpha}{2}} \iff z \leq -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$$
- One usually goes over to the standard normal distribution proposition 35.2 and thus test how far one is away from zero mean \Rightarrow Z-test.
- We thus inquire a Type I error with probability α and should be small i.e. 1%.

5.1.2. t-Test σ unknown

In reality we usually do not know the true σ 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:

$$T \sim t_{n-1} \quad (37.27)$$

Definition 37.29 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:
$$|t| \geq t_{n-1, 1-\frac{\alpha}{2}}$$
$$\iff t \in \mathcal{T}_1 = \left(-\infty, -t_{n-1, 1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1, 1-\frac{\alpha}{2}}, \infty\right)$$
$$t \geq t_{n-1, 1-\alpha} \iff t \in \mathcal{T}_1 = [t_{n-1, 1-\alpha}, \infty)$$
$$t \leq t_{n-1, \alpha} = -t_{n-1, 1-\alpha} \iff t \in \mathcal{T}_1 = (-\infty, -t_{n-1, \alpha}] = (\infty, -t_{n-1, 1-\alpha}]$$

Notes

- The t-distribution has fatter tails as the normal distribution \Rightarrow rare event become more likely
- For $n \rightarrow \infty$ 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 \Rightarrow n -samples and we are interested in one parameter μ .

5.2. Confidence Intervals

Now we are interested in the opposite of the critical region^[def. 37.19] namely the region of plausible values.

Definition 37.30 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:
$$I(D_n) = \{\theta_0 : T_n(D_n) \in \mathcal{T}_0\} = \{\theta_0 : H_0 \text{ is not rejected}\} \quad (37.28)$$

Corollary 37.4 : The confidence interval captures the unknown parameter θ with probability $1 - \alpha$:
$$\mathbb{P}_{\theta}(\theta \in I(D_n)) = \mathbb{P}(T_n(D_n) \in \mathcal{T}_0) = 1 - \alpha \quad (37.29)$$

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 37.1 ??: Let x be uniformly distributed on $[0, 1]$ (def. 35.28) with pmf $\mathsf{p}_X(x)$ then it follows:
 $\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathsf{p}_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y \mathsf{p}_Y(y) \Rightarrow x = \int_{-\infty}^y \mathsf{p}_Y(t) \, \mathrm{d}t = F_Y(x)$

Example 37.2 ??: Let

add <https://www.youtube.com/watch?v=WUUhTVIRagg>

Example 37.3 Family of Distributions: The family of normal distribution \mathcal{N} has two parameters $\{\mu, \sigma^2\}$

Example 37.4 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. 37.3) $\theta = \mu$ but assume its equal to θ_0 thus the null-and alternative hypothesis are:
 $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$

Example 37.5 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:
 $H_0 : \mathsf{p} = \mathsf{p}_0 = 0.1 \qquad \text{vs.} \qquad H_A : \mathsf{p} > 0.1$
We model the number of number of defective goods using the binomial distribution (def. 35.25)
 $X \sim \mathcal{B}(n, \mathsf{p}), n = 20 \qquad \mathbb{P}(X \geq x) = \sum_{k=x}^n \binom{n}{k} \mathsf{p}^k (1 - \mathsf{p})^{n-k}$
 $\sim \mathcal{T}(n, \mathsf{p})$
from this we find:
 $\mathbb{P}_{\mathsf{p}_0}(X \geq 4) = 1 - \mathbb{P}_{\mathsf{p}_0}(X \leq 3) = 0.13$
 $\mathbb{P}_{\mathsf{p}_0}(X \geq 5) = 1 - \mathbb{P}_{\mathsf{p}_0}(X \leq 4) = 0.04 \leq \alpha$
thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
 \Rightarrow throw away null hypothesis for the 5% niveau in favor to the alternative.
 \Rightarrow the 5% significance niveau is given by $K = \{5, 6, \dots, 20\}$

Note
If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x - 1)$

8. Proofs

Proof 37.1: (cor. 37.1):
 $\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n x_i\right] = \frac{1}{n} \mathbb{E}\left[\underbrace{\mu + \dots + \mu}_{1, \dots, n}\right]$

Proof 37.2: (cor. 37.2):
 $\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n x_i\right] \stackrel{\text{Property 35.10}}{=} \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n x_i\right]$
 $\frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n} \sigma^2$

