Definition 0.1 Predictor Variable: Is a variable used in regression to predict another variable.
Thus predictor variables are the independent variables if they are manipulated rather than just measured.

Definition 0.2 Output Variable:

Add

1. ANOVA

Definition 0.3 ANOVA: ANOVA stand for ANalysis Of VAriance. The goal of ANOVA is to compare the mean of different groups and determine if those means are significantly different from each other:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$$

It does this by comparing the variance with each group to the variance between groups.

Add picture:https://www.youtube.com/watch?v=Bmi7w-vKSCs min 11:30

1.1. ANOVA for linear Regression

Definition 0.4 ANOVA for regression: ANOVA answers the question of how much of the variability of our independent/output variable is explained by:

- the predictor/explanatory variables in our model/regression and
- how much is explained by other factors we are neglecting or are unable to account for

add sample variance cref

it does this by decomposing a measure of total sample variance (??)into two terms:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2 + \sum_{i=1}^{n}(y-\bar{y})^2$$

$$\iff \quad \|\mathbf{Y}-\bar{\mathbf{Y}}\|^2 = \underbrace{\|\hat{\mathbf{Y}}-\bar{\mathbf{Y}}\|^2}_{\text{explained}} - \underbrace{\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2}_{\text{unexplained}} \tag{0.1}$$

Definition 0.5 Explained Sum of Squares (ESS): The Explained Sum of Squares is a measure of the variability in the outcome variable that can be explained by the explanatory/predictor variables and thus by our model:

$$\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{0.2}$$

Definition 0.6 Residual Sum of Squares (RSS): The Residual Sum of Squares is a measure of the variability in the outcome variable that can not be explained by the predictor variables/our model:

$$\text{RSS} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n} r_i^2 \tag{0.3}$$

Definition 0.7 Mean Squared Error (MSE): In regression analysis often used to refer to the unbiased estimate of error variance.
Hence it is the mean sum of squares divided by the number of degrees of freedom.

Attention

Usually (also in regression analysis) Mean Squared Error (MSE) refers to the total mean squared prediction error.

Definition 0.8 Explained Mean Sum of Squares: Here we fixed the first predictor $\beta_1 = 1$ in order to represent the intercept and thus have to subtract 1 from the d.o.f.

$$\frac{\|\hat{\mathbf{Y}}-\bar{\mathbf{Y}}\|^2}{p-1}$$

Definition 0.9 Residual Mean Sum of Squares:

$$\frac{\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2}{n-p}$$

Definition 0.10 Total Mean Sum of Squares:

$$\frac{\|\mathbf{Y}-\bar{\mathbf{Y}}\|^2}{p-1+n-p} = \frac{\|\mathbf{Y}-\bar{\mathbf{Y}}\|^2}{n-1}$$

Proof 0.1: eq. (0.1):

add proof from: https://en.wikipedia.org/wiki/Partition_of_sums_of_squares

Null hypothesis

ANOVA can be used to test the null hypothesis if any of the predictor variables has an influence on the output variable i.e.

$$H_o : \beta_2 = \ldots = \beta_p = 0 \qquad \text{v.s.}$$
$$\exists j \in 2,\ldots,p \qquad \text{s.t.} \qquad H_A : \beta_j \neq 0$$

Non-Parametric Regression

Definition 1.1 Nonparametric Regression:

$$Y_i = m(x_i) + \epsilon_i \qquad i = 1,\ldots,n \qquad \begin{matrix} m(x) = \mathbb{E}[Y|x] \\ \epsilon_i \text{ i.i.d. } \mathbb{E}[\epsilon_i] = 0 \end{matrix} \tag{1.1}$$

Corollary 1.1 Nonparametric Regression Function: $m(x) = \mathbb{E}[Y|x]$ is the non-parametric regression function that we want to find and that needs to fulfill some smoothness conditions.

1. Kernel Smoothing
2. Local Linear Regression
3. Smoothing Splines

1. Kernel Regression

2. Local Nonparametric Regression

Definition 1.2 [proof 2.1],[code 2.1],[code 2.2]
Nardaya Watson Kernel Estimator:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{x-x_i}{h}\right)Y_i}{\sum_{j=1}^{n}\mathcal{K}\left(\frac{x-x_j}{h}\right)} = \frac{\sum_{i=1}^{n}\omega_i Y_i}{\sum_{j=1}^{n}\omega_j} =: \sum_{i=1}^{n} w_i Y_i \tag{1.2}$$

Note

- For $h \to \infty$ we would obtain a constant line $Y = m(x) = \mu$
- For $h \to 0$ we would obtain a function that fits each point of the training set perfectly.

Proposition 1.1 [proof 2.2]
Kernel Regression is Weighted Least Squares:
Kernel regression is equal to weighted least squares for fixed x:

$$\hat{m}(x) = \arg\min_{\beta} \sum_{i=1} \omega_i(x)(y_i-\beta)^2$$
$$= \sum_{i=1}^{n}\mathcal{K}\left(\frac{x-x_i}{h}\right)(y_i-\beta)^2 \qquad \forall x \tag{1.3}$$

2.1. The hat Matrix

Definition 1.3 [proof 2.3]
Smoother/Hat Matrix $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$**:**
Is the matrix that maps the true output values onto the predicted ones:

$$\mathbf{S}: \qquad \mathbb{R}^n \mapsto \mathbb{R}^n \tag{1.4}$$
$$(Y_1\cdots\cdots Y_n)^\intercal \mapsto (\hat{m}(x_1)\cdots\cdots\hat{m}(x_n))^\intercal =: \hat{\mathbf{m}}(x) = \hat{\mathbf{Y}}$$
$$\mathbf{S}_{j,i} = w_l(x_j) \qquad j,i \in \{1,\ldots,n\} \tag{1.5}$$
$$\mathbf{S}[Y_1\cdots\cdots Y_n]^\intercal = [\hat{m}(x_1)\cdots\cdots\hat{m}(x_n)]^\intercal = \hat{\mathbf{Y}} \tag{1.6}$$

2.1.1. Variance Estimates

Definition 1.4 [proof 2.4]
Covariance of the Estimator:

$$\text{Cov}(\hat{\mathbf{m}}(x)) = \sigma_\epsilon^2 \mathbf{S}\mathbf{S}^\intercal \tag{1.7}$$
$$\text{Cov}(\hat{m}(x_i),\hat{m}(x_j)) = \sigma_\epsilon^2(\mathbf{S}\mathbf{S}^\intercal)_{ij} \qquad \mathbb{V}[\hat{m}(x_i)] = \sigma_\epsilon^2(\mathbf{S}\mathbf{S}^\intercal)_{ii}$$

Definition 1.5 Degrees of Freedom:

$$\text{df} = \text{tr}(\mathbf{S})\sum_{i}^{n} S_{ii} \tag{1.8}$$

add derivation from muiltple linear reg. part

Proposition 1.2 Estimation of the Error Variance: We can estimate the true error variance σ_ϵ^2 by the residual sum of squares, scaled by the d.o.f.:

$$\hat{\sigma}_\epsilon^2 = \frac{\text{RSS}}{n-\text{df}} = \frac{\sum_{i=1}^{n}(Y_i-\hat{m}(x_i))^2}{n-\text{df}} \tag{1.9}$$

Our estimator $\hat{m}(x)$ is in the asymptotic limit normally distributed:

$$\hat{m}(x_i) \approx \mathcal{N}(\mathbb{E}[\hat{m}(x_i)], \mathbb{V}[\hat{m}(x_i)]) \tag{1.10}$$

s.t. we can derive *pointwise* confidence bounds:

$$I = \hat{m}(x_i) \pm 1.96 \cdot \widehat{\text{s.e.}}(\hat{m}(x_i)) \tag{1.11}$$

Attention: This confidence bound is not the true confidence interval for the expected value $\mathbb{E}[\hat{m}(x_i)]$ as we have to subtract the bias from $\mathbb{E}[\hat{m}(x_i)]$. This shifts the confidence interval:

$$I_{\text{true}} = I - \widehat{\text{bias}} \tag{1.12}$$

Definition 1.6 Erosion: Is the fact that we overestimate the valleys and underestimated the hills due to the bias.

2.2. The Optimal Bandwith

Corollary 1.2 [proof 2.5]
The Optimal Bandwith $h_{\text{opt}}(x)$**:**

$$h_{opt}(x) = n^{-1/5}\left(\frac{\sigma_\epsilon^2\int\mathcal{K}^2(z)\,\mathrm{d}z}{(m''(x)\int z^2\mathcal{K}(z)\,\mathrm{d}z)^2}\right)^{1/5} \tag{1.13}$$

Problem we do not know the real $m(x)$ thus we need some other way to estimate h_{opt} i.e. using lokernsdefinition 2.3.

Algorithm 1.1 Estimating $h_{opt}(x)$**:**

Choose some inital h_0
1: **for** $t = 1,\ldots,T$ **do**
2: estimate $m''(\cdot)$ and σ_ϵ^2
3: plug them into theoretical formulaeq. (1.13)
4: repeat with the new h_{new}
5: **end for**

2.3. Local Polynomial nonparametric Regression

Definition 1.7 Local Polynomial Regression:

$$\hat{\beta}(x) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n}\mathcal{K}\left(\frac{x-x_i}{h}\right)\left(Y_i-\beta_1-\beta_2(x_i-x)-\ldots-\beta_p(x_i-x)^{p-1}\right)^2 \tag{1.14}$$

Notes

- $p = 0$ kernel formula
- $p = 1$ intercept & slope = local linear regression i.e. *lowess* and *loess* (same thing, different implementation).

maybe add derivatives

3. Smoothing Splines

Definition 1.8 Smoothing Splines Problem:

$$\hat{m}(x) = \arg\min_{m(x)} \sum_{i=1}^{n}(Y_i-m(x_i))^2 + \lambda\int m''(z)^2\,\mathrm{d}z \tag{1.15}$$

Note

We penalize this problem with the curvature m'':

- $\lambda = 0$ we obtain interpolating natural cubic splines.
- $\lambda = \infty$ we obtain the least squares linear regression fit i.e. $m''(x) \equiv 0$.

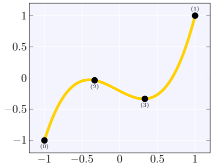
a large λ leads to a smooth function.

Theorem 1.1 Smoothing Spline Solution: The solution of eq. (1.15), (over infinite-dimensional function space) is a finite dimensional p.w. cubic polynomial – natural cubic spline[def. 1.9] and can be determined using least squares linear regression.

Definition 1.9 [code 2.4]
(Natural) Splines:
is a p.w. cubic polynomial[def. 24.6] of the form:

$$p(x) = \sum_{i=1}^{n}\beta_i b_j(x) \tag{1.16}$$

defined on the intervals $[x_i, x_{i+1})$ of a set of sorted and unique values $x_1 < x_2 < \ldots < x_n$, associated to n observations $\{y_i\}_{i=1}^{n}$.



- for the $n-1$ intervals we need to determined $(n-1)\cdot 4$ coefficients.
- we have 3 continuity conditions for the $n-2$ inner nodes:

$$\forall i = 2,\ldots,n-2 \qquad \begin{matrix} p(x_i) = p(x_{i+1}) \\ p'(x_i) = p'(x_{i+1}) \\ p''(x_i) = p''(x_{i+1}) \end{matrix} \tag{1.17}$$

- 2 *natural* conditions:

$$p''(x_1) = p''(x_n) = 0 \tag{1.18}$$

thus we have $4(n-1)-(3(n-2)+2) = n$ free parameters.

Corollary 1.3 smoothing splines solution:
problem eq. (1.15) can be written as lsq. problem:

$$\|\mathbf{Y}-b\beta\|^2 + \lambda\beta^\intercal\omega\beta \qquad \begin{matrix} \Omega_{jk} = \int\mathcal{B}_j''(z)\mathcal{B}_k''(z)\,\mathrm{d}z \\ \mathcal{B}_{:,j} = (\mathcal{B}_j(x_1)\cdots\cdots\mathcal{B}_j(x_n))^\intercal \end{matrix}$$
$$\hat{\beta} = (\mathcal{B}^\intercal\mathcal{B}+\lambda\Omega)^{-1}\mathcal{B}^\intercal\mathbf{Y} \qquad \hat{\beta}\in\mathbb{R}^n \tag{1.19}$$

Definition 1.10 Hat Matrix:

$$\hat{\mathbf{Y}} = (\hat{Y}_1\cdots\cdots\hat{Y}_n)^\intercal = \mathbf{S}_\lambda\mathbf{Y} \qquad \mathbf{S}_\lambda = \mathcal{B}(\mathcal{B}^\intercal\mathcal{B}+\lambda\Omega)^{-1}\mathcal{B}^\intercal \tag{1.20}$$

Cross Validation

1. Proofs

Proof 2.1 Derivation of Nardaya Watson Kernel Estimator[def. 1.2]: We know fromeq. (1.1) that we want to estimate:

$$\int_{\mathbb{R}} y f_{Y|X}(y|x)\,\mathrm{d}y = \frac{y f_{X,Y}(x,y)\,\mathrm{d}y}{f_X(x)}$$

plugin in the *univariate* and *bivariate* kernel densities:

$$\hat{f}_X(x) = \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{x-x_i}{h}\right)}{nh}$$
$$\hat{f}_{X,Y}(x,y) = \frac{\sum_{i=1}^{n}\mathcal{K}\left(\frac{x-x_i}{h}\right)\mathcal{K}\left(\frac{y-y_i}{h}\right)}{nh^2}$$

leads to the result.

Proof 2.2 Kernel Reg. and lsqproposition 1.1:

$$\frac{\partial}{\partial\beta}\arg\min_{\beta}\sum_{i=1}\omega_i(x)(y_i-\beta)^2 = 0$$
$$\Rightarrow \quad \sum_{i=1}^{n}\omega_i Y_i = \sum_{i=1}^{n}\omega_i\beta \quad \Rightarrow \quad \beta = \frac{\sum_{i=1}^{n}\omega_i Y_i}{\sum_{i=1}^{n}\omega_i}$$

Proof 2.3 The hat matrix[def. 1.3]: From:

$$\hat{m}(x_j) = \sum_{i=1}^{n} w_i(x_j)Y_i \qquad w_i(x_j) = \frac{\mathcal{K}\left(\frac{x_j-x_i}{h}\right)}{\sum_{k=1}^{n}\mathcal{K}\left(\frac{x_j-x_k}{h}\right)} \tag{2.1}$$

we can directly read of $\mathbf{S}$ as it maps Y_j to $\hat{m}(x_j)$.

Proof 2.4 Covariance Matrix[def. 1.4]:

$$\text{Cov}(\hat{\mathbf{m}}(\mathbf{x})) = \text{Cov}(\hat{\mathbf{Y}}) = \text{Cov}(\mathbf{S}\mathbf{Y}) = \mathbf{S}\text{Cov}(\mathbf{Y})\mathbf{S}^\intercal$$
$$= \mathbf{S}\text{Cov}(\epsilon)\mathbf{S}^\intercal = \sigma_\epsilon^2\mathbf{S}\mathbf{S}^\intercal$$

Proof 2.5 Optimal Bandwith Proof Sketch[cor. 1.2]: We want to find a $h(x)$ s.t.:

$$\min_{h(x)} \mathbb{E}\left[\left(\widehat{m}^h(x) - m(x)\right)^2\right] \qquad \forall x \in \texttt{dom}(m) \tag{2.2}$$

this can be decomposed into bias and variance i.e. lets assume a fixed x:

$$\begin{aligned}
&\mathbb{V}[\widehat{m}^h(x) - m(x)] = \\
&= \mathbb{E}\left[\left(\widehat{m}^h(x) - m(x)\right)^2\right] - \mathbb{E}\left[\widehat{m}^h(x) - m(x)\right]^2 \\
&= \mathbb{E}\left[\left(\widehat{m}^h(x) - m(x)\right)^2\right] - \left(\mathbb{E}\left[\widehat{m}^h(x)\right] - m(x)\right)^2 \\
&= \text{MSE} - (\text{Bias})^2 \\
&\frac{1}{nh}(\ldots)\sigma_\epsilon^2 = \text{MSE} - \left(h^2\left(\mathcal{K}, \sigma_x, m, m''\right)\right)^2 \\
&\frac{\mathrm{dMSE}}{\mathrm{d}h(x)} \overset{!}{=} 0
\end{aligned}$$

2. Examples

Example 2.1 Average Squared Residual(ASR)[def. 3.10]:

$$\text{ASR} = \frac{\text{RSS}}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}(x_i)\right)^2 \tag{2.3}$$

Example 2.2 Average Testing Error[def. 3.19]:

$$\text{ave}_n\text{MSEP} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}(x_i) + \sigma^2 \tag{2.4}$$

$$= \text{ave}_n\left(\text{Bias}^2(x_i) + \mathbb{V}(\widehat{m}(x_i))\right) + \sigma^2 \tag{2.5}$$

$$= \text{ave}_n\text{Bias}^2(x_i) + \text{ave}_n\mathbb{V}(\widehat{m}(x_i)) + \sigma^2 \tag{2.6}$$

3. Coding

Definition 2.1 `ksmooth`: is an implementation of eq. (1.2).

```
ksmooth(x, y, kernel = c("box", "normal"), bandwidth = 0.5)
```

Definition 2.2
locally weighted scatterplot smoothing Lowes??: is an implementation of eq. (1.2) that uses an adaptive bandwidth h:

```
lowess(x, y = NULL, f = 2/3, iter = 3, delta = 0.01 *
↪  diff(range(x)))
```

f: Adapts bandwith h s.t. always 2/3 of all the points are inside of the kernel.

Definition 2.3 `library(lokern)`
lokerns: automatically adapted local plug-in bandwidth function that adapts to non-homoscedastic variance:

```
lokerns(x, ...)
```

Definition 2.4 `library(mgcv)`
Smoothing Splines: is an implementation of eq. (1.16):

```
fm <- gam(y~s(x))
predict(fm)
```

Machine Learning Appendix

Model Assessment and Selection

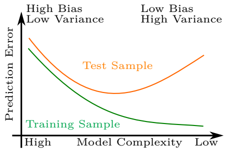
Definition 3.1 Statistical Inference: Is the process of deducing properties of an underlying probability distribution by mere analysis of data.

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.



1.1. Empirical Risk Minimization

2. Generalization Error

Definition 3.6
Generalization/Prediction Error (Risk): Is defined as the expected value of a loss function l of a given predictor m, for data drawn from a distribution $p_{\mathcal{X},\mathcal{Y}}$.

$$\begin{aligned} R_p(m) &= \mathbb{E}_{(\mathbf{x},y)\sim p}[l(y; m(\mathbf{x}))] = \int_{\mathcal{D}} p(\mathbf{x},y) l(y; m(\mathbf{x}))\,\mathrm{d}\mathbf{x}\,\mathrm{d}y \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p(\mathbf{x},y) l(y, m(\mathbf{x}))\,\mathrm{d}x\,\mathrm{d}y \\ &\overset{??}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}} l(y, m(\mathbf{x})) p(y|\mathbf{x}) p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}y \end{aligned} \tag{3.1}$$

Interpretation

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

Definition 3.7 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 $\mathbf{x}$:

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

Corollary 3.1 Note: [def. 3.6] $\iff$ [def. 3.7]:

$$R_p(m) = \mathbb{E}_{\mathbf{x}\sim p}[R_p(m,\mathbf{x})] = \int_{\mathcal{X}} p(\mathbf{x}) R_p(m,\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.2}$$

2.1. Expected Risk Minimizer

Definition 3.8 Expected Risk Minimizer (TRM) m^*:
Is the model m that minimizes the total expected risk:

$$m^* \in \arg\min_{m\in\mathcal{C}} \mathcal{R}(m) = \arg\min_{m\in\mathcal{C}} \mathbb{E}_p[l(y; m(\mathbf{x}))] \tag{3.3}$$

3. Empirical Risk

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.

Definition 3.9 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 $p_{\mathcal{X},\mathcal{Y}}(\mathbf{x}, y)$:

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

3.1. Empirical Risk Minimizer

Definition 3.10 Empirical Risk Minimizer (ERM) $\hat{m}$:
Is the model $\widehat{m}$ that minimizes the total empirical risk:

$$\hat{m} \in \arg\min_{m\in\mathcal{C}} \hat{\mathcal{R}}(m) = \arg\min_{m\in\mathcal{C}} n^{-1}\sum_{i=1}^n l(m(\mathbf{x}_i), y_i) \tag{3.4}$$

Questions

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

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

We hope that $\lim_{n\to\infty} \hat{\mathcal{R}}_n(m) = \mathcal{R}(m)$.

3.1.1. Squared Loss

Expected Squared Risk

Definition 3.11 Mean Squared Error (MSE):

$$\mathcal{R}(m) = \text{MSE}(x) = \mathbb{E}\left[(\widehat{m}(x) - m(x))^2\right] \tag{3.5}$$

Corollary 3.2 title:
add proof

$$\text{MSE}(x) = \text{Bias}^2(x) + \mathbb{V}(x) = (\mathbb{E}[\widehat{m}(x) - m(x)])^2 + \mathbb{V}(\widehat{m}(x)) \tag{3.6}$$

Definition 3.12
Integrated Means Squared Error (IMSE)/(MISE):
the integrated MSE or *Mean integrated sqquare error* (MISE) is defined as:

$$\text{IMSE} = \int_x \text{MSE}(x)\,\mathrm{d}x = \int_x \mathbb{E}\left[(\widehat{m}(x) - m(x))^2\right]\mathrm{d}x \tag{3.7}$$

Empirical Squared Risk

Definition 3.13
Mean/Average Squared Prediction Error (MSPE):
the empirical MSE or *Mean/Average Squared Error of Prediction* (MSEP)

$$\hat{\mathcal{R}}_n(m) = \text{ave}_n(\widehat{m})^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{m}(x_i) - m(x_i))^2 \tag{3.8}$$

Corollary 3.3 [proof 4.2]
MSEP for new observations: Given a new observation x_{new} distributed as:

$$Y_{\text{new}} = m(x_{\text{new}}) + \epsilon \qquad \epsilon \overset{\text{i.e.}}{\sim} \mathcal{N}(0, \sigma^2)$$

then it holds that:

$$\text{MSEP}(x_{\text{new}}) = \text{MSE}(x_{\text{new}}) + \sigma^2 \tag{3.9}$$

Explanation 3.1. *The mean squared error of prediction does not go to zero if $n \to \infty$ as it has an irreducible noise σ.*

Definition 3.14 [example 4.9],[proof 4.1]
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:

$$h^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \tag{3.10}$$

Cross Validation

Definition 3.15 Cross Validation: Is a model validation/assessment techniques in order to improve the model generalization performance.

Explanation 3.2. *Cross validation helps to increase the model ability to predict out of sample data.*

Definition 3.16 Labeled Data $\mathcal{D}/\mathcal{Z}$:

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

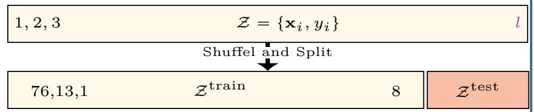
3.2. Training Set

Definition 3.17 Training Set $\mathcal{Z}^{\text{train}} \subset \mathcal{Z}$:
Is a part of the data on which we train our model $\widehat{m}$ in order to reduce the empirical

$$\mathcal{Z}^{\text{train}} = \left\{(\mathbf{x}_1^{\text{train}}, y_1^{\text{train}}), \dots, (\mathbf{x}_n^{\text{train}}, y_n^{\text{train}})\right\}$$

Definition 3.18 [example 2.1]
Training Error $\hat{\mathcal{R}}(\hat{f}, \mathcal{Z}^{\text{train}})$:
is the model that minimizes the empirical risk[def. 3.10] on the training data[def. 3.17]:

$$\begin{aligned} \widehat{m} &\in \arg\min_{\widehat{m}\in\mathcal{F}} \hat{\mathcal{R}}(\widehat{m}, \mathcal{Z}^{\text{train}}) \\ &= \arg\min_{\widehat{m}\in\mathcal{F}} n^{-1} \sum_{(\mathbf{x}_i, y_i)\in\mathcal{Z}^{\text{train}}} l(\widehat{m}(\mathbf{x}_i), y_i) \end{aligned} \tag{3.11}$$



3.3. Testing Set

Definition 3.19 [example 2.2]
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.

$$\mathcal{Z}^{\text{test}} = \left\{(\mathbf{x}_1^{\text{test}}, y_1^{\text{test}}), \dots, (\mathbf{x}_m^{\text{test}}, y_m^{\text{test}})\right\}$$

Definition 3.20 Test Error $\hat{\mathcal{R}}(\hat{f}, \mathcal{Z}^{\text{test}})$:
Is the error over the test set $\mathcal{Z}^{\text{test}}$ of a predictor $\widehat{m}$ that has been trained on the training set[def. 3.17]:

$$\hat{\mathcal{R}}(f, \mathcal{Z}^{\text{test}}) = m^{-1} \sum_{(\mathbf{x}_i, y_i)\in\mathcal{Z}^{\text{test}}} l(\widehat{m}(\mathbf{x}_i), y_i) \tag{3.12}$$

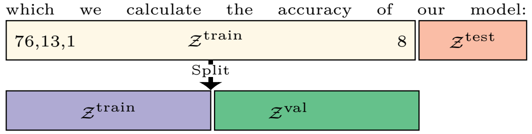
3.4. Validation Set

Definition 3.21 Validation Set $\mathcal{Z}^{\text{val}} \subset \mathcal{Z}^{\text{train}}$:
Is the part of the data that is used in order to select the our model $\widehat{m}$ from a given hypothesis class $\mathcal{F}$.

Explanation 3.3. *We want to select a model $\widehat{m}$ from $\mathcal{F}$ but in order to do so we need to determine the how well it predicts $\Rightarrow$ validation set.*

3.5. Validation Set/Split Once Approach

Definition 3.22 Hold out/Validation Set:
Split the data into a training set on which we train out model $\widehat{m}$ and a validation set on which we calculate the accuracy of our model:



Cons

- We do not use all information/data for training.
- We obtain a high variance estimate depending on the split.

Algorithm 3.1 Validation Set Approach:
Given: set of function classes $\mathcal{F}$ and a loss l
1: train the model on the training set:

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

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

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

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

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

Note: overfitting to the validation set

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

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

Definition 3.23
Leave One Out Cross-Validation (LOOCV):
Train n models on $n-1$ observations and use the left out observations for prediction:

$$\widehat{m}_{n-1}^{-i} \in \arg\min_{m\in\mathcal{F}} \frac{n-1}{n}\sum_{\substack{j=1\\ j\neq i}}^n l(y_j, m(x_j)) \qquad \forall i \in \{1, \dots, n\}$$

$$\hat{\mathcal{R}}^{\text{LOOCV}} = n^{-1}\sum_{i=1}^n l\left(y_i, \widehat{m}_{n-1}^{-i}(x_i)\right) \tag{3.13}$$

Pros

- Is basically unbiased estimator, as we use $n-1$ training samples.
- Can have a high variance due to highly correlated training sets, as the only vary in one observation.
- Can be better as K-fold cross-validation for small data sets, as small data sets have usually a higher fluctuation $\Rightarrow$ higher variance (as the are more sensitive to any noise/sampling artifacts).

Cons

- computational expensive, only for small data sets possible.
- Variance of the average can be very high due to highly correlated training sets.

3.6.1. LOOCV for Squared Loss and lin. Operator

Theorem 3.1 LOOCV Error for squared loss: For models that can be represented by a linear fitting operator $\mathbf{S}$:

$$[\widehat{m}(x_1) \cdots\cdots \widehat{m}(x_n)]^\intercal = \mathbf{S}\mathbf{Y} \tag{3.14}$$

it holds for the squared loss that:

$$n^{-1}\sum_{i=1}^n \left(y_i - \widehat{m}_{n-1}^{-i}(x_i)\right)^2 = n^{-1}\sum_{i=1}^n \left(\frac{y_i - \widehat{m}(x_i)}{1 - \mathbf{S}_{ii}}\right)^2 \tag{3.15}$$

Definition 3.24 Generalized Cross Validation (GCV):

$$\text{GCV} = n^{-1}\sum_{i=1}^n \frac{(y_i - \widehat{m}(x_i))^2}{(1 - n^{-1}\operatorname{tr}(\mathbf{S}))^2} \tag{3.16}$$

Explanation 3.4. *It holds $\overline{S_{ii}} = \frac{1}{n}\sum_{i=1}^n \mathbf{S}_{ii} = \frac{1}{n}\operatorname{tr}(\mathbf{S})$ thus we can rewirte the mean as the trace, which can efficiently calculated in $\mathcal{O}(n)$.*

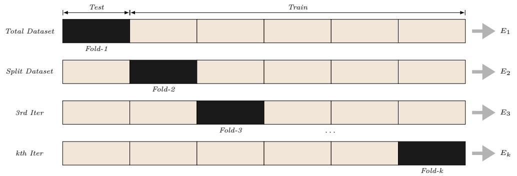
Note

GCV is a misdemeanor as it is an approximation and not a generalization.

3.7. K-Fold Cross Validation

Explanation 3.5 (K-fold Cross-Validation)**.**

1. *use all of the data by splitting the data into K random folds.*
2. *Calculate the training error K times by leaving out the k-th fold, fit the model to the other $K-1$ combined folds (training set) of size* $n \cdot \overset{\text{validation set}}{\frac{K-1}{K}}$.
3. *Do this by choosing each fold $k = 1, \ldots, K$ once as validation set and calculate cross-validation error by averaging over them.*



Definition 3.25 K-fold Cross Validation:

$$\mathcal{Z} = \mathcal{Z}_1 \cup \ldots \cup \mathcal{Z}_\nu \cup \ldots \cup \mathcal{Z}_K \qquad \forall k \in \{1, \ldots, K\}$$

$$\widehat{m}^{-\mathcal{Z}_k}_{n-|\mathcal{Z}_k|} \in \arg\min_{m \in \mathcal{F}} \frac{|\mathcal{Z}_k|}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}\setminus\mathcal{Z}_k} l(y_i, m(x_i)) \tag{3.17}$$

$$\hat{\mathcal{R}}^{\mathrm{CV}} = K^{-1} \sum_{k=1}^{K} |\mathcal{Z}_k|^{-1} \sum_{i \in \mathcal{Z}_k} l\left(y_i, \widehat{m}^{-\mathcal{Z}_k}_{n-|\mathcal{Z}_k|}(x_i)\right) \tag{3.18}$$

Note

A good heuristic for choosing K is 5, or 10 or:

$$k = \min\left(\sqrt{n}, 10\right)$$

Pros

- faster then LOOCV.

Cons

- runs $\approx K$ times slower than traing/test-split, as we need to train the model K times.
- Has higher bias then LOOCV.
  There exits systematic tendency to underfit, as each of the K-fold cross validation models uses only $n \cdot \frac{K-1}{K}$ training samples
  $\Rightarrow$ the estimates of prediction error will typically be more biased (towards simpler models), as the bias increases with a lower number of sampls/d.o.f. (see Rao Cramer).
- Depends on the explicit realization of the K subsets.

3.8. Many Random Divisions

Definition 3.26 Leave d-out CV:

Generalize LOOCV/d-fold CV by considering all possible realizationeq. (32.3) of d samples:

$$\mathcal{Z} = \mathcal{Z}_1 \cup \ldots \cup \ldots \cup \mathcal{Z}_{\binom{n}{d}} \qquad \forall k \in \left\{1, \ldots, \binom{n}{d}\right\}$$

$$\widehat{m}^{-\mathcal{Z}_k}_{n-|\mathcal{Z}_k|} \in \arg\min_{m \in \mathcal{F}} \frac{|\mathcal{Z}_k|}{|\mathcal{Z}|} \sum_{i \in \mathcal{Z}\setminus\mathcal{Z}_k} l(y_i, m(x_i)) \tag{3.19}$$

$$\hat{\mathcal{R}}^{\mathrm{CV}} = \binom{n}{d}^{-1} \sum_{k=1}^{\binom{n}{d}} |\mathcal{Z}_k|^{-1} \sum_{i \in \mathcal{Z}_k} l\left(y_i, \widehat{m}^{-\mathcal{Z}_k}_{n-|\mathcal{Z}_k|}(x_i)\right) \tag{3.20}$$

Explanation 3.6. *Is a generalization of LOOCV as it does not depend on the indexing in comparison to classical K-CV.*

Pros

- has often a smaller variance.

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:

$$I(p_0 \cdot p_1 \cdots p_n) = I(p_0) + I(p_1) + \cdots + I(p_n)$$

$\Rightarrow$ 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 $X \sim p$:

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

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

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

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

$\hat{=}$ units:

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

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

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

1.2. Entropy

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

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

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

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

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

Notes

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

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

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

1.2.1. Conditonal Entropy

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

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

Definition 4.5 proof 4.4 **Conditional Entropy** $H(Y|X)$: Is the amount of information need to determine Y if we arleady know X and is given by averagin $H(Y|X=x)$ over X.

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

Definition 4.6 proof 4.5 **Chain Rule for Entropy**:

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

Property 4.2 Monotonicity: Information/conditioning reduces the entropy $\Rightarrow$ *Information never hurts.*

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

Corollary 4.2 From eq. (4.17):

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

1.3. Cross Entropy

Definition 4.7 Cross Entropy proof 4.3: Lets say a model follows a true distribution $X \sim p$ but we model X as with a different distribution $X \sim q$. The cross entropy between p and q measure the average amount of information/bits needed to model an outcome $x \sim p$ with X:

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

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

1.4. Kullback-Leibler (KL) divergence

If we want to measure how different two distributions q and 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 p to q**: Given two probability distributions p, q of a random variable X. The Kullback–Leibler divergence is defined to be:

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

and measures how far away a distribution q is from a another distribution p.

Explanation 4.3.

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

Intuition

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

Property 4.3 Non-Symmetric:

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

Property 4.4:

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

Note

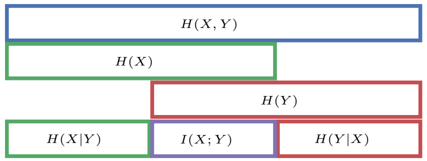
The KL-divergence is not a real distance measure as $\mathrm{KL}(\mathbb{P} \parallel Q) \neq \mathrm{KL}(Q \parallel \mathbb{P})$

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

1.5. Mutual Information

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

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.17}$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= D_{KL}(p_{X,Y} \parallel p_X p_Y)$$



Explanation 4.4 (Definition 4.9).

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

Property 4.5 Symmetry:

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

Property 4.6 Positiveness:

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

Property 4.7:

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

Property 4.8 Self-Information:

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

Property 4.9 Montone Submodularity: Mutual information is monotone submodular[def. 20.14]:

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

2. Proofs

Proof 4.1 Bayes Optimal Predictor[def. 3.14]: :

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

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

$$\stackrel{\heartsuit}{=} \mathbb{E}_{\mathbf{x}\sim p_{\mathcal{X}}}\left[\min_h(\mathbf{x})\underbrace{\mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}|\mathcal{X}}}\left[(y - h(\mathbf{x}))^2\,|\mathbf{x}\right]}_{\mathcal{R}_p(h,\mathbf{x})\ ([\text{def. } 3.7])}\right]$$

Now lets minimize the conditional executed risk:

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

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

$$= \int \frac{\mathrm{d}}{\mathrm{d}h^*}\left(y - h^*\right)^2 p(y|x)\,\mathrm{d}y = \int 2(y - h^*)p(y|x)\,\mathrm{d}y$$

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

Proof 4.2 Irreducible Error[cor. 3.3]:

$$\mathrm{MSEP}(x_n) = \mathbb{E}\left[\left(Y - \hat{Y}(x_n)\right)^2\right] = \mathbb{E}\left[(Y - \widehat{m}(x_n))^2\right]$$

$$= \mathbb{E}\left[(\epsilon + m(x_n) - \widehat{m}(x_n))^2\right]$$

$$= \mathbb{E}\left[\epsilon^2\right] + 2\mathbb{E}\left[\epsilon\cdot(m(x_n) - \widehat{m}(x_n))\right]$$

$$+ \mathbb{E}\left[(\epsilon + m(x_n) - \widehat{m}(x_n))\right]^2$$

$$= \mathbb{E}\left[\epsilon^2\right] + 2\mathbb{E}\left[\epsilon\cdot(m(x_n) - \widehat{m}(x_n))\right]$$

$$+ \mathbb{E}\left[(\epsilon + m(x_n) - \widehat{m}(x_n))\right]^2$$

$$= \mathbb{V}[\epsilon] + 2\underbrace{\mathbb{E}[\epsilon]}_{=0}\cdot\mathbb{E}\left[(m(x_n) - \widehat{m}(x_n))\right]$$

$$+ \mathbb{E}\left[(\epsilon + m(x_n) - \widehat{m}(x_n))\right]^2$$

$$= \mathbb{V}[\epsilon] + \mathrm{MSE}(x_n)$$

Proof 4.3: [def. 4.7]

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

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

Notes: ♡

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

Note

$$\mathbb{E}[X]\,\mathbb{E}[Y|X] = \int_X p_X(x)\,\mathrm{d}x\int_Y p(y|x)\,\mathrm{d}y$$

$$= \int_X\int_Y p_X(x)p(y|x)xy\,\mathrm{d}x\,\mathrm{d}y = \mathbb{E}[X, Y]$$

Proof 4.4: Definition 4.5

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

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

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

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

Proof 4.5: [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 4.6: example 4.4

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[\log(p) - \log(q)\right]$$

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

$$\left. + \frac{1}{2}(\mathbf{x} - \mu_q)^\intercal\Sigma_q^{-1}(\mathbf{x} - \mu_q)\right]$$

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

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

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

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

$$\underline{\mathbb{E}_p\left[a\right]} \stackrel{\mathrm{tr}(\mathbb{R})=\mathbb{R}}{=} \mathbb{E}_p\left[\mathrm{tr}\left\{(\mathbf{x} - \mu_p)^\intercal\Sigma_p^{-1}(\mathbf{x} - \mu_p)\right\}\right]$$

$$\stackrel{\text{eq. }(27.54)}{=} \mathbb{E}_p\left[\mathrm{tr}\left\{(\mathbf{x} - \mu_p)(\mathbf{x} - \mu_p)^\intercal\Sigma_p^{-1}\right\}\right]$$

$$= \mathbb{E}_p\left[\mathrm{tr}\left\{\Sigma_p\Sigma_p^{-1}\right\}\right]$$

$$\stackrel{\text{eq. }(27.54)}{=} \mathbb{E}_p\left[\mathrm{tr}\left\{\mathbf{I}_d\right\}\right] = \mathbb{E}_p\left[d\right] = d$$

$$\underline{\mathbb{E}_p\left[b\right]} \stackrel{\text{eq. }(33.54)}{=} (\mu_p - \mu_q)^\intercal\Sigma_q^{-1}(\mu_p - \mu_q) + \mathrm{tr}\left\{\Sigma_q^{-1}\Sigma_p\right\}$$

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:

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\ln|2\pi e\Sigma| \stackrel{\text{eq. }(27.55)}{=} \frac{1}{2}\ln\left((2\pi e)^d|\Sigma|\right)$$

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

$$\stackrel{\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)}{=} \frac{1}{2}\ln|2\pi e| + \cancel{\frac{1}{2}}\sum_{i=1}^d \ln\sigma_i^{\cancel{2}}$$

Example 4.4 proof 4.6

KL Divergence of Gaussians:

Given two Gaussian distributions:

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

$$D_{\mathrm{KL}}(p \parallel q) =$$

$$= \frac{\mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) + (\mu_q - \mu_p)^\intercal\Sigma_q^{-1}(\mu_q - \mu_p) - d + \ln\left(\frac{|\Sigma_q|}{|\Sigma_p|}\right)}{2}$$

Example 4.5 KL Divergence of Scalar Gaussians:

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

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

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

Example 4.6 KL Divergence of Diag. Gaussians:

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

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

Example 4.7 KL Divergence of Gaussians:

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

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

Example 4.8 Gaussian Mutal Information:

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

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

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

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

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

Example 4.9 Bayes Optimal Predictor and MLE[def. 3.14]**:** **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. (3.10).

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:

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

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.

Supervised Learning

Definition 4.10 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?

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

Recall: goal of supervised learning

Given: training data:

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

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

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

s.t. we minimize prediction error/empirical risk [def. 3.10].

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,...).

4. Estimators

Definition 4.11 (Sample) Statistic: A statistc is a measuarble function f that assigns a **single** value F to a sample of random variables: $\mathbf{X} = \{X_1, \ldots, X_n\}$

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

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

Note

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

Definition 4.12 Statistical/Population Parameter: Is a parameter defining a family of probabilty distributions see example 4.1

Definition 4.13 (Point) Estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$**:**

Given: n-samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathbf{X}$ an estimator

$$\hat{\theta} = h(\mathbf{x}_1, \ldots, \mathbf{x}_n) \tag{4.25}$$

is a statistic/randomn variable used to estimate a true (population) parameter θ[def. 4.12] 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).

Generalized Linear Models (GLMs)

Definition 4.14 Generalized Linear Model (GLM):

$$\mu = \mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right] = g^{-1}\left(\eta\right) \tag{4.26}$$

$$\eta = \sum_{j=0}^{p} \beta_{jm} X_j \tag{4.27}$$

$$g\left(\mathbb{E}\left[\mathbf{Y}|\mathbf{X}\right]\right) = \eta \tag{4.28}$$

Generalized Additive Models (GAMs)

Definition 4.15 Generalized Additive Models (GAMs):

$$sdf \tag{4.29}$$

Regression

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

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

Definition 5.3 Coefficients β: Are the coefficients that we are seeking.

Definition 5.4 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:

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

Note

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

Definition 5.5 Linear Regression: Refers to regression that is linear w.r.t. to the parameter vector β (but not necessarily the data):

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

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

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

Simple (linear) regression (SLR)

Definition 5.7 [example 5.1]
Simple Linear Regression: Is a *linear regression*[def. 5.8] with only one explanatory variable[def. 5.1]:

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

Multiple (linear) regression (MLR)

Definition 5.8 Multiple Linear Regression:
Is a linear regression model with multiple $\{\beta_j\}_{j=1}^p$ explanatory[def. 5.1] variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i$$
$$= \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i = \beta^\intercal x_i + \epsilon_i \qquad i = 1, \dots, n$$

Design Matrix:

$$\begin{bmatrix} \mathbf{x} \end{bmatrix}\begin{bmatrix} \beta \end{bmatrix} = \begin{bmatrix} \mathbf{Y} \end{bmatrix} \qquad \mathbf{y} = \mathbf{X}\beta \qquad \begin{aligned} &\mathbf{X} \in \mathbb{R}^{n,(p+1)} \\ &\mathbf{y} \in \mathbb{R}^n \\ &\beta \in \mathbb{R}^{p+1} \end{aligned} \tag{5.5}$$

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

Note

Eq. 5.8 is usually an over-determined system of linear equations i.e. we have more observations then predictor variables.

Multiple vs. Multivariate lin. Reg.

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

Definition 5.9 [example 5.2]
Simple Linear Quadratic Regression: Is a *linear regression*[def. 5.8] with two explanatory variables[def. 5.1] written as:

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

0.0.1. Existence

Corollary 5.1 Existence:

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

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

1. Linear/Ordinary Least Squares (OLS)

Problem: for an over determined system $n > p$ (usually) $\nexists \mathbf{y} \in \Re(\mathbf{X})$ (in particular given round off errors) s.t. there exists no parameter vector β that solves [def. 5.8].
Idea: try to find the next best solution by minimizing the residual(s)[def. 5.6].

Definition 5.10 Residual Sum of Squares:
Is the sum of residuals[def. 5.6]:

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

Definition 5.11 Least Squares Regression $\text{lsq}(\mathbf{X}, \mathbf{y})$:
Minimizes the residual sum of squares:

$$\hat{\beta} \in \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \arg\min_{\mathbf{u}\in\Re(\mathbf{X})} \|\mathbf{y} - \mathbf{u}\|_2^2 \tag{5.10}$$

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

Alternative Formulation

Sometimes people write eq. (5.10) as $\frac{1}{2} \arg\min_\beta \|\mathbf{r}\|_2^2$ which leads to the same solutioneq. (24.62).

2. Maximum Likelihood Estimate

Ridge MLE

Proposition 5.1 (Gauss Markov Assumptions)
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. 37.22]:
$$\mathbb{V}[\epsilon_i] = \sigma^2 < \infty \qquad \forall i \tag{5.11}$$
3. The errors are uncorrelated:
$$\text{Cov}[\epsilon_i, \epsilon_j] = 0 \qquad \forall i \neq j \tag{5.12}$$
4. The errors are jointly normally distributed with mean 0 and constant variance σ^2:
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i = 1, \dots, n \quad \iff \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n) \tag{5.13}$$

Definition 5.12 [proof 5.2]
Simple Linear Regression Log-Likelihood:

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

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

$$\text{l}_n(\mathbf{y}|\mathbf{X}, \theta) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^\intercal \mathbf{x}_i)^2 = -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2$$
$$\theta^* \in \arg\max_{\theta\in\mathbb{R}^{p+1}} \text{l}_n(\mathbf{y}|\mathbf{X}, \theta) = \arg\min_{\theta\in\mathbb{R}^{p+1}} -\text{l}_n(\mathbf{y}|\mathbf{X}, \theta) \tag{5.14}$$

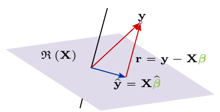
2.1. The Normal Equation

Definition 5.13 [proof 5.4]
The Normal Equations:
Is the equation we need to solve in order to solveeq. (5.10) or equivalently eq. (5.14) and is no longer an over determined system:

$$\begin{bmatrix} \mathbf{X}^\intercal\mathbf{X} \end{bmatrix}\begin{bmatrix} \beta \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\intercal \end{bmatrix}\begin{bmatrix} \mathbf{Y} \end{bmatrix} \qquad \mathbf{X}^\intercal\mathbf{X}\hat{\beta} = \mathbf{X}^\intercal\mathbf{Y} \qquad \begin{aligned} &\mathbf{X}^\intercal\mathbf{X} \in \mathbb{R}^{p\times p} \\ &\beta \in \mathbb{R}^p \\ &\mathbf{X}^\intercal \in \mathbb{R}^{p\times n} \\ &\mathbf{Y} \in \mathbb{R}^n \end{aligned} \tag{5.15}$$

Geometric Interpretation

Corollary 5.2 [proof 5.5]
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:

$$\arg\min_{\hat{\mathbf{y}}\in\{\mathbf{X}\beta:\beta\in\mathbb{R}^n\}=\Re(\mathbf{X})} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$$

but this minimum is equal to the orthogonal projection[def. 27.20] of $\mathbf{y}$ onto $\Re(\mathbf{X})$ i.e. the map:

$$\mathbf{y} \mapsto \hat{\mathbf{y}}$$

is the orthogonal projection of $\mathbf{y}$ onto $\Re(\mathbf{X})$.

Corollary 5.3 Orthogonality of residuals [proof 5.6]:
Corollary 5.2 implies that the residuals are orthogonal w.r.t. to all the column vectors of $\mathbf{X}$:

$$\mathbf{r}^\intercal \mathbf{x}^{(j)} = 0 \qquad \forall j = 1, \dots p \tag{5.16}$$

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

Proposition 5.2 Least Squares Solution:

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

Note

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

add MPPI to linear algebra appendix

2.1.2. Solving The Normal Equation

Cholesky Decomposition

Corollary 5.4 Computational Complexity: $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{y} \in \mathbb{R}^n, \mathbf{w} \in \mathbb{R}^d$ with n, the number of observations and d, the number of equations/feautres/dimension of the problem.
Assume: $d \leqslant n$, that is we have an overdetermined system, more equations than unkowns.

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

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

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

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

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

QR Decomposition

2.1.3. Simple Linear Regression Solution

Definition 5.14 [proof 5.4]
Linear Regression Solution:

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

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

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

2.1.4. Making Predictions

Definition 5.15 $\quad$ **P/H** $= \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal : y \mapsto \hat{y}$
Hat/Projection Matrix:
Is the matrix that projects the y onto the $\hat{y}$:

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

Property 5.1 Symmetry: P is trivially symmetric.

Property 5.2 Idem-potent $\quad \mathbf{P}^2 = \mathbf{P}$: **P** is idem-potent i.e. projecting multiple times by **P** is the same as projecting once.

Property 5.3 Trace:

$$\text{tr}(P) = \text{tr}(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal) = \text{tr}((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X})$$
$$= \text{tr}(I_{p\times p}) = p$$

Corollary 5.5 $\mathbf{P}: \mathbb{R}^n \mapsto \mathcal{X} \subseteq \mathbb{R}^p$: From these three properties it follows that **P** is an orthogonal projection onto a p-dim subspace.

Corollary 5.6 Residual Projection: The residual can be represented in terms ofeq. (5.20):

$$\mathbf{r} = (\mathbf{I} - \mathbf{P})\mathbf{Y} \tag{5.21}$$

it follows that $\mathbf{I} - \mathbf{P}$ is an orthogonal projection onto $(n-p)$-dim subspace $\mathcal{X}^\perp = \mathbb{R}^n \setminus \mathcal{X}$.

Uniqueness

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

$$\mathcal{N}(\mathbf{A}) = \mathcal{N}(\mathbf{A}^\intercal\mathbf{A}) \qquad \Re(\mathbf{A}^\intercal) = \Re(\mathbf{A}^\intercal\mathbf{A}) \tag{5.22}$$

add rest from formulary

Theorem 5.2 Full-Rank Condition F.R.C.:
Equation 5.13 has a unique least squares solution given by:

$$\hat{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y} \tag{5.23}$$

$$\iff \quad \mathcal{N}(\mathbf{X}) = \{0\} \quad \iff \quad \text{rank}(\mathbf{X}) = p \qquad p \geqslant p \tag{5.24}$$

2.2. Moments and Distributions

Property 5.4 Moments of $\hat{\beta}$ [proof 5.7]:

$$\mathbb{E}\left[\hat{\beta}\right] = \beta \qquad \mathbb{V}\left[\hat{\beta}\right] = \text{Cov}\left[\hat{\beta}\right] = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} \tag{5.25}$$

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\right) \tag{5.26}$$

Property 5.5 Moments of $\hat{y}$ [proof 5.9]:

$$\mathbb{E}[\hat{\mathbf{y}}] = \mathbb{E}[\mathbf{y}] = \mathbf{X}\beta \qquad \mathbb{V}[\hat{\mathbf{y}}] = \text{Cov}[\hat{\mathbf{y}}] = \sigma^2\mathbf{P} \tag{5.27}$$

$$\hat{\mathbf{y}} \sim \mathcal{N}_n\left(\mathbf{X}\beta, \sigma^2\mathbf{P}\right) \tag{5.28}$$

Property 5.6 Moments of r:

$$\mathbb{E}[\mathbf{r}] = 0 \qquad \text{Cov}[\mathbf{r}] = \sigma^2(\mathbf{I} - \mathbf{P}) \tag{5.29}$$

$$\mathbf{r} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\left(\mathbf{I} - \mathbf{P}\right)\right) \tag{5.30}$$

Property 5.7 Moments of $\hat{\sigma}$:

$$\hat{\sigma}^2 := \frac{1}{n-p} \sum_{i=1}^n r_i^2 \quad \Longrightarrow \quad \mathbb{E}[\hat{\sigma}] = \sigma \tag{5.31}$$

$$\hat{\sigma}^2 \sim \frac{\sigma}{n-p} \chi_{n-p}^2 \tag{5.32}$$

Note

The standard deviation σ^2 is given by $\epsilon \sim \mathcal{N}0, \sigma^2$. However we may not know σ^2, thus we can estimate it by using the residuals $\mathbf{r}$.

add proofs and explainations see Greens economics 2003 p48: greens

Proof 5.1 Property 5.7: $\hat{\sigma}^2$ is an unbiased estimator of σ:

2.2.1. The Gaus Markov Theorem

Theorem 5.3 Gauss–Markov theorem [proof 5.10]**:**
The BLUE of the β coefficients, of a linear regression model, satisfying the **Gauss-Markov assumptions** is given by the ordinary least squares (OLS) estimator, provided it exists (is invertible).

$$\mathbb{V}\left[\hat{\beta}\right] \leqslant \mathbb{V}\left[\tilde{\beta}\right] \quad \text{with} \quad \begin{aligned} &\hat{\beta} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal y = \mathbf{C}\mathbf{y} \\ &\tilde{\beta} \text{ any lin. unb. est. for } \beta \end{aligned} \tag{5.33}$$

add delete next section

3. MLE with linear Model & Gaussian Noise

3.1. MLE for conditional linear Gaussians

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

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

$\mathbf{X}$: set of explicative variables. 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 $h(\mathbf{x}) = \mathbf{w}^\intercal\mathbf{x}$
and Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

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

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

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

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

$$\mathscr{L}_n(Y|\mathbf{X},\theta) = \hat{\mathrm{p}}(y_1,\dots,y_n|\mathbf{x}_1,\dots,\mathbf{x}_n,\theta)$$

$$\overset{\text{i.i.d.}}{\underset{??}{=}} \prod_{i=1}^n \hat{\mathrm{p}}_{Y|\mathbf{X}}(y_i|\mathbf{x}_i,\theta) = \prod_{i=1}^n \mathcal{N}(\mathbf{w}^\intercal\mathbf{x}_i, \sigma^2)$$

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

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

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

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

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

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

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

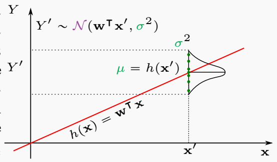
$$\frac{\partial \mathrm{l}_n(Y|\mathbf{X},\theta)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n \left(y_i - \mathbf{w}^\intercal\mathbf{x}_i\right)^2 = 0$$

$$\boxed{\theta^* = \begin{pmatrix}\mathbf{w}_* \\ \sigma^2_*\end{pmatrix} = \begin{pmatrix} \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\intercal\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i y_i\right) \\ \frac{1}{n}\sum_{i=1}^n \left(y_i - \mathbf{w}_*^\intercal\mathbf{x}_i\right)^2 \end{pmatrix}} \tag{5.34}$$

Note

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

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



3.2. Conditional MLE ≙ Least Squares

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

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

$$\arg\max_{\mathbf{w}} \mathrm{l}_n(\mathbf{w}) \iff \arg\min_{\mathbf{w}} -\mathrm{l}_n(\mathbf{w})$$

$$\arg\min_{\mathbf{w}} \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2 = \arg\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}^\intercal\mathbf{x}_i)^2 \tag{5.35}$$

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

Corollary 5.7 : 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 ??

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

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

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

$$\boxed{\theta^* = \begin{pmatrix}\mu_* \\ \sigma^2_*\end{pmatrix} = \begin{pmatrix}\frac{1}{n}\sum_{i=1}^n y_i \\ \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_i)^2\end{pmatrix}} \tag{5.36}$$

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

Note

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

3.3. MLE for general conditional Gaussians

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

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

Is given by the least squares solution:

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

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

Other distributions

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

Ridge Max Prior

Prior

Assume: prior $\mathbb{P}(\beta|\boldsymbol{\Sigma})$ on the model parameter β is gaussian as well and depends on the hyperparameter ([def. 7.7]) $\boldsymbol{\Sigma}$ (≙ co-variance matrix):

$$\beta \sim \mathrm{p}^{\text{Ridge}}(\beta|\boldsymbol{\Sigma}) = \mathcal{N}(\beta|0, \boldsymbol{\Sigma})$$

$$\overset{[\text{def. } 34.34]}{=} (2\pi)^{-\frac{d+1}{2}}\det(\Sigma)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\beta^\intercal\Sigma^{-1}\beta\right)$$

$$\mathrm{l}_n(\beta|\boldsymbol{\Sigma}) = -\frac{1}{2}\ln\det(\boldsymbol{\Sigma})^{-1} - \frac{d+1}{2}\ln 2\pi - \frac{1}{2}\beta\boldsymbol{\Sigma}^{-1}\beta \tag{5.37}$$

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

Max Prior

$$\beta^* \in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\, \mathrm{l}_n(\beta|\boldsymbol{\Sigma})$$

$$= \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max} -\frac{1}{2}\ln\det(\boldsymbol{\Sigma})^{-1} - \frac{d+1}{2}\ln 2\pi - \frac{1}{2}\beta\boldsymbol{\Sigma}^{-1}\beta$$

$$0 \overset{!}{=} \frac{\partial}{\partial\beta^*}\mathrm{l}_n(\beta^*|\boldsymbol{\Sigma}) = -\frac{\partial}{\partial\beta^*}\beta^*\boldsymbol{\Sigma}^{-1}\beta^* \overset{\text{eq. (5.46)}}{=} -2\boldsymbol{\Sigma}^{-1}\beta^*$$

$$\beta^* \in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\log\mathrm{p}(\beta|\boldsymbol{\Sigma}) = \underset{\beta\in\mathbb{R}^{d+1}}{\arg\min} -\mathrm{l}_n(\beta|\boldsymbol{\Sigma}) = 2\boldsymbol{\Sigma}^{-1}\beta^* \tag{5.38}$$

Log-MAP

$$\beta^* \in \underset{\beta\in\mathbb{R}^{d+1}}{\arg\max}\,\mathbb{P}(\beta|\mathbf{X},\mathbf{y})$$

$$= \underset{\beta\in\mathbb{R}^{d+1}}{\arg\min} -\log\overbrace{\mathbb{P}(\beta|\boldsymbol{\Sigma})}^{\text{eq. (5.38)}} - \log\overbrace{\mathbb{P}(\mathbf{X},\mathbf{y}|\beta)}^{??}$$

$$= \boldsymbol{\Sigma}^{-1}\beta^* - \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{y} + \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{X}\beta^* = 0$$

$$\iff (\boldsymbol{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X}\sigma^{-2})\beta^* = \sigma^{-2}\mathbf{X}^\intercal\mathbf{y}$$

$$(\sigma^2\boldsymbol{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X})\hat{\beta} = \mathbf{X}^\intercal\mathbf{y}$$

$$\hat{\beta}^{\text{MAP}} = \left(\sigma^2\boldsymbol{\Sigma}^{-1} + \mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}$$

Definition 5.16 Ridge MAP: For ridge regression we assume that the noise of the prior is uncorrelated/diagonal i.e.

