Probabilistic Artificial Intelligence
Markov Decision Processes
Active Learning

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

Definition 2.1 Active Learning: Is to actively choose the most information samples in order to reduce the amount of samples we need to label.

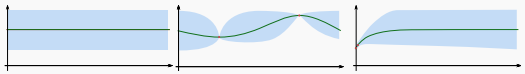
Definition 2.2 Utility Function **F:** Is a function that provide a ranking to judge uncertain situations.

1. Uncertainty Sampling for Regression

1.1. Maximizing the Information Gain

Let f be a unkown function that we can evaluate with $D = \texttt{dom}(f)$. Let S be a subset of points $S \subseteq D$ that we can choose make noisy observations y_S of f in order to maximze the *information gain*[def. 5.1]: [example 2.1]

$$F(S) := H(f) - H(f|y_s) \overset{\text{eq. (5.16)}}{=} I(f; y_s) \tag{2.1}$$



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

Definition 2.3 Optimal Set of labels:

$$\left\{x_1, \dots, x_{|S|}\right\} = \underset{S \subseteq D, |S| \leqslant T}{\arg\max} F(S) \tag{2.2}$$

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

Definition 2.4 [Proof 4]
Greedy Mutual Information Maximization Objective:
Only consider the next point that maximizes the mutual information and not all at once:

$$\begin{aligned} x_{t+1} &= \underset{x \in D}{\arg\max}\, F(S_t \cup \{x\}) \\ &= \underset{x \in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \end{aligned} \tag{2.3}$$

Corollary 2.1 [Proof 4]
Homoscedactic Gaussian:

$$x_t = \underset{x \in D}{\arg\max}\, \sigma^2_{t-1}(x) \tag{2.4}$$

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

$$\sigma^2_t(x) = \mathrm{k}(x,x) - \mathrm{k}_{x,A_t}\left(\mathcal{K}_{A_t,A_t} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{k}_{x,A_t} \tag{2.5}$$

Algorithm 2.1 Greedy Uncertainty Sampling:
Given: $S_t := \{x_1, \dots, x_t\}$
1: **for** $t+1 \dots, T$ **do**

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

2: **end for**

Corollary 2.2 Diminishing Returns Property:
Mutal information satisifies modular submodularity (Property 5.9)
$\Rightarrow$ adding a label/memasurement for some data point can only increase information:

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

Note

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

Corollary 2.3 Constant Factor Approximation: algorithm 2.1 provides a constant factor approximation of eq. (2.1):

$$F(S_T) \leqslant \underbrace{\left(1 - \frac{1}{\mathrm{e}}\right)}_{\approx .63} \max_{S \subseteq D, |S| \leqslant T} F(S) \tag{2.6}$$

Note

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

1.2. Heteroscedastic Case

So far we considered homoscedastic noise[def. 25.18] but sometimes we may have heteroscedasctic[def. 25.19] noise $\sigma_n(x) \iff$ different locations may have different noise i.e. to different sensors.
Problem: in the heteroscedas case the most uncertain outcomes are no longer necessarily the most informative.

Corollary 2.4 [Proof 4]
Heteroscedastic Gaussian:

$$x_t = \underset{x \in D}{\arg\max} \frac{\text{epistemic uncertainty}}{\text{aleatoric uncertainty}} = \underset{x \in D}{\arg\max} \frac{\sigma^2_f(x)}{\sigma^2_n(x)} \tag{2.7}$$

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

$$\sigma^2_t(x) = \mathrm{k}(x,x) - \mathrm{k}_{x,A_t}\left(\mathcal{K}_{A_t,A_t} + \sigma^2 \boldsymbol{I}\right)^{-1} \mathrm{k}_{x,A_t} \tag{2.8}$$

2. Uncertainty Sampling for Classification

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

Definition 2.5 Greedy Mutual Entropy Maximization: select the next point that maximizes the entropy over the label distribution:

$$\begin{aligned} x_{t+1} &= \underset{x \in D}{\arg\max}\, H(Y|x, x_{1:t}, y_{1:t}) = \underset{x \in D}{\arg\max} && (2.9) \\ &= \underset{x \in D}{\arg\max} - \mathrm{p}(y|x, x_{1:t}, y_{1:t}) && (2.10) \\ &= \underset{x \in D}{\arg\max} - \sum_y \mathrm{p}(y|x, x_{1:t}, y_{1:t}) && (2.11) \end{aligned}$$

Notes

The posterior $\mathrm{p}(y|x, x_{1:t}, y_{1:t})$ is usually intractable but we can using approximate inference section 9 methods:
- Approximate Inference section 1
- Markov Chain Monte Carlos section 2

2.1. Heteroscedastic Case

So far we considered homoscedastic noise[def. 25.18] but sometimes we may have heteroscedasctic[def. 25.19] noise $\sigma_n(x) \iff$ different locations may have different noise i.e. to different sensors.
Problem: in the heteroscedas case the most uncertain labels are no longer necessarily the most informative.

2.1.1. Informative Sampling for Classification

Definition 2.6 [Proof 4]
Bayesian active learning by disagreement (BALD):

$$\begin{aligned} x_{t+1} &= \underset{\hat{x} \in D}{\arg\max}\, I(\theta; \hat{y}|\hat{x}, x_{1:t}, y_{1:t}) && (2.12) \\ &= \underset{\hat{x} \in D}{\arg\max}\, H(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}) - \mathbb{E}_{\theta \sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H(\hat{y}, \hat{x}, \theta)\right] \end{aligned}$$

Explanation 2.1.
(1) $H(\hat{y}|\hat{x}, x_{1:t}, y_{1:t})$:
is the entropy of the predictive posterior distribution[def. 6.19], *approximate using approximate inference section 9.*
(2) $\mathbb{E}_{\theta \sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H(\hat{y}, \hat{x}, \theta)\right]$:
is the conditional Entropy over the labels by drawing θ from the posterior distribution and averagin over them.

3. Examples

Example 2.1 Gaussian Information Gain:

$$F(S) \overset{\text{example 5.8}}{=} \frac{1}{2} \log\left|\boldsymbol{I} + \sigma^{-2} \mathcal{K}_S\right|$$

4. Proofs

Proof. [def. 2.4]

$$\begin{aligned} x_{t+1} &= \underset{x \in D}{\arg\max}\, F(S_t \cup \{x\}) \overset{\text{eq. (14.47)}}{=} \underset{x \in D}{\arg\max}\, F(S_t \cup \{x\}) - F(S_t) \\ &= \underset{x \in D}{\arg\max}\, I(f; y_{S_t + x}) - I(f; y_{S_t}) \\ &= \underset{x \in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{\text{eq. (5.16)}}{=} \underset{x \in D}{\arg\max}\, H\left(y_{S_t+x}\right) - H\left(y_{S_t+x}|f\right) - H\left(y_{S_t}\right) + H\left(y_{S_t}|f\right) \\ &= \underset{x \in D}{\arg\max}\, H\left(y_{S_t}, x\right) - H\left(y_{S_t+x}|f\right) - H\left(y_{S_t}\right) + H\left(y_{S_t}|f\right) \\ &\overset{\text{eq. (5.7)}}{=} \underset{x \in D}{\arg\max}\, H\left(y_x|y_{S_t}\right) - H\left(y_{S_t+x}|f\right) + H\left(y_{S_t}|f\right) \\ &\overset{\clubsuit}{=} H(y_x|y_{S_t}) - H(y_x|f) \end{aligned}$$

□

Note ♣

$$\begin{aligned} H\left(y_{S_t} \cup x|f\right) &\overset{\text{eq. (5.7)}}{=} H\left(y_{s_t}|f\right) + H\left(y_x|f, y_{S_t}\right) \\ &= H\left(y_{s_t}|f\right) + H\left(y_x|f\right) \end{aligned}$$

Proof. corollary 2.1

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

$$\begin{aligned} x_{t+1} &= \underset{x \in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{\text{eq. (5.29)}}{=} \underset{x \in D}{\arg\max}\, \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2 x|S_t - \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2_n \\ &\overset{\text{eq. (14.47)}}{=} \sigma^2_{x|S_t} \end{aligned}$$

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

$$x_t = \underset{x \in D}{\arg\max}\, \sigma^2_{t-1}(x) \tag{2.13}$$

□

Proof. corollary 2.4

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

$$\begin{aligned} x_{t+1} &= \underset{x \in D}{\arg\max}\, H(y_x|y_{S_t}) - H(y_x|f) \\ &\overset{\text{eq. (5.29)}}{=} \underset{x \in D}{\arg\max}\, \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2 x|S_t - \frac{1}{2}\ln(2\pi\mathrm{e})\sigma^2_n(x) \\ &\overset{\text{eq. (14.47)}}{=} \underset{x \in D}{\arg\max}\, \ln \frac{\sigma^2_f(x)}{\sigma^2_n(x)} \overset{\text{eq. (14.59)}}{=} \underset{x \in D}{\arg\max}\, \frac{\sigma^2_f(x)}{\sigma^2_n(x)} \end{aligned}$$

□

Proof. [def. 2.6]

$$\begin{aligned} I(\theta; \hat{y}|x_{1:t}, y_{1:t}) &\overset{\text{eq. (5.16)}}{=} H(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}) - H(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t}) \\ &\overset{\text{eq. (5.6)}}{=} H(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}) \\ &\quad - \mathbb{E}_{\hat{y}|\theta \sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[\log \mathrm{p}_{\hat{y}|\theta}(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t})\right] \\ &\overset{\text{eq. (5.2)}}{=} H(\hat{y}|\hat{x}, x_{1:t}, y_{1:t}) \\ &\quad - \mathbb{E}_{\hat{y}|\theta \sim \mathrm{p}(\cdot|x_{1:t}, y_{1:t})}\left[H(\hat{y}|\theta, \hat{x}, x_{1:t}, y_{1:t})\right] \end{aligned}$$

□

Bayesian Optimizaton

In section 1 we tried to maximize our information gain about an unknown function f.
While While sequentially optimizing eqs. (2.3) and (2.4) is a provably good way to explore f globally, it is not well suited for function value optimization, where we only care about maximizing our knowledge about the maxima.

Given

- set of possible inputs $D = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$
- unknown (black-box) function/oracle:

$$f \in \mathcal{F} \qquad f : D \mapsto \mathbb{R} \tag{3.1}$$

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

$$y_t = f(\boldsymbol{x}_t) + \epsilon \tag{3.2}$$

Goal

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

$$\sum_{t=1}^{T} f(\boldsymbol{x}_t) \tag{3.3}$$

$\Rightarrow$ need a measure of performance i.e. cumulative regret [def. 3.3] as we can only draw point samples from f.

Definition 3.1
Action Set $\mathcal{A} = \{a_1, \dots, a_n\}$**:**
Is the set of possible actions from which we can choose at each step.

Corollary 3.1 : If we want to maximize a function f, then its just the set of possible inputs $\mathcal{A} = D$

Definition 3.2 Optimizing Agent/Decision Making Policy:
Is a policy on how to choose an action $a \in \mathcal{A}$ based on a objective/utility function[def. 2.2]

Definition 3.3 (Cumulative) Regret for a fixed f: Is defined as the the cumulative loss we suffer in comparison to taking the optimal value $\boldsymbol{x}^*$ if we had full knowlede of f.

$$\begin{aligned} R_T &:= \sum_{t=1}^{T} \left(\max_{\boldsymbol{x} \in D} f(\boldsymbol{x}) - f(\boldsymbol{x}_t) \right) = \sum_{t=1}^{T} r_t \\ &= T \max_{\boldsymbol{x} \in D} f(\boldsymbol{x}) - \sum_{t=1}^{T} f(\boldsymbol{x}_t) \end{aligned} \tag{3.4}$$

r_t: instantaneous regret

Definition 3.4 (Time) Average Regret:

$$\frac{R_T}{T} = \frac{1}{T} \sum_{t=1}^{T} r_t = \frac{1}{T} \sum_{t=1}^{T} \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) \right) \tag{3.5}$$

Definition 3.5
No/Sublinear Regret Algorithms:

$$\lim_{T \to \infty} \underbrace{\frac{R_T}{T}}_{\text{Average regret}} = 0 \qquad \begin{matrix} R_T = o(T) \\ \frac{R_T}{T} = o(1) \end{matrix} \qquad \forall \text{sequences } 1, \dots, T \tag{3.6}$$

Explanation 3.1. *Due to more information the instantaneous regret decreases over time and we obtain no regret in average.*

Definition 3.6
Pure Exploration/Follow the Leader Policy: Take the action with the current maximum empirical mean payoff.

Algorithm 3.1 Epsilon Greedy Algorithm:

Set: $\epsilon_t = \mathcal{O}\left(\frac{1}{t}\right)$

1: **for** $t = 1, \dots, T$ **do**
2: With probability ϵ_t *explore* unif. at randomn:

$$a_{t+1} = \mathcal{U}(a_1, \dots |\mathcal{A}|) \tag{3.7}$$

3: With probability $1 - \epsilon_t$ *take* action with highest known empirical mean payoff:

$$a_{t+1} = \arg\max_{a \in \mathcal{A}} \hat{\mu}_{a,T} \qquad \hat{\mu}_{a,T} = \frac{1}{n_{a,T}} \sum_{s=1}^{T} \mathbb{1}_{\{a_s = a\}} v_{a,s} \tag{3.8}$$

4: **end for**

Problem

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

1. Optimistic Bayesian Optimization

Problem

Picking the nex point greedily by maximizing the mean payoff[def. 3.6]

$$x_t = \arg\max_{x \in D} \mu_{t-1}(x) \tag{3.9}$$

of the posterior distribution tends to lead to local optima.

Assumption

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

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

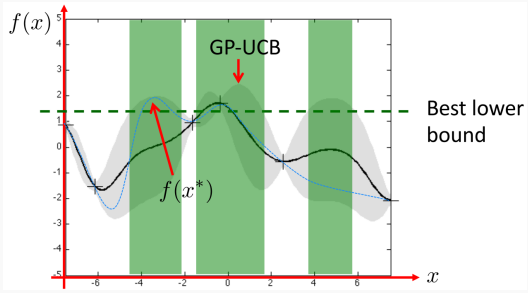
then it follows that:

$$f\left(\boldsymbol{x}^*\right) \geqslant \max \mu(x) - \beta\sigma(x) \tag{3.10}$$

this implies that we should focus only on certain regions. Because if the best predicted value of a point $\mu(x') + \sigma(x')$ is less then the *best lower confidence bound*eq. (3.10) then the maximum cannot be at x':

$$f\left(\boldsymbol{x}^*\right) \geqslant \max \mu(x) - \beta\sigma(x) \geqslant \mu(x') + \sigma(x') \tag{3.11}$$

add own figure



This idea can be utilized in various ways:

- GP-UCB section 1
- Thompson Sampling section 2

1.1. Gaussian Process-UCB

Principle 3.1 Optimization in the phase of uncertainty:
Pick the action that has the highest upper confidence bound (UCP).

Explanation 3.2 (principle 3.1)**.** *We do not pick the action that maximizes our current estimate $\mu(x)$ but the most optimistic one.*
If the guess is wrong optimism will fade quickly but if the guess is right we will maximize our utility will decreasing uncertainty.

Definition 3.7 GP-UCB:

$$\boldsymbol{x}_t = \arg\max_{x \in D} \mu_{t-1}(\boldsymbol{x}) + \beta_t \sigma_{t-1}(\boldsymbol{x}) \tag{3.12}$$

Explanation 3.3 (Definition 3.7)**.**

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

1.1.1. Maximizing the UCB

The GP-UCB[def. 3.7] is usually a non-convex function. Thus in order to maximize this objective we need to use:

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

1.1.2. Guarantees on the regret

Theorem 3.1 Bayesian Regret of GP-UCB: assuming the true function f follows a Gaussian Process $f \sim \mathcal{GP}$ then it holds that for a suitable choice of β_t (needs to slowly decay with const$\cdot \log t$):

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

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

Explanation 3.4 (γ_T)**.** *The regret depends on how much information we can gain in T steps.*

Corollary 3.2 Linear Kernel:
For a linear kernel[def. 10.9] it holds:

$$\gamma_T = \mathcal{O}(d \log T) \tag{3.13}$$

Corollary 3.3 Squared Exponential Kernel:
For a squared exponential kernel[def. 10.13] it holds:

$$\gamma_T = \mathcal{O}\left((\log T)^{d+1}\right) \tag{3.14}$$

Corollary 3.4 Matern Kernel $\nu > 2$:
For a linear kernel[def. 10.14] it holds:

$$\gamma_T = \mathcal{O}\left(T^{\frac{d(d+1)}{2\nu + d(d+1)}}\right) \tag{3.15}$$

Note: Reproducing Kernel Hilbert Space (RKHS)

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

1.2. Thompson Sampling

Definition 3.8 Thompson Sampling: Draw a function $\tilde{f}$ from the posterior and maximize it:

$$\tilde{f} \sim \mathrm{p}(f | \boldsymbol{x}_{1:n}, \boldsymbol{y}_{1:t}) \qquad \boldsymbol{x}_{t+1} \in \arg\max_{\boldsymbol{x} \in D} \tilde{f}(\boldsymbol{x}) \tag{3.16}$$

Explanation 3.5 (Definition 3.8)**.** *The randomness in $\tilde{f}$ helps to trade of exploration vs. exploitation.*

Machine Learning Appendix
Model Assessment and Selection

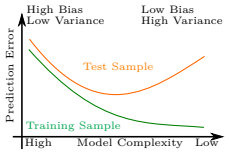
Definition 4.1 Statistical Inference: Is the process of deducing properties of an underlying probability distribution by mere analysis of data.

Definition 4.2
Model Selection:
Is the process of selecting a model f from a given or chosen class of models $\mathcal{F}$

Definition 4.3 Hyperparameter Tuning: Is the process of choosing the hyperparameters θ of a given model $f \in \mathcal{F}$

Definition 4.4 Model Assessment/Evaluation: Is the process of evaluating the performance of a model.

Definition 4.5 Overfitting:
Describes the result of training/fitting a model f to closely to the training data $\mathcal{Z}^{\text{train}}$.
That is, we are producing overly complicated model by fitting the model to the noise of the training set.
Consequences: the model will generalize poorly as the test set $\mathcal{Z}^{\text{test}}$ will not have not the same noise
$\Rightarrow$ big test error.



Definition 4.6 Training Set $\mathcal{Z}^{\text{train}}$: Is the part of the data that is used in order to train the model i.e. part of data which is used in order to update the weight according to the loss.

Definition 4.7 Validation Set $\mathcal{Z}^{\text{val}}$: Is the part of the data that is used in order to evaluate different hyperparamters.

Definition 4.8 Test Set $\mathcal{Z}^{\text{test}}$: Is part of the data that is used in order to test the performance of our model.

Move next section to statistical perspective section or merge it somehow

1.1. Core Problem of Statistical Inference

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

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

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

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

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

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

Questions

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

$$f^*(\boldsymbol{x}) \in \underset{f \in \mathcal{F}}{\arg\min}\, \mathcal{R}(f) \quad \text{vs.} \quad \hat{f}(\boldsymbol{x}) \in \underset{f \in \mathcal{F}}{\arg\min}\, \hat{\mathcal{R}}(f)$$

1.2. Empirical Risk Minimization

1. For a chosen set of function classes $\mathcal{F}$ minimize the empirical risk/loss:
$$\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min}\, \hat{\mathcal{R}}\left(f, \mathcal{Z}^{tr}\right) = \underset{f \in \mathcal{F}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, f(\boldsymbol{x}_i)\right)$$
2. Determine the best parameter θ^* by using the validation set for evaluation:
$$\underline{\hat{\theta}\left(\mathcal{Z}^{val}\right)} \in \underset{\theta : \hat{f}_\theta \in \mathcal{F}_\theta}{\arg\min}\, \hat{R}\left(\hat{f}_\theta\left(\mathcal{Z}^{tr}\right), \mathcal{Z}^{val}\right)$$
3. Use the tests set in order to test the model:
$$\hat{\mathcal{R}}\left(\hat{f}_{\underline{\hat{\theta}(\mathcal{Z}^{val})}}\left(\mathcal{Z}^{tr}\right), \mathcal{Z}^{test}\right)$$

Note: overfitting to the validation set

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

2. Cross Validation

Definition 4.9 Cross Validation: Is a model validation/assessment techniques for assessing how the results of a statistical analysis (model) will generalize to an independent data set.

2.1. Validation Set Approach

Definition 4.10 Hold out/Validation Set:

Add

2.2. Leave-One-Out Cross Validation (LOOCV)
2.3. K-Fold Cross Validation

A Statistical Perspective

1. Information Theory

1.1. Information Content

Definition 5.1 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:

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

$\Rightarrow$ can use the logarithm to satisfy this

Definition 5.2 Surprise/Self-Information/-Content: Is a measure of the information of a realization x of a random variable $X \sim \mathrm{p}$:

$$I_X(x) = \log\left(\frac{1}{\mathrm{p}(X=x)}\right) = -\log \mathrm{p}(X=x) \tag{5.1}$$

Corollary 5.1 Units of the Shannon Entropy:
The Shannon entropy can be defined for different logarithms

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

Explanation 5.1. *An uncertain event is much more informative than an expected/certain event:*

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

1.2. Entropy

Information content deals with a single event. If we want to quantify the amount of uncertainty/information of a probability distribution, we need to take the expectation over the information content[def. 5.2]:

Definition 5.3 Shannon Entropy example 5.3:
Is the expected amount of information of a random variable $X \sim \mathrm{p}$:

$$H(\mathrm{p}) = \mathbb{E}_X[I_X(x)] = \mathbb{E}_X\left[\log\frac{1}{\mathrm{p}_X(x)}\right] = -\mathbb{E}_X[\log \mathrm{p}_X(x)]$$
$$= -\sum_{i=1}^{n} \mathrm{p}(x_i)\log \mathrm{p}(x_i) \tag{5.2}$$

Definition 5.4 Differential/Continuous entropy: Is the continuous version of the Shannon entropy[def. 5.3]:

$$H(\mathrm{p}) = \int_{x\sim\mathrm{p}} -f(x)\log f(x)\,\mathrm{d}x \tag{5.3}$$

Notes

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

Property 5.1 Non negativity:
Entropy is always non-negative:

$$H(X) \geqslant 0 \qquad \text{if } X \text{ is deterministic} \qquad H(X) = 0 \tag{5.4}$$

1.2.1. Conditonal Entropy

Proposition 5.1 Conditioned Entropy $H(Y|X=x)$:
Let X and Y be two random variables with a condititional pdf $\mathrm{p}_{X|Y}$. The entropy of Y conditioned on X taking a certain value x is given as:

$$H(Y|X=x) = \mathbb{E}_{Y|X=x}\left[\log\frac{1}{\mathrm{p}_{y|X}(Y|X=x)}\right]$$
$$= -\mathbb{E}_{Y|X=x}\left[\log \mathrm{p}_{Y|X}(y|X=x)\right] \tag{5.5}$$

Definition 5.5 proof 3
Conditional Entropy $H(Y|X)$:
Is the amount of information need to determine Y if we arleady know X and is given by averagin $H(Y|X=x)$ over X.

$$H(Y|X) = [\mathbb{E}_X H(Y|X=x)] = -\mathbb{E}_{X,Y}\left[\log\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)}\right] \tag{5.6}$$
$$= \mathbb{E}_{X,Y}\left[\log\frac{\mathrm{p}(x)}{\mathrm{p}(x,y)}\right]$$

Definition 5.6 proof 3
Chain Rule for Entropy:

$$H(Y|X) = H(X,Y) - H(X)$$
$$H(X|Y) = H(X,Y) - H(Y) \tag{5.7}$$

Property 5.2 Monotonicity:
Information/conditioning reduces the entropy
$\Rightarrow$ *Information never hurts.*

$$H(X|Y) \geqslant H(X) \tag{5.8}$$

Corollary 5.2 From eq. (5.16):

$$H(X,Y) \leqslant H(X) + H(Y) \tag{5.9}$$

1.3. Cross Entropy

Definition 5.7 Cross Entropy proof 3:
Lets say a model follows a true distribution $X \sim \mathrm{p}$ but we model X as with a different distribution $X \sim q$. The cross entropy between p and q measure the average amount of information/bits needed to model an outcome $x \sim \mathrm{p}$ with X:

$$H(\mathrm{p}, q) = \mathbb{E}_{x\sim\mathrm{p}}\left[\log\left(\frac{1}{q(x)}\right)\right] = -\mathbb{E}_{x\sim\mathrm{p}}[\log q(x)] \tag{5.10}$$
$$= H(\mathrm{p}) + D_{\mathrm{KL}}(\mathrm{p} \parallel q) \tag{5.11}$$

Corollary 5.3 Kullback-Leibler Divergence: $D_{\mathrm{KL}}(\mathrm{p} \parallel q)$ measures the extra price (bits) we need to pay for using q.

1.4. Kullback-Leibler (KL) divergence

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

Definition 5.8
Kullback–Leibler divergence. examples 5.4 and 5.7
/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:

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) = \mathbb{E}_{x\sim\mathrm{p}}\left[\log\frac{\mathrm{p}(x)}{q(x)}\right] = \mathbb{E}_{x\sim\mathrm{p}}[\log \mathrm{p}(x) - \log q(x)] \tag{5.12}$$

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

Explanation 5.2.

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

Intuition

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

Property 5.3 Non-Symmetric:

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) \neq D_{\mathrm{KL}}(q \parallel \mathrm{p}) \qquad \forall \mathrm{p}, q \tag{5.13}$$

Property 5.4:

$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) \geqslant 0 \tag{5.14}$$
$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) = 0 \iff \mathrm{p}(x) = q(x) \forall x \in \mathcal{X} \tag{5.15}$$

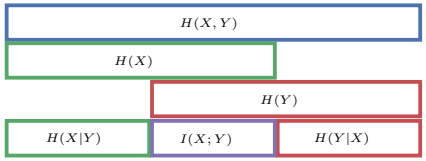
Note

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

1.5. Mutual Information

Definition 5.9 example 5.8
Mutual Information/Information Gain: Let X and Y be two random variables with a joint probability distribution. The mutual information of X and Y is the reduction in uncertainty in X if we know Y and vice versa.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{5.16}$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= D_{\mathrm{KL}}\left(\mathrm{p}_{X,Y} \parallel \mathrm{p}_X \mathrm{p}_Y\right)$$



Explanation 5.3 (Definition 5.9).

$$I(X;Y) = \begin{cases} \textit{big} & \textit{if } X \textit{ and } Y \textit{ are highly dependent} \\ 0 & \textit{if } X \textit{ and } Y \textit{ are independent} \end{cases} \tag{5.17}$$

Property 5.5 Symmetry:

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

Property 5.6 Positiveness:

$$I(X;Y) \geqslant 0 \qquad \text{if } X \perp\!\!\!\perp Y \qquad I(X;Y) = 0 \tag{5.18}$$

Property 5.7:

$$I(X;Y) \leqslant H(X) \qquad I(X;Y) \leqslant H(Y) \tag{5.19}$$

Property 5.8 Self-Information:

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

Property 5.9 Montone Submodularity: Mutual information is monotone submodular[def. 12.10]:

$$H(X,z) - H(X) \geqslant H(Y,z) - H(Y) \tag{5.20}$$
$$\overset{\text{[def. 5.6]}}{\iff} \qquad H(z|X) \geqslant H(x|Y) \tag{5.21}$$

Recall: goal of supervised learning

Given: training data:

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

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

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

s.t. we minimize prediction error/empirical risk [def. 5.14].

Fundamental assumption

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

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

Note

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

1.6. Generalization Error

Definition 5.10
Generalization/Prediction Error (Risk): Is defined as the expected value of a loss function l of a given predictor h, for data drawn from a distribution $\mathrm{p}_{\mathcal{X},\mathcal{Y}}$.

$$R_{\mathrm{p}}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathrm{p}}[l(y;h(\boldsymbol{x}))] = \int_{\mathcal{D}} \mathrm{p}(\boldsymbol{x},y) l(y;h(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}y$$
$$= \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathrm{p}(\boldsymbol{x},y) l(y,h(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}y$$
$$\overset{??}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}} l(y,h(\boldsymbol{x}))\mathrm{p}(y|\boldsymbol{x})\mathrm{p}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}y \tag{5.22}$$

Interpretation

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

Definition 5.11 Expected Conditional Risk: If we only know a certain $\boldsymbol{x}$ but not the distribution of those measurements ($\boldsymbol{x} \sim \mathrm{p}_{\mathcal{X}}(\boldsymbol{x})$), we can still calculate the expected risk given/conditioned on the known measurement $\boldsymbol{x}$:

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

Note: [def. 5.10] $\iff$ [def. 5.11]

$$R_{\mathrm{p}}(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathrm{p}}[R_{\mathrm{p}}(h,\boldsymbol{x})] = \int_{\mathcal{X}} \mathrm{p}(\boldsymbol{x}) R_{\mathrm{p}}(h,\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{5.23}$$

Definition 5.12 Expected Risk Minimizer (TRM) h^*:
Is the model h that minimizes the total expected risk:

$$h^* \in \arg\min_{h\in\mathcal{C}} \mathcal{R}(h) \tag{5.24}$$

Problem

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

1.7. Empirical Risk

Definition 5.13 Empirical Risk: Is the the average of a loss function of an estimator h over a finite set of data $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ drawn from $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}, y)$:

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

Note

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

Definition 5.14 Empirical Risk Minimizer (ERM) $\hat{h}$: Is the model h that minimizes the total empirical risk:

$$\hat{h} \in \arg\min_{h \in \mathcal{C}} \hat{\mathcal{R}}(h) \tag{5.25}$$

Objective

Given data generated *i.i.d.* from an distribution $\mathrm{p}_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}_i, y_i)$.
Goal: find the function/predictor $h : \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the expected risk [def. 5.10] i.e. we want to find the expected risk minimizer ([def. 5.12]).

Definition 5.15
Bayes' optimal predictor for the L2-Loss:
Assuming: i.i.d. generated data by $(\boldsymbol{x}_i, y_i) \sim \mathrm{p}(\mathcal{X}, \mathcal{Y})$.
Considering: the least squares risk:

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

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

$$h^*(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \tag{5.26}$$

see

Proof. proof:defn:bayesOptPredictor

Notes

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

1.7.1. Bayes Optimal Predictor

1.8. How to make use of this in Practice

In Practice

Problem: we do not know the real distribution $\mathrm{p}_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})$, which we need in order to find the bayes optimal predictor according to eq. (5.26).
Idea:

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

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

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

Definition 5.16 Statistical Inference: Goal of Inference
(1) What is a good guess of the parameters of my model?
(2) How do I quantify my uncertainty in the guess?

2. Estimators

Definition 5.17 (Sample) Statistic: A statistc is a measuarble function f that assigns a **single** value F to a sample of random variables:

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

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 5.18 Statistical/Population Parameter:
Is a parameter defining a family of probabilty distributions see example 5.1

Definition 5.19 (Point) Estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$:
Given: n-samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim \boldsymbol{X}$ an estimator

$$\hat{\theta} = h(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \tag{5.27}$$

is a statistic/randomn variable used to estimate a true (population) parameter θ[def. 5.18] see also example 5.2.

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).

3. Proofs

Proof. [def. 5.7]

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

□

Proof. [def. 5.15]:

$$\min_h R(h) = \min_h \mathbb{E}_{(\boldsymbol{x},y)\sim \mathrm{p}}[(y - h(\boldsymbol{x}))^2]$$
$$\overset{??}{=} \min_h \mathbb{E}_{\boldsymbol{x}\sim \mathrm{p}_{\mathcal{X}}}\left[\mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y - h(\boldsymbol{x}))^2 \,|\boldsymbol{x}\right]\right]$$
$$\overset{\heartsuit}{=} \mathbb{E}_{\boldsymbol{x}\sim \mathrm{p}_{\mathcal{X}}}\Big[\min_h(\boldsymbol{x}) \underbrace{\mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y - h(\boldsymbol{x}))^2 \,|\boldsymbol{x}\right]}_{\mathcal{R}_{\mathrm{p}}(h,\boldsymbol{x})\ ([\text{def. } 5.11])}\Big]$$

Now lets minimize the conditional executed risk:

$$h^*(\boldsymbol{x}) = \arg\min_h \mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\mathcal{Y}|\mathcal{X}}}\left[(y - h(\boldsymbol{x}))^2 \,|\boldsymbol{x}\right] \tag{5.28}$$

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

□

Notes: ♡

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

Note

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

Proof. Definition 5.5

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

□

Proof. [def. 5.6] We start from eq. (5.6):

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

□

Proof. example 5.4

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

$$\underline{\mathbb{E}_{\mathrm{p}}[a]} \overset{\mathrm{tr}(\mathbb{R})=\mathbb{R}}{=} \mathbb{E}_{\mathrm{p}}\left[\mathrm{tr}\left\{(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{p}})^\intercal\Sigma_{\mathrm{p}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{p}})\right\}\right]$$
$$\overset{\text{eq. (17.16)}}{=} \mathbb{E}_{\mathrm{p}}\left[\mathrm{tr}\left\{(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{p}})(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{p}})^\intercal\Sigma_{\mathrm{p}}^{-1}\right\}\right]$$
$$= \mathbb{E}_{\mathrm{p}}\left[\mathrm{tr}\left\{\Sigma_{\mathrm{p}}\Sigma_{\mathrm{p}}^{-1}\right\}\right]$$
$$\overset{\text{eq. (17.16)}}{=} \mathbb{E}_{\mathrm{p}}\left[\mathrm{tr}\left\{\boldsymbol{I}_d\right\}\right] = \mathbb{E}_{\mathrm{p}}[d] = d$$

$$\underline{\mathbb{E}_{\mathrm{p}}[b]} \overset{\text{eq. (22.56)}}{=} (\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_q)^\intercal\Sigma_q^{-1}(\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_q) + \mathrm{tr}\left\{\Sigma_q^{-1}\Sigma_{\mathrm{p}}\right\}$$

□

4. Examples

Example 5.1 : Normal distribution has two population parameters: the mean μ and the variance σ^2.