Proof 37.3: definition 37.11:
 $\mathbb{E}[\hat{\sigma}_X^2] = \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right]$
 $= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2)\right]$
 $= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2\right]$
 $= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - 2n\bar{x} \cdot n\bar{x} + n\bar{x}^2\right]$
 $= \frac{1}{n-1} \mathbb{E}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right]$
 $= \frac{1}{n-1} \left[\sum_{i=1}^n \mathbb{E}[x_i^2] - n\mathbb{E}[\bar{x}^2]\right]$
 $= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\mathbb{E}[\bar{x}^2]\right]$
 $= \frac{1}{n-1} \left[\sum_{i=1}^n (\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right]$
 $= \frac{1}{n-1} \left[(n\sigma^2 + n\mu^2) - (\sigma^2 + n\mu^2)\right]$
 $= \frac{1}{n-1} \left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1} \left[(n-1)\sigma^2\right] = \sigma^2$

Stochastic Calculus

Stochastic Processes

Definition 38.1 Random/Stochastic Process 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 $\{t_1, t_2, \dots\}$. Therefore, the random process X can be written as a function: $X : \mathcal{T} \subseteq \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}^d \iff (t, \omega) \mapsto X(t, \omega) \quad (38.1)$
Definition 38.2 Sample path/Trajector/Realization: Is the <i>stochastic/noise signal</i> $r(\cdot, \omega)$ on the index set ^[def. 20.1] \mathcal{T} , that we obtain be sampling ω from Ω .
Notation Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$
Corollary 38.1 Strictly Positive Stochastic Processes: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called strictly positive if it satisfies: $X_t > 0 \quad \text{P-a.s.} \quad \forall t \in \mathcal{T} \quad (38.2)$
Definition 38.3 Random/Stochastic Chain is a collection of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ ^[def. 34.1] . The random variables are ordered by an associated index set ^[def. 20.1] \mathcal{T} and take values in the same mathematical <i>discrete state space</i> ^[def. 38.5] S , which must be measurable w.r.t. some σ -algebra ^[def. 34.6] Σ . Therefore for a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space (S, Σ) , the random <i>chain</i> X is a collection of S -valued random variables that can be written as: $X : \mathcal{T} \times \Omega \mapsto S \iff (t, \omega) \mapsto X(t, \omega) \quad (38.3)$
Definition 38.4 Index/Parameter Set Usually represents time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \dots\}$.
Definition 38.5 State Space Is the range/possible values of the random variables of a stochastic process ^[def. 38.1] and must be measurable ^[def. 34.7] w.r.t. some σ -algebra Σ .
Sample-vs. State Space Sample space ^[def. 34.2] 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 38.6 Sample path/Trajector/Realization: Is the <i>stochastic/noise signal</i> $r(\cdot, \omega)$ on the index set \mathcal{T} , that we obtain be sampling ω from Ω .
Notation Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$
1.1. Filtrations Definition 38.7 Filtration A collection $\{\mathcal{F}_t\}_{t \geq 0}$ of sub σ -algebras ^[def. 34.6] $\{\mathcal{F}_t\}_{t \geq 0} \in \mathcal{F}$ is called filtration if it is increasing: $\mathcal{F}_s \subseteq \mathcal{F}_t \quad \forall s \leq t \quad (38.4)$
Explanation 38.1 (Definition 38.7). <i>A filtration describes the flow of information i.e. with time we learn more information.</i>
Definition 38.8 Filtered Probability Space A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t \geq 0}$ is called a <i>filtered probability space</i> .

Definition 38.9 Adapted Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called adapted <i>to a filtration</i> \mathbb{F} if: $X_t \text{ is } \mathcal{F}_t\text{-measurable} \quad \forall t \quad (38.5)$ That is the value of X_t is observable at time t
Definition 38.10 Predictable Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called predictable <i>w.r.t. a filtration</i> \mathbb{F} if: $X_t \text{ is } \mathcal{F}_{t-1}\text{-measurable} \quad \forall t \quad (38.6)$ 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 $r = (r_k)_{k=1, \dots, T}$ should be predictable.
Corollary 38.2 : The amount of information of an adapted random process is increasing see example 38.1.
2. Martingales Definition 38.11 Martingales: A stochastic process $X(t)$ is a martingale on a <i>filtered probability space</i> $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ if the following conditions hold: ① Given $s \leq t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value: $\forall s \leq t \quad \mathbb{E}[X(t) \mathcal{F}_s] = X(s) \quad \text{a.s.} \quad (38.7)$ ② The expectation is finite: $\mathbb{E}[X(t)] < \infty \quad \forall t \geq 0 \quad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geq 0} \text{ adapted} \quad (38.8)$
Interpretation <ul style="list-style-type: none">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 38.12 Auto Covariance Describes the covariance ^[def. 35.16] between two values of a stochastic process $(\mathbf{X}_t)_{t \in \mathcal{T}}$ at different time points t_1 and t_2 . $\gamma(t_1, t_2) = \text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}] = \mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})] \quad (38.9)$
For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance: $\gamma(t, t) = \text{Cov}[\mathbf{X}_t, \mathbf{X}_t] \stackrel{\text{eq. (35.35)}}{=} \mathbb{V}[\mathbf{X}_t] \quad (38.10)$
Notes <ul style="list-style-type: none">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 <i>similar</i> 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 38.13 Auto Correlation Is the scaled version of the auto-covariance ^[def. 38.12] : $\rho(t_2 - t_1) = \frac{\text{Cov}[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}[(\mathbf{X}_{t_1} - \mu_{t_1})(\mathbf{X}_{t_2} - \mu_{t_2})]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} \quad (38.11)$
3. Different kinds of Processes