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I}\sigma^{-2} \quad \textbf{and let} \quad \boldsymbol{\Lambda} := \sigma^2\boldsymbol{\Sigma}^{-1} = \mathbf{I}\frac{\sigma^2}{\sigma^2} \tag{5.39}$$

which leads to:

$$\hat{\beta}^{\text{MAP}} = (\boldsymbol{\Lambda} + \mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \quad \text{with} \quad \boldsymbol{\Lambda} = \mathbf{I}\lambda = \mathbf{I}\frac{\sigma^2}{\sigma^2} \tag{5.40}$$

Definition 5.17 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 5.18 Tikhonov regularization: Commonly used method of regularization of ill-posed problems.

$$\|\mathbf{X}\beta - \mathbf{y}\|^2 + \|\boldsymbol{\Gamma}\beta\|^2 \tag{5.41}$$

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/Liklihood MAP inference

$$\hat{\beta}^{\text{Ridge}} = \arg\min_\beta\left\{\underbrace{(\mathbf{y} - \mathbf{X}\beta)^\intercal(\mathbf{y} - \mathbf{X}\beta)}_{\text{data term}} + \underbrace{\beta^\intercal\boldsymbol{\Lambda}\beta}_{\text{regularizer/penalty}}\right\}$$

$$= \arg\min_\beta\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \beta^\intercal\boldsymbol{\Lambda}\beta\right\}$$

$$\overset{\text{eq. (5.39)}}{=} \arg\min_\beta\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2\right\}$$

$$= \arg\min_\beta\left\{\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\sum_{i=1}^d \beta_i^2\right\}$$

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

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

Note: SVD

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

Note: no feature selection

The coefficients in a ridge will go to zero as 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.

$$\text{dofs}(\lambda) = \begin{cases} d & \text{if } \lambda = 0 \text{ (no regularization)} \\ \to 0 & \text{if } \lambda \to \infty \end{cases} \tag{5.42}$$

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

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

1 $\lambda \xrightarrow[\sigma\downarrow]{\sigma\uparrow} \infty$: $\|\beta\|^2$ must be minimized:
  - $\sigma\uparrow$: model does not need to match data so perfectly as we have more noise in our data/observations $\iff$ bigger errors (recall $\epsilon \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$).
  - $\sigma\downarrow$: prior has smaller variance, thus our prior knowledge of the model is pretty exact/important (recall $\beta \sim \mathcal{N}(\beta|0, \mathbf{I}\sigma)$)

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

Note

- Often $\Lambda^{-1} = \mathbb{1} \in \mathbb{R}^{d+1\times d+1}$
- 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 $\mathbf{X}^\intercal\mathbf{X}$ in eq. (5.40) is not invertible/regular/has not full rank.
Then $(\mathbf{X}^\intercal\mathbf{X} + \Lambda)$ will still be invertible/well posed.
This was the original reason for L2/Ridge Regression.

MAP $\triangleq$ Ridge

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

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

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

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

Generalization

Regularized estimation can often be understood as MAP inference:

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

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

Priors

3.4. Laplace Prior$\triangleq$Lasso/L1-regularization

Intro

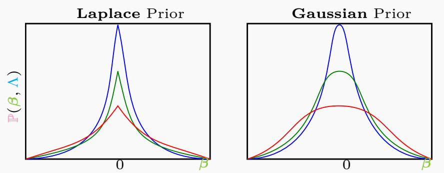
Question: what if $d \gg n$ e.g.
- bag of words with d =nb. of words$\gg n$nb. of documents.
- Genome analysis d =nb. of genes$\gg$ n patients.

Problem: we have more unkowns/parameters than observations $\Rightarrow$ no unique solution. **e.g.**: Trying to fit 1 data point with polynomimal of degree 12.
Question: can we somehow still find a good solution if $n = \mathcal{O}(\ln d) \iff$ exp. more dim. than observations
Idea: If most of the dimensions are irrelevant for the problem, then we can find a good (sparse) solution $\triangleq$ feature selection/dimensionality reduction.

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

$$\mathbb{P}^{\text{Lasso}}(\beta|\Lambda) \overset{\text{eq. (34.58)}}{=} \frac{\Lambda}{2}e^{(-\Lambda|\beta|)} = \prod_{j=1}^{d}\frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}$$

With $\Lambda^{-1} := \Sigma$ hyperparameter/covariance matrix
This leads to a L1 regularized model:



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

$$-\ln\mathbb{P}(\beta|\Lambda) = \sum_{j=1}^{d}\lambda_j|\beta_j| - d\ln\frac{\lambda_j}{2} \tag{5.43}$$

Laplacian MAP Prior Inference

$$\beta^* = \arg\min_{\beta\in\mathbb{R}^d}\left\{\|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda\|\beta\|_1\right\}$$
$$= \arg\min_{\beta\in\mathbb{R}^d}\left\{\|\mathbf{y} - (\mathbf{X}\beta + \beta_0)\|^2 + \lambda\sum_{i=1}^{d}|\beta_i|\right\} \tag{5.44}$$

$|\beta|_i$ does not change β_i while β_i^2 becomes very small for values$\in(0,1)$ thus when minimizing the L2 error $\|betac\|^2 \to 0$ **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. 24.12]) optimization problem eq. (5.44) is not differentiable at $\beta_i = 0$ and **thus** has no closed form solution as the L2 problem $\Rightarrow$ quadratic programming.

3.5. Sparsness Priors/L0-regularization

$$-\ln\mathbb{P}(\beta|s) = s\sum_{j=1}^{d}\mathbb{1}_{\beta_j\neq 0} = s\sum_{j=1}^{d}\cdot\begin{cases}1 & \text{if } \beta_j \neq 0\\ 0 & \text{otherwise}\end{cases} \tag{5.45}$$

$\Rightarrow$ measure for the number of possible non-zero dimesnions/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.

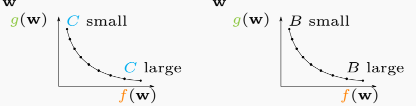
Drawback

Non-convex, non-differentiable problem $\Rightarrow$ computationally difficult combinatorics.

Scalarization vs. Constrained Optimization

Their are two equivilant ways of trading:
- $g(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$: the data term and
- $f(\beta)$: the Regularizer.

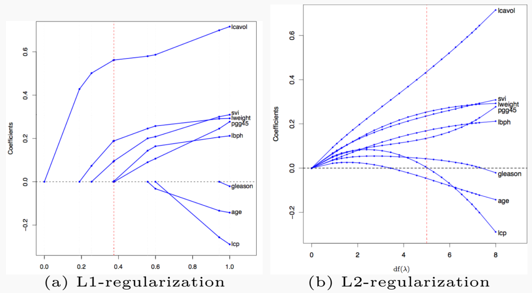
Scalariztion $\min_{\mathbf{w}} f(\beta) + Cg(\beta)$
Constraint opt. $\min_{\mathbf{w}} g(\beta)$ s.t. $f(\mathbf{w}) \leqslant B$



Note

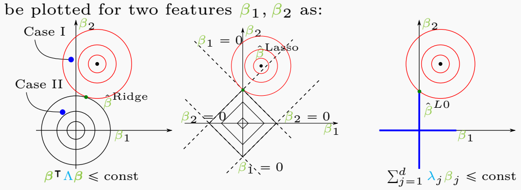
Scalarization and constrained optimzization gives the same curves $\iff$ f, g are both convex functions.
This is not necessarily for the same values of C and B **but** their exisits always a relationship $C = u(B)$ s.t. this is true.

Comparison of priors



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

The constraint formulation of the optimization problems can be plotted for two features 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 violated the constraint $\beta_1^2 + \beta_2^2 \leqslant$ const (Case I).
- **Lasso/L1-regression**: Here the constraint equals $|\beta_1| + |\beta| \leqslant$ const and leads to polyhedron. Most of the time we obtain a sparse solution $\triangleq$ corrner, due to the fact that corner regions increas much faster in volume, as the mixed regions (sparseness increases with number of dimensions).
- **Sparsness prior/L0-regression**: Leads to a super spiky geometry $\Rightarrow$ always leads to a sparse solution.

Liklihoods

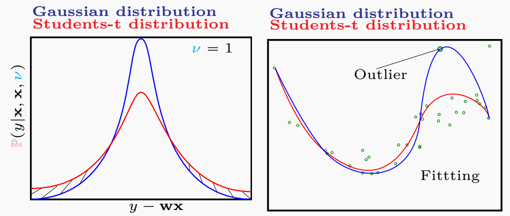
3.6. Student's-t likelihood loss function

Students-t Distribution:

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

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



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

Students-t $\mathbb{P}(|y - \mathbf{w}^\intercal\mathbf{x}| > t) = \mathcal{O}(t^{-\alpha})$ $\alpha > 0$
(Polynomial decay)
Gaussian $\mathbb{P}(|y - \mathbf{w}^\intercal\mathbf{x}| > t) = \mathcal{O}(\exp^{-\alpha t})$ $\alpha > 0$
(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 5.2 5.12: From eq. (5.12) it follows that the response variables are uncorrelated given the explanatory variables $\text{Cov}\left[Y_i, Y_j|\mathbf{X}\right] = 0$. Hence we have i.i.d. samples with a corresponding conditional (log-)likelihood given by:

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

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

Proof 5.3 Definition 5.14:

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

add to proof and linear algbra section

Note: ★

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

$$\frac{\partial}{\partial\mathbf{x}}\mathbf{M}\mathbf{x}=\mathbf{M}\quad\text{and}\quad\frac{\partial}{\partial\mathbf{x}}\mathbf{x}^\intercal\mathbf{M}\mathbf{x}=(\mathbf{M}+\mathbf{M}^\intercal)\mathbf{x}\tag{5.46}$$

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

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

Thus

$$0=\frac{\partial}{\partial\beta}(\mathbf{y}-\mathbf{X}\beta)^\intercal(\mathbf{y}-\mathbf{X}\beta)=2\mathbf{X}^\intercal(\mathbf{X}\beta-\mathbf{y})\tag{5.47}$$

combine proofs

Proof 5.4: [def. 5.13]

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

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

Note

$$\frac{\partial}{\partial\beta}\beta^\intercal\mathbf{X}^\intercal(\mathbf{X}\beta)\overset{\text{eq. (27.132)}}{=}(\mathbf{X}^\intercal\mathbf{X}+(\mathbf{X}^\intercal\mathbf{X})^\intercal)\beta=2\mathbf{X}^\intercal\mathbf{X}\beta$$

Proof 5.5: Corollary 5.2

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

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

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

Proof 5.6 Corollary 5.3: From [def. 5.13] it follows:

$$\begin{aligned}&\mathbf{X}^\intercal\mathbf{Y}=\mathbf{X}^\intercal\mathbf{X}\hat\beta=\hat\beta^\intercal\mathbf{X}^\intercal\mathbf{X}=(\mathbf{X}\hat\beta)^\intercal\mathbf{X}\\\iff&(\mathbf{Y}-\mathbf{X}\hat\beta)\mathbf{X}=\mathbf{r}^\intercal\mathbf{X}=0\end{aligned}$$

Proof 5.7 Property 5.4: $\hat\beta$ an unbiased estimator of β:

$$\begin{aligned}\hat\beta&=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal y=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\beta+\epsilon)\\&=\beta+(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\epsilon\\\mathbb{E}_\epsilon[\hat\beta]&=\mathbb{E}[\beta]+\mathbb{E}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\epsilon]\\&=\beta+(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\underbrace{\mathbb{E}[\epsilon]}_{=0}=\beta\end{aligned}$$

Proof 5.8 Property 5.4: Covariance $\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$:

$$\begin{aligned}\operatorname{Cov}\left[\hat\beta\right]&=\overbrace{\operatorname{Cov}[\beta]}^{=0}+\overbrace{\operatorname{Cov}\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\epsilon\right]}^{:=\mathbb{V}[\alpha\epsilon]}=\mathbb{E}[(\alpha\epsilon)^2]-\overbrace{\mathbb{E}[\alpha\epsilon]}^{=0}{}^2\\&=\mathbb{E}[(\alpha\epsilon)^\intercal(\alpha\epsilon)]=\mathbb{E}\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\epsilon\epsilon^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})\right]\\&=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbb{E}\left[\epsilon\epsilon^\intercal\right]\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})\\&=\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})=\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\end{aligned}$$

Proof 5.9 Property 5.5: $\hat{\mathbf{y}}$ an unbiased estimator of $\mathbf{y}$:

$$\mathbb{E}_\epsilon[\hat{\mathbf{y}}]=\mathbb{E}[\mathbf{X}\hat\beta+\epsilon]=\mathbf{X}\mathbb{E}[\hat\beta]+0\overset{\text{eq. (5.25)}}{=}\mathbf{X}\beta=\mathbb{E}[\mathbf{y}]$$

Proof 5.10 Theorem 5.3: $\hat\beta$ is a linear operator w.r.t. to $\mathbf{y}$:

$$\begin{aligned}\hat\beta&=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal y=:\mathbf{C}\mathbf{y}=\mathbf{C}(\mathbf{X}\beta)\\&=\beta+(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\epsilon=:\tilde C\epsilon+\beta\end{aligned}$$

5. Examples

Example 5.1 Simple Linear Regression:

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

Example 5.2 Simple Linear Quadratic Regression:

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

Classification

6. Intro

Definition 5.19 Training Data $\mathcal{D}$:
$$\mathcal{D} := \left\{(\mathbf{x}_i, y_i) \quad | \quad \mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathcal{Y} := \{c_1, \dots, c_K\}\right\}$$

Definition 5.20 Classifier c:
Is a mapping that maps the features into classes:
$$c : \mathcal{X} \to \mathcal{Y} \tag{5.48}$$

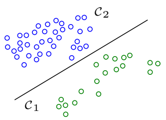
6.1. Types of Classification

Definition 5.21 Dichotomy:
Given a set $\mathcal{S} = \{s_1, \dots, s_N\}$ a dichotomy is partition of the set $\mathcal{S}$ into two subsets A, A^c that satisfy:
- *Collectively/jointly exhaustiveness*:
$$S = A \cup A^C \tag{5.49}$$
- *Mutual exclusivity*:
$$s \in A \quad \Longrightarrow \quad s \notin A^C \qquad \forall s \in S \tag{5.50}$$

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

Definition 5.22 Binaray Classification:

Is a classification problem where the labels are binary:
$$\mathcal{Y} = \{c_1, c_2\} = \{-1, 1\} \tag{5.51}$$



multiclass

6.2. Encodings

6.2.1. One Hot Encoding

Definition 5.23 One-hot encoding/representation:
Is the representation/encoding of the K categories $\{c_1, \dots, c_K\}$ by a *sparse vectors*[def. 27.68] with one non-zero entry, where the index j of the non-zero entry indicates the class c_j:
$$\mathbb{B}^n = \left\{\mathbf{y} \in \{0,1\}^n : \mathbf{y}^\intercal \mathbf{y} = \sum_{i=1}^{n} \mathbf{y} = 1\right\}$$
s.t. $\qquad \mathbf{y}_i = \mathbf{e}_j \quad \Longleftrightarrow \quad \mathbf{y}_i = c_j$

6.2.2. Soft vs. Hard Labels

Definition 5.24 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. 5.23]:
$$y = c_k \quad \Longrightarrow \quad y = \mathbf{e}_k \tag{5.52}$$

Definition 5.25 Soft Labels/Targets: Are observations $y \in \mathcal{Y}$ that are consider as noisy observations or probabilities p. We can encode them using a probabilistic vector[def. 27.69]:
$$y = [\mathrm{p}_1 \cdots \cdots \mathrm{p}_K]^\intercal \tag{5.53}$$

Corollary 5.8 Hard labels as special case: If we consider hard targets[def. 5.24] as events with probability one then we can think of them as a special case of the soft labels.

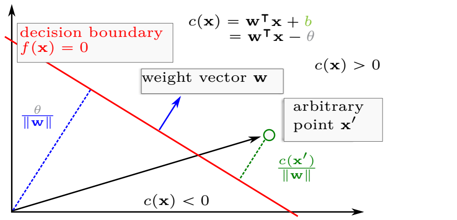
7. Binary Classification $\{-1, 1\}$

7.1. Linear Classification

Definition 5.26 Linear Dichotomy:

Definition 5.27 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. 27.13]:
$$\hat{y} = c(\mathbf{x}_i) = \begin{cases} c_1 \in \mathcal{H}^+ & \text{if } \mathbf{w}^\intercal\mathbf{x} > \theta \\ c_2 \in \mathcal{H}^- & \text{if } \mathbf{w}^\intercal\mathbf{x} < \theta \end{cases} \tag{5.54}$$

Explanation 5.2 (Definition 5.27).



- 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 $\mathbf{w}^\intercal\mathbf{x} + b$ and as threshold if it is considered to be part of the hyperplane $\theta = -b$, but its just a matter of definition.*

Definition 5.28 (Normalized) Classification Criterion:
$$\tilde{\mathbf{w}}^\intercal\mathbf{x} = \mathbf{w}^\intercal\mathbf{x}y > 0 \qquad \forall (\mathbf{x}, y) \in \mathcal{D} \tag{5.55}$$

Definition 5.29 Linear Separable Data set: A data set is *linearly separable* if there exists a separating hyperplane $\mathcal{H}$ s.t. each label can be assigned correctly:
$$\hat{y} := c(\mathbf{x}) = y \qquad \forall (\mathbf{x}, y) \in \mathcal{D} \tag{5.56}$$

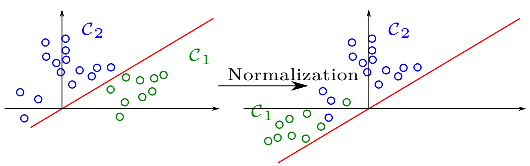
7.1.1. Normalization

Proposition 5.3 Including the Offset: In order to simplify notation the offset is usually included into the parameter vector:
$$\mathbf{w} \leftarrow \begin{pmatrix}\mathbf{w}\\ b\end{pmatrix} \qquad \mathbf{x} \leftarrow \begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix}$$
$$\Rightarrow \qquad \mathbf{w}^\intercal\mathbf{x} = (\mathbf{w}^\intercal \;\; b)\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix} = \mathbf{w}^\intercal\mathbf{x} + b$$

Proposition 5.4 Uniform Classification Criterion:
In order to avoid the case distinction in the classification criterion of eq. (5.54) we may transform the input samples by:
$$\tilde{\mathbf{x}} = \begin{cases} \mathbf{x} & \text{if } \mathbf{w}^\intercal\mathbf{x} > \theta \\ -\mathbf{x} & \text{if } \mathbf{w}^\intercal\mathbf{x} < \theta \end{cases} \tag{5.57}$$

Explanation 5.4 (proposition 5.4).

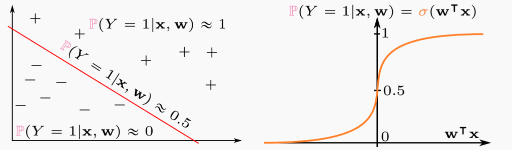
We transform the input s.t. the separating hyper-plane puts all labels on the same "positive" side $\mathbf{w}^\intercal\mathbf{x} > 0$.



Corollary 5.9 : How can we achieve this in practice?
If $\mathcal{Y} = \{-1, 1\}$ then we can simply multiply with the label y_i:
$$\left.\begin{matrix} \mathbf{w}^\intercal\mathbf{x} > 0 & \forall y = +1 \\ \mathbf{w}^\intercal\mathbf{x} < 0 & \forall y = -1 \end{matrix}\right\} \quad \Longleftrightarrow \quad \mathbf{w}^\intercal\mathbf{x} \cdot \hat{y} > 0 \qquad \forall y$$

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

Idea: in order to classify dichotomies[def. 5.21] we use a distribution that maps probabilities to a binary values 0/1
$\Rightarrow$ *Bernoulli Distribution*[def. 34.22].
Problem: we need to convert/translate distance $\mathbf{w}^\intercal\mathbf{x}$ into probability in order to use a bernouli distribution.
Idea: use a sigmoidal function to convert distances $\mathbf{z} := \mathbf{w}^\intercal\mathbf{x}$ into probabilities $\Rightarrow$ *Logistic Function*[def. 5.30].



8.1. Logistic Function

Definition 5.30 Sigmoid/Logistic Function:
$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{\text{neg. dist. from deci. boundary}}} \tag{5.58}$$

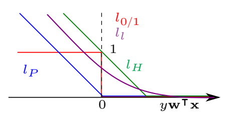
Explanation 5.5 (Sigmoid/Logistic Function).
$$\sigma(z) = \begin{cases} 0 & & -z \text{ large} \\ 1 & \text{if} & z \text{ large} \\ 0.5 & & z = 0 \end{cases}$$

8.2. Logistic Regression

Definition 5.31 Logistic Regression:
models the likelihood of the output y as a Bernoulli Distribution[def. 34.22] $y \sim \mathrm{Bern}(\mathrm{p})$, where the probability p is given by the Sigmoid function[def. 5.30] of a linear regression:
$$\mathrm{p}(y|\mathbf{x}, \mathbf{w}) = \mathrm{Bern}\left(\sigma(\mathbf{w}^\intercal\mathbf{x})\right) = \begin{cases} \frac{1}{1+e^{-\mathbf{w}^\intercal\mathbf{x}}} & \text{if } y = +1 \\ 1 - \frac{1}{1+e^{-\mathbf{w}^\intercal\mathbf{x}}} & \text{if } y = -1 \end{cases}$$
$$\overset{??\;5.11}{=} \frac{1}{1 + \exp(-y \cdot \mathbf{w}^\intercal\mathbf{x})} = \sigma\left(-y \cdot \mathbf{w}^\intercal\mathbf{x}\right) \tag{5.59}$$

8.2.1. Maximum Likelihood Estimate

Definition 5.32 Logistic Loss l_l $\qquad$ proof 5.12:
Is the objective we want to minimize when performing mle[def. 7.3] for a logistic regression likelihood and incurs higher cost for samples closer to the decision boundary:



$$l_l(\mathbf{w}; \mathbf{x}, y) := \log\left(1 + \exp(-y \cdot \mathbf{w}^\intercal\mathbf{x})\right) \tag{5.60}$$
$$\propto \log(1 + e^z) = \begin{cases} z & \text{for large } z \\ 0 & \text{for small } z \end{cases}$$

Corollary 5.10 MLE for Logistic Regression:
$$\mathrm{l}_n(\mathbf{w}) = \sum_{i=1}^{n} l_l = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \cdot \mathbf{w}^\intercal\mathbf{x}_i)\right) \tag{5.61}$$

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. 5.10].

Definition 5.33 Logistic Loss Gradient $\nabla_{\mathbf{w}} l_l(\mathbf{w})$: $\qquad$ proof 5.13
$$\nabla_{\mathbf{w}} l_l(\mathbf{w}) = \mathbb{P}(Y = -y|\mathbf{x}, \mathbf{w}) \cdot (-y\mathbf{x})$$
$$= \frac{1}{1 + \exp(y\mathbf{w}^\intercal\mathbf{x})} \cdot (-y\mathbf{x}) \tag{5.62}$$

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 beeing in the wrong class $\mathbb{P}(Y = -1|\mathbf{x}, \mathbf{w})$. Thus the more likely we are in the wrong class the bigger the step we take:
$$\mathbb{P}(Y = -y|\hat{y} = \mathbf{w}^\intercal\mathbf{x}) = \begin{cases} \uparrow & \textit{take big step} \\ \downarrow & \textit{take small step} \end{cases}$$

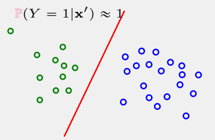
Algorithm 5.1 Vanilla SGD for Logistic Regression:
Initalize: $\mathbf{w}$
1: **for** $1, 2 \dots, T$ **do**
2: $\quad$ Pick $(\mathbf{x}, y)$ unif. at randomn from data $\mathcal{D}$
3: $\quad$ $\mathbb{P}(Y = -y|\mathbf{x}, \mathbf{w}) = \frac{1}{(1+\exp(-y \cdot \mathbf{w}^\intercal\mathbf{x}))} = \sigma(y \cdot \mathbf{w}^\intercal\mathbf{x})$
$\qquad \triangleright$ compute prob. of misclassif. with cur. model
4: $\quad$ $\mathbf{w} = \mathbf{w} + \eta_t y\mathbf{x}\sigma(y \cdot \mathbf{w}^\intercal\mathbf{x})$
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. (5.59):
$$\mathbb{P}(y|\mathbf{x}, \hat{\mathbf{w}}) = \frac{1}{1 + \exp\left(-y\hat{\mathbf{w}}^\intercal\mathbf{x}\right)} \tag{5.63}$$

Drawback

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



8.2.2. Maximum a-Posteriori Estimates

adapt

8.3. Logistic regression and regularization

Adding Priors to Logistic Liklihood

- **L2 (Gaussian prior)**:
$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda\|\mathbf{w}\|_2^2$$
- **L1 (Laplace prior)**:
$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda\|\mathbf{w}\|_1$$
- **Generalized**:
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^\intercal\mathbf{x}_i)\right) + \lambda C(\mathbf{w})$$
$$= \arg\max_{\mathbf{w}} \mathbb{P}(\mathbf{w}|\mathbf{X}, Y)$$

8.4. SGD for L2-gregularized logistic regression

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

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

9. Proofs

Proof 5.11: [def. 5.31] 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 5.12: [def. 5.32]

$$
\begin{aligned}
\mathrm{l}_n(\mathbf{w}) &= \arg\max_{\mathbf{w}} \mathrm{p}(y_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}) = \arg\min_{\mathbf{w}} -\log \mathrm{p}(Y|\mathbf{X}, \mathbf{w}) \\
&\overset{\text{i.i.d.}}{=} \arg\min_{\mathbf{w}} \sum_{i=1}^{n} -\log \mathrm{p}(y_i|\mathbf{x}_i, \mathbf{w}) \\
&\overset{\text{eq. (5.59)}}{=} -\log \frac{1}{1+\exp(-y_i \cdot \mathbf{w}^\intercal \mathbf{x}_i)} \\
&= \log\left(1+\exp(-y_i \cdot \mathbf{w}^\intercal \mathbf{x}_i)\right) =: l_l(\mathbf{w})
\end{aligned}
$$

Proof 5.13: [def. 5.33]

$$
\begin{aligned}
\nabla_{\mathrm{w}} l_l(\mathrm{w}) &= \frac{\partial}{\partial \mathrm{w}} \log\left(1+\exp(-y \cdot \mathrm{w}^\intercal \mathbf{x})\right) \\
&\overset{\text{C.R.}}{=} \frac{1}{(1+\exp(-y \cdot \mathrm{w}^\intercal \mathbf{x}))} \frac{\partial}{\partial \mathrm{w}} \left(1+\exp(-y \cdot \mathrm{w}^\intercal \mathbf{x})\right) \\
&\overset{\text{C.R.}}{=} \frac{1}{(1+\exp(-y \cdot \mathrm{w}^\intercal \mathbf{x}))} \exp(-y \cdot \mathrm{w}^\intercal \mathbf{x}) \cdot (-y\mathbf{x}) \\
&= \frac{e^{-z} \cdot (-yx)}{(1+e^{-z})} = \frac{-yx}{e^{z}(1+e^{-z})} = \frac{-yx}{(e^{z}+e^{-z+z})} \\
&= \frac{1}{\exp(y \cdot \mathrm{w}^\intercal \mathbf{x})+1} \cdot (-yx) \\
&\overset{\text{eq. (5.59)}}{=} \hat{\mathbb{P}}(Y=-y|\mathbf{x}, \mathrm{w}) \cdot (-y\mathbf{x})
\end{aligned}
$$

1. Generalized Additive Models (GAMs)

Definition 6.1 $g_{\mathbf{add}} : \mathbb{R}^p \mapsto \mathbb{R}$

Generalized Additive Models (GAM):
Are generalized linear model where the response variable depends linearly on unknown smooth functions g_j s.t.:

$$g_{\text{add}}(\mathbf{x}) = \mu + \sum_{j=1}^{p} g_j(x_j) \qquad \begin{array}{l} g_j : \mathbb{R} \mapsto \mathbb{R} \\ \mathbb{E}\left[g_j(x_j)\right] = 0 \end{array} \qquad \forall j \in \{1, \ldots, p\} \tag{6.1}$$

Pros
- Does not suffer from the curse of dimensionality.

Cons
- does not allow for interaction terms such as $g_{j,k}(x_j, x_k)$.

1.1. Backfitting

add some point check again Deviance and R squared adj from R output

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 $\{x_1, \ldots, x_n\}$. **Let**: $\theta = (\theta_1 \ \ldots \ \theta_k)^\intercal \in \Theta\mathbb{R}^k$ vector of unknown model parameters.
Consider: a probability density/mass function $f_X(\mathbf{x};\theta)$

Definition 7.1 Likelihood Function $\mathscr{L}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}_+$:
Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ be a random sample of i.i.d. data points drawn from an unknown probability distribution $\mathbf{x}_i \sim p_{\mathcal{X}}$.
The likelihood function gives the likelihood/probability of the *joint probability* of the data $\{x_1, \ldots, x_n\}$ *given* a fixed set of model parameters θ:

$$\mathscr{L}_n(\theta|\mathbf{X}) = \mathscr{L}_n(\theta;\mathbf{X}) = f(\mathbf{X}|\theta) = f(\mathbf{X};\theta) \tag{7.1}$$

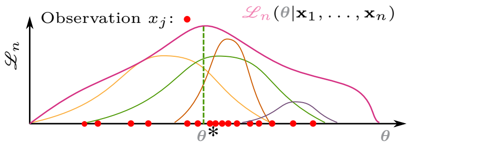


Figure 5: Possible Likelihood function in pink.
Overlayed: possible candidate functions for Gaussian model explaining the observations.

Likelihood function is not a pdf

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

Corollary 7.1 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:

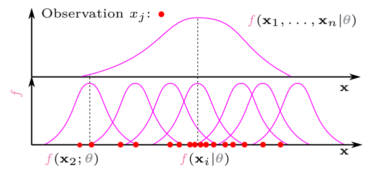


Figure 6: Bottom: probability distributions of the different data points $\mathbf{x}_i$ given a fixed θ for a Gaussian distribution
Top: joint probability distribution of the i.i.d. data points $\{\mathbf{x}_i\}_{i=1}^n$ given a fixed θ

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

Notation

- The probability density $f(\mathbf{X}|\theta)$ is considered for a fixed 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 $\mathscr{L}_n(\theta|\mathbf{X})$.
- 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 randomn variable.

1.2. Maximum Likelihood Estimation (MLE)

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

Definition 7.2 Log Likelihood Function $\mathrm{l}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}$:

$$\mathrm{l}_n(\theta|\mathbf{X}) = \log \mathscr{L}_n(\theta|\mathbf{X}) = \log f(\mathbf{X}|\theta) \tag{7.2}$$

Corollary 7.2 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. (24.65):

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

Definition 7.3 Maximum Likelihood Estimator θ^*:
Is the estimator $\theta^* \in \Theta$ that maximizes the likelihood of the model/predictor:

$$\theta^* = \operatorname*{arg\,max}_{\theta\in\Theta} \mathscr{L}_n(\theta;\mathbf{x}) \quad \text{or} \quad \theta^* = \operatorname*{arg\,max}_{\theta\in\Theta} \mathrm{l}_n(\theta;\mathbf{x}) \tag{7.3}$$

1.3. Maximization vs. Minimization

For optimization problems we minimize by convention.
The logarithm is a concave function[def. 24.28] $\cap$, thus if we calculate the extremal point we will obtain a maximum.
If we want to calculate a mimimum instead (i.e. in order to be compatible with some computer algorithm) we can convert the function into a convex functionsection 5 $\cup$ by multiplying it by minus one and consider it as a loss function instead of a likelihood.

Definition 7.4 Negative Log-likelihood $-\mathrm{l}_n(\theta|\mathbf{X})$:

$$\theta^* = \operatorname*{arg\,max}_{\theta\in\Theta} \mathrm{l}_n(\theta|\mathbf{X}) = \operatorname*{arg\,min}_{\theta\in\Theta} -\mathrm{l}_n(\theta|\mathbf{X}) \tag{7.4}$$

1.4. Conditional Maximum Likelihood Estimation

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

Note

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

Definition 7.5 Conditional (log) likelihood function:
Models the liklihood of a model with parameters θ given the data $\mathbf{Z} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

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

2. Maximum a posteriori estimation (MAP)

Idea

We have seen (??), that trading/increasing a bit of bias can lead to a big reduction of variance of the generalization error.
We also know that the least squares MLE is unbiased (??).
Thus the question arises if we can introduce a bit of bias into the MLE in turn of decreasing the variance?
$\Rightarrow$ use Bayes rule (??) to introduce a bias into our model via. a Prior distribution.

2.1. Prior Distribution

Definition 7.6 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:

$$\theta \sim \pi(\theta) = p(\theta) \tag{7.5}$$

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. 4.12]) θ.

Corollary 7.3 The prior is independent of the data: The prior $p(\theta)$ models a prior belief/bias and is thus independent of the data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$:

$$p(\theta|\mathbf{X}) = p(\theta) \tag{7.6}$$

Definition 7.7 Hyperparameters $p_{\boldsymbol{\lambda}}(\theta)$:
In most cases the prior distribution are parameterized that is the pdf $\pi(\theta|\boldsymbol{\lambda})$ depends on a set of parameters $\boldsymbol{\lambda}$.
The parameters of the prior distribution, are called hyperparameters and are supplied due to believe/prior knowledge (and do not depend on the data) see example 7.1

2.2. Posterior Distribution

Definition 7.8 Posterior Distribution $p(\theta|\mathbf{data})$:
The posterior distribution $p(\theta|\text{data})$ is a probability distribution that describes the relationship of a unknown *parameter* θ a posterior/after observing evidence of a random quantity $\mathbf{Z}$ that is in a relation with θ:

$$p(\theta|\mathbf{data}) = p(\theta|\mathbf{Z}) \tag{7.7}$$

Definition 7.9
Posterior Distribution and Bayes Theorem:
Using Bayes theorem 33.3 we can write the posterior distribution as a product of the *likelihood*[def. 7.1] weighted with our *prior*[def. 7.6] and normalized by the *evidence* $\mathbf{Z} = \{\mathbf{X}, \mathbf{y}\}$ s.t. we obtain a real probability distribution:

$$p(\theta|\mathbf{data}) = p(\theta|\mathbf{Z}) = \frac{p(\mathbf{Z}|\theta) \cdot p_{\boldsymbol{\lambda}}(\theta)}{p(\mathbf{Z})} \tag{7.8}$$

$$\text{Posterior} = \frac{\text{Liklihood} \cdot \text{Prior}}{\text{Normalization}} \tag{7.9}$$

$$\boxed{p(\theta|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\theta, \mathbf{X}) \cdot p_{\boldsymbol{\lambda}}(\theta)}{p(\mathbf{y}|\mathbf{X})}} \tag{7.10}$$

see proof ?? 7.1

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 7.10
Maximum a-Posteriori Estimates (MAP):
Is model/parameters θ that maximize the posterior probability distribution:

$$\boxed{\theta^*_{\text{MAP}} = \operatorname*{arg\,max}_{\theta} \mathbb{P}(\theta|\mathbf{X}, \mathbf{y})} \tag{7.11}$$

Log-MAP estimator:

$$\begin{aligned}
\theta^* &= \operatorname*{arg\,max}_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \qquad (7.12)\\
&= \operatorname*{arg\,max}_{\theta} \left\{\frac{p(\mathbf{y}|\mathbf{X}, \theta) \cdot p_{\boldsymbol{\lambda}}(\theta)}{p(\mathbf{y}|\mathbf{X})}\right\}\\
&\overset{\text{eq. (24.62)}}{\propto} \operatorname*{arg\,max}_{\theta} \left\{p(\mathbf{y}|\theta, \mathbf{X}) \cdot p_{\boldsymbol{\lambda}}(\theta)\right\}
\end{aligned}$$

Corollary 7.4 Negative Log MAP:

$$\begin{aligned}
\theta^* &= \operatorname*{arg\,max}_{\theta} \{p(\theta|\mathbf{X}, \mathbf{y})\} \qquad (7.13)\\
&= \operatorname*{arg\,min}_{\theta} -\log \overbrace{p(\theta)}^{\text{Prior}} - \log \overbrace{p(\mathbf{y}|\theta, \mathbf{X})}^{\text{Likelihood}} + \underbrace{\cancel{\log p(\mathbf{y}|\mathbf{X})}}_{\text{not depending on } \theta}
\end{aligned}$$

3. Proofs

Proof 7.1: 7.10:

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

$$\begin{aligned}
p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y}) &= p(\theta|\mathbf{X}, \mathbf{y})p(\mathbf{y}|\mathbf{X})p(\mathbf{X})\\
p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta) &= p(\mathbf{y}|\mathbf{X}, \theta)p(\mathbf{X}, \theta)\\
&= p(\mathbf{y}|\mathbf{X}, \theta)p(\theta|\mathbf{X})p(\mathbf{X})\\
&\overset{\text{eq. (7.6)}}{=} p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)p(\mathbf{X})
\end{aligned}$$

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

Note

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

4. Examples

Example 7.1 Hyperparameters Gaussian Prior:

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

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

Bayesian Inference/Modeling

Definition 7.11 Bayesian Inference: So far we only really looked at point estimators/estimates[def. 36.8].
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. 4.10], where uncertainty in inferences is quantified using probability. Thus we usually obtain a distribution over our parameters and not a single point estimates
⇒ can deduce statistical properties of parameters from their distributions.

Definition 7.12 $p(w|\mathbf{y}, \mathbf{X})/p(w|\mathcal{D})$
Posterior Probability Distribution:
1. Specify the prior $p_\lambda(w)$
2. Specify the likelihood $p(\mathbf{y}|w, \mathbf{X})/p(\mathcal{D}|w)$
3. Calculate the evidence $p(\mathbf{y}|\mathbf{X})/p(\mathcal{D})$
4. Calculate the posterior distribution $\mathbb{P}(w|\mathbf{y}, \mathbf{X})/p(w|\mathcal{D})$

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

Definition 7.13 $p(\mathbf{y}|\mathbf{X})/p(\mathcal{D})$
Marginal Likelihood [see proof 11.2]:
is the normalization constant that makes sure that the posterior distribution[def. 7.12] is an true probability distribution:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|w, \mathbf{X}) \cdot p_\lambda(w)\, dw = \int \text{Likelihood} \cdot \text{Prior}\, dw \tag{7.14}$$

Note
It is called marginal likelihood as we marginalize over w.

Definition 7.14 Posterior Marginal Distribution: Is the posterior distribution of single elements of our thought after parameter vector:

$$p(w_i|\mathbf{y}, \mathbf{X}) = \int p(\mathbf{y}|w, \mathbf{X})\, dw_{-i} \qquad i = 1, \dots \dim(w) \tag{7.15}$$

Definition 7.15 $p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})/p(\mathbf{f}_*|\mathbf{y})$ [see proof 11.1]
Posterior Predictive Distribution:
is the distribution of a real process $\mathbf{f}$ (i.e. $f(x) = \mathbf{x}^\mathsf{T} w$) given:
- new observation(s) $\mathbf{x}_*$
- the posterior distribution[def. 7.12] of the observed data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$
- The likelihood of a real process $\mathbf{f}_*$

$$p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{x}_*, w) \cdot p(w|\mathbf{X}, \mathbf{y})\, dw \tag{7.16}$$

it is calculated by weighting the likelihood[def. 7.1] of the new observation $\mathbf{x}_*$ with the posterior of the observed data and averaging over all parameter values w.
⇒ obtain a distribution not depending on w.

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

5. Types of Uncertainty

Definition 7.16 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 7.17 Aleatoric/Statistical Uncertainty:
Is the uncertainty of an underlying random process/model. The aleatroic uncertainty stems from the fact that we are create random process models. If we run our *trained* model multiple times with *the same* input $\mathbf{X}$ data we will end up with different outcomes $\hat{y}$.
The aleatoric noise is *irreducible* as it is an underlying part of probabilistic models.

Bayesian Filtering

Definition 8.1
Recursive Bayesian Estimation/Filtering: Is a technique for estimating the an unknown probability distribution recursively over time by a measurement-[def. 8.3] and a process-model[def. 8.2] using Bayesian inference[def. 7.11].

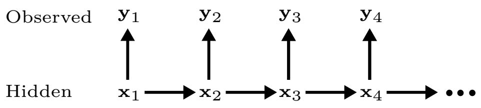


Figure 7: This problem corresponds to a *hidden Markov model (HMM)*[def. 15.1]

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

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

Definition 8.2 $\mathbf{x}_{t+1} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1})$
Process/Motion/Dynamic Model: is a model q of how our system state $\mathbf{x}_t$ evolves and is usually fraught with some uncertainty.

Corollary 8.1 Markov Property $\mathbf{x}_t \perp\!\!\!\perp \mathbf{x}_{1:t-2}|\mathbf{x}_{t-1}$: The process models[def. 8.2] is Markovian[def. 37.14] i.e. the current state depends only on the previous state:

$$p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{8.1}$$

Definition 8.3 $\mathbf{y}_t \sim p(\mathbf{y}_t|\mathbf{x}_t)$
Measurement/Sensor–Model/Likelihood: is a model h that maps observations/sensor measurements of our model $\mathbf{y}_t$ to the model state $\mathbf{x}_t$

Corollary 8.2 $\mathbf{y}_t \perp\!\!\!\perp \mathbf{y}_{1:t-1}\mathbf{x}_{1:t-1}|\mathbf{x}_t$
Conditional Independent Measurements: The measurements $\mathbf{y}_t$ are conditionally independent of the previous obser-vations $\mathbf{y}_{1:t-1}$ given the current state $\mathbf{x}_t$:

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{x}_t) = p(\mathbf{y}_t|\mathbf{x}_t) \tag{8.2}$$

Goal
We want to combine the process model[def. 8.2] and the measurement model[def. 8.3] in a recursive way to obtain a good estimate of our model state:

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

Definition 8.4 Chapman-Kolmogorov eq. $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$
Prior Update/Prediction Step [proof 11.3]:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1} \tag{8.3}$$

Prior Distribution:

$$p(\mathbf{x}_0|\mathbf{y}_{0-1}) = p(\mathbf{x}_0) = p_0 \tag{8.4}$$

Definition 8.5 $p(\mathbf{x}_t|\mathbf{y}_{1:t})$
Posterior Distribution/Update Step [proof 11.4]:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{Z_t} p(\mathbf{y}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \tag{8.5}$$

Definition 8.6 Normalization [see proof 11.5]:

$$Z_t = p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\, d\mathbf{x}_t \tag{8.6}$$

Algorithm 8.1 Optimal Bayesian Filtering:
1: **Input:** $p(\mathbf{x}_0)$
2: **while** Stopping Criterion not full-filed **do**
3: **Prediction Step:**
$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{Z_t} p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$$
4: **Update Step:**
$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}$$
with:
$$Z_t = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\, d\mathbf{x}_t$$
5: **end while**

Corollary 8.3 [proof 11.6]
Joint Probability Distribution of (HMM): we can also calculate the joint probability distribution of the (HMM):

$$p(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) = p(\mathbf{x}_1)p(\mathbf{y}_1|\mathbf{x}_1)\prod_{i=2}^{t} p(\mathbf{x}_i|\mathbf{x}_{i-1})p(\mathbf{y}_i|\mathbf{x}_i) \tag{8.7}$$

Example 8.1 Types of Bayesian Filtering:
- **Kalman Filter**: assumes a *linear* system, q, h are linear and Gaussian noise $\mathbf{v}, \mathbf{w}$.
- **Extended Kalman Filter**: assumes a *non-linear* system, q, h are non-linear and Gaussian noise $\mathbf{v}, \mathbf{w}$.
- **Particle Filter**: assumes a *non-linear* system q, h are non-linear and Non-Gaussian noise $\mathbf{v}, \mathbf{w}$, especially multi-modal distributions.

1. Kalman Filters

Definition 8.7 Kalman Filter Assumptions: Assumes a *linear*[def. 24.18] process model[def. 8.2], $\mathbf{q}$ with Gaussian model-noise $\mathbf{v}$ and a linear measurement model[def. 8.3] $\mathbf{h}$ with Gaussian process-noise $\mathbf{w}$.

add difference of prediction, filtering and smoothing for posterior marginals

Definition 8.8 Kalman Filter Model:
Process Model (8.8)

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

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

Measuremnt Model (8.9)

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

and define:

$$\hat{x}_p^{(k)} := \mathbb{E}[\mathbf{x}_p^{(k)}] \quad \textbf{and} \quad P_p^{(k)} := \mathbb{V}\left[\mathbf{x}_p^{(k)}\right] \tag{8.10}$$

$$\hat{x}_m^{(k)} := \mathbb{E}[\mathbf{x}_m^{(k)}] \quad \textbf{and} \quad P_m^{(k)} := \mathbb{V}\left[\mathbf{x}_m^{(k)}\right] \tag{8.11}$$

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

add Kalman algorithm (in slides Joesph Form I think)

Gaussian Processes (GP)

1. Gaussian Process Regression

add complexitit On3 due to inverse

1.1. Gaussian Linear Regression

Given

1. Linear Model with Gaussian Noise:

$$f(\mathbf{x}) = \mathrm{w}^\intercal \mathbf{x} \qquad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2 \mathbf{I}\right) \tag{9.1}$$
$$\mathbf{y} = f(\mathbf{x}) + \epsilon$$

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

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

Sought

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

Definition 9.1 $\quad p(\mathrm{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\bar{\mathrm{w}}, \Sigma_{\mathrm{w}}^{-1}\right)$

Posterior Distribution proof 11.7:

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

Note

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

Definition 9.2 $\quad p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \Sigma_*)$

Posterior Predictive Distribution proof 11.8:

$$\mu_* = \frac{1}{\sigma^2}\mathbf{x}_*^\intercal \Sigma_{\mathrm{w}}^{-1}\mathbf{X}\mathbf{y} \qquad \Sigma_* = \mathbf{x}_*^\intercal \Sigma_{\mathrm{w}}^{-1}\mathbf{x}_* \tag{9.2}$$

1.2. Kernelized Gaussian Linear Regression

Definition 9.3 Posterior Predictive Distribution:

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \Sigma_*) \tag{9.3}$$
$$\mu_* \tag{9.4}$$

Definition 9.4 Gaussian Process:

2. Model Selection

2.1. Marginal Likelihood

Approximate Inference

Problem

In statistical inference we often want to calculate integrals of probability distributions i.e.

- Expectations

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

- Normalization constants:

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

For non-linear distributions this integrals are in general intractable which may be due to the fact that there exist no analytic form of the distribution we want to integrate or highly dimensional latent spaces that prohibits numerical integration (curse of dimensionality).

Definition 10.1 Approximate Inference: Is the procedure of finding an probability distribution q that approximates a true probability distribution p as well as possible.

1. Variational Inference

Definition 10.2 Bayes Variational Inference: Given an unormalized (posterior) probability distribution:

$$p(\theta|y) = \frac{1}{Z}p(\theta, y) \tag{10.1}$$

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

Definition 10.3 Variational Family of Distributions $\mathcal{Q}$: a set of probability distributions $\mathcal{Q}$ that is parameterized by the same *variational parameter* λ is called a variational familiy.

1.1. Laplace Approximation

Definition 10.4 [example 11.1], [proof 11.9,11.10,11.11]
Laplace Approximation: Tries to approximate a desired probability distribution $p(\theta|\mathcal{D})$ by a Gaussian probability distribution:

$$\mathcal{Q} = \{q_\lambda(\theta) = \mathcal{N}(\lambda)\} = \mathcal{N}(\mu, \Sigma)\} \tag{10.2}$$

the distribution is given by:

$$q(\theta) = c\cdot\mathcal{N}(\theta; \lambda_1, \lambda_2) \tag{10.3}$$

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

Note

The name *Laplace Approximation* comes from its inventor *Pierre-Simon Laplace.*

Corollary 10.1 : Taylor approximation of a function $p(\theta|y) \in \mathcal{C}^k$ around its mode $\hat{\theta}$ naturally induces a Gaussian approximation. See proofs 11.9,11.10,11.11

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 p:

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

Note

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

Definition 10.5 ELBO-Optimization Problem [proof 11.12]:

$$q_\lambda^* \in \arg\min_{\{\lambda: q_\lambda\in\mathcal{Q}\}} \mathrm{KL}\left(q_\lambda(\theta)\,\|\,p(\theta|y)\right)$$
$$= \arg\max_{\{\lambda: q_\lambda\in\mathcal{Q}\}} \mathbb{E}_{\theta\sim q_\lambda}[\log p(y,\theta)] + H(q_\lambda) \tag{10.4}$$
$$= \arg\max_{\{\lambda: q_\lambda\in\mathcal{Q}\}} \mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] - \mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta)) \tag{10.5}$$
$$:= \arg\max_{\{\lambda: q_\lambda\in\mathcal{Q}\}} \mathrm{ELBO}(\lambda) \tag{10.6}$$

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

Explanation 10.1.

- *eq.* (10.4)*:*
  - *prefer uncertain approximations i.e. we maximize $H(q)$*
  - *that jointly make the joint posterior likely*
- *eq.* (10.6)*: 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 10.6 example 11.2/proof 11.13
Expected Lower Bound of Evidence (ELBO):
The evidence lower bound is a bound on the log prior:

$$\mathrm{ELBO}(q_\lambda) \leqslant \log p(y) \tag{10.7}$$

1.3.1. Maximizing The ELBO

Definition 10.7 Gradient of the ELBO Loss:

$$\nabla_\lambda L(\lambda) = \nabla_\lambda \mathrm{ELBO}(\lambda) \tag{10.8}$$
$$= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}[\log p(y,\theta)] + H(q_\lambda)\right]$$
$$= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] - \mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta))\right]$$
$$= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta))$$

Problem

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

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

however in eq. (10.8) only second term can be derived easily. For the first term we cannot move the gradient inside the expectation as the expectations depends on the parameter w.r.t. which we differentiate:

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

Solutions:

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

1.4. The Reparameterization Trick

Principle 10.1 proof 11.14
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 $\epsilon \sim \phi(\epsilon)$:

$$\theta \sim q(\theta|\lambda) = \phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1} \tag{10.9}$$

we can then write by the law of the unconscious statistician law 33.6:

$$\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] = \mathbb{E}_{\epsilon\sim\phi}[\log p(y|g(\epsilon;\lambda))] \tag{10.10}$$

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

$$\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] = \nabla_\epsilon\mathbb{E}_{\theta\sim\phi}[\log p(y|g(\epsilon;\lambda))] \tag{10.11}$$
$$= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log p(y|g((\epsilon;\lambda)))] \tag{10.12}$$

Definition 10.8 example 11.3
Reparameterized ELBO Gradient[def. 10.7]**:**
By using the reparameterization trick principle 10.1 we can write the gradient of the ELBO as:

$$\nabla_\lambda L(\lambda) = \nabla_\lambda\mathrm{ELBO}(\lambda) \tag{10.13}$$
$$= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta))$$
$$= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log p(y|g((\epsilon;\lambda)))] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta))$$

Corollary 10.2 proof 11.3
Reparameterized ELBO for Gaussians:

$$\nabla_\lambda L(\lambda) = \nabla_\lambda\mathrm{ELBO}(\lambda) \tag{10.14}$$
$$= \nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[\log p(y|\theta)] - \nabla_\lambda\mathrm{KL}(q_\lambda(\theta)\,\|\,p(\theta))$$
$$= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\nabla_{\mathbf{C},\mu}\log p(y|\mathbf{C}\epsilon + \mu)\right]$$
$$- \nabla_{\mathbf{C},\mu}\mathrm{KL}\left(q_{\mathbf{C},\mu}\,\|\,p(\theta)\right)$$
$$\approx \frac{n}{m}\sum_{j=1}^m \nabla_{\mathbf{C},\mu}\log p\left(y_{i_j}|\mathbf{C}\epsilon^j + \mu, \mathbf{x}_{i_j}\right)$$
$$- \nabla_{\mathbf{C},\mu}\mathrm{KL}\left(q_{\mathbf{C},\mu}\,\|\,p(\theta)\right)$$

2. Markov Chain Monte Carlos Methods

Definition 10.9
Markov Chain Monte Carlo (MCMC) Methods:

3. Integrated Nested Laplace Approximation

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(\mathbf{u}_{ji}) + \sum_{k=1}^{n_\beta}\beta_k z_{ki} + \epsilon_i \tag{10.15}$$
$$p(\mathbf{x},\theta)p(\mathbf{y}) = p(\mathbf{x}) \tag{10.16}$$
$$p(\mathbf{x}_i|\mathbf{y}) = \int p(x_i|\theta,\mathbf{y})\underline{p(\theta|\mathbf{y})}\,\mathrm{d}\theta$$
$$\rightarrow \tilde{p}(\mathbf{x}_i|\mathbf{y}) = \int \tilde{p}(x_i|\theta,\mathbf{y})\tilde{p}(\theta|\mathbf{y})\,\mathrm{d}\theta$$
$$p(\theta_j|\mathbf{y}) = \int p(\theta|\mathbf{y})\,\mathrm{d}\theta_{-j}$$
$$\rightarrow \tilde{p}(\theta_j|\mathbf{y}) = \int \tilde{p}(\theta|\mathbf{y})\,\mathrm{d}\theta_{-j}$$

$p(x_i|\theta,\mathbf{y})$ and $p(\theta|\mathbf{y})$ are approximated and the *posterior marginal* densities are then calculated using numerical integration:

Note

The numerical integration is possible if θ is small i.e. $m = \dim(\theta) \leqslant 5$.

4. Approximationing $p(\theta|\mathbf{y})$ and $p(x_i|\mathbf{y})$

$$p(\mathbf{x},\theta,\mathbf{y}) = p(\mathbf{x}|\theta,\mathbf{y})\mathbb{P}(\theta,\mathbf{y}) = p(\mathbf{x}|\theta,\mathbf{y})\mathbb{P}(\theta|\mathbf{y})p(\mathbf{y})$$
$$\Rightarrow \quad \tilde{p}(\theta|\mathbf{y}) = \frac{p(\mathbf{x},\theta,\mathbf{y})}{\tilde{p}(\mathbf{x}|\theta,\mathbf{y})p(\mathbf{y})} \propto \left.\frac{p(\mathbf{x},\theta,\mathbf{y})}{\tilde{p}_G(\mathbf{x}|\theta,\mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\theta)}$$

1. Marginal Posterior of the latent field $p(\mathbf{x}_i|\mathbf{y})$ are calculated by first approximating $p(\theta|\mathbf{y})$:

$$\underline{p(\theta|\mathbf{y})_G} = \mathcal{N}\left(x_i;\mu_i(\theta),\sigma_i^2(\theta)\right)$$

and then numerical integration w.r.t. θ:

$$\tilde{p}(\mathbf{x}_i|\mathbf{y}) = \sum_k \underline{p_G(\theta_k|\mathbf{y})}\tilde{p}(\theta_k|\mathbf{y})\Delta_k$$

Note

$\tilde{p}(\theta|\mathbf{y})$ is usually quiet different from a Gaussian s.t. the Gaussian approximation alone is not really sufficient.

Bayesian Neural Networks (BNN)

Definition 11.1 Bayesian Neural Networks (BNN):

1. Model the prior over our weights $\theta = \begin{bmatrix}\mathbf{W}^0 & & \mathbf{W}^L\end{bmatrix}$ by a neural network:

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

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

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

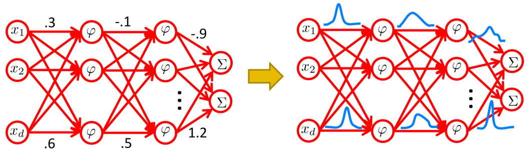


Figure 8

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

$$p(y|F(\theta, \mathbf{X})) \qquad \text{see example 11.4} \tag{11.1}$$

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 8. 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 11.2 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 $\sigma(\mathbf{x}_i; \theta)^2$:

$$\begin{aligned}\theta^* &= \arg\max_\theta \{p(\theta|\mathbf{X}, \mathbf{y})\} \overset{\text{eq. (7.13)}}{=} \arg\min_\theta \lambda \|\theta\|_2^2 \\ &- \sum_{i=1}^n \left(\frac{1}{2\sigma(\mathbf{x}_i; \theta)^2}\|y_i - \mu(\mathbf{x}_i; \theta)\|^2 + \frac{1}{2}\log \sigma(\mathbf{x}_i; \theta)^2\right)\end{aligned}$$

Explanation 11.1. [def. 11.2]

- $\frac{1}{2}\log\sigma(\mathbf{x}_i; \theta)^2$: *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 11.3 proof 11.15
MAP Gradient of BNN:

$$\theta_{t+1} = \theta_t(1 - 2\lambda\eta_t) - \eta_t \nabla \sum_{i=1}^n \log p(y_i|\mathbf{x}_i, \theta) \tag{11.2}$$

Note

- The gradients of the objective eq. (11.2) can be calculated using auto-differentiation techniques e.g. Pytorch or Tensorflow.
- The BNN MAP estimate fails to predict epistemic uncertainty[def. 7.16] $\iff$ it is overconfident in regions where we haven't 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. (10.14) as loss in order to perform back propagation.