Example 5.2 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 5.3 Entropy of a Gaussian:

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\ln|2\pi\mathrm{e}\Sigma| \overset{\text{eq. (17.2)}}{=} \frac{1}{2}\ln\left((2\pi\mathrm{e})^d|\Sigma|\right)$$
$$= \frac{d}{2}\ln(2\pi\mathrm{e})^d + \log|\Sigma| \tag{5.29}$$
$$\overset{\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)}{=} \frac{1}{2}\ln|2\pi\mathrm{e}| + \frac{1}{2}\sum_{i=1}^{d}\ln\sigma_i^2$$

Example 5.4 proof 3
KL Divergence of Gaussians:
Given two Gaussian distributions:

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

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

Example 5.5 KL Divergence of Scalar Gaussians:

$$\theta \sim q(\theta|\boldsymbol{\lambda}) = \mathcal{N}\left(\mu_q, \sigma_q^2\right) \qquad \boldsymbol{\lambda} = [\mu_q \quad \sigma_q]$$
$$\mathrm{p} = \mathcal{N}\left(\mu_{\mathrm{p}}, \sigma_{\mathrm{p}}^2\right)$$
$$D_{\mathrm{KL}}(\mathrm{p} \parallel q) = \frac{1}{2}\left(\frac{\sigma_{\mathrm{p}}^2}{\sigma_q^2}(\mu_q - \mu_{\mathrm{p}})^2\sigma_q^{-2} - 1 + \log\left(\frac{\sigma_q^2}{\sigma_{\mathrm{p}}^2}\right)\right)$$

Example 5.6 KL Divergence of Diag. Gaussians:

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

Example 5.7 KL Divergence of Gaussians:

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

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

Example 5.8 Gaussian Mutal Information:
Given $X \sim \mathcal{N}(\mu, \Sigma)$ $Y = X + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma\boldsymbol{I})$

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(\epsilon)$$
$$\overset{\text{eq. (5.29)}}{=} \frac{1}{2}\ln(2\pi\mathrm{e})^d|\Sigma + \sigma^2\boldsymbol{I}| - \frac{1}{2}\ln(2\pi\mathrm{e})^d|\sigma^2\boldsymbol{I}|$$
$$= \frac{1}{2}\ln\frac{\cancel{(2\pi\mathrm{e})^d}\cancel{\sigma^{2d}}|\sigma^{-2}\Sigma + \boldsymbol{I}|}{\cancel{(2\pi\mathrm{e})^d}|\cancel{\sigma^{2d}}\boldsymbol{I}|}$$
$$= \frac{1}{2}\ln|\boldsymbol{I} + \sigma^{-2}\Sigma|$$

Model Parameter Estimation

1. Maximum Likelihood Estimation

1.1. Likelihood Function

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

Definition 6.1 Likelihood Function $\mathscr{L}_n : \Theta \times \mathbb{R}^n \mapsto \mathbb{R}_+$:
Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a random sample of i.i.d. data points drawn from an unknown probability distribution $\boldsymbol{x}_i \sim p_{\mathcal{X}}$.
The likelihood function gives the likelihood/probability of the *joint probability* of the data $\{x_1, \dots, x_n\}$ *given* a fixed set of model parameters θ:

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

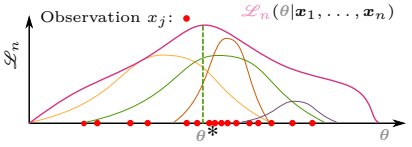


Figure 2: Possible Likelihood function in pink.
Overlayed: possible candidate functions for Gaussian model explaining the observations.

Likelihood function is not a pdf

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

Corollary 6.1 i.i.d. data: If the n-data points of our sample are i.i.d. then the likelihood function can be decomposed into a product of n-terms:

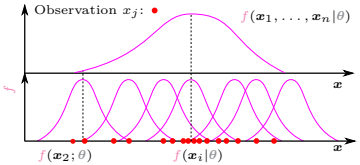


Figure 3: Bottom: probability distributions of the different data points $\boldsymbol{x}_i$ given a fixed θ for a Gaussian distribution
Top: joint probability distribution of the i.i.d. data points $\{\boldsymbol{x}_i\}_{i=1}^n$ given a fixed θ

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

Notation

- The probability density $f(\boldsymbol{X}|\theta)$ 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 $\{\boldsymbol{x}_i\}_{i=1}^n$ and thus written as $\mathscr{L}_n(\theta|\boldsymbol{X})$.
- Often the colon symbol **;** is written instead of the *is given* symbol | in order to indicate that θ resp. $\boldsymbol{X}$ is a parameter and not a randomn variable.

1.2. Maximum Likelihood Estimation (MLE)

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

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

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

Corollary 6.2 i.i.d. data: Differentiating the product of n-Terms with the help of the chain rule leads often to complex terms. As a result one usually prefers maximizing the log (especially for exponential terms), as it does not change the *argmax*–eq. (14.51):

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

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

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

1.3. Maximization vs. Minimization

For optimization problems we minimize by convention.
The logarithm is a concave function[def. 14.21] $\cap$, 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 3 $\cup$ by multiplying it by minus one and consider it as a loss function instead of a likelihood.

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

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

1.4. Conditional Maximum Likelihood Estimation

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

Note

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

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

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

2. Maximum a posteriori estimation (MAP)

Idea

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

2.1. Prior Distribution

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

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

Notes

In this section we use the terms model parameters θ and model as synonymous, as the model is fully described by its population parameters ([def. 5.18]) θ.

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

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

Definition 6.7 Hyperparameters $\mathrm{p}_{\boldsymbol{\lambda}}(\theta)$:
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 6.1

2.2. Posterior Distribution

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

$$\mathrm{p}(\theta|\textbf{data}) = \mathrm{p}(\theta|\boldsymbol{Z}) \tag{6.7}$$

Definition 6.9
Posterior Distribution and Bayes Theorem:
Using Bayes theorem 22.3 we can write the posterior distribution as a product of the *likelihood*[def. 6.1] weighted with our *prior*[def. 6.6] and normalized by the *evidence* $\boldsymbol{Z} = \{\boldsymbol{X}, \boldsymbol{y}\}$ s.t. we obtain a real probability distribution:

$$\mathrm{p}(\theta|\textbf{data}) = \mathrm{p}(\theta|\boldsymbol{Z}) = \frac{\mathrm{p}(\boldsymbol{Z}|\theta)\cdot \mathrm{p}_{\boldsymbol{\lambda}}(\theta)}{\mathrm{p}(\boldsymbol{Z})} \tag{6.8}$$

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

$$\boxed{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y}) = \frac{\mathrm{p}(\boldsymbol{y}|\theta,\boldsymbol{X})\cdot \mathrm{p}_{\boldsymbol{\lambda}}(\theta)}{\mathrm{p}(\boldsymbol{y}|\boldsymbol{X})}} \tag{6.10}$$

see proof section 3

2.2.1. Maximization –MAP

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

2.2.2. Maximization

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

$$\boxed{\theta^*_{\text{MAP}} = \arg\max_\theta \mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})} \tag{6.11}$$

Log-MAP estimator:

$$\theta^* = \arg\max_\theta \{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})\} \tag{6.12}$$

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

$$\overset{\text{eq. (14.48)}}{\propto} \arg\max_\theta \left\{\mathrm{p}(\boldsymbol{y}|\theta,\boldsymbol{X})\cdot \mathrm{p}_{\boldsymbol{\lambda}}(\theta)\right\}$$

Corollary 6.4 Negative Log MAP:

$$\theta^* = \arg\max_\theta \{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})\} \tag{6.13}$$

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

3. Proofs

Proof. 6.10:

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

$$\underline{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})\mathrm{p}(\boldsymbol{X},\boldsymbol{y})} = \underline{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})\mathrm{p}(\boldsymbol{y}|\boldsymbol{X})\mathrm{p}(\boldsymbol{X})}$$

$$\underline{\mathrm{p}(\boldsymbol{y}|\boldsymbol{X},\theta)\mathrm{p}(\boldsymbol{X},\theta)} = \mathrm{p}(\boldsymbol{y}|\boldsymbol{X},\theta)\mathrm{p}(\boldsymbol{X},\theta)$$

$$= \mathrm{p}(\boldsymbol{y}|\boldsymbol{X},\theta)\mathrm{p}(\theta|\boldsymbol{X})\mathrm{p}(\boldsymbol{X})$$

$$\overset{\text{eq. (6.6)}}{=} \mathrm{p}(\boldsymbol{y}|\boldsymbol{X},\theta)\mathrm{p}(\theta)\mathrm{p}(\boldsymbol{X})$$

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

□

Note

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

4. Examples

Example 6.1 Hyperparameters Gaussian Prior:

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

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

Regression

5. Linear/Ordinary Least Squares (OLS)

Definition 6.11 Linear Regression: Refers to regression that is linear w.r.t. to the parameter vector w (but not necessarily the data):

$$\boldsymbol{y} = \phi(\boldsymbol{x})^\intercal w \tag{6.14}$$

6. MLE with linear Model & Gaussian Noise

6.1. 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.

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

$\boldsymbol{X}$: set of explicative variables. ϵ: noise/error term.

Lemma 6.1 : The conditional distribution D of Y given $\boldsymbol{X}$ is equivilant to the unconditional distribution of the noise ϵ:

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

Example: Conditional linear Gaussian

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

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

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

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

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

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

if we consider an i.i.d. sample $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, the corresponding conditional (log-)likelihood is defined to be:

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

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

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

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

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

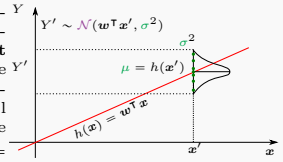
$$\theta^* = \begin{pmatrix} \boldsymbol{w}^* \\ \sigma^2{}^* \end{pmatrix} = \begin{pmatrix} \left(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\intercal\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \\ \frac{1}{n}\sum_{i=1}^n \left(y_i - \boldsymbol{w}_*^\intercal \boldsymbol{x}_i\right)^2 \end{pmatrix} \tag{6.15}$$

Note

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

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

$\Rightarrow y'$ is an sample from the distribution/model.



6.2. Conditional MLE ≙ Least Squares

Assuming that the noise is i.i.d. Gaussian with *constant* variance σ, that is $\theta = \left(\boldsymbol{w} \quad \sigma^2\right)^\intercal$

and considering the negative log likelihood in order to minimize $\arg\max \alpha = -\arg\min \alpha$:

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

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

$$\arg\min_{\boldsymbol{w}} \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i)^2 = \arg\min_{\boldsymbol{w}} \sum_{i=1}^n (y_i - \boldsymbol{w}^\intercal \boldsymbol{x}_i)^2 \tag{6.16}$$

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

Maximizing Liklihood $\iff$ Minimizing least squares

Corollary 6.5 : The Maximum Likelihood Estimate (MLE) for i.i.d. Gaussian noise (and general models) is given by the squared loss/Least squares solution, assuming that the variance is constant.

Heuristics for [def. 5.15]

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

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

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

$$\theta^* = \begin{pmatrix} \mu^* \\ \sigma^2{}^* \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n y_i \\ \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}_i)^2 \end{pmatrix} \tag{6.17}$$

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

Note

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

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

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

Is given by the least squares solution:

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

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

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.

7. Gaussian MAP

Classification

8. Logistic Regression

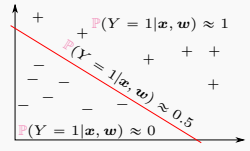
$\mathrm{Bern}(y;\sigma(w^\intercal \boldsymbol{x}, \sigma^2))$

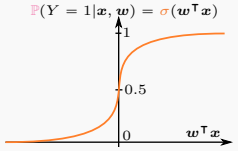
Idea: in order to classify dichometies we use a distribution that maps probabilities to a binary values 0/1

$\Rightarrow$ *Bernoulli Distribution*[def. 24.22].

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

Idea: use a sigmoidal function to convert distances $\boldsymbol{z} := \boldsymbol{w}^\intercal \boldsymbol{x}$ into probabilities $\Rightarrow$ *Logistic Function*[def. 6.12].





8.1. Logistic Function

Definition 6.12 Sigmoid/Logistic Function:

$$\sigma(z) = \frac{1}{1+\mathrm{e}^{-z}} = \frac{1}{1+\mathrm{e}^{\text{neg. dist. from deci. boundary}}} \tag{6.18}$$

Explanation 6.1 (Sigmoid/Logistic Function).

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

8.2. Logistic Regression

Definition 6.13 Logistic Regression: models the likelihood of the output y as a Bernoulli Distribution[def. 24.22] $y \sim \mathrm{Bern}(\mathrm{p})$, where the probability p is given by the Sigmoid function[def. 6.12] of a linear regression:

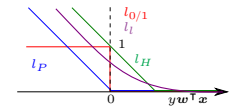
$$\mathrm{p}(y|\boldsymbol{x},\boldsymbol{w}) = \mathrm{Bern}\left(\sigma(\boldsymbol{w}^\intercal \boldsymbol{x})\right) = \begin{cases} \frac{1}{1+\mathrm{e}^{-\boldsymbol{w}^\intercal \boldsymbol{x}}} & \text{if } y = +1 \\ 1 - \frac{1}{1+\mathrm{e}^{-\boldsymbol{w}^\intercal \boldsymbol{x}}} & \text{if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

$$\overset{\text{section 9}}{=} \frac{1}{1+\exp(-y\cdot\boldsymbol{w}^\intercal \boldsymbol{x})} = \sigma\left(-y\cdot\boldsymbol{w}^\intercal \boldsymbol{x}\right) \tag{6.19}$$

8.2.1. Maximum Likelihood Estimate

Definition 6.14 Logistic Loss l_l proof 9:

Is the objective we want to minimize when performing mle[def. 6.3] for a logistic regression likelihood and incurs higher cost for samples closer to the decision boundary:



$$l_l(\boldsymbol{w};\boldsymbol{x},y) := \log\left(1+\exp(-y\cdot\boldsymbol{w}^\intercal \boldsymbol{x})\right) \tag{6.20}$$

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

Corollary 6.6 MLE for Logistic Regression:

$$\mathrm{l}_n(\boldsymbol{w}) = \sum_{i=1}^n l_l = \sum_{i=1}^n \log\left(1+\exp(-y_i\cdot\boldsymbol{w}^\intercal \boldsymbol{x}_i)\right) \tag{6.21}$$

Stochastic Gradient Descent

The logistic loss l_l is a convex function. Thus we can use convex optimization techniques s.a. SGD in order to minimize the objective corollary 6.6.

Definition 6.15 proof 9

Logistic Loss Gradient $\nabla_w l_l(w)$:

$$\begin{aligned}
\nabla_w l_l(w) &= \mathbb{P}(Y = -y|\boldsymbol{x},w)\cdot(-y\boldsymbol{x}) \\
&= \frac{1}{1+\exp(y\boldsymbol{w}^\intercal \boldsymbol{x})}\cdot(-y\boldsymbol{x})
\end{aligned} \tag{6.22}$$

Explanation 6.2.

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

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

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

Algorithm 6.1 Vanilla SGD for Logistic Regression:

Initalize: w

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

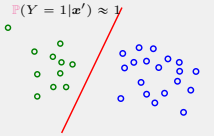
Making Predictions

Given an optimal parameter vector $\hat{w}$ found by algorithm 6.1 we can predict the output of a new label by eq. (6.19):

$$\mathbb{P}(y|\boldsymbol{x},\hat{w}) = \frac{1}{1+\exp\left(-y\hat{w}^\intercal \boldsymbol{x}\right)} \tag{6.23}$$

Drawback

Logistic regression, does not tell us anything about the liklihood $\mathrm{p}(\boldsymbol{x})$ 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)**:

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

- **L1 (Laplace prior)**:

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

- **Generalized**:

$$\begin{aligned}
\hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} \sum_{i=1}^n \log\left(1+\exp(-y_i\boldsymbol{w}^\intercal \boldsymbol{x}_i)\right) + \lambda C(\boldsymbol{w}) \\
&= \arg\max_{\boldsymbol{w}} \mathbb{P}(\boldsymbol{w}|\boldsymbol{X},Y)
\end{aligned}$$

8.4. SGD for L2-gregularized logistic regression

Initalize: $\boldsymbol{w}$

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

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

3: $\hat{\mathbb{P}}(Y = -y|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{(1+\exp(-y \cdot \boldsymbol{w}^\intercal \boldsymbol{x}))}$ ▷ compute prob. of misclassif. with cur. model

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

5: **end for**

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

9. Proofs

Proof. [def. 6.13] We need to only proof the second expression, as the first one is fulfilled anyway:

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

□

Proof. [def. 6.14]

$$\begin{aligned}
\mathrm{l}_n(\boldsymbol{w}) &= \arg\max_{\boldsymbol{w}} \mathrm{p}(y_{1:n}|\boldsymbol{x}_{1:n}, \boldsymbol{w}) = \arg\min_{\boldsymbol{w}} -\log \mathrm{p}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}) \\
&\overset{\text{i.i.d.}}{=} \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} -\log \mathrm{p}(y_i|\boldsymbol{x}_i, \boldsymbol{w}) \\
&\overset{\text{eq. (6.19)}}{=} -\log \frac{1}{1+\exp(-y_i \cdot \boldsymbol{w}^\intercal \boldsymbol{x}_i)} \\
&= \log\left(1 + \exp(-y_i \cdot \boldsymbol{w}^\intercal \boldsymbol{x}_i)\right) =: l_l(\boldsymbol{w})
\end{aligned}$$

□

Proof. [def. 6.15]

$$\begin{aligned}
\nabla_w l_l(w) &= \frac{\partial}{\partial w} \log\left(1 + \exp(-y \cdot w^\intercal \boldsymbol{x})\right) \\
&\overset{\text{C.R.}}{=} \frac{1}{(1+\exp(-y \cdot w^\intercal \boldsymbol{x}))} \frac{\partial}{\partial w} \left(1 + \exp(-y \cdot w^\intercal \boldsymbol{x})\right) \\
&\overset{\text{C.R.}}{=} \frac{1}{(1+\exp(-y \cdot w^\intercal \boldsymbol{x}))} \exp(-y \cdot w^\intercal \boldsymbol{x}) \cdot (-y\boldsymbol{x}) \\
&= \frac{e^{-z} \cdot (-yx)}{(1+e^{-z})} = \frac{-yx}{e^z(1+e^{-z})} = \frac{-yx}{(e^z + e^{-z+z})} \\
&= \frac{1}{\exp(y \cdot w^\intercal \boldsymbol{x}) + 1} \cdot (-yx) \\
&\overset{\text{eq. (6.19)}}{=} \hat{\mathbb{P}}(Y = -y|\boldsymbol{x}, w) \cdot (-y\boldsymbol{x})
\end{aligned}$$

□

Bayesian Inference/Modeling

Definition 6.16 Bayesian Inference: So far we only really looked at point estimators/estimates[def. 24.47].
But what if we are interested not only into the most likely value but also want to have a notion of the uncertainty of our prediction? Bayesian inference refers to statistical inference[def. 5.16], where uncertainty in inferences is quantified using probability. Thus we usually obtain a distribution over our parameters and not a single point estimates
⇒ can deduce statistical properties of parameters from their distributions.

Definition 6.17 $\mathrm{p}(w|\boldsymbol{y},\boldsymbol{X})/\mathrm{p}(w|\mathcal{D})$
Posterior Probability Distribution:
1. Specify the prior $\mathrm{p}_\lambda(w)$
2. Specify the likelihood $\mathrm{p}(\boldsymbol{y}|w,\boldsymbol{X})/\mathrm{p}(\mathcal{D}|w)$
3. Calculate the evidence $\mathrm{p}(\boldsymbol{y}|\boldsymbol{X})/\mathrm{p}(\mathcal{D})$
4. Calculate the posterior distribution $\mathbb{P}(w|\boldsymbol{y},\boldsymbol{X})/\mathrm{p}(w|\mathcal{D})$

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

Definition 6.18 $\mathrm{p}(\boldsymbol{y}|\boldsymbol{X})/\mathrm{p}(\mathcal{D})$
Marginal Likelihood [see proof 4]:
is the normalization constant that makes sure that the posterior distribution[def. 6.17] is an true probability distribution:

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{X}) = \int \mathrm{p}(\boldsymbol{y}|w,\boldsymbol{X})\cdot \mathrm{p}_\lambda(w)\,\mathrm{d}w = \int \text{Likelihood}\cdot\text{Prior}\,\mathrm{d}w \tag{6.24}$$

Note

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

Definition 6.19 $\mathrm{p}(\boldsymbol{f}_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y})/\mathrm{p}(\boldsymbol{f}_*|\boldsymbol{y})$ [see proof 4]
Posterior Predictive Distribution:
is the distribution of a real process $\boldsymbol{f}$ (i.e. $f(\boldsymbol{x}) = \boldsymbol{x}^\intercal w$) given:
- new observation(s) $\boldsymbol{x}_*$
- the posterior distribution[def. 6.17] of the observed data $\mathcal{D} = \{\boldsymbol{X},\boldsymbol{y}\}$
- The likelihood of a real process $\boldsymbol{f}_*$

$$\mathrm{p}(\boldsymbol{f}_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}) = \int \mathrm{p}(\boldsymbol{f}_*|\boldsymbol{x}_*,w)\cdot\mathrm{p}(w|\boldsymbol{X},\boldsymbol{y})\,\mathrm{d}w \tag{6.25}$$

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

Note $\boldsymbol{f}$ vs. $\boldsymbol{y}$

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

10. Types of Uncertainty

Definition 6.20 Epistemic/Systematic Uncertainty:
Is the uncertainty that is due to things that one could in principle know but does not i.e. only having a finite sub sample of the data. The epistemic noise will decrease the more data we have.

Definition 6.21 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 $\boldsymbol{X}$ data we will end up with different outcomes $\hat{y}$.
The aleatoric noise is *irreducible* as it is an underlying part of probabilistic models.

Bayesian Filtering

Definition 7.1
Recursive Bayesian Estimation/Filtering: Is a technique for estimating the an unknown probability distribution recursively over time by a measurement-[def. 7.3] and a process-model[def. 7.2] using Bayesian inference[def. 6.16].

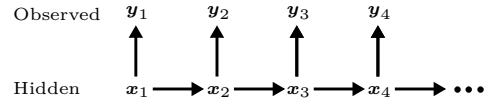


Figure 5: This problem corresponds to a *hidden Markov model (HMM)*??

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

Note

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

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

Corollary 7.1 Markov Property $\boldsymbol{x}_t \perp\!\!\!\perp \boldsymbol{x}_{1:t-2}|\boldsymbol{x}_{t-1}$**:** The process models[def. 7.2] is Markovian[def. 25.10] i.e. the current state depends only on the previous state:

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

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

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

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

Goal

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

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

Definition 7.4 Chapman-Kolmogorov eq. $\mathrm{p}(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$
Prior Update/Prediction Step [see section 4]:

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

Prior Distribution:

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

Definition 7.5 $\mathrm{p}(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$
Posterior Distribution/Update Step [see section 4]:

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

Definition 7.6 Normalization [see proof 4]:

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

Algorithm 7.1 Optimal Bayesian Filtering:
1: **Input:** $\mathrm{p}(\boldsymbol{x}_0)$
2: **while** Stopping Criterion not full-filed **do**
3: **Prediction Step:**

$$\mathrm{p}(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{1}{Z_t}\mathrm{p}(\boldsymbol{y}_t|\boldsymbol{x}_t)\mathrm{p}(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$$

4: **Update Step:**

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

with:

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

5: **end while**

Corollary 7.3 [see proof 4]
Joint Probability Distribution of (HMM): we can also calculate the joint probability distribution of the (HMM):

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

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

1. Kalman Filters

Definition 7.7 Kalman Filter Assumptions: Assumes a *linear*[def. 14.13] process model[def. 7.2], $\boldsymbol{q}$ with Gaussian model-noise $\boldsymbol{v}$ and a linear measurement model[def. 7.3] $\boldsymbol{h}$ with Gaussian process-noise $\boldsymbol{w}$.

Definition 7.8 Kalman Filter Model:
Process Model (7.8)

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

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

Measuremnt Model (7.9)

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

and define:

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

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

Note

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

Add Kalman algorithm (in slides Joesph Form I think)

Gaussian Processes (GP)

1. Gaussian Process Regression

1.1. Gaussian Linear Regression

Given

1. Linear Model with Gaussian Noise:

$$f(\boldsymbol{x}) = w^\intercal \boldsymbol{x} \qquad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2 \boldsymbol{I}\right) \tag{8.1}$$
$$\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon$$
$$\Rightarrow \text{Gaussian Likelihood:} \quad \mathrm{p}(\boldsymbol{y}|\boldsymbol{X},w) = \mathcal{N}\left(\boldsymbol{X}w, \sigma_n^2\boldsymbol{I}\right)$$

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

Sought

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

Definition 8.1 $\mathrm{p}(w|\boldsymbol{y},\boldsymbol{X}) = \mathcal{N}\left(\bar{w}, \Sigma_w^{-1}\right)$
Posterior Distribution proof 4:

$$\mu_w = \frac{1}{\sigma_n^2}\Sigma_w^{-1}\boldsymbol{X}\boldsymbol{y} \qquad \Sigma_w = \frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\intercal + \Sigma_p^{-1}$$

Note

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

Definition 8.2 $\mathrm{p}(f_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}) = \mathcal{N}(\mu_*, \Sigma_*)$
Posterior Predictive Distribution proof 4:

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

1.2. Kernelized Gaussian Linear Regression

Definition 8.3 Posterior Predictive Distribution:

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

Definition 8.4 Gaussian Process:

2. Model Selection

2.1. Marginal Likelihood

Approximate Inference

Problem

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

- Expectations
$$\mathbb{E}_{X\sim \mathrm{p}}[g(X)] = \int g(x)\mathrm{p}(x)\,\mathrm{d}x$$
- Normalization constants:
$$\mathrm{p}(\theta|y) = \frac{1}{Z}\mathrm{p}(\theta, y) \qquad Z = \int \mathrm{p}(y|\theta)\mathrm{p}(\theta)\,\mathrm{d}\theta = \int \mathrm{p}(\theta)\prod_{i=1}^{n}\mathrm{p}(y_i|\boldsymbol{x}_i,\theta)\,\mathrm{d}\theta$$

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

Definition 9.1 Approximate Inference: Is the procedure of finding an probability distribution q that approximates a true probability distribution p as well as possible.

1. Variational Inference

Definition 9.2 Bayes Variational Inference: Given an unormalized (posterior) probability distribution:
$$\mathrm{p}(\theta|y) = \frac{1}{Z}\mathrm{p}(\theta, y) \tag{9.1}$$
seeks an *approximate* probability distribution q_λ, that is parameterized by a *variational parameter* λ and approximates $\mathrm{p}(\theta|\boldsymbol{y})$ well.

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

1.1. Laplace Approximation

Definition 9.4 example 9.1/proofs 4,4,4
Laplace Approximation: Tries to approximate a desired probability distribution $\mathrm{p}(\theta|\,\mathcal{D})$ by a Gaussian probability distribution:
$$Q = \{q_\lambda(\theta) = \mathcal{N}(\lambda)\} = \mathcal{N}(\mu, \Sigma)\} \tag{9.2}$$
the distribution is given by:
$$q(\theta) = c\cdot\mathcal{N}(\theta;\lambda_1,\lambda_2) \tag{9.3}$$
$$\lambda_1 = \hat{\theta} = \arg\max_\theta \mathrm{p}(\theta|y)$$
with
$$\lambda_2 = \Sigma = H^{-1}\left(\hat{\theta}\right) = -\nabla\nabla_\theta \log \mathrm{p}(\hat{\theta}|y)$$

Corollary 9.1 : Taylor approximation of a function $\mathrm{p}(\theta|y) \in \mathcal{C}^k$ around its mode $\hat{\theta}$ naturally induces a Gaussian approximation. See proofs 4,4,4

1.2. Black Box Stochastic Variational Inference

The most common way of finding q_λ is by minimizing the KL-divergence[def. 5.8] between our approximate distribution q and our true posterior p:
$$q^* \in \arg\min_{q\in Q} \mathrm{KL}\left(q(\theta)\parallel \mathrm{p}(\theta|y)\right) = \arg\min_{\lambda\in\mathbb{R}^d} \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta|y))$$

Note

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

Definition 9.5 ELBO-Optimization Problem proof 4:
$$\begin{aligned} q_\lambda^* &\in \underset{\{\lambda: q_\lambda\in Q\}}{\arg\min}\ \mathrm{KL}\left(q_\lambda(\theta)\parallel \mathrm{p}(\theta|y)\right) \\ &= \underset{\{\lambda: q_\lambda\in Q\}}{\arg\max}\ \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y,\theta)] + H(q_\lambda) && (9.4)\\ &= \underset{\{\lambda: q_\lambda\in Q\}}{\arg\max}\ \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta)) && (9.5)\\ &:= \underset{\{\lambda: q_\lambda\in Q\}}{\arg\max}\ \mathrm{ELBO}(\lambda) && (9.6)\end{aligned}$$

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

Explanation 9.1.

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

1.3. Expected Lower Bound of Evidence (ELBO)

Definition 9.6 example 9.2/proof 4
Expected Lower Bound of Evidence (ELBO):
The evidence lower bound is a bound on the log prior:
$$\mathrm{ELBO}\left(q_\lambda\right) \leqslant \log \mathrm{p}(y) \tag{9.7}$$

1.3.1. Maximizing The ELBO

Definition 9.7 Gradient of the ELBO Loss:
$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.8)\\ &= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y,\theta)] + H(q_\lambda)\right]\\ &= \nabla_\lambda\left[\mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\right]\\ &= \nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\end{aligned}$$

Problem

In order to use SGD we need to evaluate the gradient of the loss:
$$\nabla_\lambda \mathbb{E}\left[l(\theta;\boldsymbol{x})\right] = \mathbb{E}\left[\nabla_{\boldsymbol{x}\sim \mathrm{p}} l(\theta;\boldsymbol{x})\right] = \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{x}\sim \mathrm{p}} l(\theta;\boldsymbol{x})$$
however in eq. (9.8) only second term can be derived easily. For the first term we cannot move the gradient inside the expectation as the expectations depends on the parameter w.r.t. which we differentiate:
$$\nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] = \frac{\partial}{\partial\lambda}\int q_\lambda \log \mathrm{p}(y|\theta)\,\mathrm{d}\theta$$
Solutions:

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

1.4. The Reparameterization Trick

Principle 9.1 proof 4
Reparameterization Trick: Let ϕ some base distribution from which we can sample and assume there exist an invertible function g s.t. $\theta = g(\epsilon,\lambda)$ then we can write θ in terms of a new distribution parameterized by $\epsilon\sim\phi(\epsilon)$:
$$\theta\sim q(\theta|\lambda) = \phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1} \tag{9.9}$$
we can then write by the law of the unconscious statistician law 22.6:
$$\mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] = \mathbb{E}_{\epsilon\sim\phi}[\log \mathrm{p}(y|g(\epsilon;\lambda))] \tag{9.10}$$
$\Rightarrow$ the expectations does not longer depend on λ and we can pull in the gradient!
$$\begin{aligned}\nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] &= \nabla_\epsilon \mathbb{E}_{\theta\sim\phi}[\log \mathrm{p}(y|g(\epsilon;\lambda))] && (9.11)\\ &= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))] && (9.12)\end{aligned}$$

Definition 9.8 example 9.3
Reparameterized ELBO Gradient[def. 9.7]**:**
By using the reparameterization trick principle 9.2 we can write the gradient of the ELBO as:
$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.13)\\ &= \nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\\ &= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\end{aligned}$$

Corollary 9.2 proof 9.3
Reparameterized ELBO for Gaussians:
$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.14)\\ &= \nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\nabla_{\boldsymbol{C},\mu}\log \mathrm{p}(y|\boldsymbol{C}\epsilon+\boldsymbol{\mu})\right]\\ &\quad - \nabla_{\boldsymbol{C},\mu}\mathrm{KL}\left(q_{\boldsymbol{C},\boldsymbol{\mu}}\parallel \mathrm{p}(\theta)\right)\\ &\approx \frac{n}{m}\sum_{j=1}^{m}\nabla_{\boldsymbol{C},\mu}\log \mathrm{p}\left(y_{i_j}|\boldsymbol{C}\epsilon^j+\boldsymbol{\mu},\boldsymbol{x}_{i_j}\right)\\ &\quad - \nabla_{\boldsymbol{C},\mu}\mathrm{KL}\left(q_{\boldsymbol{C},\boldsymbol{\mu}}\parallel \mathrm{p}(\theta)\right)\end{aligned}$$

2. Markov Chain Monte Carlos Methods

Principle 9.2 proof 4
Reparameterization Trick: Let ϕ some base distribution from which we can sample and assume there exist an invertible function g s.t. $\theta = g(\epsilon,\lambda)$ then we can write θ in terms of a new distribution parameterized by $\epsilon\sim\phi(\epsilon)$:
$$\theta\sim q(\theta|\lambda) = \phi(\epsilon)|\nabla_\epsilon g(\epsilon;\lambda)|^{-1} \tag{9.15}$$
we can then write by the law of the unconscious statistician law 22.6:
$$\mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] = \mathbb{E}_{\epsilon\sim\phi}[\log \mathrm{p}(y|g(\epsilon;\lambda))] \tag{9.16}$$
$\Rightarrow$ the expectations does not longer depend on λ and we can pull in the gradient!
$$\begin{aligned}\nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] &= \nabla_\epsilon \mathbb{E}_{\theta\sim\phi}[\log \mathrm{p}(y|g(\epsilon;\lambda))] && (9.17)\\ &= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))] && (9.18)\end{aligned}$$

Definition 9.9 example 9.3
Reparameterized ELBO Gradient[def. 9.7]**:**
By using the reparameterization trick principle 9.2 we can write the gradient of the ELBO as:
$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.19)\\ &= \nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\\ &= \mathbb{E}_{\epsilon\sim\phi}[\nabla_\lambda \log \mathrm{p}(y|g((\epsilon;\lambda)))] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\end{aligned}$$

Corollary 9.3 proof 9.3
Reparameterized ELBO for Gaussians:
$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathrm{ELBO}(\lambda) && (9.20)\\ &= \nabla_\lambda \mathbb{E}_{\theta\sim q_\lambda}[\log \mathrm{p}(y|\theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda(\theta)\parallel \mathrm{p}(\theta))\\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\nabla_{\boldsymbol{C},\mu}\log \mathrm{p}(y|\boldsymbol{C}\epsilon+\boldsymbol{\mu})\right]\\ &\quad - \nabla_{\boldsymbol{C},\mu}\mathrm{KL}\left(q_{\boldsymbol{C},\boldsymbol{\mu}}\parallel \mathrm{p}(\theta)\right)\\ &\approx \frac{n}{m}\sum_{j=1}^{m}\nabla_{\boldsymbol{C},\mu}\log \mathrm{p}\left(y_{i_j}|\boldsymbol{C}\epsilon^j+\boldsymbol{\mu},\boldsymbol{x}_{i_j}\right)\\ &\quad - \nabla_{\boldsymbol{C},\mu}\mathrm{KL}\left(q_{\boldsymbol{C},\boldsymbol{\mu}}\parallel \mathrm{p}(\theta)\right)\end{aligned}$$

3. Bayesian Neural Networks (BNN)

Definition 9.10 Bayesian Neural Networks (BNN):

(1) Model the prior over our weights $\theta = \left[\boldsymbol{W}^0\cdots\cdots\boldsymbol{W}^L\right]$ by a neural network:
$$\theta\sim \mathrm{p}_\lambda(\theta) = \boldsymbol{F} \quad \text{with} \quad \boldsymbol{F} = \boldsymbol{F}^L\circ\cdots\circ\boldsymbol{F}^1,\ \boldsymbol{F}^l = \varphi\circ\bar{\boldsymbol{F}}^l = \varphi\left(\boldsymbol{W}^l\boldsymbol{x}+b^l\right)$$
for each weight $w_{k,j}^{(0)}$ of input x_j with weight on the hidden variable $z_k^{(0)}$ with $a_i^0 = \varphi\left\{\boldsymbol{z}_i^{(0)}\right\}$ it follows:
$$w_{k,j}^{(0)} = \mathrm{p}_w\left(\lambda_{k,j}\right) \overset{\text{i.e.}}{=} \mathcal{N}\left(\mu_{k,j},\sigma_{k,j}^2\right)$$

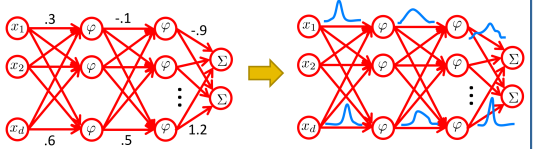


Figure 6

(2) The parameters of likelihood function are modeled by the output of the network:
$$\mathrm{p}(y|F(\theta,\boldsymbol{X})) \qquad \text{see example 9.4} \tag{9.21}$$

Note

Recall for normal Bayesian Linear regression we had:

Problem

All the weights of the prior $\mathrm{p}_\lambda(\theta) = \boldsymbol{F}$ are correlated in some complex way see Figure 6. Thus even if the prior and likelihood are simple, the posterior will be not. $\Rightarrow$ need to approximate the posterior $\mathrm{p}(\theta|\boldsymbol{y},\boldsymbol{X})$ i.e. by fitting a Gaussian distribution to each weight of the posterior neural network.