3.1. Markov Process Definition 38.14 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: $\mathbb{P}(X(t_{n+1}) \in B X(t_1), \dots, X(t_n)) = \mathbb{P}(X(t_{n+1}) \in B X(t_n))$ it thus follows for the <i>transition probability</i> – the probability of $X(t)$ lying in the set B at time t , given the value x of the process at time s : $\mathbb{P}(s, x, t, B) = P(X(t) \in B X(s) = x) \quad 0 \leq s < t \quad (38.12)$
Interpretation In order to predict the future only the current/last value counts.
Corollary 38.3 Transition Density: The transition probability of a continuous distribution \mathbf{p} can be calculated via: $\mathbb{P}(s, x, t, B) = \int_B \mathbf{p}(s, x, t, y) \, dy \quad (38.13)$
3.2. Gaussian Process Definition 38.15 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution: $X(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)) \quad \forall t \in T \quad (38.14)$
3.3. Diffusions Definition 38.16 Diffusion: Is a Markov Process ^[def. 38.14] for which it holds that: $\mu(t, X(t)) = \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[X(t + \Delta t) - X(t) X(t)] \quad (38.15)$ $\sigma^2(t, X(t)) = \lim_{t \rightarrow 0} \frac{1}{\Delta t} \mathbb{E}[(X(t + \Delta t) - X(t))^2 X(t)] \quad (38.16)$ <ul style="list-style-type: none">$\mu(t, X(t))$ is called drift$\sigma^2(t, X(t))$ is called diffusion coefficient
Interpretation There exist not discontinuities for the trajectories.
3.4. Brownian Motion/Wiener Process Definition 38.17 d-dim standard Brownian Motion/Wiener Process: Is an \mathbb{R}^d valued <i>stochastic process</i> ^[def. 38.1] $(W_t)_{t \in \mathcal{T}}$ starting at $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies: ① Normal Independent Increments: the increments are <i>normally distributed independent random variables</i> : $W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, (t_i - t_{i-1}) \mathbf{1}_{d \times d}) \quad \forall i \in \{1, \dots, T\} \quad (38.17)$ ② Stationary increments: $W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$ ③ Continuity: for <i>a.e.</i> $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous $\lim_{t \rightarrow 0} \frac{\mathbb{P}(W(t + \Delta t) - W(t) \geq \delta)}{\Delta t} = 0 \quad \forall \delta > 0 \quad (38.18)$ ④ Start $W(0) := W_0 = 0 \quad \text{a.s.} \quad (38.19)$
Notation <ul style="list-style-type: none">In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wiener process.However in some sources the Wiener process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 38.4 $W_t \sim \mathcal{N}(0, \sigma)$ [proof 38.4],[proof 38.5]: The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law $\mathbb{E}[W(t)] = \mu = 0 \quad (38.20)$ $\mathbb{V}[W(t)] = \mathbb{E}[W^2(t)] = \sigma^2 = t \quad (38.21)$
3.4.1. Properties of the Wiener Process Property 38.1 Non-Differentiable Trajectories: The sample paths of a Brownian motion are not differentiable: $\frac{dW(t)}{dt} = \lim_{t \rightarrow 0} \mathbb{E} \left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t} \right)^2 \right]$ $= \lim_{t \rightarrow 0} \frac{\mathbb{E}[W(t + \Delta t) - W(t)]}{\Delta t} = \lim_{t \rightarrow 0} \frac{\sigma^2}{\Delta t} = \infty$ <i>result</i> cannot use normal calculus anymore <i>solution</i> → Ito Calculus see section 39.
Property 38.2 Auto covariance Function: The auto-covariance ^[def. 38.12] for a Wiener process $\mathbb{E}[(W(t) - \mu t)(W(t') - \mu t')] = \min(t, t') \quad (38.22)$
Property 38.3: A standard Brownian motion is a Quadratic Variation
Definition 38.18 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as: $LV_{[a, b]}(f) = \sup_{\Pi \in S} \sum_{i=0}^{n_{\Pi}-1} f(x_{i+1}) - f(x_i) \quad (38.23)$ $S = \{\Pi\{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } ^{[\text{def. 31.12}]} \text{ of } [a, b]\}$ it is a measure of the (one dimensional) length of a function w.r.t. to the y-axis, when moving along the function. Hence it is a measure of the variation of a function w.r.t. to the y-axis.
Definition 38.19 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: $QV_{[a, b]}(f) = \sup_{\Pi \in S} \sum_{i=0}^{n_{\Pi}-1} f(x_{i+1}) - f(x_i) ^2 \quad (38.24)$ $S = \{\Pi\{x_0, \dots, x_{n_{\Pi}}\} : \Pi \text{ is a partition } ^{[\text{def. 31.12}]} \text{ of } [a, b]\}$
Corollary 38.5 Bounded (quadratic) Variation: The (quadratic) variation ^[def. 38.18] of a function is bounded if it is finite: $\exists M \in \mathbb{R}_+ : \quad LV_{[a, b]}(f) \leq M \quad (QV_{[a, b]}(f) \leq M) \quad \forall \Pi \in S \quad (38.25)$
Theorem 38.1 Variation of Wiener Process: Almost surely the total variation of a Brownian motion over an interval $[0, T]$ is infinite: $\mathbb{P}(\omega : LV(W(\omega)) < \infty) = 0 \quad (38.26)$
Theorem 38.2 Quadratic Variation of standard Brownian Motion: The quadratic variation of a standard Brownian motion over $[0, T]$ is finite: $\lim_{N \rightarrow \infty} \sum_{k=1}^N \left[W\left(k \frac{T}{N}\right) - W\left((k-1) \frac{T}{N}\right) \right]^2 = T$ with probability 1 $(dW(t))^2 = dt \quad (38.27)$
Corollary 38.6 : theorem 38.2 can also be written as: $(dW(t))^2 = dt \quad (38.28)$