0.2. Making Predictions

Proposition 11.1 Title:

1. Proofs

Proof 11.1: Definition 7.15:

$$\begin{aligned}p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \frac{p(\mathbf{f}_*, \mathbf{x}_*, \mathbf{X}, \mathbf{y})}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &= \frac{\int p(\mathbf{f}_*, \mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})\,\mathrm{d}\mathrm{w}}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &\overset{\text{eq. (33.19)}}{=} \frac{\int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})\,\mathrm{d}\mathrm{w}}{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})} \\ &\overset{\text{eq. (33.19)}}{=} \frac{\int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathrm{w}|\mathbf{x}_*, \mathbf{X}, \mathbf{y})\cancel{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})}\,\mathrm{d}\mathrm{w}}{\cancel{p(\mathbf{x}_*, \mathbf{X}, \mathbf{y})}} \\ &= \int p(\mathbf{f}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathrm{w})p(\mathrm{w}|\mathbf{x}_*, \mathbf{X}, \mathbf{y})\,\mathrm{d}\mathrm{w} \\ &\overset{\clubsuit}{=} \int p(\mathbf{f}_*|\mathbf{x}_*, \mathrm{w})p(\mathrm{w}|\mathbf{X}, \mathbf{y})\,\mathrm{d}\mathrm{w}\end{aligned}$$

Note ♣

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

Proof 11.2: Definition 7.13:

$$\begin{aligned}p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}, \mathrm{w}|\mathbf{X})\,\mathrm{d}\mathrm{w} = \int p(\mathbf{y}|\mathrm{w}, \mathbf{X})p(\mathrm{w}|\mathbf{X})\,\mathrm{d}\mathrm{w} \\ &\overset{\text{eq. (7.6)}}{=} \int p(\mathbf{y}|\mathrm{w}, \mathbf{X})p(\mathrm{w})\,\mathrm{d}\mathrm{w}\end{aligned}$$

Proof 11.3: Definition 8.4:

$$\begin{aligned}p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) &\overset{\text{eq. (33.19)}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t_1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1}) \\ &\overset{\text{independ.}}{=} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t_1})\end{aligned}$$

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

Proof 11.4: Definition 8.5:

$$p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{y}_{1:t-1}) \overset{\text{eq. (33.23)}}{=} \begin{cases} p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1})p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \\ p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\end{cases}$$

$$p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1}) \overset{\text{[cor. 8.2]}}{=} p(\mathbf{y}_t|\mathbf{x}_t)$$

from which follows immediately eq. (8.5).

Proof 11.5: Definition 8.6:

$$\begin{aligned}p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) &= \int p(\mathbf{y}_t, \mathbf{x}_t|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_t \\ &= \int p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_t \\ &\overset{\text{[cor. 8.2]}}{=} \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\,\mathrm{d}\mathbf{x}_t\end{aligned}$$

Proof 11.6: [cor. 8.3]:

$$\begin{aligned}p(\mathbf{x}_{1:t}, \mathbf{y}_{1:t}) &\overset{\text{eq. (33.19)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})p(\mathbf{x}_{1:t}) \\ &\overset{\text{law 33.2}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})p(\mathbf{x}_t|\mathbf{x}_{t-1:0})\cdots p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_1) \\ &\overset{\text{eq. (8.1)}}{=} p(\mathbf{y}_{1:t}|\mathbf{x}_{1:t})\left(p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1})\right) \\ &\overset{\substack{\text{law 33.2} \\ \text{[cor. 8.2]}}}{=} \left(p(\mathbf{y}_1|\mathbf{x}_1)\cdots p(\mathbf{y}_t|\mathbf{x}_t)\right)\left(p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{x}_i|\mathbf{x}_{i-1})\right) \\ &= p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1)\prod_{2=1}^t p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\mathbf{x}_{i-1})\end{aligned}$$

Proof 11.7: [def. 9.1]

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

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

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

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

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

Proof 11.8: [def. 9.2]

Proof 11.9: [def. 10.4] In a Bayesian setting we are usually interested in maximizing the log prior+likelihood:

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

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

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

we can no derive the distribution:

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

Notes

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

Proof 11.10: [def. 10.4] 2D:

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

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

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

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

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

Proof 11.11: [def. 10.4] k-dimensional:

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

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

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

Proof 11.12: [def. 10.5]

$$q^* \in \underset{q \in Q}{\arg\min}\, \mathrm{KL}\left(q(\theta) \parallel \mathrm{p}(\theta|y)\right)$$

$$\overset{\mathrm{p}(\theta|y)=\frac{1}{Z}\mathrm{p}(\theta,y)}{=} \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q}\left[\log \frac{q(\theta)}{\frac{1}{Z}\mathrm{p}(\theta,y)}\right]$$

$$= \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q}\left[\log q(\theta) - \log\frac{1}{Z} - \log \mathrm{p}(\theta,y)\right]$$

$$= \underset{q}{\arg\min}\, \mathbb{E}_{\theta\sim q} - \underbrace{\left[-\log q(\theta)\right]}_{H(q)} + \cancel{\mathbb{E}_{\theta\sim q}\left[\log Z\right]}$$

$$- \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta,y)\right]$$

$$= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta,y)\right] + H(q)$$

$$= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta|y) + \log \mathrm{p}(\theta) - \log q(\theta)\right]$$

$$= \underset{q}{\arg\max}\, \mathbb{E}_{\theta\sim q}\left[\log \mathrm{p}(\theta|y)\right] + \mathrm{KL}\left(q(\theta) \parallel \mathrm{p}(\theta)\right)$$

Proof 11.13: [def. 10.6]

$$\log \mathrm{p}(y) = \log \int \mathrm{p}(y,\theta)\,\mathrm{d}\theta = \log \int \mathrm{p}(y|\theta)\mathrm{p}(\theta)\,\mathrm{d}\theta$$

$$= \log \int \mathrm{p}(y|\theta)\frac{\mathrm{p}(\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\,\mathrm{d}\theta$$

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

$$\overset{eq.\ (33.55)}{\geqslant} \mathbb{E}_{\theta\sim q_\lambda}\left[\log\left(\mathrm{p}(y|\theta)\frac{\mathrm{p}(\theta)}{q_\lambda(\theta)}\right)\right]$$

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

$$= \mathbb{E}_{\theta\sim q_\lambda}\left[\log \mathrm{p}(y|\theta)\right] - \mathrm{KL}\left(q_\lambda \parallel \mathrm{p}(\cdot)\right)$$

Proof 11.14: principle 10.1 Let:

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

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

$$\theta \sim q_\lambda(\theta) = q(\theta|\lambda) = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|}$$

$$= \phi(\epsilon)\left|\nabla_\epsilon g(\epsilon;\lambda)\right|^{-1}$$

$\Rightarrow$ parameterized in terms of ϵ

Proof 11.15: [def. 11.3]

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

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

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

2. Examples

Example 11.1 Laplace Approximation Logistic Regression Likelihood + Gaussian Prior:

Example 11.2 ELBO Bayesian Logistic Regression: Suppose:

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

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

Then it follows for the terms of the ELBO:

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

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

$$= \mathbb{E}_{\theta\sim q_\lambda}\left[-\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i\theta^\intercal \mathbf{x}_i\right)\right)\right]$$

Example 11.3 ELBO Gradient Gaussian: Suppose:

$$\theta \sim q(\theta|\lambda) = \mathcal{N}(\theta;\boldsymbol{\mu},\Sigma) \quad \Rightarrow \quad \lambda = \begin{bmatrix}\boldsymbol{\mu} & \Sigma\end{bmatrix}$$

$$\boldsymbol{\epsilon} \sim \phi(\epsilon) = \mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})$$

we can reparameterize using principle 10.1 by using:

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

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

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

from this it follows:

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

$$\iff \phi(\epsilon) = q(\theta|\lambda)|C|$$

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

$$\nabla_\lambda L(\lambda)_1 = \nabla_\lambda \mathbb{E}_{\epsilon\sim\phi}\left[\log \mathrm{p}\left(y|g(\epsilon;\lambda)\right)\right]$$

$$= \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\log \mathrm{p}(y|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu})\right]$$

$$\overset{\text{i.i.d.}}{=} \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\sum_{i=1}^{n} \log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]$$

$$= \nabla_{\mathbf{C},\mu}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[n\frac{1}{n}\sum_{i=1}^{n} \log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]$$

$$= \nabla_{\mathbf{C},\mu} n\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbb{E}_{i\sim\mathcal{U}(\{1,n\})} \log \mathrm{p}(y_i|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu},\mathbf{x}_i)\right]$$

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

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

$$\nabla_\lambda L(\lambda) = \nabla_\lambda \mathrm{ELBO}(\lambda) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\nabla_{\mathbf{C},\mu} \log \mathrm{p}(y|\mathbf{C}\boldsymbol{\epsilon} + \boldsymbol{\mu})\right]$$

$$- \nabla_{\mathbf{C},\mu}\left(q_{\mathbf{C},\boldsymbol{\mu}} \parallel \mathrm{p}(\theta)\right)$$

Example 11.4 BNN Likelihood Function Examples:

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

Kernels

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

Definition 12.1 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 12.2 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 sqaured norm of two difference vectors:

$$\|\mathbf{x}-\mathbf{y}\|^2 = \mathbf{x}^\intercal\mathbf{x} + \mathbf{y}^\intercal\mathbf{y} - 2\mathbf{x}^\intercal\mathbf{y} \qquad \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d \tag{12.1}$$
$$\text{dissim}(\mathbf{x},\mathbf{y}) = \text{sim}(\mathbf{x},\mathbf{x}) + \text{sim}(\mathbf{y},\mathbf{y}) - 2\text{dissim}(\mathbf{x},\mathbf{y})$$

Attention

It is better to rely on similarity measures instead of dissimilarity measures. Dissimilarities are often not adequat from a modeling point of view, because for objects that are really dissimilar/far from each other, we usually have the biggest problem to estimate their distance.
E.g. for a bag of words it is easy to determine similar words, but it is hard to estimate which words are most dissimilar. For normed vectors the only information of a dissimilarity defined as in eq. (12.1) becomes $2\mathbf{x}^\intercal\mathbf{y} = 2\text{dissim}(\mathbf{x},\mathbf{y})$

Definition 12.3 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:

$$D = \dim(\mathcal{V}) = \binom{p+d}{d} = \mathcal{O}\left(d^p\right) \tag{12.2}$$

when using the polynomial kernel[def. 12.10], this can be reduced to the order d.

Definition 12.4 Kernel k: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the data space.
A map $\mathbf{k} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called kernel if their exists an inner product space[def. 27.76] called **feature space** $(\mathcal{V}, \langle\cdot,\cdot\rangle_{\mathcal{V}})$ and a map $\phi : \mathcal{X} \mapsto \mathcal{V}$ s.t.

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle_{\mathcal{V}} \qquad \forall \mathbf{x},\mathbf{y} \in \mathcal{X} \tag{12.3}$$

Corollary 12.1 Kernels and similarity: Kernels are defined in terms of inner product spaces and hence the have a notion of similarity between its arguments.

Example

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

Corollary 12.2 Kernels and distance: Let $\mathbf{k}(\mathbf{x},\mathbf{y})$ be a measure of similarity between $\mathbf{x}$ and $\mathbf{y}$ then $\mathbf{k}$ induces a dissimilarity/distance between $\mathbf{x}$ and $\mathbf{y}$ defined as the difference betweend the self-similarities $\mathbf{k}(\mathbf{x},\mathbf{x}) + \mathbf{k}(\mathbf{y},\mathbf{y})$ and the cross-similarities $\mathbf{k}(\mathbf{x},\mathbf{y})$:

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

Note

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

1. The Gram Matrix

Definition 12.5 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:

$$\mathcal{K} = \phi(\mathbf{X})\phi(\mathbf{X}^\intercal) = \left(\phi(\mathbf{x}_1),\dots,\phi(\mathbf{x}_n)\right)\left(\phi(\mathbf{x}_1),\dots,\phi(\mathbf{x}_n)\right)^\intercal$$
$$= \begin{pmatrix} \mathbf{k}(\mathbf{x}_1,\mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(\mathbf{x}_n,\mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_n,\mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \phi(\mathbf{x}_1)^\intercal\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_1)^\intercal\phi(\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_n)^\intercal\phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_n)^\intercal\phi(\mathbf{x}_n) \end{pmatrix}$$
$$\mathcal{K}_{ij} = \mathbf{k}(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^\intercal\phi(\mathbf{x}_j)$$

Corollary 12.3 $\mathbf{V}\Lambda\mathbf{V}^\intercal$
Kernel Eigenvector Decomposition:
For any symmetric matrix (Gram matrix $\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)|_{i,j=1}^n$) there exists an eigenvector decomposition:

$$\mathcal{K} = \mathbf{V}\Lambda\mathbf{V}^\intercal \tag{12.4}$$

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

Assuming all eigenvalues λ_t are non-negative, we can calculate the mapping:

$$\phi : \mathbf{x}_i \mapsto \left(\sqrt{\lambda_t}\mathbf{v}_{t,i}\right)_{t=1}^n \in \mathbb{R}^n, \qquad i = 1,\dots,n \tag{12.5}$$

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

$$\phi^\intercal(\mathbf{x}_i)\phi(\mathbf{x}_j) = \sum_{t=1}^n \lambda_t\mathbf{v}_{t,i}\mathbf{v}_{t,j} = \left(\mathbf{V}\Lambda\mathbf{V}^\intercal\right)_{i,j} = \mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) \tag{12.6}$$

1.1. Necessary Properties

Property 12.1 Inner Product Space:
$\mathbf{k}$ must be an *inner product* of a suitable space $\mathcal{V}$.

Property 12.2 Symmetry: $\mathbf{k}/\mathcal{K}$ must be symmetric:

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

Property 12.3 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 induces a *p.s.d. symmetric* kernel matrix $\mathbf{k}$ for any possible $S \subseteq \mathcal{X}$ see ?? 12.1.
$\iff$ all eigenvalues of the kernel gram matrix $\mathcal{K}$ for *finite* $\mathcal{V}$ must be non-negative ?? 27.2.

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 12.1 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.

$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \sum_{i=1}^\infty \lambda\phi(\mathbf{x})\phi(\mathbf{x}') \tag{12.7}$$

Theorem 12.2 General Mercers Theorem: Let Ω be a compact subset of R^n. Suppose $\mathbf{k}$ is a gernal continuous symmetric function such that the integral operator:

$$T_{\mathbf{k}} : L_2(\mathbf{X}) \mapsto L_2(\mathbf{X}) \qquad (T_{\mathbf{k}}f)(\cdot) = \int_\Omega \mathbf{k}(\cdot,\mathbf{x})f(\mathbf{x})\,d\mathbf{x} \tag{12.8}$$

is positve, that is it satisfies:

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

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

Note

All kernels satisfying mercers condtions describe an inner product in a high dimensional space.
$\Rightarrow$ can replace the inner product by the kernel function.

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

3. The Kernel Trick

Definition 12.6 Kernel Trick: If a kernel has an analytic form we do no longer need to calculate:
- the function mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$ and
- the inner product $\phi(\mathbf{x})^\intercal\phi(\mathbf{y})$

explicitly but simply us the formula for the kernel:

$$\phi(\mathbf{x})^\intercal\phi(\mathbf{x}) = \mathbf{k}(\mathbf{x},\mathbf{y}) \tag{12.9}$$

see examples 12.1 and 12.2

Note

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

4. Types of Kernels

4.1. Stationary Kernels

Definition 12.7 Stationary Kernel: A stationary kernel is a kernel that only considers vector differences:

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \mathbf{k}(\mathbf{x}-\mathbf{y}) \tag{12.10}$$

see example example 12.3

4.2. Isotropic Kernels

Definition 12.8 Isotropic Kernel: A isotropic kernel is a kernel that only considers distance differences:

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \mathbf{k}\left(\|\mathbf{x}-\mathbf{y}\|_2\right) \tag{12.11}$$

Corollary 12.4 :
Isotropic $\longrightarrow$ Stationary

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

5.1. The Linear Kernel

Definition 12.9 Linear/String Kernel:

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \mathbf{x}^\intercal\mathbf{y} \tag{12.12}$$

5.2. The Polynomial Kernel

Definition 12.10 Polynomial Kernel:
represents all monomials[def. 24.5] of degree up to m

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \left(1 + \mathbf{x}^\intercal\mathbf{y}\right)^m \tag{12.13}$$

5.3. The Sigmoid Kernel

Definition 12.11 Sigmoid/tanh Kernel:

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \tanh \kappa\mathbf{x}^\intercal\mathbf{y} - b \tag{12.14}$$

5.4. The Exponential Kernel

Definition 12.12 Exponential Kernel:
is an continuous kernel that is non-differential $\mathbf{k} \in \mathcal{C}^0$:

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_1}{\theta}\right) \tag{12.15}$$

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

5.5. The Gaussian Kernel

Definition 12.13 Gaussian/Squared Exp. Kernel/ Radial Basis Functions (RBF):
Is an inifite dimensional smooth kernel $\mathbf{k} \in \mathcal{C}^\infty$ with some usefull properties

$$\mathbf{k}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\theta^2}\right) \approx \begin{cases} 1 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ close} \\ 0 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ far away} \end{cases} \tag{12.16}$$

Explanation 12.1 (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:*

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

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

Note

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

5.6. The Matern Kernel

When looking at actual data/sample paths the smoothness of the Gaussian kernel[def. 12.13] is often a too strong assumption that does not model reality the same holds true for the non-smoothness of the exponential kernel[def. 12.12]. A solution to this dilemma is the Matern kernel.

Definition 12.14 Matern Kernel: is a kernel which allows you to specify the level of smoothness $\mathbf{k} \in \mathcal{C}^{\lfloor\nu\rfloor}$ by a positive parameter ν:

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

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

$$\tag{12.17}$$

$\mathcal{K}_\nu$ modified Bessel function of the second kind

6. Kernel Engineering

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

6.1. Closure Properties/Composite Rules

Suppose we have two kernels:

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

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

$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \mathbf{k}_1(\mathbf{x},\mathbf{x}') + \mathbf{k}_2(\mathbf{x},\mathbf{x}') \tag{12.18}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \mathbf{k}_1(\mathbf{x},\mathbf{x}') \cdot \mathbf{k}_2(\mathbf{x},\mathbf{x}') \tag{12.19}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \alpha\,\mathbf{k}_1(\mathbf{x},\mathbf{x}') \qquad \alpha \in \mathbb{R}_+ \tag{12.20}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = f(\mathbf{x})f(\mathbf{x}') \tag{12.21}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \mathbf{k}_3(\phi(\mathbf{x}),\phi(\mathbf{x}')) \tag{12.22}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = p\left(\mathbf{k}(\mathbf{x},\mathbf{x}')\right) \tag{12.23}$$
$$\mathbf{k}(\mathbf{x},\mathbf{x}') = \exp\left(\mathbf{k}(\mathbf{x},\mathbf{x}')\right) \tag{12.24}$$

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 12.1: Property 12.3**The kernel matrix is positive-semidefinite**:
Let $\phi : \mathcal{X} \mapsto \mathbb{R}^d$ and $\Phi = [\phi(\mathbf{x}_1) \;\; \dots \;\; \phi(\mathbf{x}_n)]^\intercal \in \mathbb{R}^{d\times n}$.
Thus: $\mathcal{K} = \Phi^\intercal\Phi \in \mathbb{R}^{n\times n}$.

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

Examples

Example 12.1 Calculating the Kernel by hand:

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

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

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

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

Operation Count:

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

Example 12.2
Calculating the Kernel using the Kernel Trick:

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

Operation Count:

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

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

Example 12.3 Stationary Kernels:

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

is a stationary but not an isotropic kernel.

Time Series

State Space Models

Definition 13.1 State Variables $\mathbf{x}$:
Is the smallest set of variables $\{x_1, \ldots, x_n\}$ that are fully capable of describing the state of our system which is usually *hidden* and not directly observable.

Definition 13.2 State Space $\mathcal{X}$:
Is the n-dimensional space spanned by the state variables??:
$$\mathbf{x} = [x_1 \quad x_n]^\intercal \in \mathcal{S} \subseteq \mathbb{R}^n \tag{13.1}$$

Definition 13.3 Input/Control Variables $\mathbf{u} \in \mathcal{A}$:
Are a variables $\mathbf{u}$ of the *transition model*[def. 13.5] that influence the propagation of to the state variables $\mathbf{x}$.

Definition 13.4 Output/Measurment 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 13.5 Transition Model f:
Describes the transition of the state $\mathbf{x}$ over time.

Definition 13.6 Measurment/Output/Observation Model h:
Describes the mapping of the state $\mathbf{x}$ onto the output $\mathbf{y}$.

Definition 13.7 (Discrete) State Space Model:
$$\mathbf{x}^{k+1} = f(t, \mathbf{x}^k, \mathbf{u}^k) \qquad t = 1, \ldots, K \tag{13.2}$$
$$\mathbf{y}^k = h(t, \mathbf{x}^k, \mathbf{u}^k) \tag{13.3}$$

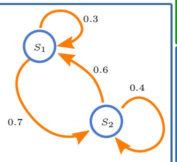
Markov Models

Definition 14.1 States $\mathcal{S} = \{s_1, \ldots, s_n\}$:
A state s_i encodes all information of the current configuration of a system.

Definition 14.2 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. 21.1]; and let $(S, \mathcal{S})$ be a measurable space[def. 33.7].
A $(S, \mathcal{S})$)-valued stochastic process $X = \{X_t : \Omega \to S\}_{t \in I}$ adapted to the filtration is said to possess the Markov property if:
$$\mathbb{P}(X_t \in A | \mathcal{F}_s) = \mathbb{P}(X_t \in A | X_s) \qquad \begin{array}{l}\forall A \in \mathcal{S} \\ s, t \in I \quad \text{s.t. } s < t\end{array} \tag{14.1}$$

1. Markov Chains

Definition 14.3 Markov Chain:
Is a sequence of random variables $\{X_i\}_{i \in \mathcal{T}}$[def. 37.3] that processes the markovian property[def. 14.2] i.e. each state X_t depend only on the previous state X_{t-1}:



$$\mathbb{P}(X_t = x | X_{t-1} = x_{t-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_t = x | X_{t-1} = x_{t-1})$$

Definition 14.4 Initial Distribution $\mathbf{q}_0$: Describes the initial distribution of states:
$$q_0(s_i) = \mathbb{P}(X_0 = s_i) \qquad \forall s_i \in S$$
$$\iff \quad \mathbf{q}_0 = [q_0(s_1) \quad \cdots \quad q_0(s_n)] \tag{14.2}$$

Definition 14.5 Transition Probability $p_{ji}(t)$:
is the probability of a random variable X_t in state s_i to transition into state s_j:
$$p_{ij}(t) = \mathbb{P}(X_{t+1} = s_j | X_t = s_i) \qquad \forall s_i, s_j \in S \tag{14.3}$$

Definition 14.6 n^{th} Transition Probability $p_{ji}^{(n)}(t)$:
denotes the probability of reaching state s_j from state s_i in n steps:
$$p_{ij}^{(n)}(t) = \mathbb{P}(X_{t+n} = s_j | X_t = s_i) \qquad \forall s_i, s_j \in S \tag{14.4}$$

Definition 14.7 Transition Matrix $\mathrm{P}(t)$:
The transition probabilities eq. (14.4) can be represented by a *row-stochastic matrix*?? $\mathrm{P}(t)$ where the i^{th} row represents the transition probabilities for the i^{th} state s_i i.e.

	To j	
From i	0.3	0.7
	0.4	0.6

Corollary 14.1 Row stochastic matrices and Graphs:
Row stochastic matrices?? represent graphs where the outgoing edges must sum to one:
$$\sum \delta^+(s_i) = 1 \tag{14.5}$$

1.1. Simulating Markov Chains

Corollary 14.2 proof 14.1
Realization of a Markov Chain:
$$\mathbb{P}(X_0 = x_0, \ldots, X_N = x_N) = q_0(x_1) \sum_{n=1}^{N} \mathrm{P}_{n-1,n}(t)$$

Algorithm 14.1 Forward Sampling:
Input: $\mathbf{q}(\mathbf{x}_0)$ and P
Output: $\mathbb{P}(X_{0:N})$
Sample $x_0 \sim \mathbb{P}(X_0)$
for $j = 1, \ldots, n$ **do**
$x_j \sim \mathbb{P}(X_j | X_{j-1} = x_{j-1})$
5: **end for**

1.2. State Distributions

Definition 14.8 Probability Distribution of the States $\mathbf{q}_{n+1}$:
$$\begin{aligned}
q_{n+1}(s_j) &= \mathbb{P}(X_{n+1} = s_j) \qquad \forall s_i \in S \\
&= \sum_{i=1}^{n} \mathbb{P}(X_n = s_i) \mathbb{P}(X_{n+1} = s_j | X_n = s_i) \\
&= \sum_{i=1}^{n} q_n(s_i) \mathrm{p}_{i,j}(t)
\end{aligned} \tag{14.6}$$
$$\begin{aligned}
\mathbf{q}_{n+1} &= [q_{n+1}(s_1) \quad \cdots \quad q_{n+1}(s_n)] \\
&= \mathbf{q}_n \mathrm{P}(t) \\
&= [q_n(s_1) \quad \cdots \quad q_n(s_n)] \begin{bmatrix} \mathrm{p}_{1,1} & \mathrm{p}_{1,2} & & \mathrm{p}_{1,n} \\ \mathrm{p}_{2,1} & \mathrm{p}_{2,2} & & \mathrm{p}_{2,n} \\ & & & \\ \mathrm{p}_{n,1} & \mathrm{p}_{n,2} & & \mathrm{p}_{n,n} \end{bmatrix} (t)
\end{aligned}$$

Corollary 14.3 [proof 14.2]
Time-homogeneous Markov Transition Probabilities:
$$\mathbf{q}_{n+1} = \mathbf{q}_0 \mathrm{P}^{n+1} \tag{14.7}$$

Definition 14.9 Stationary Distribution:
A markov chain has a stationary distribution if it satisfies:
$$\lim_{N\to\infty} q_N(s_i) = \lim_{N\to\infty} \mathbb{P}(X_N = s_i) = \pi_i \qquad \forall s_i \in S$$
$$\lim_{N\to\infty} \mathbf{q}_N = [\pi_1 \quad \cdots \quad \pi_n] \iff \mathbf{q} = \mathbf{q}\mathrm{P}(N) \tag{14.8}$$

Corollary 14.4 Existence of Stationary Distributions:
A Markov Chain has a stationary distribution if and only if at least one state is *positive recurrent*!

add matrix version with eigenvector

add recurrent and transient states

1.3. Properties of States

Definition 14.10 Absorbing State/Sink: Is a state s_i that once entered cannot be left anymore:
$$p_{ij}^{(n)}(t) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{14.9}$$

Definition 14.11 Accessible State $s_i \to s_j$:
A state s_j is accessible from state s_i iff:
$$\exists n : \quad p_{ij}^{(n)}(t) > 0 \tag{14.10}$$

Definition 14.12 Communicating States $s_i \leftrightarrow s_j$:
Two states s_j and s_i are communicating iff:
$$\exists n_1 : \quad p_{ij}^{(n_1)}(t) > 0 \quad \wedge \quad \exists n_2 : \quad p_{ji}^{(n_2)}(t) > 0 \tag{14.11}$$

Definition 14.13 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:
$$k = \gcd\{n > 0 : p_{ii}^{(n)} = \mathbb{P}(X_n = s_i \mid X_0 = s_i) > 0\} \tag{14.12}$$

Definition 14.14 Aperiodic State $k = 1$:
Is a state s_i with periodicity[def. 14.13] of one $\Leftrightarrow k = 1$

Corollary 14.5 : 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 14.6 Absorbing State: An absorbing state is an aperiodic state.

Explanation 14.1 (Defintion 14.14). *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 14.15 Time-homogeneous/Stationary Markov Chain:
are markov chains[def. 14.3] where the transition probability is independent of time:
$$p_{ji} = \mathbb{P}(X_t = s_j | X_{t-1} = s_i) = \mathbb{P}(X_{t-\tau} = s_j | X_{t-\tau} = s_i) \qquad \forall \tau \in \mathbb{N}_0 \tag{14.13}$$

Corollary 14.7 P
Transition Matrices of Stationary MCs:
Transition matrices of time-homogeneous markov chain are constant/time independent:
$$\mathrm{P}(t) = \mathrm{P} \tag{14.14}$$

Definition 14.16 Aperiodic Makrov Chain: Is a markov chain where all states are aperiodic:
$$\gcd\{n > 0 : p_{ii}^{(n)} = \mathbb{P}(X_n = s_i \mid X_0 = s_i) > 0\} = 1 \qquad \forall i \in \{1, \ldots, n\} \tag{14.15}$$

Definition 14.17 Irreducable Markov Chain: Is a Markov chain that has only *communicating states*[def. 14.12]:
$$s_j \leftrightarrow s_i \qquad \forall i, j \in \{1, \ldots n\} \tag{14.16}$$

- $\implies$ no sinks[def. 14.10]
- $\implies$ every state can be reached from every other state

Corollary 14.8 : An *irreducable*[def. 14.17] markov chain is automatically *apperiodic*[def. 14.16] if it has at least one aperiodic state[def. 14.14] $\iff$ *ergodic*[def. 14.18].

Corollary 14.9 : A markov chain is *not-irreducable* if there exist two states with different periods.

Definition 14.18 [example 14.1]
Ergodic Markov Chain: A finite markov chain is ergodic if there exist some number N s.t. any state s_j can be reached from any other state s_i in any number of steps less or equal to a N.
$\Rightarrow$ a markov chains is ergodic if it is:
1. *Irreducable*[def. 14.17]
2. *Aperiodic*[def. 14.16]

Corollary 14.10 Stationary Distribution: An erdodic markov chain has a *unique* stationary distribution[def. 14.9] and converges to it starting from any initial state $q_0(s_i)$

1.5. Types of Markov Chains

	Observable	Unobservable
Uncontrolled	MC[def. 14.3]	HMM[def. 15.1]
Controlled	MDP[def. 16.1]	POMDP[def. 17.1]

1.6. Markov Chain Monte Carlo (MCMC)

2. Proofs

Proof 14.1: [cor. 14.2]
$\mathbb{P}(X_0 = x_0, \ldots, X_N = x_N) = \mathbb{P}(X_0 = x_0) \cdot$
$\cdot \mathbb{P}(X_1 = x_1 | X_0 = x_0) \cdot \mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0) \cdot$
$\cdots \mathbb{P}(X_N = x_N | X_{N-1} = x_{N-1}, \ldots, X_0 = x_0)$
and then simply use the Markovian property

Proof 14.2: Corollary 14.3
$$\mathbf{q}_{n+1} = \mathrm{P}\mathbf{q}_n = (\mathbf{q}_{n-1}\mathrm{P})\mathrm{P} = \mathbf{q}_0 \mathrm{P}^{n+1}$$

3. Examples

Example 14.1 Ergodic Markov Chain:

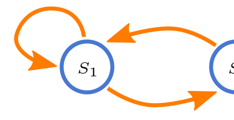
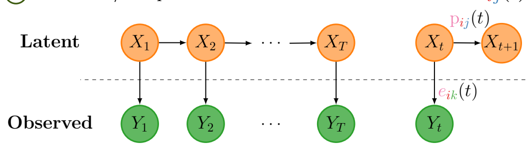


Figure 9: Ergodic for $N = 2$ (can reach s_2 at any $t \leqslant N$ after $N = 2$)

Hidden Markov Model (HMM)

Definition 15.1 $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathrm{P}, E)$
Hidden Markov Model (HMM):
Is a Markov Chain[def. 14.3] with hidden/latent states S_j that are only partially observable by noisy/indirect observations[def. 15.2]: It is characterized by the 5-tuple of:
1. States[def. 14.1] $\mathcal{S} = \{s_1, \ldots, s_n\}$
2. Actions[def. 16.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \ldots, a_m\}$
3. Observations[def. 15.2] $\mathcal{O}/\mathcal{O}_{s_j} = \{o_1, \ldots, o_m\}$
4. Transition Probabilities[def. 14.5] $\mathrm{p}(s_i, s_j)$
5. Emission/Output Probabilities[def. 15.3] $e_{ij}(t)$



Definition 15.2 Observations $\mathcal{O} = \{o_1, \ldots, o_l\}$:
Are indirect or noisy observations that are related to the true states s_j.

Definition 15.3 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:
$$e_{ij}(t) = \mathbb{P}(Y_t = o_j | X_t = s_i) \qquad \begin{cases} \forall o_i \in \mathcal{O} \\ \forall s_j \in \mathcal{S} \end{cases} \tag{15.1}$$

Markov Decision Processes (MDP)

Definition 16.1 $(\mathcal{S}, \mathcal{A}, \mathbb{P}_a, R_a)$
Markov Decision Process (MDP): A markov decision process is a *controlled* markov process/chain with an associated reward, where the transition can by steered by an actions. It is characterized by the 4-tuple of:
1. States[def. 14.1] $\mathcal{S} = \{s_1, \dots, s_n\}$
2. Actions[def. 16.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
3. Transition Probabilities[def. 16.3] $p_a(s_i, s_j)$
4. Rewards[def. 16.4] $r_a(s_i, s_j)$

Definition 16.2
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 16.3 Transition Probability $p_a(s_j, s_i)(t)$:
is the probability of a random variable X_t in state s_i to transition into state s_j and depends also on the current action a:
$$p_a(s_j, s_i) = p(s_j|s_i, a) = \mathbb{P}(x_{t+1} = s_j|x_t = s_i, a_t = a) \quad \forall s_i, s_j \in \mathcal{S}, \forall a \in \mathcal{A} \tag{16.1}$$

Definition 16.4 Reward $r_a(s_i, s_j)$:
is a function or probability distribution that measures the immediate reward and may depend on a any subset of (x_{t+1}, x_t, a):
$$(x_{t+1}, x_t, a) \mapsto R_{t+1} \in \mathcal{R} \subset \mathbb{R} \tag{16.2}$$

Markov decision processes require us to plan ahead. This is because the immediate reward[def. 16.4], that we obtain by greedily picking the best action may result in non-optimal local actions.

1. Policies and Values

Definition 16.5
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. 16.8] and can be deterministic or randomized:
$$\pi : \mathcal{S} \mapsto \mathcal{A} \qquad \text{or} \qquad \pi : \mathcal{S} \mapsto \mathbb{P}(\mathcal{A}) \tag{16.3}$$

Definition 16.6 Discounting Factor γ:
Is a factor $\gamma \in [0, 1)$ that signifies that future rewards are less valuable then current rewards.

Explanation 16.1 (Definition 16.6). *The reason for the discounting factor is that we may for example not even survive long enough to obtain future payoffs.*

Definition 16.7 Expected Discounted Value $J(\pi)$:
Is the *discounted* expected (reward) of the whole markov process:
$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t))\right] \tag{16.4}$$

Definition 16.8
Value Function $V^\pi(x)$:
Is the *discounted* expected reward[def. 16.4] of the whole markov process given an inital state $X_0 = x$:
$$V^\pi(x) = J(\pi|X_0 = x) \tag{16.5}$$
$$= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \middle| X_0 = x\right] \tag{16.6}$$
$$\tag{16.7}$$

1.1. Calculating the value of V^π

Definition 16.9 [proof 17.1]
Value Iteration:
$$V^\pi(x) = J(\pi|X_0 = x) \tag{16.8}$$
$$= \mathbb{E}_{x'|x,\pi(x)}\left[r(x, \pi(x)) + \gamma V^\pi(x')\right]$$
$$= r(x, \pi(x)) + \gamma \mathbb{E}_{x'|x,\pi(x)}\left[V^\pi(x')\right]$$
$$= \boxed{r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, \pi(x)\right) V^\pi\left(x'\right)}$$

We can now write this for all possible initial states as:
$$\mathbf{V}^\pi = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{V}^\pi \iff (\mathbf{I} - \gamma \mathbf{P}^\pi)\mathbf{V}^\pi = \mathbf{r}^\pi \tag{16.9}$$
with:
$$\mathbf{V}^\pi = \begin{bmatrix} V^\pi(s_1) \\ \cdots \\ V^\pi(s_n) \end{bmatrix} \qquad \mathbf{r}^\pi = \begin{bmatrix} r^\pi(s_1, \pi(s_1)) \\ \cdots \\ r^\pi(s_n, \pi(s_n)) \end{bmatrix}$$
$$\mathbf{P}^\pi = \begin{bmatrix} \mathbb{P}(s_1|s_1, \pi(s_1)) & \mathbb{P}(s_2|s_1, \pi(s_1)) & & \mathbb{P}(s_n|s_1, \pi(s_1)) \\ \mathbb{P}(s_1|s_2, \pi(s_2)) & \mathbb{P}(s_2|s_2, \pi(s_2)) & & \mathbb{P}(s_n|s_2, \pi(s_2)) \\ & & & \\ \mathbb{P}(s_1|s_n, \pi(s_n)) & \mathbb{P}(s_2|s_n, \pi(s_n)) & & \mathbb{P}(s_n|s_n, \pi(s_n)) \end{bmatrix}$$

1.1.1. Direct Mehtods

Corollary 16.1 LU-decomposition $\mathcal{O}(n^3)$:
The linear system from eq. (16.9):
$$(\mathbf{I} - \gamma \mathbf{P}^\pi)\mathbf{V}^\pi = \mathbf{r}^\pi \tag{16.10}$$
can be solved *directly* using Gaussian elimination in polynomial time $\mathcal{O}(n^3)$.

Note – invertebility
If $\gamma < 1$ then $(\mathbf{I} - \gamma \mathbf{P}^\pi)$ is full-rank/invertible as $\text{EVs}(\mathbf{P}^\pi) \leqslant 1$.

1.1.2. Fixed Point Iteration

Corollary 16.2 Fixed-Point Iteration $\mathcal{O}(n \cdot |\mathcal{S}|)$:
The linear system from eq. (16.9) can be solve using *fixed-point iteration*[def. 30.23] in at most $\mathcal{O}(n \cdot |\mathcal{S}|)$ (if every state s_i is connected to every other state $s_j \in \mathcal{S}$)

Algorithm 16.1 Fixed Point Iteration:
Input: Inital Guess: $V_0^\pi \overset{\text{i.e.}}{=} 0$
1: **for** $t = 1, \dots, T$ **do**
2: Use the fixed point method:
$$\mathbf{V}_t^\pi = \phi \mathbf{V}_t^\pi = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{V}_{t-1}^\pi \tag{16.11}$$
3: **end for**

Corollary 16.3
Policy Iteration Contraction [proof 17.2]:
Fixed point iteration of policy iteration is a contraction[def. 27.61] that leads to a fixed point V^π with a rate depending on the discount factor γ.
$$\|V_t^\pi - V^\pi\| = \|\phi V_{t-1}^\pi - \phi V^\pi\|$$
$$\leqslant \gamma \|V_{t-1}^\pi - V^\pi\| = \gamma^t \|V_0^\pi - V^\pi\| \tag{16.12}$$

Explanation 16.2.
- *$\gamma \downarrow$: 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 poin*

Note contraction
For a contraction:
- A unique fixed point exists
- We converge to the fixpoint

1.2. Choosing The Policy

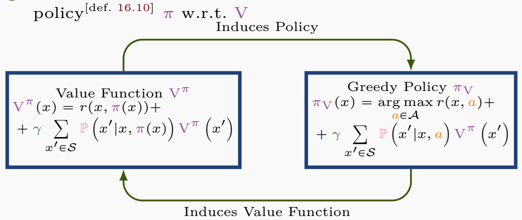
Question how should we choose the π? **Idea** compute $J(\pi)$ for *every possible* policy:
$$\pi^* = \arg\max J(\pi) \tag{16.13}$$
Problem this is unfortunately infeasible as there exist $m^n = |\mathcal{A}|^{|\mathcal{S}|}$ policies that we need to calculate the value for.

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

1.2.1. Greedy Policy

Definition 16.10 Greedy Policy:
Assuming we know $V^{\pi t-1}$ then we could choose a greedy policy:
$$a^* = \pi_t(x) \tag{16.14}$$
$$:= \arg\max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^{\pi t-1}\left(x'\right)$$

1. Given a policy π however we can calculate a value function V^π
2. Given a value function V we can induce a greedy policy[def. 16.10] π w.r.t. V



Theorem 16.1 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 16.11 Non-linear Bellman Equation: States that the optimal value is given by the action/policy that maximizes the value function eq. (16.8):
$$V^*(x) = \max_{a \in \mathcal{A}}\left[r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^*\left(x'\right)\right] \tag{16.15}$$
$$:= \max_{a \in \mathcal{A}} Q^*(x, a) \tag{16.16}$$

Note
This equation is non-linear due to the max in comparison to eq. (16.8).

1.2.2. Policy Iteration

Algorithm 16.2 Policy Iteration:
Initialize: Random Policy: π
1: **while** Not converged $t = t + 1$ **do**
2: Compute $V^{\pi t}(x)$
$$V^{\pi t}(x) = r(x, \pi(x)) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, \pi_t(x)\right) V^{\pi t}\left(x'\right)$$
3: Compute greedy policy π_G:
$$\pi_G(x) = \arg\max_{a \in \mathcal{A}} r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V^{\pi t}\left(x'\right)$$
4: Set $\pi_{t+1} \leftarrow \pi_G$
5: **end while**

Algorithm 16.2
Pros
- Monotonically improves $V^{\pi t} \geqslant V^{\pi t-1}$
- 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 16.12 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 16.3 Value Iteration [proof 17.3]:
Initialize: $V_0(x) = \max_{a \in \mathcal{A}} r(x, a)$
1: **for** $t = 1, \dots, \infty$ **do**
2: Compute:
3: $$Q_t(x, a) = r(x, a) + \gamma \sum_{x' \in \mathcal{S}} \mathbb{P}\left(x'|x, a\right) V_{t-1}\left(x'\right) \qquad \begin{matrix} \forall a \in \mathcal{A} \\ \forall x \in \mathcal{S} \end{matrix}$$
4: for all $x \in \mathcal{S}$ let:
$$V_t(x) = \max_{a \in \mathcal{A}} Q_t(x, a)$$
5: **if** $\max_{x \in \mathcal{S}} |V_t(x) - V_{t-1}(x)| \leqslant \epsilon$ **then**
6: break
7: **end if**
8: **end for**
9: Choose greedy policy π_{V_t} w.r.t. V_t

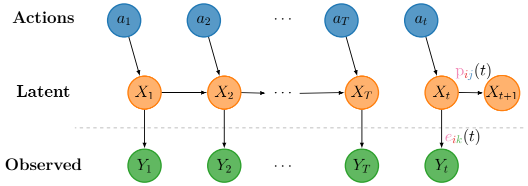
Corollary 16.4 [proof 17.4]
Value Iteration Contraction:
Algorithm 16.3 is guaranteed to converge to a ϵ optimal policy:
$$\left\|\left(V_t - V^*\right)\right\|_\infty \leqslant \gamma^t \left\|\left(V_0 - V^*\right)\right\|_\infty \tag{16.17}$$
$$\implies t \approx \ln \frac{\gamma}{\epsilon}\left\|V_0 - V^*\right\|_\infty \quad \text{for} \quad \left\|\left(V_t - V^*\right)\right\|_\infty \leqslant \epsilon$$

Algorithm 16.3
Pros
- Finds ϵ-optimal solution in polynomial *#iterrations* $\mathcal{O}(\ln \frac{1}{\epsilon})$[cor. 16.4].
- Complexity *per iteration* requires us to solve a linear system $\mathcal{O}(m \cdot n \cdot s) = \mathcal{O}(|\mathcal{A}| \cdot |\mathcal{S}| \cdot s)$ where s is the number of states we can reach.
  For small s and small m we are roughly linear w.r.t. the states $\mathcal{O}(n) = \mathcal{O}(|\mathcal{S}|)$

Cons
- Only ϵ-optimal solution.

Partially Observable MDP (POMDP)

Definition 17.1 $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{P}_a, E, R_a)$
Partially Observable Markov Decision Process:
A (POMDP) is a markov decision process[def. 16.1] with hidden markov states[def. 15.1]. It is characterized by the 6-tuple of:
1. States[def. 14.1] $\mathcal{S} = \{s_1, \dots, s_n\}$
2. Actions[def. 16.2] $\mathcal{A}/\mathcal{A}_{s_j} = \{a_1, \dots, a_m\}$
3. Observations[def. 15.2] $\mathcal{O}/\mathcal{O}_{s_j} = \{o_1, \dots, o_m\}$
4. Transition Probabilities[def. 16.3] $p_a(s_i, s_j)$
5. Emission/Output Probabilities[def. 15.3] $e_{ij}(t)$
6. Rewards[def. 16.4] $r_a(s_i, s_j)$



Explanation 17.1.
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. 16.1] by introducing a *belief state space* $\mathcal{B}$.

Definition 17.2 History H_t: Is a sequence of actions, observations and rewards:

$$H_t = \{\{a_0, o_0, r_0\}, \ldots, \{a_0, o_0, r_0\}\}$$

Definition 17.3 Belief State Space $\mathcal{B}$: Is a $|\mathcal{S}|-1$ dimensional simplex or ($|\mathcal{S}|$-dimensional probability vector[def. 27.69]) whose elements b are probabilities:

$$\mathcal{B} = \Delta(|\mathcal{S}|) = \left\{ b_t \in [0,1]^{|\mathcal{S}|} \;\middle|\; \sum_{x=1}^{n} b_t(x) = 1 \right\} \tag{17.1}$$

Definition 17.4 Belief State $b_t \in \mathcal{B}$: Is a probability distribution over the states $\mathcal{S}$ conditioned on the history H_t[def. 17.2].

1.1. Transition Model

Definition 17.5 [proof 17.5]
POMDP State/Posterior Update:

$$\begin{aligned} b_{t+1}(s_i) &= \mathbb{P}(X_{t+1} = s_i | Y_{t+1} = o_k) \\ &= \frac{1}{Z} \underline{\mathbb{P}(Y_{t+1} = o_k | X_{t+1} = s_i, a_t)} \\ &\cdot \sum_{s_j \in Pa(s_i)} b_t(s_j) \mathbb{P}(X_{t+1} = s_i | X_t = s_j, a_t) \end{aligned} \tag{17.2}$$

Definition 17.6 Stochastic Observation Model:

$$\underline{\mathbb{P}(Y_{t+1} = o_k | b_t, a_t)} = \sum_{s_i \in \mathcal{S}} b_t(s_i) \mathbb{P}(Y_{t+1} = o_k | X_t = s_i, a_t) \tag{17.3}$$

1.2. Reward Function

Definition 17.7 POMDP Reward Function:

$$r(b_t, a_t) = \sum_{s_j \in \mathcal{S}} b_t(s_i) r(s_i, a_t) \tag{17.4}$$

Note

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

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

2. Proofs

2.1. Markov Decision Processes

Proof 17.1: [def. 16.8]

$$\begin{aligned} V^\pi(x) &= \mathbb{E}_{X_{1:\infty}}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x\right] \\ &= \mathbb{E}_{\mathbf{X}}\left[\gamma^0 r(X_0, \pi(X_0)) + \sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x\right] \\ &\overset{\substack{\gamma^0=1 \\ X_0=x}}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}}\left[\sum_{t=1}^{\infty} \gamma^t r(X_t, \pi(X_t)) \mid X_0 = x\right] \\ &\overset{\text{re-index}}{=} r(x, \pi(x)) + \mathbb{E}_{\mathbf{X}}\left[\sum_{t=0}^{\infty} \underline{\gamma^{t+1}} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x\right] \\ &= r(x, \pi(x)) + \underline{\gamma}\mathbb{E}_{\mathbf{X}}\left[\sum_{t=0}^{\infty} \underline{\gamma^t} r(X_{t+1}, \pi(X_{t+1})) \mid X_0 = x\right] \\ &= \quad r(x, \pi(x)) \\ &+ \gamma\mathbb{E}_{X_1}\left[\mathbb{E}_{X_{2:\infty}}\left[\sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \middle| X_1 = x'\right] \middle| X_0 = x\right] \\ &\overset{\text{law } 33.7}{=} r(x, \pi(x))) \\ &+ \gamma \sum_{x' \in S} \mathbb{P}(x'|x, \pi(x)\mathbb{E}_{X_{2:\infty}}\left[\sum_{t=0}^{\infty} \gamma^t r(X_{t+1}, \pi(X_{t+1})) \middle| X_1 = x'\right] \\ &\overset{\text{eq. } (14.13)}{=} r(x, \pi(x))) \\ &+ \gamma \sum_{x' \in S} \mathbb{P}(x'|x, \pi(x)\underline{\mathbb{E}_{X_{2:\infty}}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, \pi(X_t)) \middle| X_0 = x'\right]} \\ &= \quad r(x, \pi(x)) + \gamma \sum_{x' \in S} \mathbb{P}\left(x'|x, \pi(x)\right) \underline{V^\pi\left(x'\right)} \end{aligned}$$

Proof 17.2 [cor. 16.3]: Consider $V, V' \in \mathbb{R}^n$ and let ϕ:

$$\phi x := r^\pi + \gamma P^\pi x \qquad \Longrightarrow \qquad \phi V^\pi = V^\pi$$

then it follows:

$$\begin{aligned} \left\|\phi V - \phi V'\right\| &= \left\|\cancel{r^\pi} + \gamma P^\pi V - \cancel{r^\pi} - \gamma P^\pi V'\right\| \\ &= \left\|\gamma P^\pi \left(V - V'\right)\right\| \\ &\overset{\text{eq. } (27.89)}{\leqslant} \gamma\left\|P^\pi\right\| \cdot \left\|\left(V - V'\right)\right\| \\ &\overset{\text{i.e. } L_2}{\leqslant} \gamma \cdot 1 \cdot \left\|\left(V - V'\right)\right\|_2 \end{aligned}$$

Proof 17.3: algorithm 16.3

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

Proof 17.4: [cor. 16.4] Let $\phi : \mathbb{R}^n \mapsto \mathbb{R}^n$, with:

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

Bellman's theorem 16.1 $\qquad \phi V^* = V^*$
and consider $V, V' \in \mathbb{R}^n$

$$\begin{aligned} \left\|\phi V - \phi V'\right\|_\infty &= \max_x \left|(\phi V)(x) - (\phi V')(x)\right| \\ &= \max_x \left|\max_a Q(x, a) - \max_{a'} Q'(x, a')\right| \\ &\overset{\text{Property } 24.8}{\leqslant} \max_x \max_a \left|Q(x, a) - Q'(x, a)\right| \end{aligned}$$

$$= \max_{x,a} |\cancel{r} + \gamma \sum_{x'} \mathbb{P}\left(x'|x, a\right) V(x') - \cancel{r} - \gamma \sum_{x'} \mathbb{P}\left(x'|x, a\right) V'(x')|$$

$$\begin{aligned} &= \gamma \max_{x,a} \left|\sum_{x'} \mathbb{P}\left(x'|x, a\right)\left(V(x') - V'(x')\right)\right| \\ &\overset{\text{eq. } (27.89)}{\leqslant} \gamma \max_{x,a} \overbrace{\left|\sum_{x'} \mathbb{P}\left(x'|x, a\right)\right|}^{\leqslant 1} \cdot \left|\left(V(x') - V'(x')\right)\right| \\ &\leqslant \gamma \cdot 1 \cdot \left\|\left(V(x') - V'(x')\right)\right\|_\infty \end{aligned}$$

Note

For the policy iteration the calculation was easier as the rewards canceled, however here we have the max.

2.2. MDPs

Proof 17.5: Defintion 17.5 Directly by definition 8.5 and its corresponding proof 11.4 with additional action a_t:

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

Reinforcement Learning

Now we are working with an *unknown* MDP[def. 16.1] meaning that:
1. we do no longer know the transition model[def. 16.3]
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 18.1 Agent:
Is the *learner/decision maker* of our *unknown* MDP.

Definition 18.2 Environment: Is the representation of the world in which our agents acts.

Definition 18.3 On-Policy Learning: At any given time the agent has full control which actions to pick.

Definition 18.4 Off-Policy Learning: The agent has to fix a policy in advance based on behavioral observations.

Definition 18.5 Trajectory τ**:**
Is a set of consecutive 3-tuples of states, actions and rewards:

$$\tau = \{s_t, a_t, r_t\} \qquad t = 1, \dots, \tau \tag{18.1}$$

Definition 18.6 Episodic Learning: Is a setting where we generate multiple K-episodes of different trajectories $\left\{\tau^{(k)}\right\}_{i=1}^{K}$ from which the agent can learn.

Explanation 18.1. *For each episode the agent starts in a random state and follows a policy.*

1. Model Based Reinforcement Learning

Proposition 18.1 Model Based RL:
Try to learn the MDP [def. 16.1] by:
1. Estimating
   - the transition probabilities[def. 16.3] $\quad p_a(s_i, s_j)$
   - the reward function[def. 16.4] $\quad r(b_t, a_t)$
2. Optimizing the policy of the estimated MDP

1.1. Estimating Transitions and Rewards

Formula 18.1 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. 34.23]:

$$N_{s_i|s_j,a} := \sum_{k=1}^{t} \delta\left(X_{k+1}{=}s_i | X_k{=}s_j, A_k{=}a\right) \tag{18.2}$$

$$N_{s_j,a} := \sum_{k=1}^{t} \delta\left(X_k{=}s_j, A_k{=}a\right) \tag{18.3}$$

$$p_a(s_i, s_j) \approx \frac{N_{s_i|s_j,a}}{N_{s_j,a}} \tag{18.4}$$

$$r(s_i, a) \approx \frac{1}{N_{s_i,a}} \sum_{k=1}^{t} \delta\left(X_k{=}s_i, A_k{=}a\right) r(X_k, A_k) \tag{18.5}$$

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 18.1 Epsilon Greedy Learning:

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

Corollary 18.1 Necessary Condition for Convergence:
If the sequence ϵ_t satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \epsilon_t < \infty, \qquad \sum_t \epsilon_t^2 < \infty \qquad (\text{i.e. } \epsilon_t = 1/t) \tag{18.6}$$

then algorithm 18.1 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 18.2 [Brafman & Tennenholz '02]
R-max Algorithm:

Initialize every state with:
$$\hat{r}(s_t, a) = R_{\max} \qquad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1 \tag{18.7}$$
Set min. number Δ of observations for policy update
Compute Policy π_1 of the MDP[def. 16.1] using $(\hat{p}, \hat{r})$:
π_t

for $k = 1, \dots, K$ **do**
  Choose $a = \pi_t(x_t)$ and observe (s, r)
  Calculate:
$$N_{\mathbf{x}_t,a}{+}{=}1 \qquad r(x_t, a){+}{=}r(x_t, a) \tag{18.8}$$
$$N_{\mathbf{x}_{t+1}|\mathbf{x}_t,a}{+}{=}1 \tag{18.9}$$
  if $k{=}{=}\Delta$ **then**
    Re-calculate (based on eqs. (18.4) and (18.5)):
    $\hat{r}(s_t, a) = R_{\max} \qquad \hat{p}_a(X_{t+1}|X_t = s_i, a) = 1$
    and update the policy $\pi_t = \pi_t(\hat{p}, \hat{r})$
  end if
end for

Note
Other ways of updating the policy at certains times exist.

Problems

Cons
- Memory: for all $a \in \mathcal{A}$, $\mathbf{x}_{t+1}, \mathbf{x}_t \in \mathcal{X}$ we need to store $\hat{p}_a(x_{t+1}|x_t, a)$ and $\hat{r}(s_t, a)$ which results in $|\mathcal{S}|^2|\mathcal{A}|$ (for dense MDP).
- Computation Time: We need to calculate the 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 18.2 [proof 18.1]
Number of Samples to bound Reward:

$$\mathbb{P}\left(\hat{r}(s, a) - r(s, a) \leqslant \epsilon\right) \geqslant 1 - \delta \iff n \in \mathcal{O}\left(\frac{R_{\max}{}^2}{\epsilon^2} \log \frac{1}{\delta}\right) \tag{18.10}$$

Theorem 18.1 : Every T timesteps, with high probability, $R_{\max}$ either:
- Obtain near optimal reward, or
- Visits at leas one unkown state-action pair

Theorem 18.2 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 18.3 Model Free RL:
Tries to estimate the value function[def. 16.8] directly in order to act greedily upon it.
- Policy Gradient Methods
- Actor Critic Methods

2.1. Temporal Difference Learning (TD)

Assume we fix a random intial policy 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 usingeq. (16.8):

$$\hat{V}^{\pi}(x_t) = \mathbb{E}_{X_{t+1},R}\left[R + \gamma \hat{V}^{\pi}(X') | X, a\right] \tag{18.11}$$

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. 35.6] with a single sample to approximate the expectation ineq. (18.11):

$$\hat{V}_{t+1}^{\pi}(s_i) = r + \gamma \hat{V}_t^{\pi}(s_j)$$

Problem: high variance of estimates $\Rightarrow$ average with previous estimate.