3.0.1. MAP estimates for BNN

Definition 9.11 BNN MAP Estimate: We need to do a forward pass for each $\boldsymbol{x}_i$ in order to obtain $\boldsymbol{\mu}(\boldsymbol{x}_i;\theta)$ and $\sigma\left(\boldsymbol{x}_i;\theta\right)^2$:
$$\theta^* = \arg\max_\theta\left\{\mathrm{p}(\theta|\boldsymbol{X},\boldsymbol{y})\right\} \overset{\text{eq. (6.13)}}{=} \arg\min_\theta \boldsymbol{\lambda}\|\theta\|_2^2 - \sum_{i=1}^{n}\left(\frac{1}{2\sigma\left(\boldsymbol{x}_i;\theta\right)^2}\|y_i-\boldsymbol{\mu}(\boldsymbol{x}_i;\theta)\|^2 + \frac{1}{2}\log\sigma\left(\boldsymbol{x}_i;\theta\right)^2\right)$$

Explanation 9.2. *[def. 9.11]*

- $\frac{1}{2}\log\sigma\left(\boldsymbol{x}_i;\theta\right)^2$*: tries to force neural network to predict small uncertainty*
- $\frac{1}{2\sigma\left(\boldsymbol{x}_i;\theta\right)^2}\|y_i-\boldsymbol{\mu}(\boldsymbol{x}_i;\theta)\|^2$*: 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 9.12 proof 4
MAP Gradient of BNN:
$$\theta_{t+1} = \theta_t\left(1-2\lambda\eta_t\right) - \eta_t\nabla\sum_{i=1}^{n}\log \mathrm{p}(y_i|\boldsymbol{x}_i,\theta) \tag{9.22}$$

Note

- The gradients of the objective eq. (9.22) can be calculated using auto-differentiation techniques e.g. Pytorch or Tensorflow.
- The BNN MAP estimate fails to predict epistemic uncertainty[def. 6.20] $\Longleftrightarrow$ it is overconfident in regions where we haven not even seen any data.
  $\Rightarrow$ need to use Bayesian approach to approximate posterior distribution.

3.1. Variational Inference For BNN

We use the objective eq. (9.20) as loss in order to perform back propagation.

3.2. Making Predictions

Proposition 9.1Title:

4. Proofs

Proof. Definition 6.19:

$$p(f_*|x_*,X,y) = \frac{p(f_*,x_*,X,y)}{p(x_*,X,y)} = \frac{\int p(f_*,x_*,X,y,w)\,dw}{p(x_*,X,y)} \overset{(22.19)}{=} \frac{\int p(f_*|x_*,X,y,w)p(x_*,X,y,w)\,dw}{p(x_*,X,y)} \overset{(22.19)}{=} \frac{\int p(f_*|x_*,X,y,w)p(w|x_*,X,y)\cancel{p(x_*,X,y)}\,dw}{\cancel{p(x_*,X,y)}} = \int p(f_*|x_*,X,y,w)p(w|x_*,X,y)\,dw \overset{\clubsuit}{=} \int p(f_*|x_*,w)p(w|X,y)\,dw$$

□

Note ♣

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

Proof. Definition 6.18:

$$p(y|X) = \int p(y,w|X)\,dw = \int p(y|w,X)p(w|X)\,dw \overset{(6.6)}{=} \int p(y|w,X)p(w)\,dw$$

□

Proof. Definition 7.4:

$$p(x_t,x_{t-1}|y_{1:t_1}) \overset{(22.19)}{=} p(x_t|x_{t-1},y_{1:t_1})p(x_{t-1}|y_{1:t_1}) \overset{\text{independ.}}{=} p(x_t|x_{t-1})p(x_{t-1}|y_{1:t_1})$$

marginalization/integration over x_{t-1} gives the desired result. □

Proof. Definition 7.5:

$$p(x_t,y_t|y_{1:t-1}) \overset{(22.23)}{=} \begin{cases} p(x_t|y_t,y_{1:t-1})p(y_t|y_{1:t-1}) \\ p(y_t|x_t,y_{1:t-1})p(x_t|y_{1:t-1}) \\ \dots\dots\dots\dots \end{cases}$$

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

from which follows immediately eq. (7.5). □

Proof. Definition 7.6:

$$p(y_t|y_{1:t-1}) = \int p(y_t,x_t|y_{1:t-1})\,dx_t = \int p(y_t|x_t,y_{1:t-1})p(x_t|y_{1:t-1})\,dx_t \overset{\text{corollary 7.2}}{=} \int p(y_t|x_t)p(x_t|y_{1:t-1})\,dx_t$$

□

Proof. corollary 7.3:

$$p(x_{1:t},y_{1:t}) \overset{(22.19)}{=} p(y_{1:t}|x_{1:t})p(x_{1:t}) \overset{\text{law 22.2}}{=} p(y_{1:t}|x_{1:t})p(x_t|x_{t-1:0})\cdots p(x_2|x_1)p(x_1) \overset{(7.1)}{=} p(y_{1:t}|x_{1:t})\left(p(x_1)\prod_{2=1}^{t}p(x_i|x_{i-1})\right) \overset{\text{law 22.2}}{\underset{\text{corollary 7.2}}{=}} \left(p(y_1|x_1)\cdots p(y_t|x_t)\right)\left(p(x_1)\prod_{2=1}^{t}p(x_i|x_{i-1})\right) = p(y_1|x_1)p(x_1)\prod_{2=1}^{t}p(y_i|x_i)p(x_i|x_{i-1})$$

□

Proof. [def. 8.1]

$$p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w) \propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}(y-Xw)^\intercal(y-Xw)\right)\exp\left(-\frac{1}{2}w^\intercal\Sigma^{-1}w\right) \propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}\left(y^\intercal y - 2w^\intercal X^\intercal y + w^\intercal X^\intercal X^\intercal w + \sigma_n^2 w^\intercal\Sigma^{-1}w\right)\right\} \propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}\left(y^\intercal y - 2w^\intercal X^\intercal y + w^\intercal\left(X^\intercal X^\intercal + \sigma_n^2\Sigma^{-1}\right)w\right)\right\}$$

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

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

Σ_w follows directly $\Sigma_w = \sigma_n^{-2}XX^\intercal + \Sigma_p$

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

Proof. [def. 8.2]

□

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

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

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

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

we can no derive the distribution:

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

□

Notes

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

Proof. [def. 9.4] 2D:

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

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

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

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

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

□

Proof. [def. 9.4] k-dimensional:

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

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

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

□

Proof. [def. 9.5]

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

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

$$= \arg\min_q \mathbb{E}_{\theta\sim q}\left[\log\frac{q(\theta)}{\frac{1}{Z}p(\theta,y)}\right] = \arg\min_q \mathbb{E}_{\theta\sim q}\left[\log q(\theta) - \log\frac{1}{Z} - \log p(\theta,y)\right] = \arg\min_q \mathbb{E}_{\theta\sim q} - \underbrace{[-\log q(\theta)]}_{H(q)} + \cancel{\mathbb{E}_{\theta\sim q}[\log Z]} - \mathbb{E}_{\theta\sim q}[\log p(\theta,y)]$$

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

□

Proof. [def. 9.6]

$$\log p(y) = \log\int p(y,\theta)\,d\theta = \log\int p(y|\theta)p(\theta)\,d\theta = \log\int p(y|\theta)\frac{p(\theta)}{q_\lambda(\theta)}q_\lambda(\theta)\,d\theta = \log\mathbb{E}_{\theta\sim q_\lambda}\left[p(y|\theta)\frac{p(\theta)}{q_\lambda(\theta)}\right] \overset{(22.54)}{\geqslant} \mathbb{E}_{\theta\sim q_\lambda}\left[\log\left(p(y|\theta)\frac{p(\theta)}{q_\lambda(\theta)}\right)\right] = \mathbb{E}_{\theta\sim q_\lambda}\left[\log p(y|\theta) - \log\frac{p(\theta)}{q_\lambda(\theta)}\right] = \mathbb{E}_{\theta\sim q_\lambda}\left[\log p(y|\theta)\right] - \mathrm{KL}\left(q_\lambda\parallel p(\cdot)\right)$$

□

Proof. principle 9.2 Let:

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

then it follows immediately with eq. (22.46):

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

$\Rightarrow$ parameterized in terms of ϵ □

Proof. [def. 9.12]

$$\theta_{t+1} = \theta_t - \eta_t\left(\nabla\log p(\theta) - \nabla\sum_{i=1}^{n}\log p(y_i|x_i,\theta)\right) = \theta_t - \eta_t\left(2\lambda\theta_t - \nabla\sum_{i=1}^{n}\log p(y_i|x_i,\theta)\right) = \theta_t(1-2\lambda\eta_t) - \eta_t\nabla\sum_{i=1}^{n}\log p(y_i|x_i,\theta)$$

□

5. Examples

Example 9.1 Laplace Approximation
Logistic Regression Likelihood + Gaussian Prior:

Example 9.2 ELBO Bayesian Logistic Regression: Suppose:

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

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

Then it follows for the terms of the ELBO:

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

$$\mathbb{E}_{\theta\sim q_\lambda}\left[p(y|\theta)\right] = \mathbb{E}_{\theta\sim q_\lambda}\left[\sum_{i=1}^{n}\log p(y_i|\theta,x_i)\right] = \mathbb{E}_{\theta\sim q_\lambda}\left[-\sum_{i=1}^{n}\log\left(1+\exp\left(-y_i\theta^\intercal x_i\right)\right)\right]$$

Example 9.3 ELBO Gradient Gaussian: Suppose:

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

we can reparameterize using principle 9.2 by using:

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

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

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

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:

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

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

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

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

Example 9.4 BNN Likelihood Function Examples:

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

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 10.1 Similarity Measure 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 10.2 Dissimilarity Measure 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:

$$\|\boldsymbol{x}-\boldsymbol{y}\|^2 = \boldsymbol{x}^\intercal\boldsymbol{x} + \boldsymbol{y}^\intercal\boldsymbol{y} - 2\boldsymbol{x}^\intercal\boldsymbol{y} \qquad \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d \tag{10.1}$$
$$\operatorname{dissim}(\boldsymbol{x},\boldsymbol{y}) = \operatorname{sim}(\boldsymbol{x},\boldsymbol{x}) + \operatorname{sim}(\boldsymbol{y},\boldsymbol{y}) - 2\operatorname{dissim}(\boldsymbol{x},\boldsymbol{y})$$

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. (10.1) becomes $2\boldsymbol{x}^\intercal\boldsymbol{y} = 2\operatorname{dissim}(\boldsymbol{x},\boldsymbol{y})$

Definition 10.3 Feature Map ϕ: is a mapping $\phi:\mathcal{X}\mapsto\mathcal{Y}$ that takes an input $\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d$ and maps it into another feature space $\mathcal{Y}\subseteq\mathbb{R}^D$.

Note

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

$$D = \dim(\mathcal{Y}) = \binom{p+d}{d} = \mathcal{O}(d^p) \tag{10.2}$$

when using the polynomial kernel[def. 10.10], this can be reduced to the order d.

Definition 10.4 Kernel k: **Let** $\mathcal{X}\subseteq\mathbb{R}^d$ be the data space. A map $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ is called kernel if their exists an inner product space[def. 17.19] called **feature space** $(\mathcal{Y},\langle\cdot,\cdot\rangle_\mathcal{Y})$ and a map $\phi:\mathcal{X}\mapsto\mathcal{Y}$ s.t.

$$k(\boldsymbol{x},\boldsymbol{y}) = \langle\phi(\boldsymbol{x}),\phi(\boldsymbol{y})\rangle_\mathcal{Y} \qquad \forall\boldsymbol{x},\boldsymbol{y}\in\mathcal{X} \tag{10.3}$$

Corollary 10.1 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 $k(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{y}$ **thus** the kernel measures the similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$ by the inner product $\boldsymbol{x}^\intercal\boldsymbol{y}$ weighted by the matrix $\boldsymbol{A}$.

Corollary 10.2 Kernels and distance: **Let** $k(\boldsymbol{x},\boldsymbol{y})$ be a measure of similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$ then k induces a dissimilarity/distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ defined as the difference betweend the self-similarities $k(\boldsymbol{x},\boldsymbol{x}) + k(\boldsymbol{y},\boldsymbol{y})$ and the cross-similarities $k(\boldsymbol{x},\boldsymbol{y})$:

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

Note

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

1. The Gram Matrix

Definition 10.5 Kernel (Gram) Matrix:
Given: a mapping $\phi:\mathbb{R}^d\mapsto\mathbb{R}^D$ and a corresponding kernel function $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ with $\mathcal{X}\subseteq\mathbb{R}^d$.
Let S be any finite subset of data $S=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}\subseteq\mathcal{X}$.
Then the kernel matrix $\mathcal{K}:\in\mathbb{R}^{n\times n}$ is defined by:

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

Corollary 10.3 $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal$
Kernel Eigenvector Decomposition:
For any symmetric matrix (Gram matrix $\mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)|_{i,j=1}^n$) there exists an eigenvector decomposition:

$$\mathcal{K} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal \tag{10.4}$$

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

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

$$\phi:\boldsymbol{x}_i\mapsto\left(\sqrt{\lambda_t}\boldsymbol{v}_{t,i}\right)_{t=1}^n\in\mathbb{R}^n, \qquad i=1,\dots,n \tag{10.5}$$

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

$$\phi^\intercal(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j) = \sum_{t=1}^n \lambda_t\boldsymbol{v}_{t,i}\boldsymbol{v}_{t,j} = (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal)_{i,j} = \mathcal{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{10.6}$$

1.1. Necessary Properties

Property 10.1 Inner Product Space:
k must be an *inner product* of a suitable space $\mathcal{Y}$.

Property 10.2 Symmetry: k/$\mathcal{K}$ must be symmetric:

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

Property 10.3 Non-negative Eigenvalues/p.s.d.s Form:
Let $S=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$ be an n-set of a *finite* input space $\mathcal{Y}$. A kernel k must induces a *p.s.d. symmetric* kernel matrix k for any possible $S\subseteq\mathcal{X}$ see section 1.
$\iff$ all eigenvalues of the kernel gram matrix $\mathcal{K}$ for *finite* $\mathcal{Y}$ must be non-negative corollary 17.2.

Notes

- The extension to infinite dimensional Hilbert Spaces might also include a non-negative weighting/eigenvalues:
$$\langle\phi(\boldsymbol{x}),\phi(\boldsymbol{z})\rangle = \sum_{i=1}^\infty \lambda_i\phi_i(\boldsymbol{x})\phi_i(\boldsymbol{z})$$
- In order to be able to use a kernel, we need to verify that the kernel is p.s.d. for all n-vectors $\mathcal{X}=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$, as well as for future unseen values.

2. Mercers Theorem

Theorem 10.1 Mercers Theorem: Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$ and $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ a kernel function.
Then one can expand k in a uniformly convergent series of bounded functions ϕ s.t.

$$k(\boldsymbol{x},\boldsymbol{x}') = \sum_{i=1}^\infty \lambda\phi(\boldsymbol{x})\phi(\boldsymbol{x}') \tag{10.7}$$

Theorem 10.2 General Mercers Theorem: Let Ω be a compact subset of R^n. Suppose k is a gernal continuous symmetric function such that the integral operator:

$$T_k : L_2(\boldsymbol{X})\mapsto L_2(\boldsymbol{X}) \qquad (T_kf)(\cdot) = \int_\Omega k(\cdot,\boldsymbol{x})f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{10.8}$$

is positve, that is it satisfies:

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

Then we can expand $k(\boldsymbol{x},\boldsymbol{z})$ 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 10.6 Kernel Trick: If a kernel has an analytic form we do no longer need to calculate:
- the function mapping $\boldsymbol{x}\mapsto\phi(\boldsymbol{x})$ and
- the inner product $\phi(\boldsymbol{x})^\intercal\phi(\boldsymbol{y})$

explicitly but simply us the formula for the kernel:

$$\phi(\boldsymbol{x})^\intercal\phi(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{y}) \tag{10.9}$$

see examples 10.1 and 10.2

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

$$k(\boldsymbol{x},\boldsymbol{y}) = k(\boldsymbol{x}-\boldsymbol{y}) \tag{10.10}$$

see example example 10.3

4.2. Isotropic Kernels

Definition 10.8 Isotropic Kernel: A isotropic kernel is a kernel that only considers distance differences:

$$k(\boldsymbol{x},\boldsymbol{y}) = k(\|\boldsymbol{x}-\boldsymbol{y}\|_2) \tag{10.11}$$

Corollary 10.4 :
Isotropic $\rightarrow$ Stationary

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

5.1. The Linear Kernel

Definition 10.9 Linear/String Kernel:

$$k(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\intercal\boldsymbol{y} \tag{10.12}$$

5.2. The Polynomial Kernel

Definition 10.10 Polynomial Kernel:
represents all monomials[def. 14.2] of degree up to m

$$k(\boldsymbol{x},\boldsymbol{y}) = (1+\boldsymbol{x}^\intercal\boldsymbol{y})^m \tag{10.13}$$

5.3. The Sigmoid Kernel

Definition 10.11 Sigmoid/tanh Kernel:

$$k(\boldsymbol{x},\boldsymbol{y}) = \tanh\kappa\boldsymbol{x}^\intercal\boldsymbol{y} - b \tag{10.14}$$

5.4. The Exponential Kernel

Definition 10.12 Exponential Kernel:
is an continuous kernel that is non-differential $k\in\mathcal{C}^0$:

$$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_1}{\theta}\right) \tag{10.15}$$

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

5.5. The Gaussian Kernel

Definition 10.13 Gaussian/Squared Exp. Kernel/ Radial Basis Functions (RBF):
Is an inifite dimensional smooth kernel $k\in\mathcal{C}^\infty$ with some usefull properties

$$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\theta^2}\right) \approx \begin{cases}1 & \text{if } \boldsymbol{x} \text{ and } \boldsymbol{y} \text{ close}\\ 0 & \text{if } \boldsymbol{x} \text{ and } \boldsymbol{y} \text{ far away}\end{cases} \tag{10.16}$$

Explanation 10.1 (Threshold θ). $2\theta\in\mathbb{R}$ *corresponds to a threshold that determines how close input values need to be in order to be considered similar:*

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

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 $k(\boldsymbol{x},\boldsymbol{x}) = e^0 = 1$.
Thus: Data points in input space are projected onto a high-(infintie-)dimensional sphere in feature space.
Classification: Cutting with hyperplanes through the sphere. **How to chose** h: good heuristics, take median of the distance all points but better is cross validation.

5.6. The Matern Kernel

When looking at actual data/sample paths the smoothness of the Gaussian kernel[def. 10.13] is often a too strong assumption that does not model reality the same holds true for the non-smoothness of the exponential kernel[def. 10.12]. A solution to this dilemma is the Matern kernel.

Definition 10.14 Matern Kernel: is a kernel which allows you to specify the level of smoothness $k\in\mathcal{C}^{\lceil\nu\rceil}$ by a positive parameter ν:

$$k(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\rho}\right)^\nu \mathcal{K}_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\rho}\right) \tag{10.17}$$

$\nu,\rho\in\mathbb{R}_+$
$\mathcal{K}_\nu$ 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:

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

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

$$k(\boldsymbol{x},\boldsymbol{x}') = k_1(\boldsymbol{x},\boldsymbol{x}') + k_2(\boldsymbol{x},\boldsymbol{x}') \tag{10.18}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_1(\boldsymbol{x},\boldsymbol{x}')\cdot k_2(\boldsymbol{x},\boldsymbol{x}') \tag{10.19}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = \alpha\,k_1(\boldsymbol{x},\boldsymbol{x}') \qquad \alpha\in\mathbb{R}_+ \tag{10.20}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = f(\boldsymbol{x})f(\boldsymbol{x}') \tag{10.21}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_3(\phi(\boldsymbol{x}),\phi(\boldsymbol{x}')) \tag{10.22}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = p\left(k(\boldsymbol{x},\boldsymbol{x}')\right) \tag{10.23}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = \exp\left(k(\boldsymbol{x},\boldsymbol{x}')\right) \tag{10.24}$$

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
k_3	a Kernel over $\mathbb{R}^e\times\mathbb{R}^e$

Proofs

Proof. Property 10.3**The kernel matrix is positive-semidefinite**:
Let $\phi:\mathcal{X}\mapsto\mathbb{R}^d$ and $\boldsymbol{\Phi} = \begin{bmatrix}\phi(\boldsymbol{x}_1) & \dots & \phi(\boldsymbol{x}_n)\end{bmatrix}^\intercal\in\mathbb{R}^{d\times n}$.
Thus: $\mathcal{K} = \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\in\mathbb{R}^{n\times n}$.

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

□

Examples

Example 10.1 Calculating the Kernel by hand:

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

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

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

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

Operation Count:

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

Example 10.2

Calculating the Kernel using the Kernel Trick:

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

Operation Count:

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

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

Example 10.3 Stationary Kernels:

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

is a stationary but not an isotropic kernel.

Math Appendix

Logic

Set Theory

Definition 12.1 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 12.2 Empty Set $\{\}/\varnothing$**:** is the unique set having no elements/cardinality[def. 12.4] zero.

Definition 12.3 Multiset/Bag: Is a set-like object in which multiplicity matters, that is we can have multiple elements of the same type.
I.e. $\{1, 1, 2, 3\} \neq \{1, 2, 3\}$

Definition 12.4 Cardinality $|S|$**:** Is the number of elements that are contained in a set.

Definition 12.5 The Power Set $\mathcal{P}(S)/2^S$**:** The power set of any set S is the set of all subsets of S, including the empty set and S itself. The cardinality of the power set is 2^S is equal to $2^{|S|}$.

Definition 12.6 Closure: A set is *closed* under an operation Ω if performance of that operations onto members of the set always produces a member of that set.

1. Number Sets

1.1. The Real Numbers $\mathbb{R}$

1.1.1. Intervals

Definition 12.7 Closed Interval $[a, b]$**:** The closed interval of a and b is the set of all real numbers that are within a and b, including a and b:

$$[a, b] = \{x \in \mathbb{R} \,|\, a \leqslant x \leqslant b\} \tag{12.1}$$

Definition 12.8 Open Interval (a, b)**:** The open interval of a and b is the set of all real numbers that are within a and b:

$$(a, b) = \{x \in \mathbb{R} \,|\, a < x \leqslant <\} \tag{12.2}$$

1.2. The Rational Numbers $\mathbb{Q}$

Example 12.1 Power Set/Cardinality of $S = \{x, y, z\}$**:** The subsets of S are:
$\{\varnothing\}$, $\{x\}$, $\{y\}$, $\{z\}$, $\{x, y\}$, $\{x, z\}$, $\{y, z\}$, $\{x, y, z\}$ and hence the power set of S is $\mathcal{P}(S) = \{\{\varnothing\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$ with a cardinality of $|S| = 2^3 = 8$.

2. Set Functions

2.1. Submoduluar Set Functions

Definition 12.9 Submodular Set Functions: A submodular function $f : 2^{\Omega} \mapsto \mathbb{R}$ is a function that satisifies:

$$f(A \cup \{x\}) - f(A) \geqslant f(B \cup \{X\}) - F(B) \qquad \begin{matrix} \forall A \subseteq B \subset \Omega \\ \{x\} \in \Omega \backslash B \end{matrix} \tag{12.3}$$

Explanation 12.1 (Definition 12.9)**.** *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 12.10 Montone Submodular Function: A *monotone* submodular function is a submodular function[def. 12.9] that satisifies:

$$f(A) \leqslant f(B) \qquad \forall A \subseteq B \subseteq \Omega \tag{12.4}$$

Explanation 12.2 (Definition 12.10)**.** *Adding more elements to a set will always increase the information/value gain.*

Sequences&Series

Definition 13.1 Index Set: Is a set[def. 12.1] 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.

$$f : A \mapsto S \qquad A \in \mathbb{N} \tag{13.1}$$

Definition 13.2 Sequence $(a_n)_{n \in A}$**:** is an by an index set A *enumerated* multiset[def. 12.3] (repetitions are allowed) of objects in which *order does matter*.

Definition 13.3 Series: is an infinite ordered set of terms combined together by addition.

1. Types of Sequences

1.1. Arithmetic Sequence

Definition 13.4 Arithmetic Sequence: Is a sequence where the *difference* between two consecutive terms constant i.e. $(2, 4, 6, 8, 10, 12, \ldots)$.

$$t_n = t_0 + nd \qquad d \text{ :difference between two terms} \tag{13.2}$$

1.2. Geometric Sequence

Definition 13.5 Geometric Sequence: Is a sequence where the *ratio* between two consecutive terms constant i.e. $(2, 4, 8, 16, 32, \ldots)$.

$$t_n = t_0 \cdot r^n \qquad r \text{ :ratio between two terms} \tag{13.3}$$

Calculus and Analysis

1. Building Blocks of Analysis

1.1. Polynomials

Definition 14.1 Polynomial: A function $\mathcal{P}_n : \mathbb{R} \mapsto \mathbb{R}$ is called *Polynomial*, if it can be represented in the form:

$$\mathcal{P}_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + a_n x^n \tag{14.1}$$

Corollary 14.1 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 14.2 Monomial: Is a polynomial with only one term.

Definition 14.3 Quadratic Formula: $ax^2 + bx + c = 0$ or in reduced form: $x^2 + px + q = 0$ with $p = b/a$ and $q = c/a$

Definition 14.4 Discriminant: $\delta = b^2 - 4ac$

Definition 14.5 Solution to [def. 14.3]**:**

$$x_\pm = \frac{-b \pm \sqrt{\delta}}{2a} \quad \textbf{or} \quad x_\pm = \frac{1}{2}\left(-p \pm \sqrt{p^2 - 4q}\right)$$

Theorem 14.1
Fist Fundamental Theorem of Calculus: Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined $\forall x \in [a, b]$ by:

$$F(X) = \int_a^x f(t)\,\mathrm{d}t \tag{14.2}$$

Then it follows:

$$F'(x) = f(x) \qquad \forall x \in (a, b) \tag{14.3}$$

Theorem 14.2
Second Fundamental Theorem of Calculus: Let f be a real-valued function on a closed interval $[a, b]$ and F an antiderivative of f in $[a, b]$: $F'(x) = f(x)$, then it follows if f is Riemann integrable on $[a, b]$:

$$\int_a^b f(t)\,\mathrm{d}t = F(b) - F(a) \iff \int_a^x \frac{\partial}{\partial x} F(t)\,\mathrm{d}t = F(x) \tag{14.4}$$

Definition 14.6 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 – $\texttt{dom}(f)$.

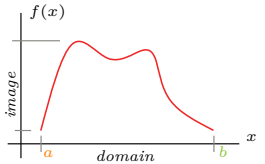
Definition 14.7
Codomain/target set of a function $\texttt{codom}(\cdot)$**:**
Given a function $f : \mathcal{X} \to \mathcal{Y}$, the codaomain of that function is the set $\mathcal{Y}$ into which all of the output of the function is **constrained** to fall.

Definition 14.8 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:

$$\{y \in \mathcal{Y} | y = f(x), \quad \forall x \in \mathcal{X}\} := f[\mathcal{X}] \tag{14.5}$$

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.



Definition 14.9 Inverse Image/Preimage $f^{-1}(\cdot)$**:**
Let $f : X \mapsto Y$ be a function, and A a subset set of its codomain Y.
Then the preimage of A under f is the set of all elements of the domain X, that map to elements in A under f:

$$f^{-1}(A) = \{x \subseteq X : f(x) \subseteq A\} \tag{14.6}$$

Example 14.1 :
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:

$$f[A] = \{y \in \mathcal{Y} | y = f(x), \quad \forall x \in A\} \tag{14.7}$$

Note: Range

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

Definition 14.10 (strictly) Increasing Functions:
A function f is called **monotonically increasing/ increasing/non-decreasing** if:

$$x \leqslant y \iff f(x) \leqslant f(y) \qquad \forall x, y \in \texttt{dom}(f) \tag{14.8}$$

And **strictly increasing** if:

$$x < y \iff f(x) < f(y) \qquad \forall x, y \in \texttt{dom}(f) \tag{14.9}$$

Definition 14.11 (strictly) Decreasing Functions:
A function f is called monotonically decreasing/decreasing or non-increasing if:

$$x \geqslant y \iff f(x) \geqslant f(y) \qquad \forall x, y \in \texttt{dom}(f) \tag{14.10}$$

And *strictly* decreasing if:

$$x > y \iff f(x) > f(y) \qquad \forall x, y \in \texttt{dom}(f) \tag{14.11}$$

Definition 14.12 Monotonic Function: A function f is called monotonic iff either f is increasing or decreasing.