3.4.2. Lévy's Characterization of BM

Theorem 38.3 [proof 38.7],[proof 38.8]
d-dim standard BM/Wiener Process by Paul Lévy:

An \mathbb{R}^d valued *adapted stochastic process*^[def. 38.1, 38.7] $(W_t)_{t \in T}$ with the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}_+}$, that satisfies:

- ① **Start**
$$W(0) := W_0 = 0 \quad \text{a.s.} \quad (38.29)$$
- ② **Continuous Martingale:** W_t is an a.s. *continuous* martingale^[def. 38.11] w.r.t. the filtration $(\mathcal{F}_t)_{t \in T}$ under \mathbb{P} .
- ③ **Quadratic Variation:**
$$W_t^2 - t \text{ is also an martingale} \iff QV(W_t) = t \quad (38.30)$$

is a standard Brownian motion^[def. 38.24].

Further Stochastic Processes

3.4.3. White Noise

understand script and add

Definition 38.20 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):
$$\mathbb{E}[\epsilon * [k]] = 0 \quad \forall k \in T_{\text{discret}} \quad (38.31)$$

- Zero autocorrelation^[def. 38.13] γ i.e. the signals of different times are in no-way correlated:
$$\begin{aligned} \gamma(\epsilon * [k], \epsilon * [k+n]) &= \mathbb{E}[\epsilon * [k] \epsilon * [k+n]^T] \\ &= \mathbb{V}[\epsilon * [k]] \delta_{\text{discret}}[n] \\ &\quad \forall k, n \in T_{\text{discret}} \end{aligned} \quad (38.32)$$

With $\delta_{\text{discret}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$

See proofs

Definition 38.21 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):
$$\mathbb{E}[\epsilon * (t)] = 0 \quad \forall t \in T_{\text{continuous}} \quad (38.33)$$

- Zero autocorrelation^[def. 38.13] γ i.e. the signals of different times are in no-way correlated:
$$\begin{aligned} \gamma(\epsilon * (t), \epsilon * (t+\tau)) &= \mathbb{E}[\epsilon * (t) \epsilon * (t+\tau)^T] \\ &\stackrel{\text{eq. (35.88)}}{=} \mathbb{V}[\epsilon * (t)] \delta(t-\tau) = \begin{cases} \mathbb{V}[\epsilon * (t)] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases} \end{aligned} \quad (38.34)$$

$$\forall t, \tau \in T_{\text{continuous}} \quad (38.35)$$

Definition 38.22 Homoscedastic Noise: Has constant variability for all observations/time-steps:

$$\mathbb{V}[\epsilon_{i,t}] = \sigma^2 \quad \begin{matrix} \forall t = 1, \dots, T \\ \forall i = 1, \dots, N \end{matrix} \quad (38.36)$$