Definition 18.7 Temporal Difference (TD) Learning:

$$\hat{V}(x_{t+1}) = (1 - \alpha_t)\hat{V}(x_t) + \alpha_t\left(r + \gamma \hat{V}(x_{t+1})\right) \tag{18.12}$$

Corollary 18.2 Necessary Condition for Convergence:
If the learning rate α_t satisfies the *Robbins Monro* (RM) conditions

$$\sum_t \alpha_t < \infty, \qquad \sum_t \alpha_t^2 < \infty \qquad (\text{i.e. } \alpha_t = 1/t) \tag{18.13}$$

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

$$\mathbb{P}\left(\hat{V} \to \hat{V}^{\pi}\right) = 1 \tag{18.14}$$

2.2. Q-Learning

Definition 18.8 Action Value/Q-Function:

$$Q \tag{18.15}$$

2.2.1. Policy Gradients

2.2.2. Actor-Critic Methods

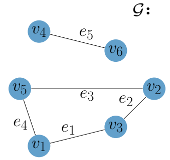
3. Proofs

Proof 18.1: proposition 18.2 using hoeffdings bound[def. 33.38] with δ and $b - a = R_{\max}$.

Graph Theory

Definition 19.1 Graph $\mathcal{G}$:

A graph $\mathcal{G}$ is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a finite set of vertices $\mathcal{V}$[def. 19.4] and a multi set[def. 20.3] of edges $\mathcal{E}$[def. 19.10].



Definition 19.2 Order $n = |\mathcal{V}|$:
The order of a graph is the cardinality of its vertix set.

Definition 19.3 Size $m = |\mathcal{E}|$:
The size of a graph is the number of its edges.

Corollary 19.1 n-Graph: Is a graph $\mathcal{G}$[def. 19.1] of order n.

Corollary 19.2 (p, q)-Graph: Is a graph $\mathcal{G}$[def. 19.1] of order p and size q.

1. Vertices

Definition 19.4 Vertices/Nodes $\mathcal{V}$:
Is a set of entities of a graph connected and related by edges in some way:

Definition 19.5 Neighborhood $N(v)$**:** The neighborhood of a vertix $v_i \in \mathcal{V}$ is the set of all adjacent vertices:

$$N(v_i) = \{v_k \in \mathcal{V} : \exists e_k = \{v_i, v_j\} \in \mathcal{E}, \forall v_j \in \mathcal{E}\} \qquad (19.1)$$

1.0.1. Adjacency Matrix

Definition 19.6 (unweighted) Adjacency Matrix A:
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its *adjacency matrix* is a square matrix $\mathbf{A} \in \mathbb{N}^{n,n}$ defined as:

$$\mathbf{A}_{i,j} := \begin{cases} 1 & \text{if } \exists e(i,j) \\ 0 & \text{otherwise} \end{cases} \qquad (19.2)$$

Definition 19.7 weighted Adjacency Matrix A:
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its *weighted adjacency matrix* is a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ defined as:

$$\mathbf{A}_{i,j} := \begin{cases} \theta_{ij} & \text{if } \exists e(i,j) \\ 0 & \text{otherwise} \end{cases} \qquad (19.3)$$

Diagonal Elements

For a graph without self-loops the diagonal elements of the adjacency are all zero.

1.0.2. Degree Matrix

Definition 19.8 Degree of a Vertix δ:
The degree of a vertix v is the cardinality of the neighborhood[def. 19.5] – the number of adjacent vertices:

$$\deg(v_i) = \delta(v) = |N(v)| = \sum_{j=1}^{j<i} \mathbf{A}_{ij} \qquad (19.4)$$

Definition 19.9 Degree Matrix D:
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its degree matrix is a diagonal matrix $\mathbf{D} \in \mathbb{N}^{n,n}$ defined as:

$$\mathbf{D}_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \qquad (19.5)$$

2. Edges

Definition 19.10 Edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$:
Represent some relation between edges[def. 19.4] and are represented by two-element subset sets of the vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ and } v_j \text{ connected} \qquad (19.6)$$

Proposition 19.1 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 19.11 Subgraph $\mathcal{H} \subseteq \mathcal{G}$:
A graph $\mathcal{H} = (U, F)$ is a *subgraph* of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ iff:

$$U \subseteq \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \qquad (19.7)$$

4. Components

Definition 19.12 Component: A connected component of a graph $\mathcal{G}$ is a *connected*[def. 19.20] subgraph[def. 19.11] 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 19.13 Walk: A walk of a graph $\mathcal{G}$ as a sequence of vertices with corresponding edges:

$$W = \{v_k, v_{k+1}\}_k^K \in \mathcal{E} \qquad (19.8)$$

Definition 19.14 Length of a Walk K**:** Is the number of edges of that Walk.

Definition 19.15 Path P**:** Is a walk of a graph $\mathcal{G}$ where all visited vertics are distinct (no-repetitions).

Attention: Some use the terms walk for paths and simple paths for paths.

Definition 19.16 Cycle: Is a path[def. 19.15] of a graph $\mathcal{G}$ where the last visited vertix is the one from which we started.

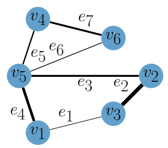
6. Different Kinds of Graphs

7. Weighted Graph

Definition 19.17 Weighted Graph:

Is a graph $\mathcal{G}$ where edges are associated with a weight:

$$\exists \theta_i := \text{weight}(e_i) \qquad \forall e_i \mathcal{E}$$

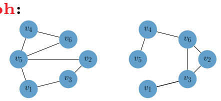


8. Spanning Graphs

Definition 19.18 Spanning Graph:

Is a subgraph[def. 19.11] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for which it holds:

$$U = \mathcal{V} \quad \text{and} \quad F \subseteq \mathcal{E} \qquad (19.9)$$



8.1. Minimum Spanning Graph

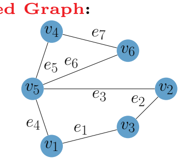
Definition 19.19 Minimum Spanning Graph: Is a spanning graph[def. 19.18] $\mathcal{H} = (U, F)$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with minimal weights/distance of the edges.

9. Connected Graphs

Definition 19.20 (Weakly) Connected Graph:

Is a graph $\mathcal{G}$[def. 19.1] where there exists a path between any two vertices:

$$\exists P(v_i, \dots, v_j) \qquad \forall v_i, v_j \in \mathcal{V} \qquad (19.10)$$



Corollary 19.3 Strongly Connected Graph: A directed Graph[def. 19.22] is called strongly connected if every nodes is *reachable* from every other node.

Corollary 19.4 Components of Connected Graphs: A connected Graph[def. 19.20] consist of one component $c(\mathcal{G}) = 1$.

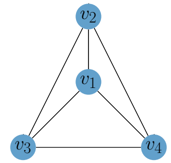
9.1. Fully Connected/Complete

Definition 19.21 Fully Connected/Complete Graph:

Is a connected graph $\mathcal{G}$[def. 19.20] where each node is connected to every other node.

$$\exists e \forall \{v_i, v_j\} \qquad \forall v_i, v_j \in \mathcal{V} \qquad (19.11)$$

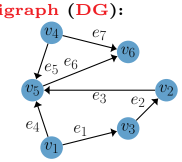
$$|\mathcal{V}| = \frac{1}{2}|\mathcal{V}|(|\mathcal{V}| - 1) \qquad (19.12)$$



9.2. Directed Graphs

Definition 19.22 Directed Graph/Digraph (DG):

A directed graph $\mathcal{G}$ is a graph where edges are direct arcs[def. 19.23].



Definition 19.23 Directed Edges/Arcs: Represent some *directional* relationship between edges[def. 19.4] and are represented by *ordered* two-element subset sets of vertices:

$$e_k = \{v_i, v_j\} \in \mathcal{E} \iff v_i \text{ goes to } v_j \qquad (19.13)$$

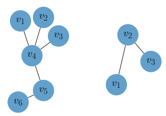
9.3. Trees And Forests

9.3.1. Acyclic Graphs

Definition 19.24 Acyclic Graphs: Are graphs[def. 19.1] where no cycles[def. 19.16] exist.

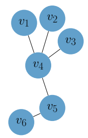
Definition 19.25 Forests:

Are acyclic graphs[def. 19.24].



Definition 19.26 Trees:

Are acyclic graphs[def. 19.24] that are connected[def. 19.20].



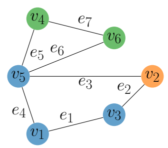
10. Graph Layering

Definition 19.27 Graph Layering:

Given a graph $\mathcal{G}$ a layering of the graph is a partition of its node set $\mathcal{V}$[def. 19.4] into subsets

$$\{\mathcal{V}_1, \dots, \mathcal{V}_L\} \subseteq \mathcal{V}$$

$$\text{s.t.} \quad \mathcal{V} = \mathcal{V}_1 \cup \dots \cup \mathcal{V}_L \qquad (19.14)$$



11. Bisection Algorithms

11.1. Local Approaches

11.2. Global Approaches

11.2.1. Spectral Decomposition

Definition 19.28 Graph Laplacian (Matrix) $\mathbf{L}(\mathcal{G})$:
Given a graph with n vertices and m edges has a graph laplacian matrix defined as:

$$\mathbf{L} = \mathbf{A} - \mathbf{D} \qquad l_{ij} := \begin{cases} -1 & \text{if } i \neq j \text{ and } e_{ij} \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } e_{ij} \notin \mathcal{E} \\ \deg(v_i) & \text{if } i = j \end{cases} \qquad (19.15)$$

Corollary 19.5 title:

11.2.2. Inertial Bisection

Set Theory

Definition 20.1 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 20.2 Empty Set $\{\}/\emptyset$**:** is the unique set having no elements/cardinality[def. 20.5] zero.

Definition 20.3 Multiset/Bag: Is a set-like object in which multiplicity[def. 20.4] matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 20.4 Multiplicity: The multiplicity n_a of a member a of a multiset[def. 20.3] $\mathcal{S}$ is the number of times it appears in that set.

Definition 20.5 Cardinality $|S|$**:** Is the number of elements that are contained in a set.

Definition 20.6 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 20.7 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 20.8 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:
  $$\|x - y\|_2 < \epsilon(x) \tag{20.1}$$
- **Metric Spaces**[def. 27.63]: a Subset U of a metric space (M, d) is open if:
  $$\exists \epsilon > 0: \quad \text{if} \quad d(x, y) < \epsilon \quad \forall y \in M, \forall x \in U \quad \Longrightarrow\ y \in U \tag{20.2}$$
- **Topological Spaces**[def. 29.2]: Let (X, τ) be a topological space. A set A is said to be open if it is contained in τ.

add image

Definition 20.9 Closed Set: Is the complement of an open set[def. 20.8].

Definition 20.10 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 20.11 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:
$$[a, b] = \{x \in \mathbb{R} \,|\, a \leqslant x \leqslant b\} \tag{20.3}$$

Definition 20.12 Open Interval (a, b)**:** The open interval of a and b is the set of all real numbers that are within a and b:
$$(a, b) = \{x \in \mathbb{R} \,|\, a < x \leqslant <\} \tag{20.4}$$

3.2. The Rational Numbers $\mathbb{Q}$

Example 20.1 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. Submoduluar Set Functions

Definition 20.13 Submodular Set Functions: A submodular function $f : 2^\Omega \mapsto \mathbb{R}$ is a function that satisifies:
$$f(A \cup \{x\}) - f(A) \geqslant f(B \cup \{X\}) - F(B) \qquad \begin{array}{l} \forall A \subseteq B \subset \Omega \\ \{x\} \in \Omega \backslash B \end{array} \tag{20.5}$$

Explanation 20.1 (Definition 20.13)**.** *Addaing an element x to the the smaller subset A yields at least as much information/value gain as adding it to the larger subset B.*

Definition 20.14 Montone Submodular Function: A *monotone* submodular function is a submodular function[def. 20.13] that satisfies:
$$f(A) \leqslant f(B) \qquad \forall A \subseteq B \subseteq \Omega \tag{20.6}$$

Explanation 20.2 (Definition 20.14)**.** *Adding more elements to a set will always increase the information/value gain.*

4.2. Complex Numbers

Definition 20.15 Complex Conjugate $\bar{z}$**:** The complex conjugate of a complex number $z = x + iy$ is defined as:
$$\bar{z} = x - iy \tag{20.7}$$

add kind of wikipedia image

Corollary 20.1 Complex Conjugate Of a Real Number: The complex conjugate of a real number $x \in \mathbb{R}$ is x:
$$\bar{x} = x \qquad \Longrightarrow \qquad x \in \mathbb{R} \tag{20.8}$$

Formula 20.1 Euler's Formula:
$$e^{\pm ix} = \cos x \pm i \sin x \tag{20.9}$$

Formula 20.2 Euler's Identity:
$$e^{\pm i} = -1 \tag{20.10}$$

Note
$$e^{n} = 1 \Leftrightarrow n = \mathrm{i}\, 2\pi k, \qquad k \in \mathbb{N} \tag{20.11}$$

Sequences&Series

Definition 21.1 Index Set: Is a set[def. 20.1] A, whose members are labels to another set S. In other words its members index member of another set. An index set is build by enumerating the members of S using a function f s.t.
$$f : A \mapsto S \qquad A \in \mathbb{N} \tag{21.1}$$

Definition 21.2 Sequence $(a_n)_{n \in A}$**:** A sequence is an by an *index set A enumerated* multiset[def. 20.3] (repetitions are allowed) of objects in which *order does matter*.

Definition 21.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 21.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \ldots)$.
$$t_n = t_0 + nd \qquad d \text{ :difference between two terms} \tag{21.2}$$

1.2. Geometric Sequence

Definition 21.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \ldots)$.
$$t_n = t_0 \cdot r^n \qquad r \text{ :ratio between two terms} \tag{21.3}$$

Property 21.1 Sum of Geometric Sequence:
$$\sum_{k=1}^{n} ar^{k-1} = \frac{a(1 - r^n)}{1 - r} \tag{21.4}$$

2. Converging Sequences

2.1. Pointwise Convergence

Definition 21.6 $\lim_{n\to\infty} f_n = f$ **pointwise Pointwise Convergence[?]:** Let (f_n) be a sequence of functions with the same domain[def. 24.11] and codomain[def. 24.12]. The sequence is said to convergence pointwise to its *pointwise limit function* f if it satisfies:
$$|\lim_{n\to\infty} f_n(x) - f(x)| = 0 \qquad \forall x \in \mathtt{dom}(f_i) \tag{21.5}$$

2.2. Uniform Convergence

Definition 21.7 $\lim_{n\to\infty} f_n = f$ **uniform/**$f_n \overset{\infty}{\rightrightarrows} f$ **Uniform Convergence[?]:** Let (g_n) be a sequence of functions with the same domain[def. 24.11] and codomain[def. 24.12]. The sequence is said to convergence uniformly to its *pointwise limit function* f if it satisfies:
$$\exists \epsilon > 0 : \exists n \geqslant 1 \quad \sup_{x \in \mathtt{dom}(f_i)} |g_n(x) - f(x)| < \epsilon \qquad \forall x \in \mathtt{dom}(f_i) \tag{21.6}$$

Note
Uniform convergence is characterized by the uniform norm??, and is stronger than pointwise convergence.

Toplogy

Definition 22.1 Topological Space[?] (X, τ)**:** Is an ordered pair (X, τ), where X is a set and τ is a topology[def. 29.1] on X.

Definition 22.2 Topological Space[?] (X, τ)**:** Is an ordered pair (X, τ), where X is a set and τ is a topology[def. 29.1] on X.

1. Weak Topologies

Definition 22.3 Weak Topology $\mathcal{C}(\mathcal{K}; \mathbb{R})$**:** Is the coresests topology s.t all cont. linear functionals w.r.t. to the strong topology are continuous.
Neighbourhood Basis:
$$\{f \,|\, |l_1| < \epsilon_1, \ldots, |l_n| < \epsilon_n, \forall \epsilon_i, \forall n, \forall \text{lin. functions} f\} \tag{22.1}$$

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 22.1 Euclidean Space: In the euclidean case, a set $X \in \mathbb{R}$ is compact iff:

- it is closed[def. 20.9]
- bounded

3. Closure

Definition 22.4 Closure of a Set[?] $\mathrm{cl}_{X,\tau}(S)/\bar{S}$**:** The closure of a subset S of a toplogical space[def. 29.2] (X, τ) is defined equivilantly 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 22.2 Uniform Closure $\bar{\cdot}^{\|\cdot\|_\infty}$**:** 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. 21.7] functions

Corollary 22.3 Weak Closure:

Logic

1. Boolean Algebra

1.1. Basic Operations

Definition 23.1 Conjunction/AND $\wedge$:

Definition 23.2 Disjunction/OR $\vee$:

Definition 23.3 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.

$$x \wedge y = xy = \min(x, y)$$
$$x \vee y = x + y = \max(x, y)$$
$$\neg x = 1 - x$$
$$x \oplus y = (x + y) \cdot (\neg x + \neg y) = x \cdot \neg y + \neg x \cdot y$$

Note: non-linearity of XOR

$$(x + y) \cdot (\neg x + \neg y) = -x^2 - y^2 - 2xy + 2x + 2y$$

1.2. Boolean Identities

Property 23.1 Idempotence:

$$x \wedge x \equiv x \quad \text{and} \quad x \vee x \equiv x \tag{23.1}$$

Property 23.2 Identity Laws:

$$x \wedge \text{true} \equiv x \quad \text{and} \quad x \vee \text{false} \equiv x \tag{23.2}$$

Property 23.3 Zero Law's:

$$x \wedge \text{false} \equiv \text{false} \quad \text{and} \quad x \vee \text{true} \equiv \text{true} \tag{23.3}$$

Property 23.4 Double Negation:

$$\neg\neg x \equiv x \tag{23.4}$$

Property 23.5 Complementation:

$$x \wedge \neg x \equiv \text{false} \quad \text{and} \quad x \vee \neg x \equiv \text{true} \tag{23.5}$$

Property 23.6 Commutativity:

$$x \vee y \equiv y \vee x \quad \text{and} \quad x \wedge y \equiv y \wedge x \tag{23.6}$$

Property 23.7 Associativity:

$$(x \vee y) \vee z \equiv x \vee (y \vee z) \tag{23.7}$$
$$(x \wedge y) \vee z \equiv x \vee (y \wedge z) \tag{23.8}$$

Property 23.8 Distributivity:

$$x \vee (y \wedge z) \equiv (x \vee y) \wedge (x \vee z) \tag{23.9}$$
$$x \wedge (y \vee z) \equiv (x \wedge y) \vee (x \wedge z) \tag{23.10}$$

Property 23.9 De Morgan's Laws:

$$\neg(x \vee z) \equiv (\neg x \wedge \neg y) \tag{23.11}$$
$$\neg(x \wedge z) \equiv (\neg x \vee \neg y) \tag{23.12}$$

Note

The algebra axioms come in pairs that can be obtained by interchanging $\wedge$ and $\vee$.

1.3. Normal Forms

Definition 23.4 Literal [example 23.1]:
Literals are atomic formulas or their negations

Definition 23.5 Negation Normal Form (NNF): A formula F is in negation normal form is the negation operator $\neg$ is only applied to literals[def. 23.4] and the only other operators are $\wedge$ and $\vee$.

Definition 23.6 Conjunctive Normal Form (CNF): An boolean algebraic expression F is in CNF if it is a *conjunction* of *clauses*, where each clause is a disjunction of *literals*[def. 23.4] $L_{i,j}$:

$$F_{\text{CNF}} = \bigwedge_{i=1}^{n} \left(\bigvee_{j=1}^{m_i} L_{i,j} \right) \tag{23.13}$$

Definition 23.7 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. 23.4] $L_{i,j}$:

$$F_{\text{DNF}} = \bigvee_{i=1}^{n} \left(\bigwedge_{j=1}^{m_i} L_{i,j} \right) \tag{23.14}$$

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 23.1:

1. Using *De Morgan's laws*Property 23.9 and double negationProperty 23.4 transform F into *Negation Normal Form*[def. 23.5]:

$\neg\neg x$	by	x
$\neg(x \wedge y)$	by	$(\neg x \vee \neg y)$
$\neg(x \vee y)$	by	$(\neg x \wedge \neg y)$
$\neg$true	by	false
$\neg$false	by	true

2. Using distributive lawsProperty 23.8 substitute all:

$x \wedge (y \vee z)$	by	$(x \wedge y) \vee (x \wedge z)$
$(y \vee z) \wedge x$	by	$(y \wedge x) \vee (z \wedge x)$
$x \wedge$ true	by	true
true $\wedge x$	by	true

3. Using the identityProperty 23.2 and zero laws Property 23.3 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 23.2]

To obtain a DNF formula from a truth table we need to have a *conjunctive*[def. 23.3] for each row where F is true.

2. Examples

Example 23.1 Literals:
Boolean literals: $x, \neg y, s$
Not boolean literals: $\neg\neg x$, $(x \wedge y)$

Example 23.2 DNF from truth tables:

Need a conjunction of:

- $(\neg x \wedge \neg y \wedge \neg z)$
- $(\neg x \wedge y \wedge z)$
- $(x \wedge \neg y \wedge \neg z)$
- $(x \wedge y \wedge z)$

x	y	z	F
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

$$(\neg x \wedge \neg y \wedge \neg z) \wedge (\neg x \wedge y \wedge z) \wedge (x \wedge \neg y \wedge \neg z) \wedge (x \wedge y \wedge z)$$

Calculus and Analysis

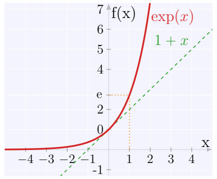
1. Functional Analysis

1.1. Elementary Functions

1.1.1. Exponential Numbers

Definition 24.1 Exponential Function $\exp : \mathbb{K} \mapsto \mathbb{K}$:

$$\exp(x) = e^x = \sum_{n=0}^{\infty} \frac{x}{n!} = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n \qquad (24.1)$$



Definition 24.2 Exponential/Euler Number e:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = 2.7182 \qquad (24.2)$$

Properties Defining the Expeontial Function

Property 24.1:

$$\exp(x+y) = \exp(x) + \exp(y) \qquad (24.3)$$

Property 24.2:

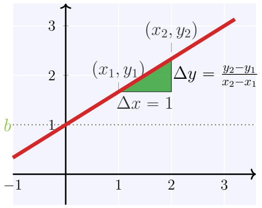
$$\exp(x) \leqslant 1 + x \qquad (24.4)$$

1.1.2. Affine Linear Functions

Definition 24.3 Affine Linear Function $f(x) = ax + b$: An affine linear function are functions that can be defined by a scaling $s_a(x) = ax$ plus a translation $t_b(x) = x + b$:

$$M = \{f : \mathbb{R} \mapsto \mathbb{R} | f(x) = (s_a \circ t_b)(x) = ax + b, \quad a, b \in \mathbb{R}\} \qquad (24.5)$$

$$f(x) = ax + b$$
$$f(0) = b$$
$$f'(x) = a$$



Formula 24.1 [proof 24.1]
Linear Function from Point and slope $f(x_0) = y_1$:
Given a point (x_1, y_1) and a slope a we can derive:

$$f(x) = a \cdot (x - x_0) + y_0 = ax + (y_1 - ax_0) \qquad (24.6)$$

Formula 24.2 Linear Function from two Points:

$$f(x) = a \cdot (x - x_p) + y_p = ax + (y_p - ax_p) \qquad (24.7)$$

$$a = \frac{y_1 - y_0}{x_1 - x_0} \qquad p = \{ \text{ or } 2\}1$$

1.1.3. Polynomials

Definition 24.4 Polynomial: A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:

$$\mathcal{P}_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + a_n x^n \qquad (24.8)$$

Corollary 24.1 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 24.5 Monomial: Is a polynomial with only one term.

Cubic Polynomials

Definition 24.6 Cubic Polynomials: Are polynomials of degree[cor. 24.1] 3 and have four coefficients:

$$f(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0 \qquad (24.9)$$

1.2. Functional Compositions

Definition 24.7 Functional Compositions $f \circ g$: Let $f : A \mapsto B$ and $g : D \mapsto C$ be to mappings s.t. $\texttt{codom}(f) \subseteq D$ then we can define a composition function $(f \circ g) A \mapsto D$ as:

$$h(\mathbf{x}) = (g \circ f)(\mathbf{x}) = g(f(\mathbf{x})) \quad \text{with} \quad \mathbf{x} \in A \qquad (24.10)$$

Corollary 24.2 Nested Functional Composition:

$$F_{k:1}(\mathbf{x}) = (F_k \circ \cdots \circ F_1)(\mathbf{x}) = F_k\Big(F_{k-1} \circ \cdots \circ (F_1(\mathbf{x}))\Big) \qquad (24.11)$$

2. Proofs

Proof 24.1 formula 24.1:

$$f(x_0) = y_0 = ax_0 + b \quad \Rightarrow \quad b = y_0 - ax_0$$

Definition 24.8 Quadratic Formula: $ax^2 + bx + c = 0$ or in reduced form:

$$x^2 + px + q = 0 \quad \text{with} \quad p = b/a \text{ and } q = c/a$$

Definition 24.9 Discriminant: $\delta = b^2 - 4ac$

Definition 24.10 Solution to [def. 24.8]**:**

$$x_\pm = \frac{-b \pm \sqrt{\delta}}{2a} \quad \text{or} \quad x_\pm = \frac{1}{2}\left(-p \pm \sqrt{p^2 - 4q}\right)$$

Theorem 24.1
Fist Fundamental Theorem of Calculus: Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined $\forall x \in [a, b]$ by:

$$F(X) = \int_a^x f(t)\,dt \qquad (24.12)$$

Then it follows:

$$F'(x) = f(x) \qquad \forall x \in (a, b) \qquad (24.13)$$

Theorem 24.2
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 $[a, b]$:

$$\int_a^b f(t)\,dt = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t)\,dt = F(x) \qquad (24.14)$$

Definition 24.11 Domain of a function $\texttt{dom}(\cdot)$: **Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the set of all possible input values $\mathcal{X}$ is called the domain of f – $\texttt{dom}(f)$.

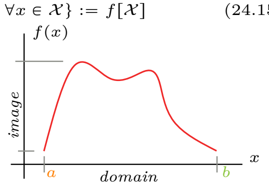
Definition 24.12
Codomain/target set of a function $\texttt{codom}(\cdot)$: **Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the codaomain of that function is the set $\mathcal{Y}$ into which all of the output of the function is **constrained** to fall.

Definition 24.13 Image (Range) of a function: $f[\cdot]$

Given a function $f : \mathcal{X} \to \mathcal{Y}$, the image of that function is the set to which the function can actually map:

$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \qquad (24.15)$$

Evaluating the function f at each element of a given subset A of its domain $\texttt{dom}(f)$ 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 24.14 Inverse Image/Preimage $f^{-1}(\cdot)$: Let $f : X \mapsto Y$ be a function, and A a subset set of its codomain Y.
Then the preimage of A under f is the set of all elements of the domain X, that map to elements in A under f:

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \qquad (24.16)$$

Example 24.1 :
Given $f : \mathbb{R} \to \mathbb{R}$
defined by $f : x \mapsto x^2 \iff f(x) = x^2$
$\texttt{dom}(f) = \mathbb{R}$, $\texttt{codom}(f) = \mathbb{R}$ **but** its image is $f[\mathbb{R}] = \mathbb{R}_+$.

Image (Range) of a subset

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

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \qquad (24.17)$$

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

$$x \leqslant y \iff f(x) \leqslant f(y) \qquad \forall x, y \in \texttt{dom}(f) \qquad (24.18)$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \qquad \forall x, y \in \texttt{dom}(f) \qquad (24.19)$$

Definition 24.16 (strictly) Decreasing Functions: A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geqslant y \iff f(x) \geqslant f(y) \qquad \forall x, y \in \texttt{dom}(f) \qquad (24.20)$$

And *strictly* decreasing if:

$$x > y \iff f(x) > f(y) \qquad \forall x, y \in \texttt{dom}(f) \qquad (24.21)$$

Definition 24.17 Monotonic Function: A function f is called monotonic iff either f is increasing or decreasing.

Definition 24.18 Linear Function:
A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

$$L(\mathbf{x} + \mathbf{y}) = L(\mathbf{x}) + L(\mathbf{y})$$
$$L(\alpha \mathbf{x}) = \alpha L(\mathbf{x}) \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

Corollary 24.3 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{d}{dx}(af(x) + bg(x)) = a\frac{d}{dx}f(x) + b\frac{d}{dx}g(x) \qquad a, b \in \mathbb{R} \qquad (24.22)$$

Definition 24.19 Quadratic Function:
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\intercal \mathbf{A} \mathbf{x} + \mathbf{b}^\intercal \mathbf{x} + c \qquad (24.23)$$

3. Norms

3.1. Infinity/Supremum Norm

Definition 24.20 Infinity/Supremum Norm:

$$\|f\|_\infty := \sup_{x \in \texttt{dom}(f)} |f(x)| \qquad (24.24)$$

Note

In order to make this a proper norm one usually considers *bounded functions* s.t.:

$$\|f\|_\infty \leqslant M < \infty$$

Corollary 24.4 Ininity Norm induced Metric: The infinty norm naturally induces a metric[def. 27.62]:

$$d := (f, g) := \|f - g\|_\infty \qquad (24.25)$$

4. Smoothness

Definition 24.21 Smoothness of a Function $\mathcal{C}^k$: **Given** a function $f : \mathcal{X} \to \mathcal{Y}$, the function is said to be of class k if it is differentiable up to order k **and** continuous, on its entire domain:

$$f \in \mathcal{C}^k(\mathcal{X}) \iff \exists f', f'', \ldots, f^{(k)} \text{ continuous} \qquad (24.26)$$

Note

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

4.0.1. Continuous Functions

Definition 24.22 Continuous Function $\mathcal{C}^0$: Functions that do not have any jumps or peaks.

4.0.2. Piece wise Continuous Functions

Definition 24.23 Piecewise Linear Functions $\mathcal{C}^0_{\text{pw}}$:

4.0.3. Continously Differentiable Function

Corollary 24.5 Continuously Differentiable Function $\mathcal{C}^1$: Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f : \mathcal{X} \to \mathcal{Y}$ of the class must satisfy:

$$f \in \mathcal{C}^1(\mathcal{X}) \iff f' \text{ continuous} \tag{24.27}$$

4.0.4. Smooth Functions

Corollary 24.6 Smooth Function $\mathcal{C}^\infty$: Is a function $f : \mathcal{X} \to \mathcal{Y}$ that has derivatives infinitely many times differerntiable.

$$f \in \mathcal{C}^\infty(\mathcal{X}) \iff f', f'', \ldots, f^{(\infty)} \tag{24.28}$$

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 differentiablility.

Definition 24.24 Lipschitz Continuity: A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Constant L:

$$|f(\mathbf{x}) - f(\mathbf{y})| \leqslant L\|\mathbf{x}-\mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y}, \qquad L > 0 \tag{24.29}$$

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 24.25 Lipschitz Continuous Gradient: A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *L-Lipschitz continuous gradient* if it satisfies:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leqslant L\|\mathbf{x}-\mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \operatorname{dom}(f), \qquad L > 0 \tag{24.30}$$

if $f \in \mathcal{C}^2$, this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leqslant L\mathbf{I} \qquad \forall \mathbf{x} \in \operatorname{dom}(f), \qquad L > 0 \tag{24.31}$$

Lemma 24.1 Descent Lemma [Poorfs 24.5,??]: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (24.30) over its domain, then it holds that:

$$|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})| \leqslant \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \tag{24.32}$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian (Definition 25.8) of f is uniformly upper bounded by L

4.2. L-Smooth Functions

Definition 24.26 L-Smoothness:
A L-smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

with $\qquad \forall \mathbf{x}, \mathbf{y} \in \operatorname{dom}(f), \qquad L > 0 \quad (24.33)$

If f is a twice differentiable this is equivalent to:

$$\nabla^2 f(\mathbf{x}) \leqslant L\mathbf{I} \qquad L > 0 \tag{24.34}$$

Theorem 24.3 [proof 24.6]
L-Smoothness of convex functions:
A *convex* and L-Smooth function ([def. 24.26]) has a *Lipschitz continuous gradient*eq. (24.30) thus it holds that:

$$f(\mathbf{x}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y}) \leqslant \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \tag{24.35}$$

Note

L-smoothnes is a weaker condition than L-Lipschitz continuous gradients

5. Convexity

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

Definition 24.27 Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:

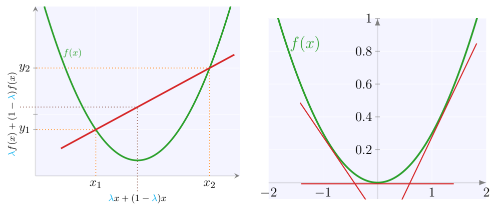
$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) \tag{24.36}$$
$$\forall \lambda \in [0,1] \qquad \forall x, y \in \operatorname{dom}(f)$$

If f is a differentiable function this is equivalent to:

$$f(x) \geqslant f(y) + \nabla f(y)^\intercal(x-y) \qquad \forall x, y \in \operatorname{dom}(f) \tag{24.37}$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(x) \geqslant 0 \qquad \forall x, y \in \operatorname{dom}(f) \tag{24.38}$$



(a) (b)

Definition 24.28 Concave Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if it satisfies:

$$f(\lambda x + (1-\lambda)y) \geqslant \lambda f(x) + (1-\lambda)f(y) \qquad \begin{matrix}\forall x, y \in \operatorname{dom}(f)\\ \forall \lambda \in [0,1]\end{matrix} \tag{24.39}$$

Corollary 24.7 Convexity → global minimima: Convexity implies that all local minima (if they exist) are global minima.

Definition 24.29 Strictly Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \qquad \begin{matrix}\forall x, y \in \operatorname{dom}(f)\\ \forall \lambda \in [0,1]\end{matrix}$$

If f is a differentiable function this is equivalent to:

$$f(x) > f(y) + \nabla f(y)^\intercal(x-y) \qquad \forall x, y \in \operatorname{dom}(f) \tag{24.40}$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(x) > 0 \qquad \forall x, y \in \operatorname{dom}(f) \tag{24.41}$$

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. (24.40) implies that $f(x)$ is above the tangent $f(x) + \nabla f(x)^\intercal(y-x)$ for all $x, y \in \operatorname{dom}(f)$
- ?? implies that $f(x)$ is flat or curved upwards

Corollary 24.8 Strict Convexity → Uniqueness:
Strict convexity implies a unique minimizer $\iff$ at most one global minimum.

Corollary 24.9 : A twice differentiable function of one variable $f : \mathbb{R} \to \mathbb{R}$ is convex on an interval $\mathcal{X} = [a,b]$ if and only if its second derivative is non-negative on that interval $\mathcal{X}$:

$$f''(x) \geqslant 0 \qquad \forall x \in \mathcal{X} \tag{24.42}$$

Definition 24.30 μ-Strong Convexity:
Let $\mathcal{X}$ be a Banach space over $\mathbb{K} = \mathbb{R}, \mathbb{C}$. A function $f : \mathcal{X} \to \mathbb{R}$ is called strongly convex iff the following equation holds:

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

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

$$f(y) \geqslant f(x) + \nabla f(x)^\intercal(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \tag{24.43}$$

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

$$\nabla^2 f(x) \geqslant \mu\mathbf{I} \qquad \forall x, y \in \mathcal{X} \qquad \mu > 0 \tag{24.44}$$

Corollary 24.10 Strong Convexity implies Strict Convexity:

https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex

Property 24.3:

$$f(\mathbf{y}) \leqslant f(\mathbf{y}) + \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y}) + \frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \tag{24.45}$$

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 24.2: eq. (24.44) analogously to **Proof** eq. (24.34)

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian ([def. 25.8]) of f is uniformly lower bounded by μ
Hence strong convexity can be considered as the analogous to smoothness

Example 24.2 Quadratic Function: A quadratic function eq. (24.23) is convex if:

$$\nabla^2_{\mathbf{x}}\text{eq. }(24.23) = \mathbf{A} \geqslant 0 \tag{24.46}$$

Corollary 24.11 :
Strong convexity $\Rightarrow$ Strict convexity $\Rightarrow$ Convexity

5.1. Properties that preserve convexity

Property 24.4 Non-negative weighted Sums: Let f be a convex function then $g(x)$ is convex as well:

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

Property 24.5 Composition of Affine Mappings: Let f be a convex function then $g(x)$ is convex as well:

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

Property 24.6 Pointwise Maxima: Let f be a convex function then $g(x)$ is convex as well:

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

Functions

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

$$f(-x) = f(x) \tag{24.47}$$

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.

$$f(-x) = -f(x) \tag{24.48}$$

Theorem 24.4 Rules:
Let f be even and f odd respectively.
$g =: f \cdot f$ is even $\qquad g =: f \cdot f$ is even
$g =: f \cdot f$ is odd $\qquad$ the same holds for division

Examples

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

$$x\text{-Shift:} \qquad \begin{matrix} f(x-c) \Rightarrow \text{shift to the right} \\ f(x+c) \Rightarrow \text{shift to the left}\end{matrix} \tag{24.49}$$
$$y\text{-Shift:} \qquad f(x) \pm c \Rightarrow \text{shift up/down} \tag{24.50}$$

Proof 24.3: **eq. (24.49)** $f(x_n - c)$ we take the x-value at x_n but take the y-value at $x_o := x_n - c$
$\Rightarrow$ we shift the function to x_n.

Euler's formula

$$e^{\pm ix} = \cos x \pm i \sin x \tag{24.51}$$

Euler's Identity

$$e^{\pm i} = -1 \tag{24.52}$$

Note

$$e^n = 1 \Leftrightarrow n = i\,2\pi k, \qquad k \in \mathbb{N} \tag{24.53}$$

Corollary 24.12 Every norm is a convex function: By using definition [def. 24.27] and the triangular inequality it follows (with the exception of the L0-norm):

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

5.2. Taylor Expansion

Definition 24.31 Taylor Expansion:

$$T_n(x) = \sum_{i=0}^n \frac{1}{n!} f^{(i)}(x_0) \cdot (x-x_0)^{(i)} \tag{24.54}$$
$$= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \mathcal{O}(x^3) \tag{24.55}$$

Definition 24.32 Incremental Taylor:
Goal: evaluate $T_n(x)$ (eq. (24.55)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^n \frac{h^i}{n!} f^{(i)}(x_0)^{i-1} \tag{24.56}$$
$$= f(x_0) \pm hf'(x_0) + \frac{h^2}{2}f''(x_0) \pm f'''(x_0)(h)^3 + \mathcal{O}(h^4)$$

Note

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

Definition 24.33 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\mathbf{x} \in X$, $f : X \mapsto \mathbb{R}$ and $f \in \mathcal{C}^2$ then it holds that

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla_{\mathbf{x}} f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^\intercal H(\mathbf{x}-\mathbf{x}_0) \tag{24.57}$$

Definition 24.34 Argmax: The argmax of a function defined on a set D is given by:

$$\arg\max_{x \in D} f(x) = \{x | f(x) \geqslant f(y), \forall y \in D\} \tag{24.58}$$

Definition 24.35 Argmin: The argmin of a function defined on a set D is given by:

$$\arg\min_{x \in D} f(x) = \{x | f(x) \leqslant f(y), \forall y \in D\} \tag{24.59}$$

Corollary 24.13 Relationship $\arg\min \leftrightarrow \arg\max$:

$$\arg\min_{x \in D} f(x) = \arg\max_{x \in D} -f(x) \tag{24.60}$$

Property 24.7 Argmax Identities:

1. **Shifting**:
$$\forall \lambda \text{ const} \qquad \arg\max f(x) = \arg\max f(x) + \lambda \tag{24.61}$$
2. **Positive Scaling**:
$$\forall \lambda > 0 \text{ const} \qquad \arg\max f(x) = \arg\max \lambda f(x) \tag{24.62}$$
3. **Negative Scaling**:
$$\forall \lambda < 0 \text{ const} \qquad \arg\max f(x) = \arg\min \lambda f(x) \tag{24.63}$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \operatorname{dom}(f)$$
$$\arg\max f(x) = \arg\min \frac{1}{f(x)} \tag{24.64}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions[def. 24.15] g it holds that:
$$\arg\max g(f(x)) = \arg\max f(x) \tag{24.65}$$

Definition 24.36 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\max_{x \in D} f(x) \tag{24.66}$$

Definition 24.37 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\min_{x \in D} f(x) \tag{24.67}$$

Corollary 24.14 Relationship $\min \leftrightarrow \max$**:**

$$\min_{x\in D} f(x) = -\max_{x\in D} -f(x) \tag{24.68}$$

Property 24.8 Max Identities:

1. **Shifting**:
$$\forall\lambda \text{ const} \qquad \max\{f(x)+\lambda\} = \lambda + \max f(x) \tag{24.69}$$
2. **Positive Scaling**:
$$\forall\lambda > 0 \text{ const} \qquad \max \lambda f(x) = \lambda \max f(x) \tag{24.70}$$
3. **Negative Scaling**:
$$\forall\lambda < 0 \text{ const} \qquad \max \lambda f(x) = \lambda \min f(x) \tag{24.71}$$
4. **Positive Functions**:
$$\forall \arg\max f(x) > 0, \forall x \in \texttt{dom}(f) \qquad \max \frac{1}{f(x)} = \frac{1}{\min f(x)} \tag{24.72}$$
5. **Stricly Monotonic Functions**: for all strictly monotonic increasing functions[def. 24.15] g it holds that:
$$\max g(f(x)) = g(\max f(x)) \tag{24.73}$$

Definition 24.38 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x\in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y|y\geqslant f(x), \forall x\in D} y \tag{24.74}$$

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

Definition 24.39 Infinmum: The infinmum of a function defined on a set D is given by:

$$\inf_{x\in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y|y\leqslant f(x), \forall x\in D} y \tag{24.75}$$

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

Corollary 24.15 Relationship $\sup \leftrightarrow \inf$**:**

$$\in_{x\in D} f(x) = -\sup_{x\in D} -f(x) \tag{24.76}$$

Note

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

Definition 24.40 Time-invariant system (TIS): A function f is called time-invariant, if shifting the input in time leads to the same output shifted in time by the same amount.

$$y(t) = f(x(t),t) \xrightarrow[\forall\tau]{\text{time-invariance}} y(t-\tau) = f(x(t-\tau),t) \tag{24.77}$$

Definition 24.41 Inverse Function $g = f^{-1}$**:**
A function g is the inverse function of the function $f : A \subset \mathbb{R} \to B \subset \mathbb{R}$ if

$$f(g(x)) = x \qquad \forall x \in \texttt{dom}(g) \tag{24.78}$$

and

$$g(f(u)) = u \qquad \forall u \in \texttt{dom}(f) \tag{24.79}$$

Property 24.9
Reflective Property of Inverse Functions: f contains (a, b) if and only if f^{-1} contains (b, a).
The line $y = x$ is a symmetry line for f and f^{-1}.

Theorem 24.5 The Existence of an Inverse Function:
A function has an inverse function if and only if it is one-to-one.

Corollary 24.16 Inverse functions and strict monotonicity: If a function f is strictly monotonic [def. 24.17] 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 24.42 The gamma function $\Gamma(\alpha)$**:** Is extension of the factorial function (**??**) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\,\mathrm{d}x \qquad \Re(z) > 0 \tag{24.80}$$

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

7. Proofs

Proof 24.4: lemma 24.1 for $\mathcal{C}^1$ functions:
Let $g(t) \equiv f(\mathbf{y} + t(\mathbf{x}-\mathbf{y}))$ from the FToC (theorem 24.2) we know that:

$$\int_0^1 g'(t)\,\mathrm{d}t = g(1) - g(0) = f(\mathbf{x}) - f(\mathbf{y})$$

It then follows from the reverse:

$$\begin{aligned}
&|f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})| \\
\overset{\substack{\text{Chain. R}\\ \text{FToC}}}{=} &\left|\int_0^1 \nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t - \nabla f(\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})\right| \\
= &\left|\int_0^1 (\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t\right| \\
= &\left|\int_0^1 (\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y}))^\intercal(\mathbf{x}-\mathbf{y})\,\mathrm{d}t\right| \\
\overset{\text{C.S.}}{\leqslant} &\left|\int_0^1 \|\nabla f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - \nabla f(\mathbf{y})\| \cdot \|\mathbf{x}-\mathbf{y}\|\,\mathrm{d}t\right| \\
\overset{\text{eq. (24.30)}}{=} &\left|\int_0^1 L\|\mathbf{y}+t(\mathbf{x}-\mathbf{y})-\mathbf{y}\| \cdot \|\mathbf{x}-\mathbf{y}\|\,\mathrm{d}t\right| \\
= &\left|L\|\mathbf{x}-\mathbf{y}\|^2\int_0^1 t\,\mathrm{d}t\right| = \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2
\end{aligned}$$

Proof 24.5: **??** for $\mathcal{C}^2$ functions:

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

Now we plug in $\nabla^2 f(\mathbf{x})$ and recover eq. (24.33):

$$f(\mathbf{y}) \leqslant f(\mathbf{x}) + \nabla f(\mathbf{x})^\intercal(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^\intercal L(\mathbf{y}-\mathbf{x})$$

Proof 24.6: theorem 24.3:
With the definition of convexity for a differentiable function (eq. (24.40)) it follows

$$\begin{aligned}
&f(x) - f(y) + \nabla f(y)^\intercal(x-y) \geqslant 0 \\
\Rightarrow &|f(x) - f(y) + \nabla f(y)^\intercal(x-y)| \\
\overset{\text{if eq. (24.40)}}{=} &f(x) - f(y) + \nabla f(y)^\intercal(x-y)
\end{aligned}$$

with lemma 24.1 and [def. 24.26] it follows theorem 24.3

Differential Calculus

1. The Chain Rule

Formula 25.1 Generalized Chain Rule:
Let $\mathbf{F}:\mathbb{R}^n\mapsto\mathbb{R}^k$ and $\mathbf{G}:\mathbb{R}^k\mapsto\mathbb{R}^m$ be to general maps then it holds:

$$\underbrace{\partial(\mathbf{G}\circ\mathbf{F})}_{\mathbb{R}^n\mapsto\mathbb{R}^{m\times n}} = \underbrace{(\partial\mathbf{G}\circ\mathbf{F})\cdot\partial\mathbf{F}}_{\mathbb{R}^n\mapsto\left(\mathbb{R}^{m\times k}\cdot\mathbb{R}^{k\times n}\right)} \qquad \begin{aligned}\partial F&:\mathbb{R}^n\mapsto\mathbb{R}^{k\times n}\\ \partial G&:\mathbb{R}^k\mapsto\mathbb{R}^{m\times k}\end{aligned} \tag{25.1}$$

2. Directional Derivative

3. Partial Differentiation

Definition 25.1 Partial Derivative:
Let $f:\mathbb{R}^n\mapsto\mathbb{R}$ be a real valued function, its partial derivative $\partial_i f:\mathbb{R}^n\mapsto\mathbb{R}$ is defined as the directional derivative?? along the coordinate axis of one of its variables:

$$\begin{aligned}\partial_i f(\mathbf{x}) &= \frac{\partial f}{\partial x_i} = D_{x_i}f = \lim_{h\to 0}\frac{f(\mathbf{x}, x_i\leftarrow x_i+h)-f(\mathbf{x})}{h}\\ &= \lim_{h\to 0}\frac{f(x_1,\dots,x_i+h,\dots,x_n)-f(x_1,\dots,x_i,\dots,x_n)}{h}\end{aligned} \tag{25.2}$$

3.1. The Gradient

3.1.1. The Nabla Operator

Definition 25.2 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:

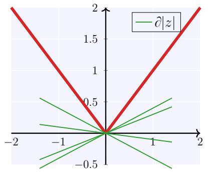
$$\nabla = \sum_{i=1}^n \frac{\partial}{\partial x_i}\tilde{\mathbf{e}}_i = \begin{bmatrix}\frac{\partial}{\partial x_1}(\mathbf{x})\\ \frac{\partial}{\partial x_2}(\mathbf{x})\\ \vdots\\ \frac{\partial}{\partial x_n}(\mathbf{x})\end{bmatrix} \tag{25.3}$$

Definition 25.3 Gradient:
Given a *scalar valued* function $f:\mathbb{R}^n\mapsto\mathbb{R}$ its gradient $\nabla f:\mathbb{R}^n\mapsto\mathbb{R}^n$ is defined as vector $\mathbb{R}^n$ of the partial derivatives[def. 25.1] w.r.t. all coordinate axes:

$$\operatorname{grad} f(\mathbf{x}) := \nabla f(\mathbf{x}) = \begin{bmatrix}\frac{\partial f}{\partial x_1}(\mathbf{x})\\ \frac{\partial f}{\partial x_2}(\mathbf{x})\\ \vdots\\ \frac{\partial f}{\partial x_n}(\mathbf{x})\end{bmatrix} = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^{\mathsf{T}} \tag{25.4}$$

3.1.2. The Subderivative

Definition 25.4 [proof 25.1]
Subgradient g:



Let $f:R^n\mapsto\mathbb{R}$ be a continuous (not necessarily differentiable) function.
$g\in\mathbb{R}^n$ is a subgradient of f at a point $\mathbf{x}_0\in\mathbb{R}^n$ if it satisfies:

$$g: f(\mathbf{x})-f(\mathbf{x}_0)\geqslant \mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0) \tag{25.5}$$

Definition 25.5 [example 25.1]
Subderivative $\partial f(\mathbf{x}_0)$:
Let $f:R^n\mapsto\mathbb{R}$ be a continuous (not necessarily differentiable) function. The subdifferential of f at a point $\mathbf{x}_0\in\mathbb{R}^n$ is defined as the set of all possible subgradients[def. 25.4] g:

$$\partial f(\mathbf{x}_0)\left\{g: f(\mathbf{x})-f(\mathbf{x}_0)\geqslant \mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0)\quad \forall\mathbf{x}\in\mathbb{R}^n\right\} \tag{25.6}$$

Heuristic

We can guess the sub derivative at a point by looking at all the slopes that are smaller then the graph.

3.2. The Jacobian

Definition 25.6
Jacobian/Jacobi Matrix $\mathbf{Df},\mathbf{J_f}$:
Given a *vector valued* function
$\mathbf{f}:\mathbb{R}^n\mapsto\mathbb{R}^m$ its derivative $\mathbf{J_f}:\mathbb{R}^n\mapsto\mathbb{R}^{m\times n}$
with components $\partial_{ij}\mathbf{f} = \partial_i f_j:\mathbb{R}^n\mapsto\mathbb{R}$ is a vector valued function defined as:

$$\mathbf{J}(\mathbf{f}(\mathbf{x})) = \mathbf{J_f}(\mathbf{x}) = \mathbf{Df} = \frac{\partial\mathbf{f}}{\partial\mathbf{x}}(\mathbf{x}) = \frac{\partial(f_1,\dots,f_m)}{\partial(x_1,\dots,x_n)}(\mathbf{x}) \tag{25.7}$$

$$= \begin{bmatrix}\frac{\partial\mathbf{f}}{\partial x_1} & \cdots & \frac{\partial\mathbf{f}}{\partial x_n}\end{bmatrix} = \begin{bmatrix}\nabla^{\mathsf{T}}f_1\\ \vdots\\ \nabla^{\mathsf{T}}f_m\end{bmatrix}$$

$$= \begin{bmatrix}\frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x})\\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & & & \frac{\partial f_2}{\partial x_n}(\mathbf{x})\\ \vdots & & & \vdots\\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x})\end{bmatrix}$$

Explanation 25.1. *Rows of the Jaccobian are transposed gradients[def. 25.3] of the component functions $f_1,\dots,f_m$.*

Corollary 25.1 :

4. Second Order Derivatives

Definition 25.7 Second Order Derivative $\frac{\partial^2}{\partial x_i\partial x_j}$:

Theorem 25.1
Symmetry of second derivatives/Schwartz's Theorem:
Given a continuous and twice differentiable function $f:\mathbb{R}^n\mapsto\mathbb{R}$ then its second order partial derivatives commute:

$$\frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}$$

4.1. The Hessian

Definition 25.8 Hessian Matrix:
Given a function $f:\mathbb{R}\mapsto\mathbb{R}^n$ its Hessian$\in\mathbb{R}^{n\times n}$ is defined as:

$$\mathbf{H}(\mathbf{f})(\mathbf{x}) = \mathbf{H}_f(\mathbf{x}) = \mathbf{J}(\nabla\mathbf{f}(\mathbf{x}))^T \tag{25.8}$$

$$= \begin{bmatrix}\frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}(\mathbf{x})\\ \frac{\partial^2 f}{\partial x_2\partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & & \frac{\partial^2 f}{\partial x_2\partial x_n}(\mathbf{x})\\ \vdots & & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_n\partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x})\end{bmatrix}$$

and it corresponds to the Jacobian of the Gradient.
Due to the differentiability and theorem 25.1 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 25.2 Eigenvector basis of the Hessian: Due to the fact that the Hessian is real and symmetric we can decompose it into a set of real eigenvalues and an orthogonal basis of eigenvectors $\{(\lambda_1,\mathbf{v}_1),\dots,\lambda_n,\mathbf{v}_n\}$.
Not let $\mathbf{d}$ be a directional unit vector then the second derivative in that direction is given by:

$$\mathbf{d}^{\mathsf{T}}\mathbf{H}\mathbf{d} \iff \mathbf{d}^{\mathsf{T}}\sum_{i=1}^n\lambda_i\mathbf{v}_i \overset{\text{if }\mathbf{d}=\mathbf{v}_j}{\iff} \mathbf{d}^{\mathsf{T}}\lambda_j\mathbf{v}_j$$

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

5. Extrema

Definition 25.9 Critical/Stationary Point: Given a function $f:\mathbb{R}^n\mapsto\mathbb{R}$, that is differentiable at a point $\mathbf{x}_0$ then it is called a critical point if the functions derivative vanishes at that point:

$$f'(\mathbf{x}_0) = 0 \iff \nabla_{\mathbf{x}}f(\mathbf{x}_0) = 0$$

Corollary 25.3 Second Derivative Test $f:\mathbb{R}\mapsto\mathbb{R}$:
Suppose $f:\mathbb{R}\mapsto\mathbb{R}$ is twice differentiable at a stationary point x [def. 25.9] then it follows that:

- $f''(x) > 0 \iff$ $f'(x+\epsilon) > 0$ slope points uphill; $f'(x-\epsilon) < 0$ slope points downhill; $f(x)$ is a local minimum
- $f''(x) < 0 \iff$ $f'(x+\epsilon) > 0$ slope points downhill; $f'(x-\epsilon) < 0$ slope points uphill; $f(x)$ is a local maximum

$\epsilon > 0$ sufficiently small enough

Corollary 25.4 Second Derivative Test $f:\mathbb{R}^n\mapsto\mathbb{R}$:
Suppose $f:\mathbb{R}^n\mapsto\mathbb{R}$ is twice differentiable at a stationary point $\mathbf{x}$ [def. 25.9] then it follows that:

- If $\mathbf{H}$ is p.d $\iff \forall\lambda_i > 0\in\mathbf{H} \quad\to\quad f(\mathbf{x})$ is a local min.
- If $\mathbf{H}$ is n.d $\iff \forall\lambda_i < 0\in\mathbf{H} \quad\to\quad f(\mathbf{x})$ is a local max.
- If $\exists\lambda_i > 0\in\mathbf{H}$ and $\exists\lambda_i < 0\in\mathbf{H}$ then $\mathbf{x}$ is a local maximum in one cross section of f but a local minimum in another
- If $\exists\lambda_i = 0\in\mathbf{H}$ and all other eigenvalues have the same sign the test is inclusive as it is inconclusive in the cross section corresponding to the zero eigenvalue.

Note

If $\mathbf{H}$ is positive definite for a minima $\mathbf{x}^*$ of a *quadratic* function f then this point must be a global minimum of that function.

6. Proofs

Proof 25.1: Definition 25.4 $f(\mathbf{x})\geqslant f(\mathbf{x}_0)+\mathbf{g}^{\mathsf{T}}(\mathbf{x}-\mathbf{x}_0)\quad\forall\mathbf{x}\in\mathbb{R}^n$ corresponds to a line (see formula 24.1) at the point $\mathbf{x}_0$ with slope $\mathbf{g}^{\mathsf{T}}$.
Thus we search for all lines with smaller slope then function graph.

7. Examples

Example 25.1 Subderivatives Absolute Value Function $|x|$: $f:\mathbb{R}\mapsto\mathbb{R}$ with $f(x)=|x|$ at the point $x=0$ it holds:

$$f(x)-f(0)\geqslant gx \implies \text{the interval } [-1;1]$$

For $x\neq 0$ the subgradient is equal to the gradient. Thus it follows for the subderivatives/differentials:

$$\partial|x| = \begin{cases}-1 & \text{if } x<0\\ [-1,1] & \text{if } x=0\\ 1 & \text{if } x>0\end{cases}$$

Theorem 26.1 Important Integral Properties:

Addition $$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x \tag{26.1}$$

Reflection $$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x \tag{26.2}$$

Translation $$\int_a^b f(x)\,\mathrm{d}x \overset{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x \mp c)\,\mathrm{d}x \tag{26.3}$$

f **Odd** $$\int_{-a}^{a} f(x)\,\mathrm{d}x = 0 \tag{26.4}$$

f **Even** $$\int_{-a}^{a} f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x \tag{26.5}$$

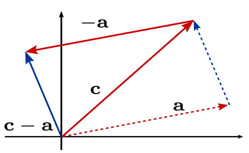
Proof 26.1: **eqs. (26.4) and (26.5)**

$$\begin{aligned}
I := \int_{-a}^{a} f(x)\,\mathrm{d}x &= \int_{-a}^{0} f(x)\,\mathrm{d}x + \int_0^a f(x)\,\mathrm{d}x \\
&\overset{\substack{t=-x \\ dt=-dx}}{=} -\int_a^0 f(-x)\,\mathrm{d}x + \int_0^a f(x)\,\mathrm{d}x \\
&= \int_0^a f(-x) + f(x)\,\mathrm{d}x = \begin{cases} 0 & \text{if} \quad f \quad \text{odd} \\ 2I & \text{if} \quad f \quad \text{even} \end{cases}
\end{aligned}$$

Linear Algebra

1. Vectors

Definition 27.1 Vector Substraction:



$$\mathbf{b} = \mathbf{c} - \mathbf{a} \tag{27.1}$$

2. Linear Systems of Equations

2.1. Gaussian Elimination

2.1.1. Rank

Definition 27.2 Matrix Rank $\mathfrak{rank}$**:** The ranks of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as the the dimension[def. 27.11] of the vector space spaned[def. 27.7] by its row or column vectors:

$$\begin{aligned}\mathfrak{rank}(\mathbf{A}) &= \dim\left(\{\mathbf{a}_{:,1},\dots,\mathbf{a}_{:,n}\}\right)\\ &= \dim\left(\{\mathbf{a}_{1,:},\dots,\mathbf{a}_{m,:}\}\right)\\ &\overset{\text{def. }27.48}{=} \dim(\mathfrak{R}(\mathbf{A}))\end{aligned} \tag{27.2}$$

Corollary 27.1 :

- The column-and row-ranks of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ are equal.
- The rank of a non-symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is limited by the smaller dimension:
$$\mathfrak{rank}(\mathbf{A}) \leqslant \min\{n, m\} \tag{27.3}$$

Property 27.1 Rank of Matrix Product: Let $\mathbf{A} \in \mathbb{R}^{m,n}$ and $\mathbf{B} \in \mathbb{R}^{n,p}$ then the rank of the matrix product is limited:

$$\mathfrak{rank}(\mathbf{AB}) \leqslant \min\{\mathfrak{rank}(\mathbf{A}), \mathfrak{rank}(\mathbf{B})\} \tag{27.4}$$

3. Sparse Linear Systems

Definition 27.3 Sparse Matrix $\mathbf{A} \in \mathbb{K}^{m,n}, m, n \in \mathbb{N}_{>0}$**:** A matrix $\mathbf{A}$ is sparse if:

$$\text{nnz}(\mathbf{A}) \ll mn \qquad \mathbf{A} \in \mathbb{K}^{m,n}, m, n \in \mathbb{N}_{>0} \tag{27.5}$$

$$\text{nnz} := \#\left\{(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\} : a_{i,j} \neq 0\right\}$$

4. Vector Spaces

4.1. Vector Space

Definition 27.4 Vector Space: TODO

4.2. Vector Subspace

Definition 27.5 Vector Subspaces: A non-empty subset U of a $\mathbb{K}$-vector space $\mathcal{V}$ is called a sub-space of $\mathcal{V}$ if it satisfies:

$$\mathbf{u}, \mathbf{v} \in U \implies \mathbf{u} + \mathbf{v} \in U \tag{27.6}$$

$$\mathbf{u} \in U \implies \lambda\mathbf{u} \in U \qquad \forall\lambda \in \mathbb{K} \tag{27.7}$$

Definition 27.6 Linearcombination: Let $X = \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathcal{V}$ be a non-empty and finite subset of vectors of an $\mathbb{K}$-vector space $\mathcal{V}$. A *linear combination* of X is a combination of the vectors defined as:

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i = \alpha_1\mathbf{v}_1 + \ldots + \alpha_n\mathbf{v}_n \qquad \alpha_i \in \mathbb{K} \tag{27.8}$$

Definition 27.7 Span/Linear Hull $\langle X\rangle$**:** Is the set of all possible linear combinations[def. 27.6] of finite set $X = \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathcal{V}$ of a $\mathbb{K}$ vector space $\mathcal{V}$:

$$\langle X\rangle = \operatorname{span}(X) = \left\{\mathbf{v} \middle| \sum_{i=1}^{n} \alpha_i\mathbf{v}_i, \forall\alpha_i \in \mathbb{K}\right\} \tag{27.9}$$

Definition 27.8 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. 27.7] $\mathcal{V}$:

$$\operatorname{span}(\mathbf{v}_1 \dots, \mathbf{v}_m) = \mathcal{V} \tag{27.10}$$

Explanation 27.1 (Definition 27.8).
The generating set of vector space (or set of vectors) $\mathcal{V} \overset{i.e.}{=} \mathbb{R}^n$ *is a subset* $X = \{\mathbf{v}_1, \dots, \mathbf{v}_m\} \subset \mathcal{V}$ *s.t. every element of* $\mathcal{V}$ *can be produced by* span(X).

Definition 27.9 Linear Independence: A set of vector $\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \in \mathcal{V}$ is called linear independent if the satisfy:

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i\mathbf{v}_i = \mathbf{0} \iff \alpha_1 = \ldots = \alpha_n = 0 \tag{27.11}$$

Corollary 27.2 : A set of vector $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathcal{V}$ is called linear independent, if for every subset $X = \mathbf{x}_1, \dots, \mathbf{x}_m \subseteq \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ it holds that:

$$\langle X\rangle \subsetneq \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \tag{27.12}$$

4.3. Basis

Definition 27.10 Basis $\mathfrak{B}$**:** A subset $\mathfrak{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of a $\mathbb{K}$-vector space $\mathcal{V}$ is called a basis of $\mathcal{V}$ if:

$$\langle\mathfrak{B}\rangle = \mathcal{V} \quad \text{and} \quad \mathfrak{B} \text{ is a linear independent generating set} \tag{27.13}$$

Corollary 27.3 : The unit vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ build a standard basis of the $\mathbb{R}^n$.

Corollary 27.4 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. 27.6] of $\mathfrak{B}$ by a unique set of coefficients α_i:

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i\mathbf{b}_i \qquad \begin{matrix}\alpha_1, \dots, \alpha_n \in \mathbb{K}\\ \mathbf{b}_1, \dots, \mathbf{b}_n \in \mathfrak{B}\end{matrix} \tag{27.14}$$

4.3.1. Dimensionality

Definition 27.11 Dimension of a vector space $\dim(\mathcal{V})$**:** Let $\mathcal{V}$ be a vector space. The dimension of $\mathcal{V}$ is defined as the number of necessary basis vectors $\mathfrak{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ in order to span $\mathcal{V}$:

$$\dim(\mathcal{V}) := |\mathfrak{B}| = n \in \mathbb{N}_0 \tag{27.15}$$

Corollary 27.5 : 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 27.12 Affine Subspaces: Given a $\mathbb{K}$-vector space $\mathcal{V}$ of dimension $\dim(\mathcal{V}) \geqslant 2$ a sub vector space[def. 27.5] U of $\mathcal{V}$ defined as:

$$\mathcal{W} := \mathbf{v} + U = \{\mathbf{v} + \mathbf{x} | \mathbf{x} \in U\} \qquad \mathbf{v} \in \mathcal{V} \tag{27.16}$$

Corollary 27.6 Direction: The sub vector spaces U are called *directions* of $\mathcal{V}$ and it holds:

$$\dim(\mathcal{W}) := \dim(U) \tag{27.17}$$

4.4.1. Hyperplanes

Definition 27.13 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. 27.14]:

$$\mathcal{H} = \left\{\mathbf{x} \in \mathbb{R}^d \middle| \hat{\mathbf{n}}^\mathsf{T}\mathbf{x} - d = 0\right\} \tag{27.18}$$

Corollary 27.7 Half spaces: A hyperplane $\mathcal{H} \in \mathbb{R}^{d-1}$ separates its d-dimensional ambient space into two half spaces:

$$\mathcal{H}^+ = \left\{x \in \mathbb{R}^d \middle| \hat{\mathbf{n}}^\mathsf{T}\mathbf{x} + b > 0\right\} \tag{27.19}$$

$$\mathcal{H}^- = \left\{x \in \mathbb{R}^d \middle| \hat{\mathbf{n}}^\mathsf{T}\mathbf{x} + b < 0\right\} = \mathbb{R}^d - \mathcal{H}^+ \tag{27.20}$$

Notes

Hyperplanes in $\mathbb{R}^2$ are lines and hyperplanes in R^3 are lines.

Hess Normal Form

Definition 27.14 Hess Normal Form: Is an equation to describe hyperplanes[def. 27.13] in $\mathbb{R}^d$:

$$\mathbf{r}^\mathsf{T}\tilde{\mathbf{n}} - d = 0 \iff \tilde{\mathbf{n}}^\mathsf{T}(\mathbf{r} - \mathbf{r}_0) \qquad \mathbf{r}_0 := \mathbf{r}^\mathsf{T}d \geqslant 0 \tag{27.21}$$

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. $\mathbf{r}^\mathsf{T}\tilde{\mathbf{n}} \geqslant 0$.

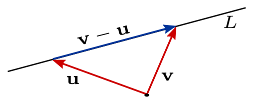
4.4.2. Lines

Definition 27.15 Lines: Lines are a set[def. 20.1] of the form:

$$L = \mathbf{u} + \mathbb{K}\,\mathbf{v} = \{\mathbf{u} + \lambda\mathbf{v} | \lambda \in \mathbb{K}\} \qquad \mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{v} \neq 0 \tag{27.22}$$

Two Point Formula

Definition 27.16 Two Point Formula:



$$L = \mathbf{u} + \mathbb{K}\,\mathbf{v} \tag{27.23}$$

4.4.3. Planes

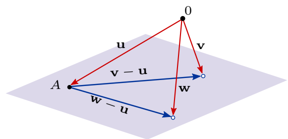
Definition 27.17 Planes: Planes are sets defined as:

$$E = \mathbf{u} + \mathbb{K}\,\mathbf{v} + \mathbb{K}\,\mathbf{w} = \{\mathbf{u} + \lambda\mathbf{v} + \mu\mathbf{w} | \lambda, \mu \in \mathbb{K}\} \tag{27.24}$$

$$\mathbf{u}, \mathbf{w} \in \mathcal{V} \qquad \text{s.t. } \mathbf{v}, \mathbf{u} \neq \mathbf{0} \qquad \text{and} \qquad \mathbf{v}, \mathbf{w} \text{ lin. indep.}$$

Parameterform

Definition 27.18 Two Point Formula:

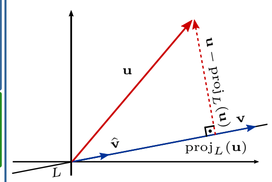


$$E = \mathbf{u} + \mathbb{K}(\mathbf{v} - \mathbf{u}) + \mathbb{K}(\mathbf{w} - \mathbf{u}) \tag{27.25}$$

4.4.4. Minimal Distance of Vector Subspaces

Projections in 2D

Definition 27.19 2D Vector Projection [Proof 27.17,27.18]**:**



$$\begin{aligned}\mathbf{u}_\mathbf{v} &= \operatorname{proj}_L(\mathbf{u})\\ &= u_v\tilde{\mathbf{v}} = (\mathbf{u}^\mathsf{T}\tilde{\mathbf{v}})\,\tilde{\mathbf{v}}\\ &= \frac{\mathbf{u}^\mathsf{T}\mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} = \frac{\mathbf{u}^\mathsf{T}\mathbf{v}}{\mathbf{v}^\mathsf{T}\mathbf{v}}\mathbf{v}\end{aligned} \tag{27.26}$$

Corollary 27.8 [proof 27.8]
2D Projection Matrix P: Is the matrix that satisfies:

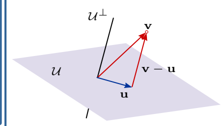
$$\mathbf{Pu} = \operatorname{proj}_L(\mathbf{u}) \qquad \mathbf{P} = \frac{\mathbf{v}\mathbf{v}^\mathsf{T}}{\mathbf{v}^T\mathbf{v}} = \frac{\mathbf{v}\mathbf{v}^\mathsf{T}}{\|\mathbf{v}\|^2} \tag{27.27}$$

Proof 27.1: [Corollary 27.8]

$$\frac{1}{\mathbf{v}^\mathsf{T}\mathbf{v}}\mathbf{u}^\mathsf{T}\mathbf{v}\mathbf{v} = \frac{1}{\mathbf{v}^\mathsf{T}\mathbf{v}}\mathbf{v}\left(\mathbf{v}^\mathsf{T}\mathbf{u}\right) = \frac{1}{\mathbf{v}^\mathsf{T}\mathbf{v}}\left(\mathbf{v}\mathbf{v}^\mathsf{T}\right)\mathbf{u}$$

General Projections

Definition 27.20 [proof 27.19]
General Vector Projection:
Is the orthogonal projection $\mathbf{u}$ of a vector $\mathbf{v}$ onto a sub-vector space $\mathcal{U}$



$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i\mathbf{b}_i \tag{27.28}$$

$$\mathbf{A}\mathbf{A}^\mathsf{T}\alpha_i = \mathbf{A}^\mathsf{T}\mathbf{v} \qquad \mathbf{A} = \begin{pmatrix}\mathbf{b}_1\\ \vdots\\ \mathbf{b}_n\end{pmatrix}$$

where $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a basis of the vector subspace $\mathcal{U}$.

Theorem 27.1 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. 27.65] projection

$$p: \begin{cases}\mathcal{V} \to \mathcal{U}\\ \mathbf{v} \mapsto \mathbf{u}\end{cases} \tag{27.29}$$

the vector $u' := \mathbf{v} - \mathbf{u}$ representing the distance between $\mathbf{u}$ and $\mathbf{v}$ and is minimal:

$$\|\mathbf{u}'\| = \|\mathbf{v} - \mathbf{u}\| \leqslant \|\mathbf{v} - \mathbf{w}\| \qquad \forall\mathbf{w} \in \mathcal{U} \qquad \mathbf{u}' \in \mathcal{U}^\perp \tag{27.30}$$

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 27.21 Symmetric Matrices: A matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ is called *symmetric* if it satisfies:

$$\mathbf{A} = \mathbf{A}^{\mathsf{T}} \tag{27.31}$$

Property 27.2 [proof ??]
Eigenvalues of real symmetric Matrices: The eigenvalues of a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ are real:

$$\operatorname{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geqslant 0}\}_{i=1}^{n} \tag{27.32}$$

Property 27.3 [proof ??]
Orthogonal Eigenvector basis: Eigenvectors of real symmetric matrices with distinct eigenvalues are orthogonal.

Corollary 27.9
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{R}^{n,n}$ is a real *symmetric*[def. 27.21] matrix then its eigenvectors are *orthogonal* and its eigen-decomposition[def. 27.84] is given by:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathsf{T}} \tag{27.33}$$

5.2. Orthogonal Matrices

Definition 27.22 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. 27.66] basis:

$$\langle \mathbf{q}_{:i}, \mathbf{q}_{:j} \rangle = \delta_{ij} \qquad \text{and} \qquad \langle \mathbf{q}_{i:}, \mathbf{q}_{j:} \rangle = \delta_{ij} \tag{27.34}$$

This is exactly true if the inverse of $\mathbf{Q}$ equals its transpose:

$$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathsf{T}} \iff \mathbf{Q}\mathbf{Q}^{\mathsf{T}} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}_n \tag{27.35}$$

Attention: *Orthogonal* matrices are sometimes also called *orthonormal matrices.*

5.3. Hermitian Matrices

Definition 27.23 Conjugate Transpose $\mathbf{A}^{\mathsf{H}}/\mathbf{A}^{*}$
Hermitian Conjugate/Adjoint Matrix:
The conjugate transpose of a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ is defined as:

$$\mathbf{A}^{\mathsf{H}} := (\overline{\mathbf{A}}^{\mathsf{T}}) = \overline{\mathbf{A}^{\mathsf{T}}} \iff \mathbf{a}^{\mathsf{H}}_{i,j} = \bar{\mathbf{a}}_{j,i} \qquad \begin{aligned} 1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m \end{aligned} \tag{27.36}$$

Definition 27.24
Hermitian/Self-Adjoint Matrices $\mathbf{A} = \mathbf{A}^{\mathsf{H}}$:
A hermitian matrix is complex square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ who is equal to its own *conjugate transpose*[def. 27.23]:

$$\mathbf{A} = \mathbf{A}^{\mathsf{H}} = \overline{\mathbf{A}^{\mathsf{T}}} \iff \mathbf{a}_{i,j} = \bar{\mathbf{a}}_{j,i} \qquad i \in \{1,\dots,n\} \tag{27.37}$$

Corollary 27.10 : [def. 27.23] implies that $\mathbf{A}$ must be a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Corollary 27.11 Real Hermitian Matrices: From [cor. 20.1] it follows:

$$\mathbf{A} \in \mathbb{R}^{n\times n} \text{ hermitian} \implies \mathbf{A} \text{ real symmetric}^{[\text{def. } 27.21]} \tag{27.38}$$

Property 27.4 [proof 27.15]
Eigenvalues of Hermitan Matrices: The eigenvalues of a hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ are real:

$$\operatorname{spectrum}(\mathbf{A}) \in \{\mathbb{R}_{\geqslant 0}\}_{i=1}^{n} \tag{27.39}$$

Property 27.5 [proof 27.16]
Orthogonal Eigenvector basis: Eigenvectors of hermitian matrices with distinct eigenvalues are orthogonal.

Corollary 27.12
Eigendecomposition Symmetric Matrices: If $\mathbf{A} \in \mathbb{C}^{n,n}$ is a hermitian matrix[def. 27.24] then its eigendecomposition[def. 27.84] is given by:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathsf{H}} \tag{27.40}$$

5.4. Unitary Matrices

Definition 27.25 Unitary Matrix $\mathbf{U}\mathbf{U}^{\mathsf{H}}$:
is a complex square matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$ whose inverse[def. 27.39] is equal to its *conjugate transpose*[def. 27.23]:

$$\mathbf{U}^{\mathsf{H}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathsf{H}} = \mathbf{I} \tag{27.41}$$

Corollary 27.13 Real Unitary Matrix: A real matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ that is unitary is an *orthogonal matrix*[def. 27.22].

Property 27.6
Preservation of Euclidean Norm [proof 27.14]:
Orthogonal and unitary matrices $\mathbf{Q} \in \mathbb{K}^{n,n}$ do not affect the 2-norm:

$$\|\mathbf{Q}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \qquad \forall \mathbf{x} \in \mathbb{K}^{n} \tag{27.42}$$

5.5. Similar Matrices

Definition 27.26 Similar Matrices: Two square matrices $\mathbf{A} \in \mathbb{K}^{n\times n}$ and $\mathbf{B} \in \mathbb{K}^{n\times n}$ are called *similar* if there exists a invertible matrix $\mathbf{S} \in \mathbb{K}^{n\times n}$ s.t.:

$$\exists \mathbf{S}: \qquad \mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{27.43}$$

Corollary 27.14
Similarity Transformation/Conjugation:
The mapping:

$$\mathbf{A} \mapsto \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \tag{27.44}$$

is called *similarity transformation*

Corollary 27.15
Eigenvalues of Similar Matrices [proof 27.13]:

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. (27.44) $\mathbf{B}$ has the same eigenvalues with transformed eigenvectors:

$$\{\{\lambda_i, \mathbf{u}_i\}\}_{i=1}^{n} \qquad \mathbf{u}_i := \mathbf{S}^{-1}\mathbf{v}_i \tag{27.45}$$

5.6. Skew Symmetric Matrices

Definition 27.27
Skey Symmetric/Antisymmetric Matrices:

$$\mathbf{A}^{\mathsf{T}} = -\mathbf{A} \tag{27.46}$$

5.7. Triangular Matrix

Definition 27.28 Triangular Matrix: An upper (lower) triangular matrix, is a matrix whose element's below (above) the main diagonal are all zero:

$$\begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ & & \ddots & \\ l_{n1} & l_{n2} & & l_{nn} \end{pmatrix} \qquad \begin{pmatrix} u_{11} & u_{12} & & u_{1n} \\ & u_{22} & & u_{2n} \\ & & \ddots & \\ & & & u_{nn} \end{pmatrix}$$

Figure 11: Lower Tri. Mat. Figure 12: Upper Tri. Mat.

5.7.1. Unitriangular Matrix

Definition 27.29 Unitriangular Matrix: An upper (lower) unitriangular matrix, is a upper (lower) triangular matrix[def. 27.28] whose diagonal elements are all ones.

5.7.2. Strictly Triangular Matrix

Definition 27.30 Strictly Triangular Matrix: An upper (lower) strictly triangular matrix, is a upper (lower) triangular matrix[def. 27.28] whose diagonal elements are all zero.

5.8. Block Partitioned Matrices

Definition 27.31 Block Partitioned Matrix:
A matrix $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ can be partitioned into a *block partitioned matrix*:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \qquad \mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \tag{27.47}$$

Definition 27.32 Block Partitioned Linear System:
A linear system $\mathbf{M}\mathbf{x} = \mathbf{b}$ with $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^{k+l}$ can be partitioned into a *block partitioned system*:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \qquad \begin{aligned} &\mathbf{A} \in \mathbb{R}^{k,k}, \mathbf{B} \in \mathbb{R}^{k,l}, \mathbf{C} \in \mathbb{R}^{l,k}, \mathbf{D} \in \mathbb{R}^{l,l} \\ &\mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^{k}, \mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^{l} \end{aligned} \tag{27.48}$$

5.8.1. Schur Complement

Definition 27.33 Schur Complement: Given a block partitioned matrix[def. 27.31] $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ its Schur complements are given by:

$$\mathbf{S}_A = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \qquad \mathbf{S}_D = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \tag{27.49}$$

5.8.2. Inverse of Block Partitioned Matrix

Definition 27.34 proof 27.3
Inverse of a Block Partitioned Matrix:
Given a block partitioned matrix[def. 27.31] $\mathbf{M} \in \mathbb{R}^{k+l,k+l}$ its inverse $\mathbf{M}^{-1}$ can be partitioned as well:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \qquad \mathbf{M}^{-1} = \begin{bmatrix} \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \widetilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{bmatrix} \tag{27.50}$$

$$\widetilde{\mathbf{A}} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}_{\mathbf{A}}^{-1}\mathbf{C}\mathbf{A}^{-1} \qquad \widetilde{\mathbf{C}} = -\mathbf{S}_{\mathbf{A}}^{-1}\mathbf{C}\mathbf{A}^{-1}$$

$$\widetilde{\mathbf{B}} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}_{\mathbf{A}}^{-1} \qquad \widetilde{\mathbf{D}} = \mathbf{S}_{\mathbf{A}}^{-1}$$

where $\mathbf{S}_{\mathbf{A}} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is the Schur complement of $\mathbf{A}$.

5.9. Properties of Matrices

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

Definition 27.35 Square Root:

5.9.2. Trace

Definition 27.36 Trace: The trace of an $\mathbf{A} \in \mathbb{R}^{n\times n}$ matrix is defined as:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn} \tag{27.51}$$

Property 27.7 Trace of a Scalar:

$$\operatorname{tr}(\mathbb{R}) = \mathbb{R} \tag{27.52}$$

Property 27.8 Trace of Transpose:

$$\operatorname{tr}(\mathbf{A}^{\mathsf{T}}) = \operatorname{tr}(\mathbf{A}) \tag{27.53}$$

Property 27.9 Trace of multiple Matrices:

$$\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA}) = \operatorname{tr}(\mathbf{CBA}) \tag{27.54}$$

6. Matrices and Determinants

6.1. Determinants

6.1.1. Laplace/Cofactor Expansion

Definition 27.37 Minor:

Definition 27.38 Cofactors:

Properties

Property 27.10 Determinant times Scalar $\det(\alpha\mathbf{A})$:
Given a matirx $\mathbf{A} \in \mathbb{R}^{n\times n}$ it holds:

$$\det(\alpha \cdot \mathbf{A}) = \alpha^{n}\mathbf{A} \tag{27.55}$$

6.2. Inverese of Matrices

Definition 27.39 Inverse Matrix $\mathbf{A}^{-1}$:

6.2.1. Invertability

Definition 27.40
Singular/Non-Invertible Matrix $\det(\mathbf{A}) = 0$:
A square matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ is singular or non-invertible if it satisfies the following and equal conditions:

- $\det(\mathbf{A}) = 0$
- $\dim(\mathbf{A}) < n$
- $\nexists \mathbf{B}: \mathbf{B} = \mathbf{A}^{-1}$
- $\mathbf{A}\mathbf{x} = \mathbf{b}$ has either
  - no solution $\mathbf{x}$
  - infinitely many solutions $\mathbf{x}$

Transformations And Mapping

7. Linear & Affine Mappings/Transformations

7.1. Linear Mapping

Definition 27.41
Linear Mapping: A linear mapping, function or transformation is a map $l : V \mapsto W$ between two $\mathbb{K}$-vector spaces[def. 27.4] V and W if it satisfies:

$$l(\mathbf{x}+\mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y}) \qquad \text{(Additivity)} \tag{27.56}$$

$$l(\alpha\mathbf{x}) = \alpha l(\mathbf{x}) \quad \forall\alpha \in \mathbb{K} \qquad \text{(Homogenitivity)} \tag{27.57}$$

$$\forall \mathbf{x}, \mathbf{y} \in V$$

Proposition 27.1 [proof 27.8]
Equivalent Formulations: Definition 27.41 is equivalent to:

$$l(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha l(\mathbf{x}) + \beta l(\mathbf{y}) \qquad \begin{matrix}\forall \alpha,\beta \in \mathbb{K}\\ \forall \mathbf{x},\mathbf{y} \in V\end{matrix} \tag{27.58}$$

Corollary 27.16 Superposition Principle:
Definition 27.41 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 27.17 [proof 27.10]
A linear mapping $\iff$ Ax:
For every matrix $\mathbf{A} \in \mathbb{K}^{m\times n}$ the map:

$$l_{\mathbf{A}} : \begin{cases} \mathbb{K}^n & \to \quad \mathbb{K}^m \\ \mathbf{x} & \mapsto \quad \mathbf{Ax} \end{cases} \tag{27.59}$$

is a *linear map* and every linear map l can be represented by a matrix vector product:

$$l \text{ is linear} \iff \exists \mathbf{A} \in \mathbb{K}^{n\times m} : f(x) = \mathbf{Ax} \quad \forall \mathbf{x} \in \mathbb{K}^m \tag{27.60}$$

Principle 27.1 [proof 27.9]
Principle of linear continuation: A linear mapping $l : \mathcal{V} \mapsto \mathcal{W}$ is determined by the image of the basis $\mathfrak{B}$ of $\mathcal{V}$:

$$l(\mathbf{v}) = \sum_{i=1}^{n} \beta_i l(b_i) \qquad \mathfrak{B}(\mathcal{V}) = \{b_1, \ldots, b_n\} \tag{27.61}$$

Property 27.11 [proof 27.11]
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. 27.41].

Definition 27.42 Level Sets:

7.2. Affine Mapping

Definition 27.43 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:

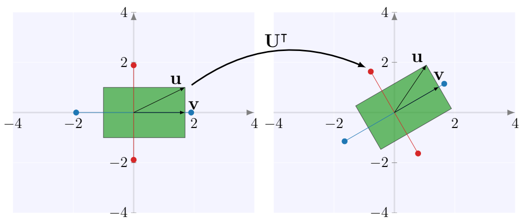
$$\mathbf{Y} = \mathbf{Ax} + \mathbf{b} \tag{27.62}$$

is called an affine transformation of $\mathbf{x}$.

7.3. Orthogonal Transformations

Definition 27.44 Orthogonal Transformation:
A linear transformation $T : \mathcal{V} \mapsto \mathcal{V}$ of an inner product space[def. 27.76] is an orthogonal transformation if preserves the inner product:

$$T(\mathbf{u}) \cdot T(\mathbf{v})\mathbf{u} \cdot \mathbf{v} \qquad \forall \mathbf{u}, \mathbf{v} \in \mathcal{V} \tag{27.63}$$



Corollary 27.18 Orthogonal Matrix Transformation:
An orthogonal matrix[def. 27.22] $\mathbf{Q}$ provides an orthogonal transformation:

$$(\mathbf{Qu})^\intercal(\mathbf{Qv}) = \mathbf{uv} \tag{27.64}$$

Explanation 27.2 (Improper Rotations).
Orthogonal transformations in two or three dimensional euclidean space[def. 27.44] represent improper rotations:

- *Stiff Rotations*
- *Reflections*
- *Reflections+Rotations*

Corollary 27.19 Preservation of Orthogonality: Orthogonal transformation preserver orthogonality.

Corollary 27.20 [proof 27.6]
Preservation of Norm:
An orthogonal transformation $\mathbf{Q} : \mathcal{V} \mapsto \mathcal{V}$ preservers the *length/norm*:

$$\|\mathbf{u}\|_{\mathcal{V}} = \|\mathbf{Qu}\|_{\mathcal{V}} \tag{27.65}$$

Corollary 27.21 Preservation of Angle:
An orthogonal transformation T preservers the *angle*[def. 27.64] of its vectors:

$$\angle(\mathbf{u}, \mathbf{v}) = \angle(T(\mathbf{u}), T(\mathbf{v})) \tag{27.66}$$

7.4. Kernel & Image

7.4.1. Kernel

Definition 27.45 Kernel/Null Space $\mathtt{N}/\varphi^{-1}(\{0\})$:
Let φ be a linear mapping[def. 27.41] between two a $\mathbb{K}$-vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The *kernel* of φ is defined as:

$$\mathtt{N}(\varphi) := \varphi^{-1}(\{\mathbf{0}\}) = \{\mathbf{v} \in \mathcal{V} \mid \varphi(\mathbf{v}) = \mathbf{0}\} \subseteq \mathcal{V} \tag{27.67}$$

Definition 27.46 Right Null Space $\mathtt{N}(\mathbf{A})$:
If $\varphi = \mathbf{A} = \in \mathbb{K}^{m\times n}$ then the eq. (27.67) is equal to:

$$\mathtt{N}(\mathbf{A}) = \varphi_{\mathbf{A}}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^n \mid \mathbf{Av} = \mathbf{0}\} \in \mathbb{K}^m \tag{27.68}$$

Definition 27.47 Left Null Space $\mathtt{N}(\mathbf{A}^\intercal)$:
If $\varphi = \mathbf{A} = \in \mathbb{K}^{m\times n}$ then the *left* null space is defined as:

$$\mathtt{N}(\mathbf{A}^\intercal) = \varphi_{\mathbf{A}^\intercal}^{-1}(\{0\}) = \{\mathbf{v} \in \mathbb{K}^m \mid \mathbf{A}^\intercal\mathbf{v} = \mathbf{0}\} \in \mathbb{K}^n \tag{27.69}$$

Note
The term *left* null space stems from the fact that:

$$(\mathbf{A}^\intercal\mathbf{x})^\intercal = \mathbf{0} \qquad \text{is equal to} \qquad \mathbf{x}^\intercal\mathbf{A} = \mathbf{0}$$

7.4.2. Image

Definition 27.48 Image/Range $\Re/\varphi$:
Let φ be a linear mapping[def. 27.41] between two a $\mathbb{K}$-vector spaces $\varphi : \mathcal{V} \mapsto \mathcal{W}$.
The *imgae* of φ is defined as:

$$\Re(\varphi) := \varphi(\mathcal{V}) = \{\varphi(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}\} \subseteq \mathcal{W} \tag{27.70}$$

Definition 27.49 Column Space $\mathbf{Ax}$:
If $\varphi = \mathbf{A} = (\mathbf{c}_1 \quad \mathbf{c}_n) \in \mathbb{K}^{m\times n}$ then eq. (27.70) is equal to:

$$\begin{aligned}\Re(\mathbf{A}) = \varphi_{\mathbf{A}}(\mathbb{K}^n) &= \left\{\mathbf{Ax}\middle|\forall\mathbf{x} \in \mathbb{K}^n\right\} = \langle(\mathbf{c}_1 \quad \mathbf{c}_n)\rangle \\ &= \left\{\mathbf{v}\middle| \sum_{i=1}^{n} \alpha_i\mathbf{c}_i, \forall\alpha_i \in \mathbb{K}\right\}\end{aligned} \tag{27.71}$$

Definition 27.50 Row Space $\mathbf{A}^\intercal\mathbf{x}$:
If $\varphi = \mathbf{A} = (\mathbf{r}_1^\intercal \quad \mathbf{r}_m^\intercal) \in \mathbb{K}^{m\times n}$ then the column space is defined as:

$$\begin{aligned}\Re(\mathbf{A}^\intercal) = \varphi_{\mathbf{A}}(\mathbb{K}^m) &= \left\{\mathbf{A}^\intercal\mathbf{x}\middle|\forall\mathbf{x} \in \mathbb{K}^m\right\} = \langle(\mathbf{r}_1 \quad \mathbf{r}_m)\rangle \\ &= \left\{\mathbf{v}\middle| \sum_{i=1}^{m} \alpha_i\mathbf{r}_i, \forall\alpha_i \in \mathbb{K}\right\}\end{aligned} \tag{27.72}$$

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

Corollary 27.22 Orthogonality [proof 27.12]:
The *right* (*left*) null space[def. 27.45] is *orthogonal*[def. 27.65] to the *row*[def. 27.50] (*column*[def. 27.49]) space:

$$\mathtt{N}(\mathbf{A}) \perp \Re(\mathbf{A}^\intercal) \qquad \text{and} \qquad \mathtt{N}(\mathbf{A}^\intercal) \perp \Re(\mathbf{A}) \tag{27.73}$$

7.4.3. Rank Nullity Theorem

Theorem 27.2 Rank-Nullity theorem:
Let $\mathcal{V}$ be a finite vector space and let φ be a linear mapping $\varphi : \mathcal{V} \mapsto \mathcal{W}$ then it holds:

$$\dim(\mathcal{V}) = \dim\underbrace{\left(\varphi^{-1}(\{\mathbf{0}\})\right)}_{\text{Kernel}} + \dim\underbrace{\left(\varphi(\mathcal{V})\right)}_{\text{Image}} \tag{27.74}$$

Corollary 27.23 Representation as Standardbases:
For every linear mapping $\varphi : \mathbb{K}^n \mapsto \mathbb{K}^m$ there exists a matrix $\mathbf{A}$ that represents this mapping:

$$\varphi = \varphi_{\mathbf{A}} = (\varphi(\mathbf{e}_1) \qquad \varphi(\mathbf{e}_n)) \in \mathbb{K}^{m\times n} \tag{27.75}$$

where

8. Eigenvalues and Vectors

Definition 27.51 Eigenvalues: Given a square matrix $\mathbf{A} \in \mathbb{K}^{n,n}$ the eigenvalues

finish

Definition 27.52 Spectrum: The spectrum of a square matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ is the set of its eigenvalues[def. 27.51]:

$$\operatorname{spectrum}(\mathbf{A}) = \lambda(\mathbf{A}) = \{\lambda_1, \ldots, \lambda_n\} \tag{27.76}$$

Formula 27.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix $\mathbf{A}$ its eigenvalues can be calculated by:

$$\{\lambda_1, \lambda_2\} \in \frac{\operatorname{tr}(\mathbf{A}) \pm \sqrt{\operatorname{tr}(\mathbf{A})^2 - 4\det(\mathbf{A})}}{2} \tag{27.77}$$

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

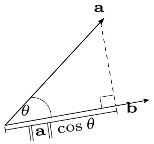
9. Vector Algebra

9.1. Dot/Standard Scalar Product

Definition 27.53 Scalar Projection a_b:

The scalar projection of a vector $\mathbf{a}$ onto a vector $\mathbf{b}$ is the *scalar* magnitude of the shadow/projection of the vector $\mathbf{a}$ onto $\mathbf{b}$:

$$a_b = \|\mathbf{a}\| \cos \theta_{a,b} = \mathbf{a}\tilde{\mathbf{b}} \tag{27.78}$$



Definition 27.54 [proof 27.4]
Standard Scalar/Dot Product:
Given two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ the standard scalar product is defined as:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^\intercal \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} u_i v_i = u_1 v_1 + \cdots + u_n v_n$$
$$= \|a\|\|b\| \cos \theta = u_v \hat{\mathbf{v}} = v_u \hat{\mathbf{u}} \qquad \theta \in [0, \pi] \tag{27.79}$$

Explanation 27.3 (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 27.12 Orthogonal Direction $\perp$:
For $\theta \in [-\pi, \pi/2]$ rad $\cos \theta = 0$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = 0 \iff \mathbf{u} \perp \mathbf{v} \tag{27.80}$$

Note: Perpendicular

Perpendicular corresponds to orthogonality of two lines.

Property 27.13 Maximizing Direction:
For $\theta = 0$ rad $\cos \theta = 1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \tag{27.81}$$

Property 27.14 Minimizing Direction:
For $\theta = \pi$rad $\cos \theta = -1$ and it follows:

$$\mathbf{u} \cdot \mathbf{v} = -\|\mathbf{u}\|\|\mathbf{v}\| \tag{27.82}$$

Definition 27.55 Vector Projecion:

General Projection via normal equation into inner product stuff i.e. with projection theorem

9.2. Cross Product

9.3. Outer Product

Definition 27.56 Outer Product $\mathbf{u}\mathbf{v}^\intercal = \mathbf{u} \otimes \mathbf{v}$:
Given two vectors $\mathbf{u} \in \mathbb{K}^m$, $\mathbf{v} \in \mathbb{K}^n$ their outer product is defined as:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^\mathsf{H} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \bar{\mathbf{v}}_1 \\ \vdots \\ \bar{\mathbf{v}}_n \end{bmatrix} \tag{27.83}$$
$$= \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 & \cdots \\ \vdots & \ddots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1\bar{\mathbf{v}}_1 & \mathbf{u}_1\bar{\mathbf{v}}_2 & \cdots & \mathbf{u}_1\bar{\mathbf{v}}_n \\ \mathbf{u}_2\bar{\mathbf{v}}_1 & \mathbf{u}_2\bar{\mathbf{v}}_2 & \cdots & \mathbf{u}_2\bar{\mathbf{v}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_m\bar{\mathbf{v}}_1 & \mathbf{u}_m\bar{\mathbf{v}}_2 & \cdots & \mathbf{u}_m\bar{\mathbf{v}}_n \end{bmatrix}$$

Proposition 27.2 [proof 27.5]
Rank of Outer Product: The outer product of two vectors is of rank one:

$$\mathrm{rank}(\mathbf{u} \otimes \mathbf{v}) = 1 \tag{27.84}$$

9.4. Vector Norms

Definition 27.57 Norm $\|\cdot\|_{\mathcal{V}}$:
Let $\mathcal{V}$ be a vector space over a field F, a norm on $\mathcal{V}$ is a map:

$$\|\cdot\|_{\mathcal{V}} : \mathcal{V} \mapsto \mathbb{R}_+ \tag{27.85}$$

that satisfies:

$$\|\mathbf{x}\|_{\mathcal{V}} = 0 \iff \mathbf{x} = 0 \qquad \text{(Definitness)} \tag{27.86}$$
$$\|\alpha\mathbf{x}\|_{\mathcal{V}} = |\alpha|\|\mathbf{x}\|_{\mathcal{V}} \qquad \text{(Homogenity)} \tag{27.87}$$
$$\|\mathbf{x} + \mathbf{y}\|_{\mathcal{V}} \leqslant \|\mathbf{x}\|_{\mathcal{V}} + \|\mathbf{y}\|_{\mathcal{V}} \qquad \text{(Triangular Inequality)} \tag{27.88}$$
$$\alpha \in \mathbb{K} \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{V}$$

Explanation 27.4 (Definition 27.57).
A norm is a measures of the size of its argument.

Corollary 27.24 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 27.58 [proof 27.21]
Cauchy Schwartz Inequality:

$$|\mathbf{u}^\intercal \mathbf{v}| \leqslant \|\mathbf{u}\|\|\mathbf{v}\| \tag{27.89}$$

9.4.2. Triangular Inequality

Definition 27.59 [proof 27.22]
Triangular Inequality: States that the length of the sum of two vectors is lower or equal to the sum of their individual lengths:

$$\|\mathbf{u} + \mathbf{v}\| \leqslant \|\mathbf{u}\| + \|\mathbf{v}\| \tag{27.90}$$

Corollary 27.25 Reverse Triangular Inequality:

$$-\|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}} \leqslant \|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}} \leqslant \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}}$$
$$\text{resp.} \qquad \big|\|\mathbf{x}\|_{\mathcal{V}} - \|\mathbf{y}\|_{\mathcal{V}}\big| \leqslant \|\mathbf{x} - \mathbf{y}\|_{\mathcal{V}}$$

9.5. Distances

Definition 27.60
Distance Function/Measure $d : S \times S \mapsto \mathbb{R}_+$:
Let S be a set, a distance functions is a mapping d that satisfies:

$$d(x, x) = 0 \qquad \text{(Zero Identity Distance)} \tag{27.91}$$
$$d(x, y) = d(y, x) \qquad \text{(Symmetry)} \tag{27.92}$$
$$d(x, z) \leqslant d(x, y) + d(y, z) \qquad \text{(Triangular Identiy)} \tag{27.93}$$
$$\forall x, y, z \in S$$

Explanation 27.5 (Definition 27.60).
Is measuring the distance between two things.

9.5.1. Contraction

Definition 27.61 Contraction: Given a metric space (M, d) is a mapping $f : M \mapsto M$ that satisfies:

$$d(f(x), f(y)) \leqslant \lambda d(x, y) \qquad \lambda \in [0, 1) \tag{27.94}$$

add metric spaces

9.6. Metrics

Definition 27.62 Metric $d : S \times S \mapsto \mathbb{R}_+$:
Is a distance measure[def. 27.60] that additionally satisfies the identity of indiscernibles:

$$d(x, y) = 0 \iff x = y \qquad \forall x, y \in S$$

Corollary 27.26 Metric→Norm: Every norm $\|\cdot\|_{\mathcal{V}}$ on a vector space $\mathcal{V}$ over a field F induces a metric by:

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

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

1. Homogenity/Scaling: $d(\alpha x, \alpha y)_{\mathcal{V}} = |\alpha| d(x, y)_{\mathcal{V}}$
2. Translational Invariance: $d(x + \alpha, y + \alpha) = d(x, y)$

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

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

Note

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

Note

(bilinear form $\xrightarrow{\text{induces}}$)
inner product $\xrightarrow{\text{induces}}$ norm $\xrightarrow{\text{induces}}$ metric.

9.6.1. Metric Space

Definition 27.63 Metric Space (M, d):
A *metric space* is a pair (M, d) of a set M and a metric[def. 27.62] d defined on M:

$$d : M \times M \mapsto \mathbb{R}_+ \tag{27.95}$$

10. Angles

Definition 27.64 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. 27.76] $\mathcal{V}$. The angle $\alpha \in [0, \pi]$ between $\mathbf{u}, \mathbf{v}$ is defined by:

$$\angle(\mathbf{u}, \mathbf{v}) := \alpha \qquad \cos \alpha = \frac{\mathbf{u}^\intercal \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \qquad \begin{matrix} \mathbf{u}, \mathbf{v} \in \mathcal{V} \\ \alpha \in [0, \pi] \end{matrix} \tag{27.96}$$

11. Orthogonality

Definition 27.65 Orthogonal Vectors: Let $\mathcal{V}$ be an inner-product space[def. 27.76]. A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \in \mathcal{V}$ is called *orthogonal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \qquad \forall i \neq \tag{27.97}$$

11.1. Orthonormality

Definition 27.66 Orthonormal Vectors: Let $\mathcal{V}$ be an inner-product space[def. 27.76]. A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n, \ldots\} \in \mathcal{V}$ is called *orthonormal* iff:

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \forall i, j \tag{27.98}$$

12. Special Kind of Vectors

12.1. Binary/Boolean Vectors

Definition 27.67
Binary/Boolean Vectors/Bit Maps $\mathbb{B}^n$: Are vectors that contain only zero or one values:

$$\mathbb{B}^n = \{0, 1\}^n \tag{27.99}$$

Definition 27.68
R-Sparse Boolean Vectors $\mathbb{B}^n_r$:
Are boolean vectors that contain exact r one values:

$$\mathbb{B}^n_r = \left\{ \mathbf{x} \in \{0, 1\}^n : \mathbf{x}^\intercal \mathbf{x} = \sum_{i=1}^{n} \mathbf{x} = r \right\} \tag{27.100}$$

12.2. Probablistic Vectors

Definition 27.69 Probabilistic Vectors: Are vectors that represent probabilities and satisfy:

$$\left\{ \mathbf{x} \in [0, 1]^n : \sum_{i=1}^{n} x_i = 1 \right\} \tag{27.101}$$

13. Vector Spaces and Measures

13.1. Bilinear Forms

13.2. Quadratic Forms

13.2.1. Min/Max Value

Corollary 27.27 [proof 27.20]
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 $\mathbf{A}$:

$$\mathbf{v}_1 \in \arg\min_{\mathbf{x}^\intercal \mathbf{x} = 1} \mathbf{x}^\intercal \mathbf{A} \mathbf{x} \qquad \mathbf{v}_1 \in \arg\max_{\mathbf{x}^\intercal \mathbf{x} = 1} \mathbf{x}^\intercal \mathbf{A} \mathbf{x} \tag{27.102}$$

Note

$$(\mathbf{Q}^\intercal \tilde{\mathbf{n}})^\intercal \mathbf{Q}^\intercal \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^\intercal \mathbf{Q}\mathbf{Q}^\intercal \tilde{\mathbf{n}} = \tilde{\mathbf{n}}^\intercal \tilde{\mathbf{n}} = 1$$

13.2.2. Skew Symmetric Matirx

Corollary 27.28
Quadratic Form of Skew Symmetric matrix: The quadratic form of a skew symmetric matrix[def. 27.27] vanishes:

$$\alpha = \mathbf{x}^\intercal \mathbf{A}_{\text{skew}} \mathbf{x} = \left(\mathbf{x}^\intercal \mathbf{A}^\intercal_{\text{skew}} \mathbf{x}\right)^\intercal = \left(\mathbf{x}^\intercal \mathbf{A}_{\text{skew}} \mathbf{x}\right)^\intercal = -\alpha \tag{27.103}$$

Which can only hold iff $\alpha = 0$.

13.3. Inner Product – Generalization of the dot product

Definition 27.70 Bilinear Form/Functional:
Is a mapping $a : \mathcal{V} \times \mathcal{V} \mapsto F$ on a field of scalars $F \subseteq \mathbb{K}$, $K = \mathbb{R}$ or $\mathbb{C}$ that satisfies:

$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$
$$\forall u, v, w \in \mathcal{V}, \quad \forall \alpha, \beta \in \mathbb{K}$$

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

Definition 27.71 Symmetric bilinear form: A bilinear form a on $\mathcal{V}$ is symmetric if and only if:

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

Definition 27.72 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:

$$a(u, u) > 0 \qquad \forall u \in \mathcal{V} \setminus \{0\} \tag{27.104}$$
$$\text{And positive semidefinte} \iff \geqslant \tag{27.105}$$

Corollary 27.29 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:

$$a(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\intercal \mathbf{A} \mathbf{x}' = (\mathbf{A}\mathbf{x}')\mathbf{x},$$

Definition 27.73 Positive (semi) definite Matrix $\succ$:
A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is positive definite if and only if:

$$\mathbf{x}^\intercal \mathbf{A} \mathbf{x} > 0 \iff \mathbf{A} \succ 0 \qquad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\} \tag{27.106}$$
$$\text{And positive semidefinte} \iff \succcurlyeq \tag{27.107}$$

Corollary 27.30 [proof 27.2]
Eigenvalues of positive (semi) definite matrix:
A positive definite matrix is a matrix where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive*.

$$\forall \lambda_i \in \text{eigenv}(\mathbf{A}) > 0 \tag{27.108}$$
$$\text{And positive semidefinte} \iff \geqslant \tag{27.109}$$

Note

Positive definite matrices are often assumed to be symmetric but that is not necessarily true.

Proof 27.2: ?? 27.2 (for real matrices):
Let $\mathbf{v}$ be an eigenvector of $\mathbf{A}$ then it follows:

$$0 \overset{??\ 27.2}{<} \mathbf{v}^\intercal \mathbf{A} \mathbf{v} = \mathbf{v}^\intercal \lambda \mathbf{v} = \|\mathbf{v}\| \lambda$$

Corollary 27.31 Positive Definiteness and Determinant: The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

Definition 27.74 Negative (semi) definite Matrix $\prec$:
A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is negative defintie if and only if:
$$\mathbf{x}^\intercal \mathbf{A}\mathbf{x} < 0 \iff \mathbf{A} \prec 0 \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\} \tag{27.110}$$
$$\text{And negative semidefinte} \iff \preccurlyeq \tag{27.111}$$

Theorem 27.3 Sylvester's criterion: Let $\mathbf{A}$ be *symmetric/Hermitian* matrix and denote by $\mathbf{A}^{(k)}$ the $k \times k$ upper left sub-matrix of $\mathbf{A}$.
Then it holds that:
- $$\mathbf{A} \succ 0 \iff \det\left(\mathbf{A}^k\right) > 0 \qquad k = 1, \dots, n \tag{27.112}$$
- $$\mathbf{A} \prec 0 \iff (-1)^k \det\left(\mathbf{A}^k\right) > 0 \qquad k = 1, \dots, n \tag{27.113}$$
- $\mathbf{A}$ is indefinite if the first $\det\left(\mathbf{A}^k\right)$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive ($\mathbf{A}$ can be anything of the previous three) if the first $\det\left(\mathbf{A}^k\right)$ that breaks both patterns is 0.

14. Inner Products

Definition 27.75 Inner Product: Let $\mathcal{V}$ be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on $\mathcal{V}$ is a map:
$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \mapsto F \subseteq \mathbb{K} \qquad K = \mathbb{R} \text{ or } \mathbb{C} \tag{27.114}$$
that satisfies: $\quad \forall x, y, z \in \mathcal{V}, \quad \alpha, \beta \in F$
1. (Conjugate) Stmmetry: $\langle x, y \rangle = \overline{\langle x, y \rangle}$.
2. Linearity in the first argument:
$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$$
3. Positve-definiteness:
$$\langle x, x \rangle \geqslant 0 : x = 0 \iff \langle x, x \rangle = 0$$

Definition 27.76 Inner Product Space $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$: Let $F \in \mathbb{K}$ be a field of scalars.
An inner product space $\mathcal{V}$ is a vetor space over a field F together with an **inner product** $\langle \cdot, \cdot \rangle_{\mathcal{V}}$).

Corollary 27.32 Inner product$\mapsto$S.p.d. Bilinear Form: **Let** $\mathcal{V}$ be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on $\mathcal{V}$ is a positive definite symmetric bilinear form on $\mathcal{V}$.

Example: scalar prodct

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

Note

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

Corollary 27.33 Inner product induced norm $\langle \cdot, \cdot \rangle_{\mathcal{V}} \to \|\cdot\|_{\mathcal{V}}$: Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ induces a norm of the form:
$$\|\mathbf{x}\|_{\mathcal{V}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \qquad \mathbf{x} \in \mathcal{V}$$
Thus We can define function spaces by their associated norm $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and inner product spaces lead to normed vector spaces and vice versa.

Corollary 27.34 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{V} \times \mathcal{V} \mapsto F$ induces an energy norm:
$$\|\mathbf{x}\|_a := (a(\mathbf{x}, \mathbf{x}))^{\frac{1}{2}} = \sqrt{a(\mathbf{x}, \mathbf{x})} \qquad \mathbf{x} \in \mathcal{V}$$

15. Matrix Algebra

16. Matrix Norms

16.1. Operator Norm

Definition 27.77 Operator/Induced Norm:
Let $\|\cdot\|_\mu : \mathbb{K}^m \mapsto \mathbb{R}$ and $\|\cdot\|_\nu : \mathbb{K}^n \mapsto \mathbb{R}$ be vector norms.
The operator norm is defined as:
$$\|\mathbf{A}\|_{\mu,\nu} := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_\mu}{\|\mathbf{x}\|_\nu} = \sup_{\|\mathbf{x}\|_\nu = 1} \|\mathbf{A}\mathbf{x}\|_\mu \qquad \|\cdot\|_\mu : \mathbb{K}^m \mapsto \mathbb{R} \tag{27.115}$$

Explanation 27.6 (Definition 27.77). *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 27.35 Induced Norms: Let $\|\cdot\|_p : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ defined as:
$$\|\mathbf{A}\|_p := \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{y}\|_p = 1} \|\mathbf{A}\mathbf{y}\|_p \tag{27.116}$$

Explanation 27.7 ([Corollary 27.35]).
Induced norms are matrix norms induced by vector norms as we:
- *Only work with vectors* $\mathbf{A}\mathbf{x}$
- *And use the normal p-vector norms* $\|\cdot\|_p$

Note supremum

The set of vectors $\{\mathbf{y} | \|\mathbf{y}\| = 1\}$ is compact, thus if we consider finite matrices the supremum is attained and we may replace it by the max.

16.3. Induced Norms

16.3.1. 1-Norm

Definition 27.78 Column Sum Norm $\quad \|\mathbf{A}\|_1$:
$$\|\mathbf{A}\|_1 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^m |a_{ij}| \tag{27.117}$$

16.3.2. ∞-Norm

Definition 27.79 Row Sum Norm $\quad \|\mathbf{A}\|_\infty$:
$$\|\mathbf{A}\|_\infty = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \mathbf{x} \neq 0}} \frac{\|\mathbf{A}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^n |a_{ij}| \tag{27.118}$$

16.3.3. Spectral Norm L2-Norm

Spectral Radius & Singular Value

Definition 27.80 Spectral Radius $\quad \rho(\mathbf{A})$:
The spectral radius is defined as the largest eigenvalue of a matrix:
$$\rho(\mathbf{A}) = \max\{\lambda | \lambda \in \text{eigenval}(\mathbf{A})\} \tag{27.119}$$

Definition 27.81 Singular Value $\quad \sigma_i$:
Given a matrix $\mathbf{A} \in \mathbb{K}^{m \times n}$ its n real and positive singular values are defined as:
$$\sigma(\mathbf{A}) := \left\{ \left\{\sqrt{\lambda_i}\right\}_{i=1}^n | \lambda_i \in \text{eigenval}\left(\mathbf{A}^\intercal \mathbf{A}\right) \right\} \tag{27.120}$$

Spectral Norm

Definition 27.82 L2/Spectral Norm $\quad \|\mathbf{A}\|_2$:
$$\|\mathbf{A}\|_2 = \sup_{\substack{\mathbf{x} \in \mathbb{K}^n \\ \|\mathbf{x}\|_2 = 1}} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \sqrt{\mathbf{x}^\intercal \mathbf{A}^\intercal \mathbf{A}\mathbf{x}} \tag{27.121}$$
$$= \max_{\|\mathbf{x}\|_2 = 1} \sqrt{\rho\left(\mathbf{A}^\intercal \mathbf{A}\right)} =: \sigma_{\max}(\mathbf{A}) \tag{27.122}$$

16.4. Energy Norm

16.5. Forbenius Norm

Definition 27.83 Forbenius Norm $\quad \|\mathbf{A}\|_F$:
The *Forbenius norm* $\|\cdot\|_F : \mathbb{K}^{m \times n} \mapsto \mathbb{R}$ is defined as:
$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}^2|} = \sqrt{\operatorname{tr}\left(\mathbf{A}\mathbf{A}^\mathsf{H}\right)} \tag{27.123}$$

16.6. Distance

17. Decompositions

17.1. Eigen/Spectral decomposition

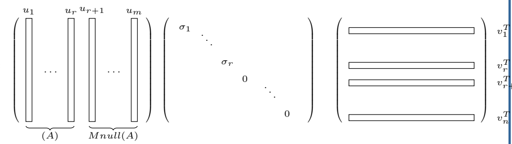
Definition 27.84 $\quad \mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}$, [proof 27.25]
Eigendecomposition/ Spectral Decomposition :
Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a *diagonalizable* square matrix and define by $\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_n] \in \mathbb{R}^{n \times n}$ a non-singular matrix whose columm vectors are the eigenvectors of $\mathbf{A}$ with associated eigenvalue matrix $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Then $\mathbf{A}$ can be represented as:
$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \tag{27.124}$$

Proposition 27.3 Diagonalization: If non of $\mathbf{A}$ eigenvalues are zero it can be diagonalized:
$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \boldsymbol{\Lambda} \tag{27.125}$$

Proposition 27.4 Existence:
$$\exists \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \iff \mathbf{A} \text{ diagonalizable} \tag{27.126}$$

17.2. QR-Decompositions

17.3. Singular Value Decomposition

Definition 27.85
Singular Value Decomposition (SVD) $\quad \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{H}$:
For any matrix $\mathbf{A} \in \mathbb{K}^{m,n}$ there exist unitary matrices[def. 27.25]
$$\mathbf{U} \in \mathbb{K}^{m,m} \qquad \mathbf{V} \in \mathbb{K}^{n,n}$$
and a (generalized) digonal matrix:
$$\boldsymbol{\Sigma} \in \mathbb{R}^{m,n} \qquad p := \min\{m, n\}$$
$$\boldsymbol{\Sigma} = \operatorname{gendiag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m,n}$$
such that:
$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathsf{H} \tag{27.127}$$
=


Images full, economical, alternative representation, range and kernel
https://math.mit.edu/classes/18.095/2016IAP/lec2/SVD_Notes.pdf

17.3.1. Eigenvalues

Proposition 27.5 [proof 27.23]:
The eigenvalues of a matrix $\mathbf{A}^\intercal \mathbf{A}$ are positive.

Proposition 27.6 [proof 27.24]
Similarity Transformation: The unitary matrix $\mathbf{V}$ provides a *similarity transformation*[cor. 27.14] of $\mathbf{A}^\intercal \mathbf{A}$ into a diagonal matrix $\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma}$:
$$\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma} \mapsto \mathbf{V}^\mathsf{H} \mathbf{A}^\intercal \mathbf{A} \mathbf{V} \tag{27.128}$$

Corollary 27.36 eigenval$(\mathbf{A}^\intercal \mathbf{A})$ = eigenval$(\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma})$:
From proposition 27.6 and [cor. 27.15] it follows that:
$$\text{eigenval}(\mathbf{A}^\intercal \mathbf{A}) = \text{eigenval}(\boldsymbol{\Sigma}^\intercal \boldsymbol{\Sigma}) \tag{27.129}$$
$$\implies \|\mathbf{A}\|_2 = \sqrt{\rho\left(\mathbf{A}^\intercal \mathbf{A}\right)} = \sqrt{\lambda_{\max}} = \sigma_{\max}$$

Note

λ and *singularvalue* corresponds to the eigenvalues/singularvalues of $\mathbf{A}^\intercal \mathbf{A}$ and not $\mathbf{A}$

17.3.2. Best Lower Rank Approximation

Theorem 27.4 Eckart Yound Theorem: Given a matrix $\mathbf{X} \in \mathbb{K}^{m,n}$ the *reduced* SVD $\mathbf{X}$ defined as:
$$\mathbf{U}_k := [\mathbf{u}_{:,1} \quad \mathbf{u}_{:,k}] \in \mathbb{K}^{m,k}$$
$$\mathbf{X}_k := \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\mathsf{H} \qquad \boldsymbol{\Sigma}_k = \operatorname{diag}(\sigma_1, \dots, \sigma_k) \in \mathbb{R}^{k,k}, \quad k \leqslant \min\{m, n\}$$
$$\mathbf{V}_k = [\mathbf{v}_{:,1} \quad \mathbf{v}_{:,k}] \in \mathbb{K}^{n,k}$$
provides the best lower k rank approximation of $\mathbf{X}$:
$$\min_{\mathbf{Y} \in \mathbb{K}^{n,m} : \operatorname{rank}(\mathbf{Y}) \leqslant k} \|\mathbf{X} - \mathbf{Y}\|_F = \|\mathbf{X} - \mathbf{X}_k\|_F \tag{27.130}$$

18. Matric Calculus

18.1. Derivatives

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{b}) = \mathbf{b}$$
$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$$
$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A} \tag{27.131}$$
$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\intercal \mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^\intercal)\mathbf{x} \tag{27.132}$$
$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top \mathbf{A}\mathbf{x}) = \mathbf{A}^\top \mathbf{b} \qquad \frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top \mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top \qquad \frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|_2}$$
$$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x} \qquad \frac{\partial}{\partial \mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$$
$$\frac{\partial}{\partial \mathbf{x}}\|\mathbf{x}\|_1 = \frac{\mathbf{x}}{|\mathbf{x}|}$$
$$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2) = \mathbf{2}(\mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b}) \qquad \frac{\partial}{\partial \mathbf{X}}(|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1}$$
$$\frac{\partial}{\partial x}(\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1} \frac{\partial \mathbf{Y}}{\partial x} \mathbf{Y}^{-1}$$

19. Proofs

Proof 27.3: [def. 27.34]
$$\mathbf{M}\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{I}_{k,k} & \mathbf{0}_{k,l} \\ \mathbf{0}_{l,k} & \mathbf{I}_{l,l} \end{bmatrix} \tag{27.133}$$

19.1. Vector Algebra

Proof 27.4 Definition 27.54:
$$\textbf{(1)}: \underline{\|a - b\|} \overset{\text{eq. (28.19)}}{=} \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \cos\theta$$
$$\textbf{(2)}: \underline{\|a - b\|} = (a - b)(a - b) = \|a\|^2 + \|b\|^2 - 2(ab)$$
$$\underline{\|a - b\|} = \underline{\|a - b\|} \quad \Rightarrow \quad ab = \|a\|\|b\| \cos\theta$$

Proof 27.5 Proposition 27.2: The outer product of $\mathbf{u}$ with $\mathbf{v}$ corresponds to a scalar multiplication of $\mathbf{v}$ with elements u_i thus the rank must be that of $\mathbf{v}$, which is a vector and hence of rank 1
$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^\mathsf{H} = \begin{bmatrix} \mathbf{u}_1 \odot \bar{\mathbf{v}}_1 \\ \vdots \\ \mathbf{u}_m \odot \bar{\mathbf{v}}_n \end{bmatrix}$$

19.2. Mappings

Proof 27.6: Corollary 27.20
$$\|\mathbf{Q}\mathbf{x}\|^2 = (\mathbf{Q}\mathbf{x})^\intercal \mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal \mathbf{Q}^\intercal \mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal \mathbf{x} = \|\mathbf{x}\|^2$$

Proof 27.7: Corollary 27.21 Follows immediately from definition 27.64 in combination with eqs. (27.63) and (27.65).

Proof 27.8: Proposition 27.1:
$$\implies \quad l(\alpha\mathbf{x} + \beta\mathbf{y}) \overset{27.56}{=} l(\alpha\mathbf{x}) + l(\beta\mathbf{y}) \overset{27.57}{=} \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$$
$$\impliedby \quad l(\alpha\mathbf{x} + \mathbf{0}) = \alpha l(\mathbf{x}) \qquad l(1\mathbf{x} + 1\mathbf{y}) = l(\mathbf{x}) + l(\mathbf{y})$$

Proof 27.9 principle 27.1:
Every vector $\mathbf{v} \in \mathcal{V}$ can be represented by a basis eq. (27.14) of $\mathcal{V}$. With *homogentiyt*eq. (27.57) and *additivity*eq. (27.56) it follows for the image of all $\mathbf{v} \in \mathcal{V}$:
$$l(\mathbf{v}) = l(\alpha_1 b_1 + \dots + \alpha_n b_n) = l\alpha_1(b_1) + \dots + l(\alpha_n) b_n \tag{27.134}$$
$\Rightarrow$ the image of the basis of $\mathcal{V}$ determines the linear mapping.

Proof 27.10 Proof [Corollary 27.17]:
$$\implies \quad l_{\mathbf{A}}(\alpha\mathbf{x} + \mathbf{y}) = \mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{A}\mathbf{x} + \beta\mathbf{A}\mathbf{y} = \alpha l(\mathbf{x}) + \beta l(\mathbf{y})$$
$\impliedby$ Let $\mathfrak{B}$ be a standard normal basis of $\mathcal{V}$ with eq. (27.134):
$$l(\mathbf{x}) = \sum_{i=1}^n x_i l(\mathbf{e}_i) = \sum_{i=1}^n x_i \mathbf{A}_{:,i} = \mathbf{A}\mathbf{x} \qquad \mathbf{A}_{:,i} := \mathbf{l}(\mathbf{e_i}) \in \mathbb{R}^n$$

Proof 27.11 Proof Property 27.11:
$$(g \circ f)(\alpha\mathbf{x}) = g(f(\alpha\mathbf{x})) = g(\alpha f(\mathbf{x})) = \alpha(g \circ f)(\mathbf{x})$$
$$(g \circ f)(\mathbf{x} + \mathbf{y}) = g(f(\mathbf{x} + \mathbf{y})) = g(f(\mathbf{x}) + f(\mathbf{y})) = (g \circ f)(\mathbf{x}) + (g \circ f)(\mathbf{y})$$
or even simpler as every linear form can be represented by a matrix product:
$$f(y) = \mathbf{A}\mathbf{y} \quad g(z) = \mathbf{B}\mathbf{z} \quad \Rightarrow \quad (f \circ g)(\mathbf{x}) = \mathbf{A}\mathbf{B}\mathbf{x} := \mathbf{C}\mathbf{x}$$

Proof 27.12: [Corollary 27.22] Let $\mathbf{y} \in \mathbb{N}(\mathbf{A})$ ($\mathbf{z} \in \mathbb{N}(\mathbf{A}^\intercal)$) then it follows:

$$\mathbb{N}(\mathbf{A}) \perp \mathfrak{R}(\mathbf{A}^\intercal) \qquad (\mathbf{A}^\intercal\mathbf{x})^\intercal \mathbf{y} = \mathbf{x}^\intercal\mathbf{A}\mathbf{y} = \mathbf{x}^\intercal\mathbf{0} = 0$$
$$\mathbb{N}(\mathbf{A}^\intercal) \perp \mathfrak{R}(\mathbf{A}) \qquad (\mathbf{A}\mathbf{x})^\intercal \mathbf{z} = \mathbf{x}^\intercal\mathbf{A}^\intercal\mathbf{z} = \mathbf{x}^\intercal\mathbf{0} = 0$$

19.3. Special Matrices

Proof 27.13 [Corollary 27.15]: 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 27.14 Property 27.6:

$$\|\mathbf{Q}\mathbf{x}\|_2^2 = (\mathbf{Q}\mathbf{x})^\intercal\mathbf{Q}\mathbf{x} = \mathbf{x}^\intercal\mathbf{Q}^\intercal\mathbf{Q}\mathbf{x} = \|\mathbf{x}\|_2^2$$

Proof 27.15: Property 27.4
Let $\mathbf{A} \in \mathbb{K}^{n\times n}$ be a hermitian matrix[def. 27.24] and let $\lambda \in \mathbb{K}$ be an eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\mathbf{v} \in \mathbb{K}^n$:

$$\lambda(\bar{\mathbf{v}}^\intercal\mathbf{v}) = \bar{\mathbf{v}}^\intercal\lambda\mathbf{v} = \bar{\mathbf{v}}^\intercal\mathbf{A}\mathbf{v} = \overline{(\bar{\mathbf{v}}^\intercal\mathbf{A}\mathbf{v})} = \bar{\mathbf{A}}\mathbf{v}^\intercal\mathbf{v} = \bar{\lambda}\,(\bar{\mathbf{v}}^\intercal\mathbf{v})$$
$$\lambda\,(\bar{\mathbf{v}}^\intercal\mathbf{v}) = \bar{\lambda}\,(\bar{\mathbf{v}}^\intercal\mathbf{v})$$

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 (20.8))

Proof 27.16: ??

add

19.4. Vector Spaces

Proof 27.17 Definition 27.19: We know that $\text{proj}_L(\mathbf{u})$ must be a vector times a certain magnitude:

$$\text{proj}_L(\mathbf{u}) = \alpha\tilde{\mathbf{v}} \qquad \alpha \in \mathbb{K} \tag{27.135}$$

the magnitude follows from the scalar projection[def. 27.53] in the direction of $\mathbf{v}$ which concludes the derivation.

Proof 27.18 Definition 27.19 (via orthogonality): We know that $\mathbf{u} - \text{proj}_L(\mathbf{u})$ must be orthogonal[def. 27.65] to $\mathbf{v}$

$$(\mathbf{u} - \text{proj}_L(\mathbf{u}))^\intercal\,\mathbf{v} = (\mathbf{u} - \alpha\mathbf{v})^\intercal\,\mathbf{v} = 0 \Rightarrow \qquad \alpha = \frac{\mathbf{u}^\intercal\mathbf{v}}{\mathbf{v}^\intercal\mathbf{v}}$$

Proof 27.19: Definition 27.20 Let $\mathfrak{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ a basis of $\mathcal{U}$ s.t. by [cor. 27.4]:

$$\mathbf{u} = \sum_{i=1}^n \alpha_i\mathbf{b}_i$$

the coefficients $\{\alpha_i\}_{i=1}^n$ need to be determined. We know that:

$$\mathbf{v} - \mathbf{u} \perp \mathbf{b}_1, \dots, \mathbf{v} - \mathbf{u} \perp \mathbf{b}_n$$
$$\Longrightarrow \qquad \left(\mathbf{v} - \sum_{i=1}^n \alpha_i\mathbf{b}_i\right)\cdot\mathbf{b}_j = 0 \qquad j = 1, \dots, n$$

this linear system of equations can be rewritten as:

$$\begin{pmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_n\end{pmatrix}\begin{pmatrix}\mathbf{b}_1\\ \vdots \\ \mathbf{b}_n\end{pmatrix}\begin{pmatrix}\alpha_1\\ \vdots \\ \alpha_n\end{pmatrix} = \begin{pmatrix}\mathbf{b}_1\\ \vdots \\ \mathbf{b}_n\end{pmatrix}\mathbf{v}$$

Proof 27.20: Corollary 27.27
Let $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\intercal$ be the eigendecomposition[cor. 27.12] of $\mathbf{A}$ then it follows:

$$\begin{aligned}
\min_{\tilde{\mathbf{n}}^\intercal\tilde{\mathbf{n}}=1} \tilde{\mathbf{n}}^\intercal\mathbf{A}\tilde{\mathbf{n}} &= \min_{\|\tilde{\mathbf{n}}\|=1} \tilde{\mathbf{n}}^\intercal(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\intercal)\tilde{\mathbf{n}} \\
&= \min_{\|\tilde{\mathbf{n}}\|=1} (\mathbf{Q}^\intercal\tilde{\mathbf{n}})^\intercal\mathbf{\Lambda}(\mathbf{Q}^T\tilde{\mathbf{n}}) \\
&= \min_{\mathbf{x}=1} \mathbf{x}^\intercal\mathbf{\Lambda}\mathbf{x} \qquad \mathbf{x} := \mathbf{Q}^\intercal\tilde{\mathbf{n}} \\
&= \min_{\mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2\Lambda_{ii} = \min_{\mathbf{x}=1} \sum_{i=1}^n \mathbf{x}_i^2\lambda_i
\end{aligned}$$

Thus in order to obtain the minimum value we need to choose the eigenvector that leads to the smallest eigenvalue.

19.5. Norms

Proof 27.21: ?? 27.21

$$|\mathbf{u}\cdot\mathbf{v}| \overset{\text{eq. (27.79)}}{=} \|\mathbf{u}\|\|\mathbf{v}\||\cos\theta| \leqslant \|\mathbf{u}\|\|\mathbf{v}\|$$

Proof 27.22: Definition 27.59

$$\|\mathbf{u}+\mathbf{v}\|_2^2 = (\mathbf{u}+\mathbf{v})(\mathbf{u}+\mathbf{v}) = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\mathbf{u}\cdot\mathbf{v})$$

from cauchy schwartz we know:

$$\mathbf{u}\cdot\mathbf{v} \leqslant |\mathbf{u}\cdot\mathbf{v}| \overset{\text{eq. (27.89)}}{\leqslant} \|\mathbf{u}\|\|\mathbf{v}\|$$
$$\|\mathbf{u}+\mathbf{v}\|^2 \leqslant \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2(\|\mathbf{u}\|\|\mathbf{v}\|) = (\|\mathbf{u}\| + \|\mathbf{v}\|)^2$$

19.6. Decompositions

19.6.1. Symmetric - Antisemitic

Definition 27.86 Symmetric - Antisymmetric Decomposition: Any matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ can be decomposed into the sum of a *symmetric matrix*[def. 27.21] $\mathbf{A}^{\text{sym}}$ and a *skew-symmetric matrix*?? $\mathbf{A}^{\text{skes}}$:

$$\mathbf{A} = \mathbf{A}^{\text{sym}} + \mathbf{A}^{\text{skew}} \qquad \begin{aligned}\mathbf{A}^{\text{sym}} &= \frac{1}{2}\left(\mathbf{A} + \mathbf{A}^{\mathsf{H}}\right) \\ \mathbf{A}^{\text{skew}} &= \frac{1}{2}\left(\mathbf{A} - \mathbf{A}^{\mathsf{H}}\right)\end{aligned} \tag{27.136}$$

19.6.2. SVD

Proof 27.23 [Corollary 27.5]: $\mathbf{B} := \mathbf{A}^\intercal\mathbf{A}$ corresponds to a *symmetric positive definite* form[def. 27.73]:

$$\mathbf{x}^\intercal\mathbf{B}\mathbf{x} = \mathbf{x}^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 > 0$$

thusProposition 27.6 follows immediately form [Corollary 27.2].

Proof 27.24 Proposition 27.6:

$$\mathbf{A}^\intercal\mathbf{A} \overset{\text{SVD}}{=} \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}\right)^{\mathsf{H}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}} = \mathbf{V}\mathbf{\Sigma}^{\mathsf{H}}\underbrace{\mathbf{U}^{\mathsf{H}}\mathbf{U}}_{\mathbf{I}_m}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}} = \mathbf{V}\mathbf{\Sigma}^{\mathsf{H}}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$$
$$\Longrightarrow \qquad \mathbf{V}^{\mathsf{H}}\mathbf{A}^\intercal\mathbf{A}\mathbf{V} = \mathbf{\Sigma}^\intercal\mathbf{\Sigma}$$

19.6.3. Eigendecomposition

Proof 27.25 Definition 27.84:

$$\mathbf{A}\mathbf{X} = [\lambda_1\mathbf{x}_1 \quad \cdots \quad \lambda_n\mathbf{x}_n] = \mathbf{X}\mathbf{\Lambda}$$

Geometry

Corollary 28.1 Affine Transformation in 1D: Given: numbers $x \in \hat{\Omega}$ with $\hat{\Omega} = [a, b]$
The affine transformation of $\phi : \hat{\Omega} \to \Omega$ with $y \in \Omega = [c, d]$ is defined by:

$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{28.1}$$

Proof 28.1: [cor. 28.1] By [def. 27.43] we want a function $f : [a, b] \to [c, d]$ that satisfies:

$$f(a) = c \qquad \textbf{and} \qquad f(b) = d$$

additionally $f(x)$ has to be a linear function ([def. 24.18]), that is the output scales the same way as the input scales.
Thus it follows:

$$\frac{d-c}{b-a} = \frac{f(x) - f(a)}{x-a} \qquad \iff \qquad f(x) = \frac{d-c}{b-a}(x-a) + c$$

Trigonometry

0.1. Trigonometric Functions

0.1.1. Sine

Definition 28.1 Sine:

$$\sin\alpha = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{a}{c} \tag{28.2}$$

0.1.2. Cosine

Definition 28.2 Cosine:

$$\cos\alpha\alpha = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{b}{c} \tag{28.3}$$

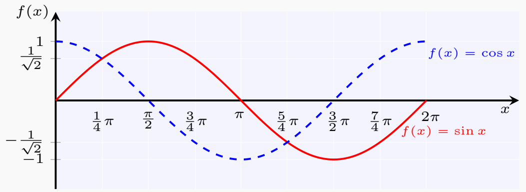
0.1.3. Tangens

Definition 28.3 Tangens:

$$\cos\alpha\alpha = \frac{\text{opposite}}{\text{adjacent}} = \frac{a}{b} = \frac{a/c}{b/c} = \frac{\sin\alpha}{\cos\alpha} \tag{28.4}$$

0.1.4. Trigonometric Functions and the Unit Circle

Sine and Cosine



$$\cos x \overset{(24.51)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{28.5}$$

$$\sin x \overset{(24.51)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{28.6}$$

Note

Using theorem 28.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{28.7}$$

0.1.5. Sinh

Definition 28.4 Sinh:

$$\sinh x \overset{(eq.\ (24.51))}{=} \frac{1}{2}\left[e^{x} - e^{-x}\right] = -\mathrm{i}\sin(\mathrm{i}\,x) \tag{28.8}$$

Property 28.1: $\sinh x = 0$ has a unique root at $x = 0$.

0.1.6. Cosh

Definition 28.5 Cosh:

$$\cosh x \overset{(24.51)}{=} \frac{1}{2}\left[e^{x} + e^{-x}\right] = \cos(\mathrm{i}\,x) \tag{28.9}$$

(28.10)

Property 28.2: $\cosh x$ is strictly positive.

Proof 28.2:

$$e^{x} = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{28.11}$$

0.2. Addition Theorems

Theorem 28.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{28.12}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{28.13}$$

0.3. Werner Formulas

Werner Formulas

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right] \tag{28.14}$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \tag{28.15}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right] \tag{28.16}$$

Note

Using theorem 28.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{28.17}$$

0.4. Law of Cosines

Law 28.1 Law of Cosines [proof 28.3]**:** relates the three side of a *general* triangle to each other.

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{b,c} \tag{28.18}$$

Law 28.2 Law of Cosines for Vectors [proof 28.4]**:** relates the length of vectors to each other.

$$\|\mathbf{a}\|^2 = \|\mathbf{c} - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2 - 2\|\mathbf{b}\|\|\mathbf{c}\|\cos\theta_{\mathbf{b},\mathbf{c}} \tag{28.19}$$

Law 28.3 Pythagorean theorem: special case of **??** for right triangle:

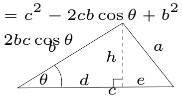
$$a^2 = b^2 + c^2 \tag{28.20}$$

1. Proofs

Proof 28.3: Law 28.1 From the defintion of the sine and cosine we know that:

$$\sin\theta = \frac{h}{b} \Rightarrow \underline{h} \qquad \text{and} \qquad \cos\theta = \frac{d}{b} \Rightarrow \underline{d}$$

$$\begin{aligned} \underline{e} &= c - \underline{d} = c - b\cos\theta \\ a^2 &= \underline{e}^2 + \underline{h}^2 = c^2 - 2cb\cos\theta + b^2\cos^2\theta + b^2\sin^2\theta \\ &= c^2 + b^2 - 2bc\cos\theta \end{aligned}$$



Proof 28.4: Law 28.2 Notice that $\mathbf{c} = \mathbf{a} + \mathbf{b} \Rightarrow \mathbf{a} = \mathbf{c} - \mathbf{b}$ and we can either use **??** 28.3 or notice that:

$$\begin{aligned} \|\mathbf{c} - \mathbf{b}\|^2 &= (\mathbf{c} - \mathbf{b}) \cdot (\mathbf{c} - \mathbf{b}) \\ &= \mathbf{c} \cdot \mathbf{c} - 2\mathbf{c} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} \\ &= \|\mathbf{c}\|^2 + \|\mathbf{b}\|^2 - 2\left(\|\mathbf{c}\|\|\mathbf{b}\|\cos\theta\right) \end{aligned}$$

Topology

Definition 29.1 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:

- $\varnothing \in \tau$ and $X \in \tau$
- Any finite or infinite union of subsets of τ is contained in τ:
$$\{U_i : i \in \mathbf{I}\} \subseteq \tau \qquad \Longrightarrow \qquad \cup_{i\in\mathbf{I}} U_i \in \tau \tag{29.1}$$
- The intersection of a finite number of elements of τ also belongs to τ:
$$\{U_i\}_{i=1}^{n} \in \tau \qquad \Longrightarrow \qquad U_1 \cap \cdots \cap U_n \in \tau \tag{29.2}$$

Definition 29.2 Topological Space[?] (X, τ)**:**
Is an ordered pair (X, τ), where X is a set and τ is a topology[def. 29.1] on X.

Numerical Methods

1. Machine Arithmetic's

1.1. Machine Numbers

Definition 30.1 Institute of Electrical and Electronics Engineers (IEEE): Is a engineering associations that defines a standard on how computers should treat machine numbers in order to have certain guarantees.

Definition 30.2 Machine/Floating Point Numbers $\mathbb{F}$:
Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{F} \subset \mathbb{R}$

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

Corollary 30.1 Closure: Machine numbers $\mathbb{F}$ are not *closed*[def. 20.7] under basic arithmetic operations:

$$\mathbb{F}\,\Omega\,\mathbb{F} \mapsto \not{\mathbb{F}} \qquad \Omega = \{+,-,*,/\} \tag{30.1}$$

Note

Corollary 30.1 provides a problem as the computer can only represent floating point number $\mathbb{F}$.

Definition 30.3 Floating Point Operation $\tilde{\Omega}$:
Is a basic arithmetic operation that obtains a number $x \in \mathbb{F}$ by applying a function rd:

$$\mathbb{F}\,\tilde{\Omega}\,\mathbb{F} \mapsto \mathbb{F} \qquad \begin{aligned}\tilde{\Omega} &:= \mathrm{rd} \circ \Omega \\ \Omega &= \{+,-,*,/\}\end{aligned} \tag{30.2}$$

Definition 30.4 Rounding Function rd:
Given a real number $x \in \mathbb{R}$ the rounding function replaces it by the nearest machine number $\tilde{x} \in \mathbb{F}$. If this is ambiguous (there are two possibilities), then it takes the larger one:

$$\mathrm{rd}: \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg \min_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases} \tag{30.3}$$

Consequence

Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

Axiom 30.1 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:

$$\begin{aligned} x\tilde{\Omega}y &= (x\Omega y)(1+\delta) \qquad \Omega = \{+,-,*,/\} \\ \tilde{f}(x) &= f(x)(1+\delta) \qquad f \in \{\exp, \sin, \cos, \log, \ldots\} \end{aligned} \tag{30.4}$$

with $|\delta| < \mathrm{EPS}$

Explanation 30.1 (axiom 30.1). *gives us a guarantee that for any two floating point numbers $x, y \in \mathbb{F}$, any operation involving them will give a floating point result which is within a factor of $1 + \delta$ of the true result $x\Omega y$.*

Finish

Definition 30.5 Overflow: Result is bigger then the biggest representable floating point number.

Definition 30.6 Underflow: Result is smaller then the smaller representable floating point number i.e. to close to zero.

1.2. Roundoff Errors

Log-Sum-Exp Trick

The sum exponential trick is at trick that helps to calculate the log-sum-exponential in a robust way by avoiding over/underflow. The log-sum-exponential[def. 30.7] is an expression that arises frequently in machine learning i.e. for the cross entropy loss or for calculating the evidence of a posterior prediction.
The root of the problem is that we need to calculate the exponential $\exp(x)$, this comes with two different problems:

- If x is large (i.e. 89 for single precision floats) then $\exp(x)$ will lead to overflow
- If x is very negative $\exp(x)$ will lead to underflow/0. This is not necessarily a problem but if $\exp(x)$ occurs in the denominator or the logarithm for example this is catastrophic.

Definition 30.7 Log sum Exponential:

$$\mathrm{LogSumExp}(x_1, \ldots, x_n) := \log\left(\sum_{i=1}^n e^{x_i}\right) \tag{30.5}$$

Formula 30.1 Log-Sum-Exp Trick:

$$\log\left(\sum_{i=1}^n e^{x_i}\right) = a + \log \sum_{i=1}^n e^{x_i - a} \qquad a := \max_{i \in \{1,\ldots,n\}} x_i \tag{30.6}$$

Explanation 30.2 (formula 30.1). *The value a can be any real value but for robustness one usually chooses the max s.t.*

- *The leading digits are preserved by pulling out the maximum a*
- *Inside the log only zero or negative numbers are exponentiated, so there can be no overflow.*
- *If there is underflow inside the log we know that at least the leading digits have been returned by the max.*

Proof 30.1:

$$\begin{aligned} \mathrm{LSE} &= \log\Big(\sum_{i=1}^n e^{x_i}\Big) = \log\Big(\sum_{i=1}^n e^{x_i - a} e^{a}\Big) \\ &= \log\Big(e^{a} \sum_{i=1}^n e^{x_i - a}\Big) = \log\Big(\sum_{i=1}^n e^{x_i - a}\Big) + \log(e^{a}) \\ &= \log\Big(\sum_{i=1}^n e^{x_i - a}\Big) + a \end{aligned}$$

Definition 30.8 Partition Π:
Given an interval $[0, T]$ a sequence of values $0 < t_0 < \cdots < t_n < T$ is called a partition $\Pi(t_0, \ldots, t_n)$ of this interval.

2. Convergence

2.1. O-Notation

2.1.1. Small $o(\cdot)$ Notation

Definition 30.9 Little o Notation:

$$f(n) = o(g(n)) \iff \lim_{n \to \infty} \frac{f(n)}{g(n)} = 0 \tag{30.7}$$

2.1.2. Big $\mathcal{O}(\cdot)$ Notation

2.2. Rate Of Convergence

Definition 30.10 Rate of Convergence: Is a way to measure the rate of convergence of a sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ to a value to $\mathbf{x}^*$. Let $\rho \in [0, 1]$ be the *rate of convergence* and define:

$$\lim_{k \to \infty} \frac{\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{k} - \mathbf{x}^*\right\|} = \rho \tag{30.8}$$

$$\iff \lim_{k \to \infty} \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \rho \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \qquad \forall k \in \mathbb{N}_0$$

Definition 30.11 Linear/Exponential Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *linearly* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho \in (0, 1) \qquad \forall k \in \mathbb{N}_0 \tag{30.9}$$

Definition 30.12 Superlinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *superlinear* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 1 \tag{30.10}$$

Definition 30.13 Sublinear Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *sublinear* to $\mathbf{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 0 \iff \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| = o\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|\right) \tag{30.11}$$

Definition 30.14 Logarithmic Convergence:
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *logarithmically* to $\mathbf{x}^*$ if it converges *sublinear*[def. 30.13] and additoinally satisfies

$$\rho = 0 \iff \left\|\mathbf{x}^{k+2} - \mathbf{x}^{k+1}\right\| = o\left(\left\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\right\|\right) \tag{30.12}$$

add explaintion why

Exponetial Convergence

Linear convergence is sometimes called exponential convergence. This is due to the fact that:

1. We often have expressions of the form:

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \underbrace{(1-\alpha)}_{:=\rho}\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$$

2. and that $(1-\alpha) = \exp(-\alpha)$ from which follows that:

$$\text{eq. (30.13)} \iff \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant e^{-\alpha}\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|$$

Definition 30.15 Convergence of order p: In order to distinguish *superlinear convergence* we define the order of convergence.
A sequence $\{\mathbf{x}^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \ldots\}$ to $\mathbf{x}^*$ if it satisfies:

$$\lim_{k \mapsto \infty} \frac{\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|}{\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^p} = C \qquad C < 1 \tag{30.13}$$

Does this even exist/check if this is ture

Definition 30.16 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 $k \to \infty$ it satisfies:

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leqslant \rho^k \left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\| \qquad \rho < 1 \tag{30.14}$$

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \in o\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|\right) \tag{30.15}$$

3. Linear Systems of Equations

3.1. Direct Methods

3.1.1. LU-Decomposition

Definition 30.17 LU Decomposition:

3.1.2. Symmetric Positive Definite Matrices

For linear systems with s.p.d.[def. 27.73] matrices $\mathbf{A}$ the LU-decomposition[def. 30.17] simplifies to the Cholesky Decomposition[def. 30.18].

Cholesky Decomposition

Definition 30.18 Cholesky Decomposition: Let $\mathbf{A}$ be a s.p.d.[def. 27.73] then it can be factorized into:

$$\mathbf{A} = \mathbf{G}\mathbf{G}^{\intercal} \qquad \text{with} \qquad \mathbf{G} := \mathbf{L}\mathbf{D}^{1/2} \tag{30.16}$$

3.2. Iterative Methods

4. Iterative Methods for Non-linear Systems

Definition 30.19
General Non-linear System of Equations (NLSE) F:
Is a system of non-linear equations F (that do **not** satisfy linearity??):

$$F :\subseteq \mathbb{R}^n \mapsto \mathbb{R}^n \qquad \text{seek to find} \qquad \mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) = \mathbf{0} \tag{30.17}$$

add linearity under linear algebra

Definition 30.20 Stationary m-point Iteration ϕ_F:
Let $n, m \in \mathbb{R}$ and let $U \subseteq (R^n)^m = \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ be a set.
The function $\phi : U \mapsto \mathbb{R}^n$, called ($m$-point) iteration function is an iterative algorithm that produces an iterative sequence $\left(\mathbf{x}^{(k)}\right)_k$ of approximate solutions to eq. (30.17), using the m most recent iterates:

$$\mathbf{x}^{(k)} = \phi_F\left(\mathbf{x}^{(k-1)}, \ldots, \mathbf{x}^{(k-m)}\right) \tag{30.18}$$

$$\text{Inital Guess} \qquad \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(m-1)}$$

Note

Stationary as ϕ does no explicitly depend on k.

Definition 30.21 Fixed Point $\mathbf{x}^*$:
Is a point $\mathbf{x}^*$ for which the sequence does not change anymore:

$$\mathbf{x}^* = \phi_F\left(\mathbf{x}^{(k-1)}, \ldots, \mathbf{x}^{(k-m)}\right) \qquad \text{with} \qquad \begin{aligned}\mathbf{x}^{(k-1)} &= \mathbf{x}^* \\ &\vdots \\ \mathbf{x}^{(k-m)} &= \mathbf{x}^*\end{aligned} \tag{30.19}$$

4.0.1. Convergence

Question

Does the sequence $\left(\mathbf{x}^{(k)}\right)_k$ converge to a limit:

$$\lim_{k \to \infty} \mathbf{x}^{(k)} = \mathbf{x}^* \tag{30.20}$$

4.0.2. Consistency

Definition 30.22 Consistent m-point Iterative Method:
A *stationary* m-point method[def. 30.20] is *consistent* with a non-linear system of equations[def. 30.19] F iff:

$$F\left(\mathbf{x}^*\right) \iff \phi_F\left(\mathbf{x}^*, \ldots, \mathbf{x}^*\right) = \mathbf{x}^* \tag{30.21}$$

4.0.3. Speed of Convergence

add cvg, consistency, speed of cvng,...

4.1. Fixed Point Iterations $\quad m = 1$

Definition 30.23 Fixed Point Iteration: Is a 1-point method $\phi_F : U \subset \mathbb{R}^n \mapsto \mathbb{R}^n$ that seeks a fixed point $\mathbf{x}^*$ to solve $F(\mathbf{x}) = 0$:

$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \qquad \text{Inital Guess: } \mathbf{x}^{(0)} \tag{30.22}$$

Corollary 30.2 Consistency: If ϕ_F is *continuous* and $\mathbf{x}^* = \lim_{k \to \infty} x^{(k)}$ then $\mathbf{x}^*$ is a fixed point[def. 30.21] of ϕ.

Algorithm 30.1 Fixed Point Iteration:

Input: Inital Guess: $\mathbf{x}^{(0)}$
1: Rewrite $F(\mathbf{x}) = 0$ into a form of $\mathbf{x} = \phi_F(\mathbf{x})$
▷ There exist many ways
2: **for** $k = 1, \ldots, T$ **do**
3: Use the fixed point method:

$$\mathbf{x}^{(k+1)} = \phi_F\left(\mathbf{x}^{(k)}\right) \tag{30.23}$$

4: **end for**

add examples and rest

5. Numerical Quadrature

Definition 30.24 Order of a Quadrature Rule:
The order of a quadrature rule $\mathcal{Q}_n : \mathcal{C}^0([a, b]) \to \mathbb{R}$ is defined as:

$$\mathrm{order}(\mathcal{Q}_n) := \max\left\{n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \int_a^b p(t)\,\mathrm{d}t \quad \forall p \in \mathcal{P}_n\right\} + 1 \tag{30.24}$$

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

Note

Is a quality measure for quadrature rules.

5.1. Composite Quadrature

Definition 30.25 Composite Quadrature:
Given a mesh $\mathcal{M} = \{a = x_0 < x_1 < \ldots < x_m = b\}$ apply a Q.R. $\mathcal{Q}_n$ to each of the mesh cells $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \ldots, m \;\widehat{=}\;$ p.w. Quadrature:

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{j=1}^m \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t = \sum_{j=1}^m \mathcal{Q}_n(f_{I_j}) \tag{30.25}$$

Lemma 30.1 Error of Composite quadrature Rules:
Given a function $f \in \mathcal{C}^k([a, b])$ with integration domain:

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

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

$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k,q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k,q\}}\right) \qquad \text{for } n \to \infty \\ &\overset{\text{[cor. 24.6]}}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{q}\right) \qquad \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{30.26}$$

Definition 30.26 Complexity W: Is the number of function evaluations $\hat{=}$ number of quadrature points.

$$W(\mathcal{Q}(f)_n) = \#\text{f-eval} \hat{=} n \tag{30.27}$$

Lemma 30.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.
Assuming and quadrature error of the form :

$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \qquad \Longleftrightarrow \qquad \epsilon_n(f) = cn^{-q} \qquad c \in \mathbb{R}_+$$

the error complexity is algebraic (**??**) and is given by:

$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{30.28}$$

Proof 30.2: lemma 30.2: **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?

$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \qquad \Rightarrow \qquad n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{30.29}$$

5.1.1. Simpson Integration

Definition 30.27 Simpson Integration:

Optimization

Definition 31.1 Fist Order Method: A first-order method is an algorithm that chooses the k-th iterate in
$$\mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \dots \nabla f(\mathbf{x}_{k-1})\} \qquad \forall k = 1, 2, \dots \tag{31.1}$$

Note
Gradient descent is a first order method

1. Linear Optimization

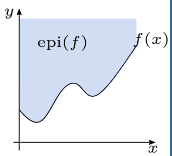
1.1. Polyhedra

Definition 31.2 Polyhedron: Is a set $P \in \mathbb{R}^n$ that can be described by the *finite* intersection of m closed *half spaces*??:
$$P = \{\mathbf{x} \in \mathbb{R}^n \,|\mathbf{A}\mathbf{x} \leqslant \mathbf{b}\} = \{\mathbf{x} \in \mathbb{R}^n \,|\mathbf{a}_j \mathbf{x} \leqslant b_j, j = 1, \dots, m\}$$
$$\mathbf{A} \in \mathbb{R}^{m \times n} \qquad \mathbf{b} \in \mathbb{R}^m \tag{31.2}$$

1.1.1. Polyhedral Function

Definition 31.3 Epigraph/Subgraph **epi(f):**

The epigraph of a function $f \in \mathbb{R}^n \mapsto \mathbb{R}$ is defined as the set of point that lie above its gaph:
$$\operatorname{epi}(f) := \{(\mathbf{x}, y) \in \mathbb{R}^n \,|y \geqslant f(\mathbf{x})\} \subseteq \mathbb{R}^{n+1} \tag{31.3}$$



Definition 31.4 Polyhedral Function: A function f is *polyhedral* if its epigraph $\operatorname{epi}(f)$[def. 31.3] is a polyhedral set[def. 31.2]:
$$f \text{ is polyhedral} \iff \operatorname{epi}(f) \text{ is polyhedral} \tag{31.4}$$

2. Lagrangian Optimization Theory

Add: derivation of lagrange function

Definition 31.5 (Primal) Constraint Optimization: **Given** an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:
$$\begin{aligned} &\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) && \\ \textbf{s.t.} \quad & g_i(\mathbf{w}) \leqslant 0 && 1 \leqslant i \leqslant k \\ & h_j(\mathbf{w}) = 0 && 1 \leqslant j \leqslant m \end{aligned}$$

Definition 31.6 Lagrange Function:
$$\mathscr{L}(\alpha, \beta, \mathbf{w}) := f(\mathbf{w}) + \alpha \mathbf{g}(\mathbf{w}) + \beta \mathbf{h}(\mathbf{w}) \tag{31.5}$$

Extremal Conditions

$\nabla \mathscr{L}(\mathbf{x}) \stackrel{!}{=} 0$ — Extremal point $\mathbf{x}^*$

$\frac{\partial}{\partial \beta} \mathscr{L}(\mathbf{x}) = h(\mathbf{x}) \stackrel{!}{=} 0$ — Constraint satisfisfaction

For the inequality constraints $g(\mathbf{x}) \leqslant 0$ we distinguish two situations:

Case I : $g(\mathbf{x}^*) < 0$ switch const. off

Case II : $g(\mathbf{x}^*) \geqslant 0$ optimze using active eq. constr.

$\frac{\partial}{\partial \alpha} \mathscr{L}(\mathbf{x}) = g(\mathbf{x}) \stackrel{!}{=} 0$ — Constraint satisfisfaction

Definition 31.7 Lagrangian Dual Problem: Is given by:
$$\begin{aligned} \text{Find} \quad & \max \theta(\alpha, \beta) = \inf_{\mathbf{w}\in\Omega} \mathscr{L}(\mathbf{w}, \alpha, \beta) \\ \textbf{s.t.} \quad & \alpha_i \geqslant 0 \qquad 1 \leqslant i \leqslant k \end{aligned}$$

Solution Strategy

1. Find the extremal point $\mathbf{w}^*$ of $\mathscr{L}(\mathbf{w}, \alpha, \beta)$:
$$\left.\frac{\partial \mathscr{L}}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^*} \stackrel{!}{=} 0 \tag{31.6}$$
2. Insert $\mathbf{w}^*$ into $\mathscr{L}$ and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left.\frac{\partial \theta}{\partial \beta}\right|_{\beta=\beta^*} \stackrel{!}{=} 0 \tag{31.7}$$
3. Calculate the solution $\mathbf{w}^*(\beta^*)$ of the constraint minimization problem.

Value of the Problem

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

Theorem 31.1 Upper Bound Dual Cost: Let $\mathbf{w} \in \Omega$ be a feasible solution of the primal problem [def. 31.5] and (α, β) a feasible solution of the respective dual problem [def. 31.7]. Then it holds that:
$$f(\mathbf{w}) \geqslant \theta(\alpha, \beta) \tag{31.8}$$

Proof 31.1:
$$\begin{aligned} \theta(\alpha, \beta) &= \inf_{\mathbf{u}\in\Omega} \mathscr{L}(\mathbf{u}, \alpha, \beta) \leqslant \mathscr{L}(\mathbf{w}, \alpha, \beta) \\ &= f(\mathbf{w}) + \sum_{i=1}^{k} \underbrace{\alpha_i}_{\geqslant 0} \underbrace{g_i(\mathbf{w})}_{\leqslant 0} + \sum_{j=}^{m} \beta_j \underbrace{h_j(\mathbf{w})}_{=0} \\ &\leqslant f(\mathbf{w}) \end{aligned}$$

Corollary 31.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:
$$\sup\{\theta(\alpha, \beta) : \alpha \geqslant 0\} \leqslant \inf\{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leqslant 0, \mathbf{h}(\mathbf{w}) = 0\} \tag{31.9}$$

Theorem 31.2 Optimality: The triple $(\mathbf{w}^*, \alpha^*, \beta^*)$ is a saddle point of the Lagrangian function for the primal problem, if and only if its components are optimal solutions of the primal and dual problems and if there is no duality gap, that is, the primal and dual problems having the same value:
$$f(\mathbf{w}^*) = \theta(\alpha^*, \beta^*) \tag{31.10}$$

Definition 31.8 Convex Optimization: **Given**: a convex function f and a convex set S solve:
$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t.} \quad \mathbf{x} \in S \end{aligned} \tag{31.11}$$
Often S is specified using linear inequalities:
$$\text{e.g.} \qquad S = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{A}\mathbf{x} \leqslant \mathbf{b}\right\}$$

Theorem 31.3 Strong Duality: Given an convex optimization problem:
$$\begin{aligned} &\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) && \\ \textbf{s.t.} \quad & g_i(\mathbf{w}) \leqslant 0 && 1 \leqslant i \leqslant k \\ & h_j(\mathbf{w}) = 0 && 1 \leqslant j \leqslant m \end{aligned}$$
where g_i, h_i can be written as affine functions: $y(\mathbf{w}) = \mathbf{A}\mathbf{w} - b$.
Then it holds that the duality gap is zero and we obtain an optimal solution.

Theorem 31.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,
$$\begin{aligned} &\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) && \\ \textbf{s.t.} \quad & g_i(\mathbf{w}) \leqslant 0 && 1 \leqslant i \leqslant k \\ & h_j(\mathbf{w}) = 0 && 1 \leqslant j \leqslant m \end{aligned}$$
with $f \in C^1$ convex and g_i, h_i affine.
Necessary and sufficient conditions for a normal point $\mathbf{w}^*$ to be an optimum are the existence of α^*, β^* s.t.:
$$\frac{\partial \mathscr{L}(\mathbf{w}, \alpha, \beta)}{\partial \mathbf{w}} \stackrel{!}{=} 0 \qquad \frac{\partial \mathscr{L}(\mathbf{w}^*, \alpha, \beta)}{\partial \beta} \stackrel{!}{=} 0 \tag{31.12}$$
under the condtions that:
- $\forall i_1, \dots, k \qquad \alpha_i^* g_i(\mathbf{w}^*) = 0$, s.t.:
  - Inactive Constraint: $g_i(\mathbf{w}^*) < 0 \to \alpha_i = 0$.
  - Active Constraint:
$$g_i(\mathbf{w}^*) \nless 0 \to \alpha_i \geqslant 0 \qquad \text{s.t.} \qquad \alpha_i^* g_i(\mathbf{w}^*) = 0$$

Consequence

We may become very sparce problems, if a lot of constraints are not actice $\iff \alpha_i = 0$.
Only a few points, for which $\alpha_i > 0$ may affact the decision surface.

Combinatorics

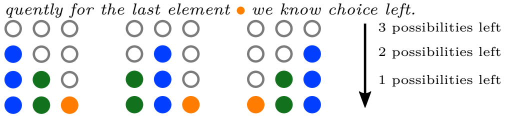
1. Permutations

Definition 32.1 Permutation: A n-Permutation is the (re)*arrangement* of n elements of a set[def. 20.1] $\mathcal{S}$ of size $n = |\mathcal{S}|$ into a sequences[def. 21.2].

Definition 32.2 Number of Permutations of a Set $n!$: Let $\mathcal{S}$ be a set[def. 20.1] $n = |\mathcal{S}|$ *distinct* objects. The number of permutations of $\mathcal{S}$ is given by:

$$P_n(\mathcal{S}) = n! = \prod_{i=0}^{n-1}(n-i) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 \tag{32.1}$$

Explanation 32.1. *If we have i.e. three distinct elements {●, ●, ●} For the first element ● that we arrange we have three possible choices where to put it. However this reduces the number of possible choices for the second element ● to only two. Consequently for the last element ● we know choice left.*



Definition 32.3
Number of Permutations of a Multiset:
Let $\mathcal{S}$ be a multi set[def. 20.3] with $n = |\mathcal{S}|$ total and k *distinct* objects. Let n_j be the multiplicity[def. 20.4] of the member $j \in \{1, \ldots, k\}$ of the multiset $\mathcal{S}$. The permutation of $\mathcal{S}$ is given by:

$$P_{n_1,\ldots,n_k}(\mathcal{S}) = \frac{n!}{n_1! \cdot \ldots \cdot n_k} \qquad \text{s.t.} \qquad \sum_{j=1}^{k} n_j \leqslant n \qquad k < n \tag{32.2}$$

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.

2. Combinations

Definition 32.4 k-Combination:
A k-combination of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is a subset $\mathcal{S}_k$ (order does not matter) of $k = |\mathcal{S}_k|$ *distinct* elements, *chosen* from $\mathcal{S}$.

Definition 32.5 Number of k-Combinations $C_{n,k}$:
The number of k-combinations of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is given by:

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{32.3}$$

3. Variation

Definition 32.6 Variation:
A k-variation of a set $\mathcal{S}$ of size $n = \mathcal{S}$ is

1. a selection/combination[def. 32.4] 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. 32.2] of that subset $\mathcal{S}_k$ (with or without repetition) into a sequence[def. 21.2]

Definition 32.7
Number of Variations without repetitions V_k^n:
Let $\mathcal{S}$ be a set[def. 20.1] $n = |\mathcal{S}|$ *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:

$$V_k^n(\mathcal{S}) = \binom{n}{k}k! = \frac{n!}{(n-k)!} \tag{32.4}$$

Note

Sometimes also denotes as P_k^n.

Definition 32.8
Number of Variations with repetitions $\bar{V}_k^n$:
Let $\mathcal{S}$ be a set[def. 20.1] $n = |\mathcal{S}|$ *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:

$$\bar{V}_k^n(\mathcal{S}) = n^k \tag{32.5}$$

»»»> d641cc7fe8996718a0ca20ca61519f18af078387

Stochastics

Definition 32.9 Stochastics: Is a collective term for the areas of *probability theory* and *statistics*.

Definition 32.10 Statistics: Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.

Definition 32.11 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 32.12 Probability: Probability is the measure of the likelihood that an event will occur in a Random Experiment. Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty.

Improve those definitions: maybe ask on quorahho

Note: Stochastics vs. Stochastic

Stochasticss is a noun and is a collective term for the areas of probability theory and statistics, while stochastic is a *adjective*, describing that a certain phenomena is governed by uncertainty i.e. a process.

Probability Theory

Definition 33.1 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 33.2 [example 33.1]
Sample Space Ω**:**
Is the set of all possible outcomes (elementary events [cor. 33.5]) of an experiment.

Definition 33.3 [example 33.2]
Event A**:**
An "event" is a subset of the sample space Ω and is a property which can be observed to hold or not to hold *after* 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 33.1 : If the outcome ω of an experiment is in the subset A, then the event A is said to "have occured".

Corollary 33.2 Complement Set A^{C}**:**
is the contrary event of A.

Corollary 33.3 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 33.4 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 33.5 The Elementary Event ω**:**
Is a "singleton", i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω.

Corollary 33.6 The Sure Event Ω**:**
Is equal to the sample space as it contains all possible elementary events.

Corollary 33.7 The Impossible Event $\varnothing$**:**
The impossible event i.e. nothing is happening is denoted by the empty set.

Definition 33.4 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 33.5 Probability $\mathbb{P}(A)$**:**
Is a number associated with every A, that measures the likelihood of the event to be realized "a priori". The bigger the number the more likely the event will happen.
1. $0 \leqslant \mathbb{P}(A) \leqslant 1$
2. $\mathbb{P}(\Omega) = 1$
3. If $A \cap B = \varnothing$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Note

We can think of the probability of an event A as the limit of the "frequency" of repeated experiments:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{\delta_n(A)}{n} \qquad \text{where} \qquad \delta(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

1. Sigma Algebras

Definition 33.6 [Proof 33.3]
Sigma Algebra σ**:**
A set $\mathcal{F}$ of subsets of Ω is called a σ-algebra on Ω if the following properties apply
- $\Omega \in \mathcal{F}$ and $\varnothing \in \mathcal{F}$
- If $A \in \mathcal{F}$ then $\Omega \backslash A = A^{\mathrm{C}} \in \mathcal{F}$:
  The complementary subset of A is also in Ω.
- For all $A_i \in \mathcal{F} : \bigcup_{i=1} A_i \in \mathcal{F}$

Explanation 33.1 ([def. 33.6]). *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.

Corollary 33.8 $\mathcal{F}_{\min}$**:** $\mathcal{F} = \{\varnothing, \Omega\}$ is the simplest σ-algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.

Corollary 33.9 $\mathcal{F}_{\max}$**:** $\mathcal{F} = 2^{\Omega}$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.

Definition 33.7 Measurable Space $\{\Omega, \mathcal{F}\}$**:**
Is the pair of a set and sigma algebra i.e. a sample space and sigma algebra $\{\Omega, \mathcal{F}\}$.

Corollary 33.10 $\mathcal{F}$**-measurable Event** $A_i \in \mathcal{F}$**:**
The measurable events A_i of $\mathcal{F}$ are called $\mathcal{F}$-*measurable* or *measurable sets*.

Definition 33.8 [Example 33.4]
Sigma Algebra generated by a subset of Ω $\sigma(\mathcal{C})$**:**
Let $\mathcal{C}$ be a class of subsets of Ω. The σ-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the *smallest* sigma algebra $\mathcal{F}$ that included all elements of $\mathcal{C}$.

Definition 33.9 [Example 33.5]
Borel σ-algebra $\mathcal{B}(\mathbb{R})$**:**
The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals in $\mathbb{R}$. The sets in contained in $\mathcal{B}(\mathbb{R})$ are called Borel sets.
The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward.
For all real numbers $a, b \in \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ contains various sets.

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 33.10 Borel Set:
add

Corollary 33.11 Generating Borel σ-Algebra[Proof 33.1]**:**
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 33.11 ($\mathbb{P}$)-trivial Sigma Algebra:
is a σ-algebra $\mathcal{F}$ for which each event has a probability of zero or one:
$$\mathbb{P}(A) \in \{0, 1\} \qquad \forall A \in \mathcal{F} \tag{33.1}$$

Interpretation

A trivial sigma algebra means that all events are almost surely constant and that there exist no non-trivial information.
An example of a trivial sigma algebra is $\mathcal{F}_{\min} = \{\Omega, \varnothing\}$.

2. Measures

Definition 33.12 Measure μ**:**
A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map:
$$\mu : \mathcal{F} \mapsto [0, \infty] \tag{33.2}$$
for which holds:
- $\mu(\varnothing) = 0$
- countable additivity [def. 33.13]

Definition 33.13 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 $(F_i)_{i\geqslant 1}$ of $\mathcal{F}$ it holds that:
$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \quad \text{for all} \quad F_j \cap F_k = \varnothing \quad \forall j \neq k \tag{33.3}$$

Corollary 33.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds:
$$F \cap G = \varnothing \quad \Longrightarrow \quad \mu(F \cup G) = \mu(F) + \mu(G) \tag{33.4}$$

Explanation 33.2. *If we take two events that cannot occur simultaneously, then the probability that at least one of the events occurs is just the sum of the measures (probabilities) of the original events.*

Definition 33.14 [Example 33.6]
Equivalent Measures $\mu \sim \nu$**:**
Let μ and ν be two measures defined on a measurable space[def. 33.7] $(\Omega, \mathcal{F})$. The two measures are said to be equivalent if it holds that:
$$\mu(A) > 0 \iff \nu(A) > 0 \qquad \forall A \subseteq \mathcal{F} \tag{33.5}$$
this is equivalent to μ and ν having equivalent null sets:
$$\mathcal{N}_\mu = \mathcal{N}_\nu \qquad \begin{aligned} \mathcal{N}_\mu &= \{A \in \mathcal{A} | \mu(A) = 0\} \\ \mathcal{N}_\nu &= \{A \in \mathcal{A} | \nu(A) = 0\} \end{aligned} \tag{33.6}$$

Definition 33.15 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 33.16 Borel Measure: A Borel Measure is any *measure*[def. 33.12] μ defined on the Borel σ-algebra[def. 33.9] $\mathcal{B}(\mathbb{R})$.

2.1.1. The Lebesgue Measure

Definition 33.17 Lebesgue Measure on $\mathcal{B}$ λ**:**
Is the Borel measure[def. 33.16] defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns for every half-open interval $(a, b]$ interval its length:
$$\lambda((a, b]) := b - a \tag{33.7}$$

Corollary 33.13 Lebesgue Measure of Atomitcs:
- The Lebesgue measure of a set containing only one point must be zero:
$$\lambda(\{a\}) = 0 \tag{33.8}$$
- The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2 \ldots, a_n\}$ must be zero:
$$\lambda(A) + \sum_{i=1}^{n} \lambda(\{a_i\}) = 0 \tag{33.9}$$
- The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2 \ldots, \}$ can be either zero, positive and finite or infinite.

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 33.1 Non-negativity: The probability of an event is a non-negative real number:
$$\text{If } A \in \mathcal{F} \qquad \text{then} \qquad \mathbb{P}(A) \geqslant 0 \tag{33.10}$$

Axiom 33.2 Unitairity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one:
$$\text{The certain event} \qquad \mathbb{P}(\Omega) = 1 \tag{33.11}$$

Axiom 33.3 σ-additivity: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are mutually disjoint, then:
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{33.12}$$

Corollary 33.14 : As a consequence of this it follows:
$$\mathbb{P}(\varnothing) = 0 \tag{33.13}$$

Corollary 33.15 Complementary Probability:
$$\mathbb{P}(A^{\mathrm{C}}) = 1 - \mathbb{P}(A) \qquad \text{with} \qquad A^{\mathrm{C}} = \Omega - A \tag{33.14}$$

Definition 33.18 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 33.19 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:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) \neq 0 \tag{33.15}$$

5. Independent Events

Theorem 33.1
Independent Events: Let A, B be two events. A and B are said to be independent iffy:
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \qquad \begin{aligned} \mathbb{P}(A|B) &= \mathbb{P}(A), & \mathbb{P}(B) > 0 \\ \mathbb{P}(B|A) &= \mathbb{P}(B), & \mathbb{P}(A) > 0 \end{aligned} \tag{33.16}$$

Note

The requirement of no impossible events follows from [def. 33.19]

Corollary 33.16 Pairwise Independent Evenest:
A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is *pairwise independent* if every pair of events is independent:
$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cap \mathbb{P}(A_j) \qquad i \neq j, \quad \forall i, j \in \mathcal{A} \tag{33.17}$$

Corollary 33.17 Mutal Independent Evenest:
A finite set of events $\{A_i\}_{i=1}^n \in \mathcal{A}$ is *mutal independent* if every event A_j is independent of any intersection of the other events:
$$\mathbb{P}\left(\bigcap_{i=i}^{k} B_i\right) = \prod_{i=1}^{k} \mathbb{P}(B_i) \qquad \begin{aligned} &\forall \{B_i\}_{i=1}^k \subseteq \{A_i\}_{i=1}^n \\ &k \leqslant n, \qquad \{A_i\}_{i=1}^n \in \mathcal{A} \end{aligned} \tag{33.18}$$

6. Product Rule

Law 33.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{33.19}$$

Law 33.2 Generalized Product Rule/Chain Rule: is the generalization of the product rule?? to n events $\{A_i\}_{i=1}^n$

$$\mathbb{P}\left(\bigcap_{i=i}^{k} E_i\right) = \prod_{k=1}^{n} \mathbb{P}\left(E_k \middle| \bigcap_{i=i}^{k-1} E_i\right) = \tag{33.20}$$

$$= \mathbb{P}(E_n|E_{n-1} \cap \ldots \cap E_1) \cdot \mathbb{P}(E_{n-1}|E_{n-2} \cap \ldots \cap E_1) \cdots \cdots \mathbb{P}(E_3|E_2 \cap E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)$$

7. Law of Total Probability

Definition 33.20 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:

$$\bigcup_{i \in I} A_i = \Omega \qquad A_i \cap A_j = \varnothing \qquad i \neq j, \forall i,j \in I \tag{33.21}$$

Theorem 33.2 Law of Total Probability/Partition Equation: Let $\{A_i : i \in I\}$ be a complete event field[def. 33.20] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{33.22}$$

8. Bayes Theorem

Law 33.3 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) > 0 \tag{33.23}$$

follows directly fromeq. (33.19).

Theorem 33.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field[def. 33.20] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{33.24}$$

proof ?? 33.2

Distributions on $\mathbb{R}$

9.1. Distribution Function

Definition 33.21 Distribution Function of $\mathbb{P}$ F: The *distribution function* F induced by a a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((\infty, x]) \tag{33.25}$$

Theorem 33.4 : A function F is the distribution function of a (unique) probability on $(\mathbb{R}, \mathcal{B})$ iff:

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

Corollary 33.18 : 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:

1. The outcome of an experiment is a random quantity $\omega \in \Omega$
2. 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 $\mathcal{E} = \mathbb{R}^n$:

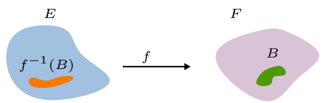
$$X : \Omega \mapsto \mathcal{E} \qquad \omega \mapsto X(\omega)$$

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

$$\underbrace{\mathbb{P}_X(E)}_{\text{Probability for an event in } E} = \mathbb{P}(\{\omega : X(\omega) \in E\}) = \mathbb{P}(X \in E) = \overbrace{\mathbb{P}\left(X^{-1}(E)\right)}^{\text{Probability for an event in } \Omega}$$

Definition 33.22 $\mathcal{E}$-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

$$\forall B \in \mathcal{F} : \qquad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \tag{33.26}$$



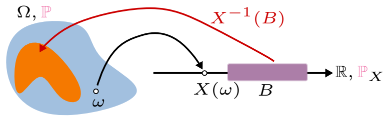
Interpretation

The pre-image[def. 24.14] 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 33.23 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 $(\mathcal{F}, \mathcal{F})$:

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \tag{33.27}$$

Since X is $\mathcal{E}$-measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 33.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ-algebra $\mathcal{B}$ of $\mathbb{R}$.

Corollary 33.20 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. 33.11]):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \tag{33.28}$$

Definition 33.24 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 33.21 Indicator Functions $I_A(\omega)$: An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{33.29}$$

We know that a probability measure $\mathbb{P}$ on $\mathbb{R}$ is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 33.22 : Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let $(E, \mathcal{E})$ be an arbitrary measurable space. Let X be a real value function on E.
Then it holds that X is measurable *if and only if*

$$\{X \leqslant a\} = \{\omega : X(\omega) \leqslant a\} = X^{-1}((-\infty, a]) \in \mathcal{E}, \forall a \in \mathbb{R}$$

or

$$\{X < a\} \in \mathcal{E}.$$

Explanation 33.3 ([cor. 33.22]). *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 33.25 Law/Distribution of X $\mathscr{L}(X)$: 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:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \tag{33.30}$$

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \qquad \forall b \in \mathcal{E}$$

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 33.5 : The law/distribution of X is a probability measure $\mathbb{P}$ on $(E, \mathcal{E})$.

Definition 33.26 (Cumulative) Distribution Function F_X: Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leqslant x) \tag{33.31}$$

Corollary 33.23 : The distribution of $\mathbb{P}^X$ of a real valued r.v. is entirely characterized by its cumulative distribution function F_X [def. 33.33].

Property 33.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \tag{33.32}$$

Property 33.2: Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leqslant B) = F_X(b) - F_X(a) \tag{33.33}$$

9.4. Probability Density Function

Definition 33.27 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 33.28 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:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \tag{33.34}$$

or alternatively:

$$f_X(x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(x \leqslant X \leqslant x + \epsilon)}{\epsilon} \tag{33.35}$$

Corollary 33.24 $\mathbb{P}(X = b) = 0, \quad \forall b \in \mathbb{R}$:

$$\mathbb{P}(X = b) = \lim_{a \to b} \mathbb{P}(a < X \leqslant b) = \lim_{a \to b} \int_a^b f(x) = 0 \tag{33.36}$$

Corollary 33.25 : From [cor. 33.24] it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leqslant X < b) \\ &= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X << b) \end{aligned}$$

Corollary 33.26 :

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{33.37}$$

Notes

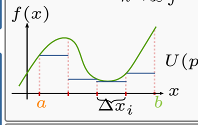
- Often the cumulative distribution function is referred to as "cdf" or simply *distribution function*.
- Often the probability density function is referred to as "pdf" or simply *density*.

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

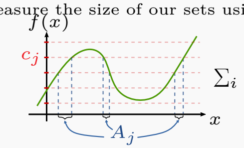
$$\lim_{n \to \infty} \int f(x)\,\mathrm{d}x \overset{\text{str. u.c.}}{=} \int \lim_{n \to \infty} f(x)\,\mathrm{d}x$$



$$U(p) = \sum_i \sup(f(x_i)) \cdot \Delta x_i \xrightarrow{n \to \infty} \int f\,\mathrm{d}x$$

Idea

Partition domain by function values of equal size i.e. values that lie within the same sets/have the same value A_j build up the partitions w.r.t. to the variable x.
Problem: we do not know how big those sets/partitions on the x-axis will be.
Solution: we can use the measure 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.



$$\sum_i c_i \cdot \mu(A_i) \xrightarrow{n \to \infty} \int f\,\mathrm{d}\mu$$

Definition 33.29 Lebesgue Integral:

$$\lim_{n \to \infty} \sum_{i=1}^{n} c_i \mu(A_i) = \int_\Omega f\,\mathrm{d}\mu \qquad \begin{array}{l} f(x) \approx c_i \\ \forall x \in A_i \end{array} \tag{33.38}$$

Definition 33.30 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:

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \qquad a_i \in \mathbb{R} \qquad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \tag{33.39}$$

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 33.1.
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 33.31 Independent Random Variables: Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leqslant x | Y \leqslant y) = \mathbb{P}(X \leqslant x) \qquad \forall x, y \in \mathbb{R} \tag{33.40}$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leqslant x\} \cap \{Y \leqslant y\}) = \mathbb{P}(X \leqslant x, Y \leqslant y) \\ &= F_X(x) F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \tag{33.41}$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad \forall A, B \in \mathcal{B} \tag{33.42}$$

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 33.32
Independent Identically Distributed:

10. Product Rule

Law 33.4 Product Rule: Let X, Y be two random variables then their jo

Law 33.5
Generalized Product Rule/Chain Rule:

11. Change Of Variables Formula

Formula 33.1
(Scalar Discret) Change of Variables: Let X be a discret rv $X \in \mathcal{X}$ with pmf p_X and define $Y \in \mathcal{Y}$ as $Y = g(x)$ s.t. $\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$. **Where** g is an arbitrary strictly monotonic ([def. 24.17]) function.
Let: $\mathcal{X}_y = x_i$ be the set of all $x_i \in \mathcal{X}$ s.t. $y = g(x_i)$.
Then the pmf of Y is given by:

$$p_Y(y) = \sum_{x_i \in \mathcal{X}_y} p_X(x_i) = \sum_{x \in \mathcal{Y}: g(x) = y} p_X(x) \tag{33.43}$$

see proof ?? 33.3

Formula 33.2
(Scalar Continuous) Change of Variables:
Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic[def. 24.17] function.
Define a new r.v. Y as

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\} \tag{33.44}$$

then the pdf of Y is given by:

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\right| \tag{33.45}$$

$$= f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|} \tag{33.46}$$

Formula 33.3
(Continuous) Change of Variables:
Let $X = \{X_1, \ldots, X_n\} \sim f_X$ be a continuous random vector and let g be an arbitrary strictly monotonic[def. 24.17] function

$$g : \mathbb{R}^n \mapsto \mathbb{R}^m$$

Define a new r.v. Y as

$$\mathcal{Y} = \{\mathbf{y}|\mathbf{y} = g(\mathbf{x}), \forall x \in \mathcal{X}\} \tag{33.47}$$

and let $h(\mathbf{x}) := g(\mathbf{x})^{-1}$ then the pdf of Y is given by:

$$\begin{aligned} f_Y(\mathbf{y}) &= f_X(x_1, \ldots, x_n) \cdot |J| \\ &= f_X(h_1(\mathbf{y}), \ldots, h_n(\mathbf{y})) \cdot |J| \\ &= f_X(\mathbf{y})|\det D_{\mathbf{x}} h(\mathbf{x})|\Big|_{\mathbf{x}=\mathbf{y}} \\ &= f_X(g^{-1}(\mathbf{y}))\left|\det\left(\frac{\partial g}{\partial \mathbf{x}}\right)^{-1}\right| \end{aligned} \tag{33.48}$$

where $J = \det Dh$ is the Jaccobian[def. 25.6].
See also proof ?? 33.6 and example 33.8

Note
A monotonic function is required in order to satisfy inevitability.

Probability Distributions on $\mathbb{R}^n$

13. Joint Distribution

Definition 33.33
Joint (Cumulative) Distribution Function $F_{\mathbf{X}}$**:**
Let $\mathbf{X} = (X_1 \quad X_n)$ be a random vector in $\mathbb{R}^n$, then its *cumulative distribution function* is defined as:

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}^X((-\infty, \mathbf{x}]) = \mathbb{P}(\mathbf{X} \leqslant \mathbf{x}) \\ &= \mathbb{P}(X_1 \leqslant x_1, \ldots X_n \leqslant x_n) \end{aligned} \tag{33.49}$$

Definition 33.34 Joint Probability Distribution:
Let $\mathbf{X} = (X_1 \quad X_n)$ be a random vector in $\mathbb{R}^n$ with associated cdf $F_{\mathbf{X}}$. If $F_{\mathbf{X}}$ is continuously integrable for all $\mathbf{x} \in \mathbb{R}$ then $\mathbf{X}$ has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\mathbf{X}}(y_1, \ldots, y_n)\, \mathrm{d}y_1\, \mathrm{d}y_n \tag{33.50}$$

or alternatively:

$$f_{\mathbf{X}}(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + \epsilon, \ldots, x_n \leqslant X_n \leqslant x_n + \epsilon)}{\epsilon} \tag{33.51}$$

13.1. Marginal Distribution

Definition 33.35 Marginal Distribution:

14. The Expectation

Definition 33.36 Expectation:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int_\Omega X\, \mathrm{d}\mathbb{P} \tag{33.52}$$

Corollary 33.27 Expectation of simple r.v.:
If X is a simple[def. 33.30] r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^n a_i \mathbb{P}(A_i) \tag{33.53}$$

14.1. Properties

14.1.1. Linear Operators

add

14.1.2. Quadratic Form

Definition 33.37 proof 33.7
Expectation of a Quadratic Form:
Let $\epsilon \in \mathbb{R}^n$ be a random vector with $\mathbb{E}[\epsilon] = \mu$ and $\mathbb{V}[\epsilon] = \Sigma$:

$$\mathbb{E}\left[\epsilon^\intercal \mathbf{A} \epsilon\right] = \operatorname{tr}(\mathbf{A}\Sigma) + \mu^\intercal \mathbf{A} \mu \tag{33.54}$$

14.2. The Jensen Inequality

Theorem 33.6 Jensen Inequality: Let X be a random variable and g be some function, then it holds:

$$\begin{aligned} g(\mathbb{E}[X]) &\leqslant \mathbb{E}[g(X)] \\ g(\mathbb{E}[X]) &\geqslant \mathbb{E}[g(X)] \end{aligned} \quad \text{if} \quad \begin{aligned} &g \text{ is convex}^{[\text{def. } 24.27]} \\ &g \text{ is concave}^{[\text{def. } 24.28]} \end{aligned} \tag{33.55}$$

14.3. Law of the Unconscious Statistician

Law 33.6 Law of the Unconscious Statistician:
Let $X \in \mathcal{X}$,$Y \in \mathcal{Y}$ be random variables where Y is defined as:

$$\mathcal{Y} = \{y|y = g(x), \forall x \in \mathcal{X}\}$$

then the expectation of Y can be calculated in terms of X:

$$\mathbb{E}_Y[y] = \mathbb{E}_X[g(x)] \tag{33.56}$$

Consequence
Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

14.4. Properties

14.5. Law of Iterated Expectation (LIE)

Law 33.7 [proof 33.8]
Law of Iterated Expectation (LIE):

$$\mathbb{E}[X] = \mathbb{E}_Y \mathbb{E}[X|Y] \tag{33.57}$$

14.6. Hoeffdings Bound

Definition 33.38 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:

$$\mathbb{P}(|\mu_{\mathbf{X}} - \mathbb{E}[X]| \geqslant \epsilon) \leqslant 2\exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \overset{[0,1]}{=} 2\mathrm{e}^{-2n\epsilon^2} \tag{33.58}$$

Explanation 33.4. *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 33.39 Moment of Random Variable: The i-th moment of a random variable X is defined as (if it exists):

$$m_i := \mathbb{E}\left[X^i\right] \tag{33.59}$$

Definition 33.40 ψ_X
Moment Generating Function (MGF):

$$\psi_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \qquad t \in \mathbb{R} \tag{33.60}$$

Corollary 33.28 Sum of MGF: The moment generating function of a sum of n independent variables $(X_j)_{1 \leqslant j \leqslant n}$ is the product of the moment generating functions of the components:

$$\psi_{S_n}(t) = \psi_{X_1}(t) \cdots \psi_{X_n}(t) \qquad S_n := X_1 + \ldots X_n \tag{33.61}$$

Corollary 33.29 : The i-th moment of a random variable is the i-th derivative of its associated moment generating function evaluated zero:

$$\mathbb{E}\left[X^i\right] = \psi_X^{(i)}(0) \tag{33.62}$$

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 33.41 $\hat{\mu}$
Fourier Transformed Probability Measure:

$$\hat{\mu} = \int \mathrm{e}^{i\langle u, x \rangle} \mu(\mathrm{d}x) \tag{33.63}$$

Corollary 33.30 : As $e^{i\langle u, x \rangle}$ can be rewritten using formulaeqs. (20.9) and (20.10) it follows:

$$\hat{\mu} = \int \cos(\langle u, x \rangle)\, \mu(\mathrm{d}x) + i \int \sin(\langle u, x \rangle)\, \mu(\mathrm{d}x) \tag{33.64}$$

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 33.42 Characteristic Function φ_X**:** Let $\mathbf{X}$ be an $\mathbb{R}^n$-valued random variable. Its characteristic function $\varphi_{\mathbf{X}}$ is defined on $\mathbb{R}^n$ as:

$$\varphi_{\mathbf{X}}(u) = \int \mathrm{e}^{i\langle \mathbf{u}, \mathbf{x} \rangle} \mathbb{P}^X(\mathrm{d}\mathbf{x}) = \widehat{\mathbb{P}^X}(\mathbf{u}) \tag{33.65}$$

$$= \mathbb{E}\left[\mathrm{e}^{i\langle \mathbf{u}, \mathbf{x} \rangle}\right] \tag{33.66}$$

Corollary 33.31 : 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 33.7 : 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 33.8 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 33.32 : Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an $\mathbb{R}^n$-valued random variable. Then the real valued r.v.'s $(X_j)_{1 \leqslant j \leqslant n}$ are independent if and only if:

$$\varphi_X(u_1, \ldots, u_n) = \prod_{j=1}^n \varphi_{X_j}(u_j) \tag{33.67}$$

Proofs

Proof 33.1: [cor. 33.11]: Let $\mathcal{C}$ denote all open intervals.
Since every open set in $\mathbb{R}$ is the countable union of open intervals[def. 20.12], 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

$$(a, b) = \cup_{n=1}^\infty (a_n, b_n] = \cup_{n=1}^\infty \left(-\infty, b_n\right] \cap (-\infty, a_n]^{\mathrm{C}}\Big)$$

Thus $\mathcal{C} \subset \sigma(\mathcal{D})$, whence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$ **but** as each element of $\mathcal{D}$ is a closed subset, $\sigma(\mathcal{D})$ must also be contained in the Borel sets $\mathcal{B}$ with

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma((D) \subset \mathcal{B}$$

Proof 33.2: theorem 33.3 Plug eq. (33.22) into the denominator and ?? into the nominator and then use [def. 33.19]:

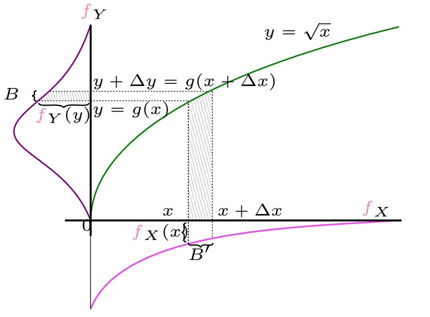
$$\frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \mathbb{P}(A_j|B)$$

Proof 33.3: ??:

$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = p_Y(y)$$

Proof 33.4: ?? (non-formal): The probability contained in a differential area must be invariant under a change of variables that is:

$$|f_Y(y)\,\mathrm{d}y| = |f_x(x)\,\mathrm{d}x|$$



Proof 33.5: ?? from CDF:

$$\mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \begin{cases} \mathbb{P}(X \leqslant g^{-1}(y)) & \text{if } g \text{ is increas.} \\ \mathbb{P}(X \geqslant g^{-1}(y)) & \text{if } g \text{ is decreas.} \end{cases}$$

If g is monotonically increasing:

$$F_Y(y) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

If g is monotonically decreasing:

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = -f_X(x) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

Proof 33.6: ??: Let $B = [x, x+\Delta x]$ and $B' = [y, y+\Delta y] = [g(x), g(x+\Delta x)]$ we know that the probability of equal events is equal:

$$y = g(x) \qquad \Rightarrow \qquad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t)\,dt \xrightarrow{\Delta x \to 0} |\Delta x \cdot f_x(x)|$$

For y we use Taylor (??)

$$g(x+\Delta x) \overset{\text{eq. (24.55)}}{=} g(x) + \frac{dg}{dx}\Delta y \qquad \text{for } \Delta x \to 0$$

$$= \qquad y + \Delta y \qquad \text{with } \Delta y := \frac{dg}{dx}\cdot \Delta x \tag{33.68}$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\mathbb{P}(X \in B') = \int_y^{y+\Delta y} f_Y(t)\,dt \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)| = \left|\frac{dg}{dx}(x)\Delta x \cdot f_Y(y)\right|$$

Now we simply need to related the surface of the two pdfs:

$$B = [x, x+\Delta x] \overset{\text{same surfaces}}{\propto} [y, y+\Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\overset{\Delta y \to 0}{\iff} |f_Y(y)\cdot \Delta y| = \left|f_Y(y)\cdot \frac{dg}{dx}(x)\Delta x\right| = |f_X(x)\cdot \Delta x|$$

$$f_Y(y)\left|\frac{dg}{dx}(x)\right||\Delta x| = f_X(x)\cdot|\Delta x|$$

$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left|\frac{dg}{dx}(x)\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{dg}{dx}g^{-1}(y)\right|}$$

Proof 33.7: [def. 33.37]

$$\begin{aligned}
\mathbb{E}\left[\epsilon^\intercal \mathbf{A}\epsilon\right] &\overset{\text{eq. (27.52)}}{=} \mathbb{E}\left[\operatorname{tr}(\epsilon^\intercal \mathbf{A}\epsilon)\right] \\
&\overset{\text{eq. (27.54)}}{=} \mathbb{E}\left[\operatorname{tr}(\mathbf{A}\epsilon\epsilon^\intercal)\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[\mathbf{A}\epsilon\epsilon^\intercal\right]\right) \\
&= \operatorname{tr}\left(\mathbf{A}\mathbb{E}\left[\epsilon\epsilon^\intercal\right]\right) \\
&= \operatorname{tr}\left(\mathbf{A}\left(\Sigma + \mu\mu^\intercal\right)\right) \\
&= \operatorname{tr}(\mathbf{A}\Sigma) + \operatorname{tr}(\mathbf{A}\mu\mu^\intercal) \\
&\overset{\text{eq. (27.52)}}{=} \operatorname{tr}(\mathbf{A}\Sigma) + \mathbf{A}\mu\mu^\intercal
\end{aligned}$$

Proof 33.8: law 33.7

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x\cdot p_X(x) = \sum_x x \cdot \sum_y p_{X,Y}(x,y) \\
&= \sum_x x\cdot \sum_y p_{X|Y}(x|y)\cdot p_Y(y) \\
&= \sum_y p_Y(y)\cdot \sum_x x\cdot p_{X|Y}(x|y) \\
&= \sum_y p_Y(y)\cdot \mathbb{E}[X|Y] = \mathbb{E}_Y\left[\mathbb{E}[X|Y]\right]
\end{aligned}$$

Examples

Example 33.1 :
- 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 33.2 :
- 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 33.3 : If the sample space is a die toss $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, the sample space may be that we are only told whether an even or odd number has been rolled:

$$\mathcal{F} = \{\varnothing, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$$

Example 33.4 : If we are only interested in the subset $A \in \Omega$ of our experiment, then we can look at the corresponding generating σ-algebra $\sigma(A) = \left\{\varnothing, A, A^C, \Omega\right\}$.

Example 33.5 :
- open half-lines: $(-\infty, a)$ and (a, ∞),
- union of open half-lines: $(a, b) = (-\infty, a) \cup (b, \infty)$,
- closed interval: $[a, b] = \overline{(-\infty, \cup a) \cup (b, \infty)}$,
- closed half-lines: $(-\infty, a] = \bigcup_{n=1}^\infty [a-n, a]$ and $[a, \infty) = \bigcup_{n=1}^\infty [a, a+n]$,
- half-open and half-closed $(a, b] = (-\infty, b] \cup (a, \infty)$,
- every set containing only one real number: $\{a\} = \bigcap_{n=1}^\infty (a - \frac{1}{n}, a + \frac{1}{n})$,
- every set containing finitely many real numbers: $\{a_1, \dots, a_n\} = \bigcup_{k=1}^n a_k$.

Example 33.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\} \qquad \begin{aligned}\mathbb{P}(\{1,2,3\}) &= \{2/3, 1/6, 1/6\} \\ \widetilde{\mathbb{P}}(\{1,2,3\}) &= \{1/3, 1/3, 1/3\}\end{aligned}$$

Example 33.7 :

add example fat book p.1286

add example prob th book 4

Example 33.8 ??: Let $X, Y \overset{\text{ind.}}{\sim} \mathcal{N}(0,1)$.
Question: proof that:

$$U = X + Y \qquad V = X - 1$$

are indepdent and normally distributed:

$$h(u,v) = \begin{cases} h_1(u,v) = \frac{u+v}{2} \\ h_2(u,v) = \frac{u-v}{2}\end{cases} \qquad J = \det\begin{bmatrix}\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2}\end{bmatrix} = -\frac{1}{2}$$

$$\begin{aligned}
f_{U,V} &= f_{X,Y}(\underline{x}, \underline{y})\cdot \frac{1}{2} \\
&\overset{\text{indp.}}{=} f_X(\underline{x})\cdot f_X(\underline{y}) \\
&= \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} \\
&= \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\left\{\left(\frac{u+v}{2}\right)^2 + \left(\frac{u-v}{2}\right)^2/2\right\}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{u^2}{4}}\cdot \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{v^2}{4}}
\end{aligned}$$

Thus U, V are independent r.v. distributed as $\mathcal{N}(0, 2)$.

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:

$$p_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapet

Definition 34.1 Almost Surely $\mathbb{P}$-(a.s.):
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \qquad \iff \qquad \omega \text{ happens a.s.} \tag{34.1}$$

Definition 34.2 Probability Mass Function (PMF):

Definition 34.3 Discrete Random Variable (DVR): The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \dots, 8\} \qquad \mathcal{X} = \mathbb{N} \tag{34.2}$$

Definition 34.4 Probability Density Function (PDF): Is real function $f: \mathbb{R}^n \to [0, \infty)$ that satisfies:

Non-negativity: $f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n$ (34.3)

Normalization: $\int_{-\infty}^{\infty} f(x)\,dx \overset{!}{=} 1$ (34.4)

Must be integrable (34.5)

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:

$$\mathbb{P}(X = x) = 0 \qquad \text{a.s.}$$

To have a sensible measure of the magnitude of these infinitesimal quantities, we use the concept of probability density, which yields a probability mass when integrated over an interval.

Definition 34.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf ([def. 34.4]) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx \tag{34.6}$$

Property 34.1 Zero Probability: If X is a continuous rrv ([def. 34.5]), then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \tag{34.7}$$

Property 34.2 Open vs. Closed Intervals: For any real numbers a and b, with $a < b$ it holds:

$$\begin{aligned}\mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b)\end{aligned} \tag{34.8}$$

$\iff$ 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 34.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_x(x)\,dx$$

Proof 34.1: Property 34.1:

$$\mathbb{P}(X = a) = \lim_{\Delta x \to 0}\mathbb{P}(X \in [a, a+\Delta x]) = \lim_{\Delta x \to 0}\int_a^{a+\Delta x} f_X(x)\,dx = 0$$

Proof 34.2: Property 34.2:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,dx$$

Definition 34.6 Support of a probability density function: The support of the density of a pdf $f_X(.)$ is the set of values of the random variable X s.t. its pdf is non-zero:

$$\operatorname{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{34.9}$$

Note: this is not a rigorous definition.

Theorem 34.1 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 34.2 Identically Distributed: From theorem 34.1 it follows that to RV X and Y that have exactly the same pdf follow the same distribution.
We say X and Y are identically distributed.

0.1. Cumulative Distribution Fucntion

Definition 34.7 Cumulative distribution function (CDF): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
The (cumulative) distribution function of a real-valued random variable X is the function given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) \qquad \forall x \in \mathbb{R}$$

Property 34.3:

Monotonically Increasing $\quad x \leqslant y \iff F_X(x) \leqslant F_X(y) \quad \forall x, y \in \mathbb{R}$ (34.10)

Upper Limit $\quad \lim_{x\to\infty} F_X(x) = 1$ (34.11)

Lower Limit $\quad \lim_{x\to-\infty} F_X(x) = 0$ (34.12)

Definition 34.8 CDF of a discret rv X: **Let** X be discret rv with pdf p_X, then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{t=-\infty}^{x} p_X(t)$$

Definition 34.9 CDF of a continuous rv X: **Let** X be continuous rv with pdf f_X, then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t)\,dt \qquad \iff \qquad \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 34.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = F_X(b) - F_X(a) \tag{34.13}$$

Proof 34.3: [def. 34.9]:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^x f_X(t)\,dt$$

Proof 34.4: lemma 34.1:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$

or by the fundamental theorem of calculus (theorem 24.2):

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,dt = \int_a^b \frac{\partial F_X(t)}{\partial t}\,dt = \left[F_X(t)\right]_a^b$$

Theorem 34.2 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 34.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{x}\in\mathcal{X}} \bar{x} p_x(\bar{x}) \tag{34.14}$$

Definition 34.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{x}\in\mathcal{X}} \bar{x} f_x(\bar{x})\,d\bar{x} \tag{34.15}$$

Law 34.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{34.16}$$

Property 34.4 Translation and scaling: If $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \tag{34.17}$$

Thus $\mathbb{E}$ is a linear operator ([def. 24.18]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 34.6 it follows:

$$\mathbb{E}\left[\mathbb{E}[X]\right] = \mathbb{E}[X] \tag{34.18}$$

Property 34.5 Matrix×Expectation: If $\mathbf{X} \in \mathbb{R}^n$ is a randomn vector and $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{AXB}] = \mathbf{A}\mathbb{E}[(\mathbf{XB})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\,\mathbf{B} \tag{34.19}$$

Proof 34.5: eq. (34.24):

$$\mathbb{E}[XY] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)xy \overset{??}{=} \sum_{x\in\mathcal{X}} p_X(x)x \sum_{y\in\mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Definition 34.12
Autocorrelation/Crosscorelation $\gamma(t_1,t_2)$: Describes the covariance ([def. 34.16]) between the two values of a stochastic process $(\mathbf{X}_t)_{t\in T}$ at different time points t_1 and t_2.

$$\gamma(t_1,t_2) = \mathrm{Cov}\left[\mathbf{X}_{t_1},\mathbf{X}_{t_2}\right] = \mathbb{E}\left[\left(\mathbf{X}_{t_1}-\mu_{t_1}\right)\left(\mathbf{X}_{t_2}-\mu_{t_2}\right)\right] \tag{34.20}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t,t) = \mathrm{Cov}\left[\mathbf{X}_t,\mathbf{X}_t\right] \overset{\text{eq. (34.35)}}{=} \mathbb{V}\left[\mathbf{X}_t\right] \tag{34.21}$$

Notes

- **Hence** the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- **Given** a random time dependent variable $\mathbf{x}(t)$ the autocorrelation function $\gamma(t, t-\tau)$ describes how *similar* the time translated function $\mathbf{x}(t-\tau)$ and the original function $\mathbf{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is non-zero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

move to prob theory maybe

Definition 34.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\mathbf{x}}\in\mathcal{X}} \bar{\mathbf{x}}\, p_x(\bar{\mathbf{x}}) \tag{34.22}$$

Definition 34.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\mathbf{x}}\in\mathcal{X}} \bar{\mathbf{x}}\, f_x(\bar{\mathbf{x}})\,\mathrm{d}\bar{\mathbf{x}} \tag{34.23}$$

Law 34.2 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{34.24}$$

Property 34.6 Translation and scaling: If $\mathbf{X}\in\mathbb{R}^n$ and $\mathbf{Y}\in\mathbb{R}^n$ are random vectors, and $a, b, a\in\mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\mathbf{X} + c\mathbf{Y}] = a + b\mathbb{E}[\mathbf{X}] + c\mathbb{E}[\mathbf{Y}] \tag{34.25}$$

Thus $\mathbb{E}$ is a linear operator[def. 24.18].

Property 34.7
Affine Transformation of the Expectation:
If $\mathbf{X}\in\mathbb{R}^n$ is a randomn vector, $\mathbf{A}\in\mathbb{R}^{m\times n}$ a constant matrix and $b\in\mathbb{R}^n$ then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mu + b \tag{34.26}$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 34.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{34.27}$$

Property 34.8 Matrix×Expectation: If $\mathbf{X}\in\mathbb{R}^n$ is a randomn vector and $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{B}\in\mathbb{R}^{n\times m}$ are constant matrices then it holds:

$$\mathbb{E}[\mathbf{A}\mathbf{X}\mathbf{B}] = \mathbf{A}\mathbb{E}[(\mathbf{X}\mathbf{B})] = \mathbf{A}\mathbb{E}[\mathbf{X}]\,\mathbf{B} \tag{34.28}$$

Proof 34.6: eq. (34.24):

$$\mathbb{E}[XY] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)xy \overset{??}{=} \sum_{x\in\mathcal{X}} p_X(x)x \sum_{y\in\mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]$$

2.2. The Variance

Definition 34.15 Variance $\mathbb{V}[X]$: The variance of a random variable X is the expected value of the squared deviation from the expectation of X ($\mu = \mathbb{E}[X]$).
It is a measure of how much the actual values of a random variable X fluctuate around its executed value $\mathbb{E}[X]$ and is defined by:

$$\mathbb{V}[X] := \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \overset{\text{see ?? 34.7}}{=} \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{34.29}$$

2.2.1. Properties

Property 34.9 Variance of a Constant: If $a\in\mathbb{R}$ is a constant then it follows that its expected value is deterministic $\Rightarrow$ we have no uncertainty $\Rightarrow$ no variance:

$$\mathbb{V}[a] = 0 \qquad \textbf{with} \qquad a\in\mathbb{R} \tag{34.30}$$

see shift and scaling for proof ?? 34.8

Property 34.10 Shifting and Scaling:

$$\mathbb{V}[a + bX] = a^2\sigma^2 \qquad \textbf{with} \qquad a\in\mathbb{R} \tag{34.31}$$

see ?? 34.8

Property 34.11 [proof 34.9]
Affine Transformation of the Variance:
If $\mathbf{X}\in\mathbb{R}^n$ is a randomn vector, $\mathbf{A}\in\mathbb{R}^{m\times n}$ a constant matrix and $b\in\mathbb{R}^n$ then it holds:

$$\mathbb{V}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal \tag{34.32}$$

Definition 34.16 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned} \tag{34.33}$$

see ?? 34.10

Definition 34.17 Covariance Matrix: The variance of a k-dimensional random vector $\mathbf{X} = (X_1 \ \ldots \ X_k)$ is given by a p.s.d. eq. (27.107) matrix called Covariance Matrix.
The Covariance is a measure of how much two or more random variables vary linearly with each other and the Variance on the diagonal is again a measure of how much a variable varies:

$$\begin{aligned}\mathbb{V}[\mathbf{X}] :=& \Sigma(\mathbf{X}) := \mathrm{Cov}[\mathbf{X},\mathbf{X}] := \\ =& \mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-\mathbb{E}[\mathbf{X}])^\intercal\right] \\ =& \mathbb{E}\left[\mathbf{X}\mathbf{X}^\intercal\right] - \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}]^\intercal \in [-\infty,\infty] \\ =& \begin{bmatrix}\mathbb{V}[X_1] & \cdots & \mathrm{Cov}[X_1,X_k] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_k,X_1] & \cdots & \mathbb{V}[X_k]\end{bmatrix} \\ =& \begin{bmatrix}\mathbb{E}[(X_1-\mu_1)(X_1-\mu_1)] & \cdots & \mathbb{E}[(X_1-\mu_1)(X_k-\mu_k)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_k-\mu_k)(X_1-\mu_1)] & \cdots & \mathbb{E}[(X_k-\mu_k)(X_k-\mu_k)]\end{bmatrix}\end{aligned} \tag{34.34}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\mathrm{Cov}[X,X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{34.35}$$

add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Property 34.12 Translation and Scaling:

$$\mathrm{Cov}(a + bX, c + dY) = bd\,\mathrm{Cov}(X,Y) \tag{34.36}$$

Property 34.13
Affine Transformation of the Covariance:
If $\mathbf{X}\in\mathbb{R}^n$ is a randomn vector, $\mathbf{A}\in\mathbb{R}^{m\times n}$ a constant matrix and $b\in\mathbb{R}^n$ then it holds:

$$\mathrm{Cov}[\mathbf{A}\mathbf{X} + b] = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal = \mathbf{A}\Sigma(\mathbf{X})\mathbf{A}^\intercal \tag{34.37}$$

Definition 34.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}\mathrm{Corr}[\mathbf{X}] :=& \frac{\mathrm{Cov}[\mathbf{X}]}{\sigma_{X_1}\cdots\sigma_{X_k}} \in [-1,1] \\ =& \begin{cases}+1 & \text{if } Y = aX + b \text{ with } a > 0, b\in\mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b\in\mathbb{R}\end{cases}\end{aligned} \tag{34.38}$$

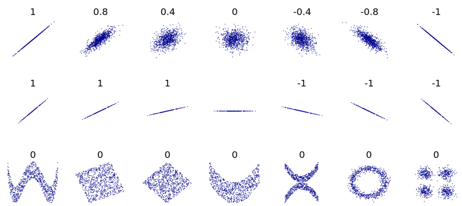


Figure 13: Several sets of (x, y) points, with their correlation coefficient

Law 34.3 Translation and Scaling:

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \tag{34.39}$$

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 $\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$ implies that there does not exist a **linear** relationship between the random variables X and Y.

Difference Covariance&Correlation

1. Variance is affected by scaling and covariance not ?? and law 34.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 34.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (??) is zero:

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \overset{\text{eq. (34.24)}}{=} \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

Zero covariance/correlation$\nRightarrow$ independence

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0 \nRightarrow p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

For example: let $X \sim \mathcal{U}([-1,1])$ and let $Y = X^2$.

1. Clearly X and Y are dependent
2. **But** the covariance/correlation between X and Y is non-zero:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X,X^2) = \mathbb{E}\left[X\cdot X^2\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \\ &= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \underset{\text{eq. (34.52)}}{\overset{\text{eq. (34.63)}}{=}} 0 - 0\cdot\mathbb{E}\left[X^2\right]\end{aligned}$$

$\Rightarrow$ the relationship between Y and X must be non-linear.

Definition 34.19 Quantile: Are specific values q_α in the range[def. 24.13] of a random variable X that are defined as the value for which the cumulative probability is less then q_α with probability $\alpha\in(0,1)$:

$$q_\alpha : \mathbb{P}(X \leqslant x) = F_X(q_\alpha) = \alpha \xrightarrow{F \text{ invert.}} q_\alpha = F_X^{-1}(\alpha) \tag{34.40}$$

add figure

3. Proofs

Proof 34.7: eq. (34.29)

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &\overset{\text{Property 34.6}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

Proof 34.8: Property 34.10

$$\begin{aligned}\mathbb{V}[a+bX] &= \mathbb{E}\left[(a + bX - \mathbb{E}[a+bX])^2\right] \\ &= \mathbb{E}\left[(\cancel{a} + bX - \cancel{a} - b\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[(bX - b\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[b^2(X-\mathbb{E}[X])^2\right] \\ &= b^2\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = b^2\sigma^2\end{aligned}$$

Proof 34.9: Property 34.11

$$\begin{aligned}\mathbb{V}(\mathbf{A}\mathbf{X}+b) &= \mathbb{E}\left[(\mathbf{A}\mathbf{X}-\mathbb{E}[\mathbf{X}\mathbf{A}])^2\right] + 0 = \\ &= \mathbb{E}\left[(\mathbf{A}\mathbf{X}-\mathbb{E}[\mathbf{A}\mathbf{X}])(\mathbf{A}\mathbf{X}-\mathbb{E}[\mathbf{A}\mathbf{X}])^\intercal\right] \\ &= \mathbb{E}\left[\mathbf{A}(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{A}(\mathbf{X}-(\mathbb{E}[\mathbf{X}]))^\intercal\right] \\ &= \mathbb{E}\left[\mathbf{A}(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-(\mathbb{E}[\mathbf{X}])^\intercal\mathbf{A}^\intercal\right] \\ &= \mathbf{A}\mathbb{E}\left[(\mathbf{X}-\mathbb{E}[\mathbf{X}])(\mathbf{X}-(\mathbb{E}[\mathbf{X}])^\intercal\right]\mathbf{A}^\intercal = \mathbf{A}\mathbb{V}[\mathbf{X}]\,\mathbf{A}^\intercal\end{aligned}$$

Proof 34.10: eq. (34.33)

$$\begin{aligned}
&\mathrm{Cov}[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

Discrete Distributions

Definition 34.20 Multivariate Distribution: the variate refers to the number of input variables i.e. a m-variate distribution has m-input variables whereas a uni-variate distribution has only one.

Dimensional vs. Multivariate

The dimension refers to the number of dimensions we need to embed the function. If the variables of a function are independent than the dimension is the same as the number of inputs but the number of input variables can also be less.

4.1. Bernoulli Distribution $\mathrm{Bern}(\mathrm{p})$

Definition 34.21 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 34.22 Bernoulli Distribution $X \sim \mathrm{Bern}(\mathrm{p})$:
X is a binary variable i.e. can only attain the values 0 (failure) or 1 (success) with a parameter p that signifies the success probability:

$$\mathrm{p}(x;\mathrm{p}) = \begin{cases}\mathrm{p} & \text{for } x=1\\ 1-\mathrm{p} & \text{for } x=0\end{cases} \iff \begin{cases}\mathbb{P}(X=1)=\mathrm{p}\\ \mathbb{P}(X=0)=1-\mathrm{p}\end{cases}$$
$$= \mathrm{p}^x\cdot(1-\mathrm{p})^{1-x} \quad \text{for } x\in\{0,1\}$$

$$\mathbb{E}[X]=\mathrm{p} \quad (34.41) \qquad \mathbb{V}[X]=\mathrm{p}(1-\mathrm{p}) \quad (34.42)$$

4.2. Multinoulli/Categorical Distribution $\mathrm{Cat}(n,\mathrm{p})$

Definition 34.23
Multinoulli/Categorical Distribution $X \sim \mathrm{Cat}(\mathrm{p})$:
Is the generalization of the Bernoulli distribution[def. 34.22] to a sample space[def. 33.2] of k individual items $\{c_1,\dots,c_c\}$ with probabilities $\mathrm{p}=\{\mathrm{p}_1,\dots,\mathrm{p}_k\}$:

$$p(x=c_i|\mathrm{p}) = \mathrm{p}_i \iff p(x|\mathrm{p}) = \prod_i^k \mathrm{p}_i^{\delta[x=c_i]}$$

$$\sum_{j=1}^k \mathrm{p}_j = 1 \qquad \mathrm{p}_j\in[0,1] \qquad \forall j=1,\dots,k \quad (34.43)$$

$$\mathbb{E}[X]=\mathrm{p} \qquad \mathbb{V}[X]_{i,j}=\Sigma_{i,j}=\begin{cases}\mathrm{p}_i(1-\mathrm{p}_i) & \text{if } i=j\\ -\mathrm{p}_i\mathrm{p}_j & \text{if } i\neq j\end{cases}$$

Corollary 34.3
One-hot encoded Categorical Distribution:
If we encode the k categories by a *sparse vectors*[def. 27.68] with norm one:

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

s.t. $\qquad \mathbf{x}_j=\mathbf{e}_j \iff \mathbf{x}=c_j$

then we can rewrite eq. (34.43) as:

$$\mathrm{p}(\mathbf{x}|\mathrm{p}) = \prod_i^k \mathbf{x}_i\cdot\mathrm{p}_i \qquad \sum_{j=1}^k \mathrm{p}_j = 1 \quad (34.44)$$

4.3. Binomial Distribution $\mathcal{B}(n,\mathrm{p})$

Definition 34.24 Binomial Coefficient: The binomial coefficient occurs inside the binomial distribution?? and signifies the different combinations/order that x out of n successes can happen.

Definition 34.25 Binomial Distribution [proof ??]:
Models the probability of exactly X success given a fixed number n-*Bernoulli experiments*[def. 34.21], where the probability of success of a single experiment is given by p:

$$\mathrm{p}(x) = \binom{n}{x}\mathrm{p}^x(1-\mathrm{p})^{n-x}$$

n :nb. of repetitions
x :nb. of successes
p :probability of success

$$\mathbb{E}[X]=n\mathrm{p} \quad (34.45) \qquad \mathbb{V}[X]=n\mathrm{p}(1-\mathrm{p}) \quad (34.46)$$

Note: Binomial Coefficient

The Binomial Coefficient corresponds to the permutation of two classes and not the variations as it seems from the formula.
Lets consider a box of n balls consisting of black and white balls. If we want to know the probability of drawing first x white and then $n-x$ black balls we can simply calculate:

$$\underbrace{(\mathrm{p}\cdots\mathrm{p})}_{x\text{-times}}\cdot\underbrace{(q\cdots q)}_{n-x\text{-times}} = \mathrm{p}^x q^{n-x}$$

4.4. Geometric Distribution $\mathrm{Geom}(\mathrm{p})$

Definition 34.26 Geometric Distribution $\mathrm{Geom}(\mathrm{p})$:
Models the probability of the number X of Bernoulli trials[def. 34.21] *until the first success*

$$\mathrm{p}(x) = \mathrm{p}(1-\mathrm{p})^{x-1}$$

x :nb. of repetitions *until first success*
p :success probability *of single Bernoulli experiment*

$$F(x) = \sum_{i=1}^x \mathrm{p}(1-\mathrm{p})^{i-1} \overset{\text{eq. (21.4)}}{=} 1-(1-\mathrm{p})^x$$

$$\mathbb{E}[X]=\frac{1}{\mathrm{p}} \quad (34.47) \qquad \mathbb{V}[X]=\frac{1-\mathrm{p}}{\mathrm{p}^2} \quad (34.48)$$

Notes

- $\mathbb{E}[X]$ is the mean waiting time until the first success
- the number of trials x in order to have at least one success with a probability of $\mathrm{p}(x)$:
$$x \geqslant \frac{\mathrm{p}(x)}{1-\mathrm{p}}$$
- $\log(1-\mathrm{p}) \approx -\mathrm{p}$ for small p

4.5. Poisson Distribution $\mathrm{Pois}(\lambda)$

Definition 34.27 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geqslant 0}$.
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.

$$\mathrm{p}(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \qquad \begin{matrix}\lambda>0\\ x\in\mathbb{Z}_{\geqslant 0}\end{matrix} \quad (34.49)$$

Event Rate λ: describes the average number of events in a single interval.

$$\mathbb{E}[X]=\lambda \quad (34.50) \qquad \mathbb{V}[X]=\lambda \quad (34.51)$$

Continuous Distributions

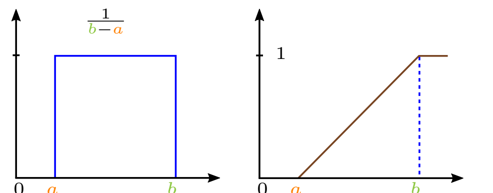
5.1. Uniform Distribution $\mathcal{U}(a,b)$

Definition 34.28 Uniform Distribution $\mathcal{U}(a,b)$:
Is probability distribution, where all intervals of the **same** length on the distribution's support[def. 34.6] $\mathrm{supp}(\mathcal{U}[a,b]) = [a,b]$ are equally probable/likely.

$$\mathrm{f}(x) = \frac{1}{b-a}\mathbb{1}_{x\in[a;b)} = \begin{cases}\frac{1}{b-a} = \text{const} & \text{if } a\leqslant x\leqslant b\\ 0 & \text{else}\end{cases} \quad (34.52)$$

$$F(x) = \begin{cases}0 & x<a\\ \frac{x-a}{b-a} & \text{if } a\leqslant x\leqslant b\\ 1 & x>b\end{cases} \quad (34.53)$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{V}(X) = \frac{(b-a)^2}{12} \quad (34.54)$$



(a) $f_X(x)$ (b) $F_X(x)$

5.2. Exponential Distribution $\exp(\lambda)$

Definition 34.29 Exponential Distribution $X \sim \exp(\lambda)$:
Is the continuous analogue to the geometric distribution [def. 34.26].
It describes the probability $\mathrm{f}(x;\lambda)$ that a continuous Poisson process (i.e., a process in which events occur continuously and independently at a constant average rate) will succeed/change state after a time interval x.

$$\mathrm{f}(x;\lambda) = \begin{cases}\lambda\mathrm{e}^{-\lambda x} & \text{if } x\geqslant 0\\ 0 & x<0\end{cases} \quad (34.55)$$

$$F(x;\lambda) = \begin{cases}1-\mathrm{e}^{-\lambda x} & \text{if } x\geqslant 0\\ 0 & x<0\end{cases} \quad (34.56)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathbb{V}(X) = \frac{1}{\lambda^2} \quad (34.57)$$

5.3. Laplace Distribution

Definition 34.30 Laplace Distribution:

Laplace Distibution $\qquad f(\mathbf{x};\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|\mathbf{x}-\mu|}{\sigma}\right) \quad (34.58)$

5.4. The Normal Distribution $\mathcal{N}(\mu,\sigma)$

Definition 34.31 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:

$$\mathbb{E}[X]=\mu \qquad \mathbb{V}(X)=\sigma^2 \quad (34.59)$$

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad (34.60)$$

$$F(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^x \exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\}\mathrm{d}u \quad (34.61)$$

$$x\in\mathbb{R} \quad \text{or} \quad -\infty<x<\infty$$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \quad (34.62)$$

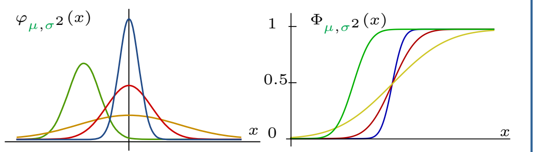


Figure 15: $\mu=0$, $\sigma^2=0.2$; $\mu=0$, $\sigma^2=1.0$; $\mu=0$, $\sigma^2=5.0$; $\mu=-2$, $\sigma^2=0.5$

Property 34.14: $\mathbb{P}_X(\mu-\sigma\leqslant x\leqslant\mu+\sigma) = 0.66$

Property 34.15: $\mathbb{P}_X(\mu-2\sigma\leqslant x\leqslant\mu+2\sigma) = 0.95$

5.5. The Standard Normal distribution $\mathcal{N}(0,1)$

Historic Problem: the cumulative distribution eq. (34.61) 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 34.32
Standard Normal Distribution $\mathbf{X}\sim\mathcal{N}(0,1)$:

$$\mathbb{E}[X]=0 \qquad \mathbb{V}(X)=1 \quad (34.63)$$

$$f(x;0,1) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2} \quad (34.64)$$

$$F(x;0,1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x \mathrm{e}^{-\frac{1}{2}u^2}\,\mathrm{d}u \quad (34.65)$$

$$x\in\mathbb{R} \quad \text{or} \quad -\infty<x<\infty$$

$$\psi_X(u) = \mathrm{e}^{\frac{u^2}{2}} \qquad \varphi_X(u) = \mathrm{e}^{-\frac{u^2}{2}} \quad (34.66)$$

Corollary 34.4
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. 34.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \qquad \alpha\in(0,1) \quad (34.67)$$

5.5.1. Calculating Probabilities

Property 34.16 Symmetry: Let $z>0$

$$\begin{aligned}
\mathbb{P}(Z\leqslant z) &= \Phi(z) && (34.68)\\
\mathbb{P}(Z\leqslant -z) &= \Phi(-z) = 1-\Phi(z) && (34.69)\\
\mathbb{P}(-a\leqslant Z\leqslant b) &= \Phi(b)-\Phi(-a) = \Phi(b)-(1-\Phi(a))\\
&\overset{a=b=z}{=} 2\Phi(z)-1 && (34.70)
\end{aligned}$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 34.1 Linear Transformation [proof 34.12]: 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:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \tag{34.71}$$

Proposition 34.2 Standardization [proof 34.13]: 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:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z = \frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0,1) \tag{34.72}$$

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 34.3 [proof 34.14]
Standardization of the CDF: Let $F_X(X)$ be the cumulative distribution function of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function $\Phi_Z(z)$ of the standardized random normal variable $Z \sim \mathcal{N}(0,1)$ is related to $F_X(X)$ by:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{34.73}$$

6. The Multivariate Normal distribution

Definition 34.33 Multivariate Normal/Gaussian: An $\mathbb{R}^n$-valued random variable $\mathbf{X} = (X_1 \quad \ldots, X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma: \quad \mathcal{L}\left(\sum_{i=1}^{n} \alpha_i X_j\right) = \mathcal{N}(\mu, \sigma^2) \quad \forall \alpha_i \in \mathbb{R} \tag{34.74}$$

(possible degenerated $\mathcal{N}(0,0)$ for $\forall \alpha_j = 0$)

Note
- **Joint** vs. **multivariate**: a joint normal distribution can be a multivariate normal distribution or a product of univariate normal distributions **but**
- Multivariate refers to the number of variables that are placed as inputs to a function.

Definition 34.34
Multivariate Normal distribution $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:
A k-dimensional random vector

$$\mathbf{X} = (X_1 \quad \ldots \quad X_n)^\intercal \quad \text{with} \quad \boldsymbol{\mu} = (\mathbb{E}[\mathbf{x}_1] \quad \ldots \quad \mathbb{E}[\mathbf{x}_k])^\intercal$$

and $k \times k$ **p.s.d.**covariance matrix:

$$\boldsymbol{\Sigma} := \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\intercal] = [\mathrm{Cov}[\mathbf{x}_i, \mathbf{x}_j], 1 \leqslant i, j \leqslant k]$$

follows a k-dim multivariate normal/Gaussian distribution if its law[def. 33.25] satisfies:

$$f_{\mathbf{X}}(X_1, \ldots, X_k) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{34.75}$$

$$= \frac{1}{\underbrace{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma})}}_{\text{Normalisation}}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})\right)$$

$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{i\mathbf{u}^\intercal \boldsymbol{\mu} - \frac{1}{2}\mathbf{u}\boldsymbol{\Sigma}\mathbf{u}\right\} \tag{34.76}$$

6.1. Joint Gaussian Distributions

Definition 34.35 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 $\mathbf{Z} = [X \quad Y]^\intercal$ is again a GRV.

Property 34.17 proof 34.16
Joint Independent Gaussian Random Variables: Let $X_1, \ldots, X_n$ be $\mathbb{R}$-valued *independent* random variables with laws $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Then the law of $\mathbf{X} = (X_1 \quad \ldots \quad X_n)$ is a (multivariate) Gaussian distribution $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{34.77}$$

Corollary 34.5 Quadratic Form:
If $\mathbf{x}$ and $\mathbf{y}$ are both independent GRVs

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

then they are jointly Gaussian[def. 34.35] given by:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{34.78}$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\mathbf{x} - \boldsymbol{\mu}_x)^\intercal \boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) + (\mathbf{y} - \boldsymbol{\mu}_y)^\intercal \boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_x)^\intercal \quad (\mathbf{y} - \boldsymbol{\mu}_y)^\intercal]\begin{bmatrix} \boldsymbol{\Sigma}_x^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_y^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{x} - \boldsymbol{\mu}_x \\ \mathbf{y} - \boldsymbol{\mu}_y \end{bmatrix}\right)$$

$$\mathrel{\hat=} \exp -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^\intercal \boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z)$$

Property 34.18
Marginal Distribution of Multivariate Gaussian: Let $\mathbf{X} = (X_1 \quad \ldots \quad X_n)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a an $\mathbb{R}^n$ valued Gaussian and let $V = \{1, 2, \ldots, n\}$ be the index set of its variables. The k-variate marginal distribution of the Gaussian indexed by a subset of the variables:

$$A = \{i_1, \ldots, i_k\} \qquad i_j \in V \tag{34.79}$$

is given by:

$$\mathbf{X} = \left(X_{i_1} \quad \ldots \quad X_{i_k}\right)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA}) \tag{34.80}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{i_1,i_1} & \cdots & \sigma^2_{i_1,i_k} \\ \vdots & \ddots & \vdots \\ \sigma^2_{i_k,i_1} & \cdots & \sigma^2_{i_k,i_k} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

6.2. Conditional Gaussian Distributions

Property 34.19 Conditional Gaussian Distribution: Let $\mathbf{X} = (X_1 \quad \ldots \quad X_n)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a an $\mathbb{R}^n$ valued Gaussian and let $V = \{1, 2, \ldots, n\}$ be the index set of its variables. Suppose we take two disjoint subsets of V:

$$A = \{i_1, \ldots, i_k\} \qquad B = \{j_1, \ldots, j_m\} \qquad i_l, j_{l'} \in V$$

then the conditional distribution of the random vector $\mathbf{X}_A$, conditioned on $\mathbf{X}_B$ given by $p(\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B)$ is:

$$\mathbf{X}_A = \left(X_{i_1} \quad \ldots \quad X_{i_k}\right)^\intercal \sim \mathcal{N}\left(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B}\right) \tag{34.81}$$

$$\boxed{\begin{aligned} \boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B) \\ \boldsymbol{\Sigma}_{A|B} &= \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA} \end{aligned}}$$

Note
Can be proofed using the matrix inversion lemma but is a very tedious computation.

maybe add sometime

Corollary 34.6
Conditional Distribution of Joint Gaussian's: Let $\mathbf{X}$ and $\mathbf{Y}$ be jointly Gaussian random vectors:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\intercal & \mathbf{B} \end{bmatrix}\right) \tag{34.82}$$

then the *marginal* distribution of $\mathbf{x}$ conditioned on $\mathbf{y}$ can be written as:

$$X \sim \mathcal{N}\left(\boldsymbol{\mu}_{X|Y}, \boldsymbol{\Sigma}_{X|Y}\right)$$

$$\boxed{\begin{aligned} \boldsymbol{\mu}_{X|Y} &= \boldsymbol{\mu}_X + \mathbf{C}\mathbf{B}^{-1}(\mathbf{y} - \boldsymbol{\mu}_Y) \\ \boldsymbol{\Sigma}_{X|Y} &= \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\intercal \end{aligned}} \tag{34.83}$$

add proofs

6.3. Transformations

Property 34.20 Multiples of Gaussian's $\mathbf{A}X$: Let $\mathbf{X} = (X_1 \quad \ldots \quad X_n)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a an $\mathbb{R}^n$ valued Gaussian and let $\mathbf{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$Y = \mathbf{A}\mathbf{X} \in \mathbb{R} \qquad Y \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\intercal\right) \tag{34.84}$$

Property 34.21 Affine Transformation of GRVs: Let $\mathbf{y} \in \mathbb{R}^n$ be GRV, $\mathbf{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$ and let $\mathbf{x}$ be defined by the affine transformation[def. 27.43]:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + b \qquad \mathbf{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$$

Then $\mathbf{x}$ is a GRV (see ?? 34.15).

Property 34.22 Linear Combination of jointly GRVs: **Let** $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ two jointly GRVs, and let $\mathbf{z}$ be defined as:

$$\mathbf{z} = \mathbf{A}_x\mathbf{x} + \mathbf{A}_y\mathbf{y} \qquad \mathbf{A}_x \in \mathbb{R}^{d \times n}, \mathbf{A}_x \in \mathbb{R}^{d \times m}$$

Then $\mathbf{z}$ is GRV (see ?? 34.17).

Definition 34.36 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 34.37 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:

$$f(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if} \quad x > 0 \\ 0 & \phantom{\text{if}} \quad x \leqslant 0 \end{cases} \tag{34.85}$$

$$\Gamma(\alpha) \overset{\text{eq. (24.80)}}{=} \int_0^\infty t^{\alpha-1} e^{-t} \, \mathrm{d}t \tag{34.86}$$

with $\alpha, \beta \in \mathbb{R}_{>0}$

6.5. Chi-Square Distribution χ_k^2
6.6. Student's t-distribution

Definition 34.38 Student' t-distribution:

add

6.7. Delta Distribution

Definition 34.39 The delta function $\delta(\mathbf{x})$:
The delta/dirac function $\delta(\mathbf{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x} = f(0)$$

for any integrable function f on $\mathbb{R}$.
Or alternativly by:

$$\delta(x - x_0) = \lim_{\sigma \to 0} \mathcal{N}(x | x_0, \sigma) \tag{34.87}$$

$$\approx \infty \mathbb{1}_{\{x = x_0\}} \tag{34.88}$$

Property 34.23 Properties of δ:
- **Normalization**: The delta function integrates to 1:
$$\int_{\mathbb{R}} \delta(x) \, \mathrm{d}x = \int_{\mathbb{R}} \delta(x) \cdot c_1(x) \, \mathrm{d}x = c_1(0) = 1$$
where $c_1(x) = 1$ is the constant function of value 1.
- **Shifting**:
$$\int_{\mathbb{R}} \delta(x - x_0) f(x) \, \mathrm{d}x = f(x_0) \tag{34.89}$$
- **Symmetry**: $\int_{\mathbb{R}} \delta(-x) f(x) \, \mathrm{d}x = f(0)$
- **Scaling**: $\int_{\mathbb{R}} \delta(\alpha x) f(x) \, \mathrm{d}x = \frac{1}{|\alpha|} f(0)$

Note
- In mathematical terms δ is not a function but a **gernalized function**.
- We may regard $\delta(x - x_0)$ as a density with all its probability mass centered at the signle point x_0.
- Using a box/indicator function s.t. its surface is one and its width goes to zero, instead of a normaldistribution eq. (34.87) would be a non-differentiable/discret form of the dirac measure.

Definition 34.40 Heaviside Step Function:

$$H(x) := \frac{\mathrm{d}}{\mathrm{d}x} \max\{x, 0\} \qquad x \in \mathbb{R}_{\neq 0} \tag{34.90}$$

or alternatively:

$$H(x) := \int_{-\infty}^{x} \delta(s) \, \mathrm{d}s \tag{34.91}$$

Proofs

Proof 34.11 Definition 34.25: Consider a sequence of n random $\{X_i\}_{i=1}^n$ Bernoulli experiments[def. 34.22] with success probability p.
Define the r.v. Y_n to be the sum of the n Bernoulli variables:

$$Y_n = \sum_{i=1}^{n} X_i \qquad n \in \mathbb{N}$$

i.e. the total number of successes. Now lets calculate the probability density function f_n of Y_n. First let $(x_1 \quad x_n) \in \{0,1\}^n$ and let $y = \sum_{i=1}^n x_i$ a bit sting of zeros and ones, with one occuring y times.

$$\mathbb{P}((X_1, X_2, \cdots, X_n) = (x_1, x_2, \ldots, x_n))$$

$$= \underbrace{(p \cdots p)}_{y} \cdot \underbrace{(q \cdots q)}_{n-y\text{-times}} = p^y(1-p)^{n-y}$$

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 $\frac{n!}{y!(n-y)!} = \binom{n}{y}$:

$$f_n(y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad y \in \{0, 1, \ldots, n\}$$

Proof 34.12: proposition 34.1: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_Y(y) \overset{y \geqslant 0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \leqslant \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

$$F_Y(y) \overset{y \leqslant 0}{=} \mathbb{P}_Y(Y \leqslant y) = \mathbb{P}(a + bX \leqslant y) = \mathbb{P}_X\left(X \geqslant \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right)$$

Differentiating both expressions w.r.t. y leads to:

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \begin{cases} \frac{1}{b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \\ \frac{1}{-b}\frac{\mathrm{d}F_X\left(\frac{y-a}{b}\right)}{\mathrm{d}y} \end{cases} = \frac{1}{|b|} f_X(x)\left(\frac{y-a}{b}\right)$$

eq. (34.71)).
in order to prove that $Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right)$ we simply plug f_X in the previous expression:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma|b|} \exp\left\{-\frac{1}{2}\left(\frac{\frac{y-a}{b} - \mu}{\sigma}\right)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma|b|} \exp\left\{-\frac{1}{2}\left(\frac{y - (a + b\mu)}{\sigma|b|}\right)^2\right\}$$

Proof 34.13: proposition 34.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$

$$\overset{\text{eq. (34.71)}}{\sim} \mathcal{N}\left(a\mu + b, a^2\sigma^2\right) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0,1)$$

Proof 34.14: proposition 34.3: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leqslant x) \overset{-\mu, \div\sigma}{=} \mathbb{P}\left(\frac{X - \mu}{\sigma} \leqslant \frac{x - \mu}{\sigma}\right) = \mathbb{P}\left(Z \leqslant \frac{x - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Proof 34.15: Property 34.21 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 $\mathbf{x} = ay + b$ $\quad a \in \mathbb{R}_+,\ b \in \mathbb{R}$

Using the Change of variables formula it follows:

$$p_x(\bar{x}) \overset{\text{eq. (33.46)}}{=} \frac{p_y(\bar{y})}{\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|} \qquad \left[\ \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = a\ \right]$$

$$\overset{\bar{y}=\frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2}\Big(\overbrace{\frac{\bar{x}-b}{a}}^{\bar{y}(\bar{x})} - \mu\Big)^2\right)$$

$$= \frac{1}{\sqrt{2\pi a^2\mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2}\big(\bar{x} \underbrace{-b - a\mu}_{\mu_x}\big)^2\right)$$

Hence $\quad x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

Note

We can also verify that we have calculated the right mean and variance by:

$$\mathbb{E}[x] = \mathbb{E}[ay+b] = a\mathbb{E}[y] + b = a\mu + b$$
$$\mathbb{V}[x] = \mathbb{V}[ay+b] = a^2\mathbb{V}[y] = a^2\sigma^2$$

Proof 34.16: **??**

$$p_{\mathbf{X}}(\mathbf{u}) = \prod_i^n p_{X_i}(u_i)$$
$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$
$$\varphi_{\mathbf{X}}(\mathbf{u}) = \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1 u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_n u_n^2\right\}$$
$$= \exp\left\{i\sum_i^n u_n\mu_n - \frac{1}{2}\sum_i^n \sigma_n u_n^2\right\} = \exp\left\{i\mathbf{u}^{\intercal}\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}\Sigma\mathbf{u}\right\}$$

Proof 34.17: Property 34.22

From Property 34.21 it follows immediately that $\mathbf{z}$ is GRV $\mathbf{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:

$$\mathbf{z} = \mathbf{A}\xi \quad \textbf{with} \quad \mathbf{A} = [\mathbf{A}_x \quad \mathbf{A}_y] \text{ and } \xi = (\mathbf{x} \ \ \mathbf{y})$$

Knowing that $\mathbf{z}$ is a GRV it is sufficient to calculate μ_z and Σ_z in order to characterize its distribution:

$$\mathbb{E}[\mathbf{z}] = \mathbb{E}[\mathbf{A}_x x + \mathbf{A}_y y] = \mathbf{A}_x\mu_x + \mathbf{A}_y\mu_y$$
$$\mathbb{V}[\mathbf{z}] = \mathbb{V}[\mathbf{A}\xi] \overset{??}{=} \mathbf{A}\mathbb{V}[\xi]\mathbf{A}^{\intercal}$$
$$= [\mathbf{A}_x \quad \mathbf{A}_y]\begin{bmatrix} \mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y] \end{bmatrix}[\mathbf{A}_x \quad \mathbf{A}_y]^{\intercal}$$
$$= [\mathbf{A}_x \quad \mathbf{A}_y]\begin{bmatrix} \mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y] \end{bmatrix}\begin{bmatrix} \mathbf{A}_x^{\intercal} \\ \mathbf{A}_y^{\intercal} \end{bmatrix}$$
$$= \mathbf{A}_x\mathbb{V}[x]\mathbf{A}_x^{\intercal} + \mathbf{A}_y\mathbb{V}[y]\mathbf{A}_y^{\intercal}$$
$$+ \underbrace{\mathbf{A}_y\mathrm{Cov}[y,x]\mathbf{A}_x^{\intercal}}_{=0\text{by independence}} + \underbrace{\mathbf{A}_x\mathrm{Cov}[x,y]\mathbf{A}_y^{\intercal}}_{=0\text{by independence}}$$
$$= \mathbf{A}_x\Sigma_x\mathbf{A}_x^{\intercal} + \mathbf{A}_y\Sigma_y\mathbf{A}_y^{\intercal}$$

Note

Can also be proofed by using the normal definition of [def. 34.15] and tedious computations.

Proof 34.18: Equation (34.43) If $\mathbf{x} = c_i$ i.e. the outcome c_i has occurred then it follows:

$$\prod_j^k p_i^{\delta[x=c_i]} = p_1^0 \cdots p_i^1 \cdots p_k^0 = 1 \cdots p_i \cdots 1 = p(\mathbf{x} = c_i|p)$$

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. (34.52)).
Furthermore repeated calls to these RNG are independent, that is:

$$\mathrm{p}_{U_1,U_2}(u_1,u_2) \overset{??}{=} \mathrm{p}_{U_1}(u_1)\cdot \mathrm{p}_{U_2}(u_2) = \begin{cases} 1 & \text{if } u_1,u_2 \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

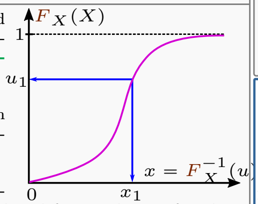
Question: using samples $\{u_1,\dots,u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discreet or continuous PDFs?

2. Inverse-transform Technique

Idea

Can make use of section 1 and the fact that CDF are increasing functions ([def. 24.15]). **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 35.1 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1,u_2,\dots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X:
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \tag{35.1}$$
2. Set $F_X(X) \overset{!}{=} U$ on the range of X with $U \sim \mathcal{U}[0,1]$.
3. Invert this equation/find the inverse $F_X^{-1}(U)$ i.e. solve:
$$U = F_X(X) = F_X\Big(\underbrace{F_X^{-1}(U)}_{X}\Big) \tag{35.2}$$
4. Plug in the uniformly distributed rn:
$$x_i = F_X^{-1}(u_i) \qquad \textbf{s.t.} \qquad x_i \sim f_X \tag{35.3}$$

Definition 35.2 Multiple Continuous Variable:
Given: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) \tag{35.4}$$
2. Use [def. 35.3] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 35.3] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof 35.1: [def. 35.3]:
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. 24.15]).
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:

$$F_X(a) = \mathbb{P}(x \leqslant a) = \mathbb{P}(F_X^{-1}(U) \leqslant a)$$

Since F_X is strictly increasing:

$$\mathbb{P}\left(F_X^{-1}(U) \leqslant a\right) = \mathbb{P}(U \leqslant F_X(a)) \overset{\text{eq. (34.52)}}{=} \int_0^{F_X(a)} 1\,\mathrm{d}t = F_X(a)$$

Note

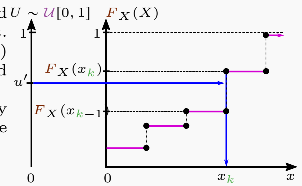
Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing ([def. 24.15]) we may always define an auxiliary function by its infimum:

$$\hat{F}_X^{-1} := \inf\{x | F_X(X) \geqslant 0\} \qquad u \in [0,1] \tag{35.5}$$

2.2. Discret Case

Idea

Given: a desired discret pmf p_X s.t. $\mathbb{P}(X = x_i) = p_X(x_i)$ and uniformly distributed rn $\{u_1,u_2,\dots\}$.
Goal: given a uniformly distributed rn u determine k s.t.:



$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff F_X(x_{k-1}) < u \leqslant F_X(x_k) \tag{35.6}$$

and return x_k.

Definition 35.3 One Discret Variable:

1. Compute the CDF of p_X ([def. 34.8])
$$F_X(x) = \sum_{t=-\infty}^{x} \mathrm{p}_X(t) \tag{35.7}$$
2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i ($\hat{=}$ inversion) s.t.:
$$F_X\left(x_{k(i)-1}\right) < u_i \leqslant F_X\left(x_{k(i)}\right) \qquad \forall u_i \tag{35.8}$$

Proof 35.2: **??**: First of all notice that we can always solve for an unique x_k.

Ask: why, are Discret CRV always strictly increasing/unique?

Given a fixed x_k determine the values of u for which:

$$F_X(x_{k-1}) < u \leqslant F_X(x_k) \tag{35.9}$$

Now observe that:

$$u \leqslant F_X(x_k) = F_X(x_{k-1}) + \mathrm{p}_X(x_k)$$
$$\Rightarrow F_X(x_{k-1}) < u \leqslant F_X(x_{k-1}) + \mathrm{p}_X(x_k)$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\mathbb{P}\left(U \in [F_X(x_{k-1}), F_X(x_k)]\right) = \int_{F_X(x_{k-1})}^{F_X(x_k)} \mathrm{p}_U(t)\,\mathrm{d}t$$
$$= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1\,\mathrm{d}t = \int_{F_X(x_{k-1})}^{F_X(x_{k-1})+\mathrm{p}_X(x_k)} 1\,\mathrm{d}t = \mathrm{p}_X(x_k)$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X.

Definition 35.4
Multiple Continuous Variables (Option 1):
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $\mathrm{p}_{X,Y}$:
$$\mathrm{p}_{X,Y} = \mathrm{p}_{X,Y}(x,y) = \mathrm{p}_{X|Y}(x|y)\mathrm{p}_Y(y) \tag{35.10}$$
2. Use **??** to first get a rv for y of $Y \sim \mathrm{p}_Y(y)$.
3. Then with this fixed y use **??** again to get a value for x of $X \sim \mathrm{p}_{X|Y}(x|y)$.

Definition 35.5
Multiple Continuous Variables (Option 2):
Note: this only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite.
Given: a pdf of multiple rvs $\mathrm{p}_{X,Y}$ **let** $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in $\mathcal{X}$ and $\mathcal{Y}$.

Define
$$\mathrm{p}_Z(1) = \mathrm{p}_{X,Y}(1,1), \mathrm{p}_Z(2) = \mathrm{p}_{X,Y}(1,2), \dots$$
$$\dots, \mathrm{p}_Z(N_x \cdot N_y) = \mathrm{p}_{X,Y}(N_x, N_y)$$

Then simply apply **??** to the auxillary pdf p_Z

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) \tag{35.11}$$
2. Use [def. 35.3] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 35.3] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

nice examples see comment in code text

3. Monte Carlo Methods

3.1. Monte Carlo (MC) Integration

Integration methods s.a. Simpson integration[def. 30.27] suffer heavily from the curse of dimensionality.
An n-order[def. 30.24] quadrature scheme $\mathcal{Q}_n$ in 1-dimension is usually of order n/d in d-dimensions.
Idea estimate an integral stochastically by drawing sample from some distribution.

Definition 35.6 Monte Carlo Integration:

$$3 + 4 \tag{35.12}$$

3.2. Rejection Sampling

3.3. Importance Sampling

Descriptive Statistics

1. Populations and Distributions

Definition 36.1 Population $\{x_i\}_{i=1}^N$:
is the entire set of entities from which we can draw sample.

Definition 36.2
Families of Probability Distributions p_θ:
Are probability distributions that vary only by a set of hyper parameters θ[def. 36.1].

Definition 36.3 [example 36.3]
Population/Statistical Parameter θ:
Are the parameters defining families of probability distributions[def. 36.2]

Explanation 36.1 (Definition 36.1). *Such hyper parameters are often characterized by populations following a certain family of distributions with the help of a stastistc. Hence they are called population or statistical parameters.*

1.1. Characteristics of Populations

Definition 36.4 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^N x_i \tag{36.1}$$

Definition 36.5 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2 \tag{36.2}$$

Note
The population variance and mean are equally to the mean derived from the true distribution of the population.

2. Sample Statistics

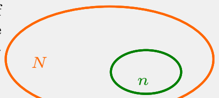
Definition 36.6 (Sample) Statistic: A statistc is a measuarble function T that assigns a **single** value t to a sample of random variables or population:

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

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

Definition 36.7 Degrees of freedom of a Statistic: Is the number of values in the final calculation of a statistic that are free to vary.

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



3. Point and Interval Estimation

Assume a population X with a given sample $\{x_i\}_{i=1}^n$ follows some family of distributions:

$$X \sim p_X(;\theta) \tag{36.3}$$

how can we estimate the correct value of the parameter θ or some function of that parameter $\tau(\theta)$?

3.1. Point Estimates

Definition 36.8 (Point) Estimator $\hat{\theta}$:
Is a statistic[def. 36.6] that tries estimates an unknown parameter θ of an underlying family of distributions[def. 36.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution:

$$\hat{\theta} = t(\mathbf{x}_1, \dots, \mathbf{x}_n) \tag{36.4}$$

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 36.9 Sample/Empirical Mean $\bar{x}$:
The sample mean is an estimate/statistic of the population mean[def. 36.4] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n x_i \tag{36.5}$$

Corollary 36.1 [proof 36.1]
Unbiased Sample Mean:
The sample mean estimator is unbiased:

$$\mathbb{E}[\hat{\mu}_X] = \mu \tag{36.6}$$

Corollary 36.2 [Proof 36.2]
Variance of the Sample Mean:
The variance of the sample mean estimator is given by:

$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n}\sigma_X^2 \tag{36.7}$$

3.1.2. Empirical Variance

Definition 36.10 Biased Sample Variance:
The sample variance is an estimate/statistic of the population variance[def. 36.5] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2 \tag{36.8}$$

Definition 36.11 [proof 36.3]
(Unbiased) Sample Variance:
The unbiased form of the sample variance[def. 36.10] is given by:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)^2 \tag{36.9}$$

Definition 36.12 Bessel's Correction: The factor

$$\frac{n}{n-1} \tag{36.10}$$

is called Bessel's correction. Multiplying the uncorrected population variance eq. (36.8) 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 36.13 Interval Estimator $\hat{\theta}$:
Is an estimator that tries to bound an unknown parameter θ of an underlying family of distributions[def. 36.2] for a given sample $\{\mathbf{x}_i\}_{i=1}^n$ of that distribution.
Let $\theta \in \Theta$ and define two point statistics[def. 36.6] g and h then an interval estimate is defined as:

$$\mathbb{P}(L_n < \theta < U_n) = \gamma \qquad \begin{matrix} \forall\theta \in \Theta & L_n = g(\mathbf{x}_1, \dots, \mathbf{x}_n) \\ \gamma \in [0,1] & U_n = h(\mathbf{x}_1, \dots, \mathbf{x}_n) \end{matrix} \tag{36.11}$$

Statistical Tests

4. Parametric Hypothesis Testing

Definition 36.14 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 36.15 Null Hypothesis H_0:
A null hypothesis H_0 is an *assumption* on a population[def. 36.1] parameter[def. 36.3] θ:

$$H_0 : \theta = \theta_0 \tag{36.12}$$

Note
Often, a null hypothesis cannot be verified, but can only be falsified.

Definition 36.16 Alternative Hypothesis H_A/H_1:
The alternative hypothesis H_1 is an *assumption* on a population[def. 36.1] parameter[def. 36.3] θ that is opposite to the null hypothesis.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{36.13}$$

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. 36.6] as this is easier to use.

Definition 36.17 [example 36.4]
Test Statistic/Testing Parameter T:
Is a sample statistic[def. 36.6] used for hypothesis tests in order to give evidence for or against a hypothesis:

$$t_n = T(D_n) = T(\{X_1, \dots, X_n\}) \tag{36.14}$$

4.3. Sampling Distribution

Definition 36.18 $\mathcal{T}_{\theta_0}(t)$
Null Distribution/Sampling Distribution under H_0:
Let $D_n = \{X_1, \dots X_n\}$ be a random sample from the true population p_{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*:

$$t \sim \mathcal{T}_{\theta_0} = \mathcal{T}(t|H_0 \text{ true}) \qquad X_i \sim p_{\text{pop}} \tag{36.15}$$

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 be the set of all possible values that the sample statistic T can map to. Now lets split $\mathcal{T}$ in two disjunct sets $\mathcal{T}_0$ and $\mathcal{T}_1$:

$$\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \qquad \mathcal{T}_0 \cap \mathcal{T}_1 = \emptyset$$

- 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 36.19 Critical/Rejection Region $\mathcal{T}_1$: Is the set of all values of the test statistic[def. 36.17] t_n that causes us to reject the Null Hypothesis in favor for the alternative hypothesis H_A:

$$K = \mathcal{T}_1 = \{\mathcal{T} : H_0 \text{ rejected}\} \tag{36.16}$$

Definition 36.20 Acceptance Region $\mathcal{T}_0$: Is the region where we accept the null hypothesis H_0.

$$\mathcal{T}_0 = \{\mathcal{T} : H_0 \text{ accepted}\} \tag{36.17}$$

Definition 36.21 Critical Value c:
Is the value of *the critical region* $c \in \mathcal{T}_1$ which is closest to the *region of acceptance*[def. 36.20]:

4.5. Type I&II Errors

Definition 36.22
False Positive **Type I Error:**
Is the rejection of the null hypothesis H_0, even-tough it is true

$$\begin{aligned} &\text{Test rejects } H_0 | H_0 \text{ true} \\ \iff\ & t_n \in \mathcal{T}_1 | H_0 \text{ true} \end{aligned} \tag{36.18}$$

Definition 36.23
False Negative **Type II Error:**
Is the acceptance of a null hypothesis H_0, even-tough its false:

$$\begin{aligned} &\text{Test accepts } H_0 | H_A \text{ true} \\ \iff\ & t_n \in \mathcal{T}_0 | H_A \text{ true} \end{aligned} \tag{36.19}$$

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 36.24 [example 36.5]
(Statistical) Significance α:
A study's defined significance level α denotes the probability to incur a *Type I Error*[def. 36.22]:

$$\mathbb{P}(t_n \in \mathcal{T}_1 | H_0 \text{ true}) = \mathbb{P}(\text{test rejects } H_0 | H_0 \text{ true}) \leqslant \alpha \tag{36.20}$$

Definition 36.25 Probability Type II Error β:
A test probability to for a *false negative*[def. 36.23] is defined as:

$$\beta(t_n) = \mathbb{P}(t_n \in \mathcal{T}_0 | H_1 \text{ true}) = \mathbb{P}(\text{test accepts } H_0 | H_1 \text{ true}) \tag{36.21}$$

Definition 36.26 (Statistical) Power $1-\beta$:
A study's power $1-\beta$ denotes a tests probability for a *true positive*:

$$\begin{aligned} 1 - \beta(t_n) &= \mathbb{P}(t_n \in \mathcal{T}_1 | H_1 \text{ true}) && (36.22) \\ &= \mathbb{P}(\text{test rejects } H_0 | H_1 \text{ true}) && (36.23) \end{aligned}$$

Corollary 36.3 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. 36.16]:

$$\begin{aligned} &\mathbb{P}(c_2 \leqslant X \leqslant c_1) \leqslant \alpha && \text{two-sided} \\ \text{or}\quad &\mathbb{P}(c_2 \leqslant X) \leqslant \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leqslant c_1) \leqslant \frac{\alpha}{2} \\ &\mathbb{P}(c_2 \leqslant X) \leqslant \alpha && \text{one-sided} \\ &\mathbb{P}(X \leqslant c_1) \leqslant \alpha && \text{one-sided} \end{aligned}$$

Decission \ Truth	H_0 true	H_0 false
H_0 accept	$1-\alpha$	$1-\beta$
H_0 rejected	α	β

4.7. P-Value

Definition 36.27 P-Value p:
Given a test statistic $t_n = T(X_1, \dots, X_n)$ the p-value $p \in [0,1]$ is the smallest value s.t. we reject the null hypothesis:

$$p := \inf_\alpha \{\alpha | t_n \in \mathcal{T}_1\} \qquad t_n = T(X_1, \dots, X_n) \tag{36.24}$$

Explanation 36.2.

- *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. 36.17] T.
2. Define the null hypothesis H_0 and the alternative hypothesis H_1 for T.
3. Find the sampling distribution[def. 36.18] $T_{\theta_0}(t)$ for T, given H_0 true.
4. Chose the significance level α
5. Evaluate the test statistic $t_n = T(X_1, \ldots, X_n)$ 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. (36.6) and (36.7) it follows that the *mean of the sample* is distributed as:

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

thus the mean of the sample $\overline{X}_n$ should equal the mean μ of the population. We now want to test the null hypothesis:

$$H_0 : \mu = \mu_0 \quad \iff \quad \overline{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n) \tag{36.25}$$

This is obviously only likely if the realization $\bar{x}_n$ is close to μ_0.

5.1.1. Z-Test $\qquad\sigma$ known

Definition 36.28 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. 36.24] α if:

$$\begin{aligned}
|z| \geqslant z_{1-\frac{\alpha}{2}} \quad &\iff z \leqslant z_{\frac{\alpha}{2}} \vee z \geqslant z_{1-\frac{\alpha}{2}} \\
&\iff z \in \mathcal{T}_1 = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}, \infty\right) \\
z \geqslant z_{1-\alpha} \quad &\iff z \in \mathcal{T}_1 = [z_{1-\alpha}, \infty) \\
z \leqslant z_{\alpha} = -z_{1-\alpha} \quad &\iff z \in \mathcal{T}_1 = (-\infty, -z_{\alpha}] = (\infty, -z_{1-\alpha}]
\end{aligned} \tag{36.26}$$

Notes

- Recall from [def. 34.19] and [cor. 34.4] that:
  $z_\alpha \overset{\text{i.e. } \alpha=0.05}{=} z_{0.05} = \Phi^{-1}(\alpha) \iff \mathbb{P}(Z \leqslant z_{0.05}) = 0.05$
- $|z| \geqslant z_{1-\frac{\alpha}{2}}$ which stands for:
  $$\mathbb{P}(Z \leqslant z_{0.05}) + \mathbb{P}(Z \geqslant z_{0.95}) = \mathbb{P}(Z \leqslant -z_{1-0.05}) + \mathbb{P}(Z \geqslant z_{0.95}) = \mathbb{P}(|Z| \geqslant z_{0.95})$$
  can be rewritten as:
  $$z \geqslant z_{1-\frac{\alpha}{2}} \quad \vee \quad -z \geqslant z_{1-\frac{\alpha}{2}} \iff z \leqslant -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$$
- One usually goes over to the standard normal distribution proposition 34.2 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 $\qquad\sigma$ 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:

$$T \sim t_{n-1} \tag{36.27}$$

Definition 36.29 t-Test:
For a realization of T with $\{x_i\}_{i=1}^n$ and mean $\bar{x}_n$:

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

we reject the null hypothesis $H_0 : \mu = \mu_0$ for the alternative H_A if:

$$\begin{aligned}
&|t| \geqslant t_{n-1,1-\frac{\alpha}{2}} \\
&\quad\iff t \in \mathcal{T}_1 = \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\
&t \geqslant t_{n-1,1-\alpha} \\
&\quad\iff t \in \mathcal{T}_1 = [t_{n-1,1-\alpha}, \infty) \\
&t \leqslant t_{n-1,\alpha} = -t_{n-1,1-\alpha} \\
&\quad\iff t \in \mathcal{T}_1 = (-\infty, -t_{n-1,\alpha}] = (\infty, -t_{n-1,1-\alpha}]
\end{aligned}$$

Notes

- The t-distribution has fatter tails as the normal distribution $\Rightarrow$ rare event become more likely
- For $n \to \infty$ the t-distribution goes over into the normal distribution
- The t-distribution gains a degree of foredoom for each sample and loses one for each parameter we are interested in $\Rightarrow$ 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. 36.19] namely the region of plaussible values.

Definition 36.30 Confidence Interval I:
Let $D_n = \{X_1, \ldots, X_n\}$ be a *sample* of observations and T_n a sample statistic of that sample. The confidence interval is defined as:

$$I(D_n) = \{\theta_0 : T_n(D_n) \in \mathcal{T}_0\} = \{\theta_0 : H_0 \text{ is not rejected}\} \tag{36.28}$$

Corollary 36.4 : The confidence interval captures the unkown parameter θ with probability $1-\alpha$:

$$\mathbb{P}_\theta\left(\theta \in I(D_n)\right) = \mathbb{P}\left(T_n(D_n) \in \mathcal{T}_0\right) = 1 - \alpha \tag{36.29}$$

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 36.1 ??: Let x be uniformly distributed on $[0,1]$ ([def. 34.28]) with pmf $p_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{p_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y\, p_y(y) \Rightarrow x = \int_{-\infty}^{y} p_y(t)\,\mathrm{d}t = F_Y(x)$$

Example 36.2 ??: Let

add https://www.youtube.com/watch?v=WUUb7VIRzgg

Example 36.3 Family of Distributions: The family of normal distribution $\mathcal{N}$ has two parameters $\left\{\mu, \sigma^2\right\}$

Example 36.4 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. 36.3] $\theta = \mu$ 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 36.5 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:

$$H_0 : p = p_0 = 0.1 \qquad \text{vs.} \qquad H_A : p > 0.1$$

We model the number of number of defective goods using the binomial distribution[def. 34.25]

$$\begin{aligned} X \sim &\mathcal{B}(n, p) \\ \sim &T(n, p) \end{aligned}, n = 20 \qquad \mathbb{P}(X \geqslant x) = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

from this we find:

$$\mathbb{P}_{p_0}(X \geqslant 4) = 1 - \mathbb{P}_{p_0}(X \leqslant 3) = 0.13$$
$$\mathbb{P}_{p_0}(X \geqslant 5) = 1 - \mathbb{P}_{p_0}(X \leqslant 4) = 0.04 \leqslant \alpha$$

thus the probability that equal 5 or more then 5 parts out of the 20 are rejects is less then 4%.
$\Rightarrow$ throw away null hypothesis for the 5% niveau in favor to the alternative.
$\Rightarrow$ the 5% significance niveau is given by $K = \{5, 6, \ldots, 20\}$

Note

If $x < n/2$ it is faster to calculate $\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X \leqslant x-1)$

8. Proofs

Proof 36.1: [cor. 36.1]:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}[\underbrace{\mu + \cdots + \mu}_{1,\ldots,n}]$$

Proof 36.2: [cor. 36.2]:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \overset{\text{Property 34.10}}{=} \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right]$$
$$\frac{1}{n^2} n \mathbb{V}[X] = \frac{1}{n}\sigma^2$$

Proof 36.3: definition 36.11:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_X^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\cdot n\bar{x} + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \left(\sigma^2 + n\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}$$

Stochastic Calculus

Stochastic Processes

Definition 37.1
Random/Stochastic Process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$:
An ($\mathbb{R}^d$-valued) stochastic process is a collection of ($\mathbb{R}^d$ *−valued*) random variables X_t on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The index set $\mathcal{T}$ is usually representing time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \ldots\}$. Therefore, the random process X can be written as a function:

$$X : \mathcal{T} \subseteq \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}^d \quad \iff \quad (t, \omega) \mapsto X(t, \omega) \tag{37.1}$$

Definition 37.2 Sample path/Trajector/Realization: Is the *stochastic/noise signal* $r(\cdot, \omega)$ on the index set[def. 21.1] $\mathcal{T}$, 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 37.1 $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\} > 0$
Strictly Positive Stochastic Processes: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called strictly positive if it satisfies:

$$X_t > 0 \qquad \mathbb{P}\text{-a.s.} \qquad \forall t \in \mathcal{T} \tag{37.2}$$

Definition 37.3
Random/Stochastic Chain $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$:
is a collection of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$[def. 33.1]. The random variables are ordered by an associated index set[def. 21.1] $\mathcal{T}$ and take values in the same mathematical *discrete state space*[def. 37.5] S, which must be measurable w.r.t. some σ-algebra[def. 33.6] Σ.
Therefore for a given probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space (S, Σ), the random *chain* X is a collection of S-valued random variables that can be written as:

$$X : \mathcal{T} \times \Omega \mapsto S \quad \iff \quad (t, \omega) \mapsto X(t, \omega) \tag{37.3}$$

Definition 37.4 Index/Parameter Set $\mathcal{T}$:
Usually represents time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \ldots\}$.

Definition 37.5 State Space S:
Is the range/possible values of the random variables of a stochastic process[def. 37.1] and must be measurable[def. 33.7] w.r.t. some σ-algebra Σ.

Sample-vs. State Space

Sample space[def. 33.2] 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 37.6 Sample path/Trajector/Realization: Is the *stochastic/noise signal* $r(\cdot, \omega)$ on the index set $\mathcal{T}$, that we obtain be sampling ω from Ω.

Notation

Even though the r.v. X is a function of two variables, most books omit the argument of the sample space $X(t, \omega) := X(t)$

1.1. Filtrations

Definition 37.7 Filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t\geqslant 0}$:
A collection $\{\mathcal{F}_t\}_{t\geqslant 0}$ of sub σ-*algebras*[def. 33.6] $\{\mathcal{F}_t\}_{t\geqslant 0} \in \mathcal{F}$ is called filtration if it is *increasing*:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \qquad \forall s \leqslant t \tag{37.4}$$

Explanation 37.1 (Definition 37.7). *A filtration describes the flow of information i.e. with time we learn more information.*

Definition 37.8
Filtered Probability Space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$:
A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$ is called a *filtered probability* space.

Definition 37.9 Adapted Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called adapted *to a* filtration $\mathbb{F}$ if:

$$X_t \text{ is } \mathcal{F}_t\text{-measurable} \qquad \forall t \tag{37.5}$$

That is the value of X_t is observable at time t

Definition 37.10 Predictable Process: A stochastic process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$ is called predictable *w.r.t. a* filtration $\mathbb{F}$ if:

$$X_t \text{ is } \mathcal{F}_{t-1}\text{-measurable} \qquad \forall t \tag{37.6}$$

That is the value of X_t is known at time $t-1$

Note

The price of a stock will usually be adapted since date k prices are known at date k.
On the other hand the interest rate of a bank account is usually already known at the beginning $k-1$, s.t. the interest rate r_t ought to be $\mathcal{F}_{k-1}$ measurable, i.e. the process $r = (r_k)_{k=1,\ldots,T}$ should be predictable.

Corollary 37.2 : The amount of information of an adapted random process is increasing see example 37.1.

2. Martingales

Definition 37.11 Martingales: A stochastic process $X(t)$ is a martingale on a *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$ if the following conditions hold:

1. Given $s \leqslant t$ the best prediction of $X(t)$, with a filtration $\{\mathcal{F}_s\}$ is the current expected value:
$$\forall s \leqslant t \qquad \mathbb{E}[X(t)|\mathcal{F}_s] = X(s) \quad \text{a.s.} \tag{37.7}$$
2. The expectation is finite:
$$\mathbb{E}[|X(t)|] < \infty \quad \forall t \geqslant 0 \quad X(t) \text{ is } \{\mathcal{F}_t\}_{t\geqslant 0} \text{ adapted} \tag{37.8}$$

Interpretation

- For any $\mathcal{F}_s$-adapted process the best prediction of $X(t)$ is the currently known value $X(s)$ i.e. if $\mathcal{F}_s = \mathcal{F}_{t-1}$ then the best prediction is $X(t-1)$
- A martingale models fair games of limited information.

Definition 37.12 Auto Covariance $\gamma(t_2 - t_1)$:
Describes the covariance[def. 34.16] between two values of a stochastic process $(\mathbf{X}_t)_{t\in\mathcal{T}}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \mathrm{Cov}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right] = \mathbb{E}\left[\left(\mathbf{X}_{t_1} - \mu_{t_1}\right)\left(\mathbf{X}_{t_2} - \mu_{t_2}\right)\right] \tag{37.9}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \mathrm{Cov}\left[\mathbf{X}_t, \mathbf{X}_t\right] \overset{\text{eq. (34.35)}}{=} \mathbb{V}\left[\mathbf{X}_t\right] \tag{37.10}$$

Notes

- **Hence** the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- **Given** a random time dependent variable $\mathbf{x}(t)$ the autocorrelation function $\gamma(t, t-\tau)$ describes how *similar* the time translated function $\mathbf{x}(t-\tau)$ and the original function $\mathbf{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random, then the autocorrelation is non-zero.
- The auto covariance is maximized/most similar for no translation $\tau = 0$ at all.

Definition 37.13 Auto Correlation $\rho(t_2 - t_1)$:
Is the scaled version of the auto-covariance[def. 37.12]:

$$\rho(t_2 - t_1) = \mathrm{Corr}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right] \tag{37.11}$$
$$= \frac{\mathrm{Cov}\left[\mathbf{X}_{t_1}, \mathbf{X}_{t_2}\right]}{\sigma_{X_{t_1}}\sigma_{X_{t_2}}} = \frac{\mathbb{E}\left[\left(\mathbf{X}_{t_1} - \mu_{t_1}\right)\left(\mathbf{X}_{t_2} - \mu_{t_2}\right)\right]}{\sigma_{X_{t_1}}\sigma_{X_{t_2}}}$$

3. Different kinds of Processes

3.1. Markov Process

Definition 37.14 Markov Process: A continuous-time stochastic process $X(t), t \in T$, is called a Markov process if for any finite parameter set $\{t_i : t_i < t_{i+1}\} \in T$ it holds:

$$\mathbb{P}\left(X(t_{n+1}) \in B | X(t_1), \ldots, X(t_n)\right) = \mathbb{P}\left(X(t_{n+1}) \in B | X(t_n)\right)$$

it thus follows for the *transition probability* – the probability of $X(t)$ lying in the set B at time t, given the value x of the process at time s:

$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \qquad 0 \leqslant s < t \tag{37.12}$$

Interpretation

In order to predict the future only the current/last value counts.

Corollary 37.3 Transition Density: The transition probability of a continuous distribution p can be calculated via:

$$\mathbb{P}(s, x, t, B) = \int_B \mathrm{p}(s, x, t, y)\, \mathrm{d}y \tag{37.13}$$

3.2. Gaussian Process

Definition 37.15 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:

$$X(t) \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right) \qquad \forall t \in T \tag{37.14}$$

3.3. Diffusions

Definition 37.16 [proof 37.1],[proof 37.2]
Diffusion:
Is a Markov Process[def. 37.14] for which it holds that:

$$\mu(t, X(t)) = \lim_{t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[X(t + \Delta t) - X(t) | X(t)\right] \tag{37.15}$$
$$\sigma^2(t, X(t)) = \lim_{t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[\left(X(t + \Delta t) - X(t)\right)^2 | X(t)\right] \tag{37.16}$$

- $\mu(t, X(t))$ 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 37.17
d-dim **standard Brownian Motion/Wiener Process:**
Is an $\mathbb{R}^d$ valued *stochastic process*[def. 37.1] $(W_t)_{t\in\mathcal{T}}$ starting at $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies:

1. **Normal Independent Increments**: the increments are *normally distributed independent random variables*:
$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}\left(0, (t_i - t_{i-1})\mathbb{1}_{d\times d}\right)$$
$$\forall i \in \{1, \ldots, T\} \tag{37.17}$$
2. **Stationary increments**:
$W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$
3. **Continuity**: for *a.e.* $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous
$$\lim_{t\to 0} \frac{\mathbb{P}\left(|W(t + \Delta t) - W(t)| \geqslant \delta\right)}{\Delta t} = 0 \qquad \forall \delta > 0 \tag{37.18}$$
4. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{37.19}$$

Notation

- In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wiener process.
- **However** in some sources the Wiener process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 37.4 $W_t \sim \mathcal{N}(0, \sigma)$ [proof 37.4],[proof 37.5]:
The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law

$$\mathbb{E}\left[W(t)\right] = \mu = 0 \tag{37.20}$$
$$\mathbb{V}\left[W(t)\right] = \mathbb{E}\left[W^2(t)\right] = \sigma^2 = t \tag{37.21}$$

3.4.1. Properties of the Wiener Process

Property 37.1 Non-Differentiable Trajectories:
The sample paths of a Brownian motion are not differentiable:

$$\frac{\mathrm{d}W(t)}{t} = \lim_{t\to 0}\mathbb{E}\left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t}\right)^2\right]$$
$$= \lim_{t\to 0}\frac{\mathbb{E}\left[W(t + \Delta t) - W(t)\right]}{\Delta t} = \lim_{t\to 0}\frac{\sigma^2}{\Delta t} = \infty$$

$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
$\xrightarrow{\text{solution}}$ Ito Calculus see section 38.

Property 37.2 Auto covariance Function:
The auto-covariance[def. 37.12] for a Wiener process

$$\mathbb{E}\left[(W(t) - \mu t)(W(t') - \mu t')\right] = \min(t, t') \tag{37.22}$$

Property 37.3: A standard Brownian motion is a

Quadratic Variation

Definition 37.18 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$LV_{[a,b]}(f) = \sup_{\Pi\in\mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)| \tag{37.23}$$
$$\mathcal{S} = \left\{\Pi\{x_0, \ldots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{\text{[def. 30.8]}} \text{ of } [a, b]\right\}$$

it is a measure of the (one dimensional) length of a function w.r.t. to the y-axis, when moving alone the function.
Hence it is a measure of the variation of a function w.r.t. to the y-axis.

Definition 37.19
Total Quadratic Variation/"sum of squares":
The total quadratic variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$QV_{[a,b]}(f) = \sup_{\Pi\in\mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)|^2 \tag{37.24}$$
$$\mathcal{S} = \left\{\Pi\{x_0, \ldots, x_{n_\Pi}\} : \Pi \text{ is a partition }^{\text{[def. 30.8]}} \text{ of } [a, b]\right\}$$

Corollary 37.5 Bounded (quadratic) Variation:
The (quadratic) variation[def. 37.18] of a function is bounded if it is finite:

$$\exists M \in \mathbb{R}_+ : \quad LV_{[a,b]}(f) \leqslant M \quad \left(QV_{[a,b]}(f) \leqslant M\right) \quad \forall \Pi \in \mathcal{S} \tag{37.25}$$

Theorem 37.1 Variation of Wiener Process: Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:

$$\mathbb{P}\left(\omega : LV(W(\omega)) < \infty\right) = 0 \tag{37.26}$$

Theorem 37.2 [proof 37.6]
Quadratic Variation of standard Brownian Motion:
The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:

$$\lim_{N\to\infty} \sum_{k=1}^{N} \left[W\left(k\frac{T}{N}\right) - W\left((k-1)\frac{T}{N}\right)\right]^2 = T$$
$$\text{with probability } 1 \tag{37.27}$$

Corollary 37.6 : theorem 37.2 can also be written as:

$$(\mathrm{d}W(t))^2 = \mathrm{d}t \tag{37.28}$$

3.4.2. Lévy's Characterization of BM

Theorem 37.3 [proof 37.7],[proof 37.8]
d-dim standard BM/Wiener Process by Paul Lévy:
An $\mathbb{R}^d$ valued *adapted stochastic process*[def's. 37.1, 37.7] $(W_t)_{t\in\mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{R}_+}$, that satisfies:

1. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{37.29}$$
2. **Continuous Martingale:** W_t is an a.s. *continuous* martingale[def. 37.11] w.r.t. the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ under $\mathbb{P}$.
3. **Quadratic Variation:**
$$W_t^2 - t \text{ is also an martingale} \iff QV(W_t) = t \tag{37.30}$$

is a standard Brownian motion[def. 37.24].

Further Stochastic Processes

3.4.3. White Noise

undestand script and add

Definition 37.20 Discrete-time white noise: Is a random signal $\{\epsilon_t\}_{t\in T_{\text{discrete}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * [k]] = 0 \qquad \forall k \in T_{\text{discret}} \tag{37.31}$$
- Zero autocorrelation[def. 37.13] γ i.e. the signals of different times are in no-way correlated:
$$\begin{aligned}\gamma(\epsilon * [k], \epsilon * [k+n]) &= \mathbb{E}\left[\epsilon * [k]\epsilon * [k+n]^\intercal\right]\\ &= \mathbb{V}[\epsilon * [k]]\,\delta_{\text{discret}}[n]\\ &\forall k, n \in T_{\text{discret}}\end{aligned} \tag{37.32}$$

With $\qquad \delta_{\text{discret}}[n] := \begin{cases}1 & \text{if } n = 0\\ 0 & \text{else}\end{cases}$

See proofs

Definition 37.21 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):
$$\mathbb{E}[\epsilon * (t)] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{37.33}$$
- Zero autocorrelation[def. 37.13] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * (t), \epsilon * (t+\tau)) = \mathbb{E}\left[\epsilon * (t)\epsilon * (t+\tau)^\intercal\right] \tag{37.34}$$
$$\overset{\text{eq. (34.88)}}{=} \mathbb{V}[\epsilon * (t)]\,\delta(t-\tau) = \begin{cases}\mathbb{V}[\epsilon * (t)] & \text{if } \tau = 0\\ 0 & \text{else}\end{cases}$$
$$\forall t, \tau \in T_{\text{continuous}} \tag{37.35}$$

Definition 37.22 Homoscedastic Noise: Has constant variability for all observations/time-steps:
$$\mathbb{V}[\epsilon_{i,t}] = \sigma^2 \qquad \begin{aligned}\forall t &= 1,\dots,T\\ \forall i &= 1,\dots,N\end{aligned} \tag{37.36}$$

Definition 37.23 Heteroscedastic Noise: Is noise whose variability may vary with each observation/time-step:
$$\mathbb{V}[\epsilon_{i,t}] = \sigma(i,t)^2 \qquad \begin{aligned}\forall t &= 1,\dots,T\\ \forall i &= 1,\dots,N\end{aligned} \tag{37.37}$$

3.4.4. Generalized Brownian Motion

Definition 37.24 Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 37.17], and define:
$$X_t = \mu t + \sigma W_t \qquad t \in \mathbb{R}_+ \qquad \begin{aligned}\mu &\in \mathbb{R} &&: \text{drift parameter}\\ \sigma &\in \mathbb{R}_+ &&: \text{scale parameter}\end{aligned} \tag{37.38}$$
then $\{X_t\}_{t\in\mathbb{R}_+}$ is normally distributed with mean μt and variance $t\sigma^2$ $X_t \sim \mathcal{N}\left(\mu t, \sigma^2 t\right)$.

Theorem 37.4 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:
$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}}\exp\left\{-\frac{(x-\mu t)^2}{2\sigma^2 t}\right\} \tag{37.39}$$

Corollary 37.7 : More generally we may define the process:
$$t \mapsto f(t) + \sigma W_t \tag{37.40}$$
which corresponds to a noisy version of f.

Corollary 37.8
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:
$$\mathrm{d}X(t) = \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) \tag{37.41}$$
$$X(0) = 0 \tag{37.42}$$

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 37.25 Geometric Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 37.17] the stochastic process $\mathbf{S}_t^1 \triangleq \mathbf{S}^1(t)$ with drift parameter μ and scale σ satisfying the SDE:
$$\begin{aligned}\mathrm{d}\mathbf{S}_t^1 &= \mathbf{S}_t^1(\mu\,\mathrm{d}t + \sigma\,\mathrm{d}W_t)\\ &= \mu\mathbf{S}_t^1\,\mathrm{d}t + \sigma\mathbf{S}_t^1\,\mathrm{d}W_t\end{aligned} \tag{37.43}$$
is called geometric Brownian motion and is given by:
$$\mathbf{S}_t^1 = \mathbf{S}_0^1\exp\left(\sigma W_t + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) \qquad t \in \mathbb{R}_+ \tag{37.44}$$

Corollary 37.9 Log-normal Returns:
For a geometric BM we obtain log-normal returns:
$$\ln\left(\frac{S_t}{S_0}\right) = \bar{\mu}t + \sigma W(t) \iff \bar{\mu}t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t)$$
$$\text{with} \qquad \bar{\mu} := \mu - \frac{1}{2}\sigma^2 \tag{37.45}$$

3.4.6. Locally Brownian Motion

Definition 37.26 Locally Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 37.17] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:
$$\mathrm{d}X(t) = \mu(X(t), t)\,\mathrm{d}t + \sigma(X(t), t)\,\mathrm{d}W(t) \tag{37.46}$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

3.4.7. Ornstein-Uhlenbeck Process

Definition 37.27 Ornstein-Uhlenbeck Process:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 37.17] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:
$$\mathrm{d}X(t) = -aX(t)\,\mathrm{d}t + b\sigma\,\mathrm{d}W(t) \qquad a > 0 \tag{37.47}$$

3.5. Poisson Processes

Definition 37.28 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 37.29 Poisson Process: A Poisson Process with *rate* $\lambda \in \mathbb{R}_{\geqslant 0}$ is a collection of random variables $X(t)$, $t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$, having a discrete *state space* $N = \{0, 1, 2, \dots\}$ and satisfies:

1. $X_0 = 0$
2. The increments follow a Poisson distribution[def. 34.27]:
$$\mathbb{P}((X_t - X_s) = k) = \frac{\lambda(t-s)}{k!}e^{-\lambda(t-s)} \qquad \begin{aligned}0 \leqslant s < t < \infty\\ \forall k \in \mathbb{N}\end{aligned}$$
3. No correlation of (non-overlapping) increments:
$\forall t_0 < t_1 < \dots < t_n$: the increments are independent
$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}} \tag{37.48}$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete, positive* realizations in $\in \mathbb{N}_{\geqslant 0}$

Corollary 37.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:
$$\mathbb{P}\left(X_{(t+\Delta t)} - X_t \neq 0\right) = \lambda\Delta t + o(\Delta t^2) \qquad t \text{ small i.e. } t \to 0 \tag{37.49}$$

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 37.30 Differential of a Poisson Process: The differential of a Poisson Process is defined as:
$$\mathrm{d}X_t = \lim_{\Delta t \to \mathrm{d}t}\left(X_{(t+\Delta t)} - X_t\right) \tag{37.50}$$

Property 37.4 Probability of Events for differential: With the definition of the differential and using the previous results from the Taylor expansion it follows:
$$\mathbb{P}(\mathrm{d}X_t = 0) = 1 - \lambda \tag{37.51}$$
$$\mathbb{P}(|\mathrm{d}X_t| = 1) = \lambda \tag{37.52}$$

Proofs

Proof 37.1: eq. (37.15):
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:
$$\begin{aligned}\mathbb{E}[x(n)] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N x_i(n)\right] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta]\\ &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}[x_i(n-1)]\\ &\overset{\text{induction}}{=} \mathbb{E}[x_{n-1}] = \dots\mathbb{E}[x(0)] = 0\end{aligned}$$
Thus in expectation the particles goes nowhere.

Proof 37.2: eq. (37.16):
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:
$$\begin{aligned}\mathbb{E}\left[x(n)^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N x_i(n)^2\right] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[x_i(n-1) \pm \delta]^2\\ &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2\right]\\ &\overset{\text{ind.}}{=} \mathbb{E}\left[x_{n-1}^2\right] + \delta^2 = \mathbb{E}\left[x_{n-2}^2\right] + 2\delta^2 = \dots\\ &= \mathbb{E}[x(0)] + n\delta^2 = n\delta^2\end{aligned}$$
as $n = \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x}$ it follows:
$$\sigma^2 = \mathbb{E}\left[x^2(n)\right] - \mathbb{E}[x(n)]^2 = \mathbb{E}\left[x^2(n)\right] = \frac{\delta^2}{\Delta x}t \tag{37.53}$$
Thus in expectation the particles goes nowhere.

Proof 37.3: eq. (37.34):
$$\begin{aligned}\gamma(\epsilon * [k], \epsilon * [k+n]) &= \operatorname{Cov}[\epsilon * [k], \epsilon * [k+1]]\\ &= \mathbb{E}\left[(\epsilon * [k] - \mathbb{E}[\epsilon * [k]])\,(\epsilon * [k+n] - \mathbb{E}[\epsilon * [k+n]])^\intercal\right]\\ &\overset{\text{eq. (37.31)}}{=} \mathbb{E}[(\epsilon * [k])\,(\epsilon * [k+n])]\end{aligned}$$

Proof 37.4: [cor. 37.4]:
Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the increement over the interval $[s-s, t-s] = [0, t-s]$
$$\text{Thus} \qquad B_t - B_s \sim B_{t-s} - B_0$$
but as B_0 is a.s. zero by definition eq. (37.19) it follows:
$$B_t - B_s \sim B_{t-s} \qquad B_{t-s} \sim \mathcal{N}(0, t-s)$$

Proof 37.5: [cor. 37.4]:
$$W(t) = W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t)$$
$$\Rightarrow \qquad \mathbb{E}[X] = 0 \qquad \mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = t$$

Proof 37.6: theorem 37.2:
$$\begin{aligned}&\sum_{k=0}^{N-1}[W(t_k) - W(t_{k-1})]^2 && t_k = k\frac{T}{N}\\ &= \sum_{k=0}^{N-1} X_k^2 && X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right)\\ &= \sum_{k=0}^{N-1} Y_k = n\left(\frac{1}{n}\sum_{k=0}^{N-1} Y_k\right) && \mathbb{E}[Y_k] = \frac{T}{N}\\ &\overset{\text{S.L.L.N}}{=} n\frac{T}{n} = T\end{aligned}$$

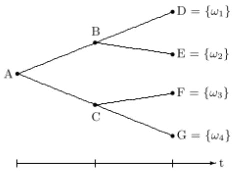
Proof 37.7: theorem 37.3 2:
1. first we need to show eq. (37.7): $\qquad \mathbb{E}[W_t|\mathcal{F}_s] = W_s$
Due to the fact that W_t is $\mathcal{F}_t$ measurable i.e. $W_t \in \mathcal{F}_t$ we know that:
$$\mathbb{E}[W_t|\mathcal{F}_t] = W_t \tag{37.54}$$
$$\begin{aligned}\mathbb{E}[W_t|\mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s|\mathcal{F}]\\ &= \mathbb{E}[W_t - W_s|\mathcal{F}_s] + \mathbb{E}[W_s|\mathcal{F}_s]\\ &\overset{\text{eq. (37.54)}}{=} \mathbb{E}[W_t - W_s] + W_s\\ &\overset{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s\end{aligned}$$
2. second we need to show eq. (37.8): $\qquad \mathbb{E}[|X(t)|] < \infty$
$$\mathbb{E}[|W(t)|]^2 \overset{??}{\leqslant} \mathbb{E}\left[|W(t)|^2\right] = \mathbb{E}\left[W^2(t)\right] = t \leqslant \infty$$

Proof 37.8: theorem 37.3 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:
$$\begin{aligned}\mathbb{E}\left[W_t^2|\mathcal{F}_s\right] &= \mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}_s\right] + \mathbb{E}\left[2W_s(W_t - W_s)|\mathcal{F}_s\right]\\ &\quad + \mathbb{E}\left[W_s^2|\mathcal{F}_s\right]\\ &\overset{\text{eq. (37.54)}}{=} \mathbb{E}\left[(W_t - W_s)^2\right] + 2W_s\mathbb{E}[(W_t - W_s)] + W_s^2\\ &\overset{\text{eq. (37.21)}}{=} \mathbb{V}[W_t - W_s] + 0 + W_s^2\\ &\quad t - s + W_s^2\end{aligned}$$
from this it follows that:
$$\mathbb{E}\left[W_t^2 - t|\mathcal{F}_s\right] = W_s^2 - s \tag{37.55}$$

understand why $\mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}\right] = \mathbb{E}\left[(W_t - W_s)^2\right]$

Examples

Example 37.1 :



Suppose we have a sample space of four elements: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. At time zero, we do not have any information about which ω has been chosen. At time $T/2$ we know whether we have $\{\omega_1, \omega_2\}$ or $\{\omega_3, \omega_4\}$. At time T, we have full information.

$$\mathcal{F} = \begin{cases} \{\varnothing, \Omega\} & t \in [0, T/2) \\ \{\varnothing, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T) \\ \mathcal{F}_{\max} = 2^{\Omega} & t = T \end{cases} \tag{37.56}$$

Thus, $\mathcal{F}_0$ represents initial information whereas $\mathcal{F}_\infty$ represents full information (all we will ever know). Hence, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geqslant 0}, \mathbb{P})$.

Ito Calculus