Definition 14.13 Linear Function:
A function $L : \mathbb{R}^n \mapsto \mathbb{R}^m$ is linear if and only if:

$$L(\boldsymbol{x} + \boldsymbol{y}) = L(\boldsymbol{x}) + L(\boldsymbol{y})$$
$$L(\alpha \boldsymbol{x}) = \alpha L(\boldsymbol{x}) \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}$$

Corollary 14.2 Linearity of Differentiation: The derivative of **any** linear combination of functions equals the same linear combination of the derivatives of the functions:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(af(x) + bg(x)\right) = a\frac{\mathrm{d}}{\mathrm{d}x}f(x) + b\frac{\mathrm{d}}{\mathrm{d}x}g(x) \qquad a, b \in \mathbb{R} \tag{14.12}$$

Definition 14.14 Quadratic Function:
A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is quadratic if it can be written in the form:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\intercal \boldsymbol{x} + c \tag{14.13}$$

2. Continuity and Smoothness

Definition 14.15 Continuous Function:

Definition 14.16 Smoothness of a Function
tcblack$\mathcal{C}^k$**: 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:

$$f \in \mathcal{C}^k(\mathcal{X}) \iff \exists f', f'', \ldots, f^{(k)} \text{ continuous} \tag{14.14}$$

Note

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

Corollary 14.3 Smooth Function $\mathcal{C}^\infty$**:** Is a function $f : \mathcal{X} \to \mathcal{Y}$ that has derivatives infinitely many times differentiable.

$$f \in \mathcal{C}^\infty(\mathcal{X}) \iff f', f'', \ldots, f^{(\infty)} \tag{14.15}$$

Corollary 14.4 Continuously Differentiable Function $\mathcal{C}^1$**:** Is the class of functions that consists of all differentiable functions whose derivative is continuous.
Hence a function $f : \mathcal{X} \to \mathcal{Y}$ of the class must satisfy:

$$f \in \mathcal{C}^1(\mathcal{X}) \iff f' \text{ continuous} \tag{14.16}$$

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

Definition 14.17 Lipschitz Continuity: A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz Contant L:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y}, \quad L > 0 \tag{14.17}$$

Note

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

Definition 14.18 Lipschitz Continuous Gradient:
A *continuously differentiable* function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has L-*Lipschitz continuous gradient* if it satisfies:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \texttt{dom}(f), \quad L > 0 \tag{14.18}$$

if $f \in \mathcal{C}^2$, this is equivalent to:

$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad \forall \boldsymbol{x} \in \texttt{dom}(f), \quad L > 0 \tag{14.19}$$

Lemma 14.1 Descent Lemma: If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ has *Lipschitz continuous gradient* eq. (14.18) over its domain, then it holds that:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})| \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{14.20}$$

Note

If f is twice differentiable then the largest eigenvalue of the Hessian ([def. 15.5]) of f is uniformly upper bounded by L

Proof. lemma 14.1 for $\mathcal{C}^1$ functions:
Let $g(t) \equiv f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))$ from the FToC (theorem 14.2) we know that:

$$\int_0^1 g'(t)\,\mathrm{d}t = g(1) - g(0) = f(\boldsymbol{x}) - f(\boldsymbol{y})$$

It then follows from the reverse:

$$\begin{aligned}
&|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})| \\
&\overset{\text{Chain. R}}{\underset{\text{FToC}}{=}} \left|\int_0^1 \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t - \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y})\right| \\
&= \left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
&= \left|\int_0^1 \left(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\right)^\intercal(\boldsymbol{x} - \boldsymbol{y})\,\mathrm{d}t\right| \\
&\overset{\text{C.S.}}{\leqslant} \left|\int_0^1 \|\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y})\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
&\overset{\text{eq. (14.18)}}{=} \left|\int_0^1 L\|\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}) - \boldsymbol{y}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|\,\mathrm{d}t\right| \\
&= \left|L\|\boldsymbol{x} - \boldsymbol{y}\|^2 \int_0^1 t\,\mathrm{d}t\right| = \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2
\end{aligned}$$

□

Proof. lemma 14.1 for $\mathcal{C}^2$ functions:

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

Now we plug in $\nabla^2 f(\boldsymbol{x})$ and recover eq. (14.21):

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\intercal L(\boldsymbol{y} - \boldsymbol{x})$$

□

Definition 14.19 L-Smoothness: A L-smooth function is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies:

$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$

with $\forall \boldsymbol{x}, \boldsymbol{y} \in \texttt{dom}(f), \quad L > 0$ (14.21)

If f is a twice differentiable this is equivalent to:

$$\nabla^2 f(\boldsymbol{x}) \preccurlyeq L\boldsymbol{I} \qquad L > 0 \tag{14.22}$$

Theorem 14.3
L-Smoothness of convex functions:
A *convex* and L-Smooth function ([def. 14.19]) has a Lipschitz continuous gradient (eq. (14.18)) thus it holds that:

$$f(\boldsymbol{x}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{x} - \boldsymbol{y}) \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{14.23}$$

Proof. theorem 14.3:
With the definition of convexity for a differentiable function (eq. (14.26)) it follows

$$\begin{aligned}
&f(x) - f(y) + \nabla f(y)^\intercal(x - y) \geqslant 0 \\
\Rightarrow &|f(x) - f(y) + \nabla f(y)^\intercal(x - y)| \\
&\overset{\text{if eq. (14.26)}}{=} f(x) - f(y) + \nabla f(y)^\intercal(x - y)
\end{aligned}$$

with lemma 14.1 and [def. 14.19] it follows theorem 14.3 □

Corollary 14.5 : L-smoothnes is a weaker condition than L-Lipschitz continuous gradients

3. Convexity

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

Definition 14.20 Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if it satisfies:

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{matrix}\forall x, y \in \texttt{dom}(f) \\ \forall \lambda \in [0, 1]\end{matrix} \tag{14.24}$$

include figure from tikz/convexity

Definition 14.21 Concave Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if it satisfies:

$$f(\lambda x + (1 - \lambda)y) \geqslant \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{matrix}\forall x, y \in \texttt{dom}(f) \\ \forall \lambda \in [0, 1]\end{matrix} \tag{14.25}$$

Corollary 14.6 Convexity $\to$ global minimima: Convexity implies that all local minima (if they exist) are global minima.

Definition 14.22 Stricly Convex Functions:
A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if it satisfies:

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \qquad \begin{matrix}\forall x, y \in \texttt{dom}(f) \\ \forall \lambda \in [0, 1]\end{matrix}$$

add plot

If f is a differentiable function this is equivalent to:

$$f(x) \geqslant f(y) + \nabla f(y)^\intercal(x - y) \qquad \forall x, y \in \texttt{dom}(f) \tag{14.26}$$

If f is a twice differentiable function this is equivalent to:

$$\nabla^2 f(x) \geqslant 0 \qquad \forall x, y \in \texttt{dom}(f) \tag{14.27}$$

Intuition

- Convexity implies that a function f is bound by/below a linear interpolation from x to y and strong convexity that f is strictly bound/below.
- eq. (14.26) implies that $f(x)$ is above the tangent $f(x) + \nabla f(x)^\intercal(y - x)$ for all $x, y \in \texttt{dom}(f)$
- **??** implies that $f(x)$ is flat or curved upwards

Corollary 14.7 Strict Convexity → Uniqueness:
Strict convexity implies a unique minimizer $\iff$ at most one global minimum.

Corollary 14.8 : A twice differentiable function of one variable $f : \mathbb{R} \to \mathbb{R}$ is convex on an interval $\mathcal{X} = [a, b]$ if and only if its second derivative is non-negative on that interval $\mathcal{X}$:

$$f''(x) \geqslant 0 \qquad \forall x \in \mathcal{X} \tag{14.28}$$

Definition 14.23 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:

$$f(y) \geqslant f(x) + \nabla f(x)^\intercal (y-x) + \frac{\mu}{2}\|y-x\|_2^2 \tag{14.29}$$

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

$$\nabla^2 f(x) \geqslant \mu \boldsymbol{I} \qquad \forall x, y \in \mathcal{X} \quad \mu > 0 \tag{14.30}$$

Corollary 14.9 Strong Convexity implies Strict Convexity:
https://math.stackexchange.com/questions/2090991/proof-for-strongly-convex-function-is-strictly-convex

Property 14.1:

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\intercal (\boldsymbol{x}-\boldsymbol{y}) + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \tag{14.31}$$

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. eq. (14.30) analogously to **Proof** eq. (14.22) □

Note

If f is twice differentiable then the smallest eigenvalue of the Hessian ([def. 15.5]) of f is uniformly lower bounded by μ
Hence strong convexity can be considered as the analogous to smoothness

Example 14.2 Quadratic Function: A quadratic function eq. (14.13) is convex if:

$$\nabla^2_{\boldsymbol{x}} \text{eq. (14.13)} = \boldsymbol{A} \succcurlyeq 0 \tag{14.32}$$

Corollary 14.10 :
Strong convexity $\Rightarrow$ Strict convexity $\Rightarrow$ Convexity

3.1. Properties that preserve convexity

Property 14.2 Non-negative weighted Sums: Let f be a convex function then $g(x)$ is convex as well:

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

Property 14.3 Composition of Affine Mappings: Let f be a convex function then $g(x)$ is convex as well:

$$g(x) = f(\boldsymbol{A}\boldsymbol{x} + b)$$

Property 14.4 Pointwise Maxima: Let f be a convex function then $g(x)$ is convex as well:

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

Functions

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

$$f(-x) = f(x) \tag{14.33}$$

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

$$f(-x) = -f(x) \tag{14.34}$$

Theorem 14.4 Rules:
Let f be even and f odd respectively.

$g =: f \cdot f$ is even	$g =: f \cdot f$ is even
$g =: f \cdot f$ is odd	the same holds for division

Examples

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

x-Shift:	$f(x-c) \Rightarrow$ shift to the right	
	$f(x+c) \Rightarrow$ shift to the left	(14.35)
y-Shift:	$f(x) \pm c \Rightarrow$ shift up/down	(14.36)

Proof. **eq. (14.35)** $f(x_n - c)$ 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

$$e^{\pm ix} = \cos x \pm i \sin x \tag{14.37}$$

Euler's Identity

$$e^{\pm i} = -1 \tag{14.38}$$

Note

$$e^n = 1 \Leftrightarrow n = \mathrm{i}\, 2\pi k, \qquad k \in \mathbb{N} \tag{14.39}$$

Corollary 14.11 Every norm is a convex function: By using definition [def. 14.20] and the triangular inequality it follows (with the exception of the L0-norm):

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

3.2. Taylor Expansion

Definition 14.24 Taylor Expansion:

$$T_n(x) = \sum_{i=0}^{n} \frac{1}{n!} f^{(i)}(x_0) \cdot (x-x_0)^{(i)} \tag{14.40}$$
$$= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \mathcal{O}(x^3) \tag{14.41}$$

Definition 14.25 Incremental Taylor:
Goal: evaluate $T_n(x)$ (eq. (14.41)) at the point $x_0 + \Delta x$ in order to propagate the function $f(x)$ by $h = \Delta x$:

$$T_n(x_0 \pm h) = \sum_{i=0}^{n} \frac{h^i}{n!} f^{(i)}(x_0) i^{-1} \tag{14.42}$$
$$= f(x_0) \pm hf'(x_0) + \frac{h^2}{2}f''(x_0) \pm f'''(x_0)(h)^3 + \mathcal{O}(h^4)$$

Note

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

Definition 14.26 Multidimensional Taylor: Suppose $X \in \mathbb{R}^n$ is open, $\boldsymbol{x} \in X$, $f : X \mapsto \mathbb{R}$ and $f \in \mathcal{C}^2$ then it holds that

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\intercal H(\boldsymbol{x}-\boldsymbol{x}_0) \tag{14.43}$$

Definition 14.27 Argmax: The argmax of a function defined on a set D is given by:

$$\arg\max_{x \in D} f(x) = \{x | f(x) \geqslant f(y), \forall y \in D\} \tag{14.44}$$

Definition 14.28 Argmin: The argmin of a function defined on a set D is given by:

$$\arg\min_{x \in D} f(x) = \{x | f(x) \leqslant f(y), \forall y \in D\} \tag{14.45}$$

Corollary 14.12 Relationship arg min $\leftrightarrow$ arg max:

$$\arg\min_{x \in D} f(x) = \arg\max_{x \in D} -f(x) \tag{14.46}$$

Property 14.5 Argmax Identities:

1. **Shifting:**
$\forall \lambda$ const $\quad \arg\max f(x) = \arg\max f(x) + \lambda$ (14.47)
2. **Positive Scaling:**
$\forall \lambda > 0$ const $\quad \arg\max f(x) = \arg\max \lambda f(x)$ (14.48)
3. **Negative Scaling:**
$\forall \lambda < 0$ const $\quad \arg\max f(x) = \arg\min \lambda f(x)$ (14.49)
4. **Positive Functions:**
$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f)$
$$\arg\max f(x) = \arg\min \frac{1}{f(x)} \tag{14.50}$$
5. **Stricly Monotonic Functions:** for all strictly monotonic increasing functions[def. 14.10] g it holds that:
$$\arg\max g(f(x)) = \arg\max f(x) \tag{14.51}$$

Definition 14.29 Max: The maximum of a function f defined on the set D is given by:

$$\max_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\max_{x \in D} f(x) \tag{14.52}$$

Definition 14.30 Min: The minimum of a function f defined on the set D is given by:

$$\min_{x \in D} f(x) = f(x^*) \quad \text{with} \quad \forall x^* \in \arg\min_{x \in D} f(x) \tag{14.53}$$

Corollary 14.13 Relationship min $\leftrightarrow$ max:

$$\min_{x \in D} f(x) = -\max_{x \in D} -f(x) \tag{14.54}$$

Property 14.6 Max Identities:

1. **Shifting:**
$\forall \lambda$ const $\quad \max\{f(x) + \lambda\} = \lambda + \max f(x)$ (14.55)
2. **Positive Scaling:**
$\forall \lambda > 0$ const $\quad \max \lambda f(x) = \lambda \max f(x)$ (14.56)
3. **Negative Scaling:**
$\forall \lambda < 0$ const $\quad \max \lambda f(x) = \lambda \min f(x)$ (14.57)
4. **Positive Functions:**
$\forall \arg\max f(x) > 0, \forall x \in \mathtt{dom}(f) \quad \max \frac{1}{f(x)} = \frac{1}{\min f(x)}$ (14.58)
5. **Stricly Monotonic Functions:** for all strictly monotonic increasing functions[def. 14.10] g it holds that:
$$\max g(f(x)) = g(\max f(x)) \tag{14.59}$$

Definition 14.31 Supremum: The supremum of a function defined on a set D is given by:

$$\sup_{x \in D} f(x) = \{y | y \geqslant f(x), \forall x \in D\} = \min_{y | y \geqslant f(x), \forall x \in D} y \tag{14.60}$$

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

Definition 14.32 Infimum: The infimum of a function defined on a set D is given by:

$$\inf_{x \in D} f(x) = \{y | y \leqslant f(x), \forall x \in D\} = \max_{y | y \leqslant f(x), \forall x \in D} y \tag{14.61}$$

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

Corollary 14.14 Relationship sup $\leftrightarrow$ inf:

$$\in_{x \in D} f(x) = -\sup_{x \in D} -f(x) \tag{14.62}$$

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

$$y(t) = f(x(t), t) \xrightarrow[\forall \tau]{\text{time-invariance}} y(t-\tau) = f(x(t-\tau), t) \tag{14.63}$$

Definition 14.34 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

$$f(g(x)) = x \qquad \forall x \in \mathtt{dom}(g) \tag{14.64}$$

and

$$g(f(u)) = u \qquad \forall u \in \mathtt{dom}(f) \tag{14.65}$$

Property 14.7
Reflective Property of Inverse Functions: f contains (a, b) if and only if f^{-1} contains (b, a).
The line $y = x$ is a symmetry line for f and f^{-1}.

Theorem 14.5 The Existence of an Inverse Function:
A function has an inverse function if and only if it is one-to-one.

Corollary 14.15 Inverse functions and strict monotonicity: If a function f is strictly monotonic [def. 14.12] on its entire domain, then it is one-to-one and therefore has an inverse function.

4. Special Functions

4.1. The Gamma Function

Definition 14.35 The gamma function $\Gamma(\alpha)$: Is extension of the factorial function (??) to the real and complex numbers (with a positive real part):

$$\Gamma(z) = \int_0^\infty x^{z-1} \mathrm{e}^{-x}\, \mathrm{d}x \qquad \Re(z) > 0 \tag{14.66}$$

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

Differential Calculus

Definition 15.1 Critical/Stationary Point: Given a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that is differentiable at a point $\boldsymbol{x}_0$ then it is called a critical point if the functions derivative vanishes at that point:

$$f'(\boldsymbol{x}_0) = 0 \qquad \iff \qquad \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = 0$$

Definition 15.2 Second Derivative $\frac{\partial^2}{\partial x_i \partial x_j}$**:**

Corollary 15.1 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. 15.1] then it follows that:

- $f''(x) > 0 \iff \begin{matrix} f'(x+\epsilon) > 0 & \text{slope points uphill} \\ f'(x-\epsilon) < 0 & \text{slope points downhill} \\ f(x) \text{ is a local minimum} \end{matrix}$
- $f''(x) < 0 \iff \begin{matrix} f'(x+\epsilon) > 0 & \text{slope points downhill} \\ f'(x-\epsilon) < 0 & \text{slope points uphill} \\ f(x) \text{ is a local maximum} \end{matrix}$

$\epsilon > 0$ sufficiently small enough

Definition 15.3 Gradient: Given $f : n \mapsto \mathbb{R}$ its gradient is defined as:

$$\text{grad}_{\boldsymbol{x}}(f) = \nabla_{\boldsymbol{x}} f := \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{15.1}$$

Definition 15.4 Jacobi Matrix: Given a vector valued function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ its derivative/Jacobian is defined as:

$$\boldsymbol{J}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{J}_f(\boldsymbol{x}) = \boldsymbol{D}\boldsymbol{f} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)}(\boldsymbol{x}) = \tag{15.2}$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_1}{\partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}) & \ddots & & \frac{\partial f_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_m}{\partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

Theorem 15.1
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:

$$\frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}$$

Definition 15.5 Hessian Matrix:
Given a function $f : \mathbb{R} \mapsto \mathbb{R}^n$ its Hessian$\in \mathbb{R}^{n \times n}$ is defined as:

$$\boldsymbol{H}(\boldsymbol{f})(\boldsymbol{x}) = \boldsymbol{H}_f(\boldsymbol{x}) = \boldsymbol{J}(\nabla \boldsymbol{f}(\boldsymbol{x}))^T \tag{15.3}$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{x}) \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix}$$

and it corresponds to the Jacobian of the Gradient.
Due to the differentiability and theorem 15.1 it follows that the Hessian is (if it exists):

- Symmetric
- Real

Corollary 15.2 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 $\{(\lambda_1, \boldsymbol{v}_1), \dots, \lambda_n, \boldsymbol{v}_n\}$.
Not let $\boldsymbol{d}$ be a directional unit vector then the second derivative in that direction is given by:

$$\boldsymbol{d}^\intercal \boldsymbol{H}\boldsymbol{d} \qquad \iff \qquad \boldsymbol{d}^\intercal \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \qquad \overset{\text{if } \boldsymbol{d}=\boldsymbol{v}_j}{\iff} \qquad \boldsymbol{d}^\intercal \lambda_j \boldsymbol{v}_j$$

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

Corollary 15.3 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 $\boldsymbol{x}$ [def. 15.1] then it follows that:

- If $\boldsymbol{H}$ is p.d $\iff \forall \lambda_i > 0 \in \boldsymbol{H} \quad \rightarrow \quad f(\boldsymbol{x})$ is a local min.
- If $\boldsymbol{H}$ is n.d $\iff \forall \lambda_i < 0 \in \boldsymbol{H} \quad \rightarrow \quad f(\boldsymbol{x})$ is a local max.
- If $\exists \lambda_i > 0 \in \boldsymbol{H}$ and $\exists \lambda_i < 0 \in \boldsymbol{H}$ then $\boldsymbol{x}$ is a local maximum in one cross section of f but a local minimum in another
- If $\exists \lambda_i = 0 \in \boldsymbol{H}$ 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 $\boldsymbol{H}$ is positive definite for a minima $\boldsymbol{x}^*$ of a *quadratic* function f then this point must be a global minimum of that function.

Integral Calculus

Theorem 16.1 Important Integral Properties:

Addition
$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x \tag{16.1}$$

Reflection
$$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x \tag{16.2}$$

Translation
$$\int_a^b f(x)\,\mathrm{d}x \overset{u:=x\pm c}{=} \int_{a\pm c}^{b\pm c} f(x\mp c)\,\mathrm{d}x \tag{16.3}$$

f Odd
$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 0 \tag{16.4}$$

f Even
$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 2\int_0^a f(x)\,\mathrm{d}x \tag{16.5}$$

Proof. **eqs. (16.4) and (16.5)**

$$\begin{aligned} I := \int_{-a}^{a} f(x)\,\mathrm{d}x &= \int_{-a}^{0} f(x)\,\mathrm{d}x + \int_0^a f(x)\,\mathrm{d}x \\ &\overset{\substack{t=-x\\ \mathrm{d}t=-\mathrm{d}x}}{=} -\int_a^0 f(-x)\,\mathrm{d}x + \int_0^a f(x)\,\mathrm{d}x \\ &= \int_0^a f(-x)+f(x)\,\mathrm{d}x = \begin{cases} 0 & \text{if } f \text{ odd} \\ 2I & \text{if } f \text{ even} \end{cases} \end{aligned}$$

□

Linear Algebra

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

Rank: $\operatorname{rank}(\boldsymbol{A}) = \dim(\mathfrak{R}(\boldsymbol{A})$
of a matrix is the dimension of the vector space generated (or spanned) by its columns/rows.
Span/Linear Hull: $\operatorname{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n) =$
$\{\lambda_1 \boldsymbol{v}_1, \lambda_2 \boldsymbol{v}_2, \ldots, \lambda_n \boldsymbol{v}_n)\} = \{\boldsymbol{v} \mid \boldsymbol{v} = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i), \lambda_i \in \mathbb{R}\}$
Is the set of vectors tha can be expressed as a linear combination of the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.
Note these vectors may be linearly independent.
Generatring Set: Is the set of vectors which span the $\mathbb{R}^n$ that is: $\operatorname{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m) = \mathbb{R}^n$.
e.g. $(4,0)^\top, (0,5)^\top$ span the $\mathbb{R}^n$.
Basis $\mathfrak{B}$: A lin. indep. generating set of the $\mathbb{R}^n$ is called basis of the $\mathbb{R}^n$.
The unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ build a standard basis of the $\mathbb{R}^n$
Vector Space
Image/Range: $\mathfrak{R}(\boldsymbol{A}) := \{\boldsymbol{A}\boldsymbol{x} \mid x \in \mathbb{K}^n\} \subset \mathbb{K}^n$
Null-Space/Kernel: $\mathbb{N} := \{z \in \mathbb{K}^n \mid \boldsymbol{A}\boldsymbol{z} = 0\}$
Dimension theorem:

Theorem 17.1 Rank-Nullity theorem: For any $\boldsymbol{A} \in \mathbb{Q}^{m \times n}$
$n = \dim(\mathbb{N}[\boldsymbol{A}]) + \dim(\mathfrak{R}[\boldsymbol{A}])$

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

1. Transformations

1.1. Affine Transformations

Definition 17.1 Affine Transfromation/Map:
Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ then:
$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{x} + b \tag{17.1}$$
is called an affine transformation of $\boldsymbol{x}$.

2. Determinants

Property 17.1 Determinant times Scalar $\det(\alpha \boldsymbol{A})$:
Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ it holds:
$$\det(\alpha \cdot \boldsymbol{A}) = \alpha^n \boldsymbol{A} \tag{17.2}$$

3. Eigenvalues and Vectors

Formula 17.1 Eigenvalues of a 2x2 matrix: Given a 2x2-matrix $\boldsymbol{A}$ its eigenvalues can be calculated by:
$$\{\lambda_1, \lambda_2\} \in \frac{\operatorname{tr}(\boldsymbol{A}) \pm \sqrt{\operatorname{tr}(\boldsymbol{A})^2 - 4\det(\boldsymbol{A})}}{2} \tag{17.3}$$
with $\operatorname{tr}(\boldsymbol{A}) = a + d$ $\det(\boldsymbol{A}) = ad - bc$

4. Special Kind of Vectors

Definition 17.2 Orthogonal Vectors: Let $\mathcal{Y}$ be an inner-product space[def. 17.19]. A set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots\} \in \mathcal{Y}$ is called *orthogonal* iff:
$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0 \qquad \forall i \neq \tag{17.4}$$

Definition 17.3 Orthonormal Vectors: Let $\mathcal{Y}$ be an inner-product space[def. 17.19]. A set of vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots\} \in \mathcal{Y}$ is called *orthonormal* iff:
$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad \forall i, j \tag{17.5}$$

5. Special Kind of Matrices

Definition 17.4 Orthogonal Matrix: A real valued square matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is said to be orthogonal if its row vectors (and respectively its column vectors) build an orthonormal basis:
$$\langle \boldsymbol{q}_{:i}, \boldsymbol{q}_{:j} \rangle = \delta_{ij} \qquad \text{and} \qquad \langle \boldsymbol{q}_{i:}, \boldsymbol{q}_{j:} \rangle = \delta_{ij} \tag{17.6}$$
This is exactly true if the inverse of $\boldsymbol{Q}$ equals its transpose:
$$\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\intercal \iff \boldsymbol{Q}\boldsymbol{Q}^\intercal = \boldsymbol{Q}^\intercal\boldsymbol{Q} = \boldsymbol{I} \tag{17.7}$$

Definition 17.5 Unitary/Hermitian Matrices:
$$\boldsymbol{A} = \boldsymbol{A}^{\mathsf{H}} \tag{17.8}$$

6. Block Partitioned Matrices

Definition 17.6 Block Partitioned Matrix:
A matrix $\boldsymbol{M} \in \mathbb{R}^{k+l,k+l}$ can be partitioned into a *block partitioned matrix*:
$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \qquad \boldsymbol{A} \in \mathbb{R}^{k,k}, \boldsymbol{B} \in \mathbb{R}^{k,l}, \boldsymbol{C} \in \mathbb{R}^{l,k}, \boldsymbol{D} \in \mathbb{R}^{l,l} \tag{17.9}$$

Definition 17.7 Block Partitioned Linear System:
A linear system $\boldsymbol{M}\boldsymbol{x} = b$ with $\boldsymbol{M} \in \mathbb{R}^{k+l,k+l}$ and $\boldsymbol{x}, b \in \mathbb{R}^{k+l}$ can be partitioned into a *block partitioned system*:
$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \qquad \boldsymbol{A} \in \mathbb{R}^{k,k}, \boldsymbol{B} \in \mathbb{R}^{k,l}, \boldsymbol{C} \in \mathbb{R}^{l,k}, \boldsymbol{D} \in \mathbb{R}^{l,l}$$
$$\boldsymbol{x}_1, b_1 \in \mathbb{R}^k, \boldsymbol{x}_2, b_2 \in \mathbb{R}^l \tag{17.10}$$

6.1. Schur Complement

Definition 17.8 Schur Complement: Given a block partitioned matrix[def. 17.6] $\boldsymbol{M} \in \mathbb{R}^{k+l,k+l}$ its Schur complements are given by:
$$\boldsymbol{S}_A = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} \qquad \boldsymbol{S}_D = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} \tag{17.11}$$

6.2. Inverse of Block Partitioned Matrix

Definition 17.9 proof 13
Inverse of a Block Partitioned Matrix:
Given a block partitioned matrix[def. 17.6] $\boldsymbol{M} \in \mathbb{R}^{k+l,k+l}$ its inverse $\boldsymbol{M}^{-1}$ can be partitioned as well:
$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \qquad \boldsymbol{M}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{B}} \\ \tilde{\boldsymbol{C}} & \tilde{\boldsymbol{D}} \end{bmatrix} \tag{17.12}$$
$$\tilde{\boldsymbol{A}} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{A}}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \qquad \tilde{\boldsymbol{C}} = -\boldsymbol{S}_{\boldsymbol{A}}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}$$
$$\tilde{\boldsymbol{B}} = -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{S}_{\boldsymbol{A}}^{-1} \qquad \tilde{\boldsymbol{D}} = \boldsymbol{S}_{\boldsymbol{A}}^{-1}$$
where $\boldsymbol{S}_{\boldsymbol{A}} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$ is the Schur complement of $\boldsymbol{A}$.

6.3. Properties of Matrices

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

Definition 17.10 Square Root:

7. Matrix Operations

7.1. Trace

Definition 17.11 Trace: The trace of an $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ matrix is defined as:
$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn} \tag{17.13}$$

Property 17.2:
$$\operatorname{tr}(\mathbb{R}) = \mathbb{R} \tag{17.14}$$

Property 17.3:
$$\operatorname{tr}(\boldsymbol{A}^\intercal) = \operatorname{tr}(\boldsymbol{A}) \tag{17.15}$$

Property 17.4:
$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{C}\boldsymbol{B}\boldsymbol{A}) \tag{17.16}$$

8. Decompositions

8.0.1. Eigendecomposition

Definition 17.12 Eigendecomposition $\boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^{-1}$:

8.0.2. Cholesky Decomposition

9. Spaces and Measures

Definition 17.13 Bilinear Form/Functional:
Is a mapping $a : \mathcal{Y} \times \mathcal{Y} \mapsto F$ on a field of scalars $F \subseteq \mathbb{K}$, $K = \mathbb{R}$ or $\mathbb{C}$ that satisfies:
$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$$
$$\forall u, v, w \in \mathcal{Y}, \quad \forall \alpha, \beta \in \mathbb{K}$$
Thus: a is linear w.r.t. each argument.

Definition 17.14 Symmetric bilinear form: A bilinear form a on $\mathcal{Y}$ is symmetric if and only if:
$$a(u, v) = a(v, u) \qquad \forall u, v \in \mathcal{Y}$$

Definition 17.15 Positive (semi) definite bilinear form:
A symmetric bilinear form a on a vector space $\mathcal{Y}$ over a field F is positive defintie if and only if:
$$a(u, u) > 0 \qquad \forall u \in \mathcal{Y} \backslash \{0\} \tag{17.17}$$
And positive semidefinte $\iff \geqslant$ (17.18)

Corollary 17.1 Matrix induced Bilinear Form:
For finite dimensional inner product spaces $\mathcal{X} \in \mathbb{K}^n$ any *symmetric* matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ induces a bilinear form:
$$a(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x}' = (\boldsymbol{A}\boldsymbol{x}')\boldsymbol{x},$$

Definition 17.16 Positive (semi) definite Matrix $\succ$:
A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive defintie if and only if:
$$\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} > 0 \iff \boldsymbol{A} \succ \quad \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\} \tag{17.19}$$
And positive semidefinte $\iff \succcurlyeq$ (17.20)

Corollary 17.2
Eigenvalues of positive (semi) definite matrix:
A positive definite matrix is a *symmetric matrix* where every eigenvalue is *strictly* positive and positive semi definite if every eigenvalue is *positive*.
$$\forall \lambda_i \in \operatorname{eigenv}(\boldsymbol{A}) > 0 \tag{17.21}$$
And positive semidefinte $\iff \succcurlyeq$ (17.22)

Proof. corollary 17.2 (for real matrices):
Let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{A}$ then it follows:
$$0 \overset{\text{corollary } 17.2}{<} \boldsymbol{v}^\intercal \boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^\intercal \lambda \boldsymbol{v} = \|\boldsymbol{v}\|\lambda$$
□

Corollary 17.3 Positive Definiteness and Determinant:
The determinant of a positive definite matrix is always positive. **Thus** a positive definite matrix is always *nonsingular*

Definition 17.17 Negative (semi) definite Matrix $\prec$:
A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is negative defintie if and only if:
$$\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x} < 0 \iff \boldsymbol{A} \prec 0 \quad \forall \boldsymbol{x} \in \mathbb{R}^n \backslash \{0\} \tag{17.23}$$
And negative semidefinte $\iff \preccurlyeq$ (17.24)

Theorem 17.2 Sylvester's criterion: Let $\boldsymbol{A}$ be *symmetric/Hermitian* matrix and denote by $\boldsymbol{A}^{(k)}$ the $k \times k$ upper left sub-matrix of $\boldsymbol{A}$.
Then it holds that:
- $\boldsymbol{A} \succ 0 \iff \det\left(\boldsymbol{A}^k\right) > 0 \qquad k = 1, \ldots, n$ (17.25)
- $\boldsymbol{A} \prec 0 \iff (-1)^k \det\left(\boldsymbol{A}^k\right) > 0 \qquad k = 1, \ldots, n$ (17.26)
- $\boldsymbol{A}$ is indefinite if the first $\det\left(\boldsymbol{A}^k\right)$ that breaks both of the previous patterns is on the wrong side.
- Sylvester's criterion is inconclusive ($\boldsymbol{A}$ can be anything of the previous three) if the first $\det\left(\boldsymbol{A}^k\right)$ that breaks both patterns is 0.

10. Inner Products

Definition 17.18 Inner Product: Let $\mathcal{Y}$ be a vector space over a field $F \in \mathbb{K}$ of scalars. An inner product on $\mathcal{Y}$ is a map:
$$\langle \cdot, \cdot \rangle : \mathcal{Y} \times \mathcal{Y} \mapsto F \subseteq \mathbb{K} \qquad K = \mathbb{R} \text{ or } \mathbb{C} \tag{17.27}$$
that satisfies: $\forall x, y, z \in \mathcal{Y}, \quad \alpha, \beta \in F$
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 17.19 Inner Product Space $(\mathcal{Y}, \langle \cdot, \cdot \rangle_{\mathcal{Y}})$: Let $F \in \mathbb{K}$ be a field of scalars.
An inner product space $\mathcal{Y}$ is a vetor space over a field F together with an an inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$).

Corollary 17.4 Inner product↦S.p.d. Bilinear Form:
Let $\mathcal{Y}$ be a vector space over a field $F \in \mathbb{K}$ of scalar.
An **inner product** on $\mathcal{Y}$ is a positive definite symmetric bilinear form on $\mathcal{Y}$.

Example: scalar prodct

Let $a(u, v) = u^\intercal \boldsymbol{I} v$ 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 $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geqslant 0$, whereas bilinear forms must not.

Definition 17.20 Norm $\|\cdot\|_{\mathcal{Y}}$:
A norm measures the **size** of its argument.
Formally let $\mathcal{Y}$ be a vector space over a field F, a norm on $\mathcal{Y}$ is a map:
$$\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \mapsto \mathbb{R}_+ \tag{17.28}$$
that satisfies: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Y}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R}$ or $\mathbb{C}$
1. Definitness: $\|\boldsymbol{x}\|_{\mathcal{Y}} = 0 \iff \boldsymbol{x} = 0.$
2. Homogenity: $\|\alpha \boldsymbol{x}\|_{\mathcal{Y}} = |\alpha| \|\boldsymbol{x}\|_{\mathcal{Y}}$
3. Triangular Inequality: $\|\boldsymbol{x} + \boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x}\|_{\mathcal{Y}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$

Meaning: Triangular Inequality

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

Corollary 17.5 Reverse Triangular Inequality:
$$-\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}} \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$$
resp. $\left| \|\boldsymbol{x}\|_{\mathcal{Y}} - \|\boldsymbol{y}\|_{\mathcal{Y}} \right| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{Y}}$

Semi-norm

add

Corollary 17.6 Normed vector space: Is a vector space $\mathcal{Y}$ over a field F, on which a norm $\|\cdot\|_{\mathcal{Y}}$ can be defined.

Corollary 17.7 Inner product induced norm $\langle \cdot, \cdot \rangle_{\mathcal{Y}} \rightarrow \|\cdot\|_{\mathcal{Y}}$: Every inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ induces a norm of the form:
$$\|\boldsymbol{x}\|_{\mathcal{Y}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \qquad \boldsymbol{x} \in \mathcal{Y}$$
Thus We can define function spaces by their associated norm $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ and inner product spaces lead to normed vector spaces and vice versa.

Corollary 17.8 Energy Norm: A *s.p.d.* bilinear form $a : \mathcal{Y} \times \mathcal{Y} \mapsto F$ induces an energy norm:
$$\|\boldsymbol{x}\|_a := (a(\boldsymbol{x}, \boldsymbol{x}))^{\frac{1}{2}} = \sqrt{a(\boldsymbol{x}, \boldsymbol{x})} \qquad \boldsymbol{x} \in \mathcal{Y}$$

Definition 17.21 Distance Function/Measure: Is measuring the **distance** between two things.
Formally: on a set S is a mapping:
$$d(\cdot, \cdot) : S \times S \mapsto \mathbb{R}_+$$
that satisfies: $\forall x, y, z \in S$
1. ?: $d(x, x) = 0$
2. Symmetry: $d(x, y) = d(y, x)$
3. Triangular Identiy: $d(x, z) \leqslant d(x, y) + d(y, z)$

Definition 17.22 Metric: Is a distance measure that additionally satisfies: $\forall x, y \in S$
identity of indiscernibles : $d(x, y) = 0 \iff x = y$

Corollary 17.9 Metric→Norm: Every norm $\|\cdot\|_{\mathcal{Y}}$ on a vector space $\mathcal{Y}$ over a field F induces a metric by:
$$d(x, y) = \|x - y\|_{\mathcal{Y}} \qquad \forall x, y \in \mathcal{Y}$$
metric induced by norms additionally satisfy: $\forall x, y \in \mathcal{Y}, \quad \alpha \in F \subseteq \mathbb{K} \quad K = \mathbb{R}$ or $\mathbb{C}$
1. Homogenity/Scaling: $d(\alpha x, \alpha y)_{\mathcal{Y}} = |\alpha| d(x, y)_{\mathcal{Y}}$
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 $\mathcal{Y}$ satisfies the properties then it induces a norm of the form:
$$\|\boldsymbol{x}\|_{\mathcal{Y}} := d(\boldsymbol{x}, 0)_{\mathcal{Y}}$$

Note

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

Note

(bilinear form $\xrightarrow{\text{induces}}$)
inner product $\xrightarrow{\text{induces}}$ norm $\xrightarrow{\text{induces}}$ metric.

11. Vector Algebra

11.1. Planes

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

12. Derivatives

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

13. Proofs

Proof. [def. 17.9]

$$\boldsymbol{M}\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{I}_{k,k} & \mathbf{0}_{k,l} \\ \mathbf{0}_{l,k} & \boldsymbol{I}_{l,l} \end{bmatrix} \tag{17.29}$$

□

Geometry

Corollary 18.1 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:

$$y = \phi(x) = \frac{d-c}{b-a}(x-a) + c \tag{18.1}$$

Proof. **corollary 18.1** By [def. 17.1] we want a function $f : [a, b] \to [c, d]$ that satisfies:

$$f(a) = c \quad \textbf{and} \quad f(b) = d$$

additionally $f(x)$ has to be a linear function ([def. 14.13]), that is the output scales the same way as the input scales.
Thus it follows:

$$\frac{d-c}{b-a} = \frac{f(x)-f(a)}{x-a} \iff f(x) = \frac{d-c}{b-a}(x-a) + c$$

□

Trigonometry

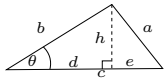
Law 18.1 Law of Cosine: relates the side of a triangle to the cosine of its angles.

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{b,c} \tag{18.2}$$

More general for vectors it holds:

$$\|\boldsymbol{x}-\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta_{\boldsymbol{x},\boldsymbol{y}} \tag{18.3}$$

Proof. eq. (18.2):
We know: $\sin\theta = \frac{h}{b} \Rightarrow \underline{h}$ and $\cos\theta = \frac{d}{b} \Rightarrow d$
Thus $\underline{e} = c - d = c - b\cos\theta \Rightarrow a^2 = \underline{e}^2 + \underline{h}^2 \Rightarrow a$ □



Proof. eq. (18.3):

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{y}\|^2 &= (\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y}) \\ &= \boldsymbol{x}\cdot\boldsymbol{x} - 2\boldsymbol{x}\cdot\boldsymbol{y} + \boldsymbol{y}\cdot\boldsymbol{y} \\ &= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\left(\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta\right)\end{aligned}$$

□

Law 18.2 Pythagorean theorem: special case of **??** for right triangle:

$$a^2 = b^2 + c^2 \tag{18.4}$$

Formula 18.1 Euler's Formula:

$$e^{\pm ix} = \cos x \pm i\sin x \tag{18.5}$$

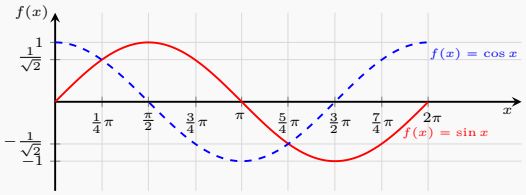
Formula 18.2 Euler's Identity:

$$e^{\pm i} = -1 \tag{18.6}$$

Note

$$e^n = 1 \Leftrightarrow n = \mathrm{i}\,2\pi k, \qquad k \in \mathbb{N} \tag{18.7}$$

Sine and Cosine



$$\cos x \overset{(14.37)}{=} \frac{1}{2}\left[e^{ix} + e^{-ix}\right] \tag{18.8}$$

$$\sin x \overset{(14.37)}{=} \frac{1}{2i}\left[e^{ix} - e^{-ix}\right] = -\frac{i}{2}\left[e^{ix} - e^{-ix}\right] \tag{18.9}$$

Sinh and Cosh

$$\cosh x \overset{(14.37)}{=} \frac{1}{2}\left[e^{x} + e^{-x}\right] = \cos(\mathrm{i}\,x) \tag{18.10}$$

$$\sinh x \overset{(14.37)}{=} \frac{1}{2}\left[e^{x} - e^{-x}\right] = -\,\mathrm{i}\sin(\mathrm{i}\,x) \tag{18.11}$$

Note

$$e^{x} = \cosh x + \sinh x \qquad e^{-x} = \cosh x - \sinh x \tag{18.12}$$

Note

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

Theorem 18.1 Addition Theorems:

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{18.13}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{18.14}$$

Werner Formulas

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha+\beta) + \sin(\alpha-\beta)\right] \tag{18.15}$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right] \tag{18.16}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha+\beta) + \cos(\alpha-\beta)\right] \tag{18.17}$$

Note

Using theorem 18.1 if follows:

$$\cos(\alpha \pm \pi) = -\cos\alpha \qquad \text{and} \qquad \sin(\alpha \pm \pi) = -\sin\alpha \tag{18.18}$$

Topology

Numerics

1. Machine Arithmetic's

1.1. Machine Numbers

Definition 20.1 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 20.2 Machine/Floating Point Numbers $\mathbb{F}$: Computers are only capable to represent a *finite, discrete* set of the real numbers $\mathbb{F} \subset \mathbb{R}$

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

Corollary 20.1 Closure: Machine numbers $\mathbb{F}$ are not *closed*[def. 12.6] under basic arithmetic operations:

$$\mathbb{F}\,\Omega\,\mathbb{F} \mapsto \not{\mathbb{F}} \qquad \Omega = \{+, -, *, /\} \tag{20.1}$$

Note

Corollary 20.1 provides a problem as the computer can only represent floating point number $\mathbb{F}$.

Definition 20.3 Floating Point Operation $\tilde{\Omega}$: Is a basic arithmetic operation that obtains a number $x \in \mathbb{F}$ by applying a function rd:

$$\mathbb{F}\,\tilde{\Omega}\,\mathbb{F} \mapsto \mathbb{F} \qquad \begin{aligned} \tilde{\Omega} &:= \mathrm{rd} \circ \Omega \\ \Omega &= \{+, -, *, /\} \end{aligned} \tag{20.2}$$

Definition 20.4 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:

$$\mathrm{rd} : \begin{cases} \mathbb{R} \mapsto \mathbb{F} \\ x \mapsto \max \arg\min_{\tilde{x} \in \mathbb{F}} |x - \tilde{x}| \end{cases} \tag{20.3}$$

Consequence

Basic arithmetic rules such as associativity do no longer hold for operations such as addition and subtraction.

Axiom 20.1 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:

$$\begin{aligned} x\tilde{\Omega}y &= (x \Omega y)\,(1+\delta) \qquad \Omega = \{+, -, *, /\} \\ \tilde{f}(x) &= f(x)(1+\delta) \qquad f \in \{\exp, \sin, \cos, \log, \dots\} \end{aligned} \tag{20.4}$$

with $|\delta| < \mathrm{EPS}$

Explanation 20.1 (axiom 20.1)**.** *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 20.5 Overflow: Result is bigger then the biggest representable floating point number.

Definition 20.6 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. 20.7] 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 20.7 Log sum Exponential:

$$\mathrm{LogSumExp}\,(x_1, \dots, x_n) := \log\left(\sum_{i=1}^{n} \mathrm{e}^{x_i}\right) \tag{20.5}$$

Formula 20.1 Log-Sum-Exp Trick:

$$\log\left(\sum_{i=1}^{n} \mathrm{e}^{x_i}\right) = a + \log \sum_{i=1}^{n} \mathrm{e}^{x_i - a} \qquad a := \max_{i \in \{1,\dots,n\}} x_i \tag{20.6}$$

Explanation 20.2 (formula 20.1)**.** *The value a can be any real value but for robustness one usually chooses the* max *s.t.*

- *The leading digits are preserved by pulling out the maximum a*
- *Inside the* log *only zero or negative numbers are exponentiated, so there can be no overflow.*
- *If there is underflow inside the* log *we know that at least the leading digits have been returned by the max.*

Proof.

$$\begin{aligned} \mathrm{LSE} &= \log\Big(\sum_{i=1}^{n} e^{x_i}\Big) = \log\Big(\sum_{i=1}^{n} e^{x_i - a} e^{a}\Big) \\ &= \log\big(e^{a} \sum_{i=1}^{n} e^{x_i - a}\big) = \log\Big(\sum_{i=1}^{n} e^{x_i - a}\Big) + \log(e^{a}) \\ &= \log\Big(\sum_{i=1}^{n} e^{x_i - a}\Big) + a \end{aligned}$$

□

Definition 20.8 Partition Π: Given an interval $[0, T]$ a sequence of values $0 < t_0 < \cdots < t_n < T$ is called a partition $\Pi\,(t_0, \dots, t_n)$ of this interval.

2. Convergence

2.1. O-Notation

2.1.1. Small $o(\cdot)$ Notation

Definition 20.9 Little o Notation:

$$f(n) = o(g(n)) \iff \lim_{n \to \infty} \frac{f(n)}{g(n)} = 0 \tag{20.7}$$

2.1.2. Big $\mathcal{O}(\cdot)$ Notation

2.2. Rate Of Convergence

Definition 20.10 Rate of Convergence: Is a way to measure the rate of convergence of a sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ to a value $\boldsymbol{x}^*$. Let $\rho \in [0, 1]$ be the *rate of convergence* and define:

$$\begin{aligned} &\lim_{k \mapsto \infty} \frac{\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{k} - \boldsymbol{x}^*\right\|} = \rho \\ \iff &\lim_{k \mapsto \infty} \left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \rho \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \qquad \forall k \in \mathbb{N}_0 \end{aligned} \tag{20.8}$$

Definition 20.11 Linear/Exponential Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *linearly* to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho \in (0, 1) \qquad \forall k \in \mathbb{N}_0 \tag{20.9}$$

Definition 20.12 Superlinear Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *superlinear* to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 1 \tag{20.10}$$

Definition 20.13 Sublinear Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *sublinear* to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ if it satisfies:

$$\rho = 0 \iff \left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| = o\left(\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|\right) \tag{20.11}$$

Definition 20.14 Logarithmic Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges *logarithmically* to $\boldsymbol{x}^*$ if it converges *sublinear*[def. 20.13] and additoinally satisfies

$$\rho = 0 \iff \left\|\boldsymbol{x}^{k+2} - \boldsymbol{x}^{k+1}\right\| = o\left(\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^{k}\right\|\right) \tag{20.12}$$

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:
$$\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \underbrace{(1-\alpha)}_{:=\rho} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$$
2. and that $(1-\alpha) = \exp(-\alpha)$ from which follows that:
$$\text{eq. (20.13)} \iff \left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \mathrm{e}^{-\alpha} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|$$

Definition 20.15 Convergence of order p: In order to distinguish *superlinear convergence* we define the order of convergence.

A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges superlinear with order $p \in \{2, \dots\}$ to $\boldsymbol{x}^*$ if it satisfies:

$$\lim_{k \mapsto \infty} \frac{\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\|}{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\|^{p}} = C \qquad C < 1 \tag{20.13}$$

Does this even exist/check if this is ture

Definition 20.16 Exponential Convergence: A sequence $\{\boldsymbol{x}^{(k)}\}_k \in \mathbb{R}^n$ converges exponentially with rate ρ to $\boldsymbol{x}^*$ if in the asymptotic limit $k \to \infty$ it satisfies:

$$\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \leqslant \rho^{k} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| \qquad \rho < 1 \left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\right\| \in o(\tag{20.14}$$

3. Numerical Quadrature

Definition 20.17 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:

$$\mathrm{order}(\mathcal{Q}_n) := \max\left\{n \in \mathbb{N}_0 : \mathcal{Q}_n(p) = \epsilon_a^b\, p(t)\,\mathrm{d}t \quad \forall p \in \mathcal{P}_n\right\} + 1 \tag{20.15}$$

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.

3.1. Composite Quadrature

Definition 20.18 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 $I_j := [x_{j-1}, x_j] \quad \forall j = 1, \dots, m \triangleq$ p.w. Quadrature:

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{j=1}^{m} \int_{x_{j-1}}^{x_j} f(t)\,\mathrm{d}t = \sum_{j=1}^{m} \mathcal{Q}_n(f_{I_j}) \tag{20.16}$$

Lemma 20.1 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:

$$\begin{aligned} \epsilon_n(f) &= \mathcal{O}\left(n^{-\min\{k,q\}}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{\min\{k,q\}}\right) \qquad \text{for } n \to \infty \\ &\overset{\text{corollary } 14.3}{=} \mathcal{O}\left(n^{-q}\right) = \mathcal{O}\left(h_{\mathcal{M}}^{q}\right) \qquad \text{with } h_{\mathcal{M}} = \frac{1}{n} \end{aligned} \tag{20.17}$$

Definition 20.19 Complexity W: Is the number of function evaluations $\triangleq$ number of quadrature points.

$$W(\mathcal{Q}(f)_n) = \#\text{f-eval} \triangleq n \tag{20.18}$$

Lemma 20.2 Error-Complexity $W(\epsilon_n(f))$: Relates the complexity to the quadrature error.

Assuming and quadrature error of the form :

$$\epsilon_n(f) = \mathcal{O}(n^{-q}) \iff \epsilon_n(f) = cn^{-q} \qquad c \in \mathbb{R}_+$$

the error complexity is algebraic (??) and is given by:

$$W(\epsilon_n(f)) = \mathcal{O}(\epsilon_n^{1/q}) = \mathcal{O}\left(\sqrt[q]{\epsilon_n}\right) \tag{20.19}$$

Proof. lemma 20.2: **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?

$$\frac{c \cdot n_n^q}{c \cdot n_o^q} = \frac{1}{\epsilon_n} \quad \Rightarrow \quad n_n = n_o \cdot \sqrt[q]{\epsilon_n} = \mathcal{O}(\sqrt[q]{\epsilon_n}) \tag{20.20}$$

□

Optimization

Definition 21.1 Fist Order Method: A first-order method is an algorithm that chooses the k-th iterate in

$$\boldsymbol{x}_0 + \mathrm{span}\{\nabla f(\boldsymbol{x}_0), \dots \nabla f(\boldsymbol{x}_{k-1})\} \qquad \forall k = 1, 2, \dots \tag{21.1}$$

Note

Gradient descent is a first order method

1. Lagrangian Optimization Theory

Add: derivation of lagragne function

Definition 21.2 (Primal) Constraint Optimization: **Given** an optimization problem with domain $\Omega \subseteq \mathbb{R}^d$:

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

Definition 21.3 Lagrange Function:

$$\mathscr{L}(\alpha, \beta, \boldsymbol{w}) := f(\boldsymbol{w}) + \alpha \boldsymbol{g}(\boldsymbol{w}) + \beta \boldsymbol{h}(\boldsymbol{w}) \tag{21.2}$$

Extremal Conditions

$$\nabla \mathscr{L}(\boldsymbol{x}) \overset{!}{=} 0 \qquad \text{Extremal point } \boldsymbol{x}^*$$

$$\frac{\partial}{\partial \beta} \mathscr{L}(\boldsymbol{x}) = h(\boldsymbol{x}) \overset{!}{=} 0 \qquad \text{Constraint satisfisfaction}$$

For the inequality constraints $g(\boldsymbol{x}) \leqslant 0$ we distinguish two situations:

Case I : $g(\boldsymbol{x}^*) < 0$ switch const. off

Case II : $g(\boldsymbol{x}^*) \geqslant 0$ optimze using active eq. constr.

$$\frac{\partial}{\partial \alpha} \mathscr{L}(\boldsymbol{x}) = g(\boldsymbol{x}) \overset{!}{=} 0 \qquad \text{Constraint satisfisfaction}$$

Definition 21.4 Lagrangian Dual Problem: Is given by:

$$\begin{aligned} &\text{Find} \quad \max \theta(\alpha, \beta) = \inf_{\boldsymbol{w} \in \Omega} \mathscr{L}(\boldsymbol{w}, \alpha, \beta) \\ &\textbf{s.t.} \quad \alpha_i \geqslant 0 \qquad 1 \leqslant i \leqslant k \end{aligned}$$

Solution Strategy

1. Find the extremal point $\boldsymbol{w}^*$ of $\mathscr{L}(\boldsymbol{w}, \alpha, \beta)$:
$$\left.\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\boldsymbol{w}^*} \overset{!}{=} 0 \tag{21.3}$$
2. Insert $\boldsymbol{w}^*$ into $\mathscr{L}$ and find the extremal point β^* of the resulting dual Lagrangian $\theta(\alpha, \beta)$ for the active constraints:
$$\left.\frac{\partial \theta}{\partial \beta}\right|_{\beta=\beta^*} \overset{!}{=} 0 \tag{21.4}$$
3. Calculate the solution $\boldsymbol{w}^*(\beta^*)$ 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 21.1 Upper Bound Dual Cost: Let $\boldsymbol{w} \in \Omega$ be a feasible solution of the primal problem [def. 21.2] and (α, β) a feasible solution of the respective dual problem [def. 21.4].
Then it holds that:

$$f(\boldsymbol{w}) \geqslant \theta(\alpha, \beta) \tag{21.5}$$

Proof.

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

□

Corollary 21.1 Duality Gap Corollary: The value of the dual problem is upper bounded by the value of the primal problem:

$$\sup\{\theta(\alpha, \beta) : \alpha \geqslant 0\} \leqslant \inf\{f(\boldsymbol{w}) : \boldsymbol{g}(\boldsymbol{w}) \leqslant 0, \boldsymbol{h}(\boldsymbol{w}) = 0\} \tag{21.6}$$

Theorem 21.2 Optimality: The triple $(\boldsymbol{w}^*, \alpha^*, \beta^*)$ 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:

$$f(\boldsymbol{w}^*) = \theta(\alpha^*, \beta^*) \tag{21.7}$$

Definition 21.5 Convex Optimization: **Given**: a convex function f and a convex set S solve:

$$\begin{aligned}&\min f(\boldsymbol{x}) \\ &\text{s.t.} \quad \boldsymbol{x} \in S\end{aligned} \tag{21.8}$$

Often S is specified using linear inequalities:

$$\text{e.g.} \qquad S = \left\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{x} \leqslant b\right\}$$

Theorem 21.3 Strong Duality: Given an convex optimization problem:

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

where g_i, h_i can be written as affine functions: $y(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w} - b$.
Then it holds that the duality gap is zero and we obtain an optimal solution.

Theorem 21.4 Kuhn-Tucker Conditions: Given an optimization problem with convex domain $\Omega \subseteq \mathbb{R}^d$,

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

with $f \in C^1$ convex and g_i, h_i affine.
Necessary and sufficient conditions for a normal point $\boldsymbol{w}^*$ to be an optimum are the existence of α^*, β^* s.t.:

$$\frac{\partial \mathscr{L}(\boldsymbol{w}, \alpha, \beta)}{\partial \boldsymbol{w}} \overset{!}{=} 0 \qquad \frac{\partial \mathscr{L}(\boldsymbol{w}^*, \alpha, \beta)}{\partial \beta} \overset{!}{=} 0 \tag{21.9}$$

under the condtions that:

- $\forall i_1, \ldots, k \qquad \alpha_i^* g_i(\boldsymbol{w}^*) = 0$, s.t.:
  - Inactive Constraint: $g_i(\boldsymbol{w}^*) < 0 \rightarrow \alpha_i = 0$.
  - Active Constraint:
    $g_i(\boldsymbol{w}^*) \not< 0 \rightarrow \alpha_i \geqslant 0 \qquad \text{s.t.} \qquad \alpha_i^* g_i(\boldsymbol{w}^*) = 0$

Consequence

We may become very sparce problems, if a lot of constraints are not actice $\iff \alpha_i = 0$.
Only a few points, for which $\alpha_i > 0$ may affact the decision surface.

Stochastics

Definition 21.6 Stochastics: Is a collective term for the areas of *probability theory* and *statistics*.

Definition 21.7 Statistics: Is concerned with the analysis of data/experiments in order to draw conclusion of the underlying governing models that describe these experiments.

Definition 21.8 Probability: Is concerned with the quantification of the uncertainty of random experiments by use of statistical models. Hence it is the opposite of statistics.

Definition 21.9 Probability: Probability is the measure of the likelihood that an event will occur in a Random Experiment. Probability is quantified as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty.

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 22.1 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 22.2 Sample Space Ω: Is the set of all possible outcomes (elementary events corollary 22.5) of an experiment see example 22.1

Definition 22.3 Event A**:** An "event" is a subset of the sample space Ω and is a property which can be observed to hold or not to hold *after* the experiment is done (example 22.2).
Mathematically speaking not every subset of Ω is an event and has an associated probability.
Only those subsets of Ω that are part of the corresponding σ-algebra $\mathcal{F}$ are events and have their assigned probability.

Corollary 22.1 : If the outcome ω of an experiment is in the subset A, then the event A is said to "have occured".

Corollary 22.2 Complement Set A^{C}**:** is the contrary event of A.

Corollary 22.3 The Union Set $A \cup B$**:** Let A, B be to evenest. The event "A or B" is interpreted as the union of both.

Corollary 22.4 The Intersection Set $A \cap B$**:** Let A, B be to evenest. The event "A and B" is interpreted as the intersection of both.

Corollary 22.5 The Elementary Event ω**:** Is a "singleton", i.e. a subset $\{\omega\}$ containing a single outcome ω of Ω.

Corollary 22.6 The Sure Event Ω**:** Is equal to the sample space as it contains all possible elementary events.

Corollary 22.7 The Impossible Event $\varnothing$**:** The impossible event i.e. nothing is happening is denoted by the empty set.

Definition 22.4 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 22.5 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:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{\delta(A)}{n} \qquad \text{where} \qquad \delta(A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases}$$

0.1. Sigma Algebras

Definition 22.6 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 $A_i \in \mathcal{F} : \bigcup_{i=1} A_i \in \mathcal{F}$

See example 22.3.

Corollary 22.8 $\mathcal{F}_{\min}$**:** $\mathcal{F} = \{\varnothing, \Omega\}$ is the simplest σ-algebra, telling us only if an event happened $\omega \in \Omega$ happened or not but not which one.

Corollary 22.9 $\mathcal{F}_{\max}$**:** $\mathcal{F} = 2^{\Omega}$ consists of all subsets of Ω and thus corresponds to full information i.e. we know if and which event happened.

Definition 22.7 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 22.10 $\mathcal{F}$-measurable Event: The elements $A_i \in \mathcal{F}$ are called *measurable sets* or *$\mathcal{F}$-measurable.*

Interpretation

The σ-algebra represents all of possible events of the experiment that we can detect.
Thus we call the sets in $\mathcal{F}$ measurable sets/events.
The sigma algebra is the mathematical construct that tells us how much information we obtain once we conduct some experiment.

Definition 22.8
Sigma Algebra generated by a subset of Ω $\sigma(\mathcal{C})$**:** Let $\mathcal{C}$ be a class of subsets of Ω. The σ-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the *smallest* sigma algebra $\mathcal{F}$ that included all elements of $\mathcal{C}$ see example 22.4.

Definition 22.9 Borel σ-algebra $\mathcal{B}(\mathbb{R})$**:** The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals in $\mathbb{R}$. The sets in contained in $\mathcal{B}(\mathbb{R})$ are called Borel sets.
The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward.
For all real numbers $a, b \in \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ contains various sets see example 22.5.

Why do we need Borel Sets

So far we only looked at atomic events ω, with the help of sigma algebras we are now able to measure continuous events s.a. $[0, 1]$.

Corollary 22.11 : The Borel σ-algebra of $\mathbb{R}$ is generated by intervals of the form $(-\infty, a]$, where $a \in \mathbb{Q}$ ($\mathbb{Q}$ =rationals). See proof section 13.

Definition 22.10 ($\mathbb{P}$)-trivial Sigma Algebra:
is a σ-algebra $\mathcal{F}$ for which each event has a probability of zero or one:

$$\mathbb{P}(A) \in \{0, 1\} \qquad \forall A \in \mathcal{F} \qquad (22.1)$$

Interpretation

A trivial sigma algebra means that all events are almost surely constant and that there exist no non-trivial information.
An example of a trivial sigma algebra is $\mathcal{F}_{\min} = \{\Omega, \varnothing\}$.

0.2. Measures

Definition 22.11 Measure μ**:**
A measure defined on a measurable space $\{\Omega, \mathcal{F}\}$ is a function/map:

$$\mu : \mathcal{F} \mapsto [0, \infty] \qquad (22.2)$$

for which holds:

- $\mu(\varnothing) = 0$
- countable additivity [def. 22.12]

Definition 22.12 Countable/σ-Additive Function:
Given a function μ defined on a σ-algebra $\mathcal{F}$.
The function μ is said to be countable additive if for every countable sequence of pairwise disjoint elements $(F_i)_{i\geqslant 1}$ of $\mathcal{F}$ it holds that:

$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i) \qquad \text{for all} \qquad F_j \cap F_k = \varnothing \quad \forall j \neq k \qquad (22.3)$$

Corollary 22.12 Additive Function: A function that satisfies countable additivity, is also additive, meaning that for every $F, G \in \mathcal{F}$ it holds:

$$\mu(F \cup G) = \mu(F) + \mu(G) \qquad \Longleftrightarrow \qquad F \cap G = \varnothing \qquad (22.4)$$

Intuition

If we take two event that cannot occur simultaneously, then the probability that at least one vent occurs is just the sum of the measure (probabilities) of the original events.

Definition 22.13 Equivalent Measures $\mu \sim \nu$**:**
Let μ and ν be two measures defined on a measurable space[def. 22.7] $(\Omega, \mathcal{F})$. The two measures are said to be equivalent if it holds that:

$$\mu(A) > 0 \iff \nu(A) > 0 \qquad \forall A \subseteq \mathcal{F} \qquad (22.5)$$

this is equivalent to μ and ν having equivalent null sets:

$$\mathcal{N}_\mu = \mathcal{N}_\nu \qquad \begin{matrix} \mathcal{N}_\mu = \{A \in \mathcal{A} | \mu(A) = 0\} \\ \mathcal{N}_\nu = \{A \in \mathcal{A} | \nu(A) = 0\} \end{matrix} \qquad (22.6)$$

see example 22.6

Definition 22.14 Measure Space $\{\mathcal{F}, \Omega, \mu\}$**:**
The triplet of sample space, sigma algebra and a measure is called a measure space.

Definition 22.15 Lebesgue Measure on $\mathcal{B}$ λ**:**
Is the measure defined on the measurable space $\{\mathbb{R}, \mathcal{B}(\mathbb{R})\}$ which assigns the measure of each interval to be its length:

$$\lambda([a, b]) = b - a \qquad (22.7)$$

Corollary 22.13 Lebesgue Measure of Atomitcs:

- The Lebesgue measure of a set containing only one point must be zero:
  $$\lambda(\{a\}) = 0 \qquad (22.8)$$
- The Lebesgue measure of a set containing countably many points $A = \{a_1, a_2 \ldots, a_n\}$ must be zero:
  $$\lambda(A) + \sum_{i=1}^{n} \lambda(\{a_i\}) = 0 \qquad (22.9)$$
- The Lebesgue measure of a set containing uncountably many points $A = \{a_1, a_2 \ldots, \}$ can be either zero, positive and finite or infinite.

0.3. Probability/Kolomogorov's Axioms 1931

One problem we are still having is the range of μ, by standardizing the measure we obtain a well defined measure of events.

Axiom 22.1 Non-negativity: The probability of an event is a non-negative real number:

$$\text{If } A \in \mathcal{F} \qquad \text{then} \qquad \mathbb{P}(A) \geqslant 0 \qquad (22.10)$$

Axiom 22.2 Unitairity: The probability that at least one of the elementary events in the entire sample space Ω will occur is equal to one:

$$\text{The certain event} \qquad \mathbb{P}(\Omega) = 1 \qquad (22.11)$$

Axiom 22.3 σ-additivity: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are mutually disjoint, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \qquad (22.12)$$

Corollary 22.14 : As a consequence of this it follows:

$$\mathbb{P}(\varnothing) = 0 \qquad (22.13)$$

Corollary 22.15 Complementary Probability:

$$\mathbb{P}(A^{\mathrm{C}}) = 1 - \mathbb{P}(A) \qquad \text{with} \qquad A^{\mathrm{C}} = \Omega - A \qquad (22.14)$$

Definition 22.16 Probability Measure $\mathbb{P}$**:**
a probability measure is function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ defined on a σ-algebra $\mathcal{F}$ of a sample space Ω that satisfies the probability axioms.

1. Conditional Probability

Definition 22.17 Conditional Probability: Let A, B be events, with $\mathbb{P}(B) \neq 0$. Then the conditional probability of the event A given B is defined as:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) \neq 0 \qquad (22.15)$$

2. Independent Events

Theorem 22.1
Independent Events: Let A, B be two events. A and B are said to be independent iffy:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \qquad \begin{matrix} \mathbb{P}(A|B) = \mathbb{P}(A), & \mathbb{P}(B) > 0 \\ \mathbb{P}(B|A) = \mathbb{P}(B), & \mathbb{P}(A) > 0 \end{matrix} \qquad (22.16)$$

Note

The requirement of no impossible events follows from [def. 22.17]

Corollary 22.16 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:

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cap \mathbb{P}(A_j) \qquad i \neq j, \quad \forall i, j \in \mathcal{A} \qquad (22.17)$$

Corollary 22.17 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:

$$\mathbb{P}\left(\bigcap_{i=i}^{k} B_i\right) = \prod_{i=1}^{k} \mathbb{P}(B_i) \qquad \begin{matrix} \forall \{B_i\}_{i=1}^k \subseteq \{A_i\}_{i=1}^n \\ k \leqslant n, \qquad \{A_i\}_{i=1}^n \in \mathcal{A} \end{matrix} \qquad (22.18)$$

3. Product Rule

Law 22.1 Product Rule: Let A, B be two events then the probability of both events occurring simultaneously is given by:

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) \qquad (22.19)$$

Law 22.2
Generalized Product Rule/Chain Rule: is the generalization of the product rule?? to n events $\{A_i\}_{i=1}^n$

$$\mathbb{P}\left(\bigcap_{i=i}^{k} E_i\right) = \prod_{k=1}^{n} \mathbb{P}\left(E_k \middle| \bigcap_{i=i}^{k-1} E_i\right) = \qquad (22.20)$$

$$= \mathbb{P}(E_n|E_{n-1} \cap \ldots \cap E_1) \cdot \mathbb{P}(E_{n-1}|E_{n-2} \cap \ldots \cap E_1) \cdots$$
$$\cdots \mathbb{P}(E_3|E_2 \cap E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)$$

4. Law of Total Probability

Definition 22.18 Complete Event Field: A complete event field $\{A_i : i \in I \subseteq \mathbb{N}\}$ is a countable or finite partition of Ω that is the partitions $\{A_i : i \in I \subseteq \mathbb{N}\}$ are a *disjoint union* the sample space:

$$\bigcup_{i\in I} A_i = \Omega \qquad A_i \cap A_j = \varnothing \qquad i \neq j, \forall i,j \in I \tag{22.21}$$

Theorem 22.2
Law of Total Probability/Partition Equation:
Let $\{A_i : i \in I\}$ be a complete event field[def. 22.18] then it holds for $B \in \mathcal{B}$:

$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{22.22}$$

5. Bayes Theorem

Law 22.3 Bayes Rule: Let A, B be two events s.t. $\mathbb{P}(B) > 0$ then it holds:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \qquad \mathbb{P}(B) > 0 \tag{22.23}$$

follows directly fromeq. (22.19).

Theorem 22.3 Bayes Theorem: Let $\{A_i : i \in I\}$ be a complete event field[def. 22.18] and $B \in \mathcal{B}$ a random event s.t. $\mathbb{P}(B) > 0$, then it holds:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i\in I}\mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{22.24}$$

proof section 13

Distributions on $\mathbb{R}$

6.1. Distribution Function

Definition 22.19 Distribution Function of $\mathbb{P}$ F**:**
The *distribution function* F induced by a a probability $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$ is the function:

$$F(x) = \mathbb{P}((\infty, x]) \tag{22.25}$$

Theorem 22.4 : 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 22.18 : A probability $\mathbb{P}$ is uniquely determined by a distribution function F
That is if there exist another probability $\mathbb{Q}$ s.t.

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

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

6.2. Random Variables

A random variable X is a quantity that is not a variable in the classical sense but a variable with respect to the outcome of an experiment. Thus it is actually not a variable but a function/map.
Its value is determined in two steps:
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 $\mathcal{E} = \mathbb{R}^n$:

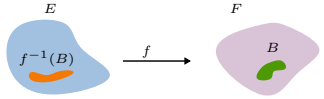
$$X : \Omega \mapsto \mathcal{E} \qquad \omega \mapsto X(\omega)$$

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 22.20 $\mathcal{E}$-measurable function: Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces. A function $f : E \mapsto F$ is called measurable (relative to $\mathcal{E}$ and $\mathcal{F}$) if

$$\forall B \in \mathcal{F} : \qquad f^{-1}(B) = \{\omega \in \mathcal{E} : f(\omega) \in B\} \in \mathcal{E} \tag{22.26}$$



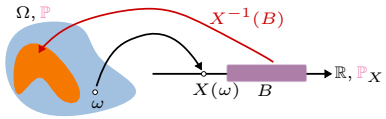
Interpretation

The pre-image[def. 14.9] of B under f i.e. $f^{-1}(B)$ maps all values of the target space F back to the sample space $\mathcal{E}$ (for all possible $B \in \mathcal{F}$).

Definition 22.21 Random Variable: A real-valued random variable (vector) X, defined on a probability space $\{\Omega, \mathcal{E}, \mathbb{P}\}$ is an $\mathcal{E}$-measurable function mapping, if it maps its sample space Ω into a target space $(F, \mathcal{F})$:

$$X : \Omega \mapsto \mathcal{F} \quad (\mathcal{F}^n) \tag{22.27}$$

Since X is $\mathcal{E}$-measurable it holds that $X^{-1} : \mathcal{F} \mapsto \mathcal{E}$



Corollary 22.19 : Usually $F = \mathbb{R}$, which usually amounts to using the Borel σ-algebra $\mathcal{B}$ of $\mathbb{R}$.

Corollary 22.20 Random Variables of Borel Sets: Given that we work with Borel σ-algebras then the definition of a random variable is equivalent to (due to corollary 22.11):

$$\begin{aligned} X^{-1}(B) &= X^{-1}((-\infty, a]) \\ &= \{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{E} \quad \forall a \in \mathbb{R} \end{aligned} \tag{22.28}$$

Definition 22.22
Realization of a Random Variable $x = X(\omega)$**:** Is the value of a random variable that is actually observed after an experiment has been conducted. In order to avoid confusion lower case letters are used to indicate actual observations/realization of a random variable.

Corollary 22.21 Indicator Functions $I_A(\omega)$**:**
An important class of measurable functions that can be used as r.v. are indicator functions:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} \tag{22.29}$$

We know that a probability measure $\mathbb{P}$ on $\mathbb{R}$ is characterized by the quantities $\mathbb{P}((-\infty, a])$. Thus the quantities.

Corollary 22.22 : Let $(F, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ and let $(E, \mathcal{E})$ and arbitrary measurable space. Let X be a real value function on E.
Then it holds that X is measurable *if and only if*
$\{X \leqslant a\} = \{\omega : X(\omega) \leqslant a\} = X^{-1}((-\infty, a]) \in \mathcal{E}$, each $a \in \mathbb{R}$
or $\{X < a\} \in \mathcal{E}$.

Explanation 22.1 (corollary 22.22). *A random variable is a function that is measurable if and only if its distribution function is defined.*

6.3. The Law of Random Variables

Definition 22.23 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:

$$\mathbb{P} : \mathcal{B} \mapsto [0, 1] \tag{22.30}$$
$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}(\omega : X(\omega) \in B) \qquad \forall b \in \mathcal{E}$$

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 22.5 : The law/distribution of X is a probability measure $\mathbb{P}$ on $(E, \mathcal{E})$.

Definition 22.24
(Cumulative) Distribution Function F_X**:**
Given a real-valued r.v. then its *cumulative distribution function* is defined as:

$$F_X(x) = \mathbb{P}^X((-\infty, x]) = \mathbb{P}(X \leqslant x) \tag{22.31}$$

Corollary 22.23 : The distribution of $\mathbb{P}^X$ of a real valued r.v. is entirely characterized by its cumulative distribution function F_X[def. 22.31].

Property 22.1:

$$\mathbb{P}(X > x) = 1 - F_X(x) \tag{22.32}$$

Property 22.2: Probability of $X \in [a, b]$

$$\mathbb{P}(a < X \leqslant B) = F_X(b) - F_X(a) \tag{22.33}$$

6.4. Probability Density Function

Definition 22.25 Continuous Random Variable: Is a r.v. for which a probability density function f_X exists.

Definition 22.26 Probability Density Function: Let X be a r.v. with associated cdf F_X. If F_X is continuously integrable for all $x \in \mathbb{R}$ then X has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \tag{22.34}$$

or alternatively:

$$f_X(x) = \lim_{\epsilon\to 0} \frac{\mathbb{P}(x \leqslant X \leqslant x + \epsilon)}{\epsilon} \tag{22.35}$$

Corollary 22.24 $\mathbb{P}(X = b) = 0, \quad \forall b \in \mathbb{R}$**:**

$$\mathbb{P}(X = b) = \lim_{a\to b} \mathbb{P}(a < X \leqslant b) = \lim_{a\to b} \int_a^b f(x) = 0 \tag{22.36}$$

Corollary 22.25 corollary 22.24: From corollary 22.24 it follows that the exact borders are not necessary:

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}(a \leqslant X < b) \\ &= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X <\leqslant b) \end{aligned}$$

Corollary 22.26 :

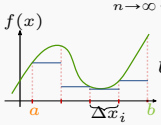
$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{22.37}$$

Notes
- Often the cumulative distribution function is referred to as "cdf" or simply *distribution function.*
- Often the probability density function is referred to as "pdf" or simply *density.*

6.5. Lebesgue Integration

Problems of Riemann Integration
- Difficult to extend to higher dimensions – general domains of definitions $f : \Omega \mapsto \mathbb{R}$
- Depends on continuity
- Integration of limit processes require strong uniform convergence in order to integrate limit processes

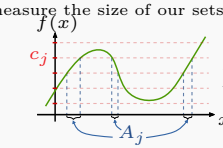
$$\lim_{n\to\infty} \int f(x)\,\mathrm{d}x \overset{\text{str. u.c.}}{=} \int \lim_{n\to\infty} f(x)\,\mathrm{d}x$$



$$U(p) = \sum_i \sup(f(x_i)) \cdot \Delta x_i \xrightarrow{n\to\infty} \int f\,\mathrm{d}x$$

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 22.27 Lebesgue Integral:

$$\lim_{n\to\infty} \sum_{i=1}^{n} c_i \mu(A_i) = \int_\Omega f\,\mathrm{d}\mu \qquad \begin{matrix} f(x) \approx c_i \\ \forall x \in A_i \end{matrix} \tag{22.38}$$

Definition 22.28
Simple Functions (Random Variables): A r.v. X is called simple if it takes on only a finite number of values and hence can be written in the form:

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} \qquad a_i \in \mathbb{R} \qquad \mathcal{A} \ni A_i = \begin{cases} 1 & \text{if } \{X = a_i\} \\ 0 & \text{else} \end{cases} \tag{22.39}$$

6.6. Independent Random Variables

We have seen that two events A and B are independent if knowledge that B has occurred does not change the probability that A will occur theorem 22.1.
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 22.29 Independent Random Variables:
Two real valued random variables X and Y are said to be independent iff:

$$\mathbb{P}(X \leqslant x | Y \leqslant y) = \mathbb{P}(X \leqslant x) \qquad \forall x, y \in \mathbb{R} \tag{22.40}$$

which amounts to:

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}(\{X \leqslant x\} \cap \{Y \leqslant y\}) = \mathbb{P}(X \leqslant x, Y \leqslant y) \\ &= F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R} \end{aligned} \tag{22.41}$$

or alternatively iff:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \quad \forall A, B \in \mathcal{B} \tag{22.42}$$

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 22.30
Independent Identically Distributed:

7. Product Rule

Law 22.4 Product Rule: Let X, Y be two random variables then their jo

Law 22.5
Generalized Product Rule/Chain Rule:

8. Change Of Variables Formula

Formula 22.1
(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. 14.12]) function.
Let: $\mathcal{X}_y = x_i$ be the set of all $x_i \in \mathcal{X}$ s.t. $y = g(x_i)$.
Then the pmf of Y is given by:

$$p_Y(y) = \sum_{x_i \in \mathcal{X}_y} p_X(x_i) = \sum_{x\in\mathcal{Y}: g(x)=y} p_X(x) \tag{22.43}$$

see proof section 13

Formula 22.2
(Scalar Continuous) Change of Variables:
Let $X \sim f_X$ be a continuous r.v. and let g be an arbitrary strictly monotonic[def. 14.12] function.
Define a new r.v. Y as

$$\mathcal{Y} = \{y | y = g(x), \forall x \in \mathcal{X}\} \tag{22.44}$$

then the pdf of Y is given by:

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(x)\left|\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\right| \tag{22.45}$$

$$= f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(g^{-1}(y))\right|} \tag{22.46}$$

Formula 22.3
(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. 14.12] function

$$g : \mathbb{R}^n \mapsto \mathbb{R}^m$$

Define a new r.v. Y as

$$\mathcal{Y} = \{\boldsymbol{y} | \boldsymbol{y} = g(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}\} \tag{22.47}$$

and let $h(\boldsymbol{x}) := g(\boldsymbol{x})^{-1}$ then the pdf of Y is given by:

$$\begin{aligned} f_Y(\boldsymbol{y}) &= f_X(x_1, \ldots, x_n) \cdot |J| \\ &= f_X(h_1(\boldsymbol{y}), \ldots, h_n(\boldsymbol{y})) \cdot |J| \\ &= f_X(\boldsymbol{y})|\det D_{\boldsymbol{x}} h(\boldsymbol{x})|\Big|_{\boldsymbol{x}=\boldsymbol{y}} \\ &= f_X(g^{-1}(\boldsymbol{y}))\left|\det\left(\frac{\partial g}{\partial \boldsymbol{x}}\right)\right|^{-1} \end{aligned} \tag{22.48}$$

where $J = \det Dh$ is the Jaccobian[def. 15.4].
See also proof section 13 and example 22.8

Note
A monotonic function is required in order to satisfy inevitability.

Probability Distributions on $\mathbb{R}^n$

10. Joint Distribution

Definition 22.31
Joint (Cumulative) Distribution Function $F_{\boldsymbol{X}}$**:**
Let $\boldsymbol{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$, then its *cumulative distribution function* is defined as:

$$\begin{aligned} F_{\boldsymbol{X}}(\boldsymbol{x}) &= \mathbb{P}^X((-\infty, \boldsymbol{x}]) = \mathbb{P}(\boldsymbol{X} \leqslant \boldsymbol{x}) \\ &= \mathbb{P}(X_1 \leqslant x_1, \ldots X_n \leqslant x_n) \end{aligned} \tag{22.49}$$

Definition 22.32 Joint Probability Distribution:
Let $\boldsymbol{X} = (X_1 \cdots\cdots X_n)$ be a random vector in $\mathbb{R}^n$ with associated cdf $F_{\boldsymbol{X}}$. If $F_{\boldsymbol{X}}$ is continuously integrable for all $\boldsymbol{x} \in \mathbb{R}$ then $\boldsymbol{X}$ has a *probability density* f_X defined by:

$$F_X(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(y_1, \ldots, y_n)\, \mathrm{d}y_1\, \mathrm{d}y_n \tag{22.50}$$

or alternatively:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + \epsilon, \ldots, x_n \leqslant X_n \leqslant x_n + \epsilon)}{\epsilon} \tag{22.51}$$

10.1. Marginal Distribution

Definition 22.33 Marginal Distribution:

11. The Expectation

Definition 22.34 Expectation:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int_\Omega X\, \mathrm{d}\mathbb{P} \tag{22.52}$$

Corollary 22.27 Expectation of simple r.v.:
If X is a simple[def. 22.28] r.v. its *expectation* is given by:

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \mathbb{P}(A_i) \tag{22.53}$$

11.1. The Jensen Inequality

Theorem 22.6 Jensen Inequality: Let X be a random variable and g some function, then it holds:

$$\begin{matrix} g(\mathbb{E}[X]) \geqslant \mathbb{E}[g(X)] \\ g(\mathbb{E}[X]) \leqslant \mathbb{E}[g(X)] \end{matrix} \quad \text{if} \quad \begin{matrix} g \text{ is convex}^{[\text{def. } 14.20]} \\ g \text{ is concave}^{[\text{def. } 14.21]} \end{matrix} \tag{22.54}$$

11.2. Law of the Unconscious Statistician

Law 22.6 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:

$$\mathbb{E}_Y[y] = \mathbb{E}_X g(x) \tag{22.55}$$

Consequence
Hence if we p_X we do not have to first calculate p_Y in order to calculate $\mathbb{E}_Y[y]$.

11.3. Properties

11.3.1. Linear Operators

add

11.3.2. Quadratic Form

Definition 22.35 proof 13
Expectation of a Quadratic Form:
Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be a random vector with $\mathbb{E}[\boldsymbol{\epsilon}] = \mu$ and $\mathbb{V}[\boldsymbol{\epsilon}] = \Sigma$:

$$\mathbb{E}[\boldsymbol{\epsilon}^\intercal \boldsymbol{A} \boldsymbol{\epsilon}] = \mathrm{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{\mu}^\intercal \boldsymbol{A} \boldsymbol{\mu} \tag{22.56}$$

12. Moment Generating Function (MGF)

Definition 22.36 Moment of Random Variable: The i-th moment of a random variable X is defined as (if it exists):

$$m_i := \mathbb{E}\left[X^i\right] \tag{22.57}$$

Definition 22.37 ψ_X
Moment Generating Function (MGF):

$$\psi_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \qquad t \in \mathbb{R} \tag{22.58}$$

Corollary 22.28 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:

$$\psi_{S_n}(t) = \psi_{X_1}(t) \cdots \psi_{X_n}(t) \qquad S_n := X_1 + ldots X_n \tag{22.59}$$

Corollary 22.29 : The i-th moment of a random variable is the i-th derivative of its associated moment generating function evaluated zero:

$$\mathbb{E}\left[X^i\right] = \psi_X^{(i)}(0) \tag{22.60}$$

13. The Characteristic Function

Transforming probability distributions using the Fourier transform is a handy tool in probability in order to obtain properties or solve problems in another space before transforming them back.

Definition 22.38 $\hat{\mu}$
Fourier Transformed Probability Measure:

$$\hat{\mu} = \int \mathrm{e}^{i\langle u, x\rangle} \mu(\mathrm{d}x) \tag{22.61}$$

Corollary 22.30 : As $e^{i\langle u,x\rangle}$ can be rewritten using formulaeqs. (18.5) and (18.6) it follows:

$$\hat{\mu} = \int \cos(\langle u, x\rangle)\, \mu(\mathrm{d}x) + i \int \sin(\langle x, u\rangle)\, \mu(\mathrm{d}x) \tag{22.62}$$

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 22.39 Characteristic Function φ_X**:** Let $\boldsymbol{X}$ be an $\mathbb{R}^n$-valued random variable. Its characteristic function $\varphi_{\boldsymbol{X}}$ is defined on $\mathbb{R}^n$ as:

$$\varphi_{\boldsymbol{X}}(u) = \int \mathrm{e}^{i\langle \boldsymbol{u}, \boldsymbol{x}\rangle} \mathbb{P}^X(\mathrm{d}\boldsymbol{x}) = \widehat{\mathbb{P}^X}(\boldsymbol{u}) \tag{22.63}$$

$$= \mathbb{E}\left[\mathrm{e}^{i\langle \boldsymbol{u}, \boldsymbol{x}\rangle}\right] \tag{22.64}$$

Corollary 22.31 : 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 22.7 : 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 22.8 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 22.32 : Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ 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:

$$\varphi_X(u_1, \ldots, u_n) = \prod_{j=1}^{n} \varphi_{X_j}(u_j) \tag{22.65}$$

Proofs

Proof. corollary 22.11: Let $\mathcal{C}$ denote all open intervals.
Since every open set in $\mathbb{R}$ is the countable union of open intervals[def. 12.8], it holds that $\sigma(\mathcal{C})$ is the Borel σ-algebra of $\mathbb{R}$.
Let $\mathcal{D}$ denote all intervals of the form $(-\infty, a]$, $a \in \mathbb{Q}$.
Let $a, b \in \mathcal{C}$, and let
- $(a_n)_{n>1}$ be a sequence of rationals *decreasing* to a and
- $(b_n)_{n>1}$ be a sequence of rationals *increasing strictly* to b

$$(a, b) = \cup_{n=1}^{\infty}(a_n, b_n] = \cup_{n=1}^{\infty}\left(-\infty, b_n] \cap (-\infty, a_n\right]^{\mathrm{C}}\right)$$

Thus $\mathcal{C} \subset \sigma(\mathcal{D})$, whence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$ **but** as each element of $\mathcal{D}$ is a closed subset, $\sigma(\mathcal{D})$ must also be contained in the Borel sets $\mathcal{B}$ with

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma((D) \subset \mathcal{B}$$

□

Proof. theorem 22.3 Plug eq. (22.22) into the denominator and **??** into the nominator and then use [def. 22.17]:

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

□

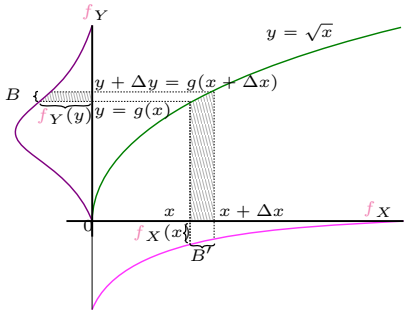
Proof. formula 22.1:

$$Y = g(X) \iff \mathbb{P}(Y = y) = \mathbb{P}(x \in \mathcal{X}_y) = \mathrm{p}_Y(y)$$

□

Proof. formula 22.2 (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. formula 22.2 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. formula 22.2: **Let** $B = [x, x + \Delta x]$ and $B' = [y, y + \Delta y] = [g(x), g(x + \Delta x)]$ we know that the probability of equal events is equal:

$$y = g(x) \qquad \Rightarrow \qquad \mathbb{P}(y) = \mathbb{P}(g(x)) \text{ (for disc. rv.)}$$

Now lets consider the probability for the continuous r.v.s:

$$\mathbb{P}(X \in B) = \int_x^{x+\Delta x} f_X(t)\,\mathrm{d}t \xrightarrow{\Delta x \to 0} |\Delta x \cdot f_x(x)|$$

For y we use Taylor (**??**)

$$\begin{aligned} g(x + \Delta x) &\overset{\text{eq. (14.41)}}{=} g(x) + \frac{\mathrm{d}g}{\mathrm{d}x}\Delta y \qquad \text{for } \Delta x \to 0 \\ &= y + \Delta y \qquad \text{with } \Delta y := \frac{\mathrm{d}g}{\mathrm{d}x} \cdot \Delta x \end{aligned} \tag{22.66}$$

Thus for $\mathbb{P}(Y \in B')$ it follows:

$$\mathbb{P}(X \in B') = \int_y^{y+\Delta y} f_Y(t)\,\mathrm{d}t \xrightarrow{\Delta y \to 0} |\Delta y \cdot f_Y(y)|$$

$$= \left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x \cdot f_Y(y)\right|$$

Now we simply need to related the surface of the two pdfs:

$$B = [x, x + \Delta x] \overset{\text{same surfaces}}{\propto} [y, y + \Delta y] = B'$$

$$\mathbb{P}(Y \in B) = \mathbb{P}(X \in B')$$

$$\overset{\Delta y \to 0}{\iff} |f_Y(y) \cdot \Delta y| = \left|f_Y(y) \cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x)\Delta x\right| = |f_X(x) \cdot \Delta x|$$

$$f_Y(y)\left|\cdot \frac{\mathrm{d}g}{\mathrm{d}x}(x)\right||\Delta x| = f_X(x) \cdot |\Delta x|$$

$$\Rightarrow f_Y(y) = \frac{f_X(x)}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right|} = \frac{f_X(g^{-1}(y))}{\left|\frac{\mathrm{d}g}{\mathrm{d}x}g^{-1}(y)\right|}$$

□

Proof. [def. 22.35]

$$\begin{aligned} \mathbb{E}[\boldsymbol{\epsilon}^\intercal \boldsymbol{A} \boldsymbol{\epsilon}] &\overset{\text{eq. (17.14)}}{=} \mathbb{E}[\mathrm{tr}(\boldsymbol{\epsilon}^\intercal \boldsymbol{A} \boldsymbol{\epsilon})] \\ &\overset{\text{eq. (17.16)}}{=} \mathbb{E}[\mathrm{tr}(\boldsymbol{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal)] \\ &= \mathrm{tr}(\mathbb{E}[\boldsymbol{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal]) \\ &= \mathrm{tr}(\boldsymbol{A}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal]) \\ &= \mathrm{tr}(\boldsymbol{A}(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal)) \\ &= \mathrm{tr}(\boldsymbol{A}\Sigma) + \mathrm{tr}(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^\intercal) \\ &\overset{\text{eq. (17.14)}}{=} \mathrm{tr}(\boldsymbol{A}\Sigma) + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^\intercal \end{aligned}$$

□

Examples

Example 22.1 :
- Toss of a coin (with head and tail): $\Omega = \{H, T\}$.
- Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$
- A cubic die: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
- The positive integers: $\Omega = \{1, 2, 3, \ldots\}$
- The reals: $\Omega = \{\omega | \omega \in \mathbb{R}\}$

Example 22.2 :
- 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 22.3 : 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 22.4 : If we are only interested in the subset-set $A \in \Omega$ of our experiment, then we can look at the corresponding generating σ-algebra $\sigma(A) = \left\{\varnothing, A, A^{\mathrm{C}}, \Omega\right\}$.

Example 22.5 :

- 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: $\{a_1, \ldots, a_n\} = \bigcup_{k=1}^{n} a_k$.

Example 22.6 Equivalent (Probability) Measures:

$$\Omega = \{1, 2, 3\} \qquad \begin{aligned} \mathbb{P}(\{1, 2, 3\}) &= \{2/3, 1/6, 1/6\} \\ \tilde{\mathbb{P}}(\{1, 2, 3\}) &= \{1/3, 1/3, 1/3\} \end{aligned}$$

Example 22.7 :

add example fat book p.1286

add example prob th book 4

Example 22.8 formula 22.2: 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:

$$h(u, v) = \begin{cases} h_1(u, v) = \frac{u+v}{2} \\ h_2(u, v) = \frac{u-v}{2} \end{cases} \qquad J = \det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = -\frac{1}{2}$$

$$\begin{aligned} f_{U,V} &= f_{X,Y}(\underline{x}, \underline{y}) \cdot \frac{1}{2} \\ &\overset{\text{indp.}}{=} f_X(\underline{x}) \cdot f_X(\underline{y}) \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\underline{x}^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\underline{y}^2}{2}} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left\{\left(\frac{\underline{u}+\underline{v}}{2}\right)^2 + \left(\frac{\underline{u}-\underline{v}}{2}\right)^2 / 2\right\}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} \mathrm{e}^{-\frac{u^2}{4}} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} \mathrm{e}^{-\frac{v^2}{4}} \end{aligned}$$

Thus U, V are independent r.v. distributed as $\mathcal{N}(0, 2)$.

Combinatorics

0.1. Permutations

Definition 23.1 Permutation $n!$: Given a set[def. 12.1] $\mathcal{S}$ of n distinct objects, into how many distinct sequences/orders can we arrange/permutate those distinct objects

$$P(\mathcal{S}) = n! \iff P(\mathcal{S}) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 \tag{23.1}$$

If there exists multiple n_j objects of the same kind within $\mathcal{S}$ with $j \in 1, \ldots, n-1$ then we need to divide by those permutations:

$$P(\mathcal{S}) = \frac{n!}{n_1! \cdot \ldots \cdot n_k} \quad \text{s.t.} \quad \sum_{i=1}^{k} n_i \leqslant n \tag{23.2}$$

Note

This is because the sequence/order does not change if we exchange objects of the same kind (e.g. red ball by red ball).

Statistics

Delete/Move the following stuff appropriately

The probability that a discret random variable x is equal to some value $\bar{x} \in \mathcal{X}$ is:

$$\mathrm{p}_x(\bar{x}) = \mathbb{P}(x = \bar{x})$$

addapet

Definition 24.1 Almost Surely (a.s.): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event $\omega \in \mathcal{F}$ happens almost surely iff

$$\mathbb{P}(\omega) = 1 \iff \omega \text{ happens a.s.} \tag{24.1}$$

Definition 24.2 Probability Mass Function (PMF):

Definition 24.3 Discrete Random Variable (DVR): The set of possible values $\bar{x}$ of $\mathcal{X}$ is countable of finite.

$$\mathcal{X} = \{0, 1, 2, 3, 4, \ldots, 8\} \qquad \mathcal{X} = \mathbb{N} \tag{24.2}$$

Definition 24.4 Probability Density Function (PDF): Is real function $f : \mathbb{R}^n \to [0, \infty)$ that satisfies:

Non-negativity: $$f(x) \geqslant 0, \quad \forall x \in \mathbb{R}^n \tag{24.3}$$

Normalization: $$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \overset{!}{=} 1 \tag{24.4}$$

Must be integrable (24.5)

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 24.5 Continuous Random Variable (CRV): A real random variable (rrv) X is said to be (absolutely) continuous if there exists a pdf ([def. 24.4]) f_X s.t. for any subset $B \subset \mathbb{R}$ it holds:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,\mathrm{d}x \tag{24.6}$$

Property 24.1 Zero Probability: If X is a continuous rrv ([def. 24.5]), then:

$$\mathbb{P}(X = a) = 0 \qquad \forall a \in \mathbb{R} \tag{24.7}$$

Property 24.2 Open vs. Closed Intervals: For any real numbers a and b, with $a < b$ it holds:

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b) \end{aligned} \tag{24.8}$$

$\iff$ 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 24.1 : In particular for any real numbers a and b with $a < b$, letting $B = [a, b]$ we obtain:

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_x(x)\,\mathrm{d}x$$

Proof. Property 24.1:

$$\begin{aligned} \mathbb{P}(X = a) &= \lim_{\Delta x \to 0} \mathbb{P}(X \in [a, a + \Delta x]) \\ &= \lim_{\Delta x \to 0} \int_a^{a+\Delta x} f_X(x)\,\mathrm{d}x = 0 \end{aligned}$$

□

Proof. Property 24.2:

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) &= \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) \\ &= \mathbb{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x \end{aligned}$$

□

Definition 24.6 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:

$$\mathtt{supp}(()f_X) := \{x \in \mathcal{X} | f(x) > 0\} \tag{24.9}$$

Note: this is not a rigorous definition.

Theorem 24.1 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 24.2 Identically Distributed: From theorem 24.1 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 24.7 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 24.3:

Monotonically Increasing $$x \leqslant y \iff F_X(x) \leqslant F_X(y) \quad \forall x, y \in \mathbb{R} \tag{24.10}$$

Upper Limit $$\lim_{x \to \infty} F_X(x) = 1 \tag{24.11}$$

Lower Limit $$\lim_{x \to -\infty} F_X(x) = 0 \tag{24.12}$$

Definition 24.8 CDF of a discret rv X: Let X be discret rv with pdf p_X, then the CDF of X is given by:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{t=-\infty}^{x} \mathrm{p}_X(t)$$

Definition 24.9 CDF of a continuous rv X: Let X be continuous rv with pdf f_X, then the CDF of X is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \iff \frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Lemma 24.1 Probability Interval: Let X be a continuous rrv with pdf f_X and cumulative distribution function F_X, then it holds that:

$$\mathbb{P}(a \leqslant X \leqslant b) = F_X(b) - F_X(a) \tag{24.13}$$

Proof. [def. 24.9]:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x)) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

□

Proof. lemma 24.1:

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a)$$

or by the fundamental theorem of calculus (theorem 14.2):

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(t)\,\mathrm{d}t = \int_a^b \frac{\partial F_X(t)}{\partial t}\,\mathrm{d}t = [F_X(t)]_a^b$$

□

Theorem 24.2 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 24.10 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} \mathrm{p}_x(\bar{\boldsymbol{x}}) \tag{24.14}$$

Definition 24.11 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} f_x(\bar{\boldsymbol{x}})\,\mathrm{d}\bar{\boldsymbol{x}} \tag{24.15}$$

Law 24.1 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{24.16}$$

Property 24.4 Translation and scaling: If $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \tag{24.17}$$

Thus $\mathbb{E}$ is a linear operator ([def. 14.13]).

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 24.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{24.18}$$

Property 24.5 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\boldsymbol{AXB}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{XB})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\,\boldsymbol{B} \tag{24.19}$$

Proof. eq. (24.24):

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{p}_{X,Y}(x, y)xy \\ &\overset{??}{=} \sum_{x \in \mathcal{X}} \mathrm{p}_X(x)x \sum_{y \in \mathcal{Y}} \mathrm{p}_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned}$$

□

Definition 24.12 Autocorrelation/Crosscorelation $\gamma(t_1, t_2)$: Describes the covariance ([def. 24.16]) between the two values of a stochastic process $(\boldsymbol{X}_t)_{t \in T}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \mathrm{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] = \mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})] \tag{24.20}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \mathrm{Cov}[\boldsymbol{X}_t, \boldsymbol{X}_t] \overset{\text{eq. (24.35)}}{=} \mathbb{V}[\boldsymbol{X}_t] \tag{24.21}$$

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 $\boldsymbol{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $\boldsymbol{x}(t - \tau)$ and the original function $\boldsymbol{x}(t)$ are.
- If there exists some relation between the values of the time series that is non-random then the autocorrelation is non-zero.
- The autocorrelation is maximized/most similar for no translation $\tau = 0$ at all.

2. Key Figures

2.1. The Expectation

Definition 24.13 Expectation (disc. case):

$$\mu_X := \mathbb{E}_x[x] := \sum_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} p_x(\bar{\boldsymbol{x}}) \tag{24.22}$$

Definition 24.14 Expectation (cont. case):

$$\mathbb{E}_x[x] := \int_{\bar{\boldsymbol{x}} \in \mathcal{X}} \bar{\boldsymbol{x}} f_x(\bar{\boldsymbol{x}}) \, \mathrm{d}\bar{\boldsymbol{x}} \tag{24.23}$$

Law 24.2 Expectation of independent variables:

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{24.24}$$

Property 24.6 Translation and scaling: If $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ are random vectors, and $a, b, a \in \mathbb{R}^n$ are constants then it holds:

$$\mathbb{E}[a + b\boldsymbol{X} + c\boldsymbol{Y}] = a + b\mathbb{E}[\boldsymbol{X}] + c\mathbb{E}[\boldsymbol{Y}] \tag{24.25}$$

Thus $\mathbb{E}$ is a linear operator[def. 14.13].

Property 24.7
Affine Transformation of the Expectation:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{X} + b] = \boldsymbol{A}\mu + b \tag{24.26}$$

Note: Expectation of the expectation

The expectation of a r.v. X is a constant hence with Property 24.6 it follows:

$$\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \tag{24.27}$$

Property 24.8 Matrix×Expectation: If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ are constant matrices then it holds:

$$\mathbb{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{X}\boldsymbol{B})] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\,\boldsymbol{B} \tag{24.28}$$

Proof. eq. (24.24):

$$\mathbb{E}[XY] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y)xy$$
$$\overset{??}{=} \sum_{x \in \mathcal{X}} p_X(x)x \sum_{y \in \mathcal{Y}} p_Y(y)y = \mathbb{E}[X]\,\mathbb{E}[Y]$$

□

2.2. The Variance

Definition 24.15 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:

$$\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \overset{\text{see section 3}}{=} \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \tag{24.29}$$

2.2.1. Properties

Property 24.9 Variance of a Constant: If $a \in \mathbb{R}$ is a constant then it follows that its expected value is deterministic $\Rightarrow$ we have no uncertainty $\Rightarrow$ no variance:

$$\mathbb{V}[a] = 0 \quad \textbf{with} \quad a \in \mathbb{R} \tag{24.30}$$

see shift and scaling for proof section 3

Property 24.10 Shifting and Scaling:

$$\mathbb{V}[a + bX] = a^2\sigma^2 \quad \textbf{with} \quad a \in \mathbb{R} \tag{24.31}$$

see section 3

Property 24.11
Affine Transformation of the Variance:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathbb{V}[\boldsymbol{A}\boldsymbol{X} + b] = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\,\boldsymbol{A}^\intercal \tag{24.32}$$

see section 3.

Definition 24.16 Covariance: The Covariance is a measure of how much two or more random variables vary linearly with each other.

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned} \tag{24.33}$$

see section 3

Definition 24.17 Covariance Matrix: The variance of a k-dimensional random vector $\boldsymbol{X} = (X_1 \ \dots \ X_k)$ is given by a p.s.d. eq. (17.20) 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:

$$\begin{aligned}\mathbb{V}[\boldsymbol{X}] :=& \Sigma(\boldsymbol{X}) := \mathrm{Cov}[\boldsymbol{X},\boldsymbol{X}] := \\ =& \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\intercal\right] \\ =& \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^\intercal\right] - \mathbb{E}[\boldsymbol{X}]\,\mathbb{E}[\boldsymbol{X}]^\intercal \in [-\infty, \infty]\end{aligned} \tag{24.34}$$

$$= \begin{bmatrix} \mathbb{V}[X_1] & \cdots & \mathrm{Cov}[X_1, X_k] \\ \vdots & \ddots & \\ \mathrm{Cov}[X_k, X_1] & \cdots & \mathbb{V}[X_k] \end{bmatrix}$$
$$= \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_k - \mu_k)] \\ \vdots & \ddots & \\ \mathbb{E}[(X_k - \mu_k)(X_1 - \mu_1)] & \cdots & \mathbb{E}[(X_k - \mu_k)(X_k - \mu_k)] \end{bmatrix}$$

Note: Covariance and Variance

The variance is a special case of the covariance in which two variables are identical:

$$\mathrm{Cov}[X,X] = \mathbb{V}[X] \equiv \sigma^2(X) \equiv \sigma_X^2 \tag{24.35}$$

add http://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/

Property 24.12 Translation and Scaling:

$$\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X,Y) \tag{24.36}$$

Property 24.13
Affine Transformation of the Covariance:
If $\boldsymbol{X} \in \mathbb{R}^n$ is a randomn vector, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ a constant matrix and $b \in \mathbb{R}^n$ then it holds:

$$\mathrm{Cov}[\boldsymbol{A}\boldsymbol{X} + b] = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\,\boldsymbol{A}^\intercal = \boldsymbol{A}\Sigma(\boldsymbol{X})\boldsymbol{A}^\intercal \tag{24.37}$$

Definition 24.18 Correlation Coefficient: Is the standardized version of the covariance:

$$\begin{aligned}\mathrm{Corr}[\boldsymbol{X}] :=& \frac{\mathrm{Cov}[\boldsymbol{X}]}{\sigma_{X_1} \cdots \sigma_{X_k}} \in [-1,1] \\ =& \begin{cases} +1 & \text{if } Y = aX + b \text{ with } a > 0, b \in \mathbb{R} \\ -1 & \text{if } Y = aX + b \text{ with } a < 0, b \in \mathbb{R} \end{cases}\end{aligned} \tag{24.38}$$

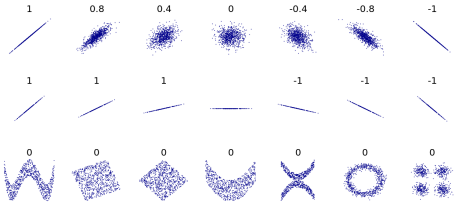


Figure 7: Several sets of (x, y) points, with their correlation coefficient

Law 24.3 Translation and Scaling:

$$\mathrm{Corr}(a + bX, c + dY) = \mathrm{sign}(b)\mathrm{sign}(d)\mathrm{Cov}(X,Y) \tag{24.39}$$

Note

- The correlation/covariance reflects the noisiness and direction of a linear relationship (top row fig. 7), **but** not the slope of that relationship (middle row fig. 7) nor many aspects of nonlinear relationships (bottom row)
- The set in the center of fig. 7 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 24.3.
2. Correlation is dimensionless, whereas the unit of the covariance is obtained by the product of the units of the two RV variables.

Law 24.4 Covariance of independent RVs: The covariance/correlation of two independent variable's (**??**) is zero:

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$
$$\overset{\text{eq. (24.24)}}{=} \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

Zero covariance/correlation⇏ 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:

$$\begin{aligned}\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,X^2) &= \mathbb{E}\left[X \cdot X^2\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \\ &= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}\left[X^2\right] \overset{\substack{\text{eq. (24.61)} \\ \text{eq. (24.50)}}}{=} 0 - 0 \cdot \mathbb{E}\left[X^2\right]\end{aligned}$$

$\Rightarrow$ the relationship between Y and X must be non-linear.

Definition 24.19 Quantile: Are specific values q_α in the range[def. 14.8] of a random variable X that are defined as the value for which the cumulative probability is less then $\alpha \in (0,1)$:

$$q_\alpha : \mathbb{P}(X \leqslant x) = F_X(q_\alpha) = \alpha \xrightarrow{F \text{ invert.}} q_\alpha = F_X^{-1}(\alpha) \tag{24.40}$$

add figure

3. Proofs

Proof. eq. (24.29)

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right]$$
$$\overset{\text{Property 24.6}}{=} \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mu^2$$

□

Proof. Property 24.10

$$\begin{aligned}\mathbb{V}[a + bX] &= \mathbb{E}\left[(a + bX - \mathbb{E}[a + bX])^2\right] \\ &= \mathbb{E}\left[(\cancel{a} + bX - \cancel{a} - b\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[(bX - b\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[b^2(X - \mathbb{E}[X])^2\right] \\ &= b^2\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = b^2\sigma^2\end{aligned}$$

□

Proof. Property 24.11

$$\begin{aligned}\mathbb{V}(\boldsymbol{A}\boldsymbol{X} + b) &= \mathbb{E}\left[(\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{A}\boldsymbol{X}])^2\right] + 0 = \\ &= \mathbb{E}\left[(\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{A}\boldsymbol{X}])(\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{A}\boldsymbol{X}])^\intercal\right] \\ &= \mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{A}(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}]))^\intercal\right] \\ &= \mathbb{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal\boldsymbol{A}^\intercal\right] \\ &= \boldsymbol{A}\mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - (\mathbb{E}[\boldsymbol{X}])^\intercal\right]\boldsymbol{A}^\intercal = \boldsymbol{A}\mathbb{V}[\boldsymbol{X}]\,\boldsymbol{A}^\intercal\end{aligned}$$

□

Proof. eq. (24.33)

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\,\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[X]\,\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

□

Discrete Distributions

Definition 24.20 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 — $\mathrm{Bern}(p)$

Definition 24.21 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 24.22 Bernoulli Distribution $X \sim \mathrm{Bern}(p)$: 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:

$$p(x;p) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases} \iff \begin{cases} \mathbb{P}(X=1) = p \\ \mathbb{P}(X=0) = 1-p \end{cases}$$
$$= p^x \cdot (1-p)^{1-x} \quad \text{for } x \in \{0,1\}$$

$$\mathbb{E}[X] = p \quad (24.41) \qquad \mathbb{V}[X] = p(1-p) \quad (24.42)$$

4.2. Binomial Distribution — $\mathcal{B}(n,p)$

Definition 24.23 Binomial Distribution: Models the probability of exactly X success given a fixed number n-*Bernoulli experiments*[def. 24.21], where the probability of success of a single experiment is given by p:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

n :nb. of repetitions
x :nb. of successes
p :probability of success

$$\mathbb{E}[X] = np \quad (24.43) \qquad \mathbb{V}[X] = np(1-p) \quad (24.44)$$

Note: Binomial Coefficient

The Binomial Coefficient corresponds to the permutation of two classes and not the variations as it seems from the formula.
Lets consider a box of n balls consisting of black and white balls. If we want to know the probability of drawing first x white and then $n-x$ black balls we can simply calculate:

$$\underbrace{(p \cdots p)}_{\text{x-times}} \cdot \underbrace{(q \cdots q)}_{n-x\text{-times}} = p^x q^{n-x}$$

But there exists obviously further realization $X = x$, that correspond to permutations of the n-drawn balls.
There exist two classes of $n_1 = x$-white and $n_2 = (n-x)$ black balls s.t.

$$P(n; n_1, n_2) = \frac{n!}{x!(n-x)!} = \binom{n}{x}$$

4.3. Geometric Distribution — $\mathrm{Geom}(p)$

Definition 24.24 Geometric Distribution$\mathrm{Geom}(p)$: Models the probability of the number X of Bernoulli trials[def. 24.21] *until the first success*

$$p(x) = p(1-p)^{x-1}$$

x :nb. of repetitions *until first success*
p :success probability *of single Bernoulli experiment*

$$F(x) = \sum_{i=1}^{x} p(1-p)^{i-1} \overset{??}{=} 1-(1-p)^x$$

$$\mathbb{E}[X] = \frac{1}{p} \quad (24.45) \qquad \mathbb{V}[X] = \frac{1-p}{p^2} \quad (24.46)$$

Notes

- $\mathbb{E}[X]$ is the mean waiting time until the first success
- the number of trials x in order to have at least one success with a probability of $p(x)$:

$$x \geqslant \frac{p(x)}{1-p}$$

- $\log(1-p) \approx -p$ for small p

4.4. Poisson Distribution — $\mathrm{Pois}(\lambda)$

Definition 24.25 Poisson Distribution: Is an extension of the binomial distribution, where the realization x of the random variable X may attain values in $\mathbb{Z}_{\geqslant 0}$.
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.

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \lambda > 0,\ x \in \mathbb{Z}_{\geqslant 0} \quad (24.47)$$

Event Rate λ: describes the average number of events in a single interval.

$$\mathbb{E}[X] = \lambda \quad (24.48) \qquad \mathbb{V}[X] = \lambda \quad (24.49)$$

Continuous Distributions

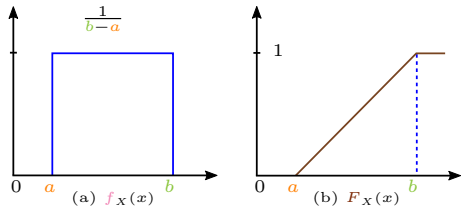
5.1. Uniform Distribution — $\mathcal{U}(a,b)$

Definition 24.26 Uniform Distribution $\mathcal{U}(a,b)$: Is probability distribution, where all intervals of the **same** length on the distribution's support ([def. 24.6]) $\mathrm{supp}(\mathcal{U}[a,b]) = [a,b]$ are equally probable/likely.

$$f(x) = \frac{1}{b-a}\mathbb{1}_{x\in[a;b)} = \begin{cases} \frac{1}{b-a} = \text{const} & \text{if } a \leqslant x \leqslant b \\ 0 & \text{else} \end{cases} \quad (24.50)$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & \text{if } a \leqslant x \leqslant b \\ 1 & x > b \end{cases} \quad (24.51)$$

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{V}(X) = \frac{(b-a)^2}{12} \quad (24.52)$$



(a) $f_X(x)$ (b) $F_X(x)$

5.2. Exponential Distribution — $\exp(\lambda)$

Definition 24.27 Exponential Distribution $X \sim \exp(\lambda)$: Is the continuous analogue to the geometric distribution[def. 24.24].

It describes the probability $f(x;\lambda)$ that a continuous Poisson process (i.e., a process in which events occur continuously and independently at a constant average rate) will succeed/change state after a time interval x.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & x < 0 \end{cases} \quad (24.53)$$

$$F(x;\lambda) = \begin{cases} 1-e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & x < 0 \end{cases} \quad (24.54)$$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathbb{V}(X) = \frac{1}{\lambda^2} \quad (24.55)$$

5.3. Laplace Distribution

Definition 24.28 Laplace Distribution:

Laplace Distibution $f(\boldsymbol{x};\mu,\sigma) = \frac{1}{2\sigma}\exp\left(-\frac{|\boldsymbol{x}-\mu|}{\sigma}\right)$ (24.56)

5.4. The Normal Distribution — $\mathcal{N}(\mu,\sigma)$

Definition 24.29 Normal Distribution $\boldsymbol{X} \sim \mathcal{N}(\mu,\sigma^2)$: Is a symmetric distribution where the population parameters μ, σ^2 are equal to the expectation and variance of the distribution:

$$\mathbb{E}[X] = \mu \qquad \mathbb{V}(X) = \sigma^2 \quad (24.57)$$

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad (24.58)$$

$$F(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left\{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right\}\mathrm{d}u \quad (24.59)$$

$x \in \mathbb{R}$ or $-\infty < x < \infty$

$$\varphi_X(u) = \exp\left\{iu\mu - \frac{u^2\sigma^2}{2}\right\} \quad (24.60)$$

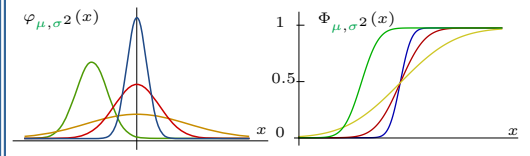


Figure 9: $\mu = 0$, $\sigma^2 = 0.2$ | $\mu = 0$, $\sigma^2 = 1.0$ | $\mu = 0$, $\sigma^2 = 5.0$ | $\mu = -2$, $\sigma^2 = 0.5$

Property 24.14: $\mathbb{P}_X(\mu-\sigma \leqslant x \leqslant \mu-\sigma) = 0.66$

Property 24.15: $\mathbb{P}_X(\mu-2\sigma \leqslant x \leqslant \mu-2\sigma) = 0.95$

5.5. The Standard Normal distribution — $\mathcal{N}(0,1)$

Historic Problem: the cumulative distribution eq. (24.59) does not have an analytical solution and numerical integration was not always computationally so easy. So how should people calculate the probability of x falling into certain ranges $\mathbb{P}(x \in [a,b])$?
Solution: use a standardized form/set of parameters (by convention) $\mathcal{N}_{0,1}$ and tabulate many different values for its cumulative distribution $\phi(x)$ s.t. we can transform all families of Normal Distributions into the standardized version $\mathcal{N}(\mu,\sigma^2) \xrightarrow{z} \mathcal{N}(0,1)$ and look up the value in its table.

Definition 24.30
Standard Normal Distribution $\boldsymbol{X} \sim \mathcal{N}(0,1)$:

$$\mathbb{E}[X] = 0 \qquad \mathbb{V}(X) = 1 \quad (24.61)$$

$$f(x;0,1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \quad (24.62)$$

$$F(x;0,1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}u^2}\,\mathrm{d}u \quad (24.63)$$

$x \in \mathbb{R}$ or $-\infty < x < \infty$

$$\psi_X(u) = e^{\frac{u^2}{2}} \qquad \varphi_X(u) = e^{-\frac{u^2}{2}} \quad (24.64)$$

Corollary 24.3
Standard Normal Distribution Notation: As the standard normal distribution is so commonly used people often use the letter Z in order to denote its the *standard* normal distribution and its α-quantile[def. 24.19] is then denoted by:

$$z_\alpha = \Phi^{-1}(\alpha) \qquad \alpha \in (0,1) \quad (24.65)$$

5.5.1. Calculating Probabilities

Property 24.16 Symmetry: Let $z > 0$

$$\mathbb{P}(Z \leqslant z) = \Phi(z) \quad (24.66)$$
$$\mathbb{P}(Z \leqslant -z) = \Phi(-z) = 1-\Phi(z) \quad (24.67)$$
$$\mathbb{P}(-a \leqslant Z \leqslant b) = \Phi(b) - \Phi(-a) \overset{a=b=z}{=} \Phi(b) - (1-\Phi(a)) = 2\Phi(z) - 1 \quad (24.68)$$

5.5.2. Linear Transformations of Normal Dist.

Proposition 24.1 **Linear Transformation** proof 1:
Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the linear transformed r.v. Y given by the *affine transformation* $Y = a + bX$ with $a \in \mathbb{R}, b \in \mathbb{R}_+$ follows:

$$Y \sim \mathcal{N}\left(a + b\mu, b^2\sigma^2\right) \iff f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) \tag{24.69}$$

Proposition 24.2 **Standardization:** Let X be a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then there exists a linear transformation $Z = a + bX$ s.t. Z is a standard normally distributed random variable:

$$X \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{Z=\frac{X-\mu}{\sigma}} Z \sim \mathcal{N}(0,1) \tag{24.70}$$

section 1

Note

If we know how many standard deviations our distribution is away from our target value then we can characterize it fully by the standard normal distribution.

Proposition 24.3 **Standardization of the CDF:** Let $F_X(X)$ be the cumulative distribution function of a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, then the cumulative distribution function $\Phi_Z(z)$ of the standardized random normal variable $Z \sim \mathcal{N}(0,1)$ is related to $F_X(X)$ by:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{24.71}$$

section 1

6. The Multivariate Normal distribution

Definition 24.31 **Multivariate Normal/Gaussian:**
An $\mathbb{R}^n$-valued random variable $\boldsymbol{X} = (X_1 \quad \dots, X_n)$ is *Multivariate Gaussian/Normal* if every linear combination of its components is a (one-dimensional) Gaussian:

$$\exists \mu, \sigma: \quad \mathscr{L}\left(\sum_{i=1}^{n} \alpha_i X_j\right) = \mathcal{N}(\mu, \sigma^2) \quad \forall \alpha_i \in \mathbb{R} \tag{24.72}$$

(possible degenerated $\mathcal{N}(0,0)$ for $\forall \alpha_j = 0$)

Note

- **Joint** vs. **multivariate**: a joint normal distribution can be a multivariate normal distribution or a product of univariate normal distributions **but**
- Multivariate refers to the number of variables that are placed as inputs to a function.

Definition 24.32
Multivariate Normal distribution $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$:
A k-dimensional random vector

$$\boldsymbol{X} = (X_1 \quad \dots \quad X_n)^\intercal \quad \text{with} \quad \boldsymbol{\mu} = (\mathbb{E}[\boldsymbol{x}_1] \quad \dots \quad \mathbb{E}[\boldsymbol{x}_k])^\intercal$$

and $k \times k$ **p.s.d.**covariance matrix:

$$\Sigma := \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\intercal] = [\mathrm{Cov}[\boldsymbol{x}_i, \boldsymbol{x}_j], 1 \leqslant i, j \leqslant k]$$

follows a k-dim multivariate normal/Gaussian distribution if its law[def. 22.23] satisfies:

$$f_{\boldsymbol{X}}(X_1, \dots, X_k) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{24.73}$$

$$= \underbrace{\frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}}}_{\text{Normalisation}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})\right)$$

$$\varphi_{\boldsymbol{X}}(\boldsymbol{u}) = \exp\left\{i\boldsymbol{u}^\intercal \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{u}\Sigma\boldsymbol{u}\right\} \tag{24.74}$$

6.1. Joint Gaussian Distributions

Definition 24.33 **Jointly Gaussian Random Variables:**
Two random variables X, Y both scalars or vectors, are said to be jointly Gaussian if the joint vector random variable $\boldsymbol{Z} = [X \quad Y]^\intercal$ is again a GRV.

Property 24.17 proof 1
Joint Independent Gaussian Random Variables: Let $X_1, \dots, X_n$ be $\mathbb{R}$-valued *independent* random variables with laws $\mathcal{N}\left(\mu_i, \sigma_i^2\right)$. Then the law of $\boldsymbol{X} = (X_1 \quad \dots \quad X_n)$ is a (multivariate) Gaussian distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{24.75}$$

Corollary 24.4 Quadratic Form:
If $\boldsymbol{x}$ and $\boldsymbol{y}$ are both independent GRVs

$$\boldsymbol{x} \sim \mathcal{N}(\mu_x, \Sigma_x) \qquad \boldsymbol{y} \sim \mathcal{N}(\mu_y, \Sigma_y)$$

then they are jointly Gaussian[def. 24.33] given by:

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{p}(x)\mathrm{p}(y) \tag{24.76}$$

$$\propto \exp\left(-\frac{1}{2}\left\{(\boldsymbol{x} - \mu_x)^\intercal \Sigma_x^{-1}(\boldsymbol{x} - \mu_x) + (\boldsymbol{y} - \mu_y)^\intercal \Sigma_y^{-1}(\boldsymbol{y} - \mu_y)\right\}\right)$$

$$= \exp\left(-\frac{1}{2}[(\boldsymbol{x} - \mu_x)^\intercal \quad (\boldsymbol{y} - \mu_y)^\intercal]\begin{bmatrix}\Sigma_x^{-1} & 0 \\ 0 & \Sigma_y^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{x} - \mu_x \\ \boldsymbol{y} - \mu_y\end{bmatrix}\right)$$

$$\triangleq \exp -\frac{1}{2}(\boldsymbol{z} - \mu_z)^\intercal \Sigma_z^{-1}(\boldsymbol{z} - \mu_z)$$

Property 24.18
Marginal Distribution of Multivariate Gaussian: Let $\boldsymbol{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a an $\mathbb{R}^n$ valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. The k-variate marginal distribution of the Gaussian indexed by a subset of the variables:

$$A = \{i_1, \dots, i_k\} \qquad i_j \in V \tag{24.77}$$

is given by:

$$\boldsymbol{X} = \begin{pmatrix} X_{i_1} & \dots & X_{i_k} \end{pmatrix}^\intercal \sim \mathcal{N}(\mu_A, \Sigma_{AA}) \tag{24.78}$$

$$\Sigma = \begin{bmatrix} \sigma^2_{i_1,i_1} & \cdots & \sigma^2_{i_1,i_k} \\ \vdots & \ddots & \vdots \\ \sigma^2_{i_k,i_1} & \cdots & \sigma^2_{i_k,i_k} \end{bmatrix} \quad \text{and} \quad \mu = \begin{bmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \vdots \\ \mu_{i_k} \end{bmatrix}$$

6.2. Conditional Gaussian Distributions

Property 24.19 **Conditional Gaussian Distribution:** Let $\boldsymbol{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a an $\mathbb{R}^n$ valued Gaussian and let $V = \{1, 2, \dots, n\}$ be the index set of its variables. Suppose we take two disjoint subsets of V:

$$A = \{i_1, \dots, i_k\} \qquad B = \{j_1, \dots, j_m\} \qquad i_l, j_{l'} \in V$$

then the conditional distribution of the random vector $\boldsymbol{X}_A$, conditioned on $\boldsymbol{X}_B$ given by $\mathrm{p}(\boldsymbol{X}_A | \boldsymbol{X}_B = \boldsymbol{x}_B)$ is:

$$\boldsymbol{X}_A = \begin{pmatrix} X_{i_1} & \dots & X_{i_k} \end{pmatrix}^\intercal \sim \mathcal{N}\left(\mu_{A|B}, \Sigma_{A|B}\right) \tag{24.79}$$

$$\boxed{\begin{aligned} \mu_{A|B} &= \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\boldsymbol{x}_B - \mu_B) \\ \Sigma_{A|B} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \end{aligned}}$$

Note

Can be proofed using the matrix inversion lemma but is a very tedious computation.

maybe add sometime

Corollary 24.5
Conditional Distribution of Joint Gaussian's: Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly Gaussian random vectors:

$$\begin{bmatrix}\boldsymbol{X} \\ \boldsymbol{Y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^\intercal & \boldsymbol{B}\end{bmatrix}\right) \tag{24.80}$$

then the *marginal* distribution of $\boldsymbol{x}$ conditioned on $\boldsymbol{y}$ can be written as: $X \sim \mathcal{N}\left(\mu_{X|Y}, \Sigma_{X|Y}\right)$

$$\boxed{\begin{aligned} \mu_{X|Y} &= \mu_X + \boldsymbol{C}\boldsymbol{B}^{-1}(\boldsymbol{y} - \mu_Y) \\ \Sigma_{X|Y} &= \boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\intercal \end{aligned}} \tag{24.81}$$

add proofs

6.3. Transformations

Property 24.20 **Multiples of Gaussian's** AX:
Let $\boldsymbol{X} = (X_1 \quad \dots \quad X_n)^\intercal \sim \mathcal{N}(\mu, \Sigma)$ be a an $\mathbb{R}^n$ valued Gaussian and let $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ then it follows:

$$Y = \boldsymbol{A}X \in \mathbb{R} \qquad Y \sim \mathcal{N}\left(\boldsymbol{A}\mu, \boldsymbol{A}\Sigma\boldsymbol{A}^\intercal\right) \tag{24.82}$$

Property 24.21 **Affine Transformation of GRVs:** **Let** $\boldsymbol{y} \in \mathbb{R}^n$ be GRV, $\boldsymbol{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$ and let $\boldsymbol{x}$ be defined by the affine transformation[def. 17.1]:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + b \qquad \boldsymbol{A} \in \mathbb{R}^{d \times n}, b \in \mathbb{R}^d$$

Then $\boldsymbol{x}$ is a GRV (see Section 1).

Property 24.22 **Linear Combination of jointly GRVs:** **Let** $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$ two jointly GRVs, and let $\boldsymbol{z}$ be defined as:

$$\boldsymbol{z} = \boldsymbol{A}_x\boldsymbol{x} + \boldsymbol{A}_y\boldsymbol{y} \qquad \boldsymbol{A}_x \in \mathbb{R}^{d \times n}, \boldsymbol{A}_x \in \mathbb{R}^{d \times m}$$

Then $\boldsymbol{z}$ is GRV (see Section 1).

Definition 24.34 **Gaussian Noise:** Is statistical noise having a probability density function (PDF) equal to that of the normal/Gaussian distribution.

6.4. Gamma Distribution $\Gamma(x, \alpha, \beta)$

Definition 24.35 **Gamma Distribution** $X \sim \Gamma(x, \alpha, \beta)$:
Is a widely used distribution that is related to the exponential distribution, Erlang distribution, and chi-squared distribution as well as Normal distribution:

$$\mathrm{f}(x; \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x} & \text{if} \quad x > 0 \\ 0 & \quad\quad x \leqslant 0 \end{cases} \tag{24.83}$$

$$\Gamma(\alpha) \overset{\text{eq. (14.66)}}{=} \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{24.84}$$

with $\alpha, \beta \in \mathbb{R}_{>0}$

7. Student's t-distribution

Definition 24.36 **Student' t-distribution:**

add

7.1. Delta Distribution

Definition 24.37 **The delta function** $\delta(\boldsymbol{x})$:
The delta/dirac function $\delta(\boldsymbol{x})$ is defined by:

$$\int_{\mathbb{R}} \delta(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = f(0)$$

for any integrable function f on $\mathbb{R}$.
Or alternativly by:

$$\delta(x - x_0) = \lim_{\sigma \to 0} \mathcal{N}(x | x_0, \sigma) \tag{24.85}$$

$$\approx \infty \mathbb{1}_{\{x = x_0\}} \tag{24.86}$$

Property 24.23 **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**:
$$\int_{\mathbb{R}} \delta(x - x_0) f(x)\,\mathrm{d}x = f(x_0) \tag{24.87}$$
- **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. (24.85) would be a non-differentiable/discret form of the dirac measure.

Proofs

Proof. proposition 24.1: 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. (24.69)).
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. proposition 24.2: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$Z := \frac{X - \mu}{\sigma} = \frac{1}{std}X - \frac{\mu}{\sigma} = aX + b \quad \text{with } a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$$
$$\overset{\text{eq. (24.69)}}{\sim} \mathcal{N}\left(a\mu + b, a^2\sigma^2\right) \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}\right) \sim \mathcal{N}(0,1)$$

□

Proof. proposition 24.3: Let X be normally distributed with $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$F_X(x) = \mathbb{P}(X \leqslant x) \overset{\div\sigma}{\overset{-\mu}{=}} \mathbb{P}\left(\frac{X - \mu}{\sigma} \leqslant \frac{x - \mu}{\sigma}\right) \mathbb{P}\left(Z \leqslant \frac{x - \mu}{\sigma}\right)$$
$$= \Phi\left(\frac{x - \mu}{\sigma}\right)$$

□

Proof. Property 24.21 scalar case
Let $y \sim \mathrm{p}(y) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$ and define $\boldsymbol{x} = ay + b \qquad a \in \mathbb{R}_+, b \in \mathbb{R}$
Using the Change of variables formula it follows:

$$\mathrm{p}_x(\bar{x}) \overset{??}{=} \frac{\mathrm{p}_y(\bar{y})}{\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|} \overset{\bar{y} = \frac{\bar{x}-b}{a}}{=} \frac{1}{a}\frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\overbrace{\frac{\bar{x}-b}{a}}^{\bar{y}(\bar{x})} - \mu\right)^2\right)$$
$$= \frac{1}{\sqrt{2\pi a^2\mu^2}} \exp\left(-\frac{1}{2\sigma^2 a^2}\left(\bar{x} \underbrace{-b - a\mu}_{\mu_x}\right)^2\right)$$

Hence $x \sim \mathcal{N}(\mu_x, \sigma_x^2) = \mathcal{N}(a\mu + b, a^2\sigma^2)$

□

Note

We can also verify that we have calculated the right mean and variance by:

$$\mathbb{E}[x] = \mathbb{E}[ay + b] = a\mathbb{E}[y] + b = a\mu + b$$
$$\mathbb{V}[x] = \mathbb{V}[ay + b] = a^2\mathbb{V}[y] = a^2\sigma^2$$

Proof. **??**

$$\mathrm{p}_{\boldsymbol{X}}(\boldsymbol{u}) = \prod_{i}^{n} \mathrm{p}_{X_i}(u_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\varphi_{\boldsymbol{X}}(\boldsymbol{u}) = \exp\left\{iu_1\mu_1 - \frac{1}{2}\sigma_1 u_1^2\right\} \cdots \exp\left\{iu_n\mu_n - \frac{1}{2}\sigma_n u_n^2\right\}$$

$$= \exp\left\{ i\sum_{i}^{n} u_n\mu_n - \frac{1}{2}\underbrace{\sum_{i}^{n} \sigma_n u_n^2}_{\cdots\cdots\cdots} \right\} = \exp\left\{ i\boldsymbol{u}^\intercal\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{u}\underbrace{\Sigma}_{\cdots}\boldsymbol{u} \right\}$$

□

Proof. Property 24.22

From Property 24.21 it follows immediately that $\boldsymbol{z}$ is GRV $\boldsymbol{z} \sim \mathcal{N}(\mu_z, \Sigma_z)$ with:

$$\boldsymbol{z} = \boldsymbol{A}\xi \qquad \textbf{with} \qquad \boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix} \text{ and } \xi = \begin{pmatrix}\boldsymbol{x} & \boldsymbol{y}\end{pmatrix}$$

Knowing that $\boldsymbol{z}$ is a GRV it is sufficient to calculate μ_z and Σ_z in order to characterize its distribution:

$$\mathbb{E}[\boldsymbol{z}] = \mathbb{E}\left[\boldsymbol{A}_x x + \boldsymbol{A}_y y\right] = \boldsymbol{A}_x\mu_x + \boldsymbol{A}_y\mu_y$$

$$\mathbb{V}[\boldsymbol{z}] = \mathbb{V}[\boldsymbol{A}\xi] \overset{??}{=} \boldsymbol{A}\mathbb{V}[\xi]\,\boldsymbol{A}^\intercal$$

$$= \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}^\intercal$$

$$= \begin{bmatrix}\boldsymbol{A}_x & \boldsymbol{A}_y\end{bmatrix}\begin{bmatrix}\mathbb{V}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathbb{V}[y]\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_x^\intercal \\ \boldsymbol{A}_y^\intercal\end{bmatrix}$$

$$= \boldsymbol{A}_x\mathbb{V}[x]\,\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\mathbb{V}[y]\,\boldsymbol{A}_y^\intercal + \underbrace{\boldsymbol{A}_y\mathrm{Cov}[y,x]\,\boldsymbol{A}_x^\intercal}_{=0\text{by independence}} + \underbrace{\boldsymbol{A}_x\mathrm{Cov}[x,y]\,\boldsymbol{A}_y^\intercal}_{=0\text{by independence}}$$

$$= \boldsymbol{A}_x\Sigma_x\boldsymbol{A}_x^\intercal + \boldsymbol{A}_y\Sigma_y\boldsymbol{A}_y^\intercal$$

□

Note

Can also be proofed by using the normal definition of [def. 24.15] and tedious computations.

8. Sampling Random Numbers

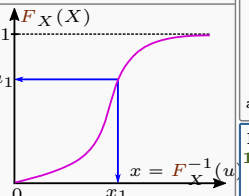
Most math libraries have uniform **random number generator** (**RNG**) i.e. functions to generate uniformly distributed random numbers $U \sim \mathcal{U}[a,b]$ (eq. (24.50)).
Furthermore repeated calls to these RNG are independent, that is:

$$\begin{aligned} p_{U_1,U_2}(u_1,u_2) &\overset{??}{=} p_{U_1}(u_1)\cdot p_{U_2}(u_2) \\ &= \begin{cases} 1 & \text{if } u_1,u_2 \in [a,b] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Question: using samples $\{u_1,\dots,u_n\}$ of these CRVs with uniform distribution, how can we create random numbers with arbitrary discreet or continuous PDFs?

9. Inverse-transform Technique

Idea



Can make use of section 1 and the fact that CDF are increasing functions ([def. 14.10]). **Advantage**:
- Simple to implement
- All discrete distributions can be generated via inverse- transform technique

Drawback:
- Not all continuous distributions can be integrated/have closed form solution for their CDF.
  E.g. Normal-,Gamma-,Beta-distribution.

9.1. Continuous Case

Definition 24.38 One Continuous Variable: Given: a desired continuous pdf f_X and uniformly distributed rn $\{u_1,u_2,\dots\}$:

1. Integrate the desired pdf f_X in order to obtain the desired cdf F_X:
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \tag{24.88}$$
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:
$$U = F_X(X) = F_X\Big(\underbrace{F_X^{-1}(U)}_{x}\Big) \tag{24.89}$$
4. Plug in the uniformly distributed rn:
$$x_i = F_X^{-1}(u_i) \qquad \textbf{s.t.} \qquad x_i \sim f_X \tag{24.90}$$

Definition 24.39 Multiple Continuous Variable:
Given: a pdf of multiple rvs $f_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) \tag{24.91}$$
2. Use [def. 24.40] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 24.40] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

Proof. [def. 24.40]:
Claim: if U is a uniform rv on $[0,1]$ then $F_X^{-1}(U)$ has F_X as its CDF.
Assume that F_X is strictly increasing ([def. 14.10]).
Then for any $u \in [0,1]$ there must exist a **unique** x s.t. $F_X(x) = u$.
Thus F_X must be invertible and we may write $x = F_X^{-1}(u)$.
Now let a arbitrary:

$$F_X(a) = \mathbb{P}(x \leqslant a) = \mathbb{P}(F_X^{-1}(U) \leqslant a)$$

Since F_X is strictly increasing:

$$\begin{aligned} \mathbb{P}\left(F_X^{-1}(U) \leqslant a\right) &= \mathbb{P}(U \leqslant F_X(a)) \\ &\overset{\text{eq. (24.50)}}{=} \int_0^{F_X(a)} 1\,\mathrm{d}t = F_X(a) \end{aligned}$$

□

Note

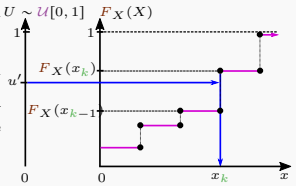
Strictly speaking we may not assume that a CDF is strictly increasing but we as all CDFs are weakly increasing ([def. 14.10]) we may always define an auxiliary function by its infinimum:

$$\hat{F}_X^{-1} := \inf\{x | F_X(X) \geqslant 0\} \qquad u \in [0,1] \tag{24.92}$$

9.2. Discrest Case

Idea

Given: a desired discrete 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.:



$$\sum_{i=1}^{k-1} < U \leqslant \sum_{i=1}^{k} \iff F_X(x_{k-1}) < u \leqslant F_X(x_k) \tag{24.93}$$

and return x_k.

Definition 24.40 One Discret Variable:

1. Compute the CDF of p_X ([def. 24.8])
$$F_X(x) = \sum_{t=-\infty}^{x} p_X(t) \tag{24.94}$$
2. Given the uniformly distributed rn $\{u_i\}_{i=1}^n$ find k^i ($\triangleq$ inversion) s.t.:
$$F_X\left(x_{k(i)-1}\right) < u_i \leqslant F_X\left(x_{k(i)}\right) \qquad \forall u_i \tag{24.95}$$

Proof. **??**: First of all notice that we can always solve for an unique x_k.

Ask: why, are Discret CRV always strictly increasing/unique?

Given a fixed x_k determine the values of u for which:

$$F_X(x_{k-1}) < u \leqslant F_X(x_k) \tag{24.96}$$

Now observe that:

$$\begin{aligned} u &\leqslant F_X(x_k) = F_X(x_{k-1}) + p_X(x_k) \\ \Rightarrow & F_X(x_{k-1}) < u \leqslant F_X(x_{k-1}) + p_X(x_k) \end{aligned}$$

The probability of U being in $(F_X(x_{k-1}), F_X(x_k)]$ is:

$$\begin{aligned} \mathbb{P}\left(U \in [F_X(x_{k-1}), F_X(x_k)]\right) &= \int_{F_X(x_{k-1})}^{F_X(x_k)} p_U(t)\,\mathrm{d}t \\ &= \int_{F_X(x_{k-1})}^{F_X(x_k)} 1\,\mathrm{d}t = \int_{F_X(x_{k-1})}^{F_X(x_{k-1})+p_X(x_k)} 1\,\mathrm{d}t = p_X(x_k) \end{aligned}$$

Hence the random variable $x_k \in \mathcal{X}$ has the pdf p_X. □

Definition 24.41
Multiple Continuous Variables (Option 1):
Given: a pdf of multiple rvs $p_{X,Y}$:

1. Use the product rule (**??**) in order to decompose $p_{X,Y}$:
$$p_{X,Y} = p_{X,Y}(x,y) = p_{X|Y}(x|y) p_Y(y) \tag{24.97}$$
2. Use **??** to first get a rv for y of $Y \sim p_Y(y)$.
3. Then with this fixed y use **??** again to get a value for x of $X \sim p_{X|Y}(x|y)$.

Definition 24.42
Multiple Continuous Variables (Option 2):
Note: this only works if $\mathcal{X}$ and $\mathcal{Y}$ are finite.
Given: a pdf of multiple rvs $p_{X,Y}$ **let** $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ the number of elements in $\mathcal{X}$ and $\mathcal{Y}$.

Define

$$\begin{aligned} & p_Z(1) = p_{X,Y}(1,1), p_Z(2) = p_{X,Y}(1,2), \dots \\ & \dots, p_Z(N_x \cdot N_y) = p_{X,Y}(N_x, N_y) \end{aligned}$$

Then simply apply **??** to the auxillary pdf p_Z

1. Use the product rule (**??**) in order to decompose $f_{X,Y}$:
$$f_{X,Y} = f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) \tag{24.98}$$
2. Use [def. 24.40] to first get a rv for y of $Y \sim f_Y(y)$.
3. Then with this fixed y use [def. 24.40] again to get a value for x of $X \sim f_{X|Y}(x|y)$.

nice examples see comment in code text

10. Descriptive Statistics

10.1. Population Parameters

Definition 24.43 Population/Statistical Parameter: Are parameters defining families of probability distributions and thus characteristics of population following such distributions i.e. the normal distribution has two parameters $\{\mu, \sigma^2\}$

Definition 24.44 Population Mean: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{24.99}$$

Definition 24.45 Population Variance: Given a population $\{x_i\}_{i=1}^N$ of size N its variance is defined as: $\{x_i\}_{i=1}^N$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{24.100}$$

Note

The population variance and mean are equally to the mean derived from the true distribution of the population.

10.2. Sample Estimates

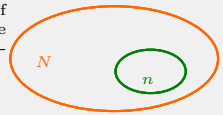
Definition 24.46 (Sample) Statistic: A statistc is a measuarble function f that assigns a **single** value F to a sample of random variables or population:

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

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

Note

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



Definition 24.47 (Point) Estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$:
Given: n-samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim \boldsymbol{X}$ an estimator

$$\hat{\theta} = h(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \tag{24.101}$$

is a statistic/randomn variable used to estimate a true (population) parameter θ[def. 24.43].

Note

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

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

Definition 24.48 Degrees of freedom of a Statistic: Is the number of values in the final calculation of a statistic that are free to vary.

10.2.1. Empirical Mean

Definition 24.49 Sample/Empirical Mean $\bar{x}$:
The sample mean is an estimate/statistic of the population mean[def. 24.44] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{24.102}$$

Corollary 24.6 Expectation: The sample mean estimator is unbiased (see section 14):

$$\mathbb{E}[\hat{\mu}_X] = \mu \tag{24.103}$$

Corollary 24.7 Variance: For the variance of the sample mean estimator it holds (see section 14):

$$\mathbb{V}[\hat{\mu}_X] = \frac{1}{n}\sigma_X^2 \tag{24.104}$$

10.2.2. Empirical Variance

Definition 24.50 Biased Sample Variance: The sample mean is an estimate/statistic of the population variance[def. 24.45] and can be calculated from an observation/sample of the total population $\{x_i\}_{i=1}^n \subset \{x_i\}_{i=1}^N$:

$$s_n^2 = \hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{24.105}$$

Definition 24.51 (Unbiased) Sample Variance:

$$s^2 = \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 \tag{24.106}$$

see section 14

Definition 24.52 Bessel's Correction: The factor

$$\frac{n}{n-1} \tag{24.107}$$

as multiplying the uncorrected population varianceeq. (24.105) by this term yields an unbiased estimated of the variance (not the standard deviation). The reason for this is that are

Attention: Usually only unbiased variance is used and also sometimes denoted by s_n^2

Proof.
finish this
□

11. Statistical Tests

Definition 24.53 Null Hypothesis: A Null Hypothesis H_0 is usually a commonly accepted fact/view/base hypothesis that researchers try to nullify or disprove.

$$H_0 : \theta = \theta_0 \tag{24.108}$$

Definition 24.54 Alternative Hypothesis: The Alternative Hypothesis H_A/H_1 is the opposite of the Null Hypotheses/contradicts it and is what we try to test against the Null Hypothesis.

$$H_A : \theta \begin{cases} > \theta_0 & \text{(one-sided)} \\ < \theta_0 & \text{(one-sided)} \\ \neq \theta_0 & \text{(two-sided)} \end{cases} \tag{24.109}$$

Definition 24.55 Testing Parameters:
Given: a parameter θ that we want to test.
Let Θ be the set of all possible values that θ can achieve.
We now split Θ in two disjunct sets Θ_0 and Θ_1.

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \Theta_0 \cap \Theta_1 = \varnothing$$

Null Hypothesis	$H_0 : \theta \in \Theta_0$	(24.110)
Alternative Hypothesis	$H_A : \theta \in \Theta_1$	(24.111)

11.1. Type I&II Errors

Definition 24.56 Type I Error: Is the rejection of a Null Hypothesis, even-tough its true (also known as a "false positive").

Definition 24.57 Type II Error: Is the acceptance of a Null Hypothesis, even-tough its false (also known as a "false negative").

Decision	H_0 **true**	H_0 **false**
Accept	**TN**	**Type II (FN)**
Reject	**Type I (FP)**	**TP**

Definition 24.58 Critical Value c: Value from which on the Null-hypothesis H_0 gets rejected.

Definition 24.59 Statistical significance α: A study's defined significance level, denoted α, is the **probability** of the study rejecting the null hypothesis, given that the null hypothesis were true (Type I Error).

Definition 24.60 Critical Region K_α: Is the set of all values that causes us to reject the Null Hypothesis in favor for the Alternative Hypothesis H_A.
The Critical region is usually chosen s.t. we incur a Type I Error with probability less than α.

$$\boxed{K_\alpha \in \Theta : \mathbb{P}(\text{Type I Error}) \leqslant \alpha} \tag{24.112}$$

$$\begin{aligned} &\mathbb{P}(c_2 \leqslant X \leqslant c_1) \leqslant \alpha && \text{two-sided} \\ \text{or} \quad &\mathbb{P}(c_2 \leqslant X) \leqslant \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}(X \leqslant c_1) \leqslant \frac{\alpha}{2} \\ &\mathbb{P}(c_2 \leqslant X) \leqslant \alpha && \text{one-sided} \\ &\mathbb{P}(X \leqslant c_1) \leqslant \alpha && \text{one-sided} \end{aligned}$$

Definition 24.61 Acceptance Region: Is the region where we accept the null hypothesis H_0.

Note

see example 24.3.

11.2. Normally Distributed Data

Let us consider a sample of $\{x_i\}_{i=1}^n$ i.i.d. observations, that follow a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

11.2.1. Z-Test — σ known

11.2.2. t-Test — σ unknown

12. Inferential Statistics

Goal of Inference

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

13. Examples

Example 24.1 ??: Let x be uniformly distributed on $[0,1]$ ([def. 24.26]) with pmf $\mathrm{p}_X(x)$ then it follows:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{\mathrm{p}_Y(y)} \Rightarrow \mathrm{d}x = \mathrm{d}y\mathrm{p}_y(y) \Rightarrow x = \int_{-\infty}^{y} \mathrm{p}_y(t)\,\mathrm{d}t = F_Y(x)$$

Example 24.2 ??: Let

add https://www.youtube.com/watch?v=WUUb7VIRzgg

Example 24.3 Binomialtest:
Given: a manufacturer claims that a maximum of 10% of its delivered components are substandard goods.
In a sample of size $n = 20$ we find $x = 5$ goods that do not fulfill the standard and are skeptical that the what the manufacture claims is true, so we want to test:

$$H_0 : \mathrm{p} = \mathrm{p}_0 = 0.1 \qquad \text{vs.} \qquad H_A : \mathrm{p} > 0.1$$

We model the number of number of defective goods using the binomial distribution[def. 24.23]

$$X \sim \mathcal{B}(n, \mathrm{p}), n = 20 \qquad \mathbb{P}(X \geqslant x) = \sum_{k=x}^{n} \binom{n}{k} \mathrm{p}^k (1-\mathrm{p})^{n-k}$$

from this we find:

$$\mathbb{P}_{\mathrm{p}_0}(X \geqslant 4) = 1 - \mathbb{P}_{\mathrm{p}_0}(X \leqslant 3) = 0.13$$
$$\mathbb{P}_{\mathrm{p}_0}(X \geqslant 4) = 1 - \mathbb{P}_{\mathrm{p}_0}(X \leqslant 3) = 0.04 \leqslant \alpha$$

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 $\mathbb{P}(X \geqslant x) = 1 - \mathbb{P}(X \leqslant x-1)$

14. Proofs

Proof. corollary 24.6:

$$\mathbb{E}[\hat{\mu}_X] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\mathbb{E}[\underbrace{\mu + \cdots + \mu}_{1,\ldots,n}]$$

□

Proof. corollary 24.7:

$$\mathbb{V}[\hat{\mu}_X] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \overset{\text{Property 24.10}}{=} \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right]$$
$$\frac{1}{n^2} n\mathbb{V}[X] = \frac{1}{n}\sigma^2$$

□

Proof. definition 24.51:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_X^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\cdot n\bar{x} + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\mathbb{E}\left[\bar{x}^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{1}{n}\sigma^2 + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[(n\sigma^2 + n\mu^2) - \left(\sigma^2 + n\mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2
\end{aligned}$$

□

Stochastic Calculus

Stochastic Processes

Definition 25.1
Random/Stochastic Process $\{X_t, t \in \mathcal{T} \subseteq \mathbb{R}_+\}$:
is a collection of random variables on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The index set $\mathcal{T}$ is usually representing time and can be either an interval $[t_1, t_2]$ or a discrete set $\{t_1, t_2, \dots\}$. Therefore, the random process X can be written as a function:

$$X : \mathbb{R} \times \Omega \mapsto \mathbb{R} \qquad \Longleftrightarrow \qquad (t, \omega) \mapsto X(t, \omega) \tag{25.1}$$

Definition 25.2 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 $X(t, \omega) := X(t)$

Definition 25.3 Filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \geqslant 0}$:
A collection $\{\mathcal{F}_t\}_{t \geqslant 0}$ of sub σ-*algebras* $\{\mathcal{F}_t\}_{t \geqslant 0} \in \mathcal{F}$ is called filtration if is *increasing*:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \qquad \forall s \leqslant t \tag{25.2}$$

Definition 25.4 Adapted Process: A stochastic process $\{X_t : 0 \leqslant t \leqslant \infty\}$ is called adapted to a filtration $\mathbb{F}$ if, X_t is $\mathcal{F}_t$-measurable, i.e. observable at time t.

Definition 25.5 Predictable Process: A stochastic process $\{X_t : 0 \leqslant t \leqslant \infty\}$ is called predictable w.r.t. a filtration $\mathbb{F}$ if, X_t is $\{\mathcal{F}_{t-1}\}$-measurable, i.e. the value of X_t is known at time $t - 1$.

Note

The price of a stock will usually be adapted since date k prices are known at date k.
On the other hand the interest rate of a bank account is usually already known at the beginning $k - 1$, s.t. the interest rate r_t ought to be $\mathcal{F}_{k-1}$ measurable, i.e. the process $r = (r_k)_{k=1,\dots,T}$ should be predictable.

Definition 25.6
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.

Corollary 25.1 : The amount of information of an adapted random process is increasing see example 25.1.

Definition 25.7 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:
$$\forall s \leqslant t \qquad \mathbb{E}[X(t)|\mathcal{F}_s] = X(s) \quad \text{a.s.} \tag{25.3}$$
2. The expectation is finite:
$$\mathbb{E}[|X(t)|] < \infty \quad \forall t \geqslant 0 \qquad X(t) \text{ is } \{\mathcal{F}_t\}_{t \geqslant 0} \text{ adapted} \tag{25.4}$$

Interpretation

- For any $\mathcal{F}_s$-adapted process the best prediction of $X(t)$ is the currently known value $X(s)$ i.e. if $\mathcal{F}_s = \mathcal{F}_{t-1}$ then the best prediction is $X(t-1)$
- A martingale models fair games of limited information.

Definition 25.8 Auto Covariance $\gamma(t_2 - t_1)$:
Describes the covariance[def. 24.16] between two values of a stochastic process $(\boldsymbol{X}_t)_{t \in \mathcal{T}}$ at different time points t_1 and t_2.

$$\gamma(t_1, t_2) = \operatorname{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] = \mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})] \tag{25.5}$$

For zero time differences $t_1 = t_2$ the autocorrelation functions equals the variance:

$$\gamma(t, t) = \operatorname{Cov}[\boldsymbol{X}_t, \boldsymbol{X}_t] \overset{\text{eq. (24.35)}}{=} \mathbb{V}[\boldsymbol{X}_t] \tag{25.6}$$

Notes

- **Hence** the autocorrelation describes the correlation of a function or signal with itself at a previous time point.
- **Given** a random time dependent variable $\boldsymbol{x}(t)$ the autocorrelation function $\gamma(t, t - \tau)$ describes how *similar* the time translated function $\boldsymbol{x}(t - \tau)$ and the original function $\boldsymbol{x}(t)$ 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 25.9 Auto Correlation $\rho(t_2 - t_1)$:
Is the scaled version of the auto-covariance[def. 25.8]:

$$\rho(t_2 - t_1) = \operatorname{Corr}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}] \tag{25.7}$$
$$= \frac{\operatorname{Cov}[\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}} = \frac{\mathbb{E}[(\boldsymbol{X}_{t_1} - \mu_{t_1})(\boldsymbol{X}_{t_2} - \mu_{t_2})]}{\sigma_{X_{t_1}} \sigma_{X_{t_2}}}$$

1. Different kinds of Processes

1.1. Markov Process

Definition 25.10 Markov Process: A continuous-time stochastic process $X(t), t \in T$, is called a Markov process if for any finite parameter set $\{t_i : t_i < t_{i+1}\} \in T$ it holds:

$$\mathbb{P}(X(t_{n+1}) \in B | X(t_1), \dots, X(t_n)) = \mathbb{P}(X(t_{n+1}) \in B | X(t_n))$$

it thus follows for the *transition probability* – the probability of $X(t)$ lying in the set B at time t, given the value x of the process at time s:

$$\mathbb{P}(s, x, t, B) = P(X(t) \in B | X(s) = x) \quad 0 \leqslant s < t \tag{25.8}$$

Interpretation

In order to predict the future only the current/last value counts.

Corollary 25.2 Transition Density: The transition probability of a continuous distribution p can be calculated via:

$$\mathbb{P}(s, x, t, B) = \int_B \mathrm{p}(s, x, t, y)\, dy \tag{25.9}$$

1.2. Gaussian Process

Definition 25.11 Gaussian Process: Is a stochastic process $X(t)$ where the random variables follow a Gaussian distribution:

$$X(t) \sim \mathcal{N}\left(\mu(t), \sigma^2(t)\right) \qquad \forall t \in T \tag{25.10}$$

1.3. Diffusions

Definition 25.12 Diffusion: Is a Markov Process[def. 25.10] for which it holds that:

$$\mu(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}[X(t + \Delta t) - X(t) | X(t)] \tag{25.11}$$
$$\sigma^2(t, X(t)) = \lim_{t \to 0} \frac{1}{\Delta t} \mathbb{E}\left[(X(t + \Delta t) - X(t))^2 | X(t)\right] \tag{25.12}$$

See **??**/eq. (25.12) for simple proof of eq. (25.11)/**??**.

- $\mu(t, X(t))$ is called **drift**
- $\sigma^2(t, X(t))$ is called **diffusion coefficient**

Interpretation

There exist not discontinuities for the trajectories.

1.4. Brownian Motion/Wienner Process

Definition 25.13
***d*-dim standard Brownian Motion/Wienner Process:**
Is an $\mathbb{R}^d$ valued *stochastic process*[def. 25.1] $(W_t)_{t \in \mathcal{T}}$ starting at $\boldsymbol{x}_0 \in \mathbb{R}^d$ that satisfies:

1. **Normal Independent Increments**: the increments are *normally distributed independent random variables*:
$$W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, (t_i - t_{i-1}) \mathbb{1}_{d \times d}) \qquad \forall i \in \{1, \dots, T\} \tag{25.13}$$
2. **Stationary increments**:
$W(t + \Delta t) - W(t)$ is independent of $t \in \mathcal{T}$
3. **Continuity**: for *a.e.* $\omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous
$$\lim_{t \to 0} \frac{\mathbb{P}(|W(t + \Delta t) - W(t)| \geqslant \delta)}{\Delta t} = 0 \qquad \forall \delta > 0 \tag{25.14}$$
4. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{25.15}$$

Notation

- In many source the Brownian motion is a synonym for the standard Brownian Motion and it is the same as the Wienner process.
- **However** in some sources the Wienner process is the standard Brownian Motion, while the Brownian motion denotes a general form $\alpha W(t) + \beta$.

Corollary 25.3 $W_t \sim \mathcal{N}(0, \sigma)$:
The random variable W_t follows the $\mathcal{N}(0, \sigma)$ law

$$\mathbb{E}[W(t)] = \mu = 0 \tag{25.16}$$
$$\mathbb{V}[W(t)] = \mathbb{E}\left[W^2(t)\right] = \sigma^2 = t \tag{25.17}$$

See section 5

1.4.1. Properties of the Wienner Process

Property 25.1 Non-Differentiable Trajectories:
The sample paths of a Brownian motion are not differentiable:

$$\frac{dW(t)}{t} = \lim_{t \to 0} \mathbb{E}\left[\left(\frac{W(t + \Delta t) - W(t)}{\Delta t}\right)^2\right]$$
$$= \lim_{t \to 0} \frac{\mathbb{E}[W(t + \Delta t) - W(t)]}{\Delta t} = \lim_{t \to 0} \frac{\sigma^2}{\Delta t} = \infty$$

$\xrightarrow{\text{result}}$ cannot use normal calculus anymore
$\xrightarrow{\text{solution}}$ Ito Calculus see section 26.

Property 25.2 Auto covariance Function:
The auto-covariance[def. 25.8] for a Wienner process

$$\mathbb{E}\left[(W(t) - \mu t)(W(t') - \mu t')\right] = \min(t, t') \tag{25.18}$$

Property 25.3: A standard Brownian motion is a

Quadratic Variation

Definition 25.14 Total Variation: The total variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$LV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)| \tag{25.19}$$

$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition } {}^{[\text{def. 20.8}]} \text{ of } [a, b]\right\}$$

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 25.15
Total Quadratic Variation/"sum of squares":
The total quadratic variation of a function $f : [a, b] \subset \mathbb{R} \mapsto \mathbb{R}$ is defined as:

$$QV_{[a,b]}(f) = \sup_{\Pi \in \mathcal{S}} \sum_{i=0}^{n_\Pi - 1} |f(x_{i+1}) - f(x_i)|^2 \tag{25.20}$$

$$\mathcal{S} = \left\{\Pi\{x_0, \dots, x_{n_\Pi}\} : \Pi \text{ is a partition } {}^{[\text{def. 20.8}]} \text{ of } [a, b]\right\}$$

Corollary 25.4 Bounded (quadratic) Variation:
The (quadratic) variation[def. 25.14] of a function is bounded if it is finite:

$$\exists M \in \mathbb{R}_+ : \quad LV_{[a,b]}(f) \leqslant M \qquad \left(QV_{[a,b]}(f) \leqslant M\right) \qquad \forall \Pi \in \mathcal{S} \tag{25.21}$$

Theorem 25.1 Variation of Wienner Process: Almost surely the total variation of a Brownian motion over a interval $[0, T]$ is infinite:

$$\mathbb{P}(\omega : LV(W(\omega)) < \infty) = 0 \tag{25.22}$$

Theorem 25.2
Quadratic Variation of standard Brownian Motion:
The quadratic variation of a standard Brownian motion over $[0, T]$ is finite:

$$\lim_{N \to \infty} \sum_{k=1}^{N} \left[W\left(k \frac{T}{N}\right) - W\left((k-1)\frac{T}{N}\right)\right]^2 = T$$
with probability 1 (25.23)

See **??**

Corollary 25.5 : theorem 25.2 can also be written as:

$$(dW(t))^2 = dt \tag{25.24}$$

1.4.2. Lévy's Characterization of BM

Theorem 25.3
***d*-dim standard BM/Wienner Process by Paul Lévy:**
An $\mathbb{R}^d$ valued *adapted stochastic process*[def's. 25.1, 25.3] $(W_t)_{t \in \mathcal{T}}$ with the filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}_+}$, that satisfies:

1. **Start**
$$W(0) := W_0 = 0 \qquad a.s. \tag{25.25}$$
2. **Continuous Martingale**: W_t is an a.s. *continuous* martingale[def. 25.7] w.r.t. the filtration $(\mathcal{F}_t)_{t \in \mathcal{T}}$ under $\mathbb{P}$.
3. **Quadratic Variation**:
$$W_t^2 - t \text{ is also an martingale} \qquad \Longleftrightarrow \qquad QV(W_t) = t \tag{25.26}$$

is a standard Brownian motion[def. 25.20]. Proof see section 5

Further Stochastic Processes

1.4.3. White Noise

understand script and add

Definition 25.16 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):
$$\mathbb{E}[\epsilon * [k]] = 0 \qquad \forall k \in T_{\text{discrete}} \tag{25.27}$$
- Zero autocorrelation[def. 25.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * [k], \epsilon * [k + n]) = \mathbb{E}[\epsilon * [k] \epsilon * [k + n]^\intercal] = \mathbb{V}[\epsilon * [k]]\, \delta_{\text{discrete}}[n] \qquad \forall k, n \in T_{\text{discrete}} \tag{25.28}$$

With $\delta_{\text{discrete}}[n] := \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{else} \end{cases}$

See proofs

Definition 25.17 Continuous-time white noise: Is a random signal $(\epsilon_t)_{t \in T_{\text{continuous}}}$ having equal intensity at different frequencies and is defined by:

- Having zero tendencies/expectation (otherwise the signal would not be random):
$$\mathbb{E}[\epsilon * (t)] = 0 \qquad \forall t \in T_{\text{continuous}} \tag{25.29}$$
- Zero autocorrelation[def. 25.9] γ i.e. the signals of different times are in no-way correlated:
$$\gamma(\epsilon * (t), \epsilon * (t + \tau)) = \mathbb{E}[\epsilon * (t) \epsilon * (t + \tau)^\intercal] \tag{25.30}$$
$$\overset{\text{eq. (24.86)}}{=} \mathbb{V}[\epsilon * (t)]\, \delta(t - \tau) = \begin{cases} \mathbb{V}[\epsilon * (t)] & \text{if } \tau = 0 \\ 0 & \text{else} \end{cases}$$
$$\forall t, \tau \in T_{\text{continuous}} \tag{25.31}$$

Definition 25.18 Homoscedastic Noise: Is constant for all observations/time-steps:

$$\mathbb{V}[\epsilon_t] = \sigma^2 \qquad \forall t = 1, \dots, T \tag{25.32}$$

Definition 25.19 Heteroscedastic Noise: Is noise that can vary with each observation/time-step:

$$\mathbb{V}[\epsilon_t] = \sigma(t)^2 \qquad \forall t = 1, \dots, T \tag{25.33}$$

1.4.4. Generalized Brownian Motion

Definition 25.20 Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 25.13], and define:

$$X_t = \mu t + \sigma W_t \qquad t \in \mathbb{R}_+ \qquad \begin{matrix}\mu \in \mathbb{R} & \text{: drift parameter}\\ \sigma \in \mathbb{R}_+ & \text{: scale parameter}\end{matrix} \tag{25.34}$$

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 25.4 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:

$$f_{W(t)}(x) \sim \mathcal{N}(\mu t, \sigma^2 t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left\{-\frac{(x-\mu t)^2}{2\sigma^2 t}\right\} \tag{25.35}$$

Corollary 25.6 : More generally we may define the process:

$$t \mapsto f(t) + \sigma W_t \tag{25.36}$$

which corresponds to a noisy version of f.

Corollary 25.7
Brownian Motion as a Solution of an SDE: A stochastic process X_t follows a BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} \mathrm{d}X(t) &= \mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t) && (25.37)\\ X(0) &= 0 && (25.38)\end{aligned}$$

1.4.5. Geometric Brownian Motion (GBM)

For many processes $X(t)$ it holds that:
- there exists an (exponential) growth
- that the values may not be negative $X(t) \in \mathbb{R}_+$

Definition 25.21 Geometric Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 25.13] the exponential transform:

$$X(t) = \exp(W(t)) = \exp(\mu t + \sigma W(t)) \qquad t \in \mathbb{R}_+ \tag{25.39}$$

is called geometric Brownian motion

Corollary 25.8 Log-normal Returns: For a geometric BM we obtain log-normal returns:

$$\ln\left(\frac{S_t}{S_0}\right) = \mu t + \sigma W(t) \quad \Longleftrightarrow \quad \mu t + \sigma W(t) \sim \mathcal{N}(\mu t, \sigma^2 t) \tag{25.40}$$

meaning that the mean and the variance of the process (stock) *log-returns* grow over time linearly.

Corollary 25.9
Geometric BM as a Solution of an SDE:
A stochastic process X_t follows a geometric BM with drift μ and scale σ if it satisfies the following SDE:

$$\begin{aligned} \mathrm{d}X(t) &= X(t)\left(\mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)\right)\\ &= \mu X(t)\,\mathrm{d}t + \sigma X(t)\,\mathrm{d}W(t) && (25.41)\\ X(0) &= 0 && (25.42)\end{aligned}$$

1.4.6. Locally Brownian Motion

Definition 25.22 Locally Brownian Motion:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 25.13] a local Brownian motion is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = \mu\left(X(t), t\right)\mathrm{d}t + \sigma\left(X(t), t\right)\mathrm{d}W(t) \tag{25.43}$$

Note

A local Brownian motion is an generalization of a geometric Brownian motion.

1.4.7. Ornstein-Uhlenbeck Process

Definition 25.23 Ornstein-Uhlenbeck Process:
Let $\{W_t\}_{t\in\mathbb{R}_+}$ be a standard Brownian motion[def. 25.13] a Ornstein-Uhlenbeck Process or exponentially correlated noise is a stochastic process $X(t)$ that satisfies the SDE:

$$\mathrm{d}X(t) = -aX(t)\,\mathrm{d}t + b\sigma\,\mathrm{d}W(t) \qquad a > 0 \tag{25.44}$$

1.5. Poisson Processes

Definition 25.24 Rare/Extreme Events: Are events that lead to discontinuous in stochastic processes.

Problem

A Brownian motion is not sufficient as model in order to describe extreme events s.a. crashes in financial market time series. Need a model that can describe such discontinuities/jumps.

Definition 25.25 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. 24.25]:
$$\mathbb{P}\left((X_t - X_s) = k\right) = \frac{\lambda(t-s)}{k!}\mathrm{e}^{-\lambda(t-s)} \qquad \begin{matrix}0 \leqslant s < t < \infty\\ \forall k \in \mathbb{N}\end{matrix}$$
3. No correlation of (non-overlapping) increments:
$$\begin{aligned}&\forall t_0 < t_1 < \dots < t_n : \text{the increments are independent}\\ &X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}} \qquad (25.45)\end{aligned}$$

Interpretation

A Poisson Process is a *continuous-time* process with *discrete, positive* realizations in $\in \mathbb{N}_{\geqslant 0}$

Corollary 25.10 Probability of events: Using Taylor in order to expand the Poisson distribution one obtains:

$$\mathbb{P}\left(X_{(t+\Delta t)} - X_t \neq 0\right) = \lambda\Delta t + o(\Delta t^2) \qquad t \text{ small i.e. } t \to 0 \tag{25.46}$$

1. Thus the probability of an event happening during Δt is proportional to time period and the rate λ
2. The probability of two or more events to happen *during* Δt is of order $o(\Delta t^2)$ and thus extremely small (as *Deltat* is small).

Definition 25.26 Differential of a Poisson Process: The differential of a Poisson Process is defined as:

$$\mathrm{d}X_t = \lim_{\Delta t \to \mathrm{d}t}\left(X_{(t+\Delta t)} - X_t\right) \tag{25.47}$$

Property 25.4 Probability of Events for differential:
With the definition of the differential and using the previous results from the Taylor expansion it follows:

$$\begin{aligned}\mathbb{P}(\mathrm{d}X_t = 0) &= 1 - \lambda && (25.48)\\ \mathbb{P}(|\mathrm{d}X_t| = 1) &= \lambda && (25.49)\end{aligned}$$

Proofs

Proof. eq. (25.11):
Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned}\mathbb{E}[x(n)] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x_i(n-1) \pm \delta]\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x_i(n-1)]\\ &\overset{\text{induction}}{=} \mathbb{E}\left[x_{n-1}\right] = \dots \mathbb{E}[x(0)] = 0\end{aligned}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (25.12):
Let by δ denote the displacement of a particle at each step, and assume that the particles start at the center i.e. $x(0) = 0$, then we have:

$$\begin{aligned}\mathbb{E}\left[x(n)^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i(n)^2\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[x_i(n-1) \pm \delta]^2\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i(n-1)^2 \pm 2\delta x_i(n-1) + \delta^2\right]\\ &\overset{\text{ind.}}{=} \mathbb{E}\left[x_{n-1}^2\right] + \delta^2 = \mathbb{E}\left[x_{n-2}^2\right] + 2\delta^2 = \dots\\ &= \mathbb{E}[x(0)] + n\delta^2 = n\delta^2\end{aligned}$$

as $n = \frac{\text{time}}{\text{step-size}} = \frac{t}{\Delta x}$ it follows:

$$\sigma^2 = \mathbb{E}\left[x^2(n)\right] - \mathbb{E}[x(n)]^2 = \mathbb{E}\left[x^2(n)\right] = \frac{\delta^2}{\Delta x}t \tag{25.50}$$

Thus in expectation the particles goes nowhere. □

Proof. eq. (25.30):

$$\begin{aligned}\gamma(\epsilon * [k], \epsilon * [k+n]) &= \mathrm{Cov}[\epsilon * [k], \epsilon * [k+1]]\\ &= \mathbb{E}[(\epsilon * [k] - \mathbb{E}[\epsilon * [k]])(\epsilon * [k+n] - \mathbb{E}[\epsilon * [k+n]]\\ &\overset{\text{eq. (25.27)}}{=} \mathbb{E}[(\epsilon * [k])(\epsilon * [k+n])]\end{aligned}$$

□

Proof. corollary 25.3:
Since $B_t - B_s$ is the increment over the interval $[s, t]$, it is the same in distribution as the incremeent 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. (25.15) it follows:

$$B_t - B_s \sim B_{t-s} \qquad B_{t-s} \sim \mathcal{N}(0, t-s)$$

□

Proof. corollary 25.3:

$$W(t) = W(t) - \underbrace{W(0)}_{=0} \sim \mathcal{N}(0, t)$$

$$\Rightarrow \qquad \mathbb{E}[X] = 0 \qquad \mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = t$$

□

Proof. theorem 25.2:

$$\begin{aligned}&\sum_{k=0}^{N-1}\left[W(t_k) - W(t_{k-1})\right]^2 && t_k = k\frac{T}{N}\\ &= \sum_{k=0}^{N-1} X_k^2 && X_k \sim \mathcal{N}\left(0, \frac{T}{N}\right)\\ &= \sum_{k=0}^{N-1} Y_k = n\left(\frac{1}{n}\sum_{k=0}^{N-1} Y_k\right) && \mathbb{E}[Y_k] = \frac{T}{N}\\ &\overset{\text{S.L.L.N}}{=} n\frac{T}{n} = T\end{aligned}$$

□

Proof. theorem 25.3 (2):
1. first we need to show eq. (25.3): $\mathbb{E}[W_t|\mathcal{F}_s] = W_s$
Due to the fact that W_t is $\mathcal{F}_t$ measurable i.e. $W_t \in \mathcal{F}_t$ we know that:

$$\mathbb{E}[W_t|\mathcal{F}_t] = W_t \tag{25.51}$$

$$\begin{aligned}\mathbb{E}[W_t|\mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s|\mathcal{F}]\\ &= \mathbb{E}[W_t - W_s|\mathcal{F}_s] + \mathbb{E}[W_s|\mathcal{F}_s]\\ &\overset{\text{eq. (25.51)}}{=} \mathbb{E}[W_t - W_s] + W_s\\ &\overset{W_t - W_s \sim \mathcal{N}(0, t-s)}{=} W_s\end{aligned}$$

2. second we need to show eq. (25.4): $\mathbb{E}[|X(t)|] < \infty$

$$\mathbb{E}[|W(t)|]^2 \overset{??}{\leqslant} \mathbb{E}\left[|W(t)|^2\right] = \mathbb{E}\left[W^2(t)\right] = t \leqslant \infty$$

□

Proof. theorem 25.3 (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:

$$\begin{aligned}\mathbb{E}\left[W_t^2|\mathcal{F}_s\right] &= \mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}_s\right] + \mathbb{E}[2W_s(W_t - W_s)|\mathcal{F}_s]\\ &\quad + \mathbb{E}\left[W_s^2|\mathcal{F}_s\right]\\ &\overset{\text{eq. (25.51)}}{=} \mathbb{E}\left[(W_t - W_s)^2\right] + 2W_s\mathbb{E}[(W_t - W_s)] + W_s^2\\ &\overset{\text{eq. (25.17)}}{=} \mathbb{V}[W_t - W_s] + 0 + W_s^2\\ &= t - s + W_s^2\end{aligned}$$

from this it follows that:

$$\mathbb{E}\left[W_t^2 - t|\mathcal{F}_s\right] = W_s^2 - s$$

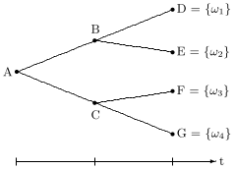
□

understand why $\mathbb{E}\left[(W_t - W_s)^2|\mathcal{F}\right] = \mathbb{E}\left[(W_t - W_s)^2\right]$

Examples

Example 25.1 :

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.



$$\mathcal{F} = \begin{cases}\{\varnothing, \Omega\} & t \in [0, T/2)\\ \{\varnothing, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\} & t \in [T/2, T)\\ \mathcal{F}_{\max} = 2^{\Omega} & t = T\end{cases} \tag{25.52}$$

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