Definition 38.23 Heteroscedastic Noise: Is noise whose variability may vary with each observation/time-step:

$$\mathbb{V}[\epsilon_{i,t}] = \sigma(i, t)^2 \quad \begin{matrix} \forall t = 1, \dots, T \\ \forall i = 1, \dots, N \end{matrix} \quad (38.37)$$

3.4.4. Generalized Brownian Motion

Definition 38.24 Brownian Motion: Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 38.17], and define:

$$X_t = \mu t + \sigma W_t \quad t \in \mathbb{R}_+ \quad \begin{matrix} \mu \in \mathbb{R} : \text{drift parameter} \\ \sigma \in \mathbb{R}_+ : \text{scale parameter} \end{matrix} \quad (38.38)$$

then $\{X_t\}_{t \in \mathbb{R}_+}$ is normally distributed with mean μt and variance $\sigma^2 X_t \sim \mathcal{N}(\mu t, \sigma^2 t)$.

Theorem 38.4 Normally Distributed Increments:

If $W(t)$ is a Brownian motion, then $W(t) - W(0)$ is a normal random variable with mean μt and variance $\sigma^2 t$, where $\mu, \sigma \in \mathbb{R}$. From this it follows that $W(t)$ is distributed as:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x - \mu t)^2}{2\sigma^2 t}\right\} \quad (38.39)$$

Corollary 38.7 : More generally we may define the process:
$$t \mapsto f(t) + \sigma W_t \quad (38.40)$$

which corresponds to a noisy version of f .

Corollary 38.8

Brownian Motion as a Solution of an SDE: A stochastic process X_t follows a BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} dX(t) &= \mu dt + \sigma dW(t) \\ X(0) &= 0 \end{aligned} \quad (38.41) \quad (38.42)$$

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 38.25 Geometric Brownian Motion:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 38.17] the stochastic process $\mathbf{S}_t^1 \triangleq \mathbf{S}^1(t)$ with drift parameter μ and scale σ satisfying the SDE:

$$\begin{aligned} d\mathbf{S}_t^1 &= \mathbf{S}_t^1 (\mu dt + \sigma dW_t) \\ &= \mu \mathbf{S}_t^1 dt + \sigma \mathbf{S}_t^1 dW_t \end{aligned} \quad (38.43)$$

is called geometric Brownian motion and is given by:

$$\mathbf{S}_t^1 = \mathbf{S}_0^1 \exp\left(\sigma W_t + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) \quad t \in \mathbb{R}_+ \quad (38.44)$$

Corollary 38.9 Log-normal Returns:

For a geometric BM we obtain log-normal returns:

$$\begin{aligned} \ln\left(\frac{S_t^1}{S_0^1}\right) &= \bar{\mu} t + \sigma W(t) \iff \bar{\mu} t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \\ \text{with} \quad \bar{\mu} &:= \mu - \frac{1}{2}\sigma^2 \end{aligned} \quad (38.45)$$

3.4.6. Locally Brownian Motion

Definition 38.26 Locally Brownian Motion:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 38.17] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = \mu(X(t), t) dt + \sigma(X(t), t) dW(t) \quad (38.46)$$

Note

A local Brownian motion is a generalization of a geometric Brownian motion.

3.4.7. Ornstein-Uhlenbeck Process

Definition 38.27 Ornstein-Uhlenbeck Process:

Let $\{W_t\}_{t \in \mathbb{R}_+}$ be a standard Brownian motion^[def. 38.17] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$dX(t) = -aX(t) dt + b\sigma dW(t) \quad a > 0 \quad (38.47)$$

3.5. Poisson Processes

Definition 38.28 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 38.29 Poisson Process: A Poisson Process with *rate* $\lambda \in \mathbb{R}_{\geq 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$, having a discrete *state space* $N = \{0, 1, 2, \dots\}$ and satisfies:

1. $X_0 = 0$
2. The increments follow a Poisson distribution^[def. 35.27]:
$$\mathbb{P}((X_t - X_s) = k) = \frac{\lambda(t-s)}{k!} e^{-\lambda(t-s)} \quad 0 \leq s < t < \infty \quad \forall k \in \mathbb{N}$$
3. No correlation of (non-overlapping) increments:
 $\forall t_0 < t_1 < \dots < t_n$: the increments are independent
$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}} \quad (38.48)$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete*, *positive* realizations in $\mathbb{N}_{\geq 0}$

Corollary 38.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}(X_{(t+\Delta t)} - X_t \neq 0) = \lambda \Delta t + o(\Delta t^2) \quad t \text{ small i.e. } t \rightarrow 0 \quad (38.49)$$

1. Thus the probability of an event happening during Δt is proportional to time period and the rate λ
2. The probability of two or more events to happen *during* Δt is of order $o(\Delta t^2)$ and thus extremely small (as *Deltat* is small).

Definition 38.30 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$dX_t = \lim_{\Delta t \rightarrow dt} (X_{(t+\Delta t)} - X_t) \quad (38.50)$$

Property 38.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\begin{aligned} \mathbb{P}(dX_t = 0) &= 1 - \lambda \\ \mathbb{P}(|dX_t| = 1) &= \lambda \end{aligned} \quad (38.51) \quad (38.52)$$

Proofs

Proof 38.1: eq. (38.15):

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)] \\ \text{induction } \mathbb{E}[x_{n-1}] &= \dots \mathbb{E}[x(0)] = 0 \end{aligned}$$

Thus in expectation the particles goes nowhere.

Proof 38.2: eq. (38.16):

Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned} \mathbb{E}[x(n)^2] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i(n)^2\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta]^2 \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2] \\ \text{ind. } &= \mathbb{E}[x_{n-1}^2] + \delta^2 = \mathbb{E}[x_{n-2}^2] + 2\delta^2 = \dots \\ &= \mathbb{E}[x(0)] + n\delta^2 = n\delta^2 \end{aligned}$$

as $n = \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x}$ it follows:

$$\sigma^2 = \mathbb{E}[x^2(n)] - \mathbb{E}[x(n)]^2 = \mathbb{E}[x^2(n)] = \frac{\delta^2}{\Delta x} t \quad (38.53)$$

Thus in expectation the particles goes nowhere.

Proof 38.3: eq. (38.34):

$$\begin{aligned} \gamma(\epsilon * [k], \epsilon * [k+n]) &= \text{Cov}[\epsilon * [k], \epsilon * [k+n]] \\ &= \mathbb{E}[(\epsilon * [k] - \mathbb{E}[\epsilon * [k]]) (\epsilon * [k+n] - \mathbb{E}[\epsilon * [k+n]])^T] \\ &\stackrel{\text{eq. (38.31)}}{=} \mathbb{E}[(\epsilon * [k]) (\epsilon * [k+n])] \end{aligned}$$

Proof 38.4: [cor. 38.4]:

Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the incremente over the interval $[s - s, t - s] = [0, t - s]$

Thus $B_t - B_s \sim B_{t-s} - B_0$
but as B_0 is a.s. zero by definition eq. (38.19) it follows:
$$B_t - B_s \sim B_{t-s} \quad B_{t-s} \sim \mathcal{N}(0, t-s)$$

Proof 38.5: [cor. 38.4]:

$$\begin{aligned} W(t) &= W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t) \\ \implies \mathbb{E}[X] &= 0 \quad \mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = t \end{aligned}$$

Proof 38.6: theorem 38.2:

$$\begin{aligned} \sum_{k=0}^{N-1} [W(t_k) - W(t_{k-1})]^2 &\quad t_k = k \frac{T}{N} \\ &= \sum_{k=0}^{N-1} X_k^2 \quad X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right) \\ &= \sum_{k=0}^{N-1} Y_k = n \left(\frac{1}{n} \sum_{k=0}^{N-1} Y_k\right) \quad \mathbb{E}[Y_k] = \frac{T}{N} \\ \text{S.L.} \stackrel{\text{L.L.N}}{=} n \frac{T}{n} &= T \end{aligned}$$

Proof 38.7: theorem 38.3 ②:

1. first we need to show eq. (38.7): $\mathbb{E}[W_t | \mathcal{F}_s] = W_s$
Due to the fact that W_t is \mathcal{F}_t measurable i.e. $W_t \in \mathcal{F}_t$ we know that:

$$\begin{aligned} \mathbb{E}[W_t | \mathcal{F}_t] &= W_t \\ \mathbb{E}[W_t | \mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s | \mathcal{F}_s] \\ &= \mathbb{E}[W_t - W_s | \mathcal{F}_s] + \mathbb{E}[W_s | \mathcal{F}_s] \\ &\stackrel{\text{eq. (38.54)}}{=} \mathbb{E}[W_t - W_s] + W_s \\ W_t - W_s &\stackrel{\sim \mathcal{N}(0, t-s)}{=} W_s \end{aligned} \quad (38.54)$$

2. second we need to show eq. (38.8): $\mathbb{E}[|X(t)|] < \infty$
$$\mathbb{E}[|W(t)|]^2 \stackrel{??}{\leq} \mathbb{E}[|W(t)|^2] = \mathbb{E}[W^2(t)] = t \leq \infty$$

Proof 38.8: theorem 38.3 ③: $W_t^2 - t$ is a martingale?
Using the binomial formula we can write and adding $W_s - W_s$:
$$W_t^2 = (W_t - W_s)^2 + 2W_s(W_t - W_s) + W_s^2$$

using the expectation:

$$\begin{aligned} \mathbb{E}[W_t^2 | \mathcal{F}_s] &= \mathbb{E}[(W_t - W_s)^2 | \mathcal{F}_s] + \mathbb{E}[2W_s(W_t - W_s) | \mathcal{F}_s] \\ &\quad + \mathbb{E}[W_s^2 | \mathcal{F}_s] \\ &\stackrel{\text{eq. (38.54)}}{=} \mathbb{E}[(W_t - W_s)^2] + 2W_s \mathbb{E}[(W_t - W_s)] + W_s^2 \\ &\stackrel{\text{eq. (38.21)}}{=} \mathbb{V}[W_t - W_s] + 0 + W_s^2 \\ &\quad = t - s + W_s^2 \end{aligned}$$

from this it follows that:

$$\mathbb{E}[W_t^2 - t | \mathcal{F}_s] = W_s^2 - s \quad (38.55)$$

understand why $\mathbb{E}[(W_t - W_s)^2 | \mathcal{F}_s] = \mathbb{E}[(W_t - W_s)^2]$

Examples

Example 38.1 :

Suppose we have a sample space of four elements: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. At time zero, we do not have any information about which ω has been chosen. At time $T/2$ we know whether we have $\{\omega_1, \omega_2\}$ or $\{\omega_3, \omega_4\}$. At time T , we have full information.

t

$$\mathcal{F} = \begin{cases} \{\emptyset, \Omega\} & t \in [0, T/2) \\ \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^\Omega & t = T \end{cases} \quad (38.56)$$

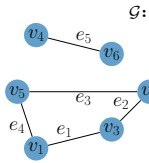
Thus, \mathcal{F}_0 represents initial information whereas \mathcal{F}_∞ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$.

Ito Calculus

Graph Theory

Definition 40.1 Graph

A graph \mathcal{G} is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a finite set of vertices \mathcal{V} ^[def. 40.4] and a multi set^[def. 19.3] of edges \mathcal{E} ^[def. 40.10].



Definition 40.2 Order $n = |\mathcal{V}|$: The order of a graph is the cardinality of its vertex set.

Definition 40.3 Size $m = |\mathcal{E}|$: The size of a graph is the number of its edges.

Corollary 40.1 n -Graph: Is a graph \mathcal{G} ^[def. 40.1] of order n .

Corollary 40.2 (p, q) -Graph: Is a graph \mathcal{G} ^[def. 40.1] of order p and size q .

1. Vertices

Definition 40.4 Vertices/Nodes \mathcal{V} : Is a set of entities of a graph connected and related by edges in some way:

Definition 40.5 Neighborhood $N(v)$: The neighborhood of a vertex $v_i \in \mathcal{V}$ is the set of all adjacent vertices:
$$N(v_i) = \{v_k \in \mathcal{V} : \exists e_k = \{v_i, v_j\} \in \mathcal{E}, \forall v_j \in \mathcal{E}\} \quad (40.1)$$

1.0.1. Adjacency Matrix

Definition 40.6 (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:

$$\mathbf{A}_{i,j} := \begin{cases} 1 & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (40.2)$$

Definition 40.7 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:

$$\mathbf{A}_{i,j} := \begin{cases} \theta_{ij} & \text{if } \exists e(i, j) \\ 0 & \text{otherwise} \end{cases} \quad (40.3)$$

Diagonal Elements

For a graph without self-loops the diagonal elements of the adjacency are all zero.

1.0.2. Degree Matrix

Definition 40.8 Degree of a Vertex δ : The degree of a vertex v is the cardinality of the neighborhood^[def. 40.5] – the number of adjacent vertices:

$$\deg(v_i) = \delta(v) = |N(v)| = \sum_{j=1}^{j < i} \mathbf{A}_{ij} \quad (40.4)$$

Definition 40.9 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:

$$\mathbf{D}_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (40.5)$$

2. Edges

Definition 40.10 Edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$: Represent some relation between edges^[def. 40.4] and are represented by two-element subset sets of the vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ and } v_j \text{ connected} \quad (40.6)$$

Proposition 40.1 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 40.11 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:
$$U \subseteq \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \quad (40.7)$$

4. Components

Definition 40.12 Component: A connected component of a graph \mathcal{G} is a *connected*^[def. 40.20] subgraph^[def. 40.11] 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 40.13 Walk: A walk of a graph \mathcal{G} as a sequence of vertices with corresponding edges:
$$W = \{v_k, v_{k+1}\}_k^K \in \mathcal{E} \quad (40.8)$$

Definition 40.14 Length of a Walk K : Is the number of edges of that Walk.

Definition 40.15 Path P : Is a walk of a graph \mathcal{G} where all visited vertices are distinct (no-repetitions).

Attention: Some use the terms walk for paths and simple paths for paths.

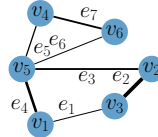
Definition 40.16 Cycle: Is a path^[def. 40.15] of a graph \mathcal{G} where the last visited vertex is the one from which we started.

6. Different Kinds of Graphs

7. Weighted Graph

Definition 40.17 Weighted Graph:

Is a graph \mathcal{G} where edges are associated with a weight:
$$\exists \theta_i := \text{weight}(e_i) \quad \forall e_i \in \mathcal{E}$$



8. Spanning Graphs

Definition 40.18 Spanning Graph:

Is a subgraph^[def. 40.11] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for which it holds:
$$U = \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \quad (40.9)$$



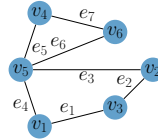
8.1. Minimum Spanning Graph

Definition 40.19 Minimum Spanning Graph: Is a spanning graph^[def. 40.18] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with minimal weights/distance of the edges.

9. Connected Graphs

Definition 40.20 (Weakly) Connected Graph:

Is a graph \mathcal{G} ^[def. 40.1] where there exists a path between any two vertices:
$$\exists P(v_i, \dots, v_j) \quad \forall v_i, v_j \in \mathcal{V} \quad (40.10)$$



Corollary 40.3 Strongly Connected Graph: A directed Graph^[def. 40.22] is called strongly connected if every nodes is *reachable* from every other node.

Corollary 40.4 Components of Connected Graphs: A connected Graph^[def. 40.20] consist of one component $c(\mathcal{G}) = 1$.

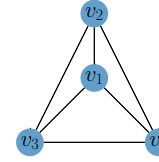
9.1. Fully Connected/Complete

Definition 40.21 Fully Connected/Complete Graph:

Is a connected graph \mathcal{G} ^[def. 40.20] where each node is connected to every other node.

$$\exists e \forall \{v_i, v_j\} \quad \forall v_i, v_j \in \mathcal{V} \quad (40.11)$$

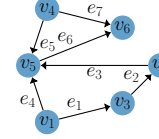
$$|\mathcal{V}| = \frac{1}{2}|\mathcal{V}|(|\mathcal{V}| - 1) \quad (40.12)$$



9.2. Directed Graphs

Definition 40.22 Directed Graph/Digraph (DG):

A directed graph \mathcal{G} is a graph where edges are direct arcs^[def. 40.23].



Definition 40.23 Directed Edges/Arcs: Represent some *directional* relationship between edges^[def. 40.4] and are represented by *ordered* two-element subset sets of vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ goes to } v_j \quad (40.13)$$

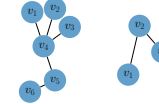
9.3. Trees And Forests

9.3.1. Acyclic Graphs

Definition 40.24 Acyclic Graphs: Are graphs^[def. 40.1] where no cycles^[def. 40.16] exist.

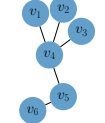
Definition 40.25 Forests:

Are acyclic graphs^[def. 40.24]:



Definition 40.26 Trees:

Are acyclic graphs^[def. 40.24] that are connected^[def. 40.20].



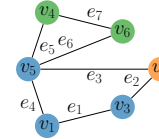
10. Graph Layering

Definition 40.27 Graph Layering:

Given a graph \mathcal{G} a layering of the graph is a partition of its node set \mathcal{V} ^[def. 40.4] into subsets

$$\{\mathcal{V}_1, \dots, \mathcal{V}_L\} \subseteq \mathcal{V}$$

$$\text{s.t.} \quad \mathcal{V} = \mathcal{V}_1 \cup \dots \cup \mathcal{V}_L \quad (40.14)$$



11. Bisection Algorithms

11.1. Local Approaches

11.2. Global Approaches

11.2.1. Spectral Decomposition

Definition 40.28 Graph Laplacian (Matrix) $\mathbf{L}(\mathcal{G})$: Given a graph with n vertices and m edges has a graph laplacian matrix defined as:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \quad l_{ij} := \begin{cases} -1 & \text{if } i \neq j \text{ and } e_{ij} \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } e_{ij} \notin \mathcal{E} \\ \deg(v_i) & \text{if } i = j \end{cases} \quad (40.15)$$

Corollary 40.5 title: