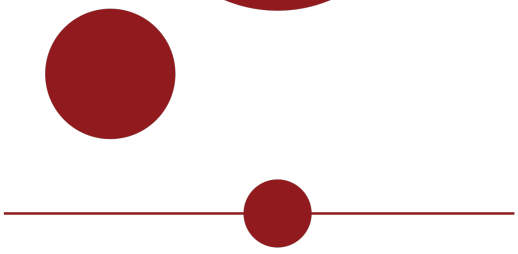
University of Copenhagen

Msc in Actuarial Mathematics

Complete Theory

Author:
Joakim Bilyk

Date:
March 28, 2023



UNIVERSITY OF
COPENHAGEN

Abstract

This document contain a comprehensive outline of theory on probability theory and mathematical statistics applied in finance, life insurance and non-life insurance.

Keywords: *probability theory, insurance mathematics, life insurance, non-life insurance, stochastic differential equations.*

Contents

1 Introduction **1**
1.1 Abbreviations . . . 1
1.2 To-do reading . . . 2

2 Basic Life Insurance Mathematics **3**

3 Stochastic Processes in Life Insurance Mathematics **5**

4 Topics in Life Insurance Mathematics **7**
4.1 Markov Jump Processes . . . 7
4.2 Phase-type distributions . . . 9
4.3 Interest rates . . . 14
4.3.1 Basic definitions and properties . . . 14
4.3.1.1 Forward rates and yields . . . 15
4.3.2 Phase-type representation of bond prices . . . 16
4.3.3 Term structure models . . . 18
4.3.4 Estimation of PH bond models . . . 19
4.3.4.1 Data for estimation . . . 20
4.3.4.2 Estimation of Phase-type models . . . 20
4.4 Survival and mortality rates . . . 20
4.4.1 Survival probabilities and forward mortality rates . . . 20
4.4.2 Forward transistion rates . . . 22
4.4.3 Reserves revisited . . . 23
4.4.4 Stochastic mortality rates . . . 23
4.5 Matrix methods in life insurance . . . 23
4.5.1 Basic setup . . . 23
4.5.2 Interest rate free analysis . . . 24
4.5.3 Transform of rewards and higher order moments . . . 25
4.5.4 Markovian interest rates . . . 27
4.5.5 Reserves . . . 28
4.5.6 Higher order moments . . . 29
4.5.7 Equivalence premium . . . 29
4.5.8 Distributions of future payments . . . 29
4.6 Financial Mathematics in Life Insurance . . . 29
4.6.1 Background and Simple Claims . . . 29
4.6.2 Payment Streams . . . 29
4.6.3 Unit-Link Insurance . . . 29
4.6.4 With-Profit Insurance and the Dynamics of the Surplus . . . 29
4.6.5 Cash Dividends and Market Reserve . . . 29
4.6.6 The Pure Case of Cash Dividends . . . 29
4.6.7 Bonus Payments and Market Reserve . . . 29
4.6.8 The Pure Case of Bonus Payments . . . 29

4.6.9 Comparison of Products . . . 29
4.7 Special Studies in Life Insurance . . . 29
4.7.1 Survival Probabilities and Forward Mortality Rates . . . 29
4.7.2 Dependent Interest and Mortality Rates . . . 29
4.7.3 Stochastic Interest and Mortality Rate Models . . . 29
4.7.4 Reserves Revisited . . . 29
4.7.5 Incidental Policy Holder Behavior . . . 29

5 Continuous Time Finance **31**
5.1 Discrete time models . . . 31
5.1.1 One-period time models . . . 31
5.1.1.1 Model description . . . 31
5.1.1.2 Portfolios and arbirtage . . . 32
5.1.1.3 Contingent Claims . . . 33
5.1.1.4 Risk Neutral Valuation . . . 34
5.1.2 Multi-period model . . . 35
5.1.3 Generelised one-period model . . . 39
5.1.3.1 Model specification . . . 39
5.1.3.2 Absence of Arbitrage . . . 39
5.1.3.3 Martingale Measures . . . 40
5.1.3.4 Martingale Pricing . . . 41
5.1.3.5 Completeness . . . 41
5.1.3.6 Stochastic Discount Factors . . . 41
5.2 Self-financing portfolios . . . 42
5.2.1 Discrete time SF portfolio . . . 42
5.2.2 Continuous time SF portfolio . . . 43
5.2.3 Portfolio weights . . . 43
5.3 Black-Scholes PDE . . . 45
5.3.1 Contingent Claims and Arbitrage . . . 45
5.3.2 Risk Neutral Valuation . . . 46
5.3.3 Black-Scholes formula . . . 47
5.4 Completeness and Hedging . . . 50
5.4.1 Completeness in Black-Scholes . . . 50
5.4.2 Absence of Arbitrage . . . 51
5.4.3 Incomplete Markets . . . 52
5.5 Parity relations . . . 54
5.5.1 Put-call Parity . . . 54
5.5.2 The Greeks . . . 55
5.6 Fundamental pricing theorem I and II . . . 56
5.6.1 Completeness . . . 58
5.6.2 Risk Neutral Valuation Formula . . . 58
5.6.3 Stochastic Discount Factors . . . 59
5.6.4 Summary . . . 59
5.7 Mathematics of the martingale approach . . . 61
5.7.1 Martingale representation theorem . . . 61
5.7.2 Girsanov theorem . . . 61
5.8 Black-Scholes model - martingale approach . . . 64
5.9 Multidimensional models . . . 66

6 Basic Non-Life Insurance Mathematics **69**

7 Stochastic Processes in Non-Life Insurance Mathematics **71**

8 Topics in Non-Life Insurance Mathematics **73**

9 Probabilistic Machine Learning **75**
9.1 Supervised Learning . . . 75
9.1.1 What is a good estimator? . . . 76
9.1.2 Excess risk . . . 77
9.2 Training, Validating and Testing . . . 79
9.2.1 Estimating risk . . . 79
9.3 Linear Models . . . 80
9.3.1 Least Squares Estimator . . . 82
9.3.2 Ridge Regression . . . 83
9.3.3 Lasso Regression . . . 84
9.3.4 Conclusion . . . 88
9.4 Nonparametric Regression . . . 88
9.4.1 Linear Smoothers . . . 88
9.4.2 Curse of dimensionality . . . 91
9.4.3 Splines . . . 91
9.4.4 Linear regression with splines. . . . 92
9.5 Trees and forests . . . 93
9.5.1 CART . . . 94
9.5.2 Pruning . . . 95
9.5.3 Bagging . . . 96
9.5.4 Random Forests . . . 97
9.6 Boosting and additive trees . . . 97
9.6.1 Gradient Boosting Machines . . . 98
9.6.2 Bayesian additive regression trees . . . 100
9.7 Some practical considerations . . . 100
9.8 Neural Networks . . . 102
9.9 Local explanations . . . 106
9.9.1 Interpretability . . . 106
9.10 Causality . . . 110
9.11 Local and Global Explanations . . . 114
9.11.1 Interpretability . . . 114
9.11.2 Partial dependence plots . . . 114
9.11.3 A functional decomposition . . . 115

10 Quantative Risk Management **119**
10.1 The Loss Variable . . . 119
10.1.1 Risk measures . . . 120
10.1.1.1 Value at Risk . . . 120

11 Measure theory **123**
11.1 Axioms of Probability . . . 123
11.2 Conditional Probability and Independence . . . 124
11.3 Probabilities on a Finite or Countable Space . . . 124
11.4 Construction of a Probability Measure on $\mathbb{R}$ . . . 125
11.5 Random Variables . . . 126
11.6 Integration with Respect to a Probability Measure . . . 126
11.7 Independent Random Variables . . . 127
11.8 Probability Distributions on $\mathbb{R}$ . . . 128
11.9 Probability Distributions on $\mathbb{R}^n$ . . . 129
11.10 Equivalent Probability Measures . . . 130
11.10.1 The Radon-Nikodym Theorem . . . 130
11.10.2 Equivalent Probability Measures . . . 131
11.10.3 Likelihood processes . . . 131

12 Random Variables **133**
12.1 Introduction . . . 133
12.2 Conditional expectation . . . 136
12.3 Independence . . . 138
12.4 Moment generating function . . . 140
12.5 Standard distributions . . . 141
12.5.1 Normal disribution . . . 141

13 Discrete Time Stochastic Processes **143**
13.1 Convergence concepts . . . 143
13.1.1 Sums and average processes . . . 146
13.1.2 Ergodic Theory . . . 148
13.1.3 Weak Convergence . . . 151
13.1.4 Central Limit Theorems . . . 156

14 Markov Chains **159**
14.1 Definition of a Markov Chain . . . 159
14.2 Classification of states . . . 160
14.3 Limit results and invariant probabilities . . . 161
14.4 Absorbing probabilities . . . 162
14.5 Markov Chains in Continuous Times . . . 162
14.6 Properties of transitionsprobabilities . . . 163
14.7 Invariant probabilies and absorption . . . 164
14.8 Birth-death processes . . . 166

15 Continuous Time Stochastic Processes **169**
15.1 Brownian Motion . . . 169
15.2 Filtration . . . 171
15.3 Martingale . . . 172

16 Stochastic calculus **173**
16.1 Stochastic Integrals . . . 173
16.1.1 Information . . . 173
16.1.2 Stochastic Integrals . . . 174
16.1.3 Martingales . . . 175
16.1.4 Stochastic Calculus and the Ito Formula . . . 175
16.1.5 The multidimensional Ito Formula . . . 176
16.1.6 Correlated Brownian motions . . . 177
16.2 Discrete Stochastic Integrals . . . 178
16.3 Stochastic Differential Equations . . . 179
16.4 Partial differential equations . . . 180
16.5 The Product Integral . . . 181
16.5.1 Properties of the Product Integral . . . 183

17 Linear Algebra **187**
17.1 Invertible matrices . . . 187

18 Coding **189**
18.1 R-Packages . . . 189
18.1.1 mlr3 . . . 189

Chapter 1

Introduction

1.1 Abbreviations

Below is given the abbreviations used when referencing to books:

Chapter	Abbreviation	Source
Basic Life Insurance Mathematics Stochastic Processes in Life Insurance Mathematics		
Topics in Life Insurance Mathematics	Asmussen	*Risk and Insurance*: A Graduate Text by Soren Asmussen and Mogens Steffensen (2020).
	Bladt	Notes from lectures in Liv2.
Continuous Time Finance	Bjork	*Arbitrage Theory in Continuous Time (Fourth edition)* by Thomas Bjork, Oxford University Press (2019).
Basic Non-Life Insurance Mathematics Stochastic Processes in Life Insurance Mathematics Topics in Non-Life Insurance Mathematics		
Probabilistic Machine Learning Quantative Risk Management	*None*	Slides from lectures.
Measure Theory	Bjork	*Arbitrage Theory in Continuous Time (Fourth edition)* by Thomas Bjork, Oxford University Press (2019).
	Protter	*Probability Essentials (2. edition)* by Jean Jacod and Philip Protter (2004).
Random Variables	Bjork	*Arbitrage Theory in Continuous Time (Fourth edition)* by Thomas Bjork, Oxford University Press (2019).
	Hansen	*Stochastic Processes* (2. edition) by Ernst Hansen (2021).
Discrete Time Stochastic Processes	Hansen	*Stochastic Processes* (2. edition) by Ernst Hansen (2021).

Chapter	Abbreviation	Source
Continuous Time Stochastic Processes	Bjork	*Arbitrage Theory in Continuous Time (Fourth edition)* by Thomas Bjork, Oxford University Press (2019).
Stochastic Calculus	Bjork	*Arbitrage Theory in Continuous Time (Fourth edition)* by Thomas Bjork, Oxford University Press (2019).
	Bladt	Notes from lectures in Liv2.
Linear Algebra	Wiki	Wikipedia

1.2 To-do reading

Chapter	Note	Progress
Liv2	Matrix representation	4 and later
	Chapter 6	Remember handouts
	Chapter 7	Remember handouts
ML	Make outlines for exam	
QRM	Finish risk measures	
StokLiv	Week 1	
StokLiv	Week 2	
SkadeStok	Week 1	
SkadeStok	Week 2	
Liv1	Week 1	
Liv1	Week 2	
Nonlife1	Week 1	
Nonlife1	Week 2	

Chapter 2

Basic Life Insurance Mathematics

Noget indhold

Chapter 3

Stochastic Processes in Life Insurance Mathematics

Noget indhold

Chapter 4

Topics in Life Insurance Mathematics

4.1 Markov Jump Processes

Life insurance mathematics revolves arrow Markov processes in continuous time on some discrete (or at least countable) state space. This may in the simplest model be the life-death model or it can be a generalized life-death model with interest levels. In any case, we use Markov processes to determine the possible payments that may occur in the future or have occurred in the past. Let us start by define the Markov jump process as follows.

Definition 2.1. (Bladt) *A continuous-time stochastic process $\{X_t\}_{t\geqslant 0}$ taking values in a countable space E is called a Markov jump process with state space E if for all $t_n > t_{n-1} > \cdots > t_1 > 0$ and $i_n, i_{n-1}, ..., i_0 \in E$ it holds that*

$$P(X_{t_n} = i_n \mid X_{t_{n-1}} = i_{n-1}, ..., X_{t_1} = i_1, X_0 = i_0) = P(X_{n-1} = i_n \mid X_{n-1} = i_{n-1}).$$

It is seen that process have some property of "forgetability" in the sense that the jump to a fixed state at some future time only depends on the current state, not the sample path of X. We may then define the transition probabilities as

$$p_{ij}(s,t) = P(X_t = j \mid X_s = i),$$

being the probability of the process being in state j at time t given that the process is in state i at time s. This gives the natural transition matrix

$$\mathbf{P}(t,s) = \big\{p_{ij}(s,t)\big\}_{i,j\in E}, \quad s \leqslant t.$$

In general, we study the time in-homogeneous models where the above is possibly different for any pair $s \leqslant t$. There does however exist the special case, where the probabilities only depends on the time $h = t - s$. We call these time homogeneous Markov jump processes and they have the form

$$p_{ij}(t,s) = p_{ij}(t-s).$$

I.e. the probability of jumping from i to j is the same so long the interval is the same length. Let us now define the transition intensities as below.

Definition 2.2. (Bladt) *Assume that the limit*

$$\mathbf{M}(s) = \big\{\mu_{ij}(s)\big\}_{i,j\in E} = \lim_{h\to 0_+} \frac{\mathbf{P}(s, s+h) + \mathbf{I}}{j}$$

exist for all $s \geqslant 0$. The matrices $\mathbf{M}(s)$ are called intensity matrices. The the special case of time-homogenous processes we have $\mathbf{M}(s) = \mathbf{M}$ is a constant matrix.

We may interpret the intensity matrices in the infinitesimal interpretation as

$$\mathbf{P}(s, s+h) = \mathbf{I} + h\mathbf{M}(s) + o(h),$$

i.e.

$$p_{ij}(s, s+h) = \delta_{ij} + \mu_{ij}(s)h + o(h),$$

where $\delta_{ij} = 1_{i=j}$ is the Kronecker delta. We furthermore see, that the matrix $\mathbf{M}(s)$ has non-positive diagonal and non-negative upper and lower triangle. Furthermore, the matrix has zero row sums and so it follows that

$$\mu_{ii}(s) = -\sum_{j \neq i} \mu_{ij}(s) := -\mu_i(t),$$

for all $i \in E$. Returning to the infinitesimal interpretation we may write the dynamics of μ_{ij} in the following manner

$$d\mu_{ij}(t) = \mu_{ij}(t)\ dt = P(X_{t+dt} = j \mid X_t = i)$$

i.e. the probability of a jump from state i to state j during $[t, t+dt)$. On the other hand we have

$$1 - \mu_{ii}(t)\ dt = P(X_{t+dt} = i \mid X_t = i),$$

i.e. the probability that no jumps occur during $[t, t+dt)$. We can then define the conditional probability that a jump, if one occurs, is to state j from i as

$$\begin{aligned} q_{ij}(t) &= P(X_{t+dt} = j \mid X_t = i, X_{t+dt} \neq i) = \frac{P(X_{t+dt} = j \mid X_t = i)}{P(X_{t+dt} \neq i \mid X_t = i)} \\ &= \frac{\mu_{ij}(t)}{\mu_i(t)} = -\frac{\mu_{ij}(t)}{\mu_{ii}(t)}. \end{aligned}$$

We can now look at the time until the next jump as the random variable

$$T(s) = \inf\{u \geqslant 0 \ : \ X_{s+u} \neq X_s\},$$

i.e. $T(s)$ is a positive random variable. If we condition on $X_s = i$ then $T(s) \mid X_s = i$ is the time until X jumps out of state i into some other state $j \neq i$. If we set $S_i(t) = P(T(s) > t \mid X_s = i)$ and $f_i(t)$ is the density of $T(s) \mid X_s = i$, then we have

$$\mu_i(u) = \frac{f_i(u)}{S_i(u)} = -\frac{S_i'(u)}{S_i(u)} = -\frac{d}{du}\log(S_i(u))$$

which is solved for

$$S_i(t) = P(T(s) > t \mid X_s = i) = \exp\left(-\int_s^t \mu_i(u)\ du\right)$$

and then

$$f_i(t) = \exp\left(-\int_s^t \mu_i(u)\ du\right)\mu_i(t).$$

Theorem 2.3. (Bladt) (Kolmogorov' Differential Equations) *Relation between* $\mathbf{M}(s)$ *and* $\mathbf{P}(s,t)$ *are given by Kolmogorov' forward and backward differential equations:*

$$\frac{\partial}{\partial s}\mathbf{P}(s,t) = -\mathbf{M}(s)\mathbf{P}(s,t)$$

and

$$\frac{\partial}{\partial t}\mathbf{P}(s,t) = \mathbf{P}(s,t)\mathbf{M}(t).$$

the soluction of which is given by

$$\mathbf{P}(s,t) = \prod_s^t (\mathbf{I} + \mathbf{M}(x)\ dx).$$

In the time-homogeneous case we have that

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{M}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{M}$$

with solution

$$\mathbf{P}(t) = \exp(\mathbf{M}t).$$

Proof.

We may subdivide the interval $[s,t]$ into $[s,u]$ and $[u,t]$ and use theorem 1.5 from the chapter on Product Integrals to obtain the Chapman-Kolmogorov's equation

$$\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t)$$

and in the time homogeneous case we have

$$\mathbf{P}(s+t) = \mathbf{P}(s)\mathbf{P}(t).$$

Finally, let us define a stopping time and introduce the strong markov property.

Definition 2.4. (Bladt) (Stopping time) *Let X be a Markov jump process and let $\mathcal{F}_t = \sigma(X_s\ :\ s \leqslant t)$. A nonnegative random variable τ is called a stopping time for X of $\{\tau \leqslant t\} \in \mathcal{F}_t$ for all t.\ The σ-algebra of the process up to a stopping time τ, $\mathcal{F}_\tau$, is defined by the measurable sets A for which*

$$\forall t \geqslant 0 : A \cap \{\tau \leqslant t\} \in \mathcal{F}_t.$$

We see that a stopping time in an intuitive sense is a random variable which at any time t we know either that the stopping time is in the future i.e. $\tau > t$ or that the stopping time already occured and so τ becomes known at time t. Furthermore, we see that $\{\tau \leqslant t\}$ is $\mathcal{F}_t$ measurable and so we only need information from the Markov jump process in order to determine the above.

Theorem 2.5. (Bladt) (Strong Markov property) *Every Markov jump process satisfies the strong Markov property, i.e., for all $0 \leqslant h_1 \leqslant h_2 \leqslant \cdots \leqslant h_n$, we have that*

$$P(X_{\tau+h_1} = i_1, ..., X_{\tau+h_n} = i_n \mid \mathcal{F}_t) = P(X_{\tau+h_1} = i_1, ..., X_{\tau+h_n} = i_n \mid X_\tau)$$

on $\{\tau < \infty\}$. For time-homogeneous processes this further reduces to

$$P(X_{\tau+h_1} = i_1, ..., X_{\tau+h_n} = i_n \mid \mathcal{F}_t) = P_{X_\tau}(X_{\tau+h_1} = i_1, ..., X_{\tau+h_n} = i_n)$$

also on $\{\tau < \infty\}$.

4.2 Phase-type distributions

In life insurance models we often, that is always, have a model with one absorbing state, namely, *death*. This means that the state space E will have the form of $p \geqslant 1$ transient states (states that may interact in both ways) and one absorbing state (a state that is never moved away from). In this case we would often like to study the distribution of

$$\inf\{t \geqslant 0 : X(t) = p+1\},$$

where $p+1$ is the absorbing state and $1, ..., p$ are the transient states. We will assume that $P(X(0) = p+1) = 0$ i.e. the above is at least zero and never $-\infty$. Let us now formalise this setup.

Consider a time-inhomogeneous Markov jump process $\{X_t\}_{t\geqslant 0}$ on the finite state-space $E = \{1, ..., p, p+1\}$ where the states $1, ..., p$ are transient states and $p+1$ is the only absorbing state. This implies that the intesity matrix of $\{X_t\}_{t\geqslant 0}$ take the form

$$\Lambda(t) = \begin{bmatrix} \mathbf{T}(t) & \mathbf{t}(t) \\ \mathbf{0} & 0 \end{bmatrix}.$$

In the above $\mathbf{T}(t)$ is a $p \times p$ matrix function and $\mathbf{t}(t)$ is a $p \times 1$ matrix function. We now define the initial distribution of X_0 as

$$\{P(X_0 = i)\}_{i \in E \setminus \{p+1\}} = \pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_p \end{pmatrix}^\top.$$

By assumption we have $P(X_0 = p+1) = 0$ and so $\sum_{i=1}^p \pi_i = 1$ and π is a proper distribution.

Definition 2.10. (Bladt) (Phase-Type distribution) *Let*

$$\tau = \inf\{t \geqslant 0 : X(t) = p+1\}$$

denote the time until absorption of X. The distribution of τ is then said to be an inhomogeneous phase-type distribution with representation $(\pi, \mathbf{T}(t))$ and we write $\tau \sim IPH(\pi, \mathbf{T}(t))$.

Lemma 2.11. (Bladt) *We have the following decomposition:*

$$\mathbf{P}(s,t) = \prod_s^t (\mathbf{I} + \mathbf{\Lambda}(u)\ du) = \begin{bmatrix} \prod_s^t (\mathbf{I} + \mathbf{T}(u)\ du) & \mathbf{e} - \prod_s^t (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{e} \\ \mathbf{0} & 1 \end{bmatrix},$$

where $\mathbf{e} = (1, 1, ..., 1)^\top$.

Theorem 2.12. (Bladt) *Assume that $\tau \sim IPH(\pi, \mathbf{T}(t))$. Then the density f and the distribution function F of τ are given by*

$$f(x) = \pi \prod_0^x (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{t}(x),$$
$$F(x) = 1 - \pi \prod_0^x (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{e}.$$

Proof.

Theorem 2.13. (Bladt) *If $\tau \sim IPH(\pi, \mathbf{T}(t))$ then*

$$P(\tau > s + t \mid \tau > s) = \frac{\pi \prod_0^s (\mathbf{I} + \mathbf{T}(u)\ du)}{\pi \prod_0^s (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{e}} \prod_s^t (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{e}$$

so that

$$\tau - s \mid \{\tau > s\} \sim IPH(\alpha, \mathbf{S}(\cdot)),$$

where $\mathbf{S}(u) = \mathbf{T}(s+u)$ and

$$\alpha = \frac{\pi \prod_0^s (\mathbf{I} + \mathbf{T}(u)\ du)}{\pi \prod_0^s (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{e}}.$$

Corollary 2.14. (Bladt) *If $\tau \sim IPH(\alpha, \mathbf{T}(t))$ and $\mathbf{T}(t_1)$ and $\mathbf{T}(t_2)$ commute for all $t_1, t_2 \geqslant 0$, then the density f and the distribution function F of τ are given by*

$$f(x) = \pi \exp\left(\int_0^x \mathbf{T}(u)\ du\right)\mathbf{t}(x),$$
$$F(x) = 1 - \pi \exp\left(\int_0^x \mathbf{T}(u)\ du\right)\mathbf{e}.$$

Example (Approximation in time-inhomogeneous case).

Consider the Markov jump process on the state space $E = \{1, 2, 3\}$ where 1 is the state *"alive"* (working), 2 is *"disabled"* and 3 is the state *"death"*. We assume that $\Lambda(t)$ has the structure:

$$\begin{aligned}\Lambda(t) &= \begin{bmatrix} \mu_{11}(t) & \mu_{12}(t) & \mu_{13}(t) \\ \mu_{21}(t) & \mu_{22}(t) & \mu_{23}(t) \\ \mu_{31}(t) & \mu_{32}(t) & \mu_{33}(t) \end{bmatrix} \\ &= \begin{bmatrix} -\mu_{12}(t) - \mu_{13}(t) & 0.000015 + 10^{(4.6-10+0.015\cdot t)} & 0.00005 + 10^{(4.6-10+0.05\cdot t)} \\ 0.000005 + 10^{(4.6-10+0.015\cdot t)} & -\mu_{21}(t) - \mu_{23}(t) & 0.0001 + 10^{(4.6-10+0.05\cdot t)} \\ 0 & 0 & 0 \end{bmatrix}\end{aligned}$$

We can now implement this intensity matrix as a function in R:

```
mu12 <- function(t) {
  0.000015 + 10^(4.6-10+0.015*t)
}
mu13 <- function(t) {
  0.00005 + 10^(4.6-10+0.05*t)
}
mu11 <- function(t) {
  -mu12(t)-mu13(t)
}
mu21 <- function(t) {
  0.000005 + 10^(4.6-10+0.015*t)
}
mu23 <- function(t) {
  0.0001 + 10^(4.6-10+0.05*t)
}
mu22 <- function(t) {
  -mu21(t)-mu23(t)
}
mu31 <- function(t) {0}
mu32 <- function(t) {0}
mu33 <- function(t) {0}
M <- function(t) {
  matrix(
    c(mu11(t),mu21(t),mu31(t),
      mu12(t),mu22(t),mu32(t),
      mu13(t),mu23(t),mu33(t)),
    ncol = 3
  )
}
M(0)
```

```
##               [,1]          [,2]          [,3]
## [1,] -7.296214e-05  1.898107e-05 5.398107e-05
## [2,]  8.981072e-06 -1.129621e-04 1.039811e-04
## [3,]  0.000000e+00  0.000000e+00 0.000000e+00
```

We see that the intensities is choosen such that the following holds for any time interval $[t, t + dt)$

1. The transition $1 \to 3$ is less likely than $2 \to 3$,
2. The transition $2 \to 1$ is less likely than $1 \to 2$,
3. The transition $1 \to 2$ is less likely than $1 \to 3$.

We now wish to compute the density and distribution of

$$\tau = \inf\{t \geqslant 0 \ : \ X(t) = 3\}.$$

We assume that $\pi = (1, 0)$ i.e. the person is alive. Calculating the distribution of τ is then the distribution of the length of a newborn child lifespan. From theorem 2.12 this ammount to calculating the product integral of

$$\mathbf{T}(t) = \{\mu_{ij}(t)\}_{i,j\in\{1,2\}}$$

i.e.

$$\prod_0^x (\mathbf{I} + \mathbf{T}(u)\ du).$$

To du this we simply approximate with stepsize $h = 1/n$ (for some $n \geqslant 1$) by

$$\prod_0^{x+h} (\mathbf{I} + \mathbf{T}(u)\ du) = h \prod_0^x (\mathbf{I} + \mathbf{T}(u)\ du)\mathbf{T}(x) + \prod_0^x (\mathbf{I} + \mathbf{T}(u)\ du),$$

and

$$\prod_0^0 (\mathbf{I} + \mathbf{T}(u)\ du) = \mathbf{I}.$$

This is done in the code below

```
n <- 12 #monthly
h <- 1/n
N <- 120 #max number years
T_matrix <- function(t) {
  matrix(
    c(mu11(t),mu21(t),
      mu12(t),mu22(t)),
    ncol = 2
  )
}
T_matrix(0)
```

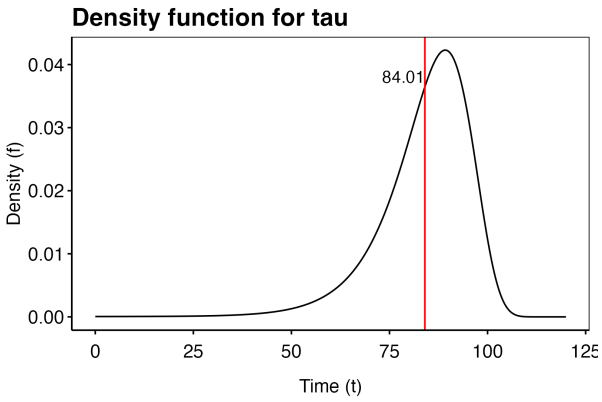
```
##               [,1]          [,2]
## [1,] -7.296214e-05  1.898107e-05
## [2,]  8.981072e-06 -1.129621e-04
```

```
t_matrix <- function(t) {
  matrix(
    c(mu13(t),mu23(t)),
    ncol=1
  )
}
t_matrix(0)
```

```
##              [,1]
## [1,] 5.398107e-05
## [2,] 1.039811e-04
```

```
library(rlist)
#Initial condition t=0
T_product_integral <- list(diag(c(1,1)))
pi <- matrix(c(1,0),ncol = 1)
f <- t(pi) %*% diag(c(1,1)) %*% t_matrix(0)
F <- 1-t(pi) %*% diag(c(1,1)) %*% matrix(c(1,1),ncol = 1)
for (i in 1:(N*n)) {
  x_1 <- i/n
```

```
  x_0 <- (i-1)/n
  T_0 <- T_matrix(x_0)
  M_0 <- T_product_integral[[i]]
  M_1 <- h*M_0 %*% T_0 + M_0
  T_product_integral <- list.append(T_product_integral,M_1)
  f <- c(f,t(pi) %*% M_1 %*% t_matrix(x_1))
  F <- c(F,1-t(pi) %*% M_1 %*% matrix(c(1,1),ncol = 1))
}
```



We see that by calculating

$$E[\tau \mid X_0 = 1] = \int_0^\infty \tau f(\tau)\, d\tau \approx \int_0^{120} \tau f(\tau)\, d\tau \approx \frac{1}{n} \sum_{i=0}^{120\cdot n} i/n \cdot f(i/n).$$

that the life expectation is 84.01 years. We have that we may approximate f and F with an arbitrary precision by choosing n appropriately. ▫

I practice we may have alot complications in calculating the product integral of $\mathbf{T}$ as the matrix may have large dimensions, be non-cummative, encounter alternating sums with growing terms and obviously being time in-homogeneous. So one has to be smart when constructing numerical methods in computing the integral. We will briefly consider some considerations an implementer may use in calculating the product integral.

1. **Applying differential equation.** Assume that X is a time in-homogeneous markov jump process then we may always calculate $\prod_s^t(\mathbf{I} + \mathbf{T}(x)\, dx)$ using a stepwise argument:

$$\prod_s^t(\mathbf{I} + \mathbf{T}(x)\, dx) = \prod_s^{s+1\cdot(t-s)/n}(\mathbf{I} + \mathbf{T}(x)\, dx) \prod_s^{s+2\cdot(t-s)/n}(\mathbf{I} + \mathbf{T}(x)\, dx) \cdots \prod_s^t(\mathbf{I} + \mathbf{T}(x)\, dx)$$

   for some $n \geqslant 1$. In particular one can simply choose n large enough such that the increments is approximately linear and so this is a brute force methods (see example above).
2. **Piece wise constant matrix.** Since data is scarce it often occur that mortality rates are constant over at least a monthly timeline hence one may assume that Λ is piecewise constant for some fine grid. This in particular means that if the grid has size $1/n$ (for instance $n = 12$ or $n = 4$) we have

$$\prod_s^{s+1/n}(\mathbf{I} + \mathbf{T}(x)\, dx) = \mathbf{I}e^{\mathbf{T}(s)\frac{1}{n}} = e^{\mathbf{T}(s)\frac{1}{n}}.$$

   and so the above reduces to

$$\prod_s^t(\mathbf{I} + \mathbf{T}(x)\, dx) = e^{\mathbf{T}(s)\frac{1}{n}} e^{\mathbf{T}(s+1/n)\frac{1}{n}} \cdots e^{\mathbf{T}(t-1/n)\frac{1}{n}},$$

assuming that s and t are integers (one could make this more general).

3. **Diagonalization.** Assume that $\mathbf{T}(x)$ is constant on some interval $[s,t]$ and that for the constant matrix $\mathbf{T} := \mathbf{T}(s)$ there exist unique eigenvalues $\lambda_1, ..., \lambda_p$. Then we can diagonalize $\mathbf{T}$ as
$$\mathbf{T} = \mathbf{B}\mathbf{D}\mathbf{B}^{-1},$$
where as usual $\mathbf{D} = \text{diag}(\lambda_1, ..., \lambda_p)$ and $\mathbf{B}$ is a $p \times p$ matrix with columns of eigenvectors for $\lambda_1, ..., \lambda_p$. In this case we can calculate
$$\begin{aligned}
\prod_s^t (\mathbf{I} + \mathbf{T}(x)\ dx) = e^{\mathbf{T}(t-s)} = e^{\mathbf{B}\mathbf{D}\mathbf{B}^{-1}(t-s)} &= \sum_{n=0}^{\infty} \frac{\left(\mathbf{B}\mathbf{D}\mathbf{B}^{-1}\right)^n (t-s)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{\mathbf{B}\mathbf{D}\mathbf{B}^{-1}\mathbf{B}\mathbf{D}\mathbf{B}^{-1}\cdots\mathbf{B}\mathbf{D}\mathbf{B}^{-1}\mathbf{B}\mathbf{D}\mathbf{B}^{-1}(t-s)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{\mathbf{B}\mathbf{D}^n\mathbf{B}^{-1}(t-s)^n}{n!} = \mathbf{B}\left(\sum_{n=0}^{\infty} \frac{\mathbf{D}^n(t-s)^n}{n!}\right)\mathbf{B}^{-1} \\
&= \mathbf{B}\left(\sum_{n=0}^{\infty} \frac{\text{diag}(\lambda_1^n(t-s)^n, ..., \lambda_p^n(t-s)^n)}{n!}\right)\mathbf{B}^{-1} \\
&= \mathbf{B}\text{diag}\left(\sum_{n=0}^{\infty} \frac{\lambda_1^n(t-s)^n}{n!}, ..., \sum_{n=0}^{\infty} \frac{\lambda_p^n(t-s)^n}{n!}\right)\mathbf{B}^{-1} \\
&= \mathbf{B}\text{diag}\left(e^{\lambda_1(t-s)}, ..., e^{\lambda_p(t-s)}\right)\mathbf{B}^{-1}.
\end{aligned}$$
which is much easier than any approximation. This is however a unrealisitic expectations to have.

4. **Univerformization.** Assume that $\mathbf{T}(x)$ is constant on some interval $[s,t]$ and define $\mathbf{T} := \mathbf{T}(s)$. We may furthermore define $\lambda = \max_i(-\lambda_{ii})$ as the largest diagonal entry in $\mathbf{T}$ ($\lambda < 0$). We now set
$$\mathbf{P} = \mathbf{I} + \lambda^{-1}\mathbf{T},$$
and see that $\mathbf{P}$ is a transition matrix i.e. rowsums is 1 and the the diagonals $0 \leqslant p_{ii} \leqslant 1$. We see this as $0 \leqslant \lambda^{-1}\lambda_{ii} \leqslant 1$ as $\lambda_{ii} \leqslant 0$ for all i and $\lambda \geqslant 0$ and dominates all diagonal entries. The rowsums is 1 as $\mathbf{T}$ is a intensity matrix i.e. rowsums is 0 and so adding one to the diagonal of the scaled matrix gives a rowsum of 1. We can now rearrange and see that
$$\begin{aligned}
\prod_s^t (\mathbf{I} + \mathbf{T}(x)\ dx) = e^{\mathbf{T}(t-s)} = e^{\lambda(\mathbf{P}-\mathbf{I})(t-s)} &= e^{\lambda\mathbf{P}(t-s)}e^{-\lambda\mathbf{I}(t-s)} \\
&= e^{\lambda\mathbf{P}(t-s)}\mathbf{I}e^{-\lambda(t-s)} = e^{-\lambda(t-s)}e^{\lambda\mathbf{P}(t-s)} \\
&= e^{-\lambda(t-s)}\sum_{n=0}^{\infty}\frac{(t-s)^n}{n!}\mathbf{P}^n.
\end{aligned}$$
This is nice since $\mathbf{P}$ has entries in the interval $[0,1]$ and so the series above converges and is monotonic, so we avoid the alternating sums from the negative diagonal in $\mathbf{T}$.

5. **Scaling and squaring argument.** Assuming the setup above. If the interval $(t-s)$ is large we may apply a scaling and squaring argument by setting
$$e^{\mathbf{T}(t-s)} = \mathbf{S}^{2^n}, \qquad \mathbf{S} = e^{\mathbf{T}(t-s)2^{-n}}.$$
This ensures small entries in the matrix S and we can simply square $\mathbf{S}$ n times after approximating $\mathbf{S}$ numerically.

4.3 Interest rates

4.3.1 Basic definitions and properties

In money markets interest is derived from the short term rate r that is a continuously compounded interest rate i.e. a bank account $B(t)$ is on the form

$$B(t) = B(0)e^{\int_0^t r(u)\ du}$$

where we will by convention assume $B(0) = 1$. Let us define what we will require for the process r.

Definition. (Interest rate process) *A stochastic process $\{r(t)\}_{t\geqslant 0}$ is an interest rate process if and only if $r(t)$ is adapted to some filtration $\mathcal{F}_t$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

This in particular means that in theory we may construct the rate as a diffusion process, piecewise constant function jumping at random times and so forth. The class of possible rate process is large but few is realistic. We will in the following think of r generally as a given process which we only assume is adapted to some filtration $\mathcal{F}$. The most important price process other than the bank account is the zero-coupon bond.

Definition. (Zero-Coupon Bond) *A zero-coupon bond with maturity T with underlying interest rate process $r(t)$ is a security which at time T pays the holder 1 and the price of the claim at time $0 \leqslant t \leqslant T$, $B(t,T)$, satisfies*

$$\left\{\frac{B(t,T)}{B_t}\right\}_{0\leqslant t\leqslant T},$$

is a $\mathbb{Q}$-martingale measure for some equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$ (including $\mathbb{Q} = \mathbb{P}$).

We can in particular deduce the price process under a $\mathbb{Q}$-expectation by the equation

$$E^{\mathbb{Q}}\left[\frac{1}{B_T}\middle|\mathcal{F}_t\right] = E^{\mathbb{Q}}\left[\frac{B(T,T)}{B_T}\middle|\mathcal{F}_t\right] = \frac{B(t,T)}{B_t},$$

since $B(T,T) = 1$ and the integrant is a $\mathbb{Q}$-martingale. Rearranging we have

$$B(t,T) = B_t E^{\mathbb{Q}}\left[\frac{1}{B_T}\middle|\mathcal{F}_t\right] = e^{\int_0^t r(u)\ du}E^{\mathbb{Q}}\left[e^{-\int_0^T r(u)\ du}\middle|\mathcal{F}_t\right] = E^{\mathbb{Q}}\left[e^{-\int_t^T r(u)\ du}\middle|\mathcal{F}_t\right]. \tag{11}$$

4.3.1.1 Forward rates and yields

There exist alot of different transformations of $r(t)$ which may be useful in simplifying complex equation and/or easing interpretation. Some of the most used include *forward rates* and *yields*. Let us start by defining a simple forward rate called the LIBOR forward rate.

Definition 3.1. (Bladt) *Consider three fixed time-points $t \leqslant S < T$ and the corresponding self-financing portfolio consisting of one short position in a zero-coupon bond with maturity S and a long position in a zero-coupon bond with maturity T of size $B(t,S)/B(t,T)$. This portfolio gives the payout: -1 at time S and $B(t,S)/B(t,T)$ at time T. This in particular results in the investment of 1 dollar at time S with a $\mathcal{F}_t$ deterministic payout*

$$\frac{B(t,S)}{B(t,T)}$$

*at time T. The **LIBOR forward rate** $L(t,S,T)$ is thus the average return in the interval i.e.*

$$L(t,S,T) = \frac{1}{T-S}\left(\frac{B(t,S)}{B(t,T)} - 1\right).$$

We could of couse define the rate as a continuously compounding rate by solving

$$e^{(T-S)R} = \frac{B(t,S)}{B(t,T)} \iff R = \frac{1}{T-S}\left(\log B(t,S) - \log B(t,T)\right).$$

Definition. (Continuously compounded forward rate) *Let $t \leqslant T$ be a fixed maturity date. The continuously compounded forward rate on $[S,T]$ is defined by*

$$f(t,S,T) = -\frac{\log B(t,T) - \log B(t,S)}{T-S}$$

being the continuously compounded rate for a portfolio consisting of $B(t,S)/B(t,T)$ zero-coupon bond with maturity T.

The yield with maturity T is the defined by.

Definition. (Yield) *Let $t \leqslant T$ be a fixed maturity date. The yield on $[t,T]$ is defined by*

$$Y(t,T) = -\frac{\log B(t,T)}{T-t},$$

being the continuous rate on the portfolio consisting of 1 zero-coupon bond with maturity T.

We obviously have that $Y(t,T) = f(t,t,T)$ is the continuous compounded forward rate on $[S,T]$ with $S = t$. Furthermore, we call $\{Y(t,T)\}_{T \geqslant t}$ the **yield curve**.

Definition. (Forward rate) *Let $t \leqslant T$ be a fixed maturity date. The forward rate on $[t,T]$ is defined by*

$$f(t,T) = -\frac{\partial}{\partial T}\log(B(t,T)).$$

We see that the forward rate is simply the continuously compounded forward rate on a infinitesimal interval of time $[T-dt,T]$ i.e.

$$f(t,T) = \lim_{S\to T} f(t,S,T) = \lim_{S\to T} -\frac{\log B(t,T) - \log B(t,S)}{T-S} = -\frac{\partial}{\partial T}\log(B(t,T)).$$

Note that this is all in expectation. The forward rate is useful in the sense that we can write

$$B(t,T) = \exp\left[-\int_t^T f(t,s)\ ds\right] = \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^T r(s)\ ds\right)\middle|\mathcal{F}_t\right].$$

We may use forward rate interchangeably as expectation-weighted discounting factor, in constructing fixed rates and arbitrage free loans.

4.3.2 Phase-type representation of bond prices

Consider the interest rate process

$$r(t) = r_{X(t)}(t),$$

where $\{X(t)\}_{t\geqslant 0}$ is a time-inhomogeneous Markov jump process on a statespace $E = \{1,...p\}$ with intensity matrix $\mathbf{M}(t) = \{\mu_{ij}\}_{i,j\in E}$ and $r_i(t)$ for $i = 1,...,p$ are deterministic functions. We call this a Markov-jump representation interest model and we summaries the model below.

Definition. (Markov-jump representation interest model) *Let $\{X(t)\}_{t\geqslant 0}$ be a p dimensional time-inhomogeneous Markov jump process. Define $\mathcal{F}_t = \sigma(X(s) : 0 \leqslant s \leqslant t)$. Then if $r_i(t)$ is bounded from below and deterministic the process*

$$r(t) = r_{X(t)}(t),$$

is an $\mathcal{F}_t$ adapted interest rate process.

We see that in the context of this interest rate model the zero-coupon bond with maturity T has price process

$$B(t,T) = \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^T r_{X(s)}(s)\ ds\right)\middle|\mathcal{F}_t\right] = \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^T r_{X(s)}(s)\ ds\right)\middle|X_t\right].$$

We can then make the matrix representation of the discounting factors.

Theorem 3.2. (Bladt) *For $i,j \in E$, let*

$$d_{ij}(s,t) = \mathbb{E}^{\mathbb{Q}}\left[1_{\{X(t)=j\}}\exp\left(-\int_s^t r_{X(u)}(u)\ du\right)\middle|X_s = i\right],$$

for $s \leqslant t$. Then the matrix $\mathbf{D}(s,t) = \{d_{ij}(s,t)\}_{i,j\in E}$ has representation

$$\mathbf{D}(s,t) = \prod_s^t\Big(\mathbf{I} + (\mathbf{M}(u) - \Delta(r(u)))\ du\Big),$$

with $\Delta(r(u)) = diag(r_1(u), ..., r_p(u))$.

We can then by setting ρ as

$$\rho = \max\left(0, -\min_{i\in E}\inf_{x\geqslant 0} r_i(x)\right),$$

use the matrix representation of $\mathbf{D}(s,t)$ to represent the zero-coupon bond and make a statement that relate $B(t,T)$ to an IPH distributed random variable.

Theorem 3.3. (Bladt) (Phase-type representation of bond prices) *Assume the Markov-jump interest model. The price process of the zero-coupon bond satisfies*

$$B(t,T) = \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^T r_{X(s)}(s)\ ds\right)\middle| X_t\right] = \pi_{X(t)}^{\top}\mathbf{D}(t,T)\mathbf{e},$$

where $\mathbf{e} = (1, ..., 1)^{\top}$ *and* $\pi_{X(t)}$ *is the distribution of* $X(t)$. *Let* $\tau \sim IPH(\pi_{X(t)}, \mathbf{M}(x+t) - \Delta(r(x+t)) - \rho\mathbf{I})$ *and define* $\overline{F}(t)$ *as the survival function for* τ. *Then we have*

$$B(t,T) = e^{\rho(T-t)}\overline{F}(T).$$

In particular if $\rho = 0$ *then* $B(t,T) = \overline{F}(T)$ *is the survival function of* τ.

We furthermore have the following result regarding the forward rate.

Corollary 3.4. (Bladt) *Assume the Markov-jump interest model and that* $X(t) = i$. *The forward rate is then the hazard rate at* T *for the random variable* $\tau \sim IPH(\pi_{X(t)}, \mathbf{M}(x+t) - \Delta(r(x+t)) - \rho\mathbf{I})$ *less* ρ *i.e.*

$$f(t,T) = \frac{f_\tau(T)}{\overline{F}(T)} - \rho.$$

where f_τ *is the density function of* τ.

We can now for explicity write the distribution function of τ in 3.3 and 3.4 above.

Corollary 3.5. (Bladt) *The stopping time* $\tau(t) \sim IPH(\pi, \mathbf{M}(x+t) - \Delta(r(x+t)) - \rho\mathbf{I})$ *with* $\pi_j = 1_{\{j=i\}}$ *has distribution function*

$$F_{\tau(t)}(T) = \mathbb{E}^{\mathbb{Q}}\left[\int_t^T r_{X(y)}(y)\exp\left(-\int_t^y r_{X(u)}(u)\ du\right)\ dy \ \middle|\ X(t) = i\right].$$

Corollary 3.6. (Bladt) *The stopping time* $\tau(t) \sim IPH(\pi, \mathbf{M}(x+t) - \Delta(r(x+t)) - \rho\mathbf{I})$, *with* π *as the initial distribution of* $X(t)$. *Then the following holds*

$$\begin{aligned}
\mathbb{P}(\tau > T) &= \mathbb{E}^{\mathbb{Q}}\left(\exp\left(-\int_0^T r_{X(u)}(u)\ du\right)\right),\\
F_{\tau(t)}(T) &= \mathbb{E}^{\mathbb{Q}}\left(\int_0^T r_{X(y)}(y)\exp\left(-\int_0^y r_{X(u)}(u)\ du\right)\ dy\right),\\
f_{\tau(t)}(t) &= \mathbb{E}^{\mathbb{Q}}\left(r_{X(t)}(t)\exp\left(-\int_0^t r_{X(u)}(u)\ du\right)\right),\\
f(0,T) &= \frac{f_{\tau(t)}(T)}{1 - F_{\tau(t)}(T)}.
\end{aligned}$$

4.3.3 Term structure models

We may introduce another rate model, where r is drifting with a locally deterministic term and a drift given by a Brownian motion. We call these models term structure models.

Definition. (Term structure interest model) *Let $\mathbb{Q}$ be an equivalent martingale measure and let $W^{\mathbb{Q}}$ be a $\mathbb{Q}$-Brownian motion. Assume that the process r has dynamics*

$$dr(t) = \alpha(t, r(t))\ dt + \sigma(t, r(t))\ dW^{\mathbb{Q}}(t),$$

with $r(0) = r_0$. Assume that $\alpha(t, r)$ and $\sigma(t, r)$ are deterministic functions. Define $\mathcal{F}_t = \sigma(W^{\mathbb{Q}}(s) : 0 \leqslant s \leqslant t)$. Then r is $\mathcal{F}_t$-adapted and we call r a term structure interest rate with local drift α and volatility σ.

One easily sees that r is indeed a Markov process and the arbitrage free price of a zero-coupon bond in this model is given by the price process

$$B(t, T) = p(t, r(t))$$

i.e. a function of time and the current rate. Furthermore, the pricing is derived from the risk neutral valuation formula:

$$p(t, r(r)) = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r(\tau)\ d\tau} \middle| \mathcal{F}_t \right] = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r(\tau)\ d\tau} \middle| r(t) \right],$$

as r is Markov. We then know that the process

$$e^{-\int_0^t r(u)\ du} B(t, T) = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_0^T r(\tau)\ d\tau} \middle| \mathcal{F}_t \right] \tag{23}$$

is a Doob $\mathbb{Q}$-martingale and so this gives rise to the term structure equation below.

Theorem 3.9. (Bladt) (Term structure equation) *Assume the term structure model. Then the bond price $B(t, T) = p(t, r(t))$ satisfies the PDE*

$$p_t(t, r) = rp(t, r) - \alpha(t, r)p_r(t, r) - \frac{1}{2}\sigma^2(t, r)p_{rr}(t, r)$$

subect to the condition

$$p(T, r) = 1.$$

The result follows from Ito's formula on the martingale in (23) (simply set the local drift equal to 0).

The equation above is generally not solvable, but there exist some restricted models, that does have an explicit solution. One of the models that does have a solution is the **affine model**.

Corollary. (Affine models) *Assume that the pricing function p of a zero-coupon bond has representation*

$$B(t, T) = p(t, r(t)) = e^{f(t)r(t)+g(t)},$$

then the term structure equation becomes

$$r = f'(t)r + g'(t) + \alpha(t, r)f(t) + \frac{1}{2}\sigma^2(t, r)f^2(t),$$

with terminal conditions

$$f(T) = g(T) = 0.$$

Assume furthermore that α and σ has representation

$$\alpha(t, r) = a(t) + b(t)r \quad \text{and} \quad \sigma(t, r) = \sqrt{c(t) + d(t)r}$$

then it follows that

$$f'(t) = 1 - b(t)f(t) - \frac{1}{2}d(t)f^2(t) \quad \text{and} \quad g'(t) = -a(t)f(t) - \frac{1}{2}c(t)f^2(t).$$

Proof.

From theorem 3.9 the term structure equation is

$$p_t(t,r) = rp(t,r) - \alpha(t,r)p_r(t,r) - \frac{1}{2}\sigma^2(t,r)p_{rr}(t,r).$$

Under assumption that the log of the price process is affine we find that

$$p_t(t,r) = (f'(t)r + g'(t)p(t,r), \quad p_r(t,r) = f(t)p(t,r), \quad p_{rr}(t,r) = f^2(t)p(t,r)$$

and so

$$(f'(t)r + g'(t)p(t,r) = rp(t,r) - \alpha(t,r)f(t)p(t,r) - \frac{1}{2}\sigma^2(t,r)f^2(t)p(t,r)$$

Hence

$$r = f'(t)r + g'(t) + \alpha(t,r)f(t) + \frac{1}{2}\sigma^2(t,r)f^2(t).$$

This shows the first part. Now assume that

$$\alpha(t,r) = a(t) + b(t)r, \quad \sqrt{c(t) + d(t)r}.$$

When inserting in the equation before we have

$$r = f'(t)r + g'(t) + (a(t) + b(t)r)f(t) + \frac{1}{2}(c(t) + d(t)r)f^2(t).$$

Then by a coefficient matching argument we have

$$\begin{pmatrix}1\\0\end{pmatrix}^\top = \begin{pmatrix}f'(t) + b(t)f(t) + \frac{1}{2}d(t)f^2(t)\\ g'(t) + a(t)f(t) + \frac{1}{2}c(t)f^2(t)\end{pmatrix}^\top$$

since by multiplying by $(r,1)^\top$ we get the equation above. Isolating f' and g' yields the result. ■

One popular model is proposed by Vasicek where

$$\alpha(t,r) = a - br(t) \quad \text{and} \quad \sigma(t,r) = \sigma,$$

with $a, b \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$ are constants. Given the initial condition $r(0) = r_0 \in \mathbb{R}$ we have the solution as

$$r(t) = r_0 e^{-bt} + \frac{a}{b}\left(1 - e^{-bt}\right) + \sigma \int_0^t e^{-(t-s)b}\, dW^{\mathbb{Q}}(t).$$

One may achieve the above integration result by differentiating the function $h(t,r) = r(t)e^{bt}$ and integrating the derivative giving

$$\begin{aligned} r(t)e^{bt} - r(0) = \int_0^t dh &= \int_0^t be^{bs}r(s) + e^{bs}(a - br(s))\, ds + \int_0^t e^{bs}\sigma\, dW^{\mathbb{Q}}(s) \\ &= \frac{a}{b}(e^{bt} - 1) + \sigma \int_0^t e^{bs}\, dW^{\mathbb{Q}}(s) \end{aligned}$$

where simpel isolation gives the equation.

4.3.4 Estimation of PH bond models

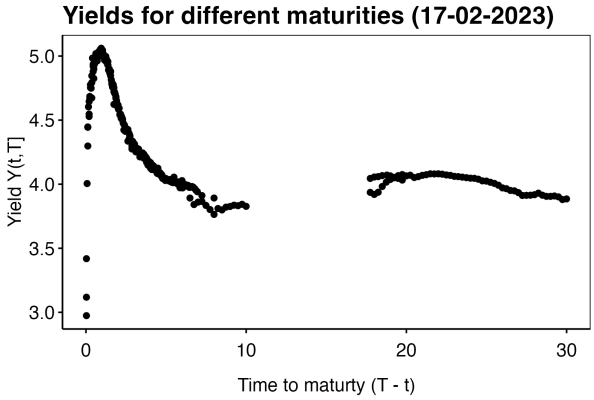
Estimation of rate models may be done by observing current spot rates on the market for zero-coupon bonds with a variety of maturity dates. One may for instance consider $T = t+1, ..., t+30$ or even up to $T = t+100$. This gives us the corresponding yield curve defined as

$$T \mapsto Y(t,T) = -\frac{\log p(t,T)}{T-t} = -\frac{\log B(t,T)}{T-t}.$$

If we make probabilistic assumptions on $r(t)$ we may in a parametrialized model estimate a parameter θ that associate a distribution function for $r(t)$ i.e. $\theta \mapsto F_{\theta,t}(r)$. Let us see how we may estimate a Markov-jump model and a term structure model below.

4.3.4.1 Data for estimation

We consider yield for non-zero coupon bonds listed on The Wall Street Journal. The data considered was pulled on the 17th of february 2023.



It must be stressed that the rates are not from zero-coupon bonds and so we in general have yields form contracts paying R continuously until maturity where the face value of 1 is payed. This in particular means that the prices in the data is prices according to

$$B_R(t,T] = \mathbb{E}^{\mathbb{Q}}\left[\int_t^T Re^{-\int_t^s r(u)\ du}\ ds + e^{-\int_0^T r(u)\ du}\middle|\mathcal{F}_t\right] = R\ \mathbb{E}^{\mathbb{Q}}\left[\int_t^T e^{-\int_t^s r(u)\ du}\ ds\middle|\mathcal{F}_t\right] + B(t,T]$$

This means that the yields are not equivalent to the zero-coupon yields. In fact it must hold that

$$Y_R(t,T] \geqslant Y(t,T]$$

with $Y_R(t,T) \approx Y(t,T] + R \cdot f(T-t)$ where $f \to 0$ for $T-t \to 0$ and $f \to 1$ for $T-t \to \infty$. We will in any case try to determine estimate a statistical model of the yield curve based on the yields in the dataset.

4.3.4.2 Estimation of Phase-type models

We restrict ourselves to the transition matrices on the form

$$\mathbf{T}(x) = \lambda_\theta(x)\mathbf{T},$$

where $\mathbf{T}$ is a constant $p \times p$ matrix and λ_θ is a real-valued function. The parameter space $\Theta \subset \mathbb{R}^d$ for some $d \geqslant 1$. This then gives a constant *base dynamic* between the rate states $1, ..., p \in E$ and λ_θ homogeneously speeds up or slows down the waiting time in each state by scaling the transitions intensities.

4.4 Survival and mortality rates

4.4.1 Survival probabilities and forward mortality rates

We consider a non-negative random variable τ on a probability space $(\Omega, \mathcal{F}, P)$. The mortality rate of τ at time t is defined through the dynamics in the following definition.

Definition. (Mortality rate) *Let $\tau \geqslant 0$ be a non-negative stochastic variable. The mortality rate function $\mu(t)$ is a, possibly stochastic, function defined as*

$$\mu(t)\ dt = \mathbb{P}(\tau \in [t, t+dt)\ |\ \tau > t) = \frac{f(t)}{\overline{F}(t)}\ dt,$$

where $\overline{F}(t) = 1 - F(t) = \mathbb{P}(\tau > t)$ is the survival function of τ.

The definition above shows that we may alternatively define μ in terms of the below derivative

$$\mu(t) = -\frac{d}{dt}\log\Big(\overline{F}(t)\Big),$$

hence we have that

$$-\int_0^t \mu(s)\ ds = \log\Big(\overline{F}(t)\Big) - \log\Big(\overline{F}(0)\Big) = \log\Big(\overline{F}(t)\Big),$$

giving the well-known formula

$$\overline{F}(t) = \exp\left(-\int_0^t \mu(s)\ ds\right).$$

This also gives the nice interpretation for the conditional distribution of τ given $\tau > t$ since.

$$\mathbb{P}(\tau > s \mid \tau > t) = \frac{\overline{F}(s)}{\overline{F}(t)} = \exp\left(-\int_t^s \mu(u)\ du\right).$$

Corollary. *Let $\tau \geqslant 0$ be a non-negative stochastic variable with mortality rate $\mu(t)$. The probability that $\tau \in [t, s)$ in the event $\tau > t$ is*

$$p(t,s) = \mathbb{P}(\tau > s \mid \tau > t) = \exp\left(-\int_t^s \mu(u)\ du\right).$$

How to deal with these random intensities. Either we could model them directly under a probability measure, $\mathbb{P}$ say, but since we do not know much about the probabilities anyway, we might as well model the consequences in price (reserve) by changes of the intensities. That is, how much does a change in intensity cost? This might been seen in a financial context, even as a derivative security, and should therefore be evaluated relative to a risk–free asset, i.e. by using an equivalent martingale measure, $\mathbb{Q}$ say. We therefore define the mortality forward rates as

Definition. (Mortality forward rate) *Let $\tau \geqslant 0$ be a non-negative stochastic variable with mortality rate $\mu(t)$. The mortality forward rate rate $m(t,s)$ is a function defined by*

$$\exp\left(-\int_t^s m(t,u)\ du\right) = q(t,s) = \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^s \mu(u)\ du\right)\middle|\, \mathcal{F}(t)\right].$$

Notice that in particular the function $q(t,s)$ solves the differential equation

$$\frac{\partial q(t,s)}{\partial s} = -m(t,s)q(t,s), \qquad q(t,t) = 1.$$

When evaluating a payment from insurance company and the insuread we also need to take into account the interest rate r and so bonds would be priced according to the quantity

$$E^{\mathbb{Q}}\left[\exp\left(-\int_t^s r(u) + \mu(u)\ du\right)\middle|\, \mathcal{F}(t)\right].$$

Notice that the filtration $\mathcal{F}(t)$ is now given by the path of $(\mu(t), r(t))$. A payment of 1 in the event the insured dies in the time interval $[s, s+dt)$ is given by

$$E^{\mathbb{Q}}\left[\exp\left(-\int_t^s r(u) + \mu(u)\ du\right)\mu(s)\middle|\, \mathcal{F}(t)\right].$$

Then means that we may price for instance a contract of a payment of 1 in the event of death as

$$E^{\mathbb{Q}}\left[\int_t^T \exp\left(-\int_t^s r(u) + \mu(u)\ du\right)\mu(s)\ ds\middle|\, \mathcal{F}(t)\right]$$

where the contract matures at time $T > t$. Furhermore if we assume indepence between the mortality and the rates (obviously satisfies) we have the following decompositions:

$$\begin{aligned}
\text{(Forward rates)}: &\quad E^{\mathbb{Q}}\left[\exp\left(-\int_t^s r(u)\ du\right)\middle|\mathcal{F}(t)\right] = e^{-\int_t^s f(t,u)\ du},\\
\text{(Bond prices)}: &\quad E^{\mathbb{Q}}\left[\exp\left(-\int_t^s r(u)+\mu(u)\ du\right)\middle|\mathcal{F}(t)\right] = e^{-\int_t^s f(t,u)+m(t,u)\ du},\\
\text{(Insurance event)}: &\quad E^{\mathbb{Q}}\left[\exp\left(-\int_t^s r(u)\ du\right)\mu(s)\middle|\mathcal{F}(t)\right] = e^{-\int_t^s f(t,u)+m(t,u)\ du}m(t,s).
\end{aligned}$$

4.4.2 Forward transistion rates

In the above sections we considered the simple Markov process with two states, one real and a absorbing state. In that context we simply model one jump. However in the general case, we would assume a Markov jump on a finite state space E i.e. $\{Z(t)\}_{t\geqslant 0}$ with random intensities $\mu_{ij}(t)$ with $i,j \in E$. We can combine these intensities into the matrix function

$$\mathbf{M}(s) = \{\mu_{ij}(s)\}_{i,j\in E}.$$

Then the natural definition of forward transition rates are the elements of the matrix, $\mathbf{F}(t,s)$, given by the product integral below

$$\mathbb{E}^{\mathbb{Q}}\left(\prod_t^s(\mathbf{I}+\mathbf{M}(u))\ du\middle|\mathcal{F}(t)\right) = \prod_t^s(\mathbf{I}+\mathbf{F}(t,u))\ du.$$

We in this context define

$$\mathbf{P}(t,s) = \prod_t^s(\mathbf{I}+\mathbf{M}(u))\ du,$$

being the transition probabilities $p_{ij}(t,s)$. We furthemore set the matrix $\mathbf{q}$ to

$$\mathbf{q}(t,s) = \mathbb{E}^{\mathbb{Q}}\left(\mathbf{P}(t,s)\middle|\mathcal{F}(t)\right) = \prod_t^s(\mathbf{I}+\mathbf{F}(t,u))\ du.$$

If we assume we may differentiate under the $\mathbb{Q}$ expectation we get

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left(\mathbf{P}(t,s)\mathbf{M}(s)\middle|\mathcal{F}(t)\right) &= \mathbb{E}^{\mathbb{Q}}\left(\frac{\partial}{\partial s}\mathbf{P}(t,s)\middle|\mathcal{F}(t)\right) = \frac{\partial}{\partial s}\mathbb{E}^{\mathbb{Q}}\left(\mathbf{P}(t,s)\middle|\mathcal{F}(t)\right)\\
&= \frac{\partial}{\partial s}\prod_t^s(\mathbf{I}+\mathbf{F}(t,u))\ du = \mathbf{q}(t,s)\mathbf{F}(t,s),
\end{aligned}$$

where we simply use the above definition and the definition of the product integral.

Lemma. (Mortality forward rate) *Let $\{Z(t)\}_{t\geqslant 0}$ be a Markov jump process on the finite state space E. Assume that the intensity matrix function $\mathbf{M}(t)$ exists, then*

$$\mathbf{F}(t,s) = \mathbf{q}(t,s)^{-1}\mathbb{E}^{\mathbb{Q}}\left(\mathbf{P}(t,s)\mathbf{M}(s)\middle|\mathcal{F}(t)\right)$$

with

$$\mathbf{q}(t,s) = \mathbb{E}^{\mathbb{Q}}\left(\mathbf{P}(t,s)\middle|\mathcal{F}(t)\right).$$

4.4.3 Reserves revisited

Consider an life insurance contract of an insured with underlying Markov jump process $\{Z(t)\}_{t\geqslant 0}$ commencing at time $T > 0$. The contract pays continuous rate $b^i(t)$ while $Z(t) = i$ and transition lump sum payments $b^{ij}(t)$ in the event Z jumps from i to j at time t. That is the payment has dynamics

$$dB(t) = dB^{Z(t)}(t) + \sum_{j\neq Z(t-)} b^{Z(t-)j} dN_{Z(t-)j}(t),$$

where $N_{ij}(t) = \#\{s \leqslant t : Z(s-) = i, Z(s) = j\}$ is the number of jumps from i to j in the interval $[0, t]$. The above is equivalent with the process

$$dB(t) = \sum_{i\in E} I^i(t)\ dB^i(t) + \sum_{i\in E}\sum_{j\in E:j\neq i} b^{ij}(t)\ dN_{ij}(t),$$

where obviously $I^i(t) = 1_{Z(t)=i}$. We assume that $\{r(t)\}_{t\geqslant 0}$ is the stochastic interest rate process which is independent of the payment process. The market reserve or third order reserve is defined under the martingale measure $\mathbb{Q}$ as

$$V(t) = \mathbb{E}^{\mathbb{Q}}\left[\int_t^T e^{-\int_t^u r(s)\ ds}\ dB(u)\middle|\mathcal{F}(t)\right],$$

i.e. the discounted expected value of the payments. The above filtration is the one given by the sample path of both r and Z. Using the independence assumption we have

$$\begin{aligned} V(t) &= \int_t^T E^{\mathbb{Q}}\left[e^{-\int_t^u r(s)\ ds}\ dB(u)\middle|\mathcal{F}(t)\right] \\ &= \int_t^T E^{\mathbb{Q}}\left[e^{-\int_t^u r(s)\ ds}\middle|\mathcal{F}(t)\right] E^{\mathbb{Q}}\left[dB(u)|\,\mathcal{F}(t)\right] \\ &= \int_t^T e^{-\int_t^u f(t,s)\ ds} E^{\mathbb{Q}}\left[dB(u)|\,\mathcal{F}(t)\right], \end{aligned}$$

by the definition of the forward rate $f(t, s)$. Notice that now we have two terms both not depending on one another. In other words, $E^{\mathbb{Q}}\left[dB(u)|\,\mathcal{F}(t)\right]$ only depends on Z not r. When evaluating the expected dynamics under $\mathbb{Q}$ we have

$$E^{\mathbb{Q}}\left[dB(u)|\,\mathcal{F}(t)\right] = E^{\mathbb{Q}}\left[dB(u)|\,Z(t)\right] = \sum_{i\in E} I^i(t) E^{\mathbb{Q}}\left[dB(u)|\,Z(t) = i\right]$$

with

$$\begin{aligned} E^{\mathbb{Q}}\left[dB(u)|\,Z(t) = i\right] &= E^{\mathbb{Q}}\left[\sum_{k\in E} I^k(u)\ dB^k(u) + \sum_{k\in E}\sum_{j\in E:j\neq k} b^{kj}(u)\ dN_{jk}(u)\middle| Z(t) = i\right] \\ &= \sum_{k\in E} E^{\mathbb{Q}}\left[I^k(u)\middle|\,Z(t) = i\right]\ dB^i(u) + \sum_{k\in E}\sum_{j\in E:j\neq k} b^{kj}(u)\ E^{\mathbb{Q}}\left[dN_{jk}(u)|\,Z(t) = i\right] \\ &= \sum_{k\in E} q_{ik}(t,u)\ dB^i(u) + \sum_{k\in E}\sum_{j\in E:j\neq k} b^{kj}(u)\ E^{\mathbb{Q}}\left[p_{ik}(t,u)\mu_{kj}(u)|\,Z(t) = i\right]\ du \\ &= \sum_{k\in E} q_{ik}(t,u)\ dB^i(u) + \sum_{k\in E}\sum_{j\in E:j\neq k} E^{\mathbb{Q}}\left[p_{ik}(t,u)b^{kj}(u)\mu_{kj}(u)\middle|\,Z(t) = i\right]\ du. \end{aligned}$$

4.4.4 Stochastic mortality rates

4.5 Matrix methods in life insurance

4.5.1 Basic setup

We consider the time-inhomogeneous Markov jump process $X = \{X(t)\}_{t\geqslant 0}$ with a finite state–space E and intensity matrix $\mathbf{\Lambda}(t) = \{\lambda_{ij}(t)\}_{i,j\in E}$. We then define the payment process as

$$dB(t) = \sum_{i \in E} 1_{X(t-)=i} \left(b_i(t)\ dt + \sum_{j \in E} b_{ij}(t)\ dN_{ij}(t) \right),$$

with $b_i(t)$ are continuous payment rates (negative if premiums) and $b_{ij}(t)$ lump sum payments, which occur according to the counting measure $N_{ij}(t)$. The intensity matrix is decomposed into

$$\mathbf{\Lambda}(t) = \mathbf{\Lambda}^0(t) + \mathbf{\Lambda}^1(t),$$

where $\mathbf{\Lambda}^1(t)$ is a non–negative matrix and, consequently, $\mathbf{\Lambda}^0(t)$ a sub–intensity matrix, i.e. row sums are non–positive. We choose this decomposition in a way such that $\mathbf{\Lambda}^1(t)$ contains the intensities with a factor $l_{ij}(t)$ being the probability upon jump from i to j at time t of recieving the payment $b_{ij}(t)$. In other words, we have the decomposition

$$\lambda_{ij}(t) = \lambda^0_{ij}(t) + \lambda^1_{ij}(t) = l_{ij}(t)\lambda^0_{ij}(t) + (1 - l_{ij}(t))\lambda^1_{ij}(t) \iff l_{ij}(t) = \frac{\lambda^1_{ij}(t)}{\lambda^0_{ij}(t) + \lambda^1_{ij}(t)}.$$

In the case $i = j$ then the counting measure N_{ii} denotes a poisson arrival process for lump sum payments arriving during the visit in the state i. This in particular means that the payment $b_{ii}(t)$ will be triggered in $[t, t + dt)$ with probability $\lambda^1_{ii}\ dt$ given that $X(t-) = i$.

Finally, we assume that the spot interest rates in state i follow a deterministic function $r_i(t)$. Hence the interest rates follow the model

$$r(u) = r_{X(u)}(u).$$

Recall that the reserve is defined as

$$V(t) = \mathbb{E}^{\mathbb{Q}} \left[\int_t^T e^{-\int_t^u r(s)\ ds}\ dB(u) \middle| \mathcal{F}(t) \right].$$

We will be using the following matrix functions when computing the reserve and higher order moments.

$$\begin{aligned} \mathbf{B}(t) &= \{b_{ij}(t)\}_{i,j \in E}, \\ \mathbf{R}(t) &= \mathbf{\Lambda}^1(t) \bullet \mathbf{B}(t) + \mathbf{\Delta}(\mathbf{b}(t)), \\ \mathbf{C}^{(k)}(t) &= \mathbf{\Lambda}^1(t) \bullet \mathbf{B}^{\bullet k}(t), \qquad k \geqslant 2, \end{aligned}$$

where $\bullet$ denotes the Schur product simply defined as $\mathbf{A} \bullet \mathbf{B} = \{a_{ij}b_{ij}\}$ with $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$. The notation $\mathbf{\Delta}(\cdot)$ operator sets $\cdot$ as the diagonal i.e.

$$\mathbf{\Delta}(\mathbf{b}(t)) = \text{diag}(b_1(t), ..., b_p(t)),$$

assuming $E = \{1, ..., p\}$. We call $\mathbf{B}(t)$ the transition payments, $\mathbf{R}(t)$ the rewards and $\mathbf{C}^{(k)}(t)$ the kth order contributions. The matrix $\mathbf{C}^{(k)}(t)$ is mostly usefull notationally rather than intuitively.

4.5.2 Interest rate free analysis

In this section, we consider the risk free context where we are not concerned with any discounting of expected payments. This make us introduce the following random process

$$U^0(s,t) = \sum_{i \in E} \int_s^t b_i(u) 1_{X(u-)=i}\ du + \sum_{i,j \in E} \int_s^t b_{ij}(u)\ dN_{ij}(u),$$

giving the total reward obtained in the time interval $[s, t]$. We are interested in higher order moments of this quantity (representing the rate free reserve). To this goal we introduce the conditional moments with

$$m^{(k)}_{ij}(s,t) = \mathbb{E}\left[1_{X(t)=j} U^0(s,t)^k \middle| X(s) = i \right],$$

representing the weighted expected moment upon start in state i weighted with the probability with end in state j. This forms the matrix

$$\mathbf{m}^{(k)}(s,t) = \left\{ m_{ij}^{(k)}(s,t) \right\}_{i,j \in E}.$$

For simplicity we simply for $k = 1$ write $\mathbf{m}^{(1)}(s,t) = \mathbf{m}(s,t)$.

Theorem 5.1. (Bladt) *With the above assumptions it holds that*

$$\mathbf{m}(s,t) = \int_s^t \mathbf{P}(s,u)\mathbf{R}(u)\mathbf{P}(u,t)\ du.$$

In particular, by Van Loans formula we have

$$\prod_s^t \left(\mathbf{I} + \begin{pmatrix} \mathbf{\Lambda}(x) & \mathbf{R}(t) \\ \mathbf{0} & \mathbf{\Lambda}(x) \end{pmatrix} dx \right) = \begin{pmatrix} \mathbf{P}(s,t) & \mathbf{m}(s,t) \\ \mathbf{0} & \mathbf{P}(s,t). \end{pmatrix}$$

Proof.

4.5.3 Transform of rewards and higher order moments

Recall that for a random variable X the moment generating function $M(\theta)$ given by

$$M(\theta) = \mathbb{E}[e^{\theta X}],$$

gives us a comprehensive insight into the moments of X. Take the taylor approximation of $e^{\theta X}$:

$$e^{\theta X} = \sum_{n=0}^{\infty} \frac{(\theta X)^n}{n!}$$

Hence taking expectation gives

$$M_X(\theta) = \mathbb{E}[e^{\theta X}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(\theta X)^n}{n!} \right] = \sum_{n=0}^{\infty} \frac{\theta^n \mathbb{E}[X^n]}{n!}.$$

Now differentiating wrt. θ yields the important result

$$\frac{\partial M_X(\theta)}{\partial \theta}(\theta) = \sum_{n=1}^{\infty} \frac{n\theta^{n-1}\mathbb{E}[X^n]}{n!} = \sum_{n=1}^{\infty} \frac{\theta^{n-1}\mathbb{E}[X^n]}{(n-1)!} = \mathbb{E}[X] + \sum_{n=1}^{\infty} \frac{\theta^n \mathbb{E}[X^{n+1}]}{n!},$$

hence by evaluating in $\theta = 0$ we get $M_X'(0) = \mathbb{E}[X]$. In general we have

$$\frac{\partial^n M_X(\theta)}{\partial \theta^n}(\theta)\Big|_{\theta=0} = \mathbb{E}[X^n], \qquad n \geqslant 1.$$

This is really the motivation for examining the moment generating function of the reward obtained in the interval $[s,t]$ i.e. the quantity $U^0(s,t)$. To this we define

$$\hat{F}_{ij}(\theta; s, t) = \mathbb{E}\left[e^{\theta U^0(s,t)} 1\{X(t) = j\} \;\middle|\; X(s) = i \right],$$

for $i, j = 1, ..., p$. Obviously we have the following identity

$$\sum_{j=1}^{p} \hat{F}_{ij}(\theta; s, t) = \mathbb{E}\left[e^{\theta U^0(s,t)} \;\middle|\; X(s) = i \right].$$

Furthermore, if we define the distribution of $X(t), U^0(s,t)$ given $X(s)$ by

$$F_{ij}(x; s, t) = \mathbb{P}(X(t) = j, U^0(s,t) \leqslant x \mid X(s) = i)$$

then we have the integral decomposition

$$\hat{F}_{ij}(\theta; s, t) = \int_{\mathbb{R}} e^{\theta x}\ dF_{ij}(x; s, t).$$

The following theorem gives a way of calculating the object $\hat{\mathbf{F}}(\theta; s, t) = \{\hat{F}_{ij}(\theta; s, t)\}$.

Theorem 5.5. (Bladt) *Let* $\hat{\mathbf{F}}(\theta;s,t) = \{\hat{F}_{ij}(\theta;s,t)\}_{i,j=1,\dots,p}$ *and*

$$\mathbf{A}(\theta;u) = \mathbf{\Lambda}^1(u) \bullet \left\{e^{\theta b_{kl}(u)}\right\}_{k,l} + \mathbf{\Lambda}^0(u) + \theta\mathbf{\Delta}(\mathbf{b}(u))$$

where $\bullet$ *denotes the Schur matrix product. Then the moment generating function of* $U^0(s,t)$ *is given by*

$$\hat{\mathbf{F}}(\theta;s,t) = \prod_s^t (\mathbf{I} + \mathbf{A}(\theta;u)\ du).$$

Proof.

By the above theorem we see that

$$\mathbb{E}\left[e^{\theta U^0(s,t)} \ \middle|\ X(s) = i\right] = e_i^\top \hat{\mathbf{F}}(\theta;s,t)e = e_i^\top \prod_s^t (\mathbf{I} + \mathbf{A}(\theta;u)\ du)e,$$

in particular

$$\prod_s^t (\mathbf{I} + \mathbf{A}(\theta;u)\ du)e = \begin{pmatrix} \mathbb{E}\left[e^{\theta U^0(s,t)} \ \middle|\ X(s) = 1\right] \\ \mathbb{E}\left[e^{\theta U^0(s,t)} \ \middle|\ X(s) = 2\right] \\ \vdots \\ \mathbb{E}\left[e^{\theta U^0(s,t)} \ \middle|\ X(s) = p\right] \end{pmatrix}.$$

Furthermore, we have from the property above that

$$\frac{\partial^k}{\partial\theta^k}\left(e_i^\top \prod_s^t (\mathbf{I} + \mathbf{A}(\theta;u)\ du)e\right)\Big|_{\theta=0} = m_i^{(k)}(s,t).$$

Lemma 5.6. (Bladt) *Given the definitions in theorem 5.5 we have*

$$\begin{aligned} \frac{\partial^0}{\partial\theta^0}\mathbf{A}(\theta;s)\Big|_{\theta=0} &= \mathbf{A}(\theta;s)\Big|_{\theta=0} = \mathbf{\Lambda}^1(s) + \mathbf{\Lambda}^0(s) = \mathbf{\Lambda}(s), \\ \frac{\partial}{\partial\theta}\mathbf{A}(\theta;s)\Big|_{\theta=0} &= \mathbf{\Lambda}^1(s) \bullet \mathbf{B}(s) + \mathbf{\Delta}(\mathbf{b}(s)) = \mathbf{R}(s), \\ \frac{\partial^k}{\partial\theta^k}\mathbf{A}(\theta;s)\Big|_{\theta=0} &= \mathbf{\Lambda}^1(s) \bullet \mathbf{B}^{\bullet k}B(s) = \mathbf{C}^{(k)}(s), \qquad k > 1. \end{aligned}$$

Recall that we defined the moments in general by

$$\mathbf{m}^{(k)}(s,t) = \left\{\mathbb{E}\left[1\{X(t) = j\}U^0(s,t)^k \ \middle|\ X(s) = i\right]\right\}_{i,j=1,\dots,p} = \left\{m_{ij}^{(k)}(s,t)\right\}_{i,j=1,\dots,p},$$

hence we can define (for ease of notation) the reduced moments by

$$\mathbf{m}_r^{(k)}(s,t) = \left\{\frac{m_{ij}^{(k)}(s,t)}{k!}\right\}_{i,j=1,\dots,p} = \frac{1}{k!}\mathbf{m}^{(k)}(s,t).$$

Furthermore, we define the reduced contributions

$$\mathbf{C}_r^{(k)}(s,t) = \frac{1}{k!}\mathbf{C}^{(k)}(s,t) = \frac{1}{k!}\mathbf{\Lambda}^1(t) \bullet \mathbf{B}^{\bullet k}(t).$$

Theorem 5.7. (Bladt) *Given the definitions above we have*

$$\mathbf{m}^{(k)}(s,t) = \int_s^t \mathbf{P}(s,x)\mathbf{R}(x)\mathbf{m}_r^{(k-1)}(x,t)\ dx + \sum_{m=2}^k \int_s^t \mathbf{P}(s,x)\mathbf{C}_r^{(m)}(x)\mathbf{m}_r^{(k-m)}(x,t)\ dx.$$

Using this result we can make a powerful statement regarding how one could calculate all of the kth moments at once. To this we have the result.

Theorem 5.8. (Bladt) *Let* $\mathbf{F}^{(k)}(x)$ *be given by*

$$\mathbf{F}^{(k)}(x) = \begin{bmatrix} \mathbf{\Lambda}(x) & \mathbf{R}(x) & \mathbf{C}_r^{(2)}(x) & \cdots & \mathbf{C}_r^{(k-1)}(x) & \mathbf{C}_r^{(k)}(x) \\ \mathbf{0} & \mathbf{\Lambda}(x) & \mathbf{R}(x) & \cdots & \mathbf{C}_r^{(k-2)}(x) & \mathbf{C}_r^{(k-1)}(x) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Lambda}(x) & \mathbf{R}(x) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Lambda}(x) \end{bmatrix},$$

and $\mathbf{H}^{(k)}(s,t)$ *defined by*

$$\mathbf{H}^{(k)}(x) = \begin{bmatrix} \mathbf{P}(s,t) & \mathbf{m}_r^{(1)}(s,t) & \mathbf{m}_r^{(2)}(s,t) & \cdots & \mathbf{m}_r^{(k-1)}(s,t) & \mathbf{m}_r^{(k)}(s,t) \\ \mathbf{0} & \mathbf{P}(s,t) & \mathbf{m}_r^{(1)}(s,t) & \cdots & \mathbf{m}_r^{(k-2)}(s,t) & \mathbf{m}_r^{(k-1)}(s,t) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}(s,t) & \mathbf{m}_r^{(1)}(s,t) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}(s,t) \end{bmatrix}.$$

Then we have the result

$$\prod_s^t (\mathbf{I} + \mathbf{F}^{(k)}(x)\, dx) = \mathbf{H}^{(k)}(s,t).$$

For ease of notation we define the Toeplitz matrix as

$$\mathcal{T}(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_n) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \cdots & \mathbf{A}_{n-1} & \mathbf{A}_n \\ \mathbf{0} & \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{n-2} & \mathbf{A}_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_1 \end{bmatrix}.$$

And hence we could write:

$$\mathbf{F}^{(k)}(x) = \mathcal{T}(\mathbf{\Lambda}(x), \mathbf{R}(x), \mathbf{C}_r^{(2)}(x), ..., \mathbf{C}_r^{(k-1)}(x), \mathbf{C}_r^{(k)}(x)),$$

and

$$\mathbf{H}^{(k)}(x) = \mathcal{T}(\mathbf{P}(s,t), \mathbf{m}_r^{(1)}(s,t), \mathbf{m}_r^{(2)}(s,t), ..., \mathbf{m}_r^{(k-1)}(s,t), \mathbf{m}_r^{(k)}(s,t)).$$

4.5.4 Markovian interest rates

We now study how we may calculate the moments of the reserves in the case with rates. To this we start by assuming the Markov-jump interest model in sync with the Markov-jump process for the policy holder. Obviously, we in practice would assume that these two processes are independent, but for generality we will not insist on this for now. We therefore define the process

$$\mathbf{r}(t) = \{r_i(t)\}_{i \in E_r}$$

and the two Markov jump processes

$$X_r = \{X_r(t)\}_{t \geqslant 0} \in E_r = \{1, ..., p\}, \qquad X_b(t) = \{X_b(t)\}_{t \geqslant 0} \in E_b = \{1, ..., q\},$$

and naturally we at time t have the interest rate $r_{X_r(t)}(t)$ and $X_v(t)$ governs the state of the policy holder. Using this we define the combined Markov jump process as

$$X(t) = \{X_r(t), X_b(t)\}_{t \geqslant 0} \in E_r \times E_b.$$

The processes X_b and X_r may or may not be independent, and the payment processes likewise may or may not be independent of X_r. In the independent case the processes X_b and X_r are defined on each their state–space, and the common state–space will be the product set of the two. If the processes are sharing

states, with the possibility of having simultaneous jumps, then we obtain dependency of the processes. Such a case could, e.g. be a rise in the interest rate causing an increased intensity of jumping to surrender or free–policy states.

In the case where we have independence we have that the transition intensities of X is given by

$$\mathbf{\Lambda}(t) = \mathbf{\Lambda}_b(t) \oplus \mathbf{\Lambda}_r(t) = \mathbf{\Lambda}_b(t) \otimes \mathbf{I}_p + \mathbf{I}_p \otimes \mathbf{\Lambda}_r(t) = \{\lambda_{ij}\}_{i,j=1,\dots,pq},$$

where $\oplus$ is the Kronecker sum and $\otimes$ is the Kronecker product given by

$$\mathbf{A} \otimes \mathbf{B} = \{a_{ij}\mathbf{B}\} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & \cdots & a_{n,n}\mathbf{B} \end{bmatrix}.$$

4.5.5 Reserves

We now consider the valuation of the payment process B. Introduce the matrix of partial state–wise prospective reserves,

$$\begin{aligned} \mathbf{V}(s,t) &= \{V_{ij}(s,t)\}_{i,j\in E_b}, \\ V_{ij}(s,t) &= \mathbb{E}\left[1\{X(t)=j\} \int_s^t e^{-\int_s^x r_{X(u)}(u)\ du}\ dB(x) \;\middle|\; X(s)=i \right]. \end{aligned}$$

From the Markov-jump representation we recall that we defines $d_{ij}(s,t)$ by

$$d_{ij}(s,t) = \mathbb{E}\left[1\{X(t)=j\} \exp\left\{ -\int_s^t r_{X(u)}(u)\ du \right\} \;\middle|\; X(s)=i \right],$$

and that the matrix $\mathbf{D}(s,t) = \{d_{ij}(s,t)\}$ has representation

$$\mathbf{D}(s,t) = \prod_s^t \Big(\mathbf{I} + [\mathbf{\Lambda}(u) - \mathbf{\Delta}(r(u))]\ du \Big).$$

Using this we have the theorem.

Theorem 5.10. (Bladt) *The matrix of partial state-wise prospective reserves $\mathbf{V}(s,t)$ has the following integral representation*

$$\mathbf{V}(s,t) = \int_s^t \mathbf{D}(s,x)\mathbf{R}(x)\mathbf{P}(x,t)\ dx.$$

Using Van loans formula we can avoid the integration and calculate the product integral from below.

Corollary 5.11. (Bladt) *The matrix of partial state-wise prospective reserves $\mathbf{V}(s,t)$ has the following product integral representation*

$$\prod_s^t \left(\mathbf{I} + \begin{pmatrix} \mathbf{\Lambda}(u) - \mathbf{\Delta}(r(u)) & \mathbf{R}(u) \\ \mathbf{0} & \mathbf{\Lambda}(u) \end{pmatrix} \right) = \begin{pmatrix} \mathbf{D}(s,t) & \mathbf{V}(s,t) \\ \mathbf{0} & \mathbf{P}(s,t) \end{pmatrix}.$$

Finally, we state and prove Thiele's differential equations for partial reserves with stochastic interest rates.

Theorem 5.12. (Bladt) The statewise reserves satisfies the differential equations

$$\frac{\partial}{\partial s}\mathbf{V}(s,t) = \int_s^t \mathbf{D}(s,x) - \mathbf{R}(s)\mathbf{P}(s,t)$$

4.5.6 Higher order moments

4.5.7 Equivalence premium

4.5.8 Distributions of future payments

4.6 Financial Mathematics in Life Insurance

4.6.1 Background and Simple Claims

4.6.2 Payment Streams

4.6.3 Unit-Link Insurance

4.6.4 With-Profit Insurance and the Dynamics of the Surplus

4.6.5 Cash Dividends and Market Reserve

4.6.6 The Pure Case of Cash Dividends

4.6.7 Bonus Payments and Market Reserve

4.6.8 The Pure Case of Bonus Payments

4.6.9 Comparison of Products

4.7 Special Studies in Life Insurance

4.7.1 Survival Probabilities and Forward Mortality Rates

4.7.2 Dependent Interest and Mortality Rates

4.7.3 Stochastic Interest and Mortality Rate Models

4.7.4 Reserves Revisited

4.7.5 Incidental Policy Holder Behavior

Chapter 5

Continuous Time Finance

This topic revolves around the theory of the Brownian motion and martingale processes. Other main topics are the binomial model and an introduction to financial derivatives. Financial derivatives is contingent on the outcome of a stochastic process at some future time $t = T$ and often is a function Φ of some assets price S_t. As such the derivative will give a stochastic payout, at time $t = T$ of the size $X_T = \Phi(S_T)$. Naturally we want to say something about the *fair* price of the derivative in the form of

$$\Pi_t(X_T) = \mathbb{E}\left[\Phi(S_T) \mid \mathcal{F}_t\right],$$

where $\mathcal{F}_t \subset \mathcal{F}$ is the available information at time t. We will by defualt intepret the times $t = 0$ as *today* and $t = T$ as *tomorrow*. This indeed require some fundamental understanding of the behaviour of the asset price S_t. This lead us over to discussing the process in center of the *Black-Scholes* model: the Brownian motion.

5.1 Discrete time models

5.1.1 One-period time models

The study of this course is the **European call** option (and *put* option). This financial derivative is an agreement between two parties where the holder of the option has the right to *"exercise"* the derivative, at a future time $t = T$. Exercising means buying an asset at a certain agreed opon price-strike K. In the case of the put-option: the holder has the right (but not obligation) to sell the asset at the strike price K. As such the derivative has the payoff

$$\text{Call option:} \quad \Phi(S_T) = (S_T - K)^+, \qquad \text{Put option:} \quad \Phi(S_T) = (K - S_T)^+.$$

Our objective is to understand when an arbitrage exist and to find the fair price of these derivative. The strategy in pricing is finding a replicating portfolio with the same payoff as the option (with probability one) and then price the derivative accordingly.

5.1.1.1 Model description

In the one-period model we consider the simplest possible market. We have two distinct times $t = 0$ (today) and $t = 1$ (tomorrow) and we may buy any portfolio as a mixture of bonds and one stock. We denote the bonds price by B_t and the stocks price by S_t and we assume the following:

$$B_0 = 1,\ B_1 = 1 + R, \qquad S_0 = s,\ S_1 = \begin{cases} s \cdot u, & \textit{with probability } p_u. \\ s \cdot d, & \textit{with probability } p_d. \end{cases}$$

We may introduce Z as the random variable

$$Z = u \cdot (I) + d \cdot (1 - I),$$

for an bernoulli variable I with succes probability p_u. Naturally, we assume $d \leqslant (1+R) \leqslant u$ (this is imperative to ensure no arbitrage as we will see).

5.1.1.2 Portfolios and arbirtage

We study any portfolio on the (B,S) market as a vector $h=(x,y)$ where x is the amount of bonds and y is the amount of stock held in the portfolio. Notice that we allow for shorting, that is $x<0$ or $y<0$. As such, we have that $h \in \mathbb{R}^2$. In this we have made some unrealistic, but attractable assumptions included in the assumptions:

- We allow short positions and fractional holding, i.e. $h \in \mathbb{R}^2$,
- We assume no spread between ask and bids,
- No transaction costs and
- A completely liquid market i.e. we may borrow and buy as much stock and bonds as wanted.

Given that we have chosen a portfolio h we may introduce the value process.

Definition 2.1. (Bjork) *The **value process** of the porfolio $h \in \mathbb{R}^2$ is the stochastic process*

$$V_t^h = xB_t + yS_t,\ t=0,1.$$

Given this notation we may define what an arbitrage is.

Definition 2.2. (Bjork) *An **arbitrage** is a portfolio h with the properties: 1) $V_0^h = 0$, 2) $P(V_1^h \geqslant 0) = 1$ and 3) $P(V_1^h > 0) > 0$.*

That is h is an deterministic money-machine where we at least never loose any money. Granted the bonds give a determinictic non-negative return, but an arbitrage does not require any money out of pocket. With the notion of an arbitrage we will show the first proposition regarding the choice of R, u, d as defined above.

Proposition 2.3. (Bjork) *The one-period binomial model is arbitrage free if and only if the following inequality hold:*

$$d \leqslant (1+R) \leqslant u. \tag{2.1}$$

Proof.

The statement is proofed by contradiction. Assume that $d > 1+R$ holds. Then by definition $u > d > 1+R$. Notice that any portfolio satisfying $V_0^h = 0$ must satisfy

$$0 = xB_0 + yS_0 = x + ys \iff x = -ys$$

That is for some choice y the only arbitrage candidate is the portfolio $h=(-ys, y)$. Calculating the value at time $t=1$ we have

$$V_1^h = -ys \cdot (1+R) + y \cdot s \cdot Z = ys(Z-1-R)$$

However since $Z \geqslant d$ we have $Z-(1+R) \geqslant 0$ and therefore an arbitrage (for $y>0$). The other inequality $1+R>u$ follows analog steps. Simply choose some $y<0$ and the result follows. ■

From inequality (2.1) we see that since $1+R$ is between u and d we may find a pair $q_d, q_u \geqslant 0$ with $q_d + q_u = 1$ such that

$$1 + R = q_u \cdot u + q_d \cdot d.$$

This yields the important risk neutral valuation formula as summed op in the following definition

Definition 2.4. (Bjork) *A probability measure Q is called a **martingale meausre** if the following condition holds:*

$$S_0 = \frac{1}{1+R} E^Q[S_1].$$

The above measure Q is the measure $Q(Z = d) = q_d$ and $Q(Z = u) = q_u$ for the binomial model. This does in fact yield the risk neautral valuation formula:

$$\begin{aligned} S_0 &= \frac{1}{1+R} E^Q[S_1] = \frac{1}{1+R}(Q(Z = d) \cdot d \cdot s + Q(Z = u) \cdot u \cdot s) \\ &= s\frac{1}{1+R}(q_d \cdot d + q_u \cdot u) = s, \end{aligned}$$

where we simply use $1 + R = q_d \cdot d + q_u \cdot u$. We call this the risk neautral valuation formula because it in some sense gives an expected discounted value of the future stock price. We end this endavour with reformulating the arbitrage proposition and determining the values of the Q-measure.

Proposition 2.5. (Bjork) *The one-period binomial model is arbitrage free if and only if there exists a martingale measure Q.*

Proposition 2.6. (Bjork) *The one-period binomial model has martingale probabilities given by:*

$$\begin{cases} q_u = \frac{(1+R)-d}{u-d}, \\ q_u = \frac{u-(1+R)}{u-d}. \end{cases}$$

5.1.1.3 Contingent Claims

This chapter revolves around the financial derivative and we start by stating the definition of the financial derivative.

Definition 2.7. (Bjork) *A **contingent claim** (financial derivative) is* any *stochastic variable X of the form $\Phi(Z)$, where Z is the stochastic varible driving the stock price process.*

We may also call the function Φ the **contract function** as it states how the contract is resolved once the stochastic variable Z has been realised. Our objective is now to study, what a buyer of said contract would have to pay at any given time t. We call the fair price of X at time t: $\Pi_t[X]$. As such it is easy to see that the fair price at the time of maturity T is simply the payout X i.e. $\Pi_T[X] = X$. Our strategy is to find a replicating portfolio h and determine the price of said portfolio.

Definition 2.8. (Bjork) *A contingent claim X can be **replicated**, or said to be **reachable** if there exist a portfolio h such that*

$$V_1^h = X,$$

*with probability one. In that case, we say that the portfolio h is a **hedging** portfolio or a **replicating** portfolio. If all claims can be replicated we say that the market is **complete**.*

Our pricing strategy is then to determine the value process of the replicating portfolio and then by the first pricing principle below we say that the price is imply the value of the replicating portfolio.

Pricing principle 1. If a clain X is reachable with replicating portfolio h, then the only reasonable price process for X is given by

$$\Pi_t[X] = V_t^h.$$

Notice, that this assumes that a replicating portfolio exist and even so we have a uniqueness statement to solve. We end this section by writing two important results.

Proposition 2.9. (Bjork) *Suppose that a claim X is reachable with replicating portfolio h. Then any price at time $t \geqslant 0$ of the claim X other than the value process of h will lead to an arbitrage on the extended market (B, S, X).*

Proposition 2.10. (Bjork) *If the one-period binomial model is free of arbitrage, then it is also complete.*

The hedging portfolio in the one-period binomial model is given by the portfolio (x, y) below

$$x = \frac{1}{1+R} \cdot \frac{u\Phi(d) - d\Phi(u)}{u-d}, \tag{2.2}$$

$$y = \frac{1}{s} \cdot \frac{\Phi(u) - \Phi(d)}{u-d}. \tag{2.3}$$

5.1.1.4 Risk Neutral Valuation

We see that since the one-period model is complete we can price any contingent claim and we see that

$$\begin{aligned}
\Pi_0[X] &= \frac{1}{1+R} \cdot \frac{u\Phi(d) - d\Phi(u)}{u-d} + s\frac{1}{s} \cdot \frac{\Phi(u) - \Phi(d)}{u-d} \\
&= \frac{1}{1+R} \left\{ \frac{u\Phi(d) - d\Phi(u)}{u-d} + (1+R)\frac{\Phi(u) - \Phi(d)}{u-d} \right\} \\
&= \frac{1}{1+R} \left\{ \frac{(1+R)-d}{u-d}\Phi(u) + \frac{u-(1+R)}{u-d}\Phi(d) \right\} \\
&= \frac{1}{1+R} E^Q[X].
\end{aligned}$$

i.e. the price at time $t = 0$ should simply be the expected discounted payout according to the martingale measure. This leads to the important pricing proposition:

Proposition 2.11. (Bjork) *If the one-period binomial model is free of arbitrage, then the arbitrage free price of a contingent claim X is given by*

$$\Pi_0[X] = \frac{1}{1+R} E^Q[X]. \tag{2.4}$$

Here the martingale measure Q is uniquely determined by the relation

$$S_0 = \frac{1}{1+R} E^Q[S_1], \tag{2.5}$$

and the explicit expressions for q_u and q_d are given in proposition 2.6. Furthermore the claim X can be replicated using the portfolio

$$x = \frac{1}{1+R} \cdot \frac{u\Phi(d) - d\Phi(u)}{u-d}, \tag{2.6}$$

$$y = \frac{1}{s} \cdot \frac{\Phi(u) - \Phi(d)}{u-d}. \tag{2.7}$$

5.1.2 Multi-period model

The one-period binomial model can easily be extended to a multi-period model, by assuming that the bond and stock pricess evolve by the processes:

$$t \geqslant 1: \ B_t = (1+R)B_{t-1} \qquad \text{and} \qquad B_0 = 1,$$

$$t \geqslant 1: \ S_t = Z_{t-1}S_{t-1} \qquad \text{and} \qquad S_0 = s,$$

where we obviously have that $B_t = (1+R)^t$ for $t \geqslant 0$. In the above Z_t is u with probability p_u and d with probability p_d. In this context, we need to define a portfolio in terms of a strategy.

Definition 2.13. (Bjork) *A **portfolio strategy** is a stochastic process on* $\{1, ..., T\}$

$$h = \{h_t = (x_t, y_t); \ t = 1, ..., T\}$$

such that h_t *is a function of* $S_0, S_1, ..., S_{t-1}$*. For a given portfolio strategy* h *we set* $h_0 = h_1$ *by convention. The associated **value process** corresponding to the portfolio* h *is defined by*

$$V_t^h = x_t(1+R) + y_t S_t.$$

Given this notation we may define what an arbitrage is, but first we introduce the notion of a self-financing portfolio. A self-financing portfolio in an intuative sense is a portfolio that is not withdrawn from or deposited into.

Definition 2.14. (Bjork) *A portfolio strategy* h *is said to be **self-financing** if the following condition holds for all* $t = 0, ..., T-1$*:*

$$x_t(1+R) + y_t S_t = x_{t+1} + y_{t+1} S_t.$$

The above equation says that the portfolio purchased at time t and helt until $t+1$ (x_{t+1}, y_{t+1}) can only be financed by the market value of the portfolio held from $[t-1, t)$ i.e. (x_t, y_t). We now define an arbitrage.

Definition 2.15. (Bjork) *An **arbitrage** is a self-financing portfolio* h *with the properties: 1)* $V_0^h = 0$*, 2)* $P(V_T^h \geqslant 0) = 1$ *and 3)* $P(V_T^h > 0) > 0$*.*

The multiperiod binomial model has an just like the oneperiod model a result regarding when an arbitrage exists.

Lemma 2.16. (Bjork) *If* $d \leqslant (1+R) \leqslant u$ *(eq. 2.8) then the multiperiod model is arbitrage-free.*

As one can see, the multiperiod model is rather similar to the one period model. We wil in the following summarise equivalent statements for the multiperiod model as the ones in the oneperiod model.

Definition 2.17. (Bjork) *The martingale probabilities* q_u *and* q_d *are defined as the probabilities for which the relation below holds.*

$$s = \frac{1}{1+R} E^Q[S_{t+1} \mid S_t].$$

Proposition 2.18. (Bjork) *The martingale probabilities* q_u *and* q_d *are given by*

$$\begin{cases} q_u = \frac{(1+R)-d}{u-d}, \\ q_u = \frac{u-(1+R)}{u-d}. \end{cases}$$

Definition 2.19. (Bjork) *A **contingent claim** is a stochastic variable* X *of the form*

$$X = \Phi(S_T),$$

*where the **contract function** Φ is some given real valued function.*

Definition 2.20. (Bjork) *A given contingent claim* X *is said to be **reachable** if there exists a self-financing portfolio* h *such that*

$$V_T^h = X,$$

*with probability one. In that case we say that the portfolio h is a **hedging** portfolio or a **replicating** portfolio. If all claims can be replicated we say that the market is* (dynamically) ***complete**.*

Pricing principle 2. (Bjork) *If a claim X is reachable with replicating portfolio h, then the only reasonable price process for X os given by*

$$\Pi_t[X] = V_t^h,\ t = 0, 1, ..., T.$$

Proposition 2.21. (Bjork) *Assume X is reachable by h, then any price other than V_t^h for some $t \geqslant 0$ leads to an arbitrage opportunity.*

Proposition 2.22. (Bjork) *The multiperiod model is complete, i.e. every claim can be replicated by a self-financing portfolio.*

Proposition 2.24. (Bjork) ***(Binomial algorithm)** Consider a T-claim $X = \Phi(S_T)$. Then this claim can be replicated using af self-financing portfolio. If $V_t(k)$ denotes the value of the portfolio at the node (t, k) (k referring to k amount of up-moves for the stock), then $V_t(k)$ can be computed recursively by the scheme*

$$\begin{cases} V_t(k) = \frac{1}{1+R}\left\{q_u V_{t+1}(k+1) + q_d V_{t+1}(k)\right\}, \\ V_T(k) = \Phi(su^k d^{T-k}). \end{cases}$$

where the martingale probabilities q_u and q_d are given by

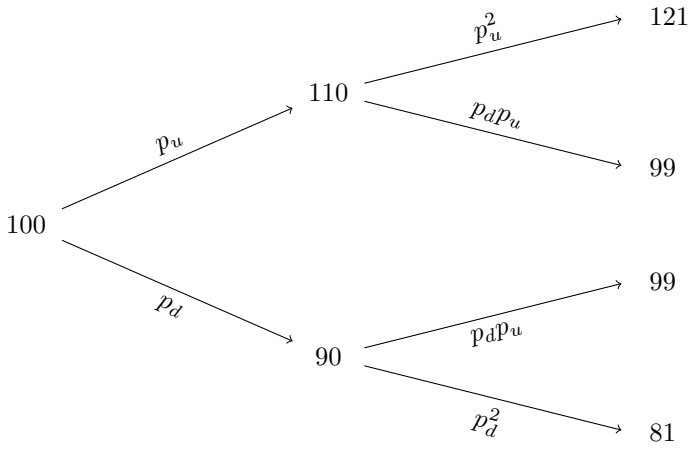
$$\begin{cases} q_u = \frac{(1+R)-d}{u-d}, \\ q_u = \frac{u-(1+R)}{u-d}. \end{cases}$$

With the notation as above, the hedging portfolio is given by

$$\begin{cases} x_t(k) = \frac{1}{1+R} \cdot \frac{uV_t(k) - dV_t(k+1)}{u-d}, \\ y_t(k) = \frac{1}{S_{t-1}} \cdot \frac{V_t(k+1) - V_t(k)}{u-d}. \end{cases}$$

In particular, the arbitrage free price of the claim at $t = 0$ is given by $V_0(0)$.

Example.



Consider $R = 0.04$, $s = 100$, $u = 1.1$, $d = 0.9$, $p_u = 0.6$ and $p_d = 0.4$. We consider a model of length $T = 2$ and we want to evaluate the price of the european call option with srike $K = 90$ that is the contingent claim

$$X = (S_T - K)^+, \qquad \Phi(s) = (s - K)^+.$$

For each time t we know the replicating portfolio, if we know the payoff the following period. Therefore we start from the leaves of the tree and work towards the root. Since the strike price is $K = 90$ the end result

will be the following payoffs:

$$\begin{aligned} u^2: &\qquad (121-90)^+ = 31 \\ ud: &\qquad (99-90)^+ = 9 \\ du: &\qquad (99-90)^+ = 9 \\ d^2: &\qquad (81-90)^+ = 0 \end{aligned}$$

Therefore by the risk neautral valuation formula with $q_u = \frac{(1+R)-d}{u-d} = 0.7$ and $q_d = \frac{u-(1+R)}{u-d} = 0.3$ we have that the cost of the replicating portfolio at time $t = 1$ is respectively

$$\begin{aligned} u: &\qquad \frac{1}{1+R}\{31 \cdot q_u + 9 \cdot q_d\} \approx 23.46 \\ d: &\qquad \frac{1}{1+R}\{9 \cdot q_u + 0 \cdot q_d\} \approx 6.06 \end{aligned}$$

To replicate this payoff at time $t = 1$ we can use the risk neutral valuation formula once more to find the base cost of the replicating portfolio i.e. the price of X at time $t = 0$

$$\frac{1}{1+R}\{23.46 \cdot q_u + 6.06 \cdot q_d\} \approx 17.54.$$

Working from the root to the leaves we can now calculate the hedging portfolio at time $t = 0, 1$ for each path. For time $t = 0$ we calculate

$$\begin{aligned} x =& \frac{1}{1+R} \cdot \frac{u \cdot 6.06 - d \cdot 23.46}{u-d} \approx -69.46, \\ y =& \frac{1}{s} \cdot \frac{23.46 - 6.06}{u-d} \approx 0.87 \end{aligned}$$

We see by calculations that this does indeed replicate the payoff at time $t = 1$:

$$\begin{aligned} u: &\qquad V_1^h = (1+R) \cdot x + 110 \cdot y \approx 23.46, \\ d: &\qquad V_1^h = (1+R) \cdot x + 90 \cdot y \approx 6.06. \end{aligned}$$

We also see by calculation that the initial portfolio does cost the expected 17.54 as

$$x \cdot 1 + y \cdot 100 = 87 - 69.46 = 17.54.$$

Following these steps at time $t = 1$ the portfolios $(-86.54, 1)$ (for the up-scenario) and $(-38.94, 0.5)$ (for the down-scenario) would arise. Notice when calculating y one has to use the current price $S_1 = S_0 \cdot Z$ not S_0. One should also check by similar calculations as above, that these portfolios does indeed replicate the payoff of the contingent claim X. □

Proposition 2.25. (Bjork) *The arbitrage free price at $t = 0$ of a T-claim X is given by*

$$\Pi_0[X] = \frac{1}{(1+R)^T} E^Q[X]$$

where Q denotes the martingale measure, or more explicitly

$$\Pi_0[X] = \frac{1}{(1+R)^T} \sum_{k=0}^{T} \binom{T}{k} q_u^k q_d^{T-k} \Phi(s u^k d^{T-k}).$$

Example.

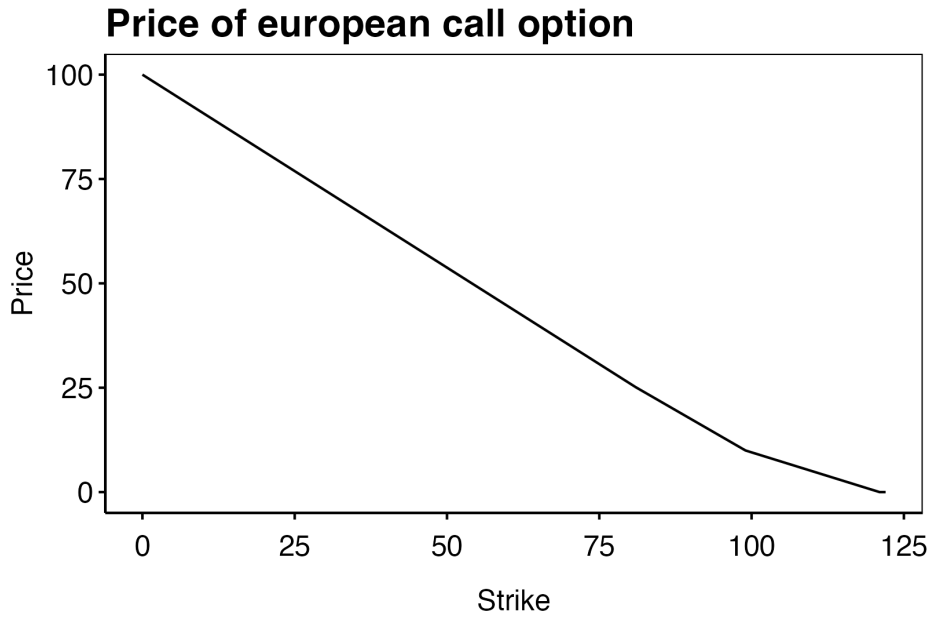


Figure 5.1: The pricing function of the European call option.

We follow an analog example as the one after proposition 2.24. Let $K = 90$ and we see that

$$\begin{aligned}
&\Pi_0[X] \\
&= \frac{1}{(1+0.04)^2} \sum_{k=0}^{2} \binom{2}{k} \cdot 0.7^k \cdot 0.3^{2-k} \cdot \Phi(100 \cdot 1.1^k \cdot 0.9^{2-k}) \\
&= 0.9245562 \cdot \left(\underbrace{1 \cdot 1 \cdot 0.09 \cdot 0}_{k=0} + \underbrace{2 \cdot 0.7 \cdot 0.3 \cdot 9}_{k=1} + \underbrace{1 \cdot 0.49 \cdot 1 \cdot 31}_{k=2} \right) \\
&= 0.9245562 \cdot (0 + 3.78 + 15.19) \\
&= 17.53883
\end{aligned}$$

Since we know that K must meaningfully range in $[0, 121]$ we could try to calculate the price of the contingent claim at time $t = 0$ for all integers in this interval. We see that the price range between S_0 and 0 as expected. One can also see that the price changes slope at the prices 99 and 121 as the function is linear in Φ and som realisations loose any effect on the price when the strike is higher than the outcome. ▫

Proposition 2.26. (Bjork) *The condition $d < (1+R) < u$ is necessary and sufficient condition for absence of arbitrage.*

5.1.3 Generelised one-period model

In the previous we had the simpel model where we only had one stochastic asset S and only one stochastic variable Z determining the future stock price. Now we will generelise this model by introducing N assets and introducing som stochastic behaviour to the system.

5.1.3.1 Model specification

We consider the market consisting of a collection of stochastic prices assets $i = 1, ..., N$ with N-dimensional price process.

$$S_t = \begin{bmatrix} S_t^1 \\ \vdots \\ S_t^N \end{bmatrix}$$

We now assume that S_t is defined on a background space with finite sample space $\Omega = \{\omega_1, ..., \omega_M\}$ with associated probabilities $p_j = P(\omega_j)$, $j = 1, ..., M$. We can then for eact time $t = 1, ..., T$ define the $N \times M$ matrix D_t as such

$$D_t = \begin{bmatrix} S_t^1(\omega_1) & \cdots & S_t^1(\omega_M) \\ \vdots & \ddots & \vdots \\ S_t^N(\omega_1) & \cdots & S_t^M(\omega_M) \end{bmatrix}.$$

We will assume that $S_0^1 > 0$ and $S_1^1(\omega_j) > 0$, $j = 1, ..., M$.

5.1.3.2 Absence of Arbitrage

We now define a **portfolio** as an N-dimensional row vector

$$h = \left[h^1, \ldots, h^N\right]$$

representing the amount of assets held at time $t = 0$ and held until $t = 1$. The **value process** is then

$$V_t^h = h \cdot S_t = \sum_{i=1}^{N} h^i S_t^i, \ t = 0, 1. \tag{3.1}$$

For a given $\omega_j \in \Omega$ we have the realisation

$$V_t^h = hS_t(\omega_j) = hd_j = (hD)_j.$$

Definition 3.1. (Bjork) *The portfolio h is an **arbitrage portfolio** fil it satisfies the conditions: $V_0^h = 0$, $P(V_1^h \geqslant 0) = 1$ and $P(V_1^h > 0) > 0$.*

Lemma 3.2. (Bjork) ***(Farkas' Lemma)*** *Suppose that $d_0, d_1, ..., d_M$ are column vectors in $\mathbb{R}^N$. Then exactly one of the following problems possesses a solution.*

- **Problem 1**: *There exist $\lambda_1, ..., \lambda_M \geqslant 0$ such that $d_0 = \sum_{j=1}^{M} \lambda_j d_j$.*
- **Problem 2**: *There exist $h \in \mathbb{R}^N$ such that $h^\top d_0 < 0$ and $h^\top d_j \geqslant 0$ for $j = 1, ..., M$.*

We now investegate this system for any possible arbitrage portfolios. However first we acknowledge that there exist a nominal price system S_t and a normalised price system Z_t. The latter we define as the nominel pricess under the numeraire S_t^1 that is

$$Z_t = \begin{bmatrix} S_t^1/S_t^1 \\ S_t^2/S_t^1 \\ \vdots \\ S_t^N/S_t^1 \end{bmatrix} = \begin{bmatrix} 1 \\ S_t^2/S_t^1 \\ \vdots \\ S_t^N/S_t^1 \end{bmatrix}.$$

The reason for introducing the normalized price system is that we can without much effort translate results in this system to the nominal system and the normalised system is easier to analize. For this, however, we need af few results.

Lemma 3.3. (Bjork) *With notation as above, the following hold.*

1. *The Z_t value process i related to the S_t value process by*

$$V_t^{h,Z} = hZ_t = \frac{1}{S_t^1} V_t^h.$$

2. *A portfolio is an arbitrage in the S_t system if and only if there is an arbitrage in the Z_t system.*
3. *In the Z_t price system, the numeraie asset Z^1 has unit constant prices i.e. $Z_t^1 = 1$ for all $t \geqslant 0$.*

One of the reason that the normalised system is attractable is that the numeraire asset is constant i.e. risk free in the normalised system. Let us formulate our first main result.

Proposition 3.4. (Bjork) *The market is arbitrage free if and only if there exists strictly positive real numbers $q_1, ..., q_M \geqslant 0$ with $q_1 + \cdots + q_M = 1$ (eq. 3.2) (probability vector) such that the following vector equality holds*

$$\begin{bmatrix} Z_0^1 \\ \vdots \\ Z_N^1 \end{bmatrix} = \begin{bmatrix} Z_1^1(\omega_1) \\ \vdots \\ Z_1^N(\omega_1) \end{bmatrix} q_1 + \cdots + \begin{bmatrix} Z_1^1(\omega_M) \\ \vdots \\ Z_1^N(\omega_M) \end{bmatrix} q_M. \tag{3.3}$$

5.1.3.3 Martingale Measures

Definition 3.5. (Bjork) *Given the objective probability measure P on $(\Omega, \mathcal{F}, P)$, we say that another probability measure Q defined on Ω is **equivalent** to P if*

$$\forall A \in \mathcal{F} : P(A) = 0 \iff Q(A) = 0,$$

or equivalently

$$\forall A \in \mathcal{F} : P(A) = 1 \iff Q(A) = 1.$$

Definition 3.7. (Bjork) *Consider the market model above and set S^1 as the numeraire asset. We say that a probability measure Q defined on Ω is a **martingale measure** if it satisfies the following conditions:*

1. *Q is equivalent to P, i.e. $Q \sim P$.*
2. *For every $i = 1, ..., N$, the normalized asset price process*

$$Z_t^i = \frac{S_t^i}{S_t^1},$$

is martingale under the measure Q.

Theorem 3.8. (Bjork) (First Fundamental Theorem) *Given a fixed numeraire, ther market is free of arbitrage possibilities if and only if there exists a martingale measure Q.*

By assuming that the numeraire asset is risk free (i.e. does not depend on ω) then by scaling we can derive the short interest rate as

$$1 + R = \frac{S_1^1}{S_0^1}.$$

With this in mind we can formulate theorem 3.8 in its more widely used form.

Theorem 3.9. (Bjork) (First Fundamental Theorem) *Assume that there exist a risk free asset, and denote the corresponding risk free interest rate by R. Then the market is arbitrage free if and only if there exist a measure $Q \sim P$ such that*

$$S_0^i = \frac{1}{1+R} E^Q[S_1^i], \qquad \text{for all } i = 1, ..., N. \tag{3.9}$$

5.1.3.4 Martingale Pricing

Moving forward we will assume that there exist a risk free asset and we will denote it by B_t ($B_t = S_t^1/S_0^1$).

Definition 3.10. (Bjork) *A **contingent claim** is any random variable X, defined on the sample space Ω.*

To ensure no arbitrage in the extended market containing the N assets and the contingent claim we can apply the first fundamental pricing theorem on the extended market.

Proposition 3.11. (Bjork) *Consider a given claim X. In order to avoid arbitrage, X must then be priced according to the formula*

$$\Pi_0[X] = \frac{1}{1+R} E^Q[X], \tag{3.10}$$

where Q is a martingale measure for the underlying market $(\Pi, S^1, ..., S^N)$.

5.1.3.5 Completeness

Given that a market is arbitrage-free we may run into a uniqueness issue when determining the price of a contingent claim. If a martingale measure exist we will very much like it to be unique as this will ensure that the price from the risk neutral valuation formula is unique. To this we need the market to be complete.

Definition 3.12. (Bjork) *Consider a contingent claim X. If there exists a portfolio h, based on the underlying assets, such that*

$$V_1^h = X, \text{ with probability 1} \tag{3.11}$$

i.e.

$$V_1^h(\omega_j) = X(\omega_j), \ j = 1, ..., M, \tag{3.12}$$

*then we say that X is **replicated**, or **hedged** by h. Such a portfolio h is called a replicating, or hedging portfolio. If every contingent claim can be replicated, we say that the market is **complete**.*

We can now formulate a proposition on when the market is complete in terms of the matrix D.

Proposition 3.13. (Bjork) *The market is complete if and only if the rows of the matrix D span $\mathbb{R}^M$, i.e. if and only if D has rank M.*

Now we formulate the second fundamental pricing theorem in terms of the martingale measure Q.

Proposition 3.14. (Bjork) (Second Fundamental Theorem) *Assume that the model is arbitrage free i.e. Q exist. Then the market is unique if and only if the martingale measure is unique.*

5.1.3.6 Stochastic Discount Factors

Definition 3.16. (Bjork) *The random variable L on Ω is defined by*

$$L(\omega_i) = \frac{q_i}{p_i}, \qquad i = 1, ..., M.$$

Definition 3.17. (Bjork) *Assume the absence of arbitrage, and fix a martingale measure Q. With notation as above, the **stochastic discount factor** (or "state price deflator") is the random variable Λ on Ω by*

$$\mathbf{M}(\omega) = \frac{1}{1+R} \cdot L(\omega). \tag{3.19}$$

Proposition 3.18. (Bjork) *The arbitrage free price of any claim X is given by the formula*

$$\Pi_0[X] = E^P[\mathbf{M} \cdot X] \tag{3.20}$$

where $\mathbf{M}$ is a stochastic discount factor.

5.2 Self-financing portfolios

We move forward in this chapter by first defining a self-financing portfolio in discrete time and then by letting the step length tend to zero obtain the continuous time analogue.

5.2.1 Discrete time SF portfolio

We consider N different adapted price processes $S^1, ..., S^N$. We use the following definition.

Definition 6.1. (Bjork) *We use the following definitions.*

- S_n^i *is th price of asset* i *at time* n,
- h_n^i *is the number of units of asset* i *held during* $[n, n+1)$, *that is bought at time* n,
- d_n^i *is the dividends from asset* i *in the time-interval* $[n-1, n)$, *that is recieved at time* n,
- h_n *is the portfolio* $(h_n^1, ..., h_n^N)$ *held during* $[n, n+1)$,
- c_n *is the consumption i.e. withdrawel at time* n *(negative being deposits/saving),*
- V_n *is the value of the portfolio just before time* n *i.e. of the portfolio* h_{n-1} *at time* n.

We are now ready to define the self-financing portfolio

Definition 6.2. (Bjork) *A* ***self-financing portfolio supporting the consumption stream*** $\mathbf{c}$ *is a portfolio adhering to the* ***budget constraint*** *given as*

$$h_{n+1}S_{n+1} + c_{n+1} = h_n S_{n+1} + h_n d_{n+1}.$$

The interpretation being, that we may only use funds obtained from selling the old portfolio h_n *and recieved in dividends to buy the new portfolio* h_{n+1} *and consume the amount* c_{n+1}.

Before studying the self-financing portfolio we define the operator Δ (in definition 6.3) as the increment $\Delta x_n = x_{n+1} - x_n$ of a countable sequence $(x_n)_{n\in\mathbb{N}_0}$. Notice that we define the increment forward so the increment n is the increment over the time period $[n, n+1)$ with the first increment being $[0, 1)$. Using this notation we can derive the lemma below.

Lemma 6.4. (Bjork) *For any pair of sequences of real numbers* $(x_n)_{n\in\mathbb{N}_0}$ *and* $(y_n)_{n\in\mathbb{N}_0}$ *we have the relations*

$$\Delta(xy)_n = x_n\Delta y_n + y_{n+1}\Delta x_n, \tag{6.5}$$
$$\Delta(xy)_n = y_n\Delta x_n + x_{n+1}\Delta y_n, \tag{6.6}$$
$$\Delta(xy)_n = x_n\Delta y_n + y_n\Delta x_n + \Delta x_n\Delta y_n. \tag{6.7}$$

This is also valid if the sequances are N*-dimensional, where we interpret the products above as scalar products* $(xy^\top)$.

Using these definitions and the lemma above we see that the dynamics of the self-financing portfolio is given below.

Proposition 6.6. (Bjork) *The dynamics of any self-financing portfolio supporting the consumption stream* c *are given by*

$$\Delta V_n = h_n\Delta S_n + h_n d_{n+1} - c_{n+1}, \tag{6.11}$$

or, in more detail

$$\Delta V_n = \sum_{i=1}^{N} h_n^i(\Delta S_n^i + d_{n+1}^i) - c_{n+1}. \tag{6.12}$$

We may rewrite the dividends as accumulating dividends $D_n^i = \sum_{k=1}^n d_k^i$ and see that $d_{n+1}^i = \Delta D_n^i$ and so the above condition is equivalent with.

Proposition 6.8. (Bjork) *The dynamics of any self-financing portfolio supporting the consumption stream* c *are given by*

$$\Delta V_n = h_n \Delta S_n + h_n \Delta D_n - c_{n+1}, \tag{6.15}$$

or, in more detail

$$\Delta V_n = \sum_{i=1}^{N} h_n^i (\Delta S_n^i + \Delta D_n^i) - c_{n+1}. \tag{6.16}$$

5.2.2 Continuous time SF portfolio

Formulating the dynamics of the self-financing portfolio in continuous time is easy work given the discrete setup above. However since we now are in continuous time we will change the n with a t and cosider the behavour $V_{t+dt} - V_t$ as we let $dt \to 0$. First we formulate some basic notation.

Definition 6.9. (Bjork) *We use the following definitions.*

- S_t^i *is th price of asset* i *at time* t*,*
- h_t^i *is the number of units of asset* i *held at time* t*,*
- D_t^i *is the cumulative dividend processs for asset* i*,*
- h_t *is the portfolio* $(h_t^1, ..., h_t^N)$ *held at time* t*,*
- c_t *is the consumption rate at time* n *(negative being deposits/saving),*
- V_t *is the value of the portfolio at time* t *i.e. of the portfolio* h_t *at time* t*.*

Given these definitions we may define a portfolio strategy that is self-financing.

Definition 6.10. (Bjork) *Let* S *be and adapted* N*-dimensional price process. We define the following*

1. *A* ***portfolio strategy*** *is any adapted* N*-dimensional process* h*.*
2. *The* ***value process*** V^h *corresponding to the portfolio* h *is given by*

$$V_t^h = \sum_{i=1}^{N} h_t^i S_t^i. \tag{6.17}$$

3. *A* ***consumption process*** *is any adapted one-dimensional process* c*.*
4. *A portfolio-consumption pair* (h, c) *is called* ***self-financing*** *if the value process* V^h *satisfies the condition*

$$dV_t^h = \sum_{i=1}^{N} h_t^i (dS_t^i + dD_t^i) - c_t \, dt, \tag{6.18}$$

i.e. if

$$dV_t^h = h_t \, dS_t + h_t \, dD_t - c_t \, dt.$$

5. *The* ***gain process*** G *is defined by*

$$G_t = S_t + D_t \tag{6.19}$$

so we can write the self-financing condition as

$$dV_t = h_t \, dG_t - c_t \, dt. \tag{6.20}$$

6. *The portfolio* h *is said to be* ***Markovian*** *if it is of the form*

$$h_t = h(t, S_t),$$

for some function $h : \mathbb{R}_+ \times \mathbb{R}^N \to \mathbb{R}^N$*.*

5.2.3 Portfolio weights

Definition 6.11. (Bjork) *For a given portfolio* h *the corresponding* ***relative portfolio*** *or* ***portfolio weights*** w *are defined by*

$$w_t^i = \frac{h_t^i S_t^i}{V_t^h}, \ i = 1, ..., N, \tag{6.21}$$

so, in particular, we have $\sum_{i=1}^{N} w_i = 1$*.*

Lemma 6.12. (Bjork) *A portfolio-consumption pair (h, c) is self-financing if and only if*

$$dV_t^h = V_t^h \sum_{i=1}^{N} w_t^i \frac{dS_t^i + dD_t^i}{S_t^i} - c_t \, dt \tag{6.22}$$

or equivalently with the absolute weights

$$dV_t^h = \sum_{i=1}^{N} h_t^i (dS_t^i + dD_t^i) - c_t \, dt.$$

Lemma 6.13. (Bjork) *Consider the case with no dividends. Let c be a consumption process, and assume that there exist a scalar process Z and a vector process $q = (q^1, ..., q^N)$ such that*

$$dZ_t = Z_t \sum_{i=1}^{N} q_t^i \frac{dS_t^i}{S_t^i} - c_t \, dt, \tag{6.23}$$

and $\sum_{i=1}^{N} qq^i = 1$ (eq. 6.24). Now define a portfolio h by

$$h_t^i = \frac{q_t^i Z_t}{S_t^i}. \tag{6.25}$$

Then the value process V^h is given by $V^h = Z$, the pair (h, c) is self-financing, and the corresponding relative portfolio w is given by $w = q$.

5.3 Black-Scholes PDE

The Black-Scholes model revolves arround SDE's as seen above. In this model we have two assets a risk free asset B and a stochastic priced asset S. We therefore start by defining what we mean by a quote-on-qoute *risk free* asset.

Definition 7.1. (Bjork) *The price process B is the price of a **risk free asset** if it has the dynamics*

$$dB_t = r_t B_t \, dt, \tag{7.1}$$

where r is any $\mathcal{F}_t$ adapted process.

We see from this definition that the meaning of "risk free" is the property, that B is priced locally deterministic in the sence that r is adapted and therefore known at time t and we therefore know the yield on a short term basis. This is also why we may call r the **short interest rate**. Given the dynamics above, we know that B in fact is represented by the process

$$B_t = B_0 e^{\int_0^t r_s \, ds},$$

for some B_0 initial value. We will moving forward assume that $B_0 = 1$. The stochastic asset S has dynamics.

$$dS_t = \mu(t, S_t) \, dt + \sigma(t, S_t) \, dW_t, \tag{7.2}$$

where as usual μ and σ are deterministic functions and W_t is a standard Brownian motion. Note that the risk free asset has a similarly process with $\sigma = 0$. We may now include this in the definition of the Black-Scholes model.

Definition 7.2. (Bjork) *The **Black-Scholes model** consists of two assets with dynamics given by*

$$dB_t = rB_t \, dt, \tag{7.3}$$
$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{7.4}$$

where r, μ, σ are deterministic constants.

Definition 7.3. (Bjork) *A **zero coupon bond** with maturity T (henceforth "T-bond") is an asset which pays the holder the face value 1 dollar at time T. The price at time n of a T-bond is denoted by $p(n, T)$.*

Definition 7.4. *The (possible stochastic) discrete **short rate** r_n, for the period $[n, n+1]$, is defined as*

$$p(n, n+1) = \frac{1}{1 + r_n}. \tag{7.6}$$

From this short rate we may derive the dynamics of the bank account recieving zero-coupon rates for each distinct time interval.

Definition 7.5. (Bjork) *The dynamics of the bank account are given by*

$$\Delta B_n = r_n B_n. \tag{7.7}$$

5.3.1 Contingent Claims and Arbitrage

Definition 7.6. (Bjork) *A **European call option** with **exercise price** (or strike price) K and **time of maturity** (exercise date) T on the **underlying asset** S is a contract defined by the following clauses:*

- *The holder of the option has, at time T, the right to buy one share of the underlying stock at the price K dollars from the underwriter of the option.*
- *The holder of the option is in no way obliged to buy the underlying stock.*
- *The right to buy the underlying stock at the price K can only be exercised at the precise time T.*

Obviously, we also have the **european put** option which gives the owner the right to sell an asset at price K at time T. Let os formally define a contingent claim.

Definition 7.7. *Consider a financial market with vector price process S. A **contingent claim** with **date of maturity** T, also called a T-claim, is any random variable $\mathcal{X} \in \mathcal{F}_T^S$. A contingent claim $\mathcal{X}$ is called a **simple** claim if it is of the form $\mathcal{X} = \Phi(S_t)$. The function Φ is called the **contract function**.*

Definition 7.8. (Bjork) *An **arbitrage** possibility on a financial market is a self-financed portfolio h such that*

$$V^h(0) = 0, \tag{7.13}$$

$$P(V_T^h \geqslant 0) = 1, \tag{7-14}$$

$$P(V_T^h > 0) > 0. \tag{7.15}$$

*We say that the market is **arbitrage free** if there are no arbitrage possibilities.*

Definition 7.9. (Bjork) *Suppose that there exists a self-financing portfolio h, such that the value process V^h has the dynamics*

$$dV_t^h = k_t V_t^h \, dt, \tag{7.16}$$

where k is an adapted process. Then it must hold that $k_t = r_t$ for all t, ore there exists an arbitrage possibility.

Theorem 7.10. (Bjork) (Black-Scholes equation) *Assume that the market is specified by the equations*

$$dB_t = rB_t \, dt, \tag{7.18}$$

$$dS_t = \mu(t, S_t)S_t \, dt + \sigma(t, S_t)S_t \, dW_t, \tag{7.19}$$

and that we want to price a contingent claim of the form $\mathcal{X} = \Phi(S_t)$ (eq. 7.20). Then the only pricing function of the form $\Pi_t[\Phi(S_t)] = F(t, S_t)$ (eq. 7.21) which is consistent with the absence of arbitrage in the market $[B_t, S_t, \Pi_t]$ is when F is the solution of the following boundary value problem in the domain $[0, T] \times \mathbb{R}_+$:

$$F_t(t,s) + rsF_s(t,s) + \frac{1}{2}s^2\sigma^2(t,s)F_{ss}(t,s) - rF(t,s) = 0,$$

$$F(T,s) = \Phi(s).$$

5.3.2 Risk Neutral Valuation

Theorem 7.11. (Bjork) (Risk Neutral Valuation) *The arbitrage free price of the claim $\Phi(S_t)$ is given by $\Pi_t[\Phi] = F(t, S_t)$, where F is given by the formula*

$$F(t,s) = e^{-r(T-t)}E_{t,s}^Q[\Phi(S_T)], \tag{7.43}$$

where the Q-dynamics of S are those of

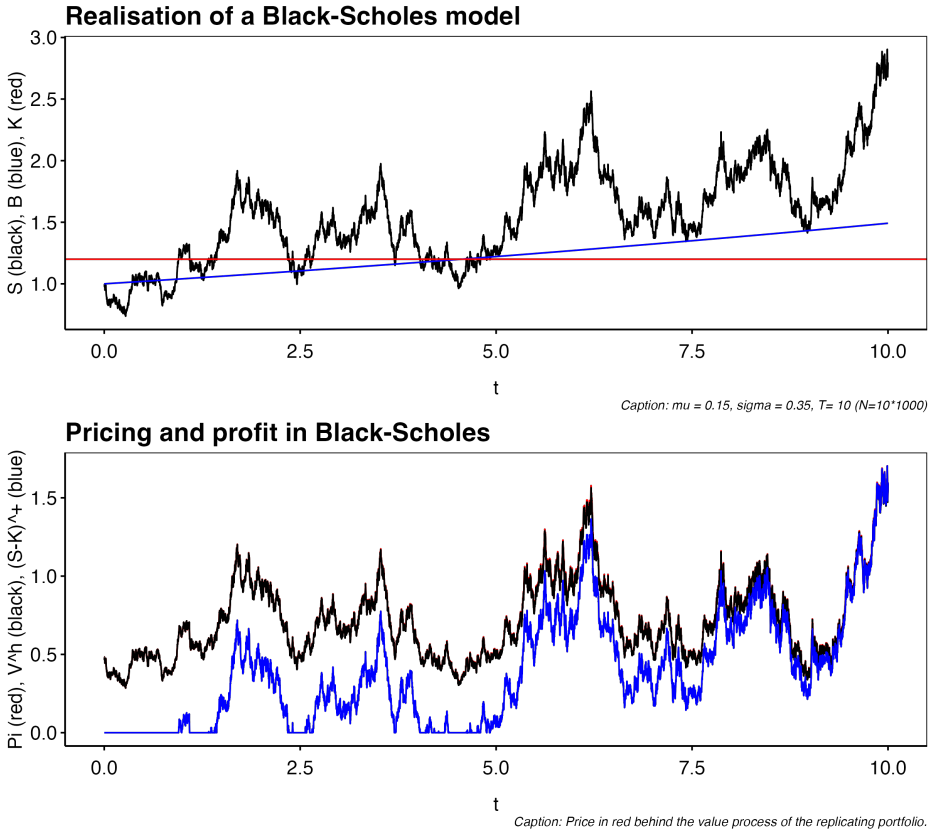
$$dS_t = rS_t \, dt + S_t\sigma(t, S_t) \, dW_t^Q. \tag{7.42}$$

Property 7.12. (Bjork) (The Martingale Property) *In the Black-Scholes model, the price process Π_t for every traded asset, be it the underlying or derivate asset, has the property the the normalized price process*

$$Z_t = \frac{\Pi_t}{B_t},$$

(including S_t/B_t) is a martingale under the measure Q.

5.3.3 Black-Scholes formula



This chapter will center on deriving the famous Black-Scholes formula. We start by laying out the assumptions of the model. We have a market consiting of two assets: a stochastic prices asset S and a risk free asset B. The prices processes have dynamics:

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{7.45}$$
$$dB_t = rB_t \, dt, \tag{7.44}$$

where $S_0 = s$ and $B_0 = 1$ (by assumption). Now from Feymann-Kac and the definition of arbitrage we know that a simple claim $\Phi(S_t)$ has the arbitrage free price given by the risk neutral valueation formula.

$$F(t,s) = e^{-r(T-t)} E^Q_{t,s}[\Phi(S_T)], \tag{7.46}$$

where Q is a probability measure, namely a Martingale measure, such that the dynamics of S under this measure is

$$dS_t = rS_t \, dt + \sigma S_t \, dW^Q_t, \tag{7.47}$$

with W^Q_t being a Brownian motion wrt. to the probability measure Q (not P). The above still has the initial condition $S_0 = s$. Given these assumptions we may formulate the Black-Scholes formula.

Theorem 7.13. (Bjork) (Black-Scholes formula) *The price of the european call option with strikeprice K and maturity T (contract function $\Phi(S_t) = (S_t - K)^+$) takes the form $\Pi_t = F(t,s)$, where*

$$F(t,s) = sN(d_1(t,s)) - e^{-r(T-t)} KN(d_2(t,s)), \tag{7.52}$$

where N is the distribution-function for an $\mathcal{N}(0,1)$-distributed random variable and

$$d_1(t,s) = \frac{1}{\sigma\sqrt{T-t}}\left(\log\left(\frac{s}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)\right), \tag{7.53}$$

$$d_2(t,s) = d_1(t,s) - \sigma\sqrt{T-t}. \tag{7.54}$$

Proof.

We let the market be given in terms of the price processes S and B with dynamics.

$$\begin{aligned} dS_t &= \mu S_t\ dt + \sigma S_t\ dW_t, \\ dB_t &= rB_t\ dt, \end{aligned}$$

with $B_t = 1$ and $S_t = s$. We assume that μ, σ, r are deterministic real numbers. Consider the contingent claim

$$\Phi(S_t) = (S_t - K)^+,$$

that is the European call option. Let Q be a martingale measure such that the dynamics of S may be written as

$$dS_t = rS_t\ dt + \sigma S_t\ dW_t^Q,$$

then S_t is clearly a GBM wrt. the measure Q. Therefore we know the solution given in terms of the increment of the Brownian motion W^Q as follows

$$S_u = s \cdot \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)(u-t) + \sigma\left(W_u^Q - W_t^Q\right)\right\},$$

for some initial condition $S_t = s$. From theorem 7.10 we know that the only pricing function which takes the form

$$\Pi_t[\Phi(S_T)] = F(t, S_t),$$

can only be consistent with the absence of arbitrage if F is the solution the the boundary value problem

$$\begin{aligned} F_t(t,s) + rsF_s(t,s) + \frac{1}{2}s^2\sigma^2 F_{ss}(t,s) - rF(t,s) = 0, \\ F(T,s) = \Phi(s). \end{aligned}$$

From Feymann-Kac we then know that the stochastic representation of such a solution take the form

$$F(t,s) = e^{-r(T-t)}E_{t,s}^Q[\Phi(S_T)].$$

Here the superscript refers to taking mean value with respect to the measure Q. This gives the solution to the pricing function

$$F(t,s) = e^{-r(T-t)}\int \Phi(S_T)\ dQ.$$

Under the measure Q we have that for $u \geqslant t$:

$$Z_u = \log(S_u/s) \sim \mathcal{N}\left(\left(r - \frac{1}{2}\sigma^2\right)(u-t), \sigma\sqrt{u-t}\right)$$

Hence we may set $u = T$ and observe that

$$\begin{aligned}
F(t,s) &= e^{-r(T-t)} \int_{-\infty}^{\infty} \Phi(se^z) f(z)\ dz \\
&= e^{-r(T-t)} \int_{-\infty}^{\infty} (se^z - K)^+ f(z)\ dz \\
&= e^{-r(T-t)} \int_{\log\left(\frac{K}{s}\right)}^{\infty} (se^z - K) f(z)\ dz \\
&= e^{-r(T-t)} \left(s \int_{\log\left(\frac{K}{s}\right)}^{\infty} e^z f(z)\ dz - K \int_{\log\left(\frac{K}{s}\right)}^{\infty} f(z)\ dz \right),
\end{aligned}$$

where we used that f is the distribution function of a normal distributed random variable with mean $(r - \sigma^2/2)(T-t)$ and variance $\sigma\sqrt{T-t}$ and that

$$(se^z - K)^+ \geqslant 0 \iff se^z \geqslant K \iff z \geqslant \log\left(\frac{K}{s}\right)$$

Using that the MGF of a $X \sim \mathcal{N}(\alpha, \beta^2)$ variable is

$$E[e^{tX}] = e^{\alpha t + \frac{1}{2}\beta^2 t^2},$$

and the shorthand $N(t)$ for the distribution function of the standard normal distribution, we have

$$\begin{aligned}
F(t,s) &= e^{-r(T-t)} \left(sE\left[e^{Z_T} 1_{Z_T \geqslant \log\left(\frac{K}{s}\right)} \right] - KP\left(Z_T \geqslant \log\left(\frac{K}{s}\right) \right) \right) \\
&= e^{-r(T-t)} s \exp\left\{ \left(r - \frac{1}{2}\sigma^2 \right)(T-t) + \frac{1}{2}\sigma^2 (T-t) \right\} E\left[1_{Z_T \geqslant \log\left(\frac{K}{s}\right)} \right] \\
&\quad - e^{-r(T-t)} KP\left(X \geqslant \frac{1}{\sigma\sqrt{T-t}} \left(\log\left(\frac{K}{s}\right) - (r - \sigma^2/2)(T-t) \right) \right) \\
&= sE\left[1_{Z_T \geqslant \log\left(\frac{K}{s}\right)} \right] - e^{-r(T-t)} KP\left(X \leqslant \frac{1}{\sigma\sqrt{T-t}} \left(\log\left(\frac{s}{K}\right) + (r - \sigma^2/2)(T-t) \right) \right) \\
&= sN(d_1(s,t)) - e^{-r(T-t)} KN\left(d_2(s,t)\right),
\end{aligned}$$

as desired. ■

5.4 Completeness and Hedging

We derived the pricing function of the european call option above and introduced the theory around boundary value problems and Feymann-Kac solution to the partial differential stochastic equation. Now we want to see if a portfolio exists such that it gives the payout $\Phi(S_T)$ with probability one.

In order to do this, we return to the concept of hedge and replication.

Definition 8.1. (Bjork) *We say that a T-claim $\mathcal{X}$ can be **replicated**, alternatively the it is **reachable** or **hedgeable**, if there exists a self-financing portfolio h such that*

$$V_T^h = \mathcal{X}, \ P - \text{a.s.} \tag{8.1}$$

*In this case we say that h is a **hedge** against $\mathcal{X}$. Alternatively, h is called a **replicating** or **hedging** portfolio. If every contingent claim is reachable we say that the market is **complete**.*

If we can find a portfolio h that reaches $\mathcal{X}$ in value over the time period $[t, T]$ it must mean, that holding the portfolio is equivalent with holding the contract itself. We therefore have the natural assumption that the price process must satisfie $\Pi_t[\mathcal{X}] = V_t^h$ for all $t \geqslant 0$. How this relates to the absence of arbitrage is given below.

Proposition 8.2. (Bjork) *Suppose $\mathcal{X}$ is hedged using the portfolio h. Then the only price process $\Pi_t[\mathcal{X}]$ which is consistent with no arbitrage is given by $\Pi_t[\mathcal{X}] = V_t^h$. Furthermore, if $\mathcal{X}$ can be hedged by both h and g then $V_t^g = V_t^h$ for all t with probability one.*

5.4.1 Completeness in Black-Scholes

The Black-Scholes model will be investegated in the following. We start by stating the important theorem.

Theorem 8.3. (Bjork) *Consider the Black-Scholes model given by*

$$dS_t = \mu(t, S_t)S_t \ dt + \sigma(t, S_t)S_t \ dW_t, \tag{8.2}$$

$$dB_t = rB_t \ dt, \tag{8.3}$$

The model above is complete.

The following lemma gives us replicability of a **simple** claim (which we will restrict ud to).

Lemma 8.4. (Bjork) *Suppose that there exist an adapted process V and an adapted process $w = [w^B, w^S]$ with $w_t^B + w_t^S = 1$ (eq. 8.4) for all $t \geqslant 0$, such that*

$$dV_t = V_t(w_t^B r + w_t^S \mu(t, S_t)) \ dt + V_t w_t^S \sigma(t, S_t) \ dW_t, \tag{8.5}$$

$$V_t = \Phi(S_t). \tag{8.5}$$

Then the claim $\mathcal{X} = \Phi(S_t)$ can be replicated using w as the relative portfolio. The corresponding value process is given by the process V and the absolute portfolio h is given by

$$h_t^B = \frac{w_t^B V_t}{B_t}, \tag{8.6}$$

$$h_t^S = \frac{w_t^S V_t}{S_t}. \tag{8.7}$$

Doing some heuristics we come up with some clever weights, which turns on to adhere to the boundary value problem formulated in the Black-Scholes equation. Given that the weights gives rise to the desired value process, we have succesfully found the portfolio weight (see lemma above).

Theorem 8.5. (Bjork) *Consider the Black-Scholes model given in (8.3)-(8.4), and a simple contingent claim $\mathcal{X} = \Phi(S_t)$. Define F as the solution to the boundary value problem*

$$F_t(t, s) + rsF_s(t, s) + \frac{1}{2}s^2\sigma^2 F_{ss}(t, s) - rF(t, s) = 0, \tag{8.17}$$

$$F(T, s) = \Phi(s). \tag{8.17}$$

Then $\mathcal{X}$ can be replicated by the relative portfolio

$$w_t^B = \frac{F(t,S_t) - S_t F_s(t,S_t)}{F(t,S_t)}, \tag{8.18}$$

$$w_t^S = \frac{S_t F_s(t,S_t)}{F(t,S_t)}. \tag{8.19}$$

The corresponding absolute portfolio is given by

$$h_t^B = \frac{F(t,S_t) - S_t F_s(t,S_t)}{B_t}, \tag{8.20}$$

$$h_t^S = F_s(t,S_t), \tag{8.21}$$

and the value process V^h is given by

$$V_t^h = F(t,S_t). \tag{8.22}$$

Proposition 8.6. (Bjork) *Consider the Black-Scholes model given in (8.3)-(8.4), and a contingent claim on the form $\mathcal{X} = \Phi(S_T, Z_T)$ (eq. 8.29). We define the process Z_t as*

$$Z_t = \int_0^t g(u,S_u)\, du, \tag{8.30}$$

for some choice of the deterministic function g. Then $\mathcal{X}$ can be replicated using a relative portfolio given by

$$w_t^B = \frac{F(t,S_t,Z_t) - S_t F_s(t,S_t,Z_t)}{F(t,S_t,Z_t)}, \tag{8.31}$$

$$w_t^S = \frac{S_t F_s(t,S_t,Z_t)}{F(t,S_t,Z_t)}. \tag{8.32}$$

where F is the solution to the boundary value problem

$$F_t(t,s,z) + rsF_s(t,s,z) + \frac{1}{2}s^2\sigma^2 F_{ss}(t,s,z) - rF(t,s,z) = 0, \tag{8.33}$$

$$F(T,s,z) = \Phi(s,z). \tag{8.33}$$

The corresponding value process is given by $V_t = F(t,S_t,Z_t)$, and F has the stochastic representation

$$F(t,s,z) = e^{-r(T-t)} E^Q_{t,s,z}[\Phi(S_T,Z_T)], \tag{8.34}$$

where the Q-dynamics are given by

$$dS_u = rS_u\, du + S_u\sigma(u,S_u)\, dW_u^Q, \tag{8.35}$$

$$S_t = s, \tag{8.36}$$

$$dZ_u = g(u,S_u)\, du, \tag{8.37}$$

$$Z_t = z. \tag{8.38}$$

5.4.2 Absence of Arbitrage

In general we have conflicting forces when evaluating when a certain market is arbitrage free and/or complete. We have in simple terms the non-rigorous theorem below.

Meta-theorem 8.3.1. (Bjork) *Let N denote the number of underlying **traded** assets in the model **excluding** the risk free asset, and let R denote the number of random sources driving the price system. Genericallly we then have the following statements.*

- *The model is arbitrage free if and only if $N \leqslant R$.*
- *The model is complete if and only if $N \geqslant R$.*
- *The model is arbitrage free and complete if and only if $N = R$.*

5.4.3 Incomplete Markets

We assume a market with a risk free asset and one risky assets with dynamics

$$dX_t = \mu(t, X_t)\ dt + \sigma(t, X_t)\ dW_t. \tag{9.1}$$

We want to find a unique price of a derivative on a functional form of the risky asset. We assume that we cannot invest in the asset representing the process X_t and so we can solely write contracts based on the observation X_T. The problem here is that we can only short or long the risk free asset and so no derivative is replicable.

The way we solve this problem is by having the market set the price of risk and universally price derivatives based on this given price process. We then have the assumptions

Assumption 9.2.1 *We have the market given with the only investable asset B with dynamics*

$$dB_t = rB_t\ dt. \tag{9.2}$$

*We furthermore, have an empirically observable stochastic process X which is **not** the price process of any traded asset. The P-dynamics of X is given by*

$$dX_t = \mu(t, X_t)\ dt + \sigma(t, X_t)\ dW_t.$$

Assumption 9.2.2 *There is a liquid market for every contingent claim.*

Assumption 9.2.3 *We assume that*

- *There is a liquid, frictionless market for each of the contingent claims $\mathcal{Y}$ and $\mathcal{Z}$.*
- *The market prices of the claims are of the form*

$$\Pi_t[\mathcal{Y}] = F(t, X_t),$$
$$\Pi_t[\mathcal{Z}] = G(t, X_t),$$

where F and G are smooth real valued function.

From Ito's formula we have the dynamics

$$dF = \mu_F F\ dt + \sigma_F F\ dW, \tag{9.4}$$
$$dG = \mu_G G\ dt + \sigma_G G\ dW. \tag{9.5}$$

Where the processes μ_F and σ_F are given by

$$\mu_F = \frac{F_t + \mu F_x + \frac{1}{2}\sigma^2 F_{xx}}{F},$$
$$\sigma_F = \frac{\sigma F_x}{F}.$$

By forming a portfolio of the two contracts we lead to the relation.

$$\frac{\mu_F - r}{\sigma_F} = \frac{\mu_G - r}{\sigma_G}.$$

This gives the important insight.

Proposition 9.1. (Bjork) *Assume that the market for derivatives is free of arbitrage. Then there exists a universal process $\lambda(t, X_t)$ such that, with probability one, and for all t, we have*

$$\frac{\mu_F(t, X_t) - r}{\sigma_F(t, X_t)} = \mu(t, X_t), \tag{9.7}$$

regardless of the specific choice of the derivative F.

Proposition 9.2. (Bjork) *Assume absence of arbitrage, the pricing function $F(t,x)$ of the T-claim $\Phi(X_T)$ solves the following boundary value problem.*

$$F_t(t,x) + \mathcal{A}F(t,x) - rF(t,x) = 0, \qquad (t,x) \in (0,T) \times \mathbb{R}, \tag{9.8}$$

$$F(T,x) = \Phi(x), \qquad x \in \mathbb{R}, \tag{9.9}$$

where

$$\mathcal{A}F(t,x) = \{\mu(t,x) - \lambda(t,x)\sigma(t,x)\}\, F_x(t,x) + \frac{1}{2}\sigma^2(t,x)F_{xx}(t,x).$$

Proposition 9.3. (Bjork) (Risk neutral valuation) *Assuming absence of arbitrage, the pricing function $F(t,x)$ of the T-claim $\Phi(X_T)$ is given by the formula*

$$F(t,x) = e^{-r(T-t)}E^Q_{t,x}[\Phi(X_T)]. \tag{9.11}$$

The dynamics of X under the martingale measure Q are given by

$$dX_t = \{\mu(t,x) - \lambda(t,x)\sigma(t,x)\}\, F_x(t,x) + \sigma(t,x)\, dW^Q_t,$$

where W^Q is a Q-Brownien motion.

5.5 Parity relations

5.5.1 Put-call Parity

The notion of continuous rebalancing the replicating portfolio require leads to problems in the real world. Trading does cost some money (typical in fractions) and so contiuous balancing would make the portfolio go to 0 rather quickly. Why? The Brownian motion has unbounded variation and so we would have to sell and buy the portfolio uncountable many time in any interval and the shift in weight is not neglible. Because of this we would like to see which claims we can replicate by buying and holding a combination of assets and derivatives.

Proposition 10.1. (Bjork) *Let Φ and Ψ be contract functions for the T-claims $\mathcal{X} = \Phi(S_T)$ and $\mathcal{Y} = \Psi(S_T)$. Then for any real numbers α and β we have the following price relation:*

$$\Pi_t[\alpha\Phi + \beta\Psi] = \alpha\Pi_t[\Phi] + \beta\Pi_t[\Psi]. \tag{10.1}$$

If we consider the basic contract functions

$$\Phi_S(x) = x, \tag{10.2}$$
$$\Phi_B(x) = 1, \tag{10.3}$$
$$\Phi_{C,K}(x) = (x-K)^+, \tag{10.4}$$
$$\Phi_{P,K}(x) = (K-x)^+.$$

That is a contract paying (respectively): the price of one stock, 1 dollar, one european call and one european put both with strike K. It is clear that the following prices are

$$\Pi_t[\Phi_S] = S_t, \tag{10.5}$$
$$\Pi_t[\Phi_B] = e^{-r(T-t)}, \tag{10.6}$$
$$\Pi_t[\Phi_{C,K}] = c(t, S_t; K, T), \tag{10.7}$$
$$\Pi_t[\Phi_{P,K}] = p(t, S_t; K, T).$$

Where $c(t, s, K, T, r, \sigma)$ and $p(t, s, K, T, r, \sigma)$ are the pricing function of the european call and put option. We see that we can replicate these payouts by: buying the stock today and selling at time T, buying a zero coupon T-bond with face value 1, buying the call and put option.

Then we can by choosing $\alpha, \beta, \gamma_1, ..., \gamma_n$ form a portfolio consisting of α stocks, β T-bonds and γ_i call options with maturity T and strike K_i. The price is then a linear combination given the choice (se proposition 10.1).

The put option is not includet in the above portfolio as we have the put-call parity below

Proposition 10.2. (Bjork) (Put-call parity) *Consider a European call and a European put, both with strike K and time of maturity T. Then we have the relation.*

$$p(t, s) = Ke^{-r(T-t)} + c(t, s) - s. \tag{10.11}$$

In particular the put option can be replicated by a constant portfolio consisting of K zero coupon T-bonds, a European call option and a single short position in the underlying stock.

We now have the pleasing proposition given the class of claims we can reach with the buy-and-hold portfolio with T-bonds, stock and call options

Proposition 10.3. (Bjork) *Fix an arbitrary continuous contract function Φ with compact support. Then the corresponding contract can be replicated with arbitrary precision (in sup-norm) using a constant portfolio consisting only of bonds, call options and the underlying stock.*

5.5.2 The Greeks

When holding a portfolio we may denote the pricing function by $P(t,s)$. Here we only have one **underlying** asset with price process S_t. We now have two types of risk:

- Price changes in the underlying asset.
- Misspecifications in the model parameters.

These two risk give rise to "the greeks" as defined below.

Definition 10.4. (Bjork) *The greeks of a portfolio is given by*

$$\Delta = \frac{\partial P}{\partial s},\ \Gamma = \frac{\partial^2 P}{\partial s^2},\ \rho = \frac{\partial P}{\partial r},\ \Theta = \frac{\partial P}{\partial t},\ \mathcal{V} = \frac{\partial P}{\partial s}.$$

For the call option in particular we have the following derivatives.

Proposition 10.5. (Bjork) *The greeks of a portfolio consisting of a single European call option with maturity T and strike price K have the following greeks (φ denoting the density function of a $\mathcal{N}(0,1)$-variable):*

$$\Delta = N(d_1), \tag{10.17}$$

$$\Gamma = \frac{\varphi(d_1)}{s\sigma\sqrt{T-t}}, \tag{10.18}$$

$$\rho = K(T-t)e^{-r(T-t)}N(d_2), \tag{10.19}$$

$$\Theta = -\frac{s\varphi(d_1)\sigma}{2\sqrt{T-t}} - rKe^{-r(T-t)}N(d_2), \tag{10.20}$$

$$\mathcal{V} = s\varphi(d_1)\sqrt{T-t}. \tag{10.21}$$

5.6 Fundamental pricing theorem I and II

We start by stating the following theorem.

Theorem 11.1. (Bjork) *If at least on of the assets $S^1, ..., S^N$ has diffusion term which is non-zero at all times, and if naive portfolio strategies are admitted, then the model admits arbitrage.*

We will go as follows. Derive the fundamental pricing theorem 1 and 2 in a setting with zero interest rate. Then we will extend the result in general by choosing a simple numeraire. We start by defining some basic notation.

Definition 11.2. (Bjork) *Define the process h as*

$$h = [h^0, h^S] := [h^0, h^1, ..., h^N]$$

We define the following.

- *For a process h, its **value process** V_t^h is defined by*

$$V_t^h = h_t^0 \cdot 1 + \sum_{i=1}^{N} h_t^i S_t^i, \tag{11.3}$$

  or in compact form

$$V_t^h = h_t^0 \cdot 1 + h_t^S S_t \tag{11.4}$$

- *An adapted process h^S is called **admissible** if there exists a non-negative real number α (which may depend on the choice of h^S) such that*

$$\int_0^t h_u^S dS_u \geqslant -\alpha, \tag{11.5}$$

  *for all $t \in [0, T]$. A process h, is called an **admissible portfolio** process if h^S is admissible.*
- *An admissible portfolio is said to be **self-financing**, if*

$$V_t^h = V_0^h + \int_0^t h_u^S \, dS_u, \tag{11.6}$$

  i.e. if

$$dV_t^h = h_t^S \, dS_t. \tag{11.7}$$

Lemma 11.3. (Bjork) *For any adapted process h^S satisfying the admissibility condition above, and for any real number x, there exists a unique adapted process h^0, such that:*

- *The process h defined by $h = [h^0, h^S]$ is self-financing.*
- *The value process is given by*

$$V_t^h = x + \int_0^t h_u^S \, dS_u. \tag{11.8}$$

In particular, the space $\mathcal{K}_0$ of portfolio values, reachable at time T by means of a self-financing portfolio with zero initial cost is given by

$$\mathcal{K}_0 = \left\{ \int_0^T h_t^S \, dS_u : \ h^S \text{ is admissible} \right\}. \tag{11.9}$$

Definition 11.4. (Bjork) *A probability measure Q on $\mathcal{F}_T$ is called **equivalent martingale measure** for the market model, the numeraire S^0, and the time interval $[0, T]$, if it has the following properties:*

- *$Q \sim P$ on $\mathcal{F}_T$, so P and Q are equivalent.*
- *All price processes $S^0, S^1, ..., S^N$ are martingales under Q on the time interval $[0, T]$.*

*An equivalent martingale measure will often be referred to as just "a martingale measure" or as "an EMM". If $Q \sim P$ has the property that $S^0, S^1, ..., S^N$ are local martingales, then Q is called a **local martingale measure**.*

Theorem 11.5. (Bjork) (The First Fundamental Theorem) *The model is arbitrage free "essentially" if and only if there existis a (local) martingale measure Q.*

Definition 11.6. (Bjork) *With the notation above, we say that the model admits*

- ***No Arbitrage*** *(NA) if*
$$\mathcal{C} \cap L^\infty_+ = \{0\}, \tag{11.21}$$
- ***No Free Lunch with Vanishing Risk*** *(NFLVR) if*
$$\tilde{\mathcal{C}} \cap L^\infty_+ = \{0\}, \tag{11.22}$$
where $\tilde{\mathcal{C}}$ denotes the closure of $\mathcal{C}$ in L^∞.

Theorem 11.7. (Bjork) (Kreps-Yan Separation Theorem) *If $\mathcal{C}$ is weak* closed, and if*
$$\mathcal{C} \cap L^\infty_+ = \{0\},$$
then there exists a random variable $L \in L^1$ such that L is P almost surely strictly positive, and
$$E^P[L \cdot X] \leqslant 0,$$
for all $X \in \mathcal{C}$.

Proposition 11.8. (Bjork) *If the asset price processes are uniformly bounded, then the condition NFLVR implies that $\mathcal{C}$ is weak* closed.*

Theorem 11.9. (Bjork) (First Fundamental Theorem) *Assume that the asset price process S is bounded. Then there exists an equivalent martingale measure if and only if the model satisfies NFLVR.*

Theorem 11.10. (Bjork) (First Fundamental Theorem) *Assume that the asset price process S is locally bounded. Then there exists an equivalent martingale measure if and only if the model satisfies NFLVR.*

Assumption 11.4.1. (Bjork) *We assume that $S^0_t > 0$ P-a.s. for all $t \geqslant 0$.*

Definition 11.11. (Bjork) *The **normalized economy** (also referred to as the "Z-economy") is defined by the price vector process Z, where*
$$Z_t = \frac{S_t}{S^0_t}.$$

Definition 11.12. (Bjork)

- *A **portfolio stragegy** is any adapted $(N+1)$-dimensional process*
$$h_t = [h^0_t, h^1_t, ..., h^N_t].$$
- *The **S-value process** V^S_t corresponding to the portfolio h is $h_t S_t$.*
- *The **Z-value process** V^Z_t corresponding to the portfolio h is $h_t Z_t$.*
- *A portfolio is said to be **admissible** if it is admissible as an Z portfolio.*
- *An admissible portfolio is **S-self-balancing** if*
$$dV^S_t = \sum_{i=0}^N h^i_t \, dS^i_t \tag{11.26}$$
- *An admissible portfolio is **Z-self-balancing** if*
$$dV^Z_t = \sum_{i=0}^N h^i_t \, dZ^i_t. \tag{11.28}$$

Lemma 11.13. (Bjork) (Invariance Lemma) *With assumptions as above, the following hold.*

i. *A portfolio h is S-self-financing if and only if it is Z-self-financing.*
ii. *The value processes V^S and V^Z are connected by*
$$V^Z_t = \frac{1}{S^0_t} \cdot V^S_t.$$

iii. *A claim $\mathcal{Y}$ is S-replical if and only if the claim*

$$\frac{\mathcal{Y}}{S_T^0}$$

is Z-replicable.

iv. *The model is S arbitrage free if and only if it is Z arbitrage free.*

Theorem 11.14. (Bjork) (The First Fundamental Theorem) *Consider the market model $S^0, S^1, ..., S^N$ where we assume that $S_t^0 > 0$, P-a.s. for all $t \geqslant 0$. Assume furthermore that $S^0, S^1, ..., S^N$ are locally bounded. Then the followin conditions are equivalent:*

- *The model satisfies NFLVR.*
- *There exists a measure $Q \sim P$ such that the processes*

$$Z^0, Z^1, ..., Z^N,$$

are local martingales under Q.

5.6.1 Completeness

Lemma 11.15. (Bjork) *Consider a given T-claim X. Fix a martingale measure Q and assume that the normalized claim X/S_T^0 is integrable. If the Q-martingale M, defined by*

$$M_t = E^Q\left[\frac{X}{S_T^0}\middle|\mathcal{F}_t\right], \tag{11.34}$$

admits an integral representation of the form

$$M_t = x + \sum_{i=1}^{N}\int_0^t h_s^i\, dZ_s^i, \tag{11.35}$$

then X can be hedged in the S-economy. Furthermore, the replicating portfolio $(h^0, h^1, ..., h^N)$ is given by the above for h^i, $i = 1, ..., N$ and $h_t^0 = M_t - \sum_{i=1}^N h_t^i Z_t^i$.

Theorem 11.16. (Bjork) (Jacod) *Let $\mathcal{M}$ denote the convex set of equivalent martingale measures. Then, for any fixed $Q \in \mathcal{M}$, the following statements are equivalent:*

- *Every Q local martingale M has dynamics of the form*

$$dM_t = \sum_{i=1}^{N} h_s^i\, dZ_s^i.$$

- *Q is an extremal point of $\mathcal{M}$.*

Theorem 11.17. (Bjork) (The Second Fundamental Theorem) *Assume that the market is arbitrage free and consider a fixed numeraire asset S^0. Then the market is complete if and onlt if the martingale measure Q, corresponding to the numeraire S^0, is unique.*

5.6.2 Risk Neutral Valuation Formula

We have the setting of a market consisting of the assets $S^0, ..., S^N$ of $N + 1$ assets. We consider the numeraire S^0 being a risk free asset. We introduce a price of contingent claim X, such that the extended market consisting of the price process of X and the $N + 1$ assets is arbitrage free. Alternatively, we can, equivalently, find a replicating portfolio h such that $V_T^h = X$ with probability one.

Theorem 11.18. (Bjork) (General Pricing Equation) *The arbitrage free price process for the T-claim X is given by*

$$\Pi_t[X] = S_t^0 E^Q\left[\frac{X}{S_T^0}\middle|\mathcal{F}_t\right], \tag{11.41}$$

where Q is the (not necessarily unique) martingale measure for the a priori given market $S^0, S^1, ..., S^N$, with S^0 as the numeraire.

If if we assume that the bank account takes the form

$$S_t^0 = S_0^0 e^{-\int_0^t r(s)\ ds},$$

where r is the short rate, then we have the familier *risk neutral valuation formula.*

Theorem 11.19. (Bjork) (Risk Neutral Valuation Formula) *Assuming the existence of a short rate, the pricing formula takes the form*

$$\Pi_t[X] = E^Q\left[e^{-\int_0^T r(s)\ ds} X \middle| \mathcal{F}_t\right], \tag{11.42}$$

where Q is the (not necessarily unique) martingale measure with the bank account as the numeraire.

Definition 11.20. (Bjork) *A **zero coupon bond** with **maturity date** T, also called a T-bond, is a contract which guarantees the holder one dollar to be paid on the date T. The price at time t of a bond with maturity date T is denoted by $p(t,T)$.*

Proposition 11.21. (Bjork) *The price of a zero coupon T-bond is given by*

$$p(t,T) = E^Q\left[e^{-\int_t^T r(s)\ ds} \middle| \mathcal{F}_t\right], \tag{11.43}$$

and in particular we have $p(T,T) = 1$ for all $T \geqslant 0$ (eq. 11.44).

5.6.3 Stochastic Discount Factors

Definition 11.22. (Bjork) *Assume the existence of a short rate r. For any fixed martingale measure Q, let the likelihood process L be defined by*

$$L_t = \frac{dQ}{dP},\ on\ \mathcal{F}_t. \tag{11.48}$$

*The **stochastic discount factor** (SDF) process $\mathbf{M}$, corresponding to Q, is defined as*

$$\mathbf{M}_t = e^{-\int_0^t r(s)\ ds} L_t \quad \left(= \frac{1}{B_t} \cdot L_t\right). \tag{11.49/50}$$

Proposition 11.23. (Bjork) *Assume absence of arbitrage. With notation as above, the following hold:*

- *For any sufficiently integrable T-claim X, the arbitrage free price is given by*

$$\Pi_t[X] = E^P\left[\frac{\mathbf{M}_T}{\mathbf{M}_t} X \ \middle|\ \mathcal{F}_t\right]. \tag{11.51}$$

- *For any arbitrage free asset price process S (derivative or underlying) the process $\mathbf{M}_t S_t$ is a (local) P-martingale.*
- *The P-dynamics of $\mathbf{M}$ are given by*

$$d\mathbf{M}_t = -r_t \mathbf{M}_t\ dt + \frac{1}{B_t}\ dL_t. \tag{11.53}$$

5.6.4 Summary

Theorem 11.24. (Bjork) (First Fundamental Theorem) *The market model is free of arbitrage if and only if there exists a **martingale measure**, i.e. a measure $Q \sim P$ such that the processes*

$$\frac{S_t^0}{S_t^0}, \frac{S_t^1}{S_t^0}, ..., \frac{S_t^N}{S_t^0}$$

are (local) martingales under Q.

Proposition 11.25. (Bjork) *If the numeraire S^0 is the money account, i.e.*

$$S_t^0 = e^{\int_0^t r(s)\ ds},$$

where r is the (possibly stochastic) short rate, and if we assume that all processes are Brownian driven, then a measure $Q \sim P$ is a martingale measure if and only if all assets $S^0, S^1, ..., S^N$ have the short rate as their local rates of return, i.e. if the Q-dynamics are of the form

$$dS_t^i = S_t^i r_t\ dt + S_t^i \sigma_t^i\ dW_t^Q, \tag{11.54}$$

where W^Q is a (multidimensional) Q-Brownian motion.

Theorem 11.26. (Bjork) (Second Fundamental Theorem) *Assuming absence of arbitrage, the market model is complete if and only if the martingale measure Q is unique.*

Proposition 11.27. (Bjork)

1. *In order to avoid arbitrage, X must be priced according to the formula*

$$\Pi_t[X] = S_t^0 E^Q\left[\frac{X}{S_T^0}\ \middle|\ \mathcal{F}_t\right], \tag{11.55}$$

   where Q is a martingale measure for $[S^0, S^1, ..., S^N]$, with S^0 as the numeraire.
2. *In particular, we can choose the bank account B_t, as the numeraire. Then B has the dynamics*

$$dB_t = r_t B_t\ dt, \tag{11.56}$$

   where r is the (possibly stochastic) short rate process. In this case the pricing formula above reduces to

$$\Pi_t[X] = E^Q\left[e^{-\int_t^T r(s)\ ds} X\ \middle|\ \mathcal{F}_t\right]. \tag{11.57}$$

3. *As a special case, the price of a zero coupon T-bond is given by*

$$p(t,T) = E^Q\left[e^{-\int_t^T r(s)\ ds}\ \middle|\ \mathcal{F}_t\right]. \tag{11.58}$$

4. *Defining the stochastic discount factor $\mathbf{M}$ by $\mathbf{M}_t = B_t^{-1} L_t$ we also have the pricing formula.*

$$\Pi_t[X] = E^Q\left[\frac{\mathbf{M}_T}{\mathbf{M}_t} X\ \middle|\ \mathcal{F}_t\right]. \tag{11.59}$$

5. *Different choices of Q will generically give rise to different price processes for a fixed claim X. However, if X is attainable then all choices of Q will produce the same price process, which then is given by*

$$\Pi_t[X] = V_t^h, \tag{11.60}$$

   where h is the hedging portfolio. Different choices of hedging portfolios (if such exist) will produce the same price process.
6. *In particular, for every replicable claim X it holds that*

$$V_t^Q = E^Q\left[e^{-\int_t^T r(s)\ ds} X\ \middle|\ \mathcal{F}_t\right]. \tag{11.61}$$

5.7 Mathematics of the martingale approach

5.7.1 Martingale representation theorem

Theorem 12.1. (Bjork) (Representation of Brownian Functionals) *Let W be a d dimensional Brownian motions, and let X be a random variable such that*

- $X \in \mathcal{F}_T^W$,
- $E[|X|] < \infty$.

Then there exist uniquely determined $\mathcal{F}_t^W$-adapted processes $h^1, ..., h^d$, such that X has the representation

$$X = E[X] + \sum_{i=1}^{d} \int_0^T h_s^i \, dW_s^i. \tag{12.2}$$

Under the additional assumption

$$E[X^2] < \infty,$$

then $h^1, ..., h^d$ are in $\mathcal{L}^2$.

Theorem 12.2. (Bjork) (The Martingale Representation Theorem) *Let W be a d dimensional Brownian motions, and assume that the filtration $\mathbf{F}$ is defined as*

$$\mathcal{F}_t = \mathcal{F}_t^W, \qquad t \in [0, T].$$

Let M be any $\mathcal{F}_t$-adapted martingale. Then there exist uniquely determined $\mathcal{F}_t$-adapted processes $h^1, ..., h^d$ such that M has the representation

$$M_t = M_0 + \sum_{i=1}^{d} \int_0^t h_s^i \, dW_s^i, \qquad t \in [0, T]. \tag{12.9}$$

If the martingale M is square integrable, then $h^1, ..., h^d$ are in $\mathcal{L}^2$.

5.7.2 Girsanov theorem

Theorem 12.3. (Bjork) (The Girsanov Theorem) *Let W be a d dimensional P-Brownian motion on $(\Omega, \mathcal{F}, P, \mathbf{F})$ and let φ be any d-dimensional adapted column vector process. Choose a fixed T and define the process L on $[0, T]$ by*

$$dL_t = \varphi_t^\top L_t \, dW_t, \tag{12.16}$$

$$L_0 = 1, \tag{12.17}$$

i.e.

$$L_t = \exp\left\{ \int_0^t \varphi_s^\top \, dW_s - \frac{1}{2} \int_0^t \|\varphi_s\|^2 \, ds \right\}.$$

Assume that

$$E^P[L_T] = 1, \tag{12.18}$$

and define the new probability measure Q on $\mathcal{F}_T$ by

$$L_T = \frac{dQ}{dP}, \qquad \text{on } \mathcal{F}_T. \tag{12.19}$$

Then

$$dW_t = \varphi \, dt + dW_t^Q, \tag{12.20}$$

where W^Q is a d dimensional Q-Brownian motion or equivalently

$$W_t^Q = W_t - \int_0^t \varphi_s \, ds \tag{12.21}$$

is a standard Q-Brownian motion.

We will often refere to φ as the **Girsanov kernel** of the measure transformation. Furthermore, we have written on component form above and the L dynamics will have the form

$$dL_t = L_t \sum_{i=1}^d \varphi_t^i \, dW_t^i,$$

and L will have the explicit form

$$L_t = \exp\left\{ \sum_{i=1}^d \int_0^t \varphi_s^i \, dW_s^i - \frac{1}{2} \int_0^d \sum_{i=1}^d (\varphi_s^i)^2 \, ds \right\}.$$

The conclusion of the Girsanov Theorem is thwn that we can write

$$dW_t^i = \varphi_t^i \, dt + dW_t^{Q,i},$$

for $i = 1, ..., d$ where $W_t^{Q,1}, ..., W_t^{Q,d}$ are independent standard Brownian motions under Q.

Definition 12.4. (Bjork) *For any Brownian motion W and any kernel process φ, the **Doleans exponential** process $\mathcal{E}$ is defined by*

$$\mathcal{E}(\varphi \bullet W)_t = \exp\left\{ \int_0^t \varphi_s^\top \, dW_s - \frac{1}{2} \int_0^t \|\varphi\|_s^2 \, ds \right\}. \tag{12.24}$$

Lemma 12.5. (Bjork) (The Novikov Condition) *Assume that the Girsanov kernel φ is such that*

$$E^P\left[e^{\frac{1}{2}\int_0^T \|\varphi_t\|^2 \, dt} \right] < \infty. \tag{12.27}$$

Then L is a martingale and in particular $E^P[L_T] = 1$.

Theorem 12.6. (Bjork) (The Converse of the Girsanov Theorem) *Let W be a d-dimensional standard P-Brownian motion on $(\Omega, \mathcal{F}, P, \mathbf{F})$ and assume that*

$$\mathcal{F}_t = \mathcal{F}_t^W, \ \forall t.$$

Assume that there exists a probability measure Q such that $Q << P$ on $\mathcal{F}_T$. Then there exists an adapted process φ such that the likelihood process L has the dynamics

$$\begin{aligned} dL_t &= L_t \varphi_t^\top \, dW_t, \\ L_0 &= 1. \end{aligned}$$

This gives us a recipe to transform dynamics of Ito processes under the measure Q as we may rewrite the dynamics of the Brownian motion. We therefore have for an Ito process X with dynamics

$$dX_t = \mu(t, X_t) X_t \, dt + \sigma(t, X_t) X_t \, dW_t,$$

may be transformed under Q as

$$\begin{aligned} dX_t &= \mu(t, X_t) X_t \, dt + \sigma(t, X_t) X_t \, dW_t \\ &= \mu(t, X_t) X_t \, dt + \sigma(t, X_t) X_t \, (\varphi_t \, dt + dW_t^Q) \\ &= (\mu(t, X_t) + \varphi_t) \, X_t \, dt + \sigma(t, X_t) X_t \, dW_t^Q. \end{aligned}$$

This may lead us into deducing that

$$\mu(t, X_t) + \sigma(t, X_t)\varphi_t = r_t \iff \varphi_t = \frac{r_t - \mu(t, X_t)}{\sigma(t, X_t)}.$$

We furthermore have the Levy characterisation of a Brownian motion.

Theorem. (Remark FinKont) (Levy Characterisation of Brownian motion) *Let X_t be an Ito process with $X_0 = 0$. Then X_t is a Brownian motion if and only if the two processes X_t and $X_t^2 - t$ are continuous martingales.*

5.8 Black-Scholes model - martingale approach

We consider the standard Black-Scholes model with a single risk free asset and risky asset with dynamics

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{13.1}$$
$$dB_t = rB_t \, dt. \tag{13.2}$$

We want check whether the model is arbitrage free on any time interval $[0, T]$, and find a (perhaps unique) martingale measure such that we may apply the fundamental pricing theorem 1 and 2. From Girsanov this endavour is equivalent with searching for a (perhaps unique) Girsanov kernel φ. We therefore define as usual the likelihood process

$$dL_t = \varphi_t L_t \, dW_t,$$

and setting $dQ = L_T \, dP$ on $\mathcal{F}_T$, we know from Girsanov theorem that

$$dW_t = \varphi_t \, dt + dW_t^Q.$$

Inserting into the Black-Scholes model we have

$$dS_t = S_t(\mu + \sigma\varphi_t) \, dt + \sigma S_t \, dW_t^Q.$$

We know that for Q to be a martingale measure we know that the local rate of return under Q of S must be the short rate r i.e. we have

$$\mu + \sigma\varphi_t = r \iff \varphi_t = \frac{r-\mu}{\sigma} = -\frac{\mu - r}{\sigma}, \tag{13.3}$$

and so we see the Girsanov kernel is **constant and deterministic**. The process has the economic interpretation that the Girsanov kernel is the risk premium per unit volatility.

Lemma 13.1. (Bjork) *The Girsanov kernel φ is given by*

$$\varphi = -\lambda$$

where the market price of risk λ is defined by

$$\lambda = \frac{r-\mu}{\sigma}.$$

We therefore have determined *a martingale* and so we have the result.

Theorem 13.2. (Bjork) *The Black-Scholes model above is arbitrage free.*

We could in general have that μ, σ, r are adapted processes. If this is the case we would have to show the Nivokov condition.

Pricing then of any T-claim X then is given by the risk neatral pricing formula

$$\Pi_t[X] = e^{-r(T-t)} E^Q[X \mid \mathcal{F}_t], \tag{13.7}$$

where the Q dynamics of S has local drift r and volatility from the Q-brownian motion W^Q.

Theorem 13.3. (Bjork) *The Black-Scholes model above is complete. This also holds for the more general model where r, μ, σ are adapted processes.*

Hedging is the possible by considiering a T claim with

$$E^Q\left[\frac{X}{B_T}\right] < \infty.$$

Notice the numeraire B_t as in the normalized Z-economy. Consider now the Q-martingale

$$M_t = E^Q\left[\frac{X}{B_T} \;\middle|\; \mathcal{F}_t\right], \tag{13.9}$$

and it now follows from lemma 11.15 the the model is complete if we can find a process h_t^1 such that

$$dM_t = h_t^1 \; dZ_t^i. \tag{13.10}$$

In order to prove existance of such a process h^1 we use the Martingale Representation Theorem 12.2, which says that there *exists* a process g_t such that

$$dM_t = g_t \; dW_t^Q. \tag{13.11}$$

We can now combine these two equation by the following Q dynamics

$$dZ_t^1 = Z_t^1 \sigma \; dW_t^Q. \tag{13.12}$$

Hence we have

$$dM_t = h_t^1 Z_t^1 \sigma \; dW_t^Q = g_t \; dW_t^Q \;\; \Rightarrow \;\; h_t^1 = \frac{g_t}{Z_t^1 \sigma}.$$

Theorem 13.4. (Bjork) *In the Black-Scholes model every T-claim X satisfying*

$$E^Q \left[\frac{X}{B_T} \right] < \infty$$

can be replicated. The replicating portfolio is given by

$$h_t^1 = \frac{g_t}{Z_t^1 \sigma}, \tag{13.13}$$

$$h_t^0 = M_t - h_t^1 Z_t^1, \tag{13.14}$$

where M is defined by the above and g is defined by above.

If the T-claim is simple that is $X = \Phi(S_T)$ we may solve a boundary value problem with Feymann-Kac to arrive at the familiar result.

Proposition 13.5. (Bjork) *In the Black-Scholes model every T-claim on the form $X = \Phi(S_T)$. Then X can be replicated by the portfolio*

$$h_t^0 = \frac{F(t, S_t) - S_t \frac{\partial F}{\partial s}(t, S_t)}{B_t}, \tag{13.15}$$

$$h_t^1 = \frac{\partial F}{\partial s}(t, S_t), \tag{13.15}$$

where F solves the ***Black-Scholes equation***

$$\frac{\partial F}{\partial t}(t,s) + rs\frac{\partial F}{\partial s}(t,s) + \frac{1}{2}\sigma^2 s^2 \frac{\partial^2 F}{\partial s^2}(t,s) - rF(t,s) = 0, \tag{13.16}$$

$$F(T,s) = \Phi(s). \tag{13.16}$$

Furthermore the value process for the replicating portfolio is given by

$$V_t = F(t, S_t).$$

5.9 Multidimensional models

We specify the general model by the assumptions below.

Assumption 14.0.1 *We assume the following:*

- *There are n risky assets $S^1, ..., S^n$.*
- *Under the objective probability measure P, the S-dynamics are given by*

$$dS_t^i = \mu_t^i S_t^i \, dt + S_t^i \sum_{j=1}^{N} \sigma_t^{ij} \, dW_t^j, \tag{14.1}$$

  for $i = 1, ..., n$.
- *The coefficients processes μ^i and σ^{ij} above are assumed to be adapted.*
- *We have a standard risk free asset with price process B with dynamics*

$$dB_t = r_t B_t \, dt, \tag{14.2}$$

  where the short rate process r is assumed to be an adapted stochastic process.

We can use the representation of μ^i, σ^{ij} and S^i as vectors and matrices on the form.

$$\mu = \begin{bmatrix} \mu^1 \\ \vdots \\ \mu^n \end{bmatrix}, \; \sigma = \begin{bmatrix} \sigma^{1,1} & \cdots & \sigma^{i,N} \\ \vdots & \ddots & \vdots \\ \sigma^{n,1} & \cdots & \sigma^{n,N} \end{bmatrix}, \; D(S) = \begin{bmatrix} S^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S^n \end{bmatrix}.$$

And so we have the model on compact form.

$$dS_t = D(S_t)\mu_t \, dt + D(S_t)\sigma_t \, dW_t, \tag{14.3}$$

$$dB_t = r_t B_t \, dt. \tag{14.4}$$

Now using Girsanov Theorem we can define the prospective likelihood process L by

$$dL_t = L_t \varphi_t^\top \, dW_t, \tag{14.5}$$

$$L_0 = 1, \tag{14.6}$$

where φ is an adapted N-dimensional process. Then our candidate martingale measure Q is given by $dQ = L_t \, dP$ on $\mathcal{F}_t$ and the Girsanov theorem give the dynamics

$$dW_t = \varphi_t \, dt + dW_t^Q, \tag{14.7}$$

where W^Q is a standard Q-Brownian motion. Inserting into the P-dynamics we obtain.

$$dS_t = D(S_t)[\mu_t + \sigma_t \varphi_t] \, dt + D(S_t)\sigma_t \, dW_t^Q. \tag{14.8}$$

The from (11.54) we know that Q is a martingale measure if and only if the local rate of return (the dt-term) is the short interest rate i.e. if and only if

$$\mu_t + \sigma_t \varphi = \mathbf{r}_t, \tag{14.9}$$

where $\mathbf{r}_t = [r, ..., r]^\top \in \mathbb{R}^n$. We then have that

$$\sigma_t \varphi_t = \mathbf{r}_t - \mu_t, \tag{14.11}$$

where we want to solve for φ. Thus we may write a condition for the absence of arbitrage in linear algebra terms.

Proposition 14.1. (Bjork) *A necessary condition for absence of arbitrage is that*

$$\mathbf{r}_t - \mu_t \in Im[\sigma_t]$$

with probability one for each t. A sufficient condition for absence of arbitrage is that there exists a process φ which solves (14.11) and such that L is a martingale.

Note that it is not enought for φ to solce (14.11). We also need that L is a martingale.

Definition 14.2. (Bjork) *A Girsanov kernel φ is said to be **admissible** if it generates a martingale measure, i.e. it solves (14.11) and L is a true martingale.*

Definition 14.3. (Bjork) *The model above is said to be **generically arbitrage free** if it is arbitrage free for every choice of μ.*

Proposition 14.4. (Bjork) *Disregarding integrability problems the model is generically arbitrage free if and only if, for each $t \leqslant T$ and P-a.s., the mapping*

$$\sigma_t : \mathbb{R}^B \to \mathbb{R}^n$$

is surjective, i.e. if and only if the volatility matrix σ_t has rank n.

Proposition 14.5. (Bjork) *Assume that the model i generically arbitrage free and that the filtration $\mathbf{F}$ is defined by*

$$\mathcal{F}_t = \mathcal{F}_t^W. \tag{14.14}$$

Then, disregarding integrability problems, the model is complete if and only if $n = N$ and the volatility matrix σ_t is invertible P-a.s. for each $t \leqslant T$.

Assumption 14.3.1. (Bjork) *We assume that the model is generically free of arbitrage, i.e. that*

$$Im[\sigma_t] = \mathbb{R}^n, \tag{14.16}$$

for all t and with probability one. We also assume that the model is purely Brownian driven, i.e. that $\mathcal{F}_t = \mathcal{F}_t^W$.

Proposition 14.6. (Bjork) *Under assumption 14.3.1 the model is complete if and only if*

$$Im[\sigma_t^\top] = \mathbb{R}^N. \tag{14.23}$$

If the model is complete then, using the notation of chapter 11, the replicating portfolio $[h^0, h^S]$ is given by

$$h_t^S = g_t \sigma_t^{-1} D^{-1}(Z_t), \tag{14.24}$$

$$h_t^2 = M_t - h_t Z_t. \tag{14.25}$$

With M_t defined by

$$M_t = E^Q\left[\frac{\mathcal{X}}{B_T} \;\middle|\; \mathcal{F}_t \right]. \tag{14.17}$$

Theorem 14.7. (Bjork) (The Second Fundamental Theorem) *Under assumptions 14.3.1 the model is complete if and only if the martingale measure is unique. This is equivalent with the statements: $Ker[\sigma_t] = \{0\}$, $Im[\sigma_t^\top] = \mathbb{R}^N$ and σ_t^{-1} exists (i.e. σ_t is invertible).*

Pricing of any T-claim $\mathcal{X}$ is now given by the risk neutral valuation formula

$$\Pi_t[\mathcal{X}] = E^Q\left[e^{-\int_t^T r_u \, du} \mathcal{X} \;\middle|\; \mathcal{F}_t \right], \tag{14.27}$$

where Q is some choice of martingale measure. Alternatively we can write the price as

$$\Pi_t[\mathcal{X}] = E^Q\left[\frac{\mathbf{M}_T}{\mathbf{M}_t} \mathcal{X} \;\middle|\; \mathcal{F}_t \right], \tag{14.29}$$

where $\mathbf{M}$ is the stochastic discount factor, defined by

$$\mathbf{M}_t = \frac{1}{B_t} L_t.$$

If we have a simple claim i.e. of the form $\mathcal{X} = \Phi(S_t)$ and if S is Markovian we have

$$e^{-r(T-t)}E^Q[\Phi(S_t) \mid \mathcal{F}_t] = e^{-r(T-t)}E^Q[\Phi(S_t) \mid S_t],$$

and then the pricing process must be of the form $\Pi_t[\Phi] = F(t, S_t)$. We then have F to be the solutions to the PDE

$$F_t(t,s) + \sum_{i=1}^{n} rs_i F_i(t,s) + \frac{1}{2}\text{tr}\left\{\sigma^\top D(S)F_{ss}D(S)\sigma\right\} - rF(t,s) = 0, \tag{14.31}$$

$$F(T,s) = \Phi(s), \tag{14.31}$$

where $F_i = \frac{\partial F}{\partial s_i}$ and F_{ss} denotes the Hessian matrix. Furthermore, $\text{tr}(A)$ denotes the trace of A i.e. the sum of the diagonal. We have that the hedging portfolio has value process $V_t^h = F(t, S_t)$ with dynamics

$$dV_t^h = \sum_{i=1}^{n} F_i(t, S_t)\ dS_t^i.$$

Then we must gave the solution

$$h_t^i = \frac{\partial F}{\partial s_i}(t, S_t), \quad i = 1, ..., n, \tag{14.32}$$

$$h_t^0 = \frac{1}{B_t}\left\{F(t, S_t) - \sum_{i=1}^{n} \frac{\partial F}{\partial s_i}(t, S_t)\ S_t^i\right\}. \tag{14.33}$$

Proposition 14.8. (Bjork) *With L-dynamics as in $dL_t = L_t\varphi_t^\top\ dW_t$, the $\mathbf{M}$-dynamics are*

$$d\mathbf{M}_t = -r_t\mathbf{M}_t\ dt + \mathbf{M}_t\varphi_t^\top\ dW_t, \tag{14.39}$$

or alternatively in terms of the market price of risk $\lambda_t = -\varphi_t$

$$d\mathbf{M}_t = -r_t\mathbf{M}_t\ dt - \mathbf{M}_t\lambda_t^\top\ dW_t. \tag{14.40}$$

Proposition 14.9. (Bjork) (The Hansen-Jagannathan Bounds) *Assume generic absence of arbitrage. Then the following holds for all assets, underlying or derivative, and for all admissible Girsanov kernels φ, and market prices of risk λ.*

$$\frac{|\mu_t^p - r_t|}{\|\sigma_t^p\|} \leqslant \|\varphi_t\|, \qquad \frac{|\mu_t^p - r_t|}{\|\sigma_t^p\|} \leqslant \|\lambda_t\|. \tag{14.42}$$

Chapter 6

Basic Non-Life Insurance Mathematics

Noget indhold

Chapter 7

Stochastic Processes in Non-Life Insurance Mathematics

Noget indhold

Chapter 8

Topics in Non-Life Insurance Mathematics

Noget indhold

Chapter 9

Probabilistic Machine Learning

9.1 Supervised Learning

In this chapter we restrict our selves to the area of *Supervised Learning* as we use numerical methods and machine learning algorithms to estimate models in a restricted framework. Take for instance the random forest, this algorithm's estimation method is perfectly capable of being written recursively and so no "leaning" is done in the sense, that the calculations are predetermined from the algorithm. We therefore call the area of study supervised learning instead of the wider area of study *machine learning*. Let us define what we mean by supervised learning.

Definition. (Supervised Learning) *Supervised learning is a field in machine learning that works with labeled data, i.e. data consisting of a set of features X, and a response Y. The goal is to learn a function m^* that maps a given input x to an output y.*

We will in this chapter only use data in the form of a spread sheet e.g.

$$\mathcal{D}_n = (X_i, Y_i)_{i=1,\dots,n} = \left[\begin{array}{cccc|c} X_{11} & X_{12} & \cdots & X_{1p} & Y_1 \\ X_{21} & X_{22} & \cdots & X_{2p} & Y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} & Y_n \end{array}\right],$$

we could however consider any data that may be interpreted by computer software.

The setting is then; assume that we have n independent copies of the random variable $D = (X, Y) \in \mathcal{X} \times \mathcal{Y}$, where X is p-dimensional and Y is one-dimensional. We make no assumption on whether X_j, $j = 1, .., p$ and Y are discrete or continuous, however in concrete cases this will be specified. We combine the sample of the n observations in the matrix $\mathcal{D}_n = (X_i, Y_i)_{i=1,\dots,n}$ being a $n \times p + 1$ matrix as in the above. We call $\mathcal{D}_n$ the **training data**.

This specification does indeed imply that $D_i = (X_i, Y_i)$ are iid. This is actually a bit controversial as we would expect that the distribution of D will shift over time. For instance, the distribution of ages in a population changes over time and have been more right skewed as humanity advances. This may be accounted for by transforming the data such that the distribution becomes the same. This may be done in a variety of ways some example include: 1) transforming to uniform variable with the time dependent distribution F_t, 2) normalizing using a price index and so forth. One should therefore start any analysis by ensuring that the data a given algorithm is trained on is iid.

The job becomes finding a good estimator such that we may predict Y given X i.e. $Y \mid X$. Let us define what an estimator is.

Definition. (Estimator) *Consider a training dataset $\mathcal{D}_n$. A estimator m is a function-valued mapping that takes $\mathcal{D}_n$ as input and associates a function $m_n : \mathcal{X} \to \mathcal{Y}$ i.e. m takes the form*

$$m(\mathcal{D}_n) = m_n : \mathcal{X} \to \mathcal{Y}$$

the class of estimators is called $\mathcal{G}$*.*

9.1.1 What is a good estimator?

It is easy to construct an estimator $\hat{m}$ for instance by maximum likelihood esitmation, bayes optimization or simply by taking conditional expectations. We are however interested in two problems:

1. What is the best class of estimators $\mathcal{G}_0 \subset \mathcal{G}$,
2. In the subset of estimators $m \in \mathcal{G}_0$, what is then the best estimator.

We will now discuss the meaning of being the *"best"* estimator. Obviously, the first type of consideration is regarding the inherent restrictions of some algorithm, where the second is the problem of error comming from the restriction that we do not have infinite observations. We therefore have two problems namely the **inductive bias** from the class $\mathcal{G}_0$ and the **estimation error** from the available data. A typical problem is that the larger a class $\mathcal{G}_0$ gives low inductive bias we then may not be able to estimate anything and so the estimation error will be large. The converse also applies.

Definition. *Let* $D = (X, Y)$ *be a random variable on the background space* $(\Omega, \mathcal{F}, P)$*. Then we define the following:*

- *A* ***decision rule*** *is a deterministic function* $m : \mathcal{X} \to \mathcal{Y}$*,*
- *A* ***loss function*** *is a deterministic function* $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$*,*
- *The* ***risk*** *of a decision rule* m *is given a loss function* L *is* $r(m) = E[L(Y, m(X))]$*.*

Notice that in the definition of risk we see that m is included inside the expectation. This means in particular that the training data $\mathcal{D}_n$ is also accounted for i.e. by the tower rule we have

$$r(m) = \mathbb{E}[L(Y, m(X))] = \mathbb{E}\left[\mathbb{E}\Big[L(Y, m_n(X)) \;\Big|\; \mathcal{D}_n\Big]\right] := \mathbb{E}\left[R(m)\right].$$

Some widely used loss function include

- Quadratic loss function: $L(y_1, y_2) = (y_1 - y_2)^2$,
- Poisson Deviance: $L(y_1, y_2) = 2\left(y_1 \log \frac{y_1}{y_2} - y_1 + y_2\right)$,
- Binary loss function: $L(y_1, y_2) = 1_{y_1 \neq y_2}$.

Given a loss function L we may find a (possibly non-unique) solution m^* that minimizes $R(m)$. We call this the **Bayes estimator**. The quantity $R(m^*)$ is called the **Bayes risk**. On some special case loss functions we may determine the unique solution.

Lemma. *Assume* Y *is* L_2 *(square integrable), then the decision function that minimized the risk for the quadratic loss function is*

$$m^* = \operatorname*{argmin}_m \mathbb{E}[(Y - m(X))^2] = E[Y \mid X = x]$$

i.e. the conditional expectation.

Proof.

Consider the loss function $L(y_1, y_2) = (y_1 - y_2)^2$ i.e. the L^2 loss function. For any estimator $m(X) \in \mathcal{G}$ we have

$$\begin{aligned} R(m(X)) &= \mathbb{E}[L(m(X), Y) \mid X] = \mathbb{E}[(m(X) - Y)^2 \mid X] \\ &= \mathbb{E}[m(X)^2 \mid X] + \mathbb{E}[Y^2 \mid X] - 2\mathbb{E}[m(X)Y \mid X] \\ &= m(X)^2 + \mathbb{E}[Y^2 \mid X] - 2m(X)\mathbb{E}[Y \mid X] \end{aligned}$$

We see that the risk is minimized for the m that minimizes $m(X)^2 - 2m(X)\mathbb{E}[Y \mid X]$. The first order condition is then

$$\frac{\partial}{\partial m} R(m(X)) = 2m(X) - 2\mathbb{E}[Y \mid X] = 0$$

hence giving that

$$m(X) = \mathbb{E}[Y \mid X]$$

as desired. ■

Remarks on the L^2-loss. The loss function $L(y_1, y_2) = (y_1 - y_2)^2$ i.e. the L^2 loss function gives some nice interpretations. We know that L as a norm on $\mathbb{R}$ forms a Hilbert Space. This means, for instance, that we have some nice geometric interpretations, but more importantly we have alot of tools from linear algebra. We know that the projection m^* onto the space $\mathcal{G}$ and so any vector $\hat{m}(X) - m^*(X)$ is orthogonal to $Y - m^*(X)$. In particular, this gives that

$$\begin{aligned} r(\hat{m}(X)) &= \mathbb{E}[L(\hat{m}(X), Y)] = \mathbb{E}\left[(\hat{m}(X) - Y)^2\right] \\ &= \mathbb{E}\left[(\hat{m}(X) - m^*(X) + m^*(X) - Y)^2\right] \\ &\overset{(\dagger)}{=} \mathbb{E}\left[(\hat{m}(X) - m^*(X))^2\right] + \mathbb{E}\left[(m^*(X) - Y)^2\right] \\ &= \mathbb{E}\left[(\hat{m}(X) - m^*(X))^2\right] + r(m^*(X)) \end{aligned}$$

Using in (†) that $\hat{m}(X) - m^*(X)$ and $m^*(X) - Y$ are orthogonal with the last equation simply being the definition of risk. Rearringing the above gives

$$r(\hat{m}(X)) - r(m^*(X)) = \mathbb{E}\left[(\hat{m}(X) - m^*(X))^2\right] = \mathrm{MSE}(\hat{m}(x))$$

We can further write out the Mean Squared Error in terms of variance and bias.

$$\begin{aligned} \mathrm{MSE}(\hat{m}(x)) &= \mathbb{E}\left[(\hat{m}(X) - m^*(X))^2\right] \\ &= \mathbb{E}\left[(\hat{m}(X) - \mathbb{E}[\hat{m}(X)|X])^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{m}(X)|X] - m^*(X))^2\right] \\ &= \mathrm{Var}(\hat{m}(X)) + \mathrm{Bias}^2(\hat{m}(X)). \end{aligned}$$

Lemma. *Assume $\mathcal{Y} = \{1, ..., K\}$, the decision function that minimizes the risk for the binary loss function satisfies*

$$m^* = \underset{m}{\operatorname{argmin}}\, \mathbb{E}[1_{Y \neq m(X)}] = \underset{m}{\operatorname{argmin}}\, \mathbb{P}(Y \neq m(X)) = \underset{k=1,..,K}{\operatorname{argmax}}\, \mathbb{P}(Y = k \mid X = x).$$

We can now define the prediction risk and the generalization error which relates to the balance of a sufficiently large class $\mathcal{G}_0$ and how effective the optimal estimator is conditional on the class $\mathcal{G}_0$.

Definition. (Conditional risk) *Let $\mathcal{D}_n$ be some training data. Given an estimator $\hat{m}_n$ we call*

$$R(\hat{m}_n) = \mathbb{E}[L(Y, \hat{m}_n(X)) \mid \mathcal{D}_n]$$

the prediction risk or conditional generalized error.

Definition. (Risk) *We call*

$$r(\hat{m}_n) = \mathbb{E}[R(\hat{m}_n)],$$

the prediction risk or generalized error.

9.1.2 Excess risk

Definition. (Excess Risk) *Consider the set $\mathcal{G}$ be the set of all measurable estimators. Fix a subset $\mathcal{G}_0 \subset \mathcal{G}$. Given some training data $\mathcal{D}_n$ consider the Bayes estimator restricted to $\mathcal{G}_0$ denoted by $\hat{m}_n$ and the unconditional Bayes estimator restricted to $\mathcal{G}$ we define the quantity*

$$R(\hat{m}_n) - r(m^*)$$

or

$$\mathbb{E}[R(\hat{m}_n) \mid X_1, ..., X_n] - r(m^*),$$

as the excess risk. This is the difference between the generalization error and the risk obtained by an optimal decision function.

In the context of the above definition we can decompose the risk associated with the optimal estimator $\hat{m}_n \in \mathcal{G}_0$ into the estimation error and the inductive bias.

$$R(\hat{m}_n) - r(m^*) = \underbrace{\left[R(\hat{m}_n) - \inf_{m \in \mathcal{G}_0} R(m)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{m \in \mathcal{G}_0} R(m) - R(m^*)\right]}_{\text{inductions bias/approximation error}}.$$

where we have to balance the trade-off with a larger $\mathcal{G}_0$ infer a lower induction bias but larger estimation error and a smaller class $\mathcal{G}_0$ infer a lower estimation error but larger induction bias.

Definition. (Empirical risk and empirical risk minimizer) *Given training data $\mathcal{D}_n$ and a loss function L, we call*

$$\hat{R}_n(m) := \sum_{i=1}^{n} L(Y_i, m(X_i))$$

*the **empirical risk**. Given an additional function class $\mathcal{G}_0$,*

$$\operatorname*{argmin}_{m \in \mathcal{G}_0} \hat{R}_n(m) = \operatorname*{argmin}_{m \in \mathcal{G}_0} \sum_{i=1}^{n} L(Y_i, m(X_i))$$

*is called **empirical risk minimizer** or (standard leaner).*

For larger function classes $\mathcal{G}_0$ the empirical risk minimizer might not be unique and possibly too noisy. In this case one sometimes adds a penalty term $J_\lambda : \mathcal{G} \to \mathbb{R}_+$ that penalizes the complexity of m and minimizes the penalized empirical risk:

$$\operatorname*{argmin}_{m \in \mathcal{G}_0} \hat{R}_{n,\lambda} := \operatorname*{argmin}_{m \in \mathcal{G}_0} \sum_{i=1}^{n} L(Y_i, m(X_i)) + J_\lambda(m).$$

If J_λ and $\hat{R}_n$ is convex one can show that

$$\operatorname*{argmin}_{m \in \mathcal{G}_0} \hat{R}_{n,\lambda} = \operatorname*{argmin}_{m \in \mathcal{G}_\eta} \hat{R}_n$$

for the class $\mathcal{G}_\eta = \{m \in \mathcal{G}_0 \mid J_\lambda(m) \leqslant \eta\}$. Some penalty terms could be

- $J_\lambda(m) = \lambda \int m''(x)\, dx$,
- $J_\lambda(m) = \lambda \int |m'(x)|\, dx$,
- $J_\lambda(m) = \lambda \int (m(x))^2\, dx$.

Proposition. (Probability bounds) *Let $\tilde{m} = \operatorname*{argmin}_m r(m)$. We have*

$$r(\hat{m}_n) - r(\tilde{m}) \leqslant 2 \sup_{m \in \mathcal{G}_0} \left|\hat{R}_n(m) - r(m)\right|,$$

and for all $\lambda \in \Lambda$,

$$r(\hat{m}_{n,\lambda}) - r(\tilde{m}) \leqslant 2 \sup_{m \in \mathcal{G}_0} \left|\hat{R}_{n,\lambda}(m) - r(m)\right| + J_\lambda(\tilde{m}) - J_\lambda(\hat{m}_n).$$

Definition. *We say $\hat{m}_n$ is ε-accurate with probabiltiy $1 - \delta$, if*

$$P\Big(R(\hat{m}_n) - \inf_{m \in \mathcal{G}} r(m) > \varepsilon\Big) < \delta.$$

9.2 Training, Validating and Testing

When deciding which method too choose for a given task, one may like to pick the method with the smallest generalization error. However, most machine learning methods depend on hyper parameters and one may first need to decide which hyper parameters are best for the given task. In short: We would like to compare the generalization error of optimally tuned machine learning methods given our data.

Definition. (Training and test set) *One often randomly divides the given data into training data and test data:*

- $\mathcal{D}_n = (X_i, Y_i)_{i=1,...,n}$ *(training data)*
- $\mathcal{T}_m = (\tilde{X}_j, \tilde{Y}_j)_{j=1,...,m}$ *(test data)*

with $n \in [0.8m, 0.95m]$.

Definition. (Training and test error) *The empirical risk on the training data is called training error and on the test data test error.*

Definition. (Validation set) *To tune the hyper parameters for an algorithm, one often randomly divides the given training data into training data (yes: also called training data.) and validation data:*

- $\mathcal{D}_{n_1} = (X_i, Y_i)_{i=\tau(1),...,\tau(n_1)}$ *(traning data)*
- $\mathcal{V}_{n_2} = (\tilde{X}_j, \tilde{Y}_j)_{j=\tau(n_1+1),...,\tau(n)}$ *(validation data)*

where τ *is a randomly picked permutation of* $\{1, ..., n\}$ *and* $n_1 + n_2 = n$ *is respectively the size of the training set and the validation set.*

One can now compare different methods via the following simple algorithm:

1. Split your data into train, validation and test set.
2. For a given method and a rich set of hyper parameter configurations train the method on the training set and compare performance via empirical risk on the validation set.
3. For every method, pick the hyper parameter with the smallest empirical risk.
4. Compare different methods with the chosen hyper parameters on the test set (trained on training+validation set) and pick the method with smallest empirical risk.

Notice that the procedure above is stochastic and has bias and variance both for selecting the optimal hyper parameters and for selecting the optimal method.

- Bias: Bias occurs because the sample sizes used for learning the hyper parameters n_1 are smaller than the actual training size n and also the full data size $n + m$.
- Variance: The results are stochastic because the validation and test set are not of infinite size.

Variance can be reduced by repeating steps 1-4 several times. The most popular method for doing so is (nested) cross validation.

9.2.1 Estimating risk

We want to estimate the generalization error of a method $\hat{m}_{n,\lambda}$ that depends on a fixed hyper parameter λ.

Definition. (M-fold Cross validation) *Given the indices of a data set* $S = \{1, ..., n\}$*, M-fold cross validation follows the following steps:*

1. *Divide the data into* M *disjoint sets* $S_1, ..., S_M$ *of same size. Define* $S_{-l} = \cup_{k \neq l} S_k$ *being the complement to* S_l*.* ($\#S_l = n/M$ *and* $\#S_{-l} = (M-1)n/M$)
2. *For each subdivision* $l = 1, ..., M$ *train the algorithm on* S_{-l} *and denote the estimator* $\hat{m}_\lambda(S_{-l})$*.*
3. *Calculate the cross validated empirical risk*

$$CV(\hat{m}_{n,\lambda}) = \frac{1}{M} \sum_{l=1}^{M} \frac{1}{|S_{-l}|} \sum_{i \in S_{-l}} L\big(Y_i, \hat{m}_\lambda(S_{-l})(X_i)\big)$$

It is a non-trivial discussion what $CV(\hat{m}_{n,\lambda})$ is estimating. Consider the heuristic

$$\begin{aligned} CV(\hat{m}_{n,\lambda}) &= \frac{1}{M}\sum_{l=1}^{M}\frac{1}{|S_{-l}|}\sum_{i\in S_{-l}} L\big(Y_i, \hat{m}_\lambda(S_{-l})(X_i)\big) \\ &\approx \frac{1}{M}\sum_{l=1}^{M} R\big(Y, \hat{m}_\lambda(S_{-l})(X)\big) \\ &\approx \mathbb{E}\left[R\big(Y, \hat{m}_\lambda(S_{-l})(X)\big)\right]. \end{aligned}$$

Where we used the law of large numbers in the first approximation. The last approximation is not so clear because the summands are dependent for $M > 2$.

The bias is minimal for large M since estimation is based on training the algorithm on $(M-1)n/M$ data points. In the case that $M = n$, M-fold cross validation is also known as leave-one-our cross validation. It is however often not practical because of the computational cost. In practice, setting $M = 5$ or $M = 10$ is common choices.

When deciding on an optimal hyper parameter we would like to pick

$$\operatorname*{argmin}_{\lambda\in\Lambda} CV(\hat{m}_{n,\lambda}).$$

But the hyper parameter space Λ is often multi-dimensional and partly continuous. Hence it is infeasible to try out all parameters. Common practice are

- Grid search
- Random search (e.g. pick 200 parameters uniformly random from the parameter space)
- Advanced optimization techniques (i.e. techniques that aim to find the minimzer of a function (here: cross validated empirical risk) without the requirement of knowing the analytical form of the function to optimize.

Above we have discussed how we can use cross validation to pick an optimal parameter for a given method and data set. But how to choose between different methods? One popular way is nested cross validation comprising an inner loop for hyper parameter selection (tuning) and an outer loop for method comparison. Assume that we want to compare J methods $\hat{m}_{n,j}$, $j = 1, ..., J$. then we may use nested cross validation.

Definition. (Nested $M_1 - M_2$ Cross-validation) *Given the indices of a data set $S = \{1, ..., n\}$, nested $M_1 - M_2$ cross validation follows the following steps:*

1. *Divide the data into M_1 disjoint sets $S_1, ..., S_{M_1}$ of same size. Define $S_{-l} = \cup_{k\neq l} S_k$ being the complement to S_l. ($\#S_l = n/M_1$ and $\#S_{-l} = (M_1 - 1)n/M_1$)*
2. *For each $l = 1, ..., M_1$, run M_2-fold cross validation on S_{-l} for all J methods, returning optimal hyper parameters $\hat{\lambda}(j, l)$, $j = 1, ..., J$ and $l = 1, ..., M_1$. (the one with lowest CV is choosen for each (j, l))*
3. *Calculate the cross validated empirical risk*

$$CV(\hat{m}_{n,j}) = \frac{1}{M_1}\sum_{l=1}^{M_1}\frac{1}{|S_{-l}|}\sum_{i\in S_l} L\big(Y_i, \hat{m}_{\lambda(j,l)}(S_{-l})(X_i)\big)$$

4. *Pick the method with the smallest risk (and possibly tune again for fitting)*

9.3 Linear Models

We may take the linear model as a case study of the methods introduced in the above chapter. As such we consider the squaed loss $L(y_1, y_2) = (y_1 - y_2)^2$ for which we already know the Bayes rule:

$$m^*(x) = \operatorname*{argmin}_m R(m) = \mathbb{E}[Y \mid X = x].$$

The linear model has the following assumptions. There exists paramaters $\beta_0^*, \beta_1^*, ..., \beta_p^*$, with

$$m^*(x) = \beta_0^* + \sum_{j=1}^{p} \beta_j^* x_j.$$

In other words, we assume that m^* is a linear function, i.e.,

$$m^* \in \mathcal{G} = \{f : \mathbb{R}^p \to \mathbb{R} \mid f(x) = \beta^\top x\}.$$

Given iid training data $(X_i, Y_i)_{i=1,...,n}$ we have an additive noise model

$$Y_i = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_{ij} + \varepsilon_i,$$

with $\varepsilon_i = Y_i - m^*(X_i)$ and hence iid with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$.

Notice, that since we are assuming $m^* \in \mathcal{G}$ we have by assumption no inductive bias and we therefore only consider estimation error i.e.

$$R(\hat{m}_n) - r(m^*).$$

Given the training data we may approximate the coefficients using the following.

Lemma. (Coefficients in the linear model) *Under the Linear model assumption we have for $j = 1, ..., p$*

$$\beta_j^* = \frac{\mathrm{Cov}\Big(X_{1j}, Y_1 - \sum_{k\in\{1,...,p\}\setminus\{j\}} \beta_k^* X_{1k}\Big)}{\mathrm{Var}(X_{1j})}$$

In particular, if the components of X are uncorrelated, we have

$$\beta_j^* = \frac{\mathrm{Cov}(X_{1j}, Y_1)}{\mathrm{Var}(X_{1j})}.$$

Lemma. (Bayes risk in the linear model) *Under the Linear model assumption we have*

1. $r(m^*) = \mathrm{Var}(\varepsilon_i)$
2. For $m(x) = \beta_0 + \sum_{j=1}^p \beta_j x_j$, $\quad r(m) - r(m^*) = \|\Sigma^{1/2}(\beta - \beta^*)\|_2^2$.

with $\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$.

Proof.

(1). We have by assumptions that $m^*(X) = \mathbb{E}[Y \mid X] = X^\top\beta^*$ is the Bayes estimator. Using that the noise is additive we have $Y = m^*(X) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$.

$$\begin{aligned} r(m^*) &= \mathbb{E}[(m^*(X) - Y)^2] = \mathbb{E}[(X^\top\beta^* - Y)^2] \\ &= \mathbb{E}[\varepsilon^2] = \mathrm{Var}(\varepsilon) - \mathbb{E}[\varepsilon]^2 \\ &= \mathrm{Var}(\varepsilon). \end{aligned}$$

(2). Take any linear estimor $m(X) = X^\top\beta$, then we have

$$\begin{aligned} r(m) &= \mathbb{E}[(m(X) - Y)^2] \\ &= \mathbb{E}[(m(X) - m^*(X) + m^*(X) - Y)^2] \\ &= \mathbb{E}[(m(X) - m^*(X))^2] + \mathbb{E}[(m^*(X) - Y)^2] + 2\mathbb{E}[(m(X) - m^*(X))(m^*(X) - Y)] \\ &= \mathbb{E}[(m(X) - m^*(X))^2] + r(m^*) \end{aligned}$$

Using that $m^*(X) - Y$ is orthogonal to $m(X) - m^*(X)$. This gives us the following

$$\begin{aligned} r(m) - r(m^*) &= \mathbb{E}[(m(X) - m^*(X))^2] \\ &= \mathbb{E}[(X^\top\beta - X^\top\beta^*)^2] \\ &= \mathbb{E}[XX^\top(\beta - \beta^*)^2] \\ &= \Sigma(\beta - \beta^*)^2 = \|\Sigma^{1/2}(\beta - \beta^*)\|_2^2 \end{aligned}$$

as desired. ■

9.3.1 Least Squares Estimator

Moving forward we will assume $\beta_0^* = 0$ since we can always translate the data and make the centered around 0. We will furthermore use the notation:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

In this cases the empirical risk takes the form

$$\hat{R}_n(m) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2.$$

Lemma. (Least squares estimator)

- *It holds that* $(\mathbf{X}^\top \mathbf{X})\hat{\beta} = \mathbf{X}^\top \mathbf{Y}$,
- *If* $\mathbf{X}$ *has full rank, then* $\hat{\beta}_n^{LS} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$.

Theorem. (Excess risk Least squares estimator) *If* $\mathbf{X}^\top \mathbf{X}$ *is invertible, then*

$$\mathbb{E}[R(\hat{m}_n^{LS}) \mid \mathbf{X}] - r(m^*) = \frac{\sigma^2}{n} \cdot \operatorname{tr}\left(\Sigma \hat{\Sigma}^{-1}\right)$$

with $\Sigma = \mathbb{E}[\mathbf{X}^\top \mathbf{X}]$ *and* $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$.

Proof.

$$\hat{\beta}^{LS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta^* + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\varepsilon = \beta^* + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\varepsilon.$$

Thus

$$\begin{aligned} R\left(\hat{m}^{LS}\right) - r\left(m^*\right) &= \left\|\Sigma^{1/2}\left(\hat{\beta}^{LS} - \beta^*\right)\right\|_2^2 = \left\|\Sigma^{1/2}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\varepsilon\right\|_2^2 = n^{-1}\left\|n^{-1/2}\Sigma^{1/2}\hat{\Sigma}^{-1}\mathbf{X}^T\varepsilon\right\|_2^2 \\ &= n^{-1}\|A\varepsilon\|_2^2, \end{aligned}$$

where $\hat{\Sigma} = n^{-1}\mathbf{X}^\mathbf{T}\mathbf{X}$, $A := n^{-1/2}\Sigma^{1/2}\hat{\Sigma}^{-1}\mathbf{X}^T$. We calculate the excess risk conditional on $\mathbf{X}$. Note that $\operatorname{tr}(\cdot)$ is linear and invariant under cyclic permutations. We have

$$\begin{aligned} \mathbb{E}\left[R\left(\hat{m}^{LS}\right) \mid \mathbf{X}\right] - r\left(m^*\right) &= n^{-1}\mathbb{E}\left[\|A\varepsilon\|_2^2 \mid \mathbf{X}\right] \\ &= n^{-1}\mathbb{E}\left[\operatorname{tr}\left(A\varepsilon\varepsilon^T A^T\right) \mid \mathbf{X}\right] = n^{-1}\operatorname{tr}(A\underbrace{\mathbb{E}\left[\varepsilon\varepsilon^T\right]}_{=\sigma^2 I_{p\times p}}A^T) = \frac{\sigma^2}{n}\|A\|_F^2 \\ &= \frac{\sigma^2}{n} \cdot \operatorname{tr}\left(\Sigma^{1/2}\hat{\Sigma}^{-1}\hat{\Sigma}\hat{\Sigma}^{-1}\Sigma^{1/2}\right) \\ &= \frac{\sigma^2}{n} \cdot \operatorname{tr}\left(\Sigma\hat{\Sigma}^{-1}\right). \end{aligned}$$

as desired. ■

From the above it follos that if $\hat{\Sigma} \approx \Sigma$ then

$$\mathbb{E}[R(\hat{m}_n^{LS}) \mid \mathbf{X}] - r(m^*) \approx \frac{\sigma^2 p}{n}.$$

This approximation does not take into account the variation of X. Due to the inverse, it is not easily possible to derive an upper bound for the expectation of $\hat{\Sigma}^{-1}$. Therefore, we only obtain a result for the excess Bayes risk which holds with high probability and under additional assumptions (which could be relaxed but would lead to much more complicated proofs).

Theorem. (PAC Least squares estimator) *We do not assume a linear model. Let* $m(x) = E[Y \mid X = x]$ *and assume* $m^*(x) = x\Sigma^{-1}E[YX]$ *is the best linear approximation. Assume that* X *has bounded support and sub-Gaussian noise, i.e., there exists a* σ *such that for all* t*:*

$$E[e^{tx} \mid X = x] \leqslant e^{t^2\sigma^2/2},$$

then for n *big enough, and* $t > \max\{0, 2.6 - \log p\}$

$$P\left(r(\hat{m}^{LS}) - r(m^*) \geqslant \frac{2A}{n}(1 + \sqrt{8t})^2 + \frac{\sigma^2(p + 2\sqrt{pt} + 2t)}{n} + o(1/n)\right) \leqslant 3e^{-t}$$

where $A = \mathbb{E}\left[\|\Sigma^{1/2}X(m(X) - \beta^\top X)\|^2\right]$ *is an approximation error.*

Up until now we have assumed that $\mathbf{X}$ has full rank. This is not very realistic for very large p ($p >> n$). Even if $\mathbf{X}^\top\mathbf{X}$ is invertible, the variance of the estimation error might be too large. This leads us to penalized models that reduce variance by adding some bias. Other alternatives (not discussed here) are dimension reduction, say via PCA, or feature selection, say forward stepwise regression.

9.3.2 Ridge Regression

Definition. (Ridge regression) *Let* $\lambda \geqslant 0$ *and*

$$J_\lambda(\beta) = \lambda\|\beta\|_2^2 = \lambda\sum_{j=1}^{p}\beta_j^2.$$

The Ridge estimator is defined as

$$\begin{aligned}\hat{\beta}_\lambda^{\text{ridge}} &= \arg\min_{\beta\in\mathbb{R}^p}\left\{\hat{R}_n(\beta) + J_\lambda(\beta)\right\} \\ &= \arg\min_{\beta\in\mathbb{R}^p}\left\{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}.\end{aligned}$$

The corresponding algorithm is

$$\hat{m}_{n,\lambda}^{\text{ridge}}(x) = \sum_{j=1}^{p}\hat{\beta}_{\lambda,j}^{\text{ridge}}x_j.$$

Lemma. (Ridge regression solution) *Let* $\lambda > 0$*. Then*

$$\hat{\beta}_\lambda^{\text{ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda n I_{p\times p}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

In ridge regression, the matrix $\mathbf{X}^T\mathbf{X}$ is 'made invertible' by adding a positive multiple of the identity matrix. Therefore, the ridge estimator also can be used in the case $p > n$. The name 'ridge' stems from the fact that the optimization problem is equivalent to

$$\min_{\beta\in\mathbb{R}^p}\hat{R}_n(X\beta) \quad \text{s.t.} \quad \|\beta\|_2 \leqslant t$$

for some suitable $t > 0$.

Theorem. (Excess risk for ridge regression estimator) *Under the linear model,*

$$\mathbb{E}[R\left(\hat{m}_\lambda^{n,\text{ridge}}\right)|X] - r\left(m^*\right) = \frac{\sigma^2}{n}\cdot\operatorname{tr}\left(\Sigma\left(\hat{\Sigma} + \lambda I_{p\times p}\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma} + \lambda I_{p\times p}\right)^{-1}\right) + \lambda^2\left\|\Sigma^{1/2}\left(\hat{\Sigma} + \lambda I_{p\times p}\right)^{-1}\beta^*\right\|_2^2.$$

Let $\Sigma = UDU^T$ *be the spectral decomposition of* Σ *with orthogonal matrix* U *and diagonal matrix* $D = \operatorname{diag}(s_1, \ldots, s_p)$ *(entries are the eigenvalues of* Σ*). By assuming* $\hat{\Sigma} = \Sigma$*, the excess risk simplifies to*

$$\frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{s_j^2}{(s_j+\lambda)^2}+\lambda^2\cdot\sum_{j=1}^{p}\frac{s_j\left(U^T\beta^*\right)_j^2}{(s_j+\lambda)^2}.$$

Proof.

$$\begin{aligned}\hat{\beta}_\lambda-\beta^* &= -\lambda n\left(\mathbf{X}^T\mathbf{X}+\lambda nI_{p\times p}\right)^{-1}\beta^*+\left(\mathbf{X}^T\mathbf{X}+\lambda nI_{p\times p}\right)^{-1}\mathbf{X}^T\varepsilon\\ &= -\lambda\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\beta^*+\frac{1}{n}\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\mathbf{X}^T\varepsilon\end{aligned}$$

. thus

$$R\left(\hat{\beta}_\lambda\right)-R\left(\beta^*\right)=\left\|B-\frac{1}{\sqrt{n}}A\varepsilon\right\|_2^2=\|B\|_2^2-\frac{2}{\sqrt{n}}\langle B,A\varepsilon\rangle+\frac{1}{n}\|A\varepsilon\|_2^2$$

where $A=\Sigma^{1/2}\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\frac{\mathbf{X}^T}{\sqrt{n}}$ and $B:=\lambda\Sigma^{1/2}\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\beta^*$. Since $\mathbb{E}\varepsilon=0$, we have

$$\begin{aligned}\mathbb{E}\left[R\left(\hat{\beta}_\lambda\right)\mid\mathbf{X}\right]-R\left(\beta^*\right) &= \frac{\sigma^2}{n}\|A\|_F^2+\|B\|_2^2\\ &= \frac{\sigma^2}{n}\cdot\operatorname{tr}\left(\Sigma\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\right)\\ &\quad+\lambda^2\left\|\Sigma^{1/2}\left(\hat{\Sigma}+\lambda I_{p\times p}\right)^{-1}\beta^*\right\|_2^2\end{aligned}$$

Furthermore, assuming $\hat{\Sigma}=\Sigma$ the above expression simplifies to

$$\begin{aligned}&\frac{\sigma^2}{n}\cdot\operatorname{tr}\left(\Sigma\left(\Sigma+\lambda I_{p\times p}\right)^{-1}\Sigma\left(\Sigma+\lambda I_{p\times p}\right)^{-1}\right)+\lambda^2\left\|\Sigma^{1/2}\left(\Sigma+\lambda I_{p\times p}\right)^{-1}\beta^*\right\|_2^2\\ =&\frac{\sigma^2}{n}\cdot\operatorname{tr}\left(D\left(D+\lambda I_{p\times p}\right)^{-1}D\left(D+\lambda I_{p\times p}\right)^{-1}\right)+\lambda^2\left\|D^{1/2}\left(D+\lambda I_{p\times p}\right)^{-1}U^T\beta^*\right\|_2^2\\ =&\frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{s_j^2}{(s_j+\lambda)^2}+\lambda^2\cdot\sum_{j=1}^{p}\frac{s_j\left(U^T\beta^*\right)_j^2}{(s_j+\lambda)^2}\end{aligned}$$

and the result follows. ■

If all eigenvalues are equal, that is, $s_j=s$ and if additionally $\left(U^T\beta^*\right)_j=b(j=1,\ldots,p)$, then the expression of the theorem simplifies to

$$\frac{\sigma^2 p}{n}\cdot\frac{s^2}{(s+\lambda)^2}+\lambda^2\frac{sb^2p}{(s+\lambda)^2}\underset{\lambda=\frac{\sigma^2/n}{b^2}}{\overset{\min}{\rightarrow}}\frac{\sigma^2p}{n}\cdot\frac{b^2s}{\frac{\sigma^2}{n}+b^2s}\leqslant\frac{\sigma^2p}{n}.$$

We see that for a suitable choice of the penalization parameter λ, the excess Bayes risk of the ridge estimator can be smaller than the corresponding upper bound of the LS estimator.

9.3.3 Lasso Regression

If we believe that some covariates are pure noise, i.e., unrelated to Y, the most obvious choice to penalize β would be of the form $\|\beta\|_0=\#\{j=1,\ldots,p:\beta_j\neq 0\}$. Then, one would simply penalize the number of non-zero entries of β. However, this leads to an NP-hard optimization problems whose solutions are not accessible in practice. One therefore uses a different norm which has similar properties but leads to a convex optimization problem.

Definition. (Lasso - Least absolute shrinkage and selection operator regression) *Let $\lambda\geqslant 0$ and*

$$J_\lambda(\beta) = \lambda \cdot \|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j| .$$

The LASSO estimator is given by

$$\begin{aligned}\hat{\beta}_\lambda^{\text{lasso}} &\in \arg\min_{\beta\in\mathbb{R}^p} \left\{ \hat{R}_n(X\beta) + J_\lambda(X\beta) \right\} \\ &= \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \cdot \|\beta\|_1 \right\}\end{aligned}$$

The corresponding algorithm reads

$$\hat{m}_{n,\lambda}^{lasso}(x) = \sum_{j=1}^{p} \hat{\beta}_{\lambda,j}^{\text{lasso}} x_j.$$

There exists no easy closed-form solution for $\hat{\beta}_\lambda^{\text{lasso}}$ besides some special cases.

For $\beta \in \mathbb{R}^p$, define

$$S(\beta) := \{j \in \{1,\dots,p\} : \beta_j \neq 0\} .$$

For $S \subset \{1,\dots,p\}$ and $v \in \mathbb{R}^p$, put $v_S := \left(v_j \mathbb{1}_{\{j\in S\}}\right)_{j=1,\dots,p}$

If $p \ll n$, then $\hat{\Sigma}$ would usually be invertible and the smallest eigenvalue (Rayleigh quotient) would satisfy

$$\lambda_{\min}(\hat{\Sigma}) := \inf_{v\in\mathbb{R}^p} \frac{v^T\hat{\Sigma}v}{\|v\|_2^2} > 0.$$

Then $\hat{\Sigma}$ would be one-to-one (injective) and the linear equation system $\hat{\Sigma}\beta = \frac{1}{n}\mathbf{X}^T\mathbf{Y}$ would lead to the (unique) least squares estimator.

For $p \gg n$, one has $\lambda_{\min}(\hat{\Sigma}) = 0$.

When employing LASSO, we are usually only interested in estimators $\hat{\beta}$ whith non-zero entries at the components $S(\beta^*)$. This means that in principle we only need injectivity of $\hat{\Sigma}$ on the set

$$\tilde{C} = \{\beta \in \mathbb{R}^p : S(\beta) = S(\beta^*)\} = \left\{\beta \in \mathbb{R}^p : \left\|\beta_{S(\beta^*)^c}\right\|_1 = 0\right\},$$

or equivalently, $\inf_{v\in\tilde{C}} \frac{v^T\hat{\Sigma}v}{\|v\|_2^2} = \inf_{v\in\tilde{C}} \frac{v^T\hat{\Sigma}v}{\left\|v_{S(\beta^*)}\right\|_2^2} > 0$.

Definition. (Restricted eigenvalue property (REP)) *We say that the restricted eigenvalue property (REP) is satisfied with $\alpha > 0$ if for*

$$C := \left\{\beta \in \mathbb{R}^p : \left\|\beta_{S(\beta^*)^c}\right\|_1 \leqslant \alpha \left\|\beta_{S(\beta^*)}\right\|_1\right\}$$

it holds that

$$\Lambda_{\min}(\Sigma) := \inf_{v\in C} \frac{v^T\Sigma v}{\left\|v_{S(\beta^*)}\right\|_2^2} > 0,$$

Theorem. *Let $\varepsilon \sim N\left(0,\sigma^2\right), X \sim N(0,\Sigma)$ and $\Sigma_{jj} = 1 (j = 1,\dots,p)$. Define $s := \#S(\beta^*)$. Then there exist universal constants $c_1, c_2 > 0$ such that the condition*

$$n \geqslant c_1 \frac{\|\Sigma\|_2}{\Lambda_{\min}(\Sigma)^2} s \log(ep/s)$$

implies: For each $t \geqslant 0$ and

$$\lambda \geqslant \frac{6\sqrt{2}\sigma}{\sqrt{n}} \sqrt{\log(p) + t},$$

it holds that

$$\mathbb{P}\left(R\left(\hat{m}_{n,\lambda}\right) - r\left(\beta^*\right) > 16\lambda^2 \frac{s}{\Lambda_{\min}(\Sigma)}\right) \leqslant e^{-t} + 2pe^{-c_2 n}.$$

The upper bound for the convergence rate of the excess risk of the LASSO estimator is minimized for $\lambda = 6\sqrt{2} \cdot \frac{\sigma}{\sqrt{n}}\sqrt{\log(p)}$. With that choice,

$$16\lambda^2 \frac{s}{\Lambda_{\min}(\Sigma)} = \frac{\mathrm{c}}{\Lambda_{\min}(\Sigma)} \cdot \frac{\sigma^2 s}{n} \cdot \log(p)$$

Interpretation: $\hat{\beta}_\lambda$ behaves like the LS estimator in a model with s instead of p dimensions.

The LASSO estimator $\hat{\beta}_\lambda$ has to 'pay' with a factor $\log(p)$ for the missing insight which components are non-zero. This is a rather small price to pay even if p is large.

One can prove similar theoretical statements without the conditions $\varepsilon \sim N\left(0, \sigma^2\right)$ and $X \sim N(0, \Sigma)$ and still can preserve the small $\log(p)$ term.

Regarding the REP: The smallest eigenvalue $\lambda_{\min}(\Sigma)$ measures how strongly the components of X are correlated. Note that a strong correlation of X is a problem for estimation of β^*, but not for the excess risk itself: In the extreme case $X_1 = X_2$, it is clear that $\hat{\beta}$ cannot distinguish the values of β_1^* and β_2^*, but it can still provide good predictions through $X\hat{\beta}$. Unfortunately, the proof technique underlying the theorem transfers the estimation quality of β^* to an upper bound of the excess risk, therefore this fact is not adequately represented in the result.

The assumption $\Sigma_{jj} = 1$ is only to provide an easier result. In practice, this normalization can be obtained by standardizing $X_1, \ldots, X_n$ before computing the LASSO estimator (that is, center X_i and divide by the empirical standard deviation).

Theorem. *We do not assume that the linear model holds. Assume that X and Y have bounded support (bounded by $B > 0$). Let*

$$\beta_* = \underset{\|\beta\|_1 \leqslant \eta}{\operatorname{argmin}}\, r(X\beta)$$

Then for any $\xi > 0$,

$$\mathbb{P}\left(r(\widehat{\beta}) - r\left(\beta_*\right) \geqslant \sqrt{\frac{2(\eta+1)^4 B^2}{n} \log\left(\frac{2p^2}{\xi}\right)}\right) \leqslant \xi.$$

Proof.

Set $Z = (Y, X)$ and $Z_i = (Y_i, X_i)$, $\gamma = \gamma(\beta) = (-1, \beta)$. Then

$$r(X\beta) = \mathbb{E}\left(Y - \beta^T X\right)^2 = \gamma^T \Lambda \gamma$$

where $\Lambda = \mathbb{E}\left[ZZ^T\right]$. Note that $\|\gamma\|_1 = \|\beta\|_1 + 1$. Let $\mathcal{B} = \{\beta : \|\beta\|_1 \leqslant \gamma\}$.

$$\hat{R}_n(X\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2 = \gamma^T \widehat{\Lambda}\gamma$$

where $\widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$. For any $\beta \in \mathcal{B}$

$$\begin{aligned}
|\hat{R}_n(\mathbf{X}\beta) - r(X\beta)| &= \left|\gamma^T(\widehat{\Lambda} - \Lambda)\gamma\right| \\
&\leqslant \sum_{j,k} |\gamma_j||\gamma_k||\widehat{\Lambda}_{jk} - \Lambda_{jk}| \\
&\leqslant (\eta+1)^2 \max_{jk} |\widehat{\Lambda}_{jk} - \Lambda_{jk}|
\end{aligned}$$

Note that $|\Lambda_{jk}| \leqslant B^2$. By Hoeffding's inequality,

$$\mathbb{P}\left(|\widehat{\Lambda}_{jk} - \Lambda_{jk}| \geqslant \epsilon\right) \leqslant 2e^{-2n\epsilon^2/B^2}$$

and so, by the union bound,

$$\mathbb{P}\left(\max_{j,k} |\widehat{\Lambda}_{jk} - \Lambda_{jk}| \geqslant \epsilon\right) \leqslant 2p^2 e^{-2n\epsilon^2/B^2} = \xi,$$

if we choose $\epsilon = \sqrt{\frac{B^2}{2n}\log\left(\frac{2p^2}{\xi}\right)}$. Hence, with probability $1-\xi$ (see slide 14, lecture 1),

$$r(X\hat{\beta}) - r\left(X\beta^*\right) \leqslant 2(\eta+1)^2\varepsilon.$$

as desired. ■

Definition. *If $P(S(\widehat{\beta}) = S(\beta)) \to 1$ we call $\hat{\beta}$ **sparsistent**.*

*We call $\hat{\beta}$ weakly **sparsistent** if, for every β as $n \to \infty$*

$$P_\beta\left(I\left(\widehat{\beta}_j = 1\right) \leqslant I\left(\beta_j = 1\right) \text{ for all } j\right) \to 1$$

In the above $S(\cdot)$ represent the covariates with non-zero covariates. Therefore we can interpret a sparsistent estimator as an estimator which for increasing information converges to choosing the correct explanatory variables, where the weak condition only ensure that we may choose the right covariates but at least not the wrong. Suppose that p is fixed. Then the least squares estimator $\widehat{\beta}_n$ is minimax and satisfies

$$\sup_\beta E_\beta\left(n\left\|\widehat{\beta}_n - \beta\right\|^2\right) = O(1).$$

But sparsistent estimators have larger risk:

Theorem. *We don't assume the linear model. Suppose that the following condiitons hold:*

- *p is fixed.*
- *The covariariates are nonstochastic and $n^{-1}\mathbf{X}^T\mathbf{X} \to \Sigma$ for some positive definite matrix Σ.*
- *The errors ϵ_i are independent with mean 0, finite variance σ^2 and have a density f satisfying*

$$0 < \int\left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty$$

If $\widehat{\beta}$ is weakly sparsistent, then

$$\sup_\beta \mathbb{E}_\beta\left(n\left\|\widehat{\beta}_n - \beta\right\|^2\right) \to \infty.$$

More generally, if ℓ is any loss function $\ell : \mathbb{R} \mapsto \mathbb{R}_{\geqslant 0}$.then

$$\sup_\beta \mathbb{E}_\beta\left(\ell\left(n^{1/2}\left(\widehat{\beta}_n - \beta\right)\right)\right) \to \sup_s \ell(s).$$

Proof.

Choose any $s \in \mathbb{R}^d$ and let $\beta_n = -s/\sqrt{n}$. Then,

$$\begin{aligned}\sup_\beta \mathbb{E}_\beta\left(\ell\left(n^{1/2}(\widehat{\beta} - \beta)\right)\right. &\geqslant \mathbb{E}_{\beta_n}\left(\ell\left(n^{1/2}(\widehat{\beta} - \beta)\right)\right. \geqslant \mathbb{E}_{\beta_n}\left(\ell\left(n^{1/2}(\widehat{\beta} - \beta)\right)I(\widehat{\beta} = 0)\right) \\ &= \ell\left(-\sqrt{n}\beta_n\right)\mathbb{P}_{\beta_n}(\widehat{\beta} = 0) = \ell(s)P_{\beta_n}(\widehat{\beta} = 0).\end{aligned}$$

Now, $\mathbb{P}_0(\widehat{\beta} = 0) \to 1$ by assumption. It can be shown (via contiguity) that we also have $\mathbb{P}_{\beta_n}(\widehat{\beta} = 0) \to 1$. Hence, with probability tending to 1,

$$\sup_\beta E_\beta\left(\ell\left(n^{1/2}(\widehat{\beta} - \beta)\right) \geqslant \ell(s).\right.$$

Since s was arbitrary the result follows. ■

9.3.4 Conclusion

Lasso and Ridge regression aim for different things. Ridge regression reduces variance of the estimator by restricting the function space. Heuristically, the smaller $||\beta^*||_2$ the more helpful is ridge regression. Lasso is useful in sparse setting, i.e, if one wishes to eliminate components/features.

Example. Assume that $\text{corr}(\text{X}_1, \text{X}_2) = 1$ and $Y = 0.5X_1 + 0.5X_2$. While Lasso will probably estimate $\beta_1 = 1, \beta_2 = 0$, Ridge will probably estimate correctly $\beta_1 = 0.5, \beta_2 = 0.5$.

Definition. (Elastic net) *For $\lambda > 0, \alpha \in (0,1)$, and $J^{elastic.net}(\beta) = \sum_{j=1}^{p} \alpha|\beta_j| + (1-\alpha)\beta_j^2$, we call*

$$\hat{\beta}_\lambda^{\text{elastic.net}} \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \hat{R}_n(\beta) + J_\lambda^{elastic.net}(\beta) \right\}$$

elastic net *estimator.*

9.4 Nonparametric Regression

Linear models are quite restrictive and one may ask how one can achieve more flexibility. In this chapter we will look at the nonparametric regression problem with squared loss $L(y_1, y_2) = (y_1 - y_2)^2$. We already know that the Bayes-rule is $m^*(x) = \mathbb{E}[Y|X = x]$.

9.4.1 Linear Smoothers

Definition. (k-nearest-neighbor) *The k-nearest-neighbor estimator is*

$$\hat{m}^{knn}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} Y_i,$$

where $\mathcal{N}_k(x)$ contains the indices of the k closest points of $\{X_1, \dots X_n\}$ to x.

Definition. (Linear Smoother) *An estimator is called linear smoother if it can be written as*

$$\hat{m}(x) = \sum_i w_i(x) Y_i,$$

where the weight function w_i can depend on $\{X_1, \dots, X_n\}$.

Example. The k-nearest-neighbor estimator is a linear smoother:

$$\hat{m}^{knn}(x) = \sum_i^n w_i(x) Y_i,$$

with

$$w_i(x) = \begin{cases} \frac{1}{k} & X_i \text{ belongs to the } k \text{ closest points to } x \\ 0 & else \end{cases}$$

In the below proposition the definition of Lipschitz continuity is used. We recall that a real-valued function $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz continuous if and only if

$$f(x_1) - f(x_2) \leqslant L\|x_1 - x_2\|_2$$

for all $(x_1, x_2) \in \mathbb{R}^2$. This in particulat means that $f(x) \in [f(x_1) - L|x_1 - x|, f(x_1) + L|x_1 - x|]$ i.e. does not on any interval grow faster that a linear function with slope L.

Proposition. (MSE k-nearest-neighbor) *Assume that*

$$E[Y|X = x] = m^*(x) \in \mathcal{G}_L = \{m : \mathbb{R}^p \mapsto \mathbb{R} \mid m \text{ is L-Lipschitz continuous}\},$$

and $Var(Y|X = x) = \sigma^2(x) \leqslant \sigma^2$. *Then*

$$\mathbb{E}[(\hat{m}^{knn}(x) - m^*(x))^2] \leqslant (cL)^2 \left(\frac{k}{n}\right)^{2/p} + \frac{\sigma^2}{k}.$$

In particular, for $k_n = O_p(n^{2/(2+p)})$, *we get*

$$\mathbb{E}[(\hat{m}^{knn}(x) - m^*(x))^2] = O_p(n^{-2/(2+p)}).$$

Proof.

Write $\mathbf{X} = (X_1, \ldots, X_n)$ and denote by $Y_i^{(x)}$ the ith closest Y to x among $Y_1, \ldots Y_n$.

$$\begin{aligned}
\mathbb{E}[(\hat{m}^{knn}(x) - m^*(x))^2] &\overset{(\dagger_1)}{=} \mathbb{E}\Big[\big(\mathbb{E}[\hat{m}^{knn}(x)|\mathbf{X}] - m^*(x)\big)^2\Big] + \mathbb{E}\Big[(\hat{m}^{knn}(x) - \mathbb{E}[\hat{m}^{knn}(x)|\mathbf{X}])^2\Big] \\
&\overset{(\dagger_2)}{=} \mathbb{E}\left[\left\{\frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} \left(m^*\left(X_i\right) - m^*(x)\right)\right\}^2\right] \\
&\quad + \frac{1}{k^2}\mathbb{E}\left[\sum_i^k \sum_j^k \{Y_i^{(x)} - \mathbb{E}[Y_i^{(x)}|\mathbf{X}]\}\{Y_j^{(x)} - \mathbb{E}[Y_j^{(x)}|\mathbf{X}]\}\right] \\
&\overset{(\dagger_3)}{\leqslant} \mathbb{E}\left[\left(\frac{L}{k}\sum_{i\in\mathcal{N}_k(x)} \|X_i - x\|_2\right)^2\right] + \frac{\sigma^2}{k} \\
&\overset{(\dagger_4)}{\leqslant} L^2 c\left(\frac{k}{n}\right)^{2/p} + \frac{\sigma^2}{k}
\end{aligned}$$

where we as usual in $(\dagger_1)$ use that $m^*(x) - Y$ is orthogonal to any element $m(x) - m^*(x)$. In the $(\dagger_2)$ we use the assumption that $m^*(x) = \mathbb{E}[Y|X = x]$ and hence

$$\mathbb{E}\left[\hat{m}^{knn}(x)\middle|X\right] = \frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} \mathbb{E}\left[Y_i\middle|X\right] = \frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} m^*(X_i).$$

Furthermore, the second term is derived from the below

$$\begin{aligned}
\mathbb{E}\left[(\hat{m}^{knn}(x) - \mathbb{E}[\hat{m}^{knn}(x)|\mathbf{X}])^2\right] &= \mathbb{E}\left[\left(\frac{1}{k}\sum_{l\in\mathcal{N}_k(x)} Y_l - \frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} \mathbb{E}\left[Y_i\middle|X\right]\right)^2\right] \\
&= \frac{1}{k^2}\mathbb{E}\left[\sum_i^k \sum_j^k \{Y_i^{(x)} - \mathbb{E}[Y_i^{(x)}|\mathbf{X}]\}\{Y_j^{(x)} - \mathbb{E}[Y_j^{(x)}|\mathbf{X}]\}\right]
\end{aligned}$$

The $(\dagger_3)$ is derived by the definition of a L-Lipschitz function and the assumption regarding the variance of the conditional variable $Y \mid X = x$. $(\dagger_4)$ is a result taken from Gyorfi et al. 2002, vol. 1, chap. 6.3. ■

Note that this is slower than the "parametric" MSE of n^{-1}. In particular the rate depends on p. Even worse: It grows exponentially in p.

We have learned that the knn estimator can be written as

$$\hat{m}^{knn}(x) = \frac{1}{k}w_i(x_i)Y_i,$$

Note that $w_i(x)$ is not smooth as a function of x. This also makes the estimator not smooth. Given that we assume that m^* is smooth, this may not be desirable. An alternative are kernel smoothers.

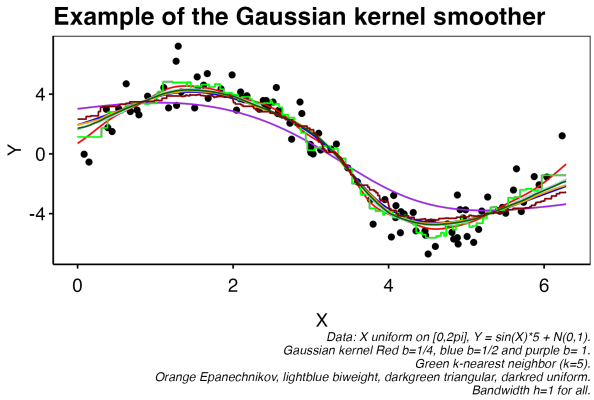
Definition. (Kernel Smoother) *The kernel smoother is a linear smoother with*

$$\hat{m}^{ks}(x) = \sum_i w_i(x_i) Y_i,$$

where

$$w_i(x) = \frac{K\left(\frac{||x-X_i||}{h}\right)}{\sum_j K\left(\frac{||x-X_j||}{h}\right)}$$

Often $K = \prod_j k_j$, *such that* $K\left(\frac{||x-X_i||}{h}\right) = \prod_j k(x - X_{ij})$. *The function* $k : \mathbb{R} \mapsto \mathbb{R}$ *is usually a symmetric density function.*



The general idea is to fit a smooth function to the data points (X, Y) such that $m(X)$ is somewhat centered between the data points. The choice of a symmetric function k ensures that the estimator weighs the datapoints closer to x higher than further away data point. In the case with $X, Y \in \mathbb{R}$ we can choose for instance

1. Gaussian kernel: $k_j(z) = \exp\left(\frac{z}{2b^2}\right)$ with $b > 0$,
2. Uniform kernel: $k_j(z) = 1_{[0,1]}(z)/2 \in \{0, 1\}$,
3. Triangular kernel: $k_j(z) = \left(1 - z\right)^+$,
4. Epanechnikov kernel: $k_j(z) = \frac{(1-z^2)^a}{2^{2a+1}\Gamma(a+1)^2\Gamma(2a+2)^{-1}} 1_{[0,1]}(z)$ with $a = 1$.
5. Biweight kernel: Above with $a = 2$.

where of course the argument is $z = |x - X_{ij}|/h$ with $h > 0$ being the bandwidth. Notice that the uniform kernel in fact is the Epanechnikov kernel with $a = 0$.

Proposition. (MSE Kernel Smoother) *Assume that* $E[Y|X = x] = m^*(x) \in \mathcal{G}_L$ *with*

$$\mathcal{G}_L = \{m : \mathbb{R}^p \mapsto \mathbb{R} \mid m \text{ is L-Lipschitz continuous}\},$$

and $Var(Y|X = x) = \sigma^2(x) \leqslant \sigma^2$. *Then*

$$\mathbb{E}[(\hat{m}^{ks}(x) - m^*(x))^2] = O_p\left(\frac{1}{nh^p} + h^2\right)$$

In particular, for $h_n = O_p(n^{-1/(2+p)})$, *we get*

$$\mathbb{E}[(\hat{m}^{ks}(x) - m^*(x))^2] = O_p(n^{-2/(2+p)}).$$

9.4.2 Curse of dimensionality

We have seen that under a Lipschitz condition, both kernel smoother and knn have an asymptotic mean squared error of order $n^{-2/(2+p)}$. One can show that under the assumption that m^* is twice continuously differentiable, the rate for both methods can be improved to

$$n^{-4/(4+p)}.$$

But this rate is still exponentially decreasing in p. Furthermore, it has been shown that no method can do better under the given assumptions.

A new observation x_0 will have very few or no observations in its neighborhood. This leads to high variance and high bias when increasing the size of the neighborhood.

Under the model in the previous section, the setting $n = 50, p = 1$ has the same expected amount of observations in a neighborhood as the setting $n = 7.5 \times 10^{110}, p = 100$.

There are two ways to tackle the curse of dimensionality.

Sparsity: Assume that the intrinsic dimension is lower. E.g. Not all variables are relevant. Or feature engineer a few highly predictive variables.

Structure: Interactions are limited and structure can be exploited e.g. an additive structure $m(x) = m_1(x_1) + m_2(x_2)$. Remember that structure is essential for interpretability.

9.4.3 Splines

We want to establish a framework to estimate additive regression functions. To this end, assume $p = 1$ until further notice.

Definition. (Splines) *A kth-order spline with l knotpoints $x_1 < \cdots < x_l$*

- *is a polynomial of degree k on each interval $(-\infty, x_1], [x_1, x_2] \ldots, [x_l, \infty)$*
- *has continuous derivatives of orders $0, ..., k-1$ on the knotpoints $x_1 < \cdots < x_l$*

Example. A kth-order spline m with l knotpoints can an be uniquely written as

$$m(x) = \sum_{j=1}^{k+1+l} \theta_j g_j(x)$$

- truncated power basis
  - For $j = 1, \ldots, k+1$: $g_j = x^{j-1}$
  - For $j = 1, \ldots, l$: $g_{k+1+j} = (x - x_j)_+^k$
    - $(x)_+ := \max(x, 0)$
- B-splines
  - more complicated, but computationally more robust and faster to compute.

Splines have high variance at the boundaries. Solution: Let the piecewise polynomial function have a lower degree at $(-\infty, x_1], [x_l, \infty)$.

Definition. (Natural Splines) *A kth-order natural spline (k = odd number) with knotpoints $x_1 < \cdots < x_l$*

- *is a polynomial of degree k on the intervals $[x_1, x_2] \ldots, [x_{l-1}, x_l)$*
- *is a polynomial of degree $(k-1)/2$ on $(-\infty, x_1], [x_l, \infty)$*
- *has continuous derivatives of orders $0, ..., k-1$ on the knotpoints $x_1 < \cdots < x_l$*

Note that natural splines have dimension l which is in particular independent of the order k (compare to dimension $k + l$ for splines). There is also a truncated power basis and a B-splines basis for natural splines.

9.4.4 Linear regression with splines.

We still assume the one-dimensional case, $p = 1$. Instead of looking at observations $(X_i, Y_i)_{i=1,\dots,n}$ we can consider a natural splines basis and look at observations $(g_1(X_i), \dots, g_l(X_i), Y_i)_{i=1,\dots,n}$. By doing so we are able to approximate any natural spline in x instead of just linear functions in x, while still being in a linear regression framework, i.e.,

$$\hat{\beta} = \arg\min_\beta \sum_i (Y_i - G_i^T \beta)^2 = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{Y}$$

where $G_i^T = (g_1(X_i), \dots, g_l(X_i))$, the rows of $\mathbf{G}$. Problem: How do we choose the number of knotpoints l and their position? First thought: Cross validation. But that would be quite expensive to run.

Let's look at the following minimization problem:

$$\hat{m} = \arg\min_m \sum_i (Y_i - m(X_i))^2 + \lambda \int_a^b m''(x)^2 \mathrm{d}x,$$

where minimization runs over all twice times differentiable functions m and observations X_i are in $[a, b]$ for all i.

Theorem (Smoothing splines) *If m is twice differentiable and the solution to*

$$\arg\min_m \sum_i (Y_i - m(X_i))^2 + \lambda \int_a^b m''(x)^2 \mathrm{d}x,$$

then m is a natural spline of order 3 (natural cubic spline).

Proof.

Given a minimizer $\tilde{m}$, we construct the unique natural cubic spline m on $[a, b]$ with knotpoints $X_1, \dots, X_n$ and $m(X_i) = \tilde{m}(X_i)$. We will show that $\tilde{m} = m$. Define $h = \tilde{m} - m$. Note that $h(X_j) = 0 (j = 1, \dots, n)$, $m'''(x) = 0$ for $x < X_1$ and $x > X_n$ as well as $m^{(4)} = 0$. Hence by applying integration by parts twice we get

$$\begin{aligned}
\int_a^b m''(x)h''(x)\mathrm{d}x &= -\int_{X_1}^{X_n} m'''(x)h'(x)\mathrm{d}x \\
&= -\sum_{j=1}^{n-1} m'''(x)h(x)\Big|_{X_j}^{X_{j+1}} \\
&= 0
\end{aligned}$$

This implies

$$\int_a^b \tilde{m}''(x)^2\mathrm{d}x = \int_a^b \{m''(x) + h(x)\}^2\mathrm{d}x = \int_a^b m''(x)^2 + h(x)^2\mathrm{d}x.$$

meaning

$$\int_a^b m''(x)^2\mathrm{d}x \leqslant \int_a^b \tilde{m}''(x)^2\mathrm{d}x.$$

with equality only for $h'' = 0$, i.e, h linear. Since $h(X_i) = 0 (i = 1, \dots n)$, we have $h = 0$ and so $\tilde{m} = m$. ■

We can conclude the following:

- Let $\mathbf{G}$ be the matrix with with rows $G_i = (g_1(X_i), \dots, g_l(X_i))$.
  - $\{g_j\}_{j=1,\dots,l}$ is a basis for natural cubic splines.
- Define
$$\hat{\beta} = \arg\min_\beta \sum_i (Y_i - G_i^T\beta)^2 + \lambda \int_0^1 \left\{ \sum_j \beta_j g_j''(x) \right\}^2 \mathrm{d}x.$$
Then,
$$\sum_j \hat{\beta}_j g_j = \arg\min_m \sum_i (Y_i - m(X_i))^2 + \lambda \int_0^1 m''(x)^2\mathrm{d}x.$$

Smoothing splines can be seen as a special case of generalized ridge regression:

- Write $\mathbf{W}_{ij} = \int_0^1 g_i''(x)g_j''(x)\mathrm{d}x$, then

$$\begin{aligned}&\arg\min_\beta \sum_i (Y_i - G_i^T\beta)^2 + \lambda \int_0^1 \left\{\sum_j \beta_j g_j''(x)\right\}^2 \mathrm{d}x \\ =&\arg\min_\beta \sum_i (Y_i - G_i^T\beta)^2 + \lambda \beta^T \mathbf{W}\beta.\end{aligned}$$

  Hence, $$\hat{\beta} = (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{W})^{-1}\mathbf{G}^T\mathbf{Y}.$$

It can be shown the smoothing splines are asymptotically equivalent to kernel smoothers with varying bandwidth and a specific choice of kernel, see Silverman (1984) or Wang, Du, and Shen (2013) for a more recent contribution. In general, smoothing splines are more practical since they can be efficiently calculated while kernel smoothers are much easier to analyze theoretically.

We have seen that fully nonparametric methods suffer from the curse of dimensionality: the optimal rate of convergence for twice continuously differentiable functions is $n^{-4/(4+p)}$. One solution is to restrict oneself to the class of additive functions

$$\mathcal{G} = \{m \mid m(x) = m_1(x_1) + \cdots + m_p(x_p)\}$$

Stone (1985) showed that the components m_k, if twice continuously differentiable, can be estimated with one-dimensional rate of $n^{-4/(4+p)}$. The components m_j are usually estimated via the so called backfitting algorithm (Hastie and Tibshirani 1990).

Backffitting comprises the following two steps:

Definition. (Backfitting Algorithm)

- *Intialize:* $\hat{m}_j^{[0]} = 0, j = 1, \ldots, p$
- *Iterate for* $r = 1, \ldots$
  1. *Residuals:* $r_{ij}^{[r]} = Y_i - \sum_{k<j} \hat{m}_k^{[r]}(x_{ik}) - \sum_{k>j} \hat{m}_k^{[r-1]}(x_{ik})$.
  2. *Smooth:* $\hat{m}_j^{[r]} = \text{Smooth}\left(\left\{X_{ij}, r_{ij}^{[r]}\right\}_{i=1,\ldots,n}\right)$.
  3. *Center:* $\hat{m}_j^{[r]} = \hat{m}_j^{[r]} - \frac{1}{n}\sum_{i=1}^n \hat{m}_j^{[r]}(X_{ij})$.

Note that Smooth is a one-dimensional regression problem. In practice Smooth is most often a smoothing spline. It has been shown backfitting via smoothing splines achieve optimal rate of $n^{-4/(4+p)}$ for each component and $pn^{-4/(4+p)}$ for the p-dimensional additive regression function.

Problem: Additive methods are still not optimal in the case of sparsity (i.e. some features being not relevant) and interactions between features.

9.5 Trees and forests

In this chapter we will look at the nonparametric regression problem with squared loss $L(y_1, y_2) = (y_1 - y_2)^2$ and also the classification problem with Binary loss function $L(y_1, y_2) = 1(y_1 \neq y_2)$ or squared loss.

We already know that the Bayes-rule is

$$m^*(x) = \mathbb{E}[Y|X = x]$$

for the squared loss and

$$m^*(x) = \underset{k=1,\ldots,K}{\text{argmax}}\ \mathbb{P}(Y = k|X = x)$$

for the binary loss.

Definition. (Decision trees) *A decision tree is an estimator, that partition the feature space $\mathcal{X}$ into sections $T = \{R_1, R_2, ..., R_k\}$ such that $R_i \cap R_j = \varnothing$ (pairwise disjoint) for all $i \neq j$ with $1 \leqslant i, j \leqslant k$ and*

$$\bigcup_{i=1}^{k} R_i = \mathcal{X}$$

The algorithm associated with the decision tree is then

$$m(T)(x) = \sum_{i=1}^{k} m_i 1_{R_i}(x).$$

From the above we notice that the tree estimate is a piecewise constant function. We call the constant areas of the tree estimate *leaves* or *terminal nodes*. The leaves partition the feature space $\mathcal{X}$.

Given the data $\mathcal{D}_n$ the value within a leave is given by averaging the responses (regression with L_2 loss) or majority vote (classification with binary loss). In particular we have

$$m_i = \left(\sum_{j=1}^{n} 1_{R_i}(X_j) \right)^{-1} \left(\sum_{j=1}^{n} 1_{R_i}(X_j) X_j \right), \qquad \text{(continuous case)}$$

i.e. the simple estimated conditional mean $\hat{\mathbb{E}}[Y \mid X \in R_i]$ and

$$m_i = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left\{ \sum_{j=1}^{n} 1_{R_i}(X_j) 1_y(Y_j) \right\}. \qquad \text{(discrete case)}$$

i.e. the mode.

Decision trees are popular because the decision making (estimate) is nicely visualizable/interpretable if not too deep. One can easily draw out the subsetting tree starting with the entire feature space $\mathcal{X}$ and then spitting in two all the way down to the terminal notes. The interior edges on this representation are called nodes and can also be interpreted as subset of $\mathcal{X}$.

Decision trees can be quite efficient in classification tasks where we are interested in 0-1 (binary) decisions. This is the case in many application but is often not the case in insurance since it is impossible and frankly not usefull having a 0/1 estimate of whether a claim may arrive. Instead the insurance company is interested in the probability that a claim may arrive during a timespan.

9.5.1 CART

A decision tree is uniquely defined by its leaves. Given a tree T, we write $m(T)$ for the corresponding regression or classification function. The naive approach of looking for leaves $R_1, \ldots, R_l$ that minimize a certain loss is practically not feasible because of the computational cost. Instead: top-down greedy approach, called CART (classification and regression trees). We now describe how to construct nodes via the CART algorithm.

Definition. (CART) *Start with $\mathcal{X}$ as initial node for splitting. Follow the process below for all subnotes made from splitting.*

1) *Let $R \subseteq \mathcal{X}$ be the input note. (In initial run $R = \mathcal{X}$)*
2) *For every dimension $j = 1, \ldots, p$ and every point $s_j \in R(j)$ where*

$$R(j) = \{x_j \mid \exists\ x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p \text{ with } (x_1, \ldots, x_p) \in R\}$$

i.e. the support of x_j under the node R. We define the following:

$$R_1(j, s_j) = \{(x_1, \ldots, x_p) \in R \mid x_j \leqslant s_j\}, \quad R_2(j, s_j) = \{(x_1, \ldots, x_p) \in R \mid x_j > s_j\}.$$

being a subdivision of $R = R_1(j, s_j) \cup R_2(j, s_j)$ since $R_1(j, s_j) \cap R_2(j, s_j) = \varnothing$. Notice that one intepret $R_1(j, s_j)$ is the lower section of R through the splitter s_j in dimension j and $R_2(j, s_j)$ being the upper section of the note.

3) *We pick the following minimizer*

$$(j^*, s^*) = \arg\min_{j,s} Q_n(R_1(j, s_j)) + Q_n(R_2(j, s_j))$$

and split the note R into $R_1(j^*, s^*)$ and $R_1(j^*, s^*)$.
4) *If either of $R_1(j^*, s^*)$ and $R_1(j^*, s^*)$ satisfy some predertermined stopping criterion the note is no longer splittet. If not it is further split as long as stopping criterion does not apply.*

Definition. (CART loss) *For regression the most common loss function is the squared loss:*

$$Q_n(R_l) = \sum_{i:X_i \in R_l} (Y_i - \overline{Y}_i(R_l))^2$$

with

$$\overline{Y}_i(R_l) = \frac{1}{|R_l|} \sum_{i:X_i \in R_l} Y_i \text{ and } |R_j| = \#\{i : X_i \in R_j\}$$

In the classification case, define

$$\hat{p}_{lk} = \frac{1}{|R_l|} \sum_{i:X_i \in R_l} 1(Y_i = k)$$

i.e. the proportion of class k observation in R_l.

Other commom loss functions include:

- Brier score: $Q_n(R_l) = \sum_{i:X_i \in R_l} (Y_i - \overline{Y}_i(R_l))^2$ (=squared loss)
- Misclassification error: $Q_n(R_l) = 1 - \max_k \hat{p}_{lk}$
- Gini index: $Q_n(R_l) = \sum_{k=1}^K \hat{p}_{lk}(1 - \hat{p}_{lk})$
- Entropy : $Q_n(R_l) = -\sum_{k=1}^K \hat{p}_{lk} \log \hat{p}_{lk}$

Note: In classification, even if binary loss is the ultimate goal, Gini index and entropy might be the better choices for splitting.

The most common stopping criterion is specifying a min-node size, i.e. stop if $|R_j| < c$. A common choice is $c = 5$. An alternative stopping criteria is the depth of R_j, i.e., the number of parent nodes of R_j.

The process described so far will probably lead to an overfit. One may think that one way out is to stop growing the tree early enough.n This is not advisable because stopping the growing early because the current split does not lead to great improvement in the loss does not mean that a split afterwards might not turn very effective.

9.5.2 Pruning

The idea of pruning is to let a tree grow very deep first (= small min node size) and then select a subtree (pruning) as final fit. Idea: We can compare different subtrees via a penalized loss.

Definition. (Subtree) *We call T' a subtree of T if the nodes of T' is a subset of the nodes of T. T' and T have the same root. If T' is a subtree of T, we also write $T' \subseteq T$.*

We can derive a subtree, by pruning a tree at any non-terminal node (i.e. by deleting all descendants of that node)

Definition. (α-pruned tree) *Consider a tree $\hat{T}_n$ with corresponding estimator $\hat{m}_n(\hat{T}_n)$. Given a parameter $\alpha > 0$, the α-pruned tree is*

$$\hat{T}_{n,\alpha} := \arg\min_{T \subseteq \hat{T}_n} \{\hat{R}_n(\hat{m}_n) + \alpha|T|\},$$

where $|T|$ is the number of leaves of T.

How do we find an optimal α (and also the optimal subtree for a fixed α)? Brute force cross-validation will be too computationally intensive even if we know the answer to the second question. The answer to both questions is the weakest link algorithm.

Proposition (weakest link) *The α-pruned tree $\hat{T}_{n,\alpha}$ is unique. Furthermore, there exist (unique) trees $T_0 \supset \cdots \supset T_l$ with $\{T_0, \dots, T_l\} = \{\hat{T}_{n,\alpha} : \alpha > 0\} := \hat{\mathbb{T}}_n$; in particular the set $\{\hat{T}_{n,\alpha} : \alpha > 0\}$ is finite.*

(Weakest link algorith): Let $\widetilde{Q}_n$ be the loss function corresponding to $\hat{R}_n$. The trees $T_0, \dots, T_l$ can be found with the following algorithm

1) *Let t_L, t_R be any two terminal nodes in $\hat{T}_n$ resulting from a split of the immediate ancestor node t. If $\widetilde{Q}_n(t) = \widetilde{Q}_n(t_L) + \widetilde{Q}_n(t_R)$, prune off t_L and t_R. Continue this process until no more pruning is possible. The resulting tree is T_0.*
2) *$k = 0$*
3) *Go through all non-terminal nodes t of T_k and calculate*

$$g(t) = \frac{\widetilde{Q}_n(t) - \widetilde{Q}_n(T_t)}{|T_t| - 1},$$

   where T_t is the tree (branch) with root t and nodes consisting of t and the descendants of t in T_k. Furthermore $\widetilde{Q}_n(T_t) = \sum_{t \text{ leaf of } T_t} \widetilde{Q}_n(t)$.
4) *$\alpha = \min_{t \text{ non-terminal node of } T_k} g(t)$.*
5) *From top to bottom in T_k prune all non-terminal nodes with $g(t) = \alpha$ and call the resulting tree T_{k+1}.*
6) *If T_{k+1} has more than one node, set $k \leftarrow k + 1$ and go to step 3.*

The above proposition gives a surprising result, that for all $\alpha \in \mathbb{R}_+$ the mapping $\alpha \mapsto \hat{T}_{n,\alpha} \in \hat{\mathbb{T}}_n$ is well defined and the set $\hat{\mathbb{T}}_n$ is finite. Intuitively one can come to this conclusion by pre-assuming that the mapping before is unique and considering that the mapping $f : \alpha \mapsto |\hat{T}_{n,\alpha}|$ is monotonic and that $f(\alpha) \to |\hat{T}_n|$ for $\alpha \to 0_+$ and $f(\alpha) \to 1$ for $\alpha \to \infty$. In particular, the tree is pruned more and more harshly for larger α and for sufficiently large α the tree $\hat{T}_{n,\alpha} \to \mathcal{X}$ for $\alpha \to \infty$.

Given l trees $\{T_0, \dots, T_l\} = \{\hat{T}_{n,\alpha} : \alpha > 0\}$, we can pick the optimal tree via cross validation.

Problem with decision trees. There are (at least) two major problems with decision trees

- Instability: Small variations in the data can lead to a very different tree
- Performance: The performance (measured via test error) is usually much weaker compared to other learning algorithms

9.5.3 Bagging

Bagging stands for Bootstrap aggregation. Idea: Generate artificial data via bootstrap, build a decision tree for each data-set and average. Hope: It reduces the variance of decision trees. In the following recall that Bootstrap sampling involves drawing m samples $(\tilde{X}_i, \tilde{Y}_i)$ from the original dataset with replacement. This gives a new dataset $\tilde{\mathcal{D}}_m$ of new data which may be analised. The process of drawing alot of dataset will reduce the variance by simply averaging over all boostrap samples $\tilde{\mathcal{D}}_m^{(1)}, ..., \tilde{\mathcal{D}}_m^{(B)}$.

Definition. *(Bagging) Given data $(X_1, Y_1), \dots, (X_n, Y_n)$, draw B bootstrap samples $\tilde{\mathcal{D}}_n^{(b)} = (\tilde{X}_1^{(b)}, \tilde{Y}_1^{(b)}), \dots, (\tilde{X}_n^{(b)}, \tilde{Y}_n^{(b)})$, for $b = 1, \dots, B$, i.e., $\tilde{\mathcal{D}}_n^{(b)}$ arises from $\mathcal{D}_n$ by drawing n times with replacement. The bagging estimator is defined as the following. In the squared loss case (average):*

$$\hat{m}_n^{bagg} = \frac{1}{B} \sum_{b=1}^{B} \hat{m}_n(\tilde{\mathcal{D}}_n^{(b)})$$

and in the binary loss case (majority vote):

$$\hat{m}_n^{bagg}(x) = \arg \max_{k \in 1, \dots K} \#\{b : m_n(\tilde{\mathcal{D}}_n^{(b)})(x) = k\}$$

Note that $\hat{m}_n^{bagg}$ is a stochastic estimator, even given $\mathcal{D}_n$. Note: Usually, bagging is applied on non-pruned trees. (See it as an alternative way to reduce variance)

While bagging improves performance of decision trees, interpretability is worsened. If interpretability is not a concern, then bagging is still inferior compared to other learning techniques. Random forests are a modification on bagging estimators.

9.5.4 Random Forests

As bagging random forest are an ensemble of decision trees. They add the following modification to bagging: - `mtry`: Each time a node is split it is only done on a subset of size `mtry` of viable variables. I.e., each time you want to split a node, turn on a random generator and draw $1 \leqslant$ `mtry` $< p$ viable variables from $\{1, \dots, p\}$. Only those variables drawn are allowed to be considered for the next split. - We denote the random forest estimator by $\hat{m}_n^{rf}$. Note that $\hat{m}_n^{rf}$, as $\hat{m}_n^{bagg}$, is a stochastic estimator, even given $\mathcal{D}_n$.

Why is `mtry` so useful? Assume that we are given l estimators $m_n(\mathcal{D}_n, U_1), \dots m_n(\mathcal{D}_n, U_l)$. Here, $U_1, \dots, U_l$ are iid variables inducing the additional stochasticity of the estimators (e.g. through random forest or bagging). Assume that $\hat{m}_n(\mathcal{D}_n, U_1), \dots \hat{m}_n(\mathcal{D}_n, U_l)$ have pairwise correlation ρ and variance σ^2. We get

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{l}\sum_{j=1}^{l}\hat{m}_n(\mathcal{D}_n, U_j)\right) &= \frac{1}{l^2}\sum_{jk}\operatorname{Cov}(m_n(\mathcal{D}_n, U_j), m_n(\mathcal{D}_n, U_k)) \\
&= \frac{1}{l^2}\left(l\sigma^2 + (l^2 - l)\rho\sigma^2\right) \\
&= \rho\sigma^2 + \frac{1-\rho}{l}\sigma^2
\end{aligned}$$

Meaning: Variance of an ensemble of trees is small the smaller the correlation between the trees. The `mtry` parameter aims to make trees less alike and hence more uncorrelated.

Random forest are competitive to many state of the art learners. While not having "best performance" in terms off accuracy they are not too far behind. Advantages of random forests are

- Default hyperparameters already lead to very strong results (i.e. without tuning hyperparameter)
  - `mtry` $= \lfloor\sqrt{p}\rfloor$
  - min node size= 1 for classification, 5 for regression
- Can be implemented (and is implemented) very fast

Random forest (and also trees) don't perform well for additive functions

Decision trees are not good with additive functions (example). Consider $m^*(x) = \sum_j^p 1(x_j \leqslant 0)$ for large p. For a perfect fit, a tree algorithm would have to grow a tree with depth p, where each leaf is the result of splitting once with respect to each covariate. Hence, we end up with 2^p leaves, which on average contain $n/(2^p)$ data points. Even if all splits are optimal for $2^p > n$, the tree will not be able to lead to a perfect decision rule.

While predictions of random forests are relatively smooth, they are not monotonic. e.g. in car insurance, if everything else is the same, more mileage should lead to higher insurance price. Random forests deal well with sparsity, i.e., when the number of features p is large, but the number of relevant features is rather small: $s << p$.

9.6 Boosting and additive trees

In the previous section we have discussed that random forests are not good in estimating additive structures. In this section we will see two tree-based estimators that can deal better with additive structures.

9.6.1 Gradient Boosting Machines

The general idea of boosting is to construct an estimator of the form

$$\hat{m}_n(x) = \sum_{j=1}^{B} \hat{m}_{n,j}(x),$$

where $\hat{m}_{n,b}$ is the result of improving on the estimator $\hat{m}_n^{(b-1)}(x) = \sum_{j=1}^{b-1} \hat{m}_{n,j}(x)$.

Definition. *(Forward Stagewise Additive Modeling) The function $\hat{m}_n(x) = \sum_{j=1}^{B} \hat{m}_{n,j}(x)$ is called Forward Stagewise Additive Modeling estimator if*

$$m_{n,0}(x) = \arg\min_{\eta \in \mathcal{G}} \sum_{i=1}^{N} L\left(Y_i, \eta\right)$$

and for $b = 1, \ldots, B$

$$\hat{m}_{n,b} = \arg\min_{\eta \in \mathcal{G}} \hat{R}_n \left(\hat{m}_n^{(b-1)}(x) + \eta(x)\right),$$

where $\hat{m}_n^{(b-1)} = \sum_{j=1}^{b-1} \hat{m}_{n,j}$.. The set $\mathcal{G}$ is chosen to be very restrictive/small. The component $m_{n,b}$ is called weak learner.

Problem: Finding the exact minimizer is only feasible (not harder than than the usual minimisation) for very specific loss functions $\hat{R}_n$. In the case of squared loss, minimization in step b boils down to minimizing the empirical squared loss with respect to the residuals $Y_i - m_n^{(b-1)}(X_i)$.

$$L\left(Y_i, \hat{m}_n^{(b-1)}(X_i) + \eta(X_i)\right) = L\left(Y_i - \hat{m}_n^{(b-1)}(X_i), \eta(X_i)\right)$$

In the classification case with $\{-1, 1\}$ valued response and an exponential loss ($L(y_1, y_2) = e^{-y_1 y_2}$), minimization in step b boils down to minimizing the weighted empirical exponential loss with observation i receiving weight $e^{-Y_i m_n^{(b-1)}(X_i)}$.

$$L\left(Y_i, \hat{m}_n^{(b-1)}(X_i) + \eta(X_i)\right) = e^{Y_i m_n^{(b-1)}(X_i)} L\left(Y_i, \eta(X_i)\right)$$

This algorithm is also known as AdaBoost.M1. Adaboost.M1 was introduced in Freund and Schapire (1997) and was only later (J. Friedman, Hastie, and Tibshirani 2000) identified as Forward Stagewise Additive Modeling with exponential loss.

Definition. *(Gradient Boosting Machine v1)* (J. H. Friedman 2001) *Gradient Boosting approximates Forward Stagewise Additive Modeling. The idea is to choose to hyperparameters: the function space $\mathcal{G}$ and the learning rate η. The algorithm for chosen hyperparameters is.*

Step 0: Initialize

$$\hat{m}_n^{(0)}(x) = \arg\min_{m \in \mathcal{G}} \sum_{i=1}^{N} L\left(Y_i, m\right).$$

Step $1, ..., B$: For $b = 1, \ldots, B$ do:

(a) *For each observation $i = 1, 2, \ldots, n$, calculate the gradient:*

$$g_{ib} = -\frac{\partial L(Y_i, y)}{\partial y}\Big|_{y = m_n^{(b-1)}(X_i)}.$$

(b) *Improve the estimator by finding:*

$$(\xi_b, \tilde{m}_{n,b}) = \arg\min_{\xi \in \mathbb{R}_+, m \in \mathcal{G}} \sum_{i=1}^{n} (g_{ib} - \xi m(X_i))^2$$

Note: ξ_b is only necessary if $\mathcal{G}$ is not closed under multiplication by constants. (i) *Search for $\tilde{m}_{n,b}$ by line search:*

$$\alpha_b = \arg\min_{\alpha} \hat{R}_n(m_n^{(b-1)} + \alpha\tilde{m}_{n,b})$$

We call $\hat{m}_{n,b}(x) = \alpha_b\tilde{m}_{n,b}(x)$ the weak learner.

(c) *Improve the estimator by setting:*

$$\hat{m}_n^{(b)}(x) = \hat{m}_n^{(b-1)}(x) + \eta\hat{m}_{n,b}(x).$$

One may tune the hyperparameters, max depth J, $\mathcal{G} = \{$trees of max depth $J\}$ and learning rate η, by:

1. Initialize $\hat{m}_n^{(0)}(x) = \arg\min_{m\in\mathcal{G}} \sum_{i=1}^N L(Y_i, m)$.
2. For $b = 1, \ldots, B$:
   (a) For $i = 1, 2, \ldots, n$, calculate the gradient:

$$g_{ib} = -\frac{\partial L(Y_i, y)}{\partial y}\Big|_{y=m_n^{(b-1)}(X_i)}.$$

   (b) Fit a regression tree with squared loss and max depth J to the targets g_{ib}. $\rightarrow$ leaves $R_{bj}, j = 1, 2, \ldots, J_b$.
   (c) For $j = 1, 2, \ldots, J_b$ calculate the value for leaf R_{bj}:

$$v_{bj} = \arg\min_{v_{bj}\in\mathbb{R}} \sum_{i:X_i\in R_{kj}} L\left(Y_i, m_n^{(b-1)}(X_i) + v_{bj}\right).$$

   (d) $\hat{m}_{n,b}(x) = \sum_{j=1}^{J_b} v_{bj} 1(X_i \in R_{bj})$
   (e) Update: $\hat{m}_n^{(b)}(x) = \hat{m}_n^{(b-1)}(x) + \eta\hat{m}_{n,b}(x)$.

The main difference between the tree gradient booster and general gradient boosting is: Least squares problem is not solved directly but greedy step-wise by growing a tree via CART-algorithm. Instead of a line search we minimize the empirical loss in each leaf.

More recently Chen and Guestrin (2016) proposed to optimize a regularized objective in the Forward Stagewise Additive Modeling with the following further changes: 1. Replace $\hat{R}_n$ by $\hat{R}_{n,\gamma,\lambda}(\hat{m}_n^{(b-1)} + m_{n,b}) = \hat{R}_n(\hat{m}_n^{(b-1)} + m_{n,b}) + J_{\gamma,\lambda}(m_{n,b})$ - $J_{\gamma,\lambda}(m_{n,b}) = \gamma J_b + \frac{1}{2}\lambda||v_b||_2^2$ - J_b= number of leaves of tree in iteration b - v_b = leaf values of tree in iteration b 2. Use second order approximation instead of gradient descent (see next slide) - i.e. gradient descent in 2b,c is replaced by a Newton-Raphson type approximation

The algorithm is implemented in the `xgboost` package. Main hyperparameters: - Penalties: γ, λ - learning rate: η - max tree depth

Gradient Boosting (Second version) *XGBoost* Use a second order approximation of $\hat{R}_{n,\gamma,\lambda}$

$$\hat{R}_{n,\gamma,\lambda}(\hat{m}_n^{(b-1)}(x)) + \hat{m}_{n,b}(x)) \tag{9.1}$$

$$\tag{9.2}$$

$$\approx \sum_{i=1}^n \left\{ L(Y_i, \hat{m}_n^{(b-1)}(X_i)) + g_i\hat{m}_{n,b}(X_i) + \frac{1}{2}h_i\hat{m}_{n,b}(X_i)^2 \right\} + J_{\gamma,\lambda}(m_{n,b}) \tag{9.3}$$

$$= \sum_{j=1}^{J_b} \left\{ \left[\sum_{i:X_i\in R_{jb}} g_i \right] v_{jb} + \frac{1}{2} \left[\sum_{i:X_i\in R_{jb}} h_i + \lambda \right] v_{jb}^2 \right\} + \gamma J_b \quad \text{(ignoring constants)} \tag{9.4}$$

with $g_i = \frac{\partial L(Y_i,y_i)}{\partial y_i}\Big|_{y_i=m_n^{(b-1)}(X_i)}$ and $h_i = \frac{\partial^2 L(Y_i,y_i)}{\partial^2 y_i}\Big|_{y_i=m_n^{(b-1)}(X_i)}$. In the last expression, for a fixed tree, the optimal leaf values v_{jb} are given by

$$v_{jb} = -\frac{\sum_{i:X_i\in R_{jb}} g_i}{\sum_{i:X_i\in R_{jb}} h_i + \lambda}$$

This leads to the objective function

$$-\frac{1}{2}\sum_{j=1}^{J_b}\frac{(\sum_{i:X_i\in R_{jb}} g_i)^2}{\sum_{i:X_i\in R_{jb}} h_i+\lambda}+\gamma J_b$$

Hence, in step b, when growing a tree the loss reduction of splitting R_{jb} in $R_{jb,L}, R_{jb,R}$ is given by

$$\frac{1}{2}\left\{\frac{(\sum_{i:X_i\in R_{jb,L}} g_i)^2}{\sum_{i:X_i\in R_{jb,L}} h_i+\lambda}+\frac{(\sum_{i:X_i\in R_{jb,R}} g_i)^2}{\sum_{i:X_i\in R_{jb,R}} h_i+\lambda}-\frac{(\sum_{i:X_i\in R_{jb}} g_i)^2}{\sum_{i:X_i\in R_{jb}} h_i+\lambda}\right\}-\gamma$$

In step b the jth node is split a dimension and position that maximises the loss reduction, given that the maximising loss reduction is positive and the current node has depth smaller J. Otherwise the node is not further split.

Gradient boosting machines are known to often provide the strongest predictive performance. `xgboost` and `lightgbm` are popular implementations. They are quite fast, but not as fast as random forests and also more reliant on optimal parameter tuning. The most relevant parameters are tree depth and learning rate

Why are gradient boosting machines so powerful? A contributing factor may be that a small max depth restricts the number of interactions fitted.

- max dept=1 corresponds to an additive model etc…. Gradient boosting machines, in contrast to random forests have also, no problem to fit additive functions. Different additive components can be fitted in different iterations b.

9.6.2 Bayesian additive regression trees

Another powerful algorithm is BART

Definition. *(Bayesian additive regression trees (BART) (for least squares))* (Chipman, George, and McCulloch 2010)

1. Let $\hat{m}^1(x)=\sum_{k=1}^K \hat{m}_k^1(x)$, $\hat{m}_1^1(x)=\hat{m}_2^1(x)=\cdots=\hat{m}_K^1(x)=\frac{1}{nK}\sum_{i=1}^n Y_i$.
2. For $b=2,\ldots,B$:
   (a) For $k=1,2,\ldots,K$:
      i. For $i=1,\ldots,n$, calculate residuals $r_i=y_i-\sum_{k'<k}\hat{m}_{k'}^b(X_i)-\sum_{k'>k}\hat{m}_{k'}^{b-1}(X_i)$
      ii. Consider a new tree, $\hat{m}_k^b$ by making ONE of the following changes to $\hat{m}_k^{b-1}$ (every change has a fixed probability).
         - GROW: split one leaf
         - PRUNE: prune sister leaves
         - CHANGE: change the decision rule of one node
         - SWAP: swap decision rule of two nodes
      iii. The change is accepted by throwing a coin with the probabilities of the coin given by comparing posterior of old tree and proposed tree. Otherwise, tree is kept: $\hat{m}_k^b=\hat{m}_k^{b-1}$
      iv. Update leaf values for $\hat{m}_k^b$ by sampling from a posterior distribution.
   (b) Set $\hat{m}^b(x)=\sum_{k=1}^K \hat{m}_k^b(x)$.
3. Compute the mean after L burn-in samples, $\hat{m}(x)=\frac{1}{B-L}\sum_{b=L+1}^B \hat{m}^b(x)$

Only a heuristic of BART is presented here. Many packages differ in the exact implementation. In general BART can be seen as a mix of random forest, boosting, additive models. BART is very expensive to train. But if one is able to tune the hyperparameters, BART has often shown to provide unmatched accuracy.

9.7 Some practical considerations

In this lecture we will discuss some practical issues that may help for your projects.

Categorical data. One point often ignored is that in many applications a significant part of features are categorical. All algorithms we have introduced so far implicitly assume, however, that features are numerical. The most "famous" ways of dealing with categorical features are one-hot encoding or dummy encoding.

Definition. *(One-hot encoding)* Given a feature with X that can take k differen categorical values, one-hot encoding entails transforming the feature into k binary features where

$$X^{(l)} = 1(X = l), \quad l = 1, \dots, k.$$

Definition. *(Dummy encoding)* Given a feature with X that can take k differen categorical values, dummy encoding entails transforming the feature into $k-1$ binary features where, given a reference l^*,

$$X^{(l)} = 1(X = l), \quad l = 1, \dots, l^* - 1, l^* + 1, \dots, k.$$

Dummy encoding solves the problem of collinearity when one-hot encoding. Both methods can easily make the dimension of the problem rapidly increase if a feature has many categories e.g. car brand, country, or postcode. Grouping variables with similar effect on the response might reduce variance. This could e.g. be done via expert knowledge. Trees are an automated alternative.

There are $B_k = \sum_{j=0}^{k} \binom{k}{j}$ ways to partition a categorical feature with k values (Bell number).

- $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, $B_6 = 203$, $B_7 = 877$, $B_8 = 4140$.

In a CART algorithm, we would only partition a current node into two child-nodes. In this case there are $2^{k-1} - 1$ possible partitions. This is still not feasible for large k. It is well known Wright and König (2019) that if optimal partition is with respect to gini index or squared loss, then the optimal partition can be found by considering only $k-1$ partitions:

1. Order feature values by mean response.
2. The optimal partition is a contiguous partition. - Search for optimal partition treating the categorical features as numerical, i.e., split into a left and right partition.

Wright and König (2019) propose to order categorical variables only once when fitting a random forest. This is the default in `ranger`. `lightgbm` and very recently (April 2022) `xgboost` offer the option to order variables before every split.

Trees with different loss functions. In many insurance applications it is common to assume that $Y|X$ has a certain (known) parameterized distribution. In this case the loss function can correspond to the negative log likelihood or deviance. When fitting a tree, leaf values then correspond to the maximum likelihood parameter estimate for observations conditioned on that leaf. One advantage is that if the distribution is chosen well this can lead to more robust/better results and additionally it provides estimates of the full distribution. The disadvantage is that: It leaves room for misspecification. If one is only interested in the mean response, then estimating the distribution first may lead to a biased estimate of the mean. Computational time might be an issue if the likelihood function is too complicated.

Example. *(Poisson loss with given exposure)* Assume that the response Y given X is Poisson distributed with mean $\lambda(X)E$. Given iid observations $(X_i, Y_i, E_i)_{i=1,\dots,n}$, the log likelihood in leaf R_1 is given by

$$l(\lambda) = \sum_{i:X_i \in R_1} -\lambda E_i + Y_i \log(\lambda E_i) - log(Y_i!)$$

The minimzer is given by

$$\hat{\lambda} = \frac{\sum_{i:X_i \in R_1} Y_i}{\sum_{i:X_i \in R_1}^{n} E_i}$$

Which has the same minimizer as when minimizing the empirical risk of the Poisson Deviance

$$\hat{R}_n(\lambda E_i) = \sum_{i=1}^{n} 2 \left(Y_i \log \frac{Y_i}{\lambda E_i} - Y_i + \lambda E_i \right).$$

`distRforest` is a recent attempt to add more distributions to random forest.

In boosting algorithms we don't directly work with the loss function but only with its gradient (boosting v1) or with a second order approximation (boosting v2). Assume $Y|X$ follows a distributions from an exponential dispersion family, i.e., the density (for fixed $X = x$) can be written as

$$f\left(Y;\theta,E^{-1}\phi\right)=\exp\left\{\frac{\theta Y-A(\theta)}{\phi/E}+c\left(Y;\phi/E\right)\right\}.$$

ϕ is dispersion parameter. E is exposure. A is twice continuously differentiable, gradient of A one-to-one. We ignore the dispersion parameter ϕ since it does not change point prediction if it does not change over different observations $i = 1, \dots, n$. Then, the negative log likelihood up to additive terms that do not depend on θ is

$$-E\times(\theta Y-A(\theta))$$

We can translate this to a generalized framework: We could say we wish to estimate the distribution $\tilde{g}(\mathbb{E}[Y])$, given canonical link function $\tilde{g}$. $\Rightarrow \tilde{g}(\mathbb{E}(Y)) = \theta$. Meaning we wish to minimize the loss $L(Y,\theta) = -E \times (\theta Y - A(\theta))$. Hence, when employing a gradient boosting machine algorithm we can use the gradient:

$$g_i=\frac{\partial L(Y_i,\theta)}{\partial\theta}\Big|_{\theta=m_n^{(b-1)}(X_i)}=-E_i\times\left(Y-\frac{\partial A(\theta)}{\partial\theta}\Big|_{\theta=m_n^{(b-1)}(X_i)}\right),$$

and the second derivative:

$$h_i=\frac{\partial^2 L(Y_i,\theta)}{\partial^2\theta}\Big|_{\theta=m_n^{(b-1)}(X_i)}=E_i\times\left(\frac{\partial^2 A(\theta)}{\partial^2\theta}\Big|_{\theta=m_n^{(b-1)}(X_i)}\right).$$

`xboost` and `lightgbm` come with many pre-configured loss functions.

Imbalanced data. When considering individual insurance claims data, one will find that most policies do not have a claim registered. In other words: The number of observed zeros will be much higher than observations of other numbers when looking at claim frequencies or claim amounts. This can be a problem when modelling under a Poisson assumptions and zero inflated distributions have been suggested. It is usually not problematic when modelling under a Bernoulli assumption (logistic regression). It can be a problem if we are instead interested in 0/1 decisions (e.g. in fraud detection). In this case, there have been loss modifications suggested that reward more detecting 1s than detecting 0s. A prominent example is focal loss (Lin et al. 2017). In pricing we are rather interested in the probability and not in 0/1 decisions.

Claim amounts are often modeled via a frequency severity approach modelling frequency and severity separately (assuming independence between frequency and severity):

$$\pi=\mathbb{E}\left(\frac{L}{E}\right)=\mathbb{E}\left(\frac{N\times Y}{E}\right)=\mathbb{E}\left(\frac{N}{E}\right)\times\mathbb{E}\left(Y\right)=\mathbb{E}(F)\times\mathbb{E}(Y).$$

where π= technical price, L= loss, E= exposure, N= frequency and Y= severity. Note that it is important here how to define a claim (is a claim with claim size zero a claim?). If it is is not counted as a claim, then modelling $E[Y]$ is easier.

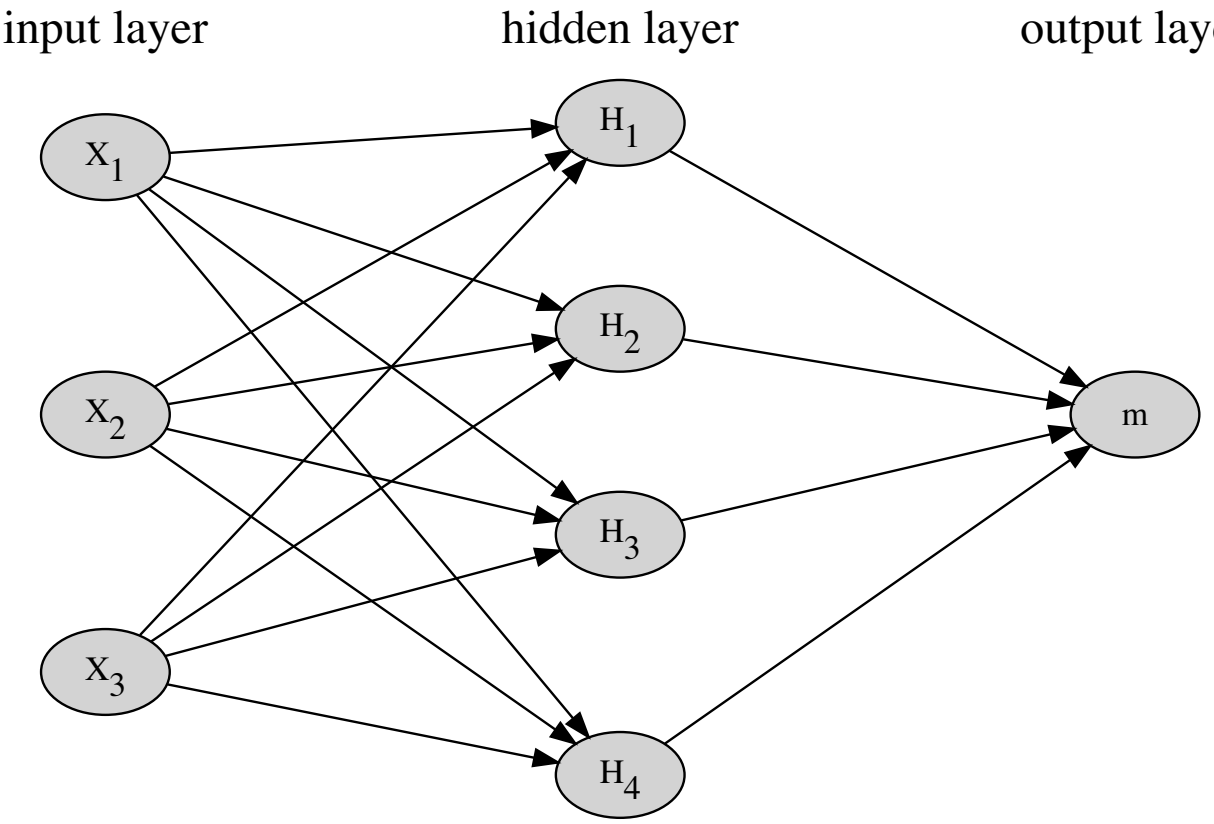
9.8 Neural Networks

In this lecture we will introduce neural networks. Neural networks with multiple layers are also known as deep learners. We will only consider feed forward neural networks. Usually used for tabular (=unstructrued) data. The recent popularity of neural networks stems mostly from modification more suitable for structured data, like

- Convolutional neural networks for image classification
- Transformers in natural learning processing

As usual, we assume that we are given an iid dataset $\mathcal{D}_n = \{X_i, Y_i\}_{i=1,\dots,n}$.

Single Layer Feed Forward. We start with a single layer neural network (= 1 hidden layer).



The architecture of a single layer neural network corresponds to the function class $\mathcal{G}$ such that $\hat{m}(D_n)(x) = \hat{m}_n(x) : \mathcal{X} \mapsto \mathcal{Y}$ can be written as

$$\hat{m}_n(x) = g\left(\beta_0 + \sum_{k=1}^{K} \beta_k H_k(x)\right) \tag{9.5}$$

$$= g\left(\beta_0 + \sum_{k=1}^{K} \beta_k \phi\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_j\right)\right). \tag{9.6}$$

In the the previous slide the illustration showed $p = 3$, $K = 4$. To ease notation we will be ignoring the intercepts, here: β_0, w_{k0}. In matrix notation, the hidden layer is $H(x) = \phi(x^t w)$ and the output is

$$\hat{m}_n(x) = g\left(H(x)^T \beta\right) = g\left(\phi(x^T w)^T \beta\right).$$

with $w \in \mathbb{R}^{p \times K}$, $\beta \in \mathbb{R}^K$. The functions ϕ and g are called activation functions and are applied element-wise. The activation functions make the estimator non-linear.

Popular activation functions.

- Sigmoid: $\phi(z) = \frac{e^z}{1+e^z}$
  - maps to $[0, 1]$
- Hyperbolic tangent: $\phi(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^x + e^{-z}}$
  - maps to $[-1, 1]$
- Rectified Linear Unit (ReLU): $\phi(z) = \max(z, 0)$
  - "standard" activation function
- Leaky ReLU: $\phi(z) = \max(z, 0) + \alpha \min(0, z)$

Backpropagation. Neural networks are fitted via gradient descent with small step sizes. We are hence interested in $\frac{\partial L(Y_i,m(\beta,w))}{\partial w}, \frac{\partial L(Y_i,m(\beta,w))}{\partial \beta}$. Calculating the gradient via the chain rule is called backpropagation.

Note that the loss is calculated via the following composition:

$$X_i \xrightarrow{w} Z_1 \xrightarrow{\phi} H \xrightarrow{\beta} Z_2 \xrightarrow{g} m \to L(Y_i, m)$$

Hence, by the chain rule we have

$$\frac{\partial L(Y_i, \partial m)}{\partial \beta} = \frac{\partial L(Y_i, m)}{m} \frac{\partial m}{Z_2} \frac{\partial Z_2}{\partial \beta},$$

as well as

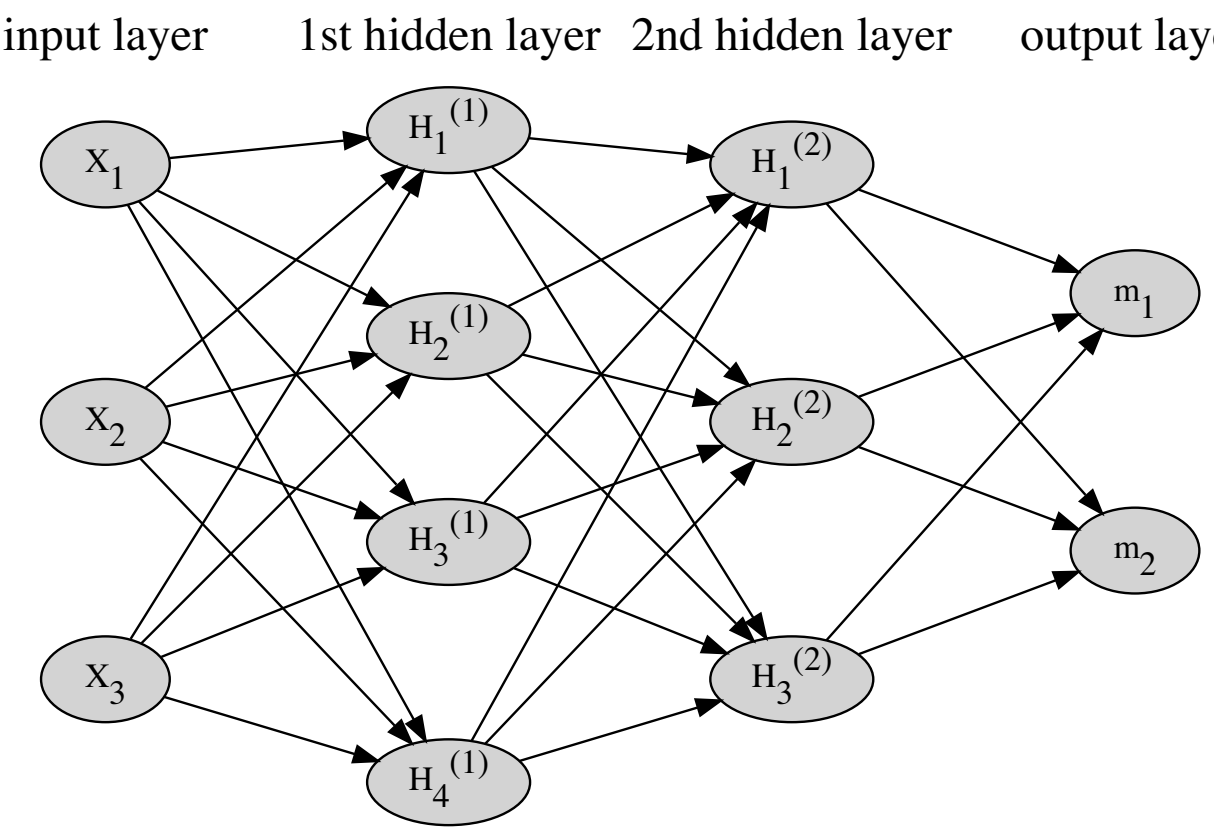
$$\frac{\partial L(Y_i, m)}{\partial w} = \frac{\partial L(Y_i, m)}{\partial m} \frac{\partial m}{\partial Z_2} \frac{\partial Z_2}{\partial H} \frac{\partial H}{\partial Z_1} \frac{\partial Z_1}{\partial w}.$$

In step r and given a learning rate γ the weights are updated as

$$\beta^{(r+1)} = \beta^{(r)} - \gamma \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L(Y_i, \hat{m}_n^{(r)}(X_i))}{\partial \beta^{(r)}},$$

$$w^{(r+1)} = w^{(r)} - \gamma \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L(Y_i, \hat{m}_n^{(r)}(X_i))}{\partial w^{(r)}}.$$

Feed foreward Neural Network. Usually one works with multiple hidden layers and possibly mutlipe outputs.



the lth hidden layer arises from the $l-1$th hidden layer from (ignoring intercepts)

$$H^{(l)} = \phi_l(H^{(l-1)} w^{(l)}).$$

The fitting is analogue to the single layer case. To avoid overfitting one can imploy penalty terms, e.g. lasso and/or ridge. In the machine learning community this is also known as weight decay. Especially ridge penalty is popular. Ridge penalty means penalizing every entry of β, w by its square.

Stochastic Gradient Descent and Early Stopping. In "deterministic" gradient descent, weights are updated via

$$\beta^{(r+1)} = \beta^{(r)} - \gamma \sum_{i=1}^{n} \frac{\partial L(Y_i, \hat{m}_n^{(r)}(X_i))}{\partial \beta^{(r)}}, \quad w^{(r+1)} = w^{(r)} - \gamma \sum_{i=1}^{n} \frac{\partial L(Y_i, \hat{m}_n^{(r)}(X_i))}{\partial w^{(r)}}.$$

For large data sets this may not be computationally feasible/too expensive and an alternative is Stochastic Gradient Descent (SGD).

Definition. *Stochastic Gradient Descent (SGD)*

Input: `batch-size`, early stopping criteria (most often whether improvement is better than some treshold on the validation set)

- For $j = 1, \ldots$STOP (=epochs)
  1. For $k = 1, \ldots, n/$`batch-size`
  2. Sample without replacement `batch-size` from $\{1, \ldots, n\} \setminus \cup_{l=1}^{k-1} B_j^j$, resulting in the kth batch, B_k^j.
  3. Perform gradient descent with using only observations $i \in B_k^j$

Dropout learning. An alternative/additional way to perform regularization is dropout, introduced in Srivastava et al. (2014). It is an analogue to random forest, but the obvious idea of averaging the outputs of many separately trained nets is prohibitively expensive. Instead: In every mini-batch, for every i, randomly remove a fraction p of the units in a layer when calculating the gradient. The surviving units get an additional weight of $1/(1p)$. I.e. in epoch j and mini-batch k, for individual i and current unit $H(j,k)$,

$$\widetilde{H}(j,k,i) = \begin{cases} 0 & \text{with probability } p \\ H(k,j)/(1p) & \text{else} \end{cases}$$

Note: $\mathbb{E}[\widetilde{H}(j,k,i)] = H(k,j)$. The heuristic is that many thinned networks are trained in parallel.

More Hyperparameters. An alterantive to a ridge penalty is max-norm regularization, with parameter c. I.e. in every learning step enforce $||\beta||_2^2, ||w||_2^2 \leqslant c$, where $||\cdot||_2^2$ denotes the squared sum of all entries.

Instead of standard SGD, alternative updates can be performed, e.g., momentum:

$$w^{(r+1)} = w^{(r)} + \alpha \Delta w^{(r)} - \gamma \sum_{i=1}^{n} \frac{\partial L(Y_i, \hat{m}_n^{(r)}(X_i))}{\partial w^{(r)}}$$
$$\Delta w^{(r)} = w^{(r)} - w^{(r-1)}$$

or Adam.

Further comments. The size of the learning rate can depend on the current epoch, also known as learning rate decay. Initial weights, i.e, starting values for w, β, do effect the outcome. Covariates are often standardized via minmax scaling.

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \in [0,1]$$

In R, implementation of neural network is provided via keras/tensorflow https://tensorflow.rstudio.com/. Given the many options to configure, initialize and stop training a neural network, training a neural network is often seen as "art".

For tabular (=unstructured) data tree-based algorithms usually outpeform neural networks with respect to test error. Neural networks are however unchallenged for structured data (especially when also extending the feed forward network accounting for the structure)

9.9 Local explanations

In this lecture we shift the perspective and we wish to understand what a predictor $\hat{m}_n(x)$ is actually doing.

9.9.1 Interpretability

Regression: $Y_i = m(X_i) + \varepsilon_i$

$$X \quad \rightarrow \quad \widehat{m}(X) \quad \rightarrow \quad m(X) \quad \text{or} \quad Y \mid X$$

Interpretability is understanding the relationship between X and $\widehat{m}(X)$. This is different to understanding the relationship between X and Y.

Quality	Linear Models	Machine Learning
interpretable	yes	no
interactions manually	yes	yes
interactions	no	yes
variable selection/sparsity	yes	yes
non-linearity	no	yes

Current machine learning algorithms are often highly flexible and can deal with interaction, non-linearity, sparsity and variable selection; but they are not interpretable.

Local explanations. Fix a value $x_0 \in \mathbb{R}^p$. A local approximation at x_0 of the function $\hat{m}_n$ is given by

$$\hat{m}\left(x_0\right) = \phi_0 + \sum_{k=1}^{p} \phi_k(x_0),$$

where $\phi_0, \phi_1(x_0), \ldots, \phi_p(x_0)$ are constants. Note: the right hand-side is not identified. Local explanations add constraints such that $\phi_k(x_0)$ is uniquely identified and best reflects the local contribution of feature k to $\hat{m}_n\left(x_0\right)$.

Definition. *(Additive Value functions)* A value function v assigns a real value $v(S)$ to each subset $S \subseteq \{1, \ldots p\}$.

Definition. *(Shapley Axioms)*

Given a vaue function v_{x_0} with $v_{x_0}(\{1, \ldots, p\}) = m(x_0)$. The four Shapley axioms are

1. **Efficiency** $m\left(x_0\right) = \phi_0 + \sum_{k=1}^{p} \phi_k(x_0)$, $\phi_0 = v(\varnothing)$.
2. **Symmetry**: Fix any $k, l \in \{1, \ldots, p\}, k \neq l$. If $v_{x_0}(S \cup k) = v_{x_0}(S \cup l)$, for all $S \subseteq \{1, \ldots p\} \backslash \{k, l\}$, then $\phi_k(x_0) = \phi_l(x_0)$.
3. **Dummy** For $k = 1, \ldots, p$: If $v_{x_0}(S \cup k) = v_{x_0}(S)$, for all $S \subseteq \{1, \ldots p\} \backslash \{k\}$, then $\phi_k = 0$.
4. **Linearity** For $k = 1, \ldots, p$: If $m(x_0) = m^1(x_0) + m^2(x_0)$, then $\phi_k(x_0) = \phi_k^1(x_0) + \phi_k^2(x_0)$, where ϕ^l is the explanation corresponding to the function m^l.

Theorem. *(Shapley Axioms)* Given a value function v_{x_0}, there exist unique constants $\phi_0, \phi_1(x_0), \ldots, \phi_p(x_0)$ such that the four Shapley axioms are satisfied. They are given by

$$\phi_k(x_0) = \frac{1}{p!} \sum_{\pi \in \Pi_p} \Delta_{v_{x_0}}\left(k, \{\pi(1), \ldots, \pi(k-1)\}\right) \tag{9.7}$$

$$= \frac{1}{p!} \sum_{S: S \subseteq \{1, \ldots, p\} \backslash \{k\}} |S|!(p - |S| - 1)! \Delta_{v_{x_0}}(k, S), \tag{9.8}$$

where $\Delta_{v_{x_0}}(k, S) = v_{x_0}(S \cup k) - v_{x_0}(S)$ and Π_p is the set of permutations of $\{1, \ldots, p\}$.

Proof.

For $R \subseteq \{1, \ldots, p\}$, and a value fucntion v, define the value function

$$v_R(S) = \begin{cases} m(x_0) & \text{if } R \subseteq S \\ 0 & \text{else} \end{cases}$$

- By the symmetry axiom, if $j, k \in R$, then $\phi_j(v_R) = \phi_k(v_R)$.
- By the dummy axiom, if $k \notin R$, then $\phi_k(v_R) = 0$.

Hence, by the efficiency axiom ($v_R(\{1, \ldots, p\}) = m(x_0)$), for $k \in R$,

$$\phi_k(v_R) = \frac{m(x_0)}{|R|}.$$

We will (later) show that

$$v(U) = \sum_{T: T \subseteq \{1,\ldots,p\}} \sum_{S: S \subseteq T} (-1)^{|T-S|} \frac{v(S)}{m(x_0)} v_T(U). \tag{1}$$

Then, by the linearity axiom

$$\phi_k(v) = \sum_{T: T \subseteq \{1,\ldots,p\}} \sum_{S: S \subseteq T} (-1)^{|T-S|} \frac{v(S)}{m(x_0)} \phi_k(v_T) \tag{9.9}$$

$$= \sum_{T: T \subseteq \{1,\ldots,p\}, k \in T} \sum_{S: S \subseteq T} (-1)^{|T-S|} \frac{v(S)}{|T|} \tag{9.10}$$

$$= \sum_{S: S \subseteq \{1,\ldots,p\}} \sum_{T: S \cup \{k\} \subseteq T} (-1)^{|T-S|} \frac{v(S)}{|T|}. \tag{9.11}$$

We can write

$$\phi_k(v) = \sum_{S: S \subseteq \{1,\ldots,p\}} \gamma_i(S) v(S), \quad \gamma_k(S) = \sum_{T: S \cup \{k\} \subseteq T} (-1)^{|T-S|} \frac{1}{|T|}$$

Observe that for $S_1 \neq S_2, S_2 = S_1 \cup \{k\}$, $\gamma_k(S_1) = -\gamma_k(S_2)$. Hence,

$$\phi_k(v) = \sum_{S: S \subseteq \{1,\ldots,p\}} \gamma_k(S) v(S) \tag{9.12}$$

$$= \sum_{S: S \subseteq \{1,\ldots,p\}, k \in S} \gamma_k(S) v(S) + \sum_{S: S \subseteq \{1,\ldots,p\}, k \notin S} \gamma_k(S) v(S) \tag{9.13}$$

$$= \sum_{S: S \subseteq \{1,\ldots,p\}, k \notin S} -\gamma_k(S) v(S \cup \{k\}) + \sum_{S: S \subseteq \{1,\ldots,p\}, k \notin S} \gamma_k(S) v(S) \tag{9.14}$$

$$= \sum_{S: S \subseteq \{1,\ldots,p\}, k \notin S} -\gamma_k(S) \left[v(S \cup \{k\}) - v(S) \right] \tag{9.15}$$

The proof follows by showing that

$$\gamma_k(S) = -\frac{|S|!(p - |S| - 1)!}{p!}, \qquad k \notin S, \tag{2}$$

as well as equation (1).

We first show (2). We will use that

$$\frac{1}{l} = \int_0^1 x^{l-1} \mathrm{d}x.$$

Such that for $k \notin S$, we have

$$\gamma_k(S) = \sum_{l=0}^{p-|S|-1} (-1)^{l+1} \binom{p-|S|-1}{l} \frac{1}{|S|+1+l} \tag{9.16}$$

$$= -\int_0^1 x^{|S|} \sum_{l=0}^{p-|S|-1} (-1)^l \binom{p-|S|-1}{l} x^l \mathrm{d}x. \tag{9.17}$$

By the binomial theorem, this simplifies to

$$\gamma_k(S) = -\int_0^1 x^{|S|}(1-x)^{p-|S|-1}\mathrm{d}x \tag{9.18}$$

$$= -\frac{|S|!(p-|S|-1)!}{p!}. \tag{9.19}$$

The last equality follows from

$$\int_0^1 x^a(1-x)^b\mathrm{d}x = \frac{a!b!}{(a+b+1)!}, \qquad a, b \geqslant 0$$

which can be proven by induction over b. It remains to show (1). We have

$$\sum_{T:T\subseteq\{1,\ldots,p\}} \sum_{S:S\subseteq T} (-1)^{|T-S|} \frac{v(S)}{m(x_0)} v_T(U) \tag{9.20}$$

$$= \sum_{T:T\subseteq U} \sum_{S:S\subseteq T} (-1)^{|T-S|} v(S) \tag{9.21}$$

$$= \sum_{S:S\subseteq U} \left[\sum_{D:D\subseteq\{U\setminus S\}} (-1)^{|D|} \right] v(S) \tag{9.22}$$

$$= v(U), \tag{9.23}$$

where the last equation follows because the expression in the square bracket is zero for $U \neq S$. That is because a non-empty set has an equal number of subsets with an odd number of elements as subsets with an even number of elements.

Note that we will slightly abuse notation by ignoring ordering in the input of the functions below. Lundberg and Lee (2017) proposed to use Shapley values with value function

$$v(S) = \mathbb{E}[\hat{m}_n(X_S, X_{-S})|X_S = x_S] = \int \hat{m}_n(x_1, \ldots, x_p) p_X(x_S, X_{-S})\mathrm{d}x_{-S}$$

for model explanation. And called it SHAP (SHapley Additive exPlanations). To simplify calculations, Lundberg and Lee (2017) proposed to calculate

$$v(S) = \mathbb{E}[\hat{m}_n(x_S, X_{-S})] = \int \hat{m}_n(x_1, \ldots, x_p) p_{X_{-S}}(X_{-S})\mathrm{d}x_{-S}$$

Janzing, Minorics, and Blöbaum (2020) argue that the latter value function should actually be the preffered value function. Chen et al. (2020) coin it interventional SHAP (and the original SHAP above observational SHAP). In the next, we give an example why interventional SHAP values might be preferred compared to observational SHAP values.

Example. *(Observational SHAP vs Interventional SHAP)* Assume

$$\hat{m}_n(x_1, x_2) = x_1$$

Let X_1, X_2 be binary with

$$p(x_1, x_2) = \begin{cases} \frac{1}{2} & \text{if } x_1 = x_2 \\ 0 & \text{else} \end{cases}$$

For observational SHAP, we have

- $v_x(\varnothing) = \mathbb{E}[\hat{m}_n(X_1, X_2)] = 0.5$
- $v_x(\{1\}) = \mathbb{E}[\hat{m}_n(X_1, X_2)|X_1 = x_1] = x_1$
- $v_x(\{2\}) = \mathbb{E}[\hat{m}_n(X_1, X_2)|X_2 = x_2] = x_2$
- $v_x(\{1, 2\}) = \hat{m}_n(x_1, x_2) = x_1$

Hence,

- $\Delta(2, \varnothing) = v_x(\{2\}) - v_x(\varnothing) = x_1 - 0.5$
- $\Delta(2, \{1\}) = v_x(\{1, 2\}) - v_x(\{1\}) = x_1 - x_1 = 0$

Such that

- $\phi_2 = \frac{1}{2}(x_1 - 0.5) \neq 0$

For interventional SHAP, we have

- $v_x(\varnothing) = \mathbb{E}[\hat{m}_n(X_1, X_2)] = 0.5$
- $v_x(\{1\}) = \mathbb{E}[\hat{m}_n(x_1, X_2)] = x_1$
- $v_x(\{2\}) = \mathbb{E}[\hat{m}_n(X_1, x_2)] = 0.5$
- $v_x(\{1, 2\}) = [\hat{m}_n(x_1, x_2) = x_1$

Hence,

- $\Delta(2, \varnothing) = v_x(\{2\}) - v_x(\varnothing) = 0$
- $\Delta(2, \{1\}) = v_x(\{1, 2\}) - v_x(\{1\}) = 0$

Such that

- $\phi_2 = 0$

We conclude, that if one uses observational SHAP for model explanation, then the contribution of a feature is a combination of its own contribution and the contribution of features it is correlated with.

Discussion Point. If $\mathbb{P}(X_1 = X_2)$, then $\tilde{m}_n(x) = x_2$ is as good in estimating the response as $\hat{m}_n(x) = x_2$. In particular the Bayes rule is not unique. With that regard, it could make sense to attribute x_1 and x_2 the same importance. An argument against assigning x_2 any importance is that if ϕ is to explain the behaviour of $\hat{m}_n(x) = x_2$, then it is not directly affected by x_2.

We will later see that observational SHAP has one further issue: One may need to extrapolate in order to calculate it.

How do we calculate/estimate Shapley values? Rember, Shapley values are:

$$\phi_k(x_0) = \frac{1}{p!} \sum_{\pi \in \Pi_p} \Delta_{v_{x_0}}(k, \{\pi(1), \ldots, \pi(k-1)\}) \tag{9.24}$$

$$= \frac{1}{p!} \sum_{S: S \subseteq \{1, \ldots, p\} \setminus \{k\}} |S|!(p - |S| - 1)! \Delta_{v_{x_0}}(k, S), \tag{9.25}$$

To calculate Shapley values exact, one would need to evaluate $p!$ or (second equation) 2^p summands. For large p that may be infeasible.

Definition. *(Estimate Shapley values via permutation sampling)*

The Shapley value ϕ_k can be approximated by the following algorithm:

For $r = 1, \ldots,$

- Sample a permutation π of $\{1, \ldots, p\}$

- Approximate $\Delta_{v_{x_0}}(k, \{\pi(1), \dots, \pi(k-1)\})$ (call the result $\Delta^{(r)}$)
  - I.e approximate $v(\{\pi(1), \dots, \pi(k-1)\} \cup \{k\})$ and $v(\{\pi(1), \dots, \pi(k-1)\})$
    - $\Delta^{(r)}$ is the difference
  - If $v(S) = \mathbb{E}[\hat{m}_n(X_S, X_{-S})|X_S = x_S]$, (observational SHAP)
    - Not clear what to do. One would need to estimate the distribution of X first.
      - Which is a high dimensional problem…
  - If $v(S) = \mathbb{E}[\hat{m}_n(x_S, X_{-S})]$ (interventional SHAP)
    - Draw m individuals from $\{1, \dots, n\}$, say $I \subseteq n$,
    - $v(S) \approx \frac{1}{I} \sum_{i \in I} \hat{m}_n(x_S, x_{i,-S})$

$\hat{\phi}_k$ is the average of all $\Delta^{(r)}$

Problem: many samples needed to get accurate approximation. Hence, slow.

Charnes et al. (1988) discussed that Shapley values can be re-written as the solution of the following constraint minimisation problem:

$$(\phi(x_0))_{0,\dots,p} = \arg \min_{(\phi_k)_{k=1,\dots,p}} \mu(S) \sum_{S: S \subseteq \{1,\dots,p\}} \left(v_{x_0}(S) - \left[\phi_0 + \sum_{k \in S} \phi_k \right] \right)^2,$$
$$\mu(S) = \frac{p-1}{\binom{p}{|S|}|S|(p-|S|)}.$$

Note that $\mu(\varnothing) = \mu(\{1, \dots, p\}) = \infty$ and so the minimisation is not well defined. The infinite weight practically enforces $\phi_0 = v_{x_0}(\varnothing)$, $\sum_{k=0}^{p} \phi_k = v_{x_0}(\{1, \dots, p\})$. One can hence, exclude the two sets $(\varnothing, \{1, \dots, p\})$ from the minimisation and add the two constraints, $\phi_0 = v_{x_0}(\varnothing)$, $\sum_{k=0}^{p} \phi_k = v_{x_0}(\{1, \dots, p\})$, to the minimisation.

This is a quadratic programing problem (good), but estimating $v_{x_0}(S)$ for all 2^p subsets might economically not be feasible.

Introduced in Lundberg and Lee (2017), Covert and Lee (2021) give a detailed explanation on how Kernel SHAP is implemented. Instead of estimating $v_{x_0}(S)$ for all 2^p subsets they only evaluate n subsets. The m subsets are drawn from the set of all subsets of $\{1, \dots, p\}$ minus the empty set and the full set, where the probability of drawing set S is proportional to $\mu(S)$

Hence the final minimisation reads

$$\min_{(\phi_k)_{k=1,\dots,p}} \frac{1}{m} \sum_{i=1}^{m} \left(v_{x_0}(S_i) - \left[v_{x_0}(\varnothing) + \sum_{k \in S_i} \phi_k \right] \right)^2, \tag{9.26}$$

$$\text{subject to} \sum_{k=1}^{p} \phi_k = v_{x_0}(\{1, \dots, p\}) - v_{x_0}(\varnothing), \tag{9.27}$$

where S_i is the subset from draw i. The value function used in Kernel SHAP is interventional SHAP, i.e., $v_{x_0}(S) = \mathbb{E}[\hat{m}_n(x_S, X_{-S})]$.

Lundberg and Lee (2017) propose a method to estimate interventional SHAP that is fast and exact (exact in the sense that it calculates the exact plug-in estimates). It is specific for tree based algoirthms. The algorithm is called TreeSHAP and it makes use of the binary tree structures that leads to a partitioning of the feature space. For calculating $\mathbb{E}[\hat{m}_n(x_S, X_{-S})]$, TreeSHAP recursively follows the decision path for x_0 if the split feature is in S, and takes the weighted average of both branches if the split feature is not in S. For efficient caulculation, TreeSHAP does not go down the tree for every S separately but only once and while doing so keeps track of all possible S.

9.10 Causality

This lecture covers a selected topic on causaliy leading to the so called adjustment formula. We mostly follow Peters, Janzing, and Schölkopf (2017).

Definition. *(Graph Terminology)*

- We are given a random variables $X = (X_1, ..., X_p)$ with index set $V := \{1, ..., p\}$.
- A graph $G = (V, \mathcal{E})$ consists of
  - nodes or vertices V and
  - edges $\mathcal{E} \subseteq V^2$ with $(v, v) \notin \mathcal{E}$ for any $v \in V$.
- A node k is called a
  - parent of j if $(k, j) \in \mathcal{E}$ and $(j, k) \notin \mathcal{E}$
    - The set of parents of k is denoted by $pa_G(k)$
  - a child if $(j, k) \in \mathcal{E}$ and $(k, j) \notin \mathcal{E}$.
    - The set of childrens of k is denoted by $ch_G(k)$
- Two nodes k and j are adjacent if either $(k, j) \in \mathcal{E}$ or $(j, k) \in \mathcal{E}$.
- We say that there is an undirected edge between two adjacent nodes k and j if $(k, j) \in \mathcal{E}$ and $(j, k) \in \mathcal{E}$.
- An edge between two adjacent nodes (k, j) is directed if $(k, j) \in \mathcal{E}$ and $(j, k) \notin \mathcal{E}$ or vice versa.
  - We write $k \to j$ for $(k, j) \in \mathcal{E}$, $(j, k) \notin \mathcal{E}$ and $j \to k$ for $(j, k) \in \mathcal{E}$, $(k, j) \notin \mathcal{E}$
- G is called directed if all its edges are directed.
- A path in G is a sequence of (at least two) distinct vertices $k_1, \ldots, k_m$, such that there is an edge between k_l and k_{l+1} for all $l = 1, \ldots, m1$.
  - If $k_{l-1} \to k_l$ and $k_{l+1} \to k_l$ ($k_{l-1} \to k_l \leftarrow k_{l+1}$), k_l is called a collider relative to this path.
  - If $k_l \to k_{l+1}$ for all l, we speak of a directed path from k_1 to k_m
    - In this case, We call k_1 an ancestor of k_m and
    - k_m a descendant of k_1.
- G is called a directed acyclic graph (DAG) if all edges are directed and there is no pair (j, k) with directed paths from j to k and from k to j.

Definition. *(Pearl's d-separation)*

In a DAG G, a path between nodes k_1 and k_m is blocked by a set S (with neither k_1 nor k_m in S) if there is a node k_l fulfilling one of the two points:

(a) $k_l \in S$ and

$$k_{l-1} \to k_l \to k_{l+1} \tag{9.28}$$
$$\text{or} \quad k_{l-1} \leftarrow k_l \leftarrow k_{l+1} \tag{9.29}$$
$$\text{or} \quad k_{l-1} \leftarrow k_l \to k_{l+1} \tag{9.30}$$

(b) neither k_l nor any of its descendants is in S, and

$$k_{l-1} \to k_l \leftarrow k_{l+1}$$

In a DAG G, we say that two disjoint subsets of vertices A and B are d-separated by a third (also disjoint) subset S if every path between nodes in A and B is blocked by S.
- We then write

$$A \perp_G B \mid S$$

Definition. *(Structural causal models)* A structural causal model (SCM) $\mathfrak{C} := (S, P_N)$ consists of a collection S of p (structural) assignments

$$X_j := f_j\left(pa(j), N_j\right), \quad j = 1, \ldots, p,$$

and a distribution $P_N = \times_{j=1}^{p} P_{N_j}$ of jontly independent noise variables $N_1 \ldots, N_p$.

Note: An SCM $\mathfrak{C}$ defines a unique graph G and a unique distribution, $P_X^{\mathfrak{C}}$, over the variables $X_1, ..., X_p$. The reverse is not true.

Example. *(Structural causal models with acyclic graph structure)*

$$\begin{aligned} X_1 &:= f_1(X_3, N_1) \\ X_2 &:= f_2(X_1, N_2) \\ X_3 &:= f_3(N_3) \\ X_4 &:= f_4(X_2, X_3, N_4) \end{aligned}$$

$N_1, \ldots, N_4$ jointly independent

Definition. *(Interventional distribution/ do-operator)*

- Consider an SCM $\mathfrak{C} := (S, P_N)$ and its entailed distribution $P_X^{\mathfrak{C}}$.
- We deifine a new SCM, $\tilde{\mathfrak{C}}$, by replacing the assignment for X_k by

  $$X_k = \tilde{f}(\tilde{pa}(k), \tilde{N}_k).$$

  The resulting entailed distribution of the new SCM is called interventional distribution. We write short:
  $$do(X_k = \tilde{f}(\tilde{pa}(k), \tilde{N}_k)), \quad P_X^{\tilde{\mathfrak{C}}} = P_X^{\mathfrak{C}, do(X_k = \tilde{f}(\tilde{pa}(k), \tilde{N}_k))}$$
- An intervention can also simply assign a fixed value a (i.e., $\tilde{N}_k = a$ (deterministic) and the set of parents is empty).
  - This is called an atomic or deterministic intervention and can be denoted by $do(X_k = a)$.

Definition. *(Total causal effect)* Given an SCM $\mathfrak{C}$, we say there is a total causal effect from X to Y if

$$X \perp\!\!\!\perp Y \quad \text{ in } P_X^{\mathfrak{C}; do\left(X = \tilde{N}_X\right)}$$

for some random variable $\tilde{N}_X$.

Note: A directed path from X to Y is a necessary but not sufficient condition for a total causal effect (effects can cancel out). In contrast, a directed path from X to Y or Y to X is NOT necessary for X and Y to be correlated.

Definition. *(Markov property)* Given a DAG G and an entailed joint absolutely continuous distribution P_X, this distribution is said to satisfy

(i) the global Markov property with respect to the DAG G if

$$A \perp\!\!\!\perp_{\mathcal{G}} B | C \Rightarrow A \perp\!\!\!\perp B | C$$

for all disjoint sets of nodes A, B, C.

(ii) the local Markov property with respect to the DAG G if each variable is independent of its non-descendants given its parents, and

(iii) the Markov factorization property with respect to the DAG G if

$$p(x) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j \mid pa(j)).$$

Theorem. *(Equivalence of Markov properties, see e.g. Lauritzen (1996))* If P_X is absolutely continuous, then all three Markov properties above are equivalent.

Remark (SCM induced Graph is Markov): Note that every SCM induced Graph satisfies the markovian properties.

Proposition. *(Adjustment formula)* Consider an SCM over variables V with $X, Y \in X$ and $Y \notin pa(X)$.

- Z is called valid adjustment set if it fulfills one of the three following conditions
  1. "parent adjustment": $$Z = pa(X)$$
  2. "backdoor criterion": $Z \subseteq V \backslash \{X, Y\}$ with
     (i) Z contains no descendant of X AND
     (ii) Z blocks all paths from X to Y entering X through the backdoor ($X \leftarrow \ldots$).
  3. "toward necessity": $Z \subseteq V \backslash \{X, Y\}$ with
     (i) Z contains no descendant of any node on a directed path from X to Y (except for descendants of X that are not on a directed path from X to Y) AND
     (ii) Z blocks all non-directed paths from X to Y
- If Z is a valid adjustment set, then

$$p^{\mathfrak{C}, do(X=x)}(y) = \int p^{\mathfrak{C}}(y \mid x, z) p^{\mathfrak{C}}(z) \mathrm{d}z.$$

Proof.

We only proof "parent adjustment" and "backdoor criterion". "towards necessity" can be looked up in Shpitser, VanderWeele, and Robins (2010)

We call the interventional density $\tilde{p}$ and the original density p.

For the parent adjustment, $Z = pa(X)$, note that

$$\tilde{p}(y|z) = \tilde{p}(y|x, z) = p(y|x, z) \tag{9.31}$$
$$\tilde{p}(z) = p(z) \tag{9.32}$$

Hence,

$$p(y|do(X = x)) = \tilde{p}(y) = \int \tilde{p}(y|z)\tilde{p}(z)\mathrm{d}z = \int p(y|x, z)p(z)\mathrm{d}z$$

- For the backdoor adjustment, let Z fulfill (i) and (ii) and let $S = pa(X)$. We have

$$p(y|do(X = x)) \overset{\text{parent adjustment}}{=} \int p(y|x, s)p(s)\mathrm{d}s \tag{9.33}$$
$$= \int p(s) \int p(y, z|x, s)\mathrm{d}s\mathrm{d}z \tag{9.34}$$
$$\overset{\text{Bayes formula}}{=} \int p(s) \int p(y|x, s, z)p(z|x, s)\mathrm{d}s\mathrm{d}z \tag{9.35}$$
$$\overset{\text{(ii):}Y \perp S|X,Z}{=} \int p(s) \int p(y|x, z)p(z|x, s)\mathrm{d}s\mathrm{d}z \tag{9.36}$$
$$\overset{\text{(i):}X \perp Z|S}{=} \int p(s) \int p(y|x, z)p(z|s)\mathrm{d}s\mathrm{d}z \tag{9.37}$$
$$\overset{p(z)=\int p(s)p(z|s)\mathrm{d}s}{=} \int p(y|x, z)\mathrm{d}z \tag{9.38}$$

Intuition behind backdoor criterion: Backdoor paths carry spurious associations from X to Y. Blocking all backdoor paths ensures that measured assoziation is causal. We don't include descendants of that are also ancestors of because this would block a causal path. We don't include descendants of that are also descendants of because this would introduce collider bias.

Questions.

a. Do we need to observe Z_1 to be able to calculate $P(Y|do(X = x))$?
b. What are all valid adjustment sets for calculating $P(Y|do(X = x))$?

Solution. All valid adjustment sets are

$$\{Z_1, Z_3\}, \{Z_2, Z_3\}, \{Z_1, Z_2, Z_3\}.$$

In particular, Z_1 is not needed need to calculate $P(Y|do(X = x))$.

Measuring Total causal effect. How can we calculate $\mathbb{E}[Y|do(X = x)] - \mathbb{E}[Y]$, i.e. the total causal effect of $X = x$ on Y. If we have a valid adjustment set Z, then

$$E[Y|do(X = x)] = \int y p^{do(X=x)}(y)\ dy \tag{9.39}$$

$$= \int\int yp(y|x, z)p(z)\ dy\ dz \tag{9.40}$$

$$= \int p(z) \int yp(y|x, z)\ dy\ dz \tag{9.41}$$

$$= \int p(z)\mathbb{E}[Y|X = x, Z = z]\ dz. \tag{9.42}$$

Which can be estimated from iid observations of (X, Z, Y).

Counterfactual fairness. Assume U is a set of protected features. For example $U = \{\text{gender, ethnicity}\}$. Let $U \cup V = \{1, \dots, p\}, U \cap V = \varnothing$. Can we debias an algorithm predicting $\mathbb{E}[Y|X = x]$ such that it does not use information contained in X_U; neither directly nor indirectly. Kusner et al. (2017) introduced the notion of counterfactual fairness. We here present a sufficient condition: Given a structural causal model of (X, Y), an estimator is counterfactual fair if it is a function of the non-descendents of X_U. A counterfactual estimator is in particular given by $\hat{m}_n^{debiased}(x_v) = E[\hat{m}_n(X)|do(X_v = x_v)]$

9.11 Local and Global Explanations

In this lecture we will introduce some global explanations and see how they are connected to local explanations. We will also see how these explanations can be used for de-biasing.

9.11.1 Interpretability

Reminder: Why interpretability.

- Algorithmic accountability
  - Are risk estimates transparent?
  - EU regulation
    - GDPR: "Individuals have the right to an explanation of the logic behind the decision."
    - EU AI Act
- If process of decision making is transparent
  - → Biases easier to detect
  - → more robustness a
    - unmeasured confounders can hunt you later under distributional shifts

9.11.2 Partial dependence plots

Partial dependence plots are popular post-hoc global explanations

Definition. *(Partial dependence plots Friedman 2001)* Given an estimator $\hat{m}_n$ and a target subset $S \subset \{1, \dots, p\}$, the partial dependence plot [], ξ_S, is defined as

$$\xi_S(x_S) = \int \hat{m}_n(x)\hat{p}_{-S}(x_{-S})\mathrm{d}x_{-S}.$$

In particular, a partial dependence plot for feature k is

$$\xi_k(x_k) = \int \hat{m}_n(x) p_{-k}(x_{-k}) \mathrm{d}x_{-k}.$$

Partial dependence can be misleading because they ignore interaction effects. Approximation is highly non-trivial and can be very unstable trough extrapolatio.

9.11.3 A functional decomposition

Assume a data set with p features. Also assume that we can approximate the regression function m by a $q-th$ order functional decomposition:

$$m(x) \approx m_0 + \sum_{k=1}^{p} m_k(x_k) + \sum_{k_1<k_2} m_{k_1 k_2}(x_{k_1}, x_{k_2}) + \cdots + \sum_{k_1<\cdots<k_q} m_{k_1,\ldots,k_q}(x_{k_1}, \ldots, x_{k_q}).$$

Optimal rates of convergence under the assumption that m has two continuous partial derivatives:

Model general	**general** p	$p = 6$	**Comparable sample sizes for p=6**
Full model	$O_p(n^{-2/(p+4)})$	$O_p(n^{-1/5})$	1 000 000
Interaction (q)	$O_p(n^{-2/(p+4)})$	$O_p(n^{-2/(p+4)})$	1 000 - 1 000 000
Interaction (q=2)	$O_p(n^{-1/3})$	$O_p(n^{-1/3})$	4 000
Additive (q=1)	$O_p(n^{-2/5})$	$O_p(n^{-2/5})$	1 000

Generalized ANOVA is probably the most considered functional decomposition identification.

Generalized ANOVA: For every$S \subseteq \{1, \ldots, d\}$ and $k \in S$,

$$\int m_S(x_S) \int w(x) \mathrm{d}x_{-S} \, \mathrm{d}x_k = 0$$

Common choices for w are

- $w \equiv 1$
- $w(x) = p(x)$, $p =$ density of X.
- $w(x) = \prod p_j(x_j)$, $p_j =$ density of X_j.

Generalized ANOVA, however, has drawbacks for interpretability

- $w = 1$ ignores the distribution of X
- $w = p$ uses the correlation structure (analog to observational SHAP)
- $w = \prod p_j$ Assumes that features are independent

Definition. *(Marginal Identification)* For every $S \subseteq \{1, \ldots, d\}$,

$$\sum_{T: T \cap S \neq \varnothing} \int \hat{m}_T(x_T) p_S(x_S) \mathrm{d}x_S = 0.$$

Theorem. Given any initial estimator $\hat{m}^{(0)} = \{\hat{m}_S^{(0)} | S \subseteq \{1, \ldots, d\}\}$, there exists exactly one set of functions $\hat{m}^* = \{\hat{m}_S^* | S \subseteq \{1, \ldots, d\}\}$ satisfying the marginal identification with $\sum_S \hat{m}_S^* = \sum_S \hat{m}_S^{(0)}$. The functions are given by

$$\hat{m}_S^*(x_S) = \sum_{T \supseteq S} \sum_{T \setminus S \subseteq U \subseteq T} (-1)^{|S|-|T \setminus U|} \times \int \hat{m}_T^{(0)}(x_T) \hat{p}_U(x_U) \mathrm{d}x_U.$$

In particular $\hat{m}^*$ does not depend on the particular identification of $\hat{m}^{(0)}$.

The are three reasons why the marginal identification is particularly interesting

- Partial dependence plots
- Interventional SHAP values
- Fairness/ discrimination-free pricing

Remember: partial dependence plot, ξ_S, is defined as

$$\xi_k(x_S) = \int \hat{m}(x) p_{-S}(x_{-S}) \mathrm{d}x_{-S}.$$

Corollary. If $\hat{m}^*$ satisfies the marginal identification, then

$$\xi_S = \sum_{U \subseteq S} \hat{m}^*_U.$$

In particular if S is only one feature, i.e., $S = \{k\}$, we have

$$\xi_k(x_k) = \hat{m}^*_0 + \hat{m}^*_k(x_k).$$

Theorem.

- $\hat{m}^*$ satisfies the marginal identification
- Then, interventional SHAP values are weighted averages components
  - interaction component is equally split between involved features:

$$\phi_k(x) = \hat{m}^*_k(x_k) + \frac{1}{2}\sum_j \hat{m}^*_{kj}(x_{kj}) + \cdots + \frac{1}{d}\hat{m}^*_{1,\ldots,d}(x_{1,\ldots,d}).$$

Assume U is a set of protected features. For example $U = \{\text{gender, ethnicity}\}$. Let $U \cup V = \{1, \ldots, d\}, U \cap V = \varnothing$. $E[m(X)|\ do(X_V = x_V))$ does not use information contained in X_U; neither directly nor indirectly. Under the assumed causal structure we have

$$E[m(X)|\ do(X_V = x_V)] = \int m(x) p_U(x_U) dx_U.$$

Under marginal identifiaction:

$$\int \hat{m}^*_n(x) \hat{p}_U(x_U) dx_U = \sum_{S \subseteq V} \hat{m}^*_S(x_S),$$

i.e., a "de-biased" estimator can be extracted from $\hat{m}_n$ by dropping all components that include features in U.

Summary (Identification) If m is identified via martginal identification,

$$\hat{m}_n(x) = \hat{m}^*_0 + \sum_j \hat{m}^*_j(x_j) + \sum_{j<k} \hat{m}^*_{j,k}(x_j, x_k) + \cdots$$

then

- **Interventional SHAP values** are:

$$\phi_k(x) = \hat{m}^*_k(x_k) + \frac{1}{2}\sum_j \hat{m}^*_{kj}(x_{kj}) + \cdots$$

- **Partial dependence plots** are: $\xi_k(x_k) = \hat{m}^*_0 + \hat{m}^*_k(x_k)$.
- **Fairness/Discrimination-free pricing:** If S are protected variables and all components that contain a subset of S are dropped,we derive a de-biased estimator.

What does discrimination free pricing actually mean?

Average claim Size

Car	Male	Female
Male Car Type	\$1	\$2
Female Car Type	\$2	\$1

Proportions in population ($p_{X,D}$)

Car	Male	Female
Male Car Type	0.45	0.05
Female Car Type	0.05	0.45

If we use $p_D(d) = 0.5$ for pricing, then the price for Male Car Type 1 and Female Car Type would be \$1.50 each.

Problem: The prices are **not calibrated** now: The insurance company is charging more than the *expected average claim size of \$1.10* . (We ignore risk loading) We could multiply the price by $\frac{1.10}{1.50} \approx 0.7333$. Hence charge everyone \$1.10.

Problem: Is the price still discrimination free? Note that \$1.10 is also the price if we would ignore gender completely, $\mathbb{E}[Y|X = x]$ (i.e. allow for indirect discrimination). The answer will most likely be around the question why we wanted to de-bias for gender in the first place; what are the risk-factors we want to use, and what are unmeasured confounders.

Chapter 10

Quantative Risk Management

10.1 The Loss Variable

We describe the financial risk with random variables defined on a filtered probability space $\left(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}\right)$. We assume that *the value process* V_t, denoting the market value of the portfolio at time t, is adapted to the filtration i.e. may be determined at time t or from information available at time t. We will be considering discrete time jumps of size Δt for simplicity and assume that

- the portfolio remains fixed over the time horizon $[t, t+\Delta t)$,
- there are no income or payments made during the time period.

This means in particular that the value

$$\Delta V_{t+\Delta t} = V_{t+\Delta t} - V_t,$$

only depends on the valuation of the components in the portfolio. We will henceforth be using the notation t and $t+1$ and so forth denoting the intervals $[t+n\Delta t, t+(n+1)\Delta t)$ i.e. t may be any time and the integer representing how many time jumps of size Δt has been made since t.

Under the assumptions above it is reasonable to assume that V_t is Markovian in the sense that there exist $d \geqslant 1$ random sources $\mathbf{Z}_t = (Z_{t,1}, ..., Z_{t,d})^\top$ such that

$$V_t = f(t, \mathbf{Z}_t) \tag{2.2}$$

for some measurable function $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$. We will call $\mathbf{Z}_t$ the *risk factors* associated with the portfolio. The could be for instance the stockprice of a stock held in the portfolio. We may now define the *loss variable* as $L_{t+1} := -(V_{t+1} - V_t)$ and *risk-factor changes* as $\mathbf{X}_{t+1} := \mathbf{Z}_{t+1} - \mathbf{Z}_t$ and see that

$$L_{t+1} = -\left(f(t+1, \mathbf{Z}_t + \mathbf{X}_{t+1}) - f(t, \mathbf{Z}_t)\right), \tag{2.3}$$

if one assume differentiability of f we may approximate L_{t+1} as

$$L^{\Delta}_{t+1} := -\left(\frac{\partial f}{\partial t}(t, \mathbf{Z}_t) + \sum_{i=1}^{d} \frac{\partial f}{\partial z_i}(t, \mathbf{Z}_t)\mathbf{X}_{t+1,i}\right). \tag{2.4}$$

This is obviously a nice linear but the approximation error may be large if Δt is large. We are interested in the the distribution of L_{t+1} such that we may determine sufficiently capital holding in realtion to the risk associated with the assets and liabilities held on the firm's balance sheet.

10.1.1 Risk measures

In the general sense, a *risk measure* is simply a measurable function $\rho : \mathbb{L} \to \mathbb{R}$ taking a loss variable $L \in \mathbb{L}$ as input and associating a number $\rho(L)$ as output. In this setting we let $\mathbb{L}$ be the set of all real-valued random variables. We may interpret this as the riskyness of the portfolio with associated loss variable L. That is say we have two loss variable from the value processes V and V' i.e. L and L' we say that V is riskier than V' if and only if $\rho(L) \geqslant \rho(L')$.

We may now consider the fundamental definition of a coherent risk measures as in *Coherent Measures of Risk* by Artzner, Delbean, Eber and Heath (1999), by first stating the definition of a risk measure.

Definition. (Risk Measure) *Let $\mathbb{L}$ be the set of all real valued random variable i.e.*

$$\mathbb{L} = \{X \mid X : (\Omega, P, \mathcal{F}) \to (\mathbb{R}, \mathbb{B})\}$$

Any mapping $\rho : \mathbb{L} \to \mathbb{R}$ is called a measure of risk.

We now want some properties to be satisfied by the mapping ρ such that it is a sensable measure of risk. To this we define four axioms as.

Definition. (Coherent Risk Measure) *Let ρ be a measure of risk. We say that ρ is a coherent risk measure if and only if ρ satisfies the axioms below.*

1. **Translation invariance:** *Let $\alpha \in \mathbb{R}$ and r be a predictable process we have $\rho(X + \alpha \cdot r) = \rho(X) + \alpha$.*
2. **Subadditivity:** *Let $X_1, X_2 \in \mathbb{L}$, then $\rho(X_1 + X_2) \leqslant \rho(X_1) + \rho(X_2)$.*
3. **Positive homogeneity:** *Let $c > 0$, then $\rho(cX) = c\rho(X)$.*
4. **Monotonicity:** *Let $X, Y \in \mathbb{L}$ such that $X \leqslant Y$ P-a.s., then $\rho(X) \leqslant \rho(Y)$.*

Remark: We see that axiom (1) gives the equation $\rho(X + \rho(X) \cdot r) = 0$ i.e. we may hedge the risk by holding $\rho(X)$ in a asset with rate of return r. The axiom (2) ensures that we take into account the correlation of multiple assets i.e. we will in general not experience the maximal loss of two sources at the same time (although this is possible). If X_1 and X_2 satisfies $\text{Corr}(X_1, X_2) = 1$ then $\rho(X_1 + X_2) = \rho(X_1) + \rho(X_2)$.

There exist alot of different tangible measures of risk, some being coherent others non-coherent. The most well studied include Value at Risk VaR and Expected Shortfall ES. These two measures are in the realm of the loss distribution approach, however before studying these we introduce a few other worthy mentions: Factor sensitivity measures and scenario based risk measures:

Factor sensitivity measures are measures on the form $\rho = g(\nabla L)$ where g is some d-dimensional measurable function. In this ∇L is the gradient i.e.

$$\nabla L = \nabla\Big(f(t, \mathbf{Z}_t) - f(t+1, \mathbf{Z}_{t+1})\Big) = \left(\frac{\partial L}{\partial z_1}, ..., \frac{\partial L}{\partial z_d}\right).$$

Scenario based risk measures are measures where one assume that a collection $\mathbf{x} = (x_1, ..., x_N)$ of events $x_i \in \Omega$ such that $\sum_{i=1}^N P(x_i) = 1 - \varepsilon$ for some small $\varepsilon > 0$. We may then associate the measures ψ as

$$\psi(L) = \max\{w_1 L(x_1), ..., w_N L(x_N)\},$$

with $\mathbf{w} \geqslant 0$ and $\sum_{i=1}^N w_i = 1$ being weights. That is ψ gives the largest weighted loss. A natural way of choosing w_i is such that $w_i \approx P(x_i)$, but one may also weight certain events with a larger weight for a more prudent measure. Notice also that the criteria $\sum_{i=1}^N P(x_i) = 1 - \varepsilon$ does not necessarily have to be satisfied when considering the worst possible outcomes.

10.1.1.1 Value at Risk

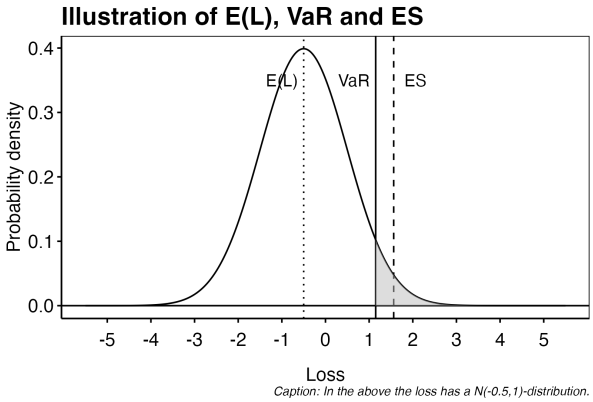
The Value at Risk, henceforth VaR, is a quantile measure of the loss distribution F_L associated with L. We define VaR as such:

Definition 2.8. (McNeil) (Value-at-Risk) *Let $\alpha \in (0,1)$ (α being large) we define the VaR of a portfolio with loss variable L at the confidence level α is a given by*

$$\begin{aligned} \mathrm{VaR}_\alpha(L) &= \inf\{l \in \mathbb{R} \ : \ P(L > l) \leqslant 1-\alpha\} \\ &= \inf\{l \in \mathbb{R} \ : \ F_L(l) \geqslant \alpha\} \\ &= F_L^{\leftarrow}(\alpha), \end{aligned}$$

where $F_L^{\leftarrow}$ is the generalized inverse of F_L.

There exist alot of different tangible measures of risk, some being coherent others non-coherent. The most well studied include Value at Risk VaR and Expected Shortfall ES. These two measures are in the realm of the loss distribution approach, however before studying these we introduce a few other worthy mentions: Factor sensitivity measures and scenario based risk measures:



Caption: In the above the loss has a N(-0.5,1)-distribution.

Chapter 11

Measure theory

11.1 Axioms of Probability

Som udgangspunkt betragtes rummet $(\Omega, \mathcal{A})$ udstyret med en brolægning $\mathcal{A} \subseteq 2^\Omega$, hvor $2^\Omega = \mathcal{P}(\Omega)$ er mængden af alle delmængder af Ω. Typisk anvendes brolægningerne 1) en algebra eller 2) en σ-algebra.

Definition 2.1. (Protter) (Algebra) *En brolægning $\mathcal{A} \subseteq 2^\Omega$ kaldes en algebra, hvis*

1. $\Omega \subseteq \mathcal{A}$,
2. $\forall A \in \mathcal{A} \Rightarrow A^c = \Omega \backslash A \in \mathcal{A}$ *(lukket under komplementer),*
3. $A_1, A_2, ..., A_n \in \mathcal{A} \Rightarrow \cup_{i=1}^n A_i \in \mathcal{A}$ *(lukket under endelige foreninger).*

Definition 2.2. (Protter) (σ-algebra) *En brolægning $\mathcal{A} \subseteq 2^\Omega$ kaldes en σ-algebra, hvis*

1. $\Omega \subseteq \mathcal{A}$,
2. $\forall A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ *(lukket under komplementer),*
3. $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \Rightarrow \cup_{i\in\mathbb{N}} A_i \in \mathcal{A}$ *(lukket under tællelige foreninger).*

Egenskaber for en σ-algebra: 1) $\varnothing \in \mathcal{A}$, 2) $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$, og 3) $(A_i)_{i\in\mathbb{N}} \subseteq \mathcal{A} \Rightarrow \cap_{i\in\mathbb{A}} A_i \in \mathcal{A}$.

Eksempel: Borel-sigma-algebraen $\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{C})$, hvor $\mathcal{C}$ kan være en af følgende frembrigersystem. ($\sigma(\mathcal{C})$ benævner den mindste σ-algebra på Ω, hvor $\mathcal{C} \in \mathcal{A}$).

i. De åbne mængder dvs. $\mathcal{C} = \mathcal{O}^n$
ii. De lukkede mængder dvs. $\mathcal{C} = \{A^c : A \in \mathcal{O}^n\}$
iii. Halvåbne mængder/bokse
iv. Uendelige intervaller som $\mathcal{C} = \{(-\infty, a] : a \in \mathbb{R}\}$

Definition 2.2. (Protter) (Sandsynlighedsmål) *Lad $\mathcal{A} \subseteq 2^\Omega$ være en σ-algebra. $P : \mathcal{A} \to [0, 1]$ kaldes et sandsynlighedsmål hvis*

1. $P(\Omega) = 1$,
2. *For $(A_n)_{n\in\mathbb{N}}$ af parvist disjunkte delmængder af Ω gælder $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$.*

Bemærkning: Husk et mål $\mu : \mathcal{A} \to [0, \infty)$ på $\mathcal{A}$ opfylder at 1) $\mu(\varnothing) = 0$ og 2) for en parvis disjunkt familie $(A_n)_{n\in\mathbb{N}}$ gælder $\mu\left(\dot{\bigcup}_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$.

Konsekvenser: Umildbare konsekvenser ved definition 2.3 er følgende

i. $\sum_{i=1}^n P(A_i) = P\left(\bigcup_{i=1}^n A_i\right)$, hvis alle A_i er pavist disjunkte
ii. $0 \leqslant P(A) \leqslant 1$, for alle $A \in \mathcal{A}$
iii. $P(A^c) = 1 - P(A)$, for alle $A \in \mathcal{A}$
iv. $P(A) \leqslant P(B)$ hvis $A \subseteq B$

Theorem 2.3. (Protter) (Opad- og nedadkontinuitet) *Lad $P : \mathcal{A} \to [0,1]$ være et ssh. mål på $(\Omega, \mathcal{A})$. Da gælder*

iii. *$P(A_n) \uparrow P(A)$ hvis $A_n \uparrow A$,*
iv. *$P(A_n) \downarrow P(A)$ hvis $A_n \downarrow A$*

11.2 Conditional Probability and Independence

Definition 3.1. (Protter) (Uafhængighed) *Lad $(\Omega, \mathcal{A}, P)$ være et ssh. rum.*

a. *Lad $A, B \in \mathcal{A}$ være to hændelser. A og B kaldes uafhængige hvis $P(A \cap B) = P(A)P(B)$.*
b. *Lad $A_i \in \mathcal{A}$ for en indeksmængde $i \in I$ (ikke et krav om endelighed eller tællelig). Hændelserne A_i kaldes uafhængige hvis $P(\cap_{i\in J} A_i) = \prod P(A_i)$ for et $J \subseteq I$ med $\#J < +\infty$.*

Bemærkning. Der gælder at 1) $\varnothing$ og Ω er uafhængige af alle $A \in \mathcal{A}$ samt $A \in \mathcal{A}$ er uafhængig med sig selv hvis og kun hvis $P(A) = \{0, 1\}$.

Theorem 3.1. (Protter) *Lad $(\Omega, \mathcal{A}, P)$ være et ssh. rum. Antag $A, B \in \mathcal{A}$ være uafhængige, så er følgende par uafhængige: $(A^c, B), (A, B^c)$ og (A^c, B^c).*

Definition 3.2. (Protter) (Betinget sandsynlighed) *Lad $A, B \in \mathcal{A}$ være hændelser og $P(B) > 0$. Den betingede sandsynlighed A givet B er $P(A|B) = P(A \cap B)/P(B)$.*

Theorem 3.2. (Protter) *Lad $A, B \in \mathcal{A}$ være hændelser og $P(B) > 0$.*

a. *A og B er uafhængige hvis og kun hvis $P(A|B) = P(A)$.*
b. *Funktionen $P(\cdot|B) : \mathcal{A} \to [0,1]$ definerer et nyt ssh. mål på $\mathcal{A}$, kaldet **den betingede sandsynlighed givet** B.*

Andet kan tilføjes f.eks. Bayes'.

11.3 Probabilities on a Finite or Countable Space

Lad Ω være et endelig eller tællelig mængde og lad σ-algebraen $\mathcal{A} = 2^\Omega$.

Theorem 4.1. (Protter) (Punktsandsynligheder) *Lad $A, B \in \mathcal{A}$ være hændelser og $P(B) > 0$.*

a. *En sandsynlighed på en tællelig eller endelig mængde Ω er givet ved sandsynlighederne for hvert atom $p_\omega = P(\{\omega\})$, $\omega \in \Omega$.*
b. *Hvis en følge af reelle tal $(p_\omega)_{\omega\in\Omega}$ indiceret over elementerne i Ω opfylder at $p_\omega \geqslant 0$ og $\sum_{k\in\Omega} p_k = 1$, så eksisterer et unikt sandsynlighedsmål P givet ved $P(\{\omega\}) = p_\omega$.*

Bemærkning. Alle sandsynlighedsmål på endelige eller tællelige mængder Ω kan således karakteriseres ved punktsandsynlighederne p_ω. Dvs. et sandsynligedsmål $P : 2^\Omega \to [0,1]$ er givet ved summen af punktsandsynligheder

$$P(A) = \sum_{k\in A} p_k, \ A \subseteq \Omega$$

Definition 4.1. (Protter) *En ssh. P på en endelig mængde Ω er uniform hvis p_ω afhænger af ω.*

Eksempler:

(Den uniforme fordeling.) Det følger direkte af definition 4.1, at den uniforme fordeling er givet ved

$$P(A) = \frac{\#A}{\#\Omega}$$

(Binomialfordelingen) Lad $\Omega = \{0, 1, 2, ..., n\}$ og $\mathcal{A} = 2^\Omega$. Givet et $q \in [0,1]$ defineres binomialfordelingen ved

$$p_k = \binom{n}{k} q^k (1-q)^{n-k}, \ k \in \Omega$$

(Geometriske fordeling) Lad $\Omega = \mathbb{N}_0 = \{0, 1, 2, ...\}$ og $\mathcal{A} = 2^{\mathbb{N}_0}$. Givet et $q \in [0, 1]$ er den geometriske fordeling givet ved

$$p_k = (1 - q)^k q, \ k \in \Omega$$

(Hypergeometrisk fordeling) Lad $N, M \in \mathbb{N}$ være givet.

(Poisson fordelingen) Lad $\Omega = \mathbb{N}_0$ og $\mathcal{A} = 2^{\mathbb{N}_0}$. Givet et parameter $\lambda > 0$ er poissonfordelingen gived ved

$$p_n = e^{-\lambda} \frac{\lambda^n}{n!}, \ n \in \Omega$$

Desuden er $K(n, k) = \binom{n}{k}$ givet ved

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

11.4 Construction of a Probability Measure on $\mathbb{R}$

Sandsynlighedsmålet kan indledelsesvis indføres på følgende vis: Lad $(\Omega, \mathcal{A}, \mu)$ være et målrum og lad $f : (\Omega, \mathcal{A}) \to [0, \infty]$ være $\mathcal{A}/\mathcal{B}$-målelig. Definer nu målet $\nu : \mathcal{A} \to [0, \infty]$ ved

$$V(A) = \int_A f d\mu = \int 1_A f d\mu = \int 1_A(x) f(x) d\mu(x) = \int 1_A(x) f(x) \mu(dx), \ A \in \mathcal{A}$$

I situationen hvor $V(\Omega) = 1$ er ν et ssh. mål. Især er vi interesseret i målrummet $(\mathbb{R}^n, \mathcal{B}_n)$. Notationen $\nu = f \cdot \mu$ bruges og betyder "ν har tæthed f mht. μ".

Eksempler. *(Lebesguemålet)* Lad $(\Omega, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{B}, m)$ været et målrum, hvor m er lebesgue-målet. Lad $f : \mathbb{R} \to (0, 1/\sqrt{2\pi}]$ være givet ved

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \ x \in \mathbb{R},$$

da er f målelig, da f er kontinuer for alle $x \in \mathbb{R}$. Da er $\nu : \mathcal{B} \to [0, 1]$, også kaldet *normalfordelingen*, et ssh. mål givet ved

$$\nu(A) = \int_A f(x) dm(x), \qquad \nu(\mathbb{R}) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dm(x) = 1$$

(Tællemålet) Lad $(\mathbb{N}_0, 2^{\mathbb{N}_0}, \mu)$ været et målrum og lad μ være tællemålet. Lad $f : \mathbb{N}_0 \to [0, \infty]$ være en vilkårlig målelig funktion. Da er ν givet ved

$$\nu(A) = \int_A f d\mu = \sum_{x \in A} f(x)$$

hvis $\nu(\mathbb{N}_0) = \sum_{x \in \mathbb{N}_0} f(x) = 1$ er ν et ssh. mål på $(\mathbb{N}_0, 2^{\mathbb{N}_0})$.

Definition 7.1. (Protter) (Fordelingsfunktionen/distribution function) *Lad P være et ssh. mål på $(\mathbb{R}, \mathcal{B})$, da er fordelingsfunktionen $F : \mathbb{R} \to [0, 1]$ defineret ved.*

$$F(x) = P((-\infty, x]).$$

Theorem 7.1. (Protter) (Entydighed) *Fordelingsfunktionen F karakteriserer P og er unik.*

Corollary 7.1. (Protter) *Lad F være en fordelingsfunktion for P på $\mathbb{R}$. Definer da $F(x-) := \lim_{u \uparrow x} F(u)$. For vilkårlige $x, y \in \mathbb{R}$ gælder*

i. $P((x, y]) = F(y) - F(x)$
ii. $P([x, y]) = F(y) - F(x-)$
iii. $P(\{x\}) = F(x) - F(x-)$

Theorem 7.2. (Protter) (Egenskaber for fordelingsfunktionen) *$F : \mathbb{R} \to [0, 1]$ er en fordelingsfunktion for et unikt sandsynlighedsmål P på $(\mathbb{R}, \mathcal{B})$, hvis*

i. *F er ikke-aftagende dvs. $F(x) \leqslant F(y)$, for $x \leqslant y$*
ii. *F er højre kontinuer dvs. $F(u) \downarrow F(x)$ for $u \downarrow x$.*
iii. *$F(x) \to 1$ når $x \to \infty$, $F(x) \to 0$ når $x \to -\infty$.*

Bemærkning. Teorem 7.2 kan fungere som et værktøj til konstruktion af et sandsynligedsmål givet en fordelingsfunktion.

11.5 Random Variables

Indledelsesvis genopfriskes målbare rum. For to rum $(E, \mathcal{E})$ og $(F, \mathcal{F})$, hvor $\mathcal{E}$ og $\mathcal{F}$ er σ-algebraer på de respektive rum. Da kaldes en en afbildning $X : (E, \mathcal{E}) \to (F, \mathcal{F})$ for en målelig afbildning, hvis og kun hvis for alle delmængder $A \in \mathcal{F}$ er $X^{-1}(A) \in \mathcal{E}$. Med andre ord for enhver billedmængde, kan vi måle urbilledet, hvorfra funktionsværdierne på billedmængden kunne være kommet fra.

Definition. (Stokastisk variabel) *En målelig afbildning udstyret med et sandsynlighed mål på domænet dvs. $X : (\Omega, \mathcal{A}, P) \to (F, \mathcal{F})$, kaldes en stokastisk variabel (stochastic/random variable).*

Bemærkning. Oftest betragtes $F = \mathbb{R}^n$ og $\mathcal{F} = \mathcal{B}_n$ især når $n = 1$.

Corollary 8.1. (Protter) (Urbilleder) *Lad X være en stokastisk variabel er $X^{-1}(A) := \{\omega \in \Omega | X(\omega) \in A\} := \{X \in A\}$. Hertil gælde specielt for eksempelvis $A = (-\infty, x]$ og $A = (x, \infty)$ følgende urbilleder $X^{-1}(A) = \{X \leqslant x\}$ og $X^{-1}(A) = \{X > x\}$.*

Definition. (Billedmål) *For en målbar funktion $X : (\Omega, \mathcal{A}, \mu) \to (F, \mathcal{F})$, hvor μ er et mål på $\mathcal{A}$ defineres $\mu^X(A) = \mu(X^{-1}(A))$ som billedmålet af A.*

Theorem 8.5. (Protter) (Billedmål for stokastiske variable) *Lad $X : (\Omega, \mathcal{A}, P) \to (F, \mathcal{F})$ være en stokastisk variabel, så gælder for alle $A \in \mathcal{F}$ at $P^X(A) = X(P)(A) = P(X^{-1}(A)) = P(X \in A)$, hvor P^X kaldes fordelingen af X (distribution of X). **Specielt er P^X et sandsynlighedsmål.**.*

Bemærkning. Fordelingsfunktionen $F_X : \mathbb{R} \to [0, 1]$ for X er givet ved $F_X(x) = P^X((-\infty, x]) = P(X \in (-\infty, x]) = P(X \leqslant x)$.

Definition. (Næsten sikker) *For en stokasitsk variabel $X : (\Omega, \mathcal{A}, P) \to (F, \mathcal{F})$ og en delmængde $A \in \mathcal{F}$ siges at*

$$X \in A \quad P\text{-n.s} \quad \Leftrightarrow \quad X \in A \quad P\text{-a.s} \quad \Leftrightarrow \quad X \in A \quad P\text{-a.e}$$

gælde hvis $P(X \in A) = 1$ dvs. ækvivalent $P(X \in A^c) = P(X \in F \backslash A) = 0$.

Ovenstående udtales n.s (næsten sikker), a.s (almost surely) og a.e (almost everywhere).

11.6 Integration with Respect to a Probability Measure

Definition 9.1. (Forventet værdi) *Lad $X : (\Omega, \mathcal{A}, P) \to (F, \mathcal{F})$ være en stokastisk variabel. Da defineres forventningen af X som følgende integrale, når dette er veldefineret.*

$$E(X) = E\{X\} = \int X dP = \int_\Omega X dP = \int_\Omega X(\omega) dP(\omega) = \int_\Omega X(\omega) P(d\omega).$$

Bemærkning. Integralet ovenfor er veldefineret hvis 1) $X \geqslant 0$ P-n.s dvs $P(X \geqslant 0) = 1$ eller, hvis 2) *(definition 9.2)* $X = X^+ - X^-$ er $E(X) = E(X^+) - E(X^-)$, hvis blot $E(X^+), E(X^-) < +\infty$ (kun en nødvendig).

Eksempler. Hvis Ω er tællelig, er P en simpel funktion, givet ved singelton mængder. Derved bestemmes $E(X)$ ved

$$E(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\})$$

Theorem 9.1. (Protter) (Sætninger fra An2) *Husk $\mathcal{L}^p$ defineres som mængden af $\mathcal{A}/\mathcal{B}$ målelige funktioner f, hvor $|f|^p$ er integrabel. Det vil sige i konteksten af forventet værdi arbejder vi med mængden*

$$\mathcal{L}^p = \{X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}) \mid E|X|^p < +\infty\}.$$

a. *$\mathcal{L}^p$ er et lineært vektorrum dvs. for $X \in \mathcal{L}^p \Rightarrow aX \in \mathcal{L}^p$, $a \in \mathbb{R}$ og for $X, Y \in \mathcal{L}^p \Rightarrow X + Y \in \mathcal{L}^p$. Desuden for $0 \leqslant X \leqslant Y$ og $Y \in \mathcal{L}^1$, så er $X \in \mathcal{L}^1$ og $E(X) \leqslant E(Y)$.*
b. *For $X \in \mathcal{L}^p$ er $E(X) \leqslant E|X|$.*
c. *Hvis $X = Y$ $P-n.s$ og $X \in \mathcal{L}^p$, så er $Y \in \mathcal{L}^p$ og $E(X) = E(Y)$ samt $E|X - Y|^p = 0$.*
d. **(Monotom konvergens teorem)** *Hvis $X_n \uparrow X$ og X_n gælder $\lim_{n\to\infty} E(X_n) = E(X)$.*
e. **(Fatou's lemma)** *Hvis $X_n \geqslant Y$ $P-n.s.$ ($Y \in \mathcal{L}^p$) eller $X_n \geqslant 0$ $P-n.s.$ for alle n, så er $E(\liminf_{n\to\infty} X_n) \leqslant \liminf_{n\to\infty} E(X_n)$.*
f. **(Lebesgue's domineret konvergens teorem)** *Hvis $X_n \uparrow X$ $P-n.s.$ og hvis $|X_n| \leqslant Y \in \mathcal{L}^1$ $P-n.s.$, så er $X_n, X \in \mathcal{L}^1$ og $E(X_n) \to E(X)$.*

Theorem 9.2. (Protter) *Lad $X_n : (\Omega, \mathcal{A}, P) \to (F, \mathcal{F})$ være en følge af stokastiske variable.*

a. *Hvis $X_n \geqslant 0$, så gælder $E\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} E(X_n)$ (begge enten uendelige eller endelige)*
b. *Hvis $\sum_{n=1}^{\infty} E|X_n| < +\infty$, så konvergerer $\sum_{n=1}^{\infty} X_n$ $P-n.s.$ og er integrabel. Desuden holder ovenstående lighed.*

Theorem 9.3. (Protter) (Cauchy-Schwarz ulighed) *Lad $L^p = \{X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}) \mid E(X^p) < +\infty\}$ dvs. mængden af p-potens integrable stokastiske variable.*

a. *Hvis $X, Y \in L^2$, så er $XY \in L^1$ og følgende ulighed gælder*

$$|E(XY)| \leqslant \sqrt{E(X^2)E(Y^2)}.$$

b. *Der gælder $L^2 \subset L^1$ og hvis $X \in L^2$, så $E(X)^2 \leqslant E(X^2)$.*
c. *Rummmet L^2 er et lineært vektor rum.*

Theorem 9.4. (Protter) (Chebyshev's/Markov's/Bienaymé-Chebyshev's ulighed) *For en stokastisk variabel X gælder*

$$P(|X| \geqslant a) \leqslant \frac{E(X^2)}{a^2}, \frac{E|X|}{a} \qquad P\{|X - E\{X\}| \geqslant a\} \leqslant \frac{\sigma^2 E\{X^2\}}{a^2}$$

Theorem 9.5. (Protter) (Forventnings reglen) *Lad $X \in \mathbb{R}$ være en stokastisk variabel, så $X : (\Omega, \mathcal{A}, P) \to (E, \mathcal{E})$ og med fordeling P^X. Lad $h : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B})$ være en målelig funktion.*

a. *Der gælder $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, P) \iff h \in \mathcal{L}^1(E, \mathcal{E}, P^X)$*
b. *Hvis (a) er opfyldt eller $h \geqslant 0$, så er*

$$E(H(X)) = \int h(x) P^X(dx) = \int h(x) dP^X(x)$$

11.7 Independent Random Variables

Definition 10.1. (Protter) (Uafhængige Stokastiske Variable)

a. *Brolægninger $(\mathcal{A}_i)_{i\in I} \subseteq \mathcal{A}$ kaldes uafhængige hvis for alle endelige $J \subseteq I$ og alle $A_i \in \mathcal{A}_i$ er*

$$P(\cap_{i\in J} A_i) = \prod_{i\in J} P(A_i)$$

dvs. for alle brolægning er alle hændelser uafhængige.
b. *tokastiske Variable $(X_i)_{i\in I}$, hvor $X_i : (\Omega, \mathcal{A}) \to (E_i, \mathcal{E}_i)$, kaldes uafhængige, hvis brolægningerne i familien givet ved $\sigma(X_i^{-1}(\mathcal{E}_i)) \subseteq \mathcal{A}$ er uafhængige.*

Theorem 10.1. (Protter) (Ækvivalensudsagn) *For to stokastiske variable $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ og $Y : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ er følgende udsagn ækvivalente.*

o. *X og Y er uafhængige.*
p. *$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for alle $A \in \mathcal{E}$ og $B \in \mathcal{F}$.*
q. *$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for alle $A \in \mathcal{C}$ og $B \in \mathcal{D}$. Hvor $\mathcal{C}$ og $\mathcal{D}$ er fællesmængdestabile mængdesystemer (brolægning) der frembringer hhv. $\mathcal{E}$ og $\mathcal{F}$.*
r. *$f(X)$ og $g(Y)$ er uafhængige for alle par (f, g) målbare funktioner.*
s. *$E(f(X)g(Y)) = E(f(X))E(g(Y))$ for alle par (f, g) begrænsede og målbar eller positive og målbare funktioner.*
t. *Hvis E og F er metriske rum med respektive Borel σ-algebraer $\mathcal{E}$ og $\mathcal{F}$. $E(f(X)g(Y)) = E(f(X))E(g(Y))$ for alle par (f, g) begrænsede og kontinuere funktioner.*

Notation. Ofte ønskes at betragte funktioner af flere variable, hvor domænerummet ønskes konstrueret fra to rum hvorpå x og y lever hhv. $(E, \mathcal{E}, P)$ og $(F, \mathcal{F}, Q)$. Vi kan da konstruere et målrum givet ved $(E \times F, \sigma(\mathcal{E} \times \mathcal{F}))$, hvor vi lader $\mathcal{E} \otimes \mathcal{F} = \sigma(\mathcal{E} \times \mathcal{F})$. Tilsvarende kan vi konstruere produktmålet $P \otimes Q(A) = P(A)Q(A)$ på $(E \times F, \mathcal{E} \otimes \mathcal{F}, P \otimes Q)$.

Theorem 10.2. (Protter) *Lad $f : (E \times F, \mathcal{E} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{R})$ være målbar. For hvert $x \in E$ og $y \in F$, er de respektive "sektionerne" $y \to f(x, y)$ og $x \to f(x, y)$ henholdsvis $\mathcal{F}$- og $\mathcal{E}$ målbare funktioner.*

Notation. Det kan ønskes at betragte mål med faste x, y og hvordan disse opfører sig. Teoremet fortæller, at man kan vælge tilfældige $x \in E$ og $y \in F$ således, at man kan betragte henholdsvis $\mathcal{F}$- og $\mathcal{E}$ målbare funktioner i én variable (enten x eller y) for sig selv.

Theorem 10.3. (Protter) (Tonelli-Fubini) *Lad $(E, \mathcal{E}, P)$ og $(F, \mathcal{F}, Q)$ være sandsynlighedsrum.*

a. *Lad $P \otimes Q(A \times B) = P(A)Q(B)$. Dette er et unikt ssh mål, som udvider til sandsynlighedsrummet $(E \times F, \mathcal{E} \otimes \mathcal{F}, P \otimes Q)$.*
b. *Lad $f : (E \times F, \mathcal{E} \otimes \mathcal{F}, P \otimes Q) \to (\mathbb{R}, \mathcal{R})$ være en målbar, positiv eller integrabel mht. $P \otimes Q$. Da er $x \mapsto \int f(x, y)Q(dy)$ en $\mathcal{E}$-målelig funktion og $y \mapsto \int f(x, y)P(dy)$ en $\mathcal{F}$-målelig funktion. Specielt er*

$$\int f dP \otimes Q = \int \left\{ \int f(x, y)Q(dy) \right\} P(dx) = \int \left\{ \int f(x, y)P(dx) \right\} Q(dy)$$

11.8 Probability Distributions on $\mathbb{R}$

Dette kapitel undersøger egenskaberne ved sandsynlighedsmål P på $(\mathbb{R}, \mathcal{B})$, hvor fordelingsfunktionen er karakteriseret ved $F(x) = P((-\infty, x])$.

Definition 11.1. (Protter) *Lebesguemålet er mængdefunktion $m : \mathcal{B} \to [0, \infty 0]$, der opfylder:*

i. *For $A_1, A_2, A_3, ... \in \mathcal{B}$ parvist disjunkte mængder gælder $m(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} m(A_i)$,*
ii. *hvis $a < b$ og $a, b \in \mathbb{R}$, så $m((a, b]) = b - a$.*

Theorem 11.1. (Protter) *Lebesguemålet er unikt.*

Theorem 11.2. (Protter) *Lebesguemålet eksisterer.*

Definition 11.2. (Protter) *Tætheden af et sandsynlighedsmål P på $(\mathbb{R}, \mathcal{B})$ er en positiv Borel målelig funktion f der opfylder for alle $x \in \mathbb{R}$:*

$$F(x) = P((-\infty, x]) = \int_{-\infty}^{x} f(y)dy = \int f(y)1_{(-\infty, x]}(y)dm(y)$$

I tilfældet $P = P^X$ (husk thm 8.5. $P^X(A) = P(X \in A)$), dvs P er fordelingsmålet af en s.v. X, så siger vi at f er tætheden af X.

Theorem 11.3. (Protter) *Lad $f \in \mathcal{M}_{\mathbb{R}}^{+}$. Da gælder ($f$ karakteriseret fuldkomment tætheden for et sandsynlighedsmål P på $(\mathbb{R}, \mathcal{B})$) $\iff$ ($\int f dm(x) = 1$). Desuden gælder hvis f' opfylder $m(f \neq f') = 0$ ($f = f'$ m-n.o.), så er f' og en tæthed for samme sandsynlighedsmål. Omvendt bestemmer et sandsynlighedsmål også den tæthed, når denne eksisterer.*

Remark 11.1. F er differentiabel m-n.o. uafhængig af f og med $f = 0$ ellers.

Corollary 11.1. (Protter) (Forventningsreglen) *Lad X være en $\mathbb{R}$-værdi stokastisk variabel med tæthed f. Lad $g \in \mathcal{M}_{\mathbb{R}}$. Så er g integrabel mht. P^X, hvis og kun hvis fg er integrabel mht. m. Hvis da er*

$$E[g(X)] = \int g(x)P^X(dx) = \int g(x)f(x)dm(x)$$

Theorem 11.4. (Protter) *Lad X have tæthed f_X og lad $g \in \mathcal{M}_{\mathbb{R}}$. Lad $Y = g(X)$. Så er*

$$F_Y(y) = P(Y \leqslant y) = P(G(X) \leqslant y) = \int_{A_y} f_X(u)dm(u),$$

hvor $A_y = \{u : g(u) \leqslant y\}$.

Corollary 11.2. (Protter) *Lad X have en kontinuer tæthed f_X. Lad $g : \mathbb{R} \to \mathbb{R}$ være C^1 og strengt monotom. Lad $h(y) = g^{-1}(y)$ være g inverse også C^1. Så har $Y = g(X)$ tæthed*

$$f_Y(y) = f_X(h(y))|h'(y)|.$$

Corollary 11.3. (Protter) *Lad X have en kontinuer tæthed f_X. Lad $g : \mathbb{R} \to \mathbb{R}$ være stykkevis C^1 og strengt monotom på intervallerne $I_1, I_2, ..., I_n$ med $\cup_{i\in I} I_i = \mathbb{R}$ (g skal kun være C^1 og strengt monotom på det indre af I_i). Lad $h_i(y) = g^{-1}(y)$ for $g : I_i \to \mathbb{R}$. Så har $Y = g(X)$ tæthed*

$$f_Y(y) = \sum_{i=1}^{n} f_X(h_i(y))|h_i'(y)|1_{g(I_i)}(y).$$

11.9 Probability Distributions on $\mathbb{R}^n$

Dette kapitel undersøger fordelinger for stokatiske variable på $(\mathbb{R}^n, \mathcal{B}^n)$ for $n = 2, 3, ...$ Særligt vil vi have interesse for særtilfældet $n = 2$. Mange sætninger i dette kapitler har en analog i kapitel 11.

Definition 12.1. (Protter) *Lebesgue målet på $(\mathbb{R}^n, \mathcal{B}^n)$ er defineret for det kartesiske produkt $A_1 \times A_2 \times ... \times A_n$ ved $m_n(\prod_{i=1}^n A_i) = \prod_{i=1}^n m(A_i)$. Dermed også $m_n(\prod_{i=1}^n (a_i, b_i]) = \prod_{i=1}^n m((a_i, b_i])$.*

Definition 12.2. (Protter) *Et sandsynlighedsmål P på $(\mathbb{R}^n, \mathcal{B}^n)$ har tæthed f hvis f er en ikke negativ Borelmålelig funktion på $\mathbb{R}^n$, der opfylder*

$$P(A) = \int_A f(x)dm_n(x) = \int f(x_1, x_2, ..., x_n)1_A(x_1, x_2, ..., x_n)dm_n(x_1, x_2, ..., x_n), \quad \forall A \in \mathcal{B}^n.$$

Theorem 12.1. (Protter) *En $f \in \mathcal{M}^+_{R^n}$ er en tæthed for et sandsynlighedsmål P på $(\mathbb{R}^n, \mathcal{B}^n)$ hvis og kun hvis $\int f(x)dm_n(x) = 1$. I det tilfælde karakterer f fuldkomment målet P og for alle $f' \in \mathcal{M}^+_{R^n}$ med $m_n(f \neq f') = 0$ er denne også tæthed for målet P. Omvendt bestemmer målet P også en tæthed f op til enhver ikke Lebesgue nulmængde.*

Theorem 12.2. (Protter) *Antag $X = (Y, Z)$ har tæthed $f = f_X = f_{Y,Z}$ på $\mathbb{R}^2$. Så gælder*

a. *Begge Y og Z har tætheder på $(\mathbb{R}, \mathcal{B})$ givet ved*

$$f_Y(y) = \int_{\mathbb{R}} f(y, z)dm(z), \qquad f_Z(z) = \int_{\mathbb{R}} f(y, z)dm(y)$$

b. *Y og Z er uafhængige hvis og kun hvis $f(y, z) = f_Y(y)f_Z(z)$ m_2-n.s. dvs. $m_2(f(y, z) \neq f_Y(y)f_Z(z) = 0$.*
c. *Følgende bestemmer en fjerde tæthed på $\mathbb{R}$ for alle $y \in \mathbb{R}$ sådan at $f_Y(y) \neq 0$: $f_{Y=y}(z) = \frac{f(y,z)}{f_Y(y)}$.*

Definition 12.3. (Protter) *Lad X, Y være to reelle stokastiske variable begge med endelig varians. Covariansen af X, Y er defineret ved $Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$. Varians er dermed også givet ved $Cov(X, X) = Var(X) = \sigma^2(X)$.*

Theorem 12.3. (Protter) *Hvis X og Y er uafhængige så er $Cov(X, Y) = 0$.*

Definition 12.4. (Protter) *Lad X og Y være stokastiske variable begge med endelig varians. Korrelationskoefficienten af X og Y er tallet $\rho_{X,Y} = Cov(X, Y)/(\sigma(X)\sigma(Y))$.*

Definition 12.5. (Protter) *Lad $X = (X_1, ..., X_n)$ være en $\mathbb{R}^n$ stokastisk variabel. Covariansmatricen for X er en $n \times n$ matrice med indgange $c_{ij} = Cov(X_i, X_j)$.*

Theorem 12.4. (Protter) *En covarians matrice er positiv semidefinit dvs. den er symmetrisk ($c_{ij} = c_{ji}$) og $\sum a_i a_j c_{ij} > 0$ for alle $\mathbf{a} \in \mathbb{R}^n$.*

Theorem 12.5. (Protter) *Lad X være en $\mathbb{R}^n$ stokastisk variabel med covarians matrice C. Lad A være en $m \times n$ matrice og set $Y = AX$. Så er Y en $\mathbb{R}^m$ stokastisk variabel med covarians matrice $C' = ACA^*$, hvor A^* er den transponerede matrice til A.*

Theorem 12.6. (Protter) (Jacobi's transformations formel) *Lad $G \subseteq \mathbb{R}^n$ være åben og lad $g : G \to \mathbb{R}^n$ være kontinuer og differentiabel. Antag g er injektiv på G og $J_g(x) \neq 0$ for alle $x \in G$. For en funktion $f \in \mathcal{M}$ med $f1_{g(G)}$ positiv eller integrabel mht. Lebesguemålet gælder: $\int_{g(G)} f(y)dm_n(y) = \int_G f(g(x))|\det(J_g(x))|dm_n(x)$, hvor $g(G)$ er mængden $\{y \in \mathbb{R}^n : \exists x \in G, g(x) = y\}$.*

Theorem 12.7. (Protter) *Lad $X = (X_1, ..., X_n)$ have simultan tæthed f og lad $g : \mathbb{R}^n \to \mathbb{R}^n$ være en kontinuer, diffenrentiabel, injektiv funktion med $J_g(x) \neq 0$. Så har $Y = g(X)$ tæthed*

$$f_Y(y) = f_X(g^{-1}(y))|\det J_{g^{-1}}(y)|1_{g(\mathbb{R}^n)}(y)$$

Corollary 12.1. (Protter) *Lad $S \in \mathcal{B}^n$ være inddelt af et endeligt indeks $I = \{0, 1, ..., m\}$, så $\cup_{i=0}^m S_i = S$, S_i parvist disjunkte, og med $m_n(S_0) = 0$ og funktionen $g : S_i \to \mathbb{R}^n$ være kontinuer, differentiabel, injektiv og $J_{g_i^{-1}}(x) \neq 0$ for alle $i \in I$. Lad X være giver som i teorem 12.7. Da har den stokastiske variabel $Y = g(X)$ tæthed*

$$f_Y(y) = \sum_{i=1}^m f_X(g_i^{-1}(y))|\det J_{g_i^{-1}}(y)|1_{g_i(S_i)}(y).$$

11.10 Equivalent Probability Measures

11.10.1 The Radon-Nikodym Theorem

Definition A.50. (Bjork) *Consider a measurable space $(X, \mathcal{F})$ on which there are defined two seperate measures μ and ν:*

- *If, for all $A \in \mathcal{F}$, it holds that*
$$\mu(A) = 0 \;\Rightarrow\; \nu(A) = 0, \tag{A.7}$$
t_hen ν is said to be **absolutely continuous** with respect to μ on $\mathcal{F}$ and we write this as $\nu << \mu$._
- *If we have both $\mu << \nu$ and $\nu << \mu$, then μ and ν said to be **equivalent** and we write $\mu \sim \nu$.*
- *If there exists two events, A and B such that:*
  - $A \cup B = X$,
  - $A \cap B = \varnothing$,
  - $\mu(B) = 0$, and $\nu(A) = 0$,

  *then ν and μ are said to be mutually **singular**, and we write $\mu \perp \nu$.*

Theorem A.52. (Bjork) (The Radon-Nikodym Theorem) *Consider the measure space $(X, \mathcal{F}, \mu)$, where we assume that μ is finite, i.e. that $\mu(X) < \infty$. Let ν be a measure on $(X, \mathcal{F})$ such that $\nu << \mu$ on $\mathcal{F}$. Then there exists a non-negative function $f : X \to \mathbb{R}$ such that:*

$$f \text{ is } \mathcal{F}\text{-measurable} \tag{A.9}$$

$$\int_X f(x)\, d\mu(x) < \infty, \tag{A.10}$$

$$\nu(A) = \int_A f(x)\, d\mu(x), \text{ for all } A \in \mathcal{F}. \tag{A.11}$$

*The function f is called the **Radon-Nikodym derivative** of ν w.r.t. μ. It is uniquely determined μ-a.e. and we write*

$$f(x) = \frac{d\nu(x)}{d\mu(x)}, \tag{A.12}$$

or alternatively

$$d\nu(x) = f(x)\ d\mu(x). \tag{A.13}$$

11.10.2 Equivalent Probability Measures

Lemma B.38. (Bjork) *For two probability measures P and Q, the relation $P \sim Q$ on $\mathcal{F}$ holds if and only if $P(A) = 1$ if and only if $Q(A) = 1$ for all $A \in \mathcal{F}$.*

Proposition B.39. (Bjork) *Assume that $Q << P$ on $\mathcal{F}$ and that $\mathcal{G} \subseteq \mathcal{F}$. Then the Radon-Nikodym derivatives $L^{\mathcal{F}}$ and $L^{\mathcal{G}}$ are related by*

$$L^{\mathcal{G}} = E^P[L^{\mathcal{F}} \mid \mathcal{G}]. \tag{B.17}$$

Proposition B.41. (Bjork) (Bayes' Theorem) *Assume that X is a random variable on $(\Omega, \mathcal{F}, P)$, and let Q be another probability measure on $(\Omega, \mathcal{F})$ the Radon-Nikodym derivative*

$$L = \frac{dQ}{dP}$$

on $\mathcal{F}$. Assume that $X \in L^1(\Omega, \mathcal{F}, Q)$ and $\mathcal{G}$ is a sigma-algebra with $\mathcal{G} \subseteq \mathcal{F}$. Then

$$E^Q[X \mid \mathcal{G}] = \frac{E^P[L \cdot X \mid \mathcal{G}]}{E^P[L \mid \mathcal{G}]},\ Q - a.s. \tag{B.18}$$

11.10.3 Likelihood processes

Proposition C.12. (Bjork) *Consider a filtered probability space $(\Omega, \mathcal{F}, P, \mathcal{F}_t)$ on a compact interval $[0, T]$. Suppose L_T is some non-negative integrable random variable in $\mathcal{F}_T$. We can then define a new measure Q on $\mathcal{F}_T$ by setting*

$$dQ = L_T\ dP$$

on $\mathcal{F}_T$ and if $E^P[L_T] = 1$ the measure Q will also be a probability measure. The likelihood process L, defined by

$$L_t = \frac{dQ}{dP},\ on\ \mathcal{F}_t, \tag{C.8}$$

is a $(P, \mathcal{F}_t)$-martingale.

Proposition C.13. (Bjork) *A process M is a Q-martingale if and only if the process $L \cdot M$ is a P-martingale.*

Chapter 12

Random Variables

12.1 Introduction

Definition 1.1. (Hansen) *A **real-valued random variable** X on a probability space $(\Omega, \mathbb{F}, P)$ is a measurable map $X : (\Omega, \mathbb{F}) \to (\mathbb{R}, \mathbb{B})$.*

We never specify the background space $(\Omega, \mathbb{F}, P)$ however we always assume X is $\mathbb{F} - \mathbb{B}$ measurable. This assumption implies $(X \in A) \in \mathbb{F}$ for every $A \in \mathbb{B}$. We may want to show measurability for constructed variables and so it surfises to show measurability for generaters for $\mathbb{B}$ such as checking $(X \leqslant a) \in \mathbb{F}$ for every $a \in \mathbb{R}$.

Definition 1.2. (Hansen) *The **distribution** of a real-valued random variable X, defined on a probability space $(\Omega, \mathbb{F}, P)$, is the collection of probability values*

$$P(X \in A) \quad \text{for } A \in \mathbb{B}. \tag{1.3}$$

In other words: the distribution of X is the image measure $X(P)$ on $(\mathbb{R}, \mathbb{B})$.

Lemma 1.3. (Hansen) *Let X and X' be two real-valued random variables on a probability space $(\Omega, \mathbb{F}, P)$. If*

$$P(X = X') = 1$$

then X and X' has the same distribution.

An often used way of summarizing the distribution is through the **distribution function** $F(x) = P(X \leqslant x)$ for some $x \in \mathbb{R}$.

Definition 1.4. (Hansen) *A real-valued random variable X has a **discrete** distribution if there is a countable set $S \subset \mathbb{R}$ such that $P(X \in S) = 1$.*

Usually S is one of $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ or a subset of these. We may in the discrete case define the distribution by the point probabilities $P(X = x) = p(x)$ for $x \in S$.

Definition 1.5. (Hansen) *A real-valued random variable X has a distribution with **density** $f : \mathbb{R} \to [0, \infty)$ if*

$$P(X \in A) = \int_A f(x)dx \quad \text{for } A \in \mathbb{B}. \tag{1.5}$$

If this is the case we will write $X(P) = f \cdot m$ or $X \sim f \cdot m$.

Definition 1.6. (Hansen) *A real-valued random variable X defined on a probability space $(\Omega, \mathbb{F}, P)$ is said to have p'th moment for som $p > 0$ if*

$$E|X|^p < \infty \tag{1.12}$$

The collection of all variables that satisfies (1.12) is denoted by $\mathcal{L}^p(\Omega, \mathbb{F}, P)$.

Recall the definition of the **expectation** of X by

$$E\,X = \int X dP \in R \cup \{-\infty, +\infty\}. \tag{1.11}$$

We recall that for any measurable function $f : \mathbb{R} \to \mathbb{R}$ that are continuous on a set $A \in \mathbb{B}$ such that $P(X \in A) = 1$ we may change variable simply by computing

$$E\,f(X) = \int f \circ X dP = \int f(x) dX(P)(x).$$

Lemma 1.7. (Hansen) *(Markov's inequality) Let X be a non-negative random variable. For any $c > 0$ it holds that*

$$P(X \geqslant c) \leqslant \frac{E\,X}{c} \left(\leqslant \frac{E\,X^n}{c^n} \text{ or } \leqslant \frac{E\,(\varphi(X))}{\varphi(c)} \right). \tag{1.14}$$

for some φ non-negative monotome increasing function.

Some other versions of Markov's inequality can be found in the form of **Chebyshev's inequality**, **Chebyshev-Cantelli's inequality** or **Jensen's inequality** repectively: Let X be a real-valued random variable in $\mathcal{L}^2(\Omega, \mathbb{F}, P)$ it holds for any $\varepsilon > 0$.

$$P\,(|X - E\,X| \geqslant \varepsilon) \leqslant \frac{V\,X}{\varepsilon^2} \tag{1.15}$$

$$P\,(X - E\,X \geqslant \varepsilon) \leqslant \frac{V\,X}{V\,X + \varepsilon^2} \qquad \text{(prob: 1.13(c))}$$

$$\varphi\,(E\;X) \leqslant E\,(\varphi(X))$$

for some convex function φ.

Lemma 1.8. (Hansen) *Let X be a non-negative random variable. It holds that*

$$E\,X = \int_0^\infty P(X > t) dt. \tag{1.16}$$

where the integral on the right hand side is with respect to Lebesgue measure.

Definition 1.9. (Hansen) *The* ***joint distribution*** *of real-valued random variables $X_1, ..., X_k$, defined on a probability space $(\Omega, \mathbb{F}, P)$, is the collection of probability values*

$$P(\mathbf{X} \in A) \qquad \text{for } A \in \mathbb{B}_k. \tag{1.21}$$

In other words: the joint distribution of $X_1, ..., X_k$ (or simply: the distribution of $\mathbf{X}$) is the image measure $\mathbf{X}(P)$ on $\left(\mathbb{R}^k, \mathbb{B}_k\right)$.

Definition 1.11. (Hansen) *Real-valued random variables $X_1, ..., X_k$, defined on a probability space $(\Omega, \mathbb{F}, P)$, are* ***jointly independent*** *if*

$$P\,(X_1 \in A_1, ..., X_k \in A_k) = \prod_{i=1}^{k} P(X_i \in A_i) \qquad \text{for } A_1, ..., A_k \in \mathbb{B}. \tag{1.23}$$

In other words: the variables are independent if the joint distribution $\mathbf{X}(P)$ equals the product measure $X_1(P) \otimes ... \otimes X_k(P)$.

Theorem 1.12. (Hansen) *Let $X_1, ..., X_k$ be real-valued random variables defined on a probability space $(\Omega, \mathbb{F}, P)$. If the variables are independent and if $E|X_i| < \infty$ for $i = 1, ..., k$, then the product $X_1 \cdot ... \cdot X_k$ has first moment and*

$$E\,(X_1 \cdot ... \cdot X_k) = \prod_{i=1}^{k} E\,X_i \tag{1.24}$$

The equality only holds for two independent variables. However the **Cauchy-Schwarz inequality** which closely resembles (1.24) holds wether or not X or Y are independent:

$$(E|XY|)^2 \leqslant E\,X^2\,E\,Y^2 \text{ or } E|XY| \leqslant \sqrt{E\ X^2}\sqrt{E\ Y^2}. \tag{1.25}$$

Furthermore the theorem give rise to a measure for dependence i.e. the **covariance** between two variables X and Y

$$\mathrm{Cov}(X,Y) = E\left((X - E\,X)(Y - E\,Y)\right) = E(XY) - (E\,X)(E\,Y) \tag{1.26}$$

with $\mathrm{Cov}(X,Y) \neq 0$ if and only if X and Y are dependent. With $\mathrm{Cov}(X,Y) = 0$ the test is inconlusive. However independence implies $\mathrm{Cov}(X,Y) = 0$.

12.2 Conditional expectation

The theory of conditional expectation is well-known from courses on the bachelor. Because of this we will only summarise the most important results.

We consider a background space $(\Omega, \mathcal{F}, P)$ and a sub-sigma algebra $\mathcal{G} \subseteq \mathcal{F}$. We assume that some stochastic variable is $\mathcal{F}$-measurable, that is the mapping $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathbb{B}, m)$ is $\mathcal{F} - \mathbb{B}$-measurable i.e. $\forall B \in \mathbb{B} : \{X \in B\} \in \mathcal{F}$. For some random variable Z defined on the subspace $(\Omega, \mathcal{G}, P)$, we say that Z is the conditional expectation of X given $\mathcal{G}$ if

$$\forall G \in \mathcal{G} : \int_G Z(\omega)\ dP(\omega) = \int_G X(\omega)\ dP(\omega).$$

This fact is summed up in the definition below.

Definition B.27. (Bjork) (Conditional expectation) *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random variable in $L^1(\Omega, \mathcal{F}, P)$ ($|X|$ is integrable). Let furthermore $\mathcal{G}$ be a sigma-algebra such that $\mathcal{G} \subseteq \mathcal{F}$. If Z is a random variable with the properties that:*

i. *Z is $\mathcal{G}$-measurable.*
ii. *For every $G \in \mathcal{G}$ it holds that*
$$\int_G Z(\omega)\ dP(\omega) = \int_G X(\omega)\ dP(\omega). \tag{B.5}$$

*Then we say that Z is the **conditional expectation of X given the sigma-algebra $\mathcal{G}$**. In that case we denote Z by the symbol $E[X \mid \mathcal{G}]$.*

We see that from the above it always holds that X satisfies (ii). It does not, however, always hold that X is $\mathcal{G}$-measurable. Given this fact it is not trivial that a random variable $E[X \mid \mathcal{G}]$ even exists. This nontriviality is fortunatly resolved by the Radon-Nikodym theorem.

Theorem B.28. (Bjork) (Existance and uniqueness of Conditional expectation) *Let $(\Omega, \mathcal{F}, P)$, X and $\mathcal{G}$ be given as in the definition above. Then the following holds:*

- *There will always exist a random variable Z satisfying conditions (i)-(ii) above.*
- *The variable Z is unique, i.e. if both Y and Z satisfy (i)-(ii) then $Y = Z$ P-a.s.*

This result ensures that we may condition on any sigma-algebra for instance $\mathcal{G} = \sigma(Y)$ in that case we (pure notation) write

$$E[X \mid \sigma(Y)] = E[X \mid Y], \qquad \sigma(Y) = \sigma\left(\{Y \in A,\ A \in \mathbb{B}\}\right).$$

In the above $\sigma(Y)$ is simply the smallest sigma-algebra containing all the pre-images of Y, that is the smallest sigma-algebra making Y measurable! Giving this foundation there are a few properties conditional expectation have which is rather useful (for instance the tower property).

Below we assume: Let $(\Omega, \mathcal{F}, P)$ be a probability space and X, Y be random variables in $L^1(\Omega, \mathcal{F}, P)$.

Proposition B.29. (Monotinicity/Linearity of Conditional expectation) *The following holds:*

$$X \leq Y \implies E[X \mid \mathcal{G}] \leq E[Y \mid \mathcal{G}], \ P-\text{a.s.} \tag{B.6}$$

$$E[\alpha X + \beta Y \mid \mathcal{G}] = \alpha E[X \mid \mathcal{G}] + \beta E[Y \mid \mathcal{G}], \ P-\text{a.s.}, \forall \alpha, \beta \in \mathbb{R}. \tag{B.7}$$

Proposition B.30. (Bjork) (Tower property) *Assume that it holds that $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$. Then the following hold:*

$$E[E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[X \mid \mathcal{H}], \tag{B.8}$$

$$E[X] = E[E[X \mid \mathcal{G}]]. \tag{B.9}$$

Proposition B.31. (Bjork) *Assume X is $\mathcal{G}$ and that both X, Y and XY are in L^1 (only assuming Y is $\mathcal{F}$-measurable), then*

$$E[XY \mid \mathcal{G}] = XE[Y \mid \mathcal{G}], \ P-\text{a.s.} \tag{B.11}$$

$$E[X \mid \mathcal{G}] = X \ P-\text{a.s.} \tag{B.12}$$

Proposition B.32. (Bjork) (Jensen inequality) *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex (measurable) function and assume $f(X)$ is in L^1. Then*

$$f(E[X|\mathcal{G}]) \leqslant E[f(X)|\mathcal{G}], \qquad P-\text{a.s.}$$

Proposition B.37. (Bjork) *Let $(\Omega, \mathcal{F}, P)$ be a given probability space, let $\mathcal{G}$ be a sub-sigma-algebra of $\mathcal{F}$, and let X be a square integrable random variable. Consider the problem of minimizing*

$$E\left[(X-Z)^2\right]$$

where Z is allowed to vary over the class of all square integrable $\mathcal{G}$ measurable random variables. The optimal solution $\hat{Z}$ is then given by.

$$\hat{Z} = E[X|\mathcal{G}].$$

Proof.

Let $X \in L^2(\Omega, \mathcal{F}, P)$ be a random variable. Now consider an arbitrary $Z \in L^2(\Omega, \mathcal{G}, P)$. Recall that $\mathcal{G} \subset \mathcal{F}$ and so X is also in $Z \in L^2(\Omega, \mathcal{G}, P)$, as it is bothe square integrable and $\mathcal{G}$-measurable. Then

$$E\left[Z \cdot (X - E[X|\mathcal{G}])\right] = E\left[Z \cdot X\right] - E\left[Z \cdot E[X|\mathcal{G}]\right].$$

Then by using the law of total expectation and secondly that Z is $\mathcal{G}$-measurable we have that

$$E\left[Z \cdot X\right] = E\left[E[Z \cdot X|\mathcal{G}]\right] = E\left[Z \cdot E[X|\mathcal{G}]\right].$$

Combining the two equations gives the desired result. Obviously, we have that

$$X - Z = X - Z + E[X|\mathcal{G}] - E[X|\mathcal{G}] = (X - E[X|\mathcal{G}]) + (E[X|\mathcal{G}] - Z).$$

Then squaring the terms gives

$$(X-Z)^2 = (X - E[X|\mathcal{G}])^2 + (E[X|\mathcal{G}] - Z)^2 + 2(X - E[X|\mathcal{G}])(E[X|\mathcal{G}] - Z)$$

Taking expectation on each side and using linearity of the expectation we have that

$$E[(X-Z)^2] = E\left[(X - E[X|\mathcal{G}])^2\right] + E\left[(E[X|\mathcal{G}] - Z)^2\right] + 2E\left[(X - E[X|\mathcal{G}])(E[X|\mathcal{G}] - Z)\right].$$

We can now use that $E[X|\mathcal{G}] - Z$ is $\mathcal{G}$-measurable with the above result on the last term.

$$E[(X-Z)^2] = E\left[(X - E[X|\mathcal{G}])^2\right] + E\left[(E[X|\mathcal{G}] - Z)^2\right].$$

Now since X is given the term $E\left[(X - E[X|\mathcal{G}])^2\right]$ is simply a constant not depending on the choice og Z. The optimal choice of Z is then $E[X|\mathcal{G}]$ since this minimizes the second term. The statement is then proved.

12.3 Independence

Definition 3.1. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space. Two events $A, B \in \mathbb{F}$ are **independent** if*

$$P(A \cap B) = P(A)P(B) \tag{3.1}$$

Definition 3.4. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{G}, \mathbb{H} \subset \mathbb{F}$ be two classes of measurable sets. We sat that $\mathbb{G}$ and $\mathbb{H}$ are independent, written $\mathbb{G} \perp\!\!\!\perp \mathbb{H}$, if*

$$P(A \cap B) = P(A)P(B) \quad \text{for all } A \in \mathbb{G}, B \in \mathbb{H}. \tag{3.2}$$

Lemma 3.5. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{G}, \mathbb{H} \subset \mathbb{F}$ be two classes of measurable sets. Let $\mathbb{G}_1 \subset \mathbb{G}$ and $\mathbb{H}_1 \subset \mathbb{H}$ be two subclasses. If $\mathbb{G} \perp\!\!\!\perp \mathbb{H}$ then it holds that $\mathbb{G}_1 \perp\!\!\!\perp \mathbb{H}_1$.*

Definition 3.6. (Hansen) *A class $\mathbb{H}$ of subsets of Ω is a **Dynkin class** if*

1. $\Omega \in \mathbb{H}$,
2. $A, B \in \mathbb{H}, A \subset B \quad \Rightarrow \quad B \backslash A \in \mathbb{H}$,
3. $A_1, A_2, ... \in \mathbb{H}, A_1 \subset A_2 \subset ... \quad \Rightarrow \quad \bigcup_{n=1}^{\infty} A_n \in \mathbb{H}$.

Lemma 3.7. (Hansen) *(Dynkin) Let $\mathbb{D} \subset \mathbb{H}_0 \subset \mathbb{H}$ be three nested classes of subsets of Ω. if*

1. $\sigma(\mathbb{D}) = \mathbb{H}$,
2. $A, B \in \mathbb{D} \quad \Rightarrow \quad A \cap B \in \mathbb{D}$
3. $\mathbb{H}_0$ *is a Dynkin class.*

then it holds that $\mathbb{H}_0 = \mathbb{H}$.

Lemma 3.8. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $A \in \mathbb{F}$ be a fixed event. The class*

$$\mathbb{H} = \{B \in \mathbb{F} \mid A \perp\!\!\!\perp B\}$$

is a Dynkin class.

Theorem 3.9. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, \mathbb{G}_2 \subset \mathbb{F}$ be two sigma-algebras. Let $\mathbb{D}_1$ and $\mathbb{D}_2$ be two classes such that $\sigma(\mathbb{D}_i) = \mathbb{G}_i$ for $i = 1, 2$. If both $\mathbb{D}_1$ and $\mathbb{D}_2$ are $\cap$-stable then it holds that*

$$\mathbb{D}_1 \perp\!\!\!\perp \mathbb{D}_2 \quad \Rightarrow \quad \mathbb{G}_1 \perp\!\!\!\perp \mathbb{G}_2.$$

Definition 3.10. (Hansen) *Two real-valued random variable X and Y on a background space $(\Omega, \mathbb{F}, P)$ are **independent**, written $X \perp\!\!\!\perp Y$, if the corresponding sigma-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.*

Definition 3.15. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, ..., \mathbb{G}_n \subset \mathbb{F}$ be finitely many classes of measurable sets. We say that $\mathbb{G}_1, ..., \mathbb{G}_n$ are **jointly independent**, written $\mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_n$, if*

$$P(A_1 \cap ... \cap A_n = \prod_{i=1}^{n} P(A_i) \quad \text{for } A_1 \in \mathbb{G}_1, ..., A_n \in \mathbb{G}_n. \tag{3.8}$$

Lemma 3.16. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, ..., \mathbb{G}_n \subset \mathbb{F}$ be finitely many classes of measurable sets. It holds that*

$$\mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_n \quad \Rightarrow \quad \mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_{n-1}$$

provided that $\Omega \in \mathbb{G}_n$.

Theorem 3.17. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, ..., \mathbb{G}_n \subset \mathbb{F}$ be sigma-algebras. Let $\mathbb{D}_1, ..., \mathbb{D}_n$ be classes such that $\sigma(\mathbb{D}_i) = \mathbb{G}_1$ for $i = 1, ..., n$. Suppose that for all lengths $k = 2, ..., n$ an all choices of indices $\leqslant j_1 < ... < j_k \leqslant n$ it holds that*

$$\mathbb{D}_{j_1} \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{D}_{j_k} \tag{3.9}$$

If all the generators $\mathbb{D}_i$ are $\cap$-stable, then it holds that $\mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_n$.

Lemma 3.18. (Hansen) *(Grouping) Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, ..., \mathbb{G}_n \subset \mathbb{F}$ be sigma-algebras. It holds that*

$$\mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_n \quad \Rightarrow \quad \mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_{n-2} \perp\!\!\!\perp \sigma(\mathbb{G}_{n-1}, \mathbb{G}_n).$$

Definition 3.19. (Hansen) *The real-valued random variables $X_1, ..., X_n$ on a background space $(\Omega, \mathbb{F}, P)$ are **jointly independent**, written $X_1 \perp\!\!\!\perp ... \perp\!\!\!\perp X_n$, if the corresponding sigma-algebras $\sigma(X_1), ..., \sigma(X_n)$ are jointly independent.*

Definition 3.20. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $(\mathbb{G}_i)_{i \in I}$ be a family of classes of measurable sets. We say that the family $(\mathbb{G}_i)_{i \in I}$ is **jointly independent** if any finite subfamily is jointly independent.*

Theorem 3.21. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, \mathbb{G}_2, ... \subset \mathbb{F}$ be sigma-algebras. Let $\mathbb{D}_1, \mathbb{D}_2, ...$ be classes such that $\sigma(\mathbb{D}_n) = \mathbb{G}_n$ for all $n \in \mathbb{N}$. Suppose that for all lengths $k \in \mathbb{N}$ and all choices of indices $1 \leqslant j_1 < ... < j_k$ it holds that*

$$\mathbb{D}_{j_1} \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{D}_{j_k}. \tag{3.14}$$

If all the generators $\mathbb{D}_n$ are $\cap$-stable, then it holds that $\mathbb{G}_1 \perp\!\!\!\perp \mathbb{G}_2 \perp\!\!\!\perp ....$

Lemma 3.22. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $\mathbb{G}_1, \mathbb{G}_2, ... \subset \mathbb{F}$ be sigma-algebras. It holds that*

$$\mathbb{G}_1 \perp\!\!\!\perp \mathbb{G}_2 \perp\!\!\!\perp ... \quad \Rightarrow \quad \mathbb{G}_1 \perp\!\!\!\perp ... \perp\!\!\!\perp \mathbb{G}_n \perp\!\!\!\perp \sigma(\mathbb{G}_{n+1}, \mathbb{G}_{n+2}, ...).$$

Definition 3.23. (Hansen) *The real-valued random variables $(X_i)_{i \in I}$ on a background space $(\Omega, \mathbb{F}, P)$ are **jointly independent** if the corresponding sigma-algebras $(\sigma(X_i))_{i \in I}$ are jointly independent.*

Definition 3.28. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space. A sigma-algebra $\mathbb{G} \subset \mathbb{F}$ satisfies a **zero-one law** if*

$$P(A) \in \{0, 1\} \quad \text{for all } A \in \mathbb{G}.$$

Theorem 3.29. (Hansen) *Let $(\Omega, \mathbb{F}, P)$ be a probability space and let $\mathbb{G} \subset \mathbb{F}$ be a sigma-algebra. The following three conditions are equivalent:*

1. *For any sigma-algebra $\mathbb{H} \subset \mathbb{F}$ it holds that $\mathbb{G} \perp\!\!\!\perp \mathbb{H}$,*
2. *It holds that $\mathbb{G} \perp\!\!\!\perp \mathbb{G}$,*
3. *$\mathbb{G}$ satisfies a 0-1 law.*

Definition 3.30. (Hansen) *Let $X_1, X_2, ...$ be real-valued random variables on a background space $(\Omega, \mathbb{F}, P)$. The **tail sigma-algebra** of the process is defined as*

$$\mathbb{J}(X_1, X_2, ...) = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, ...).$$

Theorem 3.32. (Hansen) *(Kolmogorov's zero-one law) Let $X_1, X_2, ...$ be real-valued random variables on a background space $(\Omega, \mathbb{F}, P)$. If $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp ...$ then the tail-algebra $\mathbb{J}(X_1, X_2, ...)$ satisfies a 0-1 law.*

Lemma 3.35. (Hansen) *(2nd half of Borel-Cantelli) Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $A_1, A_2, ...$ be a sequence of $\mathbb{F}$-measurable sets. If $A_1 \perp\!\!\!\perp A_2 \perp\!\!\!\perp ...$ then it holds that*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \iff P(A_n \text{ i.o.}) = 0.$$

12.4 Moment generating function

Let X be a random variable with distribution function $F(x) = P(X \leqslant x)$ and Y be a random variable with distribution function $G(y) = P(Y \leqslant y)$.

Definition. (Ex. FinKont) *The moment generating function or Laplace transform of X is*

$$\psi_X(\lambda) = E\left[e^{\lambda X}\right] = \int_{-\infty}^{\infty} e^{\lambda x} dF(x)$$

provided the expectation is finite for $|\lambda| < h$ for some $h > 0$.

The MGF uniquely determine the distribution of a random variable, due to the following result.

Theorem. (Ex. FinKont) (Uniqueness) *If $\psi_X(\lambda) = \psi_Y(\lambda)$ when $|\lambda| < h$ for some $h > 0$, then X and Y has the same distribution, that is, $F = G$.*

There is also the following result of independence for Moment generating functions.

Theorem. (Ex. FinKont) (Independence) *If*

$$E\left[e^{\lambda_1 X + \lambda_2 Y}\right] = \psi_X(\lambda_1)\psi_Y(\lambda_2)$$

for $|\lambda_i| < h$ for $i = 1, 2$ for some $h > 0$, then X and Y are independent random variables.

12.5 Standard distributions

12.5.1 Normal disribution

The following gives a comprehensive table of the standard some properties.

Description	Symbol	Normal distribution
Definition	$\sim$	$X \sim \mathcal{N}(\mu, \sigma^2)$
Parameters	$\theta \in \Theta$	$\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$
Support	$\mathrm{Im}(X)$	$x \in \mathbb{R}$
Density	f	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)^2}$
Distribution	F	$\frac{1}{2}\left(1 + N\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right)$
Mean value	$E[X]$	μ
Variance	$\mathrm{Var}(X)$	σ^2
MGF*	$\psi_X = E[e^{\lambda X}]$	$e^{\mu\lambda + \frac{1}{2}\lambda^2\sigma^2}$
Characteristic function	$\varphi_X(t) = E[e^{itX}]$	$e^{it\mu - \frac{1}{2}\sigma^2 t^2}$

In the table above we used the abbreviations: *MGF = Moment Generating function.

We also used the shorthand: N being the distribution of a standard normal distributed variable $\mathcal{N}(0, 1)$.

Chapter 13

Discrete Time Stochastic Processes

13.1 Convergence concepts

We start this chapter by refering to a sequence $X_1, X_2, ...$ of real-valued random variables as a **proces**. Consider the event $(X_n \to X) = \{\omega \in \Omega \mid X_m(\omega) \to X(\omega) \text{ for } n \to \infty\}$. We want to study such convergence in detail. However first we check measurability. Consider a family $(A_i)_{i\in I} \subset \Omega$ and observe that

$$\Big\{\omega \in \Omega \mid \forall i \in I : \omega \in A_i\Big\} = \bigcap_{i\in I} A_i, \tag{2.1}$$

$$\Big\{\omega \in \Omega \mid \exists i \in I : \omega \in A_i\Big\} = \bigcup_{i\in I} A_i. \tag{2.2}$$

From the standard N, ε definition of a convergent sequence $(x_n)_{n\in\mathbb{N}}$ we may formulate this convergens in the stochastic setting:

$$\begin{aligned}
(X_n \to X) &= \Big\{\omega \in \Omega \mid \forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \geqslant N : |X_n(\omega) - X(\omega)| < \varepsilon\Big\} \\
&= \Big(\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \geqslant N : |X_n - X| < \varepsilon\Big) \\
&= \bigcap_{\epsilon\in\mathbb{R}^+} \Big(\exists N \in \mathbb{N} \forall n \geqslant N : |X_n - X| < \varepsilon\Big) \\
&= \bigcap_{\epsilon\in\mathbb{R}^+} \bigcup_{N=1}^{\infty} \Big(\forall n \geqslant N : |X_n - X| < \varepsilon\Big) \\
&= \bigcap_{\epsilon\in\mathbb{R}^+} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \Big(|X_n - X| < \varepsilon\Big) \in \mathbb{F} \qquad \text{for all } \varepsilon > 0.
\end{aligned}$$

Hence $(X_n \to X)$ lies in $\mathbb{F}$ since $(|X_n - X| < \varepsilon)$ lies in $\mathbb{F}$ since $X_n - X$ is measurable.

Lemma 2.1. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. It holds that*

$$(X_n \to X) \in \mathbb{F}.$$

Definition 2.2. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. We say that X_n **converges** to X **almost surely**, written $X_n \overset{a.s.}{\to} X$, if*

$$P(X_n \to X) = 1. \tag{2.6}$$

Lemma 2.3. (Hansen) *Let $X, X', X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If $X_n \overset{a.s.}{\to} X$ and $X_n \overset{a.s.}{\to} X'$ then $X = X'$ almost surely.*

Lemma 2.7. (Hansen) *Let $X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. Then*

$$\Big((X_n) \text{ is Cauchy}\Big) \in \mathbb{F}.$$

Lemma 2.8. (Hansen) *Let $X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If $P\Big((X_n) \text{ is Cauchy}\Big) = 1$ then there exists and $\mathbb{F}$-measurable real-valued random variable X such that $X_n \xrightarrow{a.s.} X$.*

Theorem 2.10. (Hansen) *Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. Assume that the X-process and the Y-process have the same distribution in the sense that $(X_1, ..., X_n)$ has the same distribution ad $(Y_1, ..., Y_n)$ for all $n \in \mathbb{N}$. If $X_n \xrightarrow{a.s.} X$ for som limit variable X, there is a limit variable Y such that $Y_n \xrightarrow{a.s.} Y$.*

Definition 2.11. (Hansen) *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$ valued random variables on $(\Omega, \mathbb{F}, P)$. We say that $\mathbf{X}_n$ converges to $\mathbf{X}$ **almost surely**, written $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$, if*

$$|\mathbf{X}_n - \mathbf{X}| \xrightarrow{\text{a.s.}} 0. \tag{2.15}$$

Lemma 2.12. (Hansen) *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$ valued random variables on $(\Omega, \mathbb{F}, P)$ such that $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$. Let $f : \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map. Assume that there is a set $A \in \mathbb{B}_k$ such that f is continuous on A and such that $P(\mathbf{X} \in A) = 1$. Then it holds that $f(\mathbf{X}_n) \xrightarrow{a.s.} f(\mathbf{X})$.*

Definition 2.13. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. We say that X_n **converges** to X **in probability**, written $X_n \xrightarrow{P} X$, if*

$$\forall \varepsilon > 0 : \quad P\big(|X_n - X| \geqslant \varepsilon\big) \to 0 \quad \text{for } n \to \infty. \tag{2.17}$$

Lemma 2.14. (Hansen) *Let $X, X', X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P} X'$ then $X = X'$ almost surely.*

Lemma 2.14. (Hansen) *Let $X, X', X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.*

Definition 2.17. (Hansen) *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$ valued random variables on $(\Omega, \mathbb{F}, P)$. We say that $\mathbf{X}_n$ converges to $\mathbf{X}$ **in probability**, written $\mathbf{X}_n \xrightarrow{P} \mathbb{X}$, if*

$$|\mathbf{X}_n - \mathbf{X}| \xrightarrow{\text{P}} 0. \tag{2.23}$$

Lemma 2.18. (Hansen) *Let $X, Y, X_1, Y_1, X_2, Y_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. It holds that*

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{\text{P}} \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \Longleftrightarrow \qquad X_n \xrightarrow{\text{P}} X \text{ and } Y_n \xrightarrow{\text{P}} Y. \tag{2.24}$$

Definition 2.19. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables in $\mathcal{L}^p(\Omega, \mathbb{F}, P)$ for some $p \geqslant 1$. We say that X_n **converges** to X **in** $\mathcal{L}^p$, written $X_n \xrightarrow{\mathcal{L}^p} X$, if*

$$\|X_n - X\|_p \to 0. \tag{2.27}$$

Where the p'th norm is defined as the mapping $\|\cdot\|_p : \Omega \to [0, \infty)$ given by $X \mapsto \big(\int_\Omega |X|^p \, dP\big)^{1/p}$.

One might also define convergence in $\mathcal{L}^p$ by simply saying if $X_n \xrightarrow{\mathcal{L}^p} X$ then $E\,\|X_n - X\|_p \to 0$.

Lemma 2.20. (Hansen) *(Extended Cauchy-Schwarz inequality) Let $X, Y \in \mathcal{L}^p(\Omega, \mathbb{F}, P)$ for some $p \geqslant 1$. For any $a \in [0, p]$ it holds that*

$$E\,|X|^a|Y|^{p-a} \leqslant \Big(E\,|X|^p\Big)^{\frac{a}{p}} \Big(E\,|Y|^p\Big)^{\frac{p-a}{p}}. \tag{2.29}$$

Theorem 2.21. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables in $\mathcal{L}^p(\Omega, \mathbb{F}, P)$ for some $p \in \mathbb{N}$. If $X_n \xrightarrow{\mathcal{L}^p} X$, then it holds that $E\,X_n^p \to E\,X^p$.*

Lemma 2.22. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables in $\mathcal{L}^p(\Omega, \mathbb{F}, P)$ for some $p \geqslant 1$. If $X_n \xrightarrow{\mathcal{L}^p} X$, then $X_n \xrightarrow{P} X$.*

Lemma 2.25. (Hansen) *(Borel-Cantelli) Let $(\Omega, \mathbb{F}, P)$ be a probability space, and let $A_1, A_2, ...$ be a sequence of $\mathbb{F}$-measurable sets. It holds that*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \Rightarrow \quad P(A_n \text{ i.o.}) = 0.$$

Let $A_1, A_2, ...$ be a sequence of subsets of Ω. We define

$$(A_n \text{ i.o.}) = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m, \quad (A_n \text{ evt.}) = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m.$$

One might also define $Y = \sum_{n=1}^{\infty} 1_{A_n}$ and realise that $(A_n \text{ i.o.}) = (Y = \infty)$ and $(A_n \text{ evt.}) = (Y < \infty)$. Also by de Morgan's law it follows that $(A_n \text{ evt.})^c = (A_n^c \text{ i.o.})$.

Theorem 2.26. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If*

$$\forall \varepsilon > 0 : \quad \sum_{n=1}^{\infty} P(|X_n - X| \geqslant \varepsilon) < \infty, \tag{2.32}$$

then it holds that $X_n \xrightarrow{a.s.} X$.

Theorem 2.27. (Hansen) *Let $X, X_1, X_2, ...$ be real-valued random variables on $(\Omega, \mathbb{F}, P)$. If $X_n \xrightarrow{P} X$, then there is a subsequence $X_{n_1}, X_{n_2}, ...$ such that $X_{n_k} \xrightarrow{a.s.} X$ for $k \to \infty$.*

Lemma 2.28. (Hansen) *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$-valued random variables on $(\Omega, \mathbb{F}, P)$ such that $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$. Let $f : \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map. Assume that there is a set $A \in \mathbb{B}_k$ such that f is continuous on A and such that $P(\mathbf{X} \in A) = 1$. Then it holds that $f(\mathbf{X}_n) \xrightarrow{P} f(\mathbf{X})$.*

Lemma. *(Fatou's Lemma) Let $(\Omega, \mathbb{F}, P)$ be a measure space (here probability space). Let $f_n : \mathcal{X} \to [0, \infty]$, with $\mathcal{X} \in \mathbb{F}$, be a sequence of non-negative measurable functions. Assume f_n converge pointwise to $f : \mathcal{X} \to [0, \infty)$. Then*

$$\int_{\mathcal{X}} \liminf_{n\to\infty} f_n \; dP \leqslant \liminf_{n\to\infty} \int_{\mathcal{X}} f_n \; dP.$$

Lemma. *(Holder's Inequality) Let $(\Omega, \mathbb{F}, P)$ be a measure space (here probability space). Let f and g be real-valued (or complex-valued) functions defined on Ω. Assume f and g are measurable. For any $p, q \geqslant 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ it holds that*

$$\left(\int_\Omega |fg|^1 \; dP\right)^1 \leqslant \left(\int_\Omega |f|^p \; dP\right)^{1/p} \left(\int_\Omega |g|^q \; dP\right)^{1/q}$$

13.1.1 Sums and average processes

Lemma 4.1. (Hansen) *Let $X_1, ..., X_n$ be independent real-valued random variables with $E\,X_i^4 < \infty$ for all i. If $E\,X_i = 0$ for all i then it holds that*

$$E\left(\sum_{i=1}^{n} X_i\right)^4 = \sum_{i=1}^{n} E\,X_i^4 + 6\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E\,X_i^2\,E\,X_j^2.$$

Theorem 4.2. (Hansen) *(SLLN, weak form) Let $X_1, X_2, ...$ be a sequence of independent and identically distributed real-valued random variables. If $E\,X_1^4 < \infty$ it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \xrightarrow{\text{a.s.}} \quad E\,X_1. \tag{4.3}$$

Theorem 4.10. (Hansen) *(Etemadi's maximal inequality) Let $X_1, ..., X_n$ be independent real-valued random variables. Consider the cumulative sums*

$$S_k = \sum_{i=1}^{k} X_i \quad \text{for } k = 1, ..., n.$$

For any $\alpha > 0$ it holds that

$$P\left(\max_{j=1,...,n} |S_j| \geqslant 3\alpha\right) \leqslant 3 \max_{j=1,...,n} P(|S_j| \geqslant \alpha). \tag{4.11}$$

Theorem 4.11. (Hansen) *(Levy's maximal inequality) Let $X_1, ..., X_n$ be independent real-valued random variables, each with a symmetric distribution. Consider the cumulative sums*

$$S_k = \sum_{i=1}^{k} X_i \quad \text{for } k = 1, ..., n.$$

For any $\alpha > 0$ it holds that

$$P\left(\max_{j=1,...,n} S_j \geqslant \alpha\right) \leqslant 2P(S_j \geqslant \alpha). \tag{4.13}$$

Corollary 4.12. (Hansen) *Let $X_1, ..., X_n$ be independent real-valued random variables, each with a symmetric distribution. Consider the cumulative sums*

$$S_k = \sum_{i=1}^{k} X_i \quad \text{for } k = 1, ..., n.$$

For any $\alpha > 0$ it holds that

$$P\left(\max_{j=1,...,n} |S_j| \geqslant \alpha\right) \leqslant 2P(|S_j| \geqslant \alpha). \tag{4.14}$$

Theorem 4.13. (Hansen) *(Skorokhod) Let $X_1, X_2, ...$ be a sequence of independent real-valued random variables, and consider the cumulative sums $S_k = \sum_{i=1}^{k} X_i$. Let S be a potential limit variable. It holds that*

$$S_n \xrightarrow{\text{P}} S \quad \Rightarrow \quad S_n \xrightarrow{\text{a.s.}} S.$$

Corollary 4.14. (Hansen) *(Khintchine-Kolmogorov) Let $X_1, X_2, ...$ be a sequence of independent real-valued random variables. Assume $E\,X_n^2 < \infty$ and that $E\,X_n = 0$ for every $n \in \mathbb{N}$. Consider the cumulative sums $S_k = \sum_{i=1}^k X_i$. If*

$$\sum_{n=1}^{\infty} E\,X_n^2 < \infty \tag{4.18}$$

then there exist a limit variable S such that $S_n \to S$ almost surely and in $\mathcal{L}^2$. The limit variable satisfies that

$$E\,S = 0 \quad \text{and} \quad V\,S = \sum_{n=1}^{\infty} V\,X_n.$$

Theorem 4.17. (Hansen) *Let $X_1, X_2, ...$ be a sequence of independent real-valued random variables, and consider the cumulative sums $S_k = \sum_{i=1}^k X_i$. Let S be a potential limit variable. Assume that there is a constant $c > 0$ such that $P(|X_n| \leqslant c) = 1$ for all n, and assume that $E\,X_n = 0$ for all n. The the three statements 1. $S_n \xrightarrow{\text{P}} S$, 2. $S_n \xrightarrow{\text{a.s.}} S$ 3. $S_n \xrightarrow{\mathcal{L}^2} S$ are equivalent.*

Lemma 4.18. (Hansen) *Let $X_1, X_2, ...$ be a sequence of independent real-valued random variables. Assume that there is a constant $c > 0$ such that $P(|X_n| \leqslant c) = 1$ for all n. If the associated random walk $S_n = \sum_{i=1}^n X_i$ satisfies that $S_n \to S$ almost surely for some limit variable then it holds that*

$$1) \quad \sum_{n=1}^{N} E\,X_n \text{ converges in } \mathbb{R} \text{ for } N \to \infty, \tag{13.1}$$

$$2) \quad \sum_{n=1}^{\infty} V(X_n) < \infty. \tag{13.2}$$

Theorem 4.19. (Hansen) *(Kolmogorov's 3-series theorem) Let $X_1, X_2, ...$ be a sequence of independent real-valued random variables. Consider the assoiciated random walk $S_n = \sum_{i=1}^n X_i$. If there is a cut-off value $c > 0$ such that the capped variables $\tilde{X}_n = 1_{|X_n| \leqslant c} X_n$ satisfies that*

$$1) \quad \sum_{n=1}^{\infty} P(X_n \neq \tilde{X}_n) < \infty,$$

$$2) \quad \sum_{n=1}^{N} E\,\tilde{X}_n \text{ converges in } \mathbb{R} \text{ for } N \to \infty,$$

$$3) \quad \sum_{n=1}^{\infty} V(\tilde{X}_n) < \infty,$$

*then there is a real-valued limit variable S such that $S_n \to S$ almost surely. Conversely, if $(S_n)_{n\in\mathbb{N}}$ is almost surely convergent, then the three series above converge for **any** cut-off value $c > 0$.*

Lemma 4.20. (Hansen) *Let $(x_n)_{n\in\mathbb{N}}$ be a real-valued sequence, and let c be a real number. It holds that*

$$x_n \to c \text{ for } n \to \infty \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i \to x \text{ for } n \to \infty.$$

Lemma 4.21. (Hansen) *(Kronecker) Let $(x_n)_{n\in\mathbb{N}}$ be real-valued sequence, and let c be a real number. It holds that*

$$\sum_{i=1}^{n} \frac{x_i}{i} \to c \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} x_i \to 0$$

for $n \to \infty$.

Lemma 4.23. (Hansen) *Let $X_1, X_2, ...$ be a sequence of identically distributed real-valued random variables, and let $\tilde{X}_n = 1_{(|X_n| \leqslant n} X_n$. If $E|X_1| < \infty$ it holds that*

$$\sum_{n=1}^{\infty} \frac{E\ \tilde{X}_n^2}{n^2} < \infty.$$

Theorem 4.24. (Hansen) *(SLLN, strong version) Let $X_1, X_2, ...$ be sequence of independent and identically distributed real-valued random variables. If $E|X_1| < \infty$ it holds that*

$$\frac{1}{n} \sum_{n=1}^{\infty} X_i \overset{\text{a.s.}}{\to} E\ X_1. \tag{4.24}$$

Theorem 4.25. (Hansen) *(SLLN, $\mathcal{L}^p$-version) Let $X_1, X_2, ...$ be sequence of independent and identically distributed real-valued random variables. If $E|X_1|^p < \infty$ for some $p \geqslant 1$, then it holds that*

$$\frac{1}{n} \sum_{n=1}^{\infty} \overset{\mathcal{L}^p}{\to} E\ X_1. \tag{4.26}$$

Lemma 4.26. (Hansen) *Let $X_1, X_2, ...$ be a sequence of pairwise independent, identically distributed real-valued random variables with $E|X_1| < \infty$. Let $n_1 < n_2 < ...$ be a sequence of natural numbers. If there are constants $C_1, C_2 > 0$ and $\alpha > 1$ such that*

$$C_1 \alpha^k \leqslant n_k \leqslant C_2 \alpha^k \quad \text{for } k \to \infty \tag{4.27}$$

then it holds that

$$\frac{1}{n_k} \sum_{i=1}^{n_k} X_i \overset{\text{a.s.}}{\to} E\ X_1 \quad \text{for } k \to \infty.$$

Theorem 4.27. (Hansen) *(Etemahdi's version) Let $X_1, X_2, ...$ be a sequence of pairwise independent, identically distributed real-valued random variables. If $E|X_1| < \infty$ it holds that*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} E\ X_1. \tag{4.30}$$

13.1.2 Ergodic Theory

Definition 5.3. (Hansen) *Let $(\mathcal{X}, \mathbb{E})$ be a measurable space and let $T : \mathcal{X} \to \mathcal{X}$ be measurable. A probability measure μ on $(\mathcal{X}, \mathbb{E})$ is **invariant** under T if*

$$\mu\big(T^{-1}(A)\big) = \mu(A) \quad \text{for all } A \in \mathbb{E} \tag{5.5}$$

*In this case we call the quadruple $(\mathcal{X}, \mathbb{E}, \mu, T)$ a **measure-preserving dynamical system**.*

We say that a set $A \in \mathbb{E}$ is an **invariant set** if $T^{-1}(A) = A$ i.e. the orbit of all $x \in A$ stays in A.

Definition 5.5. (Hansen) *A measure-preserving dynamical system $(\mathcal{X}, \mathbb{E}, \mu, T)$ is **ergodic** if*

$$T^{-1}(A) = A,\ A \in \mathbb{E} \quad \Rightarrow \quad \mu(A) \in \{0, 1\}. \tag{5.7}$$

Definition 5.6. (Hansen) *A measure-preserving dynamical system $(\mathcal{X}, \mathbb{E}, \mu, T)$ is **mixing** if*

$$\mu(A \cap T^{-n}(B)) \to \mu(A)\mu(B) \quad \text{for all } A, B \in \mathbb{E}. \tag{5.8}$$

Lemma 5.7. (Hansen) *If a measure-preserving dynamical system $(\mathcal{X}, \mathbb{E}, \mu, T)$ is mixing then it is also ergodic.*

Lemma 5.8. (Hansen) *Let $(\mathcal{X}, \mathbb{E}, \mu, T)$ be a measure-preserving dynamical system. Let $\mathbb{D}$ be an $\cap$-stable generator for $\mathbb{E}$. If*

$$\mu(A \cap T^{-n}(B)) \to \mu(A)\mu(B) \quad \text{for all } A, B \in \mathbb{D}. \tag{5.10}$$

then the system is mixing. (and ergodic)

Lemma 5.9. (Hansen) *Let $(\mathcal{X}, \mathbb{E}, \mu)$ be a probability space, and let $T : \mathcal{X} \to \mathcal{X}$ be a measure-preserving map. Let $(\mathcal{Y}, \mathbb{G})$ be another measurable space, and let $S : \mathcal{Y} \to \mathcal{Y}$ be a measurable map. Suppose there is a measurable map $\gamma : \mathcal{X} \to \mathcal{Y}$ such that the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{T} & \mathcal{X} \\ \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle \gamma} \\ \mathcal{Y} & \xrightarrow{S} & \mathcal{Y} \end{array}$$

Then $(\mathcal{Y}, \mathbb{G}, \gamma(\mu), S)$ is a measure-preserving dynamical system.

Lemma 5.10. (Hansen) *Let $(\mathcal{X}, \mathbb{E}, \mu, T)$ and $(\mathcal{Y}, \mathbb{G}, \nu, S)$ be two measure-preserving dynamical systems. Suppose there is a measurable map $\gamma : \mathcal{X} \to \mathcal{Y}$ such that $\nu = \gamma(\mu)$ and such that the diagram in lemma 5.9 commutes. If $(\mathcal{X}, \mathbb{E}, \mu, T)$ is ergodic then $(\mathcal{Y}, \mathbb{G}, \nu, S)$ is also ergodic.*

Lemma 5.11. (Hansen) *(Maximal Ergodic Lemma) Let $(\mathcal{X}, \mathbb{E}, \mu, T)$ be a measure-preserving dynamical system, and let $f : \mathcal{X} \to \mathbb{R}$ be Borel measurable. If $f \in \mathcal{L}^1(\mu)$ then it holds that*

$$\int_{(M_n > 0)} f \, d\mu \geqslant 0 \tag{5.14}$$

where $M_n = \max\{0, S_1, S_2, ..., S_n\}$ from the sequence

$$\Big(f(x), f \circ T(x), f \circ T^2(x), f \circ T^3(x), ...\Big) \quad \text{with} \quad S_n = \sum_{i=0}^{n-1} f \circ T^i.$$

Theorem 5.12. (Hansen) *(Birkhoff's ergodic theorem) Let $(\mathcal{X}, \mathbb{E}, \mu, T)$ be an ergodic system. For $f \in \mathcal{L}^1(\mu)$ it holds that*

$$\frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i \overset{\text{a.s.}}{\to} \int f \, d\mu. \tag{5.16}$$

Theorem 5.13. (Hansen) *(Ergodic theorem, $\mathcal{L}^p$-version) Let $(\mathcal{X}, \mathbb{E}, \mu, T)$ be an ergodic system. If $f \in \mathcal{L}^p(\mu)$ for some $p \geqslant 1$ then it holds that*

$$\frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i \overset{\mathcal{L}^p}{\to} \int f \, d\mu. \tag{5.21}$$

Lemma 5.14. (Hansen) *Let $(\mathcal{X}, \mathbb{E})$ be a measurable space. The measurable finite dimensional product sets in $\mathcal{X}^{\mathbb{N}}$ form an $\cap$-stable generator for $\mathbb{E}^{\otimes \mathbb{N}}$.*

An element of the space $\mathcal{X}^{\mathbb{N}}$ is a countable set of coordinates x_n for $n \in \mathcal{N}$ with $x_n \in \mathcal{X}$. A **finite dimensional product set** in $\mathcal{X}^{\mathbb{N}}$ is set on the form

$$A_1 \times ... \times A_k \times \mathcal{X} \times \mathcal{X} \times ...$$

where $A_1, ..., A_k \subset \mathcal{X}$. We also define the **projection sigma-algebra** $\mathbb{E}^{\otimes\mathbb{N}}$ as $\sigma\left(\left(\hat{X}_n(\mathcal{X}^n)\right)_{n\in\mathbb{N}}\right)$ where $\hat{X}_n(x_1, ..., x_n) = x_n$.

Definition 5.15. (Hansen) *Let $X_1, X_2, ...$ be a sequence of $(\mathcal{X}, \mathbb{E})$-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$, and let $\mathbb{X} = (X_1, X_2, ...)$ be their bundling. The **distribution** of the process is the image measure $\mathbb{X}(P)$ on $(\mathcal{X}^{\mathbb{N}}, \mathbb{E}^{\otimes\mathbb{N}})$.*

Lemma 5.16. (Hansen) *Let $\mathbb{X} = (X_1, X_2, ...)$ and $\mathbb{Y} = (Y_1, Y_2, ...)$ be two $(\mathcal{X}, \mathbb{E})$-valued stochastic process, defined on a common background space . The two processes $\mathbb{X}$ and $\mathbb{Y}$ have the same distribution if and only if the have the same fidis. This can be checked by showing that*

$$P(X_1 \in A_i, ..., X_k \in A_k) = P(Y_1 \in A_1, ..., Y_k \in A_k) \tag{5.25}$$

for any $k \in \mathbb{N}$ and any choice of $A_1, ..., A_k \in \mathbb{E}$.

Definition 5.18. (Hansen) *Let $X_1, X_2, ...$ be a sequence of $(\mathcal{X}, \mathbb{E})$-valued random variable, defined on a background space $(\Omega, \mathbb{F}, P)$, and let $\mathbb{X} = (X_1, X_2, ...)$ be their bundling. Then we define:* 1. *The proces $\mathbb{X}$ is **stationary** if the distrbution $\mathbb{X}(P)$ is an S-invariant probability on $\left(\mathbb{R}^{\mathbb{N}}, \mathbb{B}^{\otimes\mathbb{N}}\right)$,* 2. *The proces $\mathbb{X}$ is **ergodic** if it is stationary and if the dynamical system $\left(\mathbb{R}^{\mathbb{N}}, \mathbb{B}^{\otimes\mathbb{N}}, \mathbb{X}(P), S\right)$ is ergodic.* 3. *The proces $\mathbb{X}$ is **mixing** if it is stationary and if the dynamical system $\left(\mathbb{R}^{\mathbb{N}}, \mathbb{B}^{\otimes\mathbb{N}}, \mathbb{X}(P), S\right)$ is mixing. with S being the **shift** map defined as $S(x_1, x_2, ...) = (x_2, x_3, ...)$.*

Theorem 5.20. (Hansen) *(Khintchine's ergodic theorem) Let $X_1, X_2, ...$ be a stationary and ergodic sequence of real-valued random variables. If $E|X_1| < \infty$ it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} E\ X_1. \tag{5.27}$$

If $E|X_1|^p < \infty$ for some $p \geqslant 1$, the convergence is also in $\mathcal{L}^p$.

Theorem 5.21. (Hansen) *(Ergodic transformation theorem) Let $X_1, X_2, ...$ be a sequence of real-valued random variables. For a measurable function $\phi : \left(\mathbb{R}^{\mathbb{N}}, \mathbb{B}^{\otimes\mathbb{N}}\right) \to (\mathbb{R}, \mathbb{B})$ we define new real-valued random variables*

$$Y_n = \phi(X_n, X_{n+1}, ...) = \phi \circ S^{n-1}(\mathbb{X}) \quad \text{for } n \in \mathbb{N}.$$

If $X_1, X_2, ...$ is stationary and ergodic then $Y_1, Y_2, ...$ is also stationary and ergodic.

Definition 5.23. (Hansen) *Let $(X_n)_{n\in\mathbb{Z}}$ be a two-sided sequence of real-valued random variables, defined on a background space $(\Omega, \mathbb{F}, P)$, and let $\mathbb{X}$ be their bundling.* 1. *The proces $\mathbb{X}$ is **stationary** if the distrbution $\mathbb{X}(P)$ is an S-invariant probability on $\left(\mathbb{R}^{\mathbb{Z}}, \mathbb{B}^{\otimes\mathbb{Z}}\right)$,* 2. *The proces $\mathbb{X}$ is **ergodic** if it is stationary and if the dynamical system $\left(\mathbb{R}^{\mathbb{Z}}, \mathbb{B}^{\otimes\mathbb{Z}}, \mathbb{X}(P), S\right)$ is ergodic.* 3. *The proces $\mathbb{X}$ is **mixing** if it is stationary and if the dynamical system $\left(\mathbb{R}^{\mathbb{Z}}, \mathbb{B}^{\otimes\mathbb{Z}}, \mathbb{X}(P), S\right)$ is mixing.*

Theorem 5.25. (Hansen) *(Khintchine's ergodic theorem, two-sided version) Let $(X_n)_{n\in\mathbb{Z}}$ be a two-sided sequence of real-valued random variables. If the sequence is stationary and ergodic and if $E|X_1| < \infty$ then it holds that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} E\ X_1. \tag{5.30}$$

If $E|X_1|^p < \infty$ for some $p \geqslant 1$, the convergence is also in $\mathcal{L}^p$.

Theorem 5.26. (Hansen) *(Ergodic transformation theorem) Let $(X_n)_{n\in\mathbb{Z}}$ be a two-sided sequence of real-valued random variables. For a measurable function $\phi : \left(\mathbb{R}^{\mathbb{Z}}, \mathbb{B}^{\otimes\mathbb{Z}}\right) \to (\mathbb{R}, \mathbb{B})$ we define new real-valued random variables*

$$Y_n = \phi \circ S^n(\mathbb{X}) \quad \text{for } n \in \mathbb{Z}.$$

If $(X_n)_{n\in\mathbb{Z}}$ is stationary and ergodic then $(Y_n)_{n\in\mathbb{Z}}$ is also stationary and ergodic.

13.1.3 Weak Convergence

Definition 6.1. (Hansen) *A sequence of probability measures $\nu_1, \nu_2, ...$ on $(\mathbb{R}, \mathbb{B})$ is said to **converge weakly** to a limit probability measure ν if*

$$\int f \, d\nu_n \to \int f \, d\nu \quad \text{for every } f \in C_b(\mathbb{R}) \tag{6.2}$$

We write $\nu_n \xrightarrow{wk} \nu$ to denote weak convergence.

Theorem 6.4. (Hansen) *(Scheffe's) Let $\nu_1, \nu_2, ...$ and ν be probability measures on $(\mathbb{R}, \mathbb{B})$. Assume that for some choice of basic measure μ it holds that $\nu_n = f_n \cdot \mu$ for every n and $\nu = f \cdot \mu$ for suitable density functions f_n and f. If*

$$f_n(x) \to f(x) \quad \mu\text{-almost surely}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

Lemma 6.8. (Hansen) *Let μ and ν be two probability measures on $(\mathbb{R}, \mathbb{B})$. If*

$$\int f \, d\mu = \int f \, d\nu \quad \text{for all } f \in C_b(\mathbb{R}) \tag{6.7}$$

then it holds that $\mu = \nu$.

Theorem 6.9. (Hansen) *Let $\nu_1, \nu_2, ...$ be a sequence of probability measures on $(\mathbb{R}, \mathbb{B})$ and let μ and ν be two extra probability measures. If*

$$\nu_n \xrightarrow{\text{wk}} \mu \quad \text{and} \quad \nu_n \xrightarrow{\text{wk}} \nu$$

then $\mu = \nu$.

Definition 6.10. (Hansen) *A sequence of real-valued variables $X_1, X_2, ...$, defined on a common background space $(\Omega, \mathbb{F}, P)$ is said to **converge in distrbution** to a limit variable X if*

$$\int f(X_n) \, dP \to \int f(X) \, dP \quad \text{for every } f \in C_b(\mathbb{R}). \tag{6.9}$$

We will write $X_n \xrightarrow{\mathcal{D}} X$ to denote convergence in distribution.

Lemma 6.11. (Hansen) *Let $X_1, X_2, ...$ and X be real-valued random variables. It holds that*

$$X_n \xrightarrow{\text{P}} X \quad \Rightarrow \quad X_n \xrightarrow{\mathcal{D}} X.$$

Lemma 6.12. (Hansen) *Let $X_1, X_2, ...$ be real-valued random variable and let $x_0 \in \mathbb{R}$. It holds that*

$$X_n \xrightarrow{\mathcal{D}} x_0 \quad \Rightarrow \quad X_n \xrightarrow{\text{P}} x_0.$$

Theorem 6.13. (Hansen) *Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$. Let $\mathcal{H}$ be a class of bounded, non-negative and measurable functions with the following approximation property: For $f \in C_b(\mathbb{R})$ with $f \geqslant 0$ there is a sequence $h_1, h_2, ...$ of $\mathcal{H}$-functions such that $h_n \nearrow f$. If*

$$\int h \, d\nu_n \to \int h \, d\nu \quad \text{for all } h \in \mathcal{H}. \tag{6.11}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

Theorem 6.14. (Hansen) *Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$. If*

$$\int f \, d\nu_n \to \int f \, d\nu \quad \text{for all } f \in C_c(\mathbb{R}) \tag{6.13}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

The class $C_c(\mathbb{R})$ is denoted as the set of all continuous real-valued functions with compact support i.e. there exist a $M > 0$ such that $f(x) = 0$ for all $|x| > M$.

Lemma 6.15. (Hansen) *Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and let $F, F_1, F_2, ...$ be the corresponding distribution functions. If $\nu_n \xrightarrow{wk} \nu$ then it holds that*

$$F_n(x_0) \to F(x_0),$$

whenever x_0 is a point of continuity for F.

Theorem 6.17. (Hansen) *(Helly-Bray) Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and let $F, F_1, F_2, ...$ be the corresponding distribution functions. It holds that $\nu_n \xrightarrow{wk} \nu$ if and only if there is a dense subset $A \subset \mathbb{R}$ such that*

$$F_n(x) \to F(x) \quad \text{for every } x \in A. \tag{6.16}$$

Theorem 6.18. (Hansen) *Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$. Let $F, F_1, F_2, ...$ be the corresponding distribution functions, and let $q, q_1, q_2, ...$ be the corresponding quantile functions. If $\nu_n \xrightarrow{wk} \nu$ then it holds that*

$$q_n(p) \to q(p)$$

for any $p \in (0, 1)$ such that the equation $F(x) = p$ hast at most one solution.

Theorem 6.19. (Hansen) *(Skorokhod's representation theorem) Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$. If $\nu_n \xrightarrow{wk} \nu$ then it is possible to find random variables $X, X_1, X_2, ...$ on a background space $(\Omega, \mathbb{F}, P)$ such that*

$$X(P) = \nu, \; X_1(P) = \nu_1, \; X_2(P) = \nu_2, ...$$

and such that $X_n \xrightarrow{a.s.} X$.

Corollary 6.20. (Hansen) *Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$ such that $\nu_n \xrightarrow{wk} \nu$. Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded and measurable function. If there is a Boral-measurable set $C \subset \mathbb{R}$ such that h is continuous in every point of C and such that $\nu(C) = 1$, then it holds that*

$$\int h \, d\nu_n \to \int h \, d\nu. \tag{6.20}$$

Corollary 6.21. (Hansen) *(Portmanteau's lemma) Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$ such that $\nu_n \xrightarrow{wk} \nu$. For any open set $G \subset \mathbb{R}$ it holds that*

$$\liminf \nu_n(G) \geqslant \nu(G) \tag{6.21}$$

Definition 6.22. (Hansen) *The **characteristic function** for a probability measure ν on $(\mathbb{R}, \mathbb{B})$ is the function $\phi : \mathbb{R} \to \mathbb{C}$ given by*

$$\phi(\theta) = \int e^{ix\theta} \, d\nu(x). \quad \text{for } \theta \in \mathbb{R}. \tag{6.23}$$

Some useful observations include $|e^{ix\theta}| = 1$ hence $\phi(\theta) \leqslant 1$ for all $\theta \in \mathbb{R}$. We may also split the ϕ into an imaginary part and a real part with Euler's cartesian form

$$\phi(\theta) = \int \cos(x\theta)\ d\nu(x) + i \int \sin(x\theta)\ d\nu(x) \tag{6.24}$$

And lastly we have the implication

$$\nu_n \xrightarrow{\text{wk}} \nu \quad \Rightarrow \quad \phi_n(\theta) \to \phi(\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Furthermore, if $Y = \xi + \sigma X$ and $X \sim \mathcal{N}(0,1)$ we have

$$\phi_Y(\theta) = e^{i\xi\theta} e^{-\sigma^2\theta^2/2}.$$

Theorem 6.28. (Hansen) *The characteristic function for any probability measure ν on $(\mathbb{R}, \mathbb{B})$ is uniformly continuous.*

Theorem 6.29. (Hansen) *Let ν be a probability measure on $(\mathbb{R}, \mathbb{B})$. If*

$$\int |x|^k\ d\nu(x) < \infty$$

for some $k \in \mathbb{N}$, then the characteristic function ϕ is C^k and it holds that

$$\phi^{(k)}(\theta) = i^k \int x^k e^{i\theta x}\ d\nu(x) \qquad \text{for } \theta \in \mathbb{R}. \tag{6.31}$$

Definition 6.30. (Hansen) *The **convolution** of two probability measures ν and ξ on $(\mathbb{R}, \mathbb{B})$ is the image measure*

$$\nu * \xi = \kappa(\nu \otimes \xi) \tag{6.33}$$

where $\kappa : \mathbb{R}^2 \to \mathbb{R}$ is the addition map $\kappa(x, y) = x + y$.

Theorem 6.31. (Hansen) *Let ν and ξ be two probability measures on $\mathbb{R}$. If $\xi = f \cdot m$, then the convolution $\nu * \xi$ will have a density with respect to m. The density is given as*

$$g(x) = \int f(x - y)\ d\nu(y) \qquad \text{for } x \in \mathbb{R}. \tag{6.35}$$

Lemma 6.31. (Hansen) *Let ν_1 and ν_2 be two probability measures on $(\mathbb{R}, \mathbb{B})$ with characteristic functions ϕ_1 and ϕ_2. The convolution $\nu_1 * \nu_2$ has characteristic function γ given by*

$$\gamma(\theta) = \phi_1(\theta)\phi_2(\theta) \qquad \text{for } \theta \in \mathbb{R}. \tag{6.37}$$

Theorem 6.34. (Hansen) *Let $\xi, \nu, \nu_1, \nu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathbb{B})$. It holds that*

$$\nu_n \xrightarrow{\text{wk}} \nu \quad \Rightarrow \quad \nu_n * \xi \xrightarrow{\text{wk}} \nu * \xi.$$

Definition 6.35. (Hansen) *A probability measure $\nu = f \cdot m$ on $(\mathbb{R}, \mathbb{B})$ with density f with respect to Lebesgue measure is of **Polya class** if $f \in C_b(\mathbb{R})$ and if the Fourier transform $\hat{f}$ is m-integrable.*

Lemma 6.39. (Hansen) *Let ν and ξ be two probability measures on $\mathbb{R}$. If ξ is of Polya class then the convolution $\nu * \xi$ is also of Polya class.*

Theorem 6.40. (Hansen) *(Inversion theorem) Let $\nu = f \cdot m$ be a probability measure on $(\mathbb{R}, \mathbb{B})$ of Polya class. It holds that*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\theta x} \hat{f}(\theta)\ d\theta,\ x \in \mathbb{R}. \tag{6.39}$$

Theorem 6.41. (Hansen) *Let ν_1 and ν_2 be two probability measures on $(\mathbb{R}, \mathbb{B})$ with characteristic functions ϕ_1 and ϕ_2. If*

$$\phi_1(\theta) = \phi_2(\theta), \ \forall \theta \in \mathbb{R}$$

then $\nu_1 = \nu_2$.

Lemma 6.42. (Hansen) *Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and uniformly continuous. For every $\varepsilon > 0$ there is a probability measure ξ of Polya class with the property that*

$$\left| f(x) - \int f(x+y) \ d\xi(y) \right| < \varepsilon, \ \forall x \in \mathbb{R}. \tag{6.43}$$

Theorem 6.43. (Hansen) *(Continuity theorem) Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $(\mathbb{R}, \mathbb{B})$ with characteristic functions $\phi, \phi_1, \phi_2, ...$ If*

$$\phi_n(\theta) \to \phi(\theta), \ \theta \in \mathbb{R}, \tag{6.45}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

Definition 6.44. (Hansen) *A sequence of probability measures $\nu_1, \nu_2, ...$ on $\left(\mathbb{R}^k, \mathbb{B}_k\right)$ is said to **converge weakly** to a limit probability measure ν if*

$$\int f(x) \ d\nu_n(x) \to \int f(x) \ d\nu(x), \forall f \in C_b(\mathbb{R}^k). \tag{6.46}$$

We will write $\nu_n \xrightarrow{wk} \nu$ to denote weak convergence.

Theorem 6.45. (Hansen) *(Continuity theorem) Let $\nu, \nu_1, \nu_2, ...$ be probability measures on $\left(\mathbb{R}^k, \mathbb{B}_k\right)$ with characteristic functions $\phi, \phi_1, \phi_2, ...$ If*

$$\phi_n(\theta) \to \phi(\theta), \ \theta \in \mathbb{R}^k, \tag{6.47}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

Lemma 6.46. (Hansen) *Let $\mathbf{X}$ be an $\mathbb{R}^k$-valued random variable with characteristic function $\phi_{\mathbf{X}}$, and let $\mathbf{Y}$ be an $\mathbb{R}^m$-valued random variable with characteristic function $\phi_{\mathbf{Y}}$. If $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ then the bundle $(\mathbf{X}, \mathbf{Y})$ is an $\mathbb{R}^{k+m}$-valued random variable with*

$$\phi_{(\mathbf{X},\mathbf{Y})}(\theta_1, \theta_2) = \phi_{\mathbf{X}}(\theta_1)\phi_{\mathbf{Y}}(\theta_2), \ \theta_1 \in \mathbb{R}^k, \theta_2 \in \mathbb{R}^m. \tag{6.49}$$

Theorem 6.47. (Hansen) *(Continuous mapping theorem) Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{X}$ be random variables with values in $\mathbb{R}^k$, and let $h : \mathbb{R}^k \to \mathbb{R}^m$ be continuous. If $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, then it holds that $h(\mathbf{X}_n) \xrightarrow{\mathcal{D}} h(\mathbf{X})$.*

Theorem 6.48. (Hansen) *(Cramer-Wold's device) Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{X}$ be random variables with values in $\mathbb{R}^k$. If*

$$\mathbf{v}^\top \mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{v}^\top \mathbf{X}, \tag{6.51}$$

for any fixed vector $\mathbf{v} \in \mathbb{R}^k$, then it holds that $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$.

Lemma 6.49. (Hansen) *Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ be random variables with values in $\mathbb{R}^k$, let $\mathbf{Y}_1, \mathbf{Y}_2, ...$ be random variables in $\mathbb{R}^m$, and let $\mathbf{y}$ be a vector in $\mathbb{R}^m$. If it holds that*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \qquad \mathbf{Y}_n \xrightarrow{\mathrm{P}} \mathbf{y}$$

then the bundle $(\mathbf{X}_n, \mathbf{Y}_n)$ in $\mathbb{R}^{k+m}$ will satisfy that

$$(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\mathcal{D}} (\mathbf{X}, \mathbf{y}).$$

Corollary 6.50. (Hansen) *(Slutsky's lemma) Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{Y}_1, \mathbf{Y}_2, ...$ be random variables with values in $\mathbb{R}^k$. It holds that*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \quad \mathbf{Y}_n \xrightarrow{\mathrm{P}} \mathbf{0} \quad \Rightarrow \quad \mathbf{X}_n + \mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{X}.$$

Corollary 6.51. (Hansen) *Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{X}$ be random variables with values in $\mathbb{R}^k$ and let $Y_1, Y_2, ...$ be real-valued random variables. It holds that*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \quad Y_n \xrightarrow{\mathrm{P}} 1 \quad \Rightarrow \quad Y_n \mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}.$$

Corollary 6.52. (Hansen) *Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{X}$ be random variables with values in $\mathbb{R}^k$ and let $Y_1, Y_2, ...$ be real-valued random variables. If*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \quad Y_n \xrightarrow{\mathrm{P}} 0 \quad \Rightarrow \quad Y_n \mathbf{X}_n \xrightarrow{\mathrm{P}} \mathbf{0}.$$

Definition 6.53. (Hansen) *The $\mathbb{R}^k$-valued random variable $\mathbf{Z} = (Z_1, ..., Z_k)$ has a **multivariate Gaussian distribution** if and only if the real-valued random variable $\sum_{j=1}^k c_j Z_j$ has a one-dimensional Gaussian distribution for every choice of $c_1, ..., c_k \in \mathbb{R}$.*

Theorem 6.54. (Hansen) *Let $\mathbf{Z} = (Z_1, ..., Z_k)$ have a multivariate Gaussian distribution with $E\ \mathbf{Z} = \xi$ and $V\ \mathbf{Z} = \Sigma$. Then the characteristic function is*

$$\phi_{\mathbf{Z}}(\theta) = e^{i\theta^\top \xi} \exp\left(-\frac{1}{2}\theta^\top \Sigma \theta\right), \ \theta \in \mathbb{R}^k. \tag{6.53}$$

Conversly, if $\mathbf{Z}$ has characteristic function given by (6.53) for some $\xi \in \mathbb{R}^k$ and some symmetric, positive semi-definite $k \times k$ matrix Σ then $\mathbf{Z}$ has a multivariate Gaussian distribution with $E\ \mathbf{Z} = \xi$ and $V\ \mathbf{Z} = \Sigma$.

Corollary 6.55. (Hansen) *Let $\mathbf{Z}$ be an $\mathbb{R}^k$-valued random variable, let $\mathbf{a} \in \mathbb{R}^m$ and let B be an $m \times k$ matrix. It holds that*

$$\mathbf{Z} \sim \mathcal{N}(\xi, \Sigma) \quad \Rightarrow \quad \mathbf{a} + B\mathbf{Z} \sim \mathcal{N}\left(\mathbf{a} + B\xi, B\Sigma B^\top\right).$$

Lemma 6.56. (Hansen) *Let $\mathbf{X} = (X_1, ..., X_k)$ and $\mathbf{Y} = (Y_1, ..., Y_m)$ be random variables with values in $\mathbb{R}^k$ respectively $\mathbb{R}^m$. If both $\mathbf{X}$ and $\mathbf{Y}$ have multivariate Gaussian distributions and if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then the $\mathbb{R}^{k+m}$-valued bundle $(\mathbf{X}, \mathbf{Y})$ has a multivariate Gaussian distribution.*

Lemma 6.58. (Hansen) *Let $\mathbf{X} = (X_1, ..., X_k)$ and $\mathbf{Y} = (Y_1, ..., Y_m)$ be random variables with values in $\mathbb{R}^k$ respectively $\mathbb{R}^m$. If the $\mathbb{R}^{k+m}$-valued bundle $(\mathbf{X}, \mathbf{Y})$ has a multivariate Gaussian distribution, and if*

$$\mathrm{Cov}(X_j, Y_l) = 0, \ \forall j \text{ and } l$$

then $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$.

Definition 6.59. (Hansen) *Let $\mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$-valued random variables. Let $\xi \in \mathbb{R}^k$ be a vector and let Σ be a symmetric, positive semi-definite $k \times k$ matrix. We sat that $\mathbf{X}_n$ has an **asymptotic normal distribution** with parameters $\left(\xi, \frac{1}{n}\Sigma\right)$, written*

$$\mathbf{X}_n \overset{\text{a.s.}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right),$$

if it holds that

$$\sqrt{n}(\mathbf{X}_n - \xi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

Lemma 6.60. (Hansen) *Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{Y}$ be random variables with values in $\mathbb{R}^k$. If it holds that $\mathbf{X}_n \overset{a.s.}{\sim} \mathcal{N}(\xi, \frac{1}{n}\Sigma)$ then it follows that $\mathbf{X}_n \xrightarrow{P} \xi$.*

Lemma 6.61. (Hansen) *Let $\mathbf{X}_1, \mathbf{X}_2, ...$ and $\mathbf{Y}$ be $\mathbb{R}^k$-valued random variables, and assume that $\sqrt{n}\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$. Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map. Assume that $g(\mathbf{0}) = \mathbf{0}$ and that g is differentiable in $\mathbf{0}$ with deriviate $Dg(\mathbf{0}) = A$. Then it holds that*

$$\sqrt{n}g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} A\ \mathbf{Y}.$$

Lemma 6.62. (Hansen) *(Delta method) Let $\mathbf{X}_1, \mathbf{X}_2, ...$ be $\mathbb{R}^k$-valued random variables, and let $f : \mathbb{R}^k \to \mathbb{R}^m$ be measurable. If f is differentiable in ξ, then it holds that*

$$\mathbf{X}_n \overset{\text{a.s.}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right) \quad \Rightarrow \quad f(\mathbf{X}_n) \overset{\text{a.s.}}{\sim} \mathcal{N}\left(f(\xi), \frac{1}{n}Df(\xi)\Sigma Df(\xi)^\top\right).$$

13.1.4 Central Limit Theorems

Lemma 7.1. (Hansen) *Let $z_1, ..., z_n$ and $w_1, ..., w_n$ be complex numbers. If $|z_i| \leqslant 1$ and $|w_i| \leqslant 1$ for all $i = 1, ..., n$ then it holds that*

$$\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| \leqslant \sum_{i=1}^n |z_i - w_i|. \tag{7.1}$$

Theorem 7.2. (Hansen) *(Basic CLT) Let $X_1, X_2, ...$ be independent and identically distributed real-valued random variables. Assume that $E\ X_1 = 0$ and $E\ X_1^2 = 1$. Then it holds that*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{7.3}$$

Theorem 7.3. (Hansen) *(Laplace's CLT) Let $X_1, X_2, ...$ be independent and identically distributed real-valued random variables with $E\ X_1^2 < \infty$. It holds that*

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathcal{D}} \mathcal{N}\left(E\ X_1, \frac{V\ X_1}{n}\right) \tag{7.4}$$

Theorem 7.7. (Hansen) *(Laplace's CLT, multivariate version) Let $\mathbf{X}_1, \mathbf{X}_2, ...$ be independent and identically distributed random variables with values in $\mathbb{R}^k$. Assume that $E|\mathbf{X}_1|^2 < \infty$. It holds that*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \xrightarrow{\mathcal{D}} \mathcal{N}\left(E\ \mathbf{X}_1, \frac{1}{n}V\ \mathbf{X}_1\right) \tag{7.7}$$

Definition 7.10. (Hansen) *Let (X_{nm}) be a centralized array of real-valued random variables.* a. *The array satisfies the **vanishing variance condition** if*

$$\max_{m=1,...,n} E\ X_{nm}^2 \to 0. \tag{7.8}$$

b. *The array satisfies **Lindeberg's condition** if*

$$\forall c > 0 : \qquad \sum_{m=1}^n \int_{(|X_{nm}|>c)} X_{nm}^2\ dP \to 0. \tag{7.9}$$

c. *The array satisfies **Lyapounov's condition** of order $\alpha > 2$ if*

$$\sum_{m=1}^n E\ |X_{nm}|^\alpha \to 0. \tag{7.10}$$

Lemma 7.11. (Hansen) *Lyapounov's condition of order $\alpha > 2$ implies Lindeberg's condition.*

Lemma 7.12. (Hansen) *Lindeberg's condition implies the vanishing variance condition.*

Theorem 7.14. (Hansen) *(Lindeberg's CLT) Let (X_{nm}) be a centralized array of real-valued random variables with $E\ X_{nm}^2 < \infty$. Assume that the array satisfies that*

$$E\ X_{nm} = 0,\ \forall n, m,$$

and that

$$\sum_{m=1}^{n} E\ X_{nm}^2 = 1. \tag{7.13}$$

Assume that the array has independence within rows i.e. $X_{i1} \perp\!\!\!\perp ... \perp\!\!\!\perp X_{ii}$ for all $i = 1, ..., n$ and satisfies Lindeberg's condition. Then it holds that

$$\sum_{m=1}^{n} X_{nm} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Theorem 7.19. (Hansen) *(Lindeberg's CLT, multivariate version) Let $(\mathbf{X}_{nm})$ be a triangular array of random variables with values in $\mathbb{R}^k$ with $E\ |\mathbf{X}_{nm}|^2 < \infty$ for all n, m. Assume that*

$$E\ \mathbf{X}_{nm} = \mathbf{0},\ \forall n, m$$

and

$$\sum_{m=1}^{n} V\ \mathbf{X}_{nm} \to \Sigma$$

for a fixed $k \times k$ matrix Σ. Assume that the array has independence within rows, and assume that the associated real-valued array $(|\mathbf{X}_{nm}|)$ satisfies Lindeberg's condition. Then it holds that

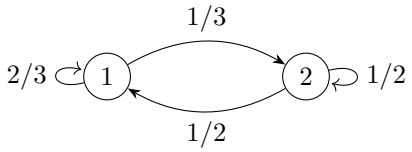
$$\sum_{m=1}^{n} \mathbf{X}_{nm} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma)$$

Chapter 14

Markov Chains

14.1 Definition of a Markov Chain

Transitionsdiagram. Et diagram, hvor der fremgår ssh. for et system overgår til et nyt stadie.



Definition. (Stokastisk process i diskret tid) *Lad $\mathcal{X} = (\mathcal{X}_n)_{n\in\mathbb{N}_0}$ være en familie af stokastiske variable og $\mathcal{S}$ et udfaldsrum . Da er $\mathcal{X}$ en funktion der opfylder $\mathcal{X} : \mathbb{N}_0 \times \Omega \to \mathcal{S}$ med $(n, w) \mapsto \mathcal{X}_n(w)$.*

Bemærkning.

i. $\mathbb{N}_0$ er tidsvariablen senere $[0, \infty)$
ii. $\mathcal{S}$ er højst tællelig

Definition. (Markov kæder i diskret tid) *En stokastisk proces i diskret tid $(X_n)_{n\in\mathbb{N}_0}$ har Markov egenskaben hvis for alle stadier $i_0, ..., i_n \in \mathcal{S}$ opfylder*

$$\mathbb{P}(\mathcal{X}_n = i_n|\mathcal{X}_0, ..., \mathcal{X}_{n-1} = i_{n-1}) = \mathbb{P}(\mathcal{X}_n = i_n|\mathcal{X}_{n-1} = i_{n-1}).$$

Da kaldes denne stokastiske proces en Markov kæde. Dvs. For hvert n er det næste udfald kun afhængig af dette forrige udfald.

Definition. (Initial fordeling) *For alle $i \in \mathcal{S}$ tilhører en initial sandsynlighed for at starte i stadiet i givet ved*

$$\phi(i) := \mathbb{P}(\mathcal{X}_0 = i),$$

hvor familien $\overline{\phi} = (\phi(i))_{i\in\mathcal{S}}$ er initial fordelingen i.e. ϕ er en vektor over alle mulige startsudfald og deres tilhørende ssh. Hvis $\mathcal{S}$ er endelig er $\overline{\phi}$ blot en rækkevektor.

Definition. (Transitionsmatricen) *For to stadier $i, j \in \mathcal{S}$ og et tidspunkt n er transitionssandsynligheden for at flytte fra i til j er*

$$P_{ij}(n) = \mathbb{P}(\mathcal{X}_{n+1} = j|\mathcal{X}_n = i).$$

*Hvis $P_{ij}(n) = P_{ij}(m)$ for alle $n, m \in \mathbb{N}_0$ så kaldes $(\mathcal{X}_n)_{n\in\mathbb{N}_0}$ **tidshomogent (definition 1)**. Den tilhørende matrice $P = (P_{ij})_{i,j\in\mathcal{S}}$ er kaldet transitionsmatricen.*

Bemærkning. En transitionsmatrice karrakteriserer et transitions diagram og omvendt. *Eksempel til figur 1*

$$P = \begin{pmatrix} 2/3 & 1/3 \\ 1/2 & 1/2 \end{pmatrix}.$$

Theorem 2. (Fordeling af $\mathcal{X}_n$) *For en Markov kæde på $\mathcal{S}$ med initial fordeling ϕ og transitionsmatrix P er*

$$(\mathbb{P}(X_n = i))_{i\in\mathcal{S}} = \overline{\phi}P^n$$

14.2 Classification of states

Definition 4. (Communication classes) *Lad $i, j \in \mathcal{S}$ være to stadier. Vi definere $i \longrightarrow j$ dvs. j er **accesible** fra i, hvis der eksisterer et $n \in \mathbb{N}$ så $P_{ij}^n > 0$. Samt $i \longleftrightarrow j$ dvs. i og j kommunikerer, hvis både $i \longrightarrow j$ og $j \longrightarrow i$.*

Bemærkning. "$i \longleftrightarrow j$" er en ækvivalens relation. Derfor sepererer denne relation $\mathcal{S}$ ind i to disjunkte ækvivalens klasser, hvad vi definerer som *communication classes*.

Definition 6. (Lukket kommunikations klasser) *En kommunikations klasse $e \subseteq \mathcal{S}$ kaldes lukket, hvis for alle $i \in e$ gælder $\sum_{i\in e} P_{ij} = 1$ (den i'te søjlesummen sum er lig 1 eller ssh. for at forblive i klassen er 1).*

Definition. (Ikke reducibel Markov kæde) *En Markov kæde kaldes irreducible, hvis der eksisterer kun en kommunikations klasse. Ellers er den reducible.*

Definition. (Hitting time) *Tiden for at ramme $i \in \mathcal{S}$ er givet ved $T_i := \inf\{n > 0 | X_n = i\}$. Hvis $X_0 = i$ kaldes denne tid **return/recurrence** tiden.*

Definition 9. (Recurrance/transients) *For en Markov kæde $(X_n)_{n\in\mathbb{N}}$ på $\mathcal{S}$ er et stadie $i \in \mathcal{S}$ kaldet **recurrent** hvis $\mathbb{P}(T_i < \infty | X_0 = i) = 1$. Ellers kaldes i **transient**.*

Bemærkning. Det kan vises at $\mathbb{P}(T_i < \infty | X_0 = i) = 1$ er ensbetydende med at $\mathbb{P}(N_i = +\infty | X_0 = i) = 1$. (se Thm 14 i noterne)

Theorem 11. (Recurrence criterium 1) *For en Markov kæde på $\mathcal{S}$ med transistionsmatrice P, da er følgende ækvivalent. Det holder at $\sum_{n=1}^{\infty}(P^n)_{ii} = +\infty$ og i er recurrent.*

Theorem 12. (Recurrence som klasseegenskab) *Alle stadier i en kommunikations klasse er enten recurrent eller transient dvs. hvis blot en af stadierne $i \in e$ er recurrent er alle recurrent og ligeledes med transient.*

Theorem 13. (Antal af besøg) *Antal af besøg er givet ved den stokastiske variabel $N_i := \sum_{n=1}^{\infty} 1(X_n = i)$. Hvis i er recurrent er $\mathbb{P}(N_i = +\infty) = 1$, hvis i er transient er $N_i \sim Geo(q)$ på N_0 dvs. $\mathbb{P}(N_i = k) = (1-q)^k q$ for $k \in \mathbb{N}_0$ og $q = P(T_i = +\infty | X(0) = i)$.*

Theorem 14. *For en endelig kommunikationsklasse $C \subseteq S$ gælder: (C er recurrent) $\Leftrightarrow$ (C er lukket.)*

Theorem 16. (Recurrence criterium 2) *For en irreducibel Markov Kæde på S gælder ($i \in S$ er recurrent) $\Leftrightarrow$ (For ligningssystemmet $\alpha(j) = \sum_{k\neq i} P_{j,k}\alpha(k)$ gælder $\alpha(j) = 0$ for alle $j \neq i$ er den eneste endelige løsning).*

Generaliseret matrix multiplikation. For $\underline{\phi} = (\phi_i)_{i\in\mathcal{S}}, \underline{\psi} = (\psi_i)_{i\in\mathcal{S}}$ *(vektorer)* og $P = (P_{ij})_{i,j\in\mathcal{S}}, Q = (Q_{ij})_{i,j\in\mathcal{S}}$ *(matricer)* kan følgende produkter udregnes ved

i. $(\underline{\phi} \cdot \underline{\psi}^T) := \sum_{i\in\mathcal{S}} \phi_i\psi_i$
ii. $(\underline{\phi} \cdot \overline{P})_j := (\sum_{i\in\mathcal{S}} \phi_i P_{ij})_j$
iii. $(P \cdot Q)_{ij} := \sum_{k,l\in\mathcal{S}} P_{i,k}Q_{l,j}$

14.3 Limit results and invariant probabilities

Sidenote. For en vektor $\overline{\phi} = (\phi_i)_{i\in\mathcal{S}}$ kaldes en fordeling, hvis for alle $i \in \mathcal{S}$ er $\phi_i \geqslant 0$ og $\sum_{i\in\mathcal{S}} \phi_i = 1$ og et mål hvis kun $\phi_i \geqslant 0$.

Definition 18. (Løkker og perioder) *For en Markov kæde i diskret tid er en mulig løkke af længde n er en følge af stadier $i_0, i_1, i_2, ..., i_n \in \mathcal{S}$ med $i_0 = i_n$ hvor*

$$P_{i_0,i_1} \cdots P_{i_{n-1},i_n} > 0$$

Perioden af et stadie $i \in \mathcal{S}$ er den største fælles divisor af

$$D_i = \{n \in \mathbb{N} \mid \exists \text{a possible loop of length } n \text{ med } i_0 = i_n = i\},$$

*dvs. $per(i) = GCD(D_i)$. Hvis perioden er 1 kaldes stadiet **aperiodisk**.*

Theorem 19. *Alle stadier i en kommunikations klasse har samme periode.*

Theorem 21. *For en irreducibel, recurrent og ikkeperiodisk Markov kæde på $\mathcal{S}$ gælder for alle $i \in \mathcal{S}$ at*

$$\lim_{n\to\infty} P(X_n = i) = \frac{1}{E[T_i|X_0 = i]}$$

Hvis $E[T_i|X_0 = i] = \infty$ defineres $\lim_{n\to\infty} P(X_n = i) := 0$.

Bemærkning. 1) Kun hvis Markov kæden er recurrent og dermed at $P(T_i = \infty|X_i = 0) = 0$ er $E[T_i|X_0 = i] = \sum_{n=1}^{\infty} nP(T_i = n|X_0 = i)$ eller er $E[T_i|X_0 = i] = +\infty$. 2) Hvis $E[T_i|X_0 = i] = \infty$ så er grænsen ikke en fordeling.

Definition 22. *Et recurrent stadie $i \in \mathcal{S}$ kaldes positivt recurrent hvis og kun hvis $E[T_i|X_0 = i] < \infty$. Ellers kaldes stadiet nul-recurrent.*

Definition. (Invariant fordeling og mål) *En ikke negativ vektor $\overline{\pi} = (\pi_j)_{j\in S}$ kaldes et invariant mål, hvis $\overline{\pi} = \overline{\pi}P$ og en invariant fordeling, hvis $|\overline{\pi}|$.*

Theorem 23. *For en irreducibel og recurrent Markov kæde på $\mathcal{S}$ findes et invariant mål ν der opfylder $\nu = \nu P$. Specielt gælder for alle fixed $i \in \mathcal{S}$ holder det at*

$$\nu_j = E\left[\sum_{n=0}^{T_i-1} 1(X_n = j|X_0 = i)\right] \text{ for alle } j \in \mathcal{S}$$

er et invariant mål. Hvis og kun hvis Markov kæden er positiv recurrent kan ovenstående nomaliseres til en invariant fordeling.

Bemærkning. 1) $i \in \mathcal{S}$ er abitrær, 2) $\nu(i) = 1$ og 3) for irreducible MC gælder (Positiv recurrent) $\Leftrightarrow$ (Eksisterer unik invariant fordeling)

Theorem 24. *or en irreducibel, aperiodisk og positiv recurrent Markov kæde gælder $\lim_{n\to\infty} P(X_n = j) = \pi_j = 1/E[T_j|X_0 = j]$. Hvor $\overline{\pi} = (\pi_j)_{j\in S}$ er en invariant fordeling.*

Theorem 25 og 26. (Grænseresultat for nul-recurrance og tranience) *For et nul-recurrent (THM 25) eller transient (THM 26) stadie $j \in S$ gælder $\lim_{n\to\infty} P(X_n = j) = 0$ for alle valg af initialfordeling $\overline{\phi}$.*

Theorem 27. (Grænseresultat for periodiske stadier) *For en irreducibel Markov kæde med periode $d > 1$ findes grænsen $\lim_{n\to\infty} P(X_n = i)$ ikke. Men gennemsnittet for en periode gør dvs*

$$\nu_j = \lim_{n\to\infty} \frac{P(X_n = i) + P(X_{n+1} = i) + ... + P(X_{n+d-1} = i)}{d}$$

$\nu = (\nu_j)_{j\in S}$ er et invariant mål og en invariant fordeling hvis $|\nu| = 1$.

14.4 Absorbing probabilities

Theorem 29. (Absorberende sandsynlighed - endelige udfaldsrum) *Lad en endelig Markov kæde være givet ved transistionsmatricen P. Antaget at transtionsmatricen kan inddeles efter kummunikations klasser så*

$$P = \left[\begin{array}{c|c} \tilde{P} & 0 \\ \hline S & Q \end{array}\right],$$

hvor $\tilde{P}$ er transistionsmatricen indeholdende kun recurrent stadier. Dertil er Q og S sub-matricer af P, hvor Q indeholder de transient stadier og S beskriver sandsynlighederne for overgangen fra et transient til et recurrent stadie. Matricen 0 repræsenterer sandsynligheden for at gå fra en recurrent til transient stadie, hvad er umuligt.

Definer dertil matricerne $M = (I - Q)^{-1}$ og $A = (I - Q)^{-1}S$. Tallet $M_{i,j}$ repræsenterer $E[N_j|X_0 = i]$ for transient stadier i, j. Tallet $A_{i,j}$ repræsenterer sandsynligheden for at j er det første recurrent stadie, der besøger hvis $X_0 = i$ for et transient stadie i.

Theorem 31. (Absorberende sandsynlighed - tællelige udfaldsrum) *For en Markov kæde på S lad $C \subseteq S$ være en recurrent klasse og $C' \in S\backslash C$ være en transient klasse. Absorberingssandsynligheden $(\alpha_j)_{j\in C'}$ givet ved $\alpha_j = P(X_n \in C|X_0 = j)$ løser ligningssystemet*

$$\alpha_i = \sum_{l\in S\backslash C} P_{i,l}\alpha_l + \sum_{l\in C} P_{i,l}$$

Der findes en endelig løsning hvis og kun hvis $\lim_{n\to\infty} P(X_n \in C'|X_o \in C') = 0$.

14.5 Markov Chains in Continuous Times

Definition. (Stokastisk process i kontinuer tid) *En stokastisk proces i kontinuer tid er en familie $X = (X_t)_{t\in[0,\infty)}$ af stokastiske variable.*

Eksempel 33. *(Poisson processen)* Lad $(W_i)_{i\in\mathbb{N}}$ være en følge af uafhængige stokastiske variable, hvor $W_i \sim Exp(\lambda)$ dvs. W_i har tæthed $f(x) = \lambda\exp(-\lambda x)$. Betragt W_i som ventetiden indtil en begivenhed af interesse indtræffer. Lad $\tau_n := W_1 + ... + W_n$ være tiden indtil den n'te begivenhed indtræffer. Den stokastiske variabel $N_t = \sum_{n=1}^{\infty} 1(\tau_n \leqslant t)$ der tæller antal begivenheder indtil tidspunkt t definerer en stokastisk proces i kontinuer tid dvs. processen $(N_t)_{t\geqslant 0}$. Vi kalder denne proces *Poisson processen*.

Definition 34. (Homogen Markov kæde i kontinuer tid) *En kontinuer Markov kæde på en højst tællelig mængde S er en familie af stokastiske variable $(X_t)_{t\geqslant 0}$ på sandsynlighedsrummet $(\Omega, \mathcal{F}, P)$ der opfylder*

$$\begin{aligned} &P(X(t_{n+1}) = j|X(t_n) = i, X(t_{n-1}) = i_{n-1}, ..., X(t_0) = i_0) \\ &= P(X(t_{n+1}) = j|X(t_n) = i) = P_{i,j}(t_{n+1} - t_n) \end{aligned}$$

for $j, i, i_{n-1}, ..., i_0 \in S$ og $t_{n+1} > t_n > ... > t_0 \geqslant 0$. Fordelingen af Markov kæden er givet ved initialfordelingen $\overline{\phi} = (P(X_0 = i))_{i\in S}$ og transistionssandsynlighederne $P_{i,j}(t) = P(X(t + s) = j|X(s) = i)$ og identiteten

$$\begin{aligned} &P(X(t_{n+1}) = j, X(t_n) = i, X(t_{n-1}) = i_{n-1}, ..., X(t_0) = i_0) \\ &= P_{i,j}(t_{n+1} - t_n) \cdot ... \cdot P_{i_0,i_1}(t_1 - t_0)\phi(i_0) \end{aligned}$$

Definition 35. (Minimal konstruktion) *Lad $\overline{\phi} = (\phi_i)_{i\in S}$ være en sandsynlighedsvektor og lad $Q = (q_{i,j})_{i,j\in S}$ være en intensitetsmatrice dvs. $q_{i,j}$ er reelle tal med egenskaberne 1) alle ikke diagonal indgange er ikke negative dvs. $q_{i,j} \geqslant 0$ for alle $i, j \in S$ og $i \neq j$ og 2) diagonal er negativ lig rækkesummen dvs. $q_{i,i} = -\sum_{j\neq i} q_{i,j}$.*

Givet $\overline{\phi}$ og Q kan en tids-homogen Markov kæde konstrueres ved følgende fem trin. Matricen Q kaldes transistionsintensiteten for den stokastiske proces $(X_t)_{t\geqslant 0}$.

1. Vælg $\gamma(0)$ ud fra initialfordelingen, så $P(\gamma(0) = i) = \phi(i)$.
2. givet $\gamma(0)$ sæt $\tau_1 := W_1 \sim Exp(q_{\gamma(0)})$ og definer $X(t) = \gamma(0), t \in [0, W_1)$
3. givet $\gamma(0)$ og W_1 vælg $\gamma(1)$ sådan at $P(\gamma(1)|\gamma(0)) = \frac{q_{\gamma(0),i}}{q_{\gamma(0)}}, i \neq \gamma(0)$
4. recursivt givet $\gamma(0), ..., \gamma(n), W_1, ..., W_n$ vælg $W_{n+1} \sim Exp(q_{\gamma(n)}$. Lad $\tau_{n+1} = \tau_n + W_{n+1}$ og definer $X(t) = \gamma(n), t \in [\tau_n, \tau_{n+1})$
5. vælg $\gamma(n+1)$ sådan at $P(\gamma(n+1)|\gamma(0), ..., \gamma(n), W_1, ..., W_n) = \frac{q_{\gamma(n),i}}{q_{\gamma(n)}}, i \neq Y(n)$.

Definition 37. (Absorbering) *Hvis Markov kæden givet ved minimal konstruktion på tidspunkt τ_n hopper til stadie $\gamma(n) = i$ med $q_i = 0$ så lader vi $X(t) = \gamma(n)$ for $t \geqslant \tau_n$ og vi siger at Markov kæden er absorberet på stadie i.*

Definition 38. (Eksplosion) *Hvis Markov kæden hopper uendeligt mange gange på et endeligt interval dvs. $\tau_\infty := \lim_{n\to\infty} \tau_n < +\infty$ og dermed $P(\tau_\infty < +\infty) > 0$. Lader vi $X(t) = \Delta$ for $t \geqslant \tau_\infty$. Vi kalder tidspunktet τ_∞ for eksplosionstidspunktet.*

Theorem 39. (Embedded Markov Chain of jumps) *For en markov kæde i kontinuert tid med transitions intensiteter (intensitetsmatrice) $Q = (q_{i,j})_{i,j\in s}$ givet ved den minimale konstruktion fra definition 35, er følgen $(\Gamma(n))_{n\in\mathbb{N}_0}$ af besøgte stadier en markov kæde i diskret tid med transitionsmatricen P givet ved*

$$\begin{cases} -\frac{q_{i,j}}{q_{i,i}} = \frac{q_{i,j}}{q_i} & i \in S\backslash A, j \notin \{i, \Delta\} \\ 0 & i \in S\backslash A, j \in \{i, \Delta\} \\ 0 & i \in A, j \neq i \\ 1 & i \in A, j = i \end{cases}$$

Hvor $A = \{i \in S | q_i = 0\}$ er delmængden af absorberende stadier.

Bemærkning: Det er klart, at de enkelte indgange $(P_{i,j}(t))_{i,j\in S}$ $(t \geqslant 0)$ i transistionsmatricen må være ikke negative. Endvidere er rækkesummerne, for en markov kæde hvor eksplosion ikke er mulig, lig 1. Specielt er transitionssandsynlighederne for et valgt $t \geqslant 0$ i overenstemmelse med dem som vi har set i kapitel 2. For skift i tid, kræves næste teorem.

14.6 Properties of transitionsprobabilities

Theorem 42. (Chapman-Kolmogorov ligninger) *Transitionssandsynlighederne for en homogen markov kæde i kontinuert tid opfylder Chapman-Kolmogorov ligningerne*

$$\forall s, t \geqslant 0, \forall i, j \in S : P_{i,j}(t+s) = \sum_{l\in S} P_{i,l}(t) \cdot P_{l,j}(s)$$

Hvis state space er endeligt, så kan $P(t) = (P_{i,j}(t))_{i,j\in S}$ ses som en matrice for hvilket som helst valgt $T \geqslant 0$ og så kan Chapman-Kolmogorov ligningerne skrives som matrice ligningen

$$P(t+s) = P(t)P(s)$$

Theorem 43. (Infinitesimal generatoren af en markov kæde) *For en markov kæde i kontinuert tid, kan transitionsintensiteterne udledes fra transitionssandsynlighederne som grænserne*

$$\lim_{t\to 0+} \frac{P_{i,i} - 1}{t} = -q_i$$
$$\lim_{t\to 0+} \frac{P_{i,j}}{t} = q_{i,j}, \;\; i \neq j$$

Theorem 44. (Backward differential equations) *For en markov kæde i kontinuert tid holder det altid, at*

$$DP_{i,j}(t) = P'_{i,j}(t) = -q_i P_{i,j}(t) + \sum_{k\neq i} q_{i,k} P_{k,j}(t)$$

Theorem 45. (Forward differential and integral equations) *For en markov kæde i kontinuert tid, holder det, at*

$$P_{i,j}(t) = \delta_{i,j}\exp(-q_j t) + \int_0^t \sum_{l\neq j} P_{i,l}(u)q_{l,j}\exp(-q_j(t-u))du$$

og at

$$DP_{i,j}(t) = -P_{i,j}(t)q_j + \sum_{l\neq j} P_{i,l}(t)q_{l,j}.$$

Theorem 47. (Transitionssandssynlighederne for et endeligt state space) *For en markov kæde i kontinuert tid med endeligt state space, kan den "backward differential equation" udtrykkes i matrice form som*

$$DP(t) = P'(t) = QP(t)$$

Hvor $P(t) = (P_{i,j}(t))_{i,j\in S}$ Ved at bruge startbetingelsen $P(0) = I$, kan transistionssandsynlighederne udtrykkes på formen af eksponentiel matricen

$$P(t) = \exp(Qt), \;\; t \geqslant 0.$$

14.7 Invariant probabilies and absorption

Definition 49. (Kommunikationsklasser og "irreducibility") *To stadier $i, j \in S$ siges at kommunikere, hvis der findes $s, t > 0$ således at*

$$P_{i,j}(s) > 0 \text{ og } P_{j,i}(t) > 0.$$

Denne definition inddeler S i disjunkte kommunikationsklasser, ligesom vi har set i kapitel 2. En markov kæde i kontinuert tid siges at være "irreducible" hvis der kun findes én kommunikationsklasse.

Bemærkning: Man kan anvende, at to stadier $i, j \in S$, hvor $i \neq j$ kommunikerer, hvis og kun hvis, der eksisterer en følge af stadier $i_1, i_2, ..., i_n \in S$ (som indeholder stadie j) således, at $q_{i,i_1}\cdot q_{i_1,i_2}\cdot ...\cdot q_{i_{n-1},i_n}\cdot q_{i_n,i} > 0$. (Er der en mulig sti frem og tilbage mellem i og j?)

Definition 50. (Recurrence og transience)

1. *En irreducible markov kæde i kontinuert tid er recurrent, hvis og kun hvis, den embedded markov kæde (se teorem. 39) er recurrent. Den er transient, hvis og kun hvis, den embedded markov kæde er transient.*
2. *Stadiet i er transient, hvis og kun hvis, den totale tid tilbragt i stadiet i*

$$V_i = \int_0^\infty 1_{X(t)=i}dt$$

er endelig med sandsynlighed 1, altså $P(V_i < +\infty | X(0) = i) = 1$. Tilsvarende er stadiet i recurrent hvis $P(V_i = +\infty | X(0) = i) = 1$.

Bemærkning: Denne definition kan anvendes på de enkelte kommunikationsklasser. Specielt er et absorberende stadie altid recurrent.

Definition 51. (Invariant fordeling) *En sandsynlighedsvektor $\overline{\pi} = (\pi(i))_{i\in S}$ er en invariant (eller stationær) fordeling for en markov kæde i kontinuert tid, hvis for alle $t \geqslant 0$ og $j \in S$*

$$\pi(j) = \sum_{i\in S} \pi(i)P_{i,j}(t).$$

Theorem 52. (Entydigheden af den invariante fordeling) *Overvej en markov kæde i kontinuert tid. En invariant fordeling er unik, hvis den eksisterer. Hvis der for et $t_0 > 0$ findes en sandsynlighed $\overline{\pi} = (\pi(i))_{i\in S}$ sådan, at*

$$\forall j \in S : \pi(j) = \sum_{i \in S} \pi(i) P_{i,j}(t_0)$$

kan vi konkludere, at

1. $\forall i \in S : \pi(i) > 0$
2. $P(t_0)$ *er en overgangssandsynlighed, dvs.*

$$\forall i \in S : \sum_{j \in S} \pi(i) P_{i,j}(t_0) = 1$$

3. $\overline{\pi}$ *er en invariant fordeling for markov kæden, dvs.*

$$\forall t \geqslant 0, \forall j \in S : \pi(j) = \sum_{j \in S} P_{i,j}(t_0) = 1$$

Bemærkning: Det følger af (3), at hvis vi kan finde en invariant fordeling for et bestemt $t_0 > 0$, så gælder denne for alle $t > 0$. Det følger også, som resultat af dette, at alle rækkesummerne $\sum_{j \in S} P_{i,j}(h) = 1$ for alle $h \geqslant 0$, hvor en invariant fordeling kan findes.

Theorem 53. (Grænseresultater for overganssandsynligheder) *For en irreducible markov kæde i kontinuert tid, med en invariant fordeling $\overline{\pi}$, gælder det for alle $i, j \in S$, at*

$$\lim_{t \to \infty} P_{i,j}(t) = \pi(j).$$

Endvidere gælder det for enhver initial fordeling $\overline{\phi}$, at

$$\lim_{t \to \infty} P(X(t) = j) = \pi(j).$$

Hvis der ikke findes en invariant fordeling, så gælder der

$$\lim_{t \to \infty} P_{i,j}(t) = 0.$$

Theorem 55. (Nødvendig betingelse for invariant fordeling) *For en markov kæde i kontinuert tid, er det nødvendigt, at den invariante fordeling $\overline{\pi}$ passer med ligningsystemet*

$$\forall j \in s : \sum_{i \in S} \pi(i) q_{i,j} = 0$$

Eller tilsvarende

$$\forall j \in S : \sum_{i \neq j} \pi(i) q_{i,j} = \pi(j)(-q_{j,j}) = \pi(j) q_j$$

Tænker man på $\overline{\pi}$ som en række vektor og Q som en matrice, har ligningssystemet en mere kompakt formulering

$$\overline{\pi} Q = 0.$$

Theorem 56. (Tilstrækkelig betingelse for en invariant fordeling) *Hvis $\overline{\pi}$ overholder betingelsen i teorem 55 og endvidere overholder*

$$\sum_{j \in S} \pi(j) q_j < \infty$$

Så er $\overline{\pi}$ en unik invariant fordeling for en irreducible markov kæde.

Theorem 58. *En irreducible markov kæde i kontinuert tid har en invariant fordeling, hvis og kun hvis, den indlejrede (embedded) markov kæde er recurrent og der ekstisterer en sandsynlighedsvektor $\overline{\pi}$ således, at teorem 55 er overholdt. ($\overline{\pi} Q = 0$)*

Theorem 60. (Tids-invariant vs. hændelses-invariant fordeling) *Antag, at der for en irreducible markov kæde i kontinuert tid findes en invariant fordeling $\overline{\nu}$. Hvis vi også har bekræftet eksistensen af en invariant fordeling for den indlejrede markov kæde, så gælder følgende forhold mellem dem.*

$$\pi(i) = \frac{\nu(i)q_i}{\sum_{j\in S}\nu(j)q_j}, \ \ i \in S.$$

Theorem 61. (Eksistens af invariant fordeling og positive recurrence) *For en irreducible og recurrent markov kæde i kontinuert tid definerer vi "escape time" fra stadiet i som*

$$W_i = \inf\{t > 0 | X(t) \neq i\}$$

Og "return time" til stadiet i som

$$R_i = \inf\{t > W_i | X(t) = i\}$$

Og der gælder, at

$$\pi(i) = \frac{E[W_i|X(0) = i]}{E[R_i|X(0) = i]} = \frac{1}{q_i E[R_i|X(0) = i]}$$

Vi siger så, at en kommunikationsklasse er "positive recurrent" hvis

$$E[R_i|X(0) = i] < +\infty$$

Bemærkning: Positive recurrence er en klasseegenskab.

Theorem 62. (Tid tilbragt i stadiet j før absorbering) *For en markov kæde i kontinuert tid findes det gennemsnitlige antal af besøg til stadiet j før det absorberende stadie i rammes, ved at betragte overgangssandsynlighederne af den indlejrede markov kæde. For et endeligt state space, kan man bruge teorem 29, mens man kan bruge teorem 31 for tælleligt uendelige state spaces.*

Hvis N_j er middelværdien af antal besøg til stadiet j før absorbering i stadiet i, så er middelværdien af tid tilbragt i stadiet j lig $\frac{N_j}{q_j}$.

Side note: Eksempel 63 har været meget brugbar til at forstå kapitel 3.3.

14.8 Birth-death processes

Definition. (Birth-and-death process) *En birth-and-death proces er en markov kæde på $S = \mathbb{N}_0$ som kun tillader hop (op eller ned) af størrelset ét. Dvs, at for overgangsintesiterne, gælder $q_{i,j} = 0, i, j \in \mathbb{N}_0, |i-j| > 1$, mens de eneste ikke-nul indgange (udover diagonalen) er lig*

$$q_{i,i+1} = \beta_i, i \in \mathbb{N}_0 \rightarrow \text{ birth intensitet}$$
$$q_{i,i+1} = \delta_i, i \in \mathbb{N} \rightarrow \text{ death intensitet}$$

Adfærden af sådan en proces er meget simpelt at beskrive. Hvis man befinder sig i stadiet i, så er ventetiden til det næste hop eksponentiel fordelt med parametren $\beta_i + \delta_i$ og med gennemsnitstiden $\frac{1}{\beta_i+\delta_i}$. Når kæden hopper, kan den hoppe op med sandsynlighed $\frac{\beta_i}{\beta_i+\delta_i}$ eller hoppe ned med sandsynlighed $\frac{\delta_i}{\beta_i+\delta_i}$.

Side note: Meget af kapitel 3.4 er eksempler på birth-and-death processer.

Theorem 66. (Birth-and-death processer: recurrence) *En birth-and-death proces er recurrent, hvis og kun hvis,*

$$\sum_{i=1}^{\infty} \frac{\delta_i \cdot \ldots \cdot \delta_1}{\beta_i \cdot \ldots \cdot \beta_1} = \infty.$$

Ækvivalent er en birth-and-death proces transient, hvis og kun hvis,

$$\sum_{i=1}^{\infty} \frac{\delta_i \cdot \ldots \cdot \delta_1}{\beta_i \cdot \ldots \cdot \beta_1} < \infty.$$

Theorem 67. (Birth-and-death processer: positive recurrence) *En birth-and-death proces er positive recurrent, hvis og kun hvis,*

$$\sum_{i=1}^{\infty} \frac{\beta_{i-1} \cdot ... \cdot \beta_0}{\delta_i \cdot ... \cdot \delta_1} < \infty \quad \text{og} \quad \sum_{i=1}^{\infty} \frac{\delta_i \cdot ... \cdot \delta_1}{\beta_i \cdot ... \cdot \beta_1} = \infty$$

Theorem 68. (Eksplosion for en birth-and-death proces) *For en birth-and-death proces med intensiteterne*

$$q_{i,i+1} = \beta_i, \quad q_{i+1,i} = \delta_{i+1}, \quad q_{i+1,i+1} = -(\delta_{i+1} + \beta_{i+1})$$

og med $q_{i,j} = 0$ ellers, $i, j \in \mathbb{N}_0$, så er eksplosion muligt, hvis og kun hvis,

$$\sum_{i=1}^{\infty} \left(\frac{1}{\beta_i} + \frac{\delta_i}{\beta_i \beta_{i-1}} + ... + \frac{\delta_i \cdot ... \cdot \delta_1}{\beta_i \cdot ... \cdot \beta_0} \right) < \infty$$

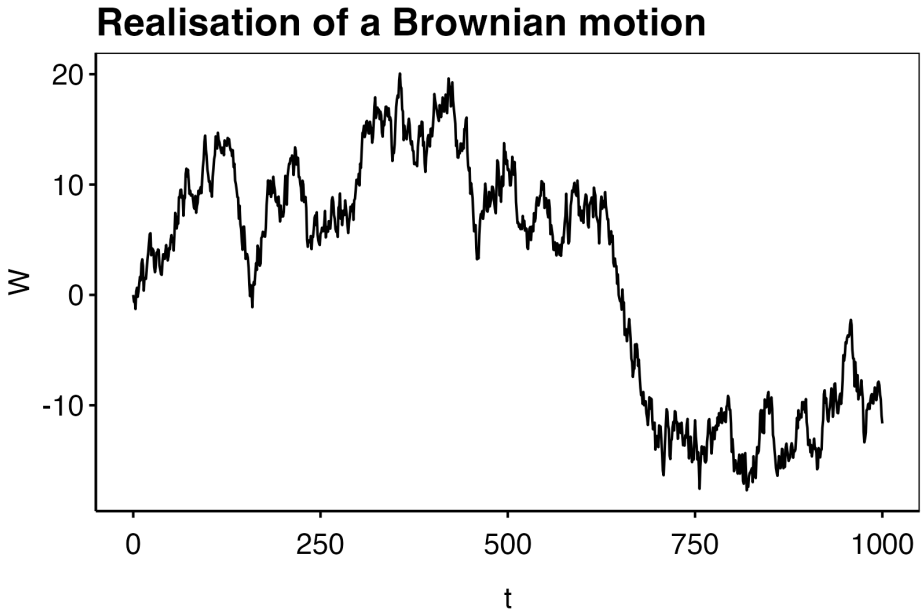
Chapter 15

Continuous Time Stochastic Processes

15.1 Brownian Motion

Definition 4.1. (Bjork) *A stochastic process W is called a* ***Brownian motion*** *or* ***Wiener process*** *if the following conditions hold*

1. $W_0 = 0$.
2. *The process W has independent increments, i.e. if $r < s \leqslant t < u$ then $W_u - W_t$ and $W_s - W_r$ are independent random variables.*
3. *For $s < t$ the random variable $W_t - W_s$ has the Gaussian distribution $\mathcal{N}(0, t-s)$.*
4. *W has continuous trajectories i.e. $s \mapsto W(s;\omega)$ i continuous for all $\omega \in \Omega$.*



As one can see from the simulated sample path on the right, the Brownian motion is rather irratic. In fact, the process varies infinitely on any interval with length greater than 0. This gives some of the characteristics of the process including that: W is continuous and W is non-differential everywhere. This irratic behaviour is summed up in the theorem.

Theorem 4.2. (Bjork) *A Brownian motions trajectory $t \mapsto W_t$ is with probability one nowhere differential,*

and it has locally infinite total variation.

15.2 Filtration

Filtrations is widely used in stochastic processes, as they allow for the concept of knowledge/information. This is useful when considering mean-values of future states but in an increasing information setting. For this we introduce the term adapted processes.

Definition B.17. (Bjork) (Adapted process) *Let $(\mathcal{F}_t)_{t\geqslant 0}$ be a filtration on the probability space $(\mathcal{F}_t)_{t\geqslant 0}$. Furthermore, let $(X_t)_{t\geqslant 0}$ be a stochastic process on the same space. We say that X_t is adapted to the filtration* **F** *if*

$$X_t \text{ is } \mathcal{F}_t - \text{measurable}, \qquad \forall t \geqslant 0.$$

Obviously, we may introduce the **natural filtration** $\mathcal{F}_t^X$ given by the tragetory of the process X_t:

$$\mathcal{F}_t^X = \sigma(\{X_s,\ s \leqslant t\}).$$

Indeed, X_t is adapted to this filtration.

15.3 Martingale

Definition. *Let M_t be a stochastic process defined on a background space $(\Omega, \mathcal{F}, P)$. Let $(\mathcal{F}_t)_{t\geqslant 0}$ be a filtration. If M_t is adapted to the filtration $\mathcal{F}_t$, $E|M_t| < \infty$ and*

$$E[M_t|\mathcal{F}_s] = M_s, \qquad P - \text{a.s.}$$

*holds for any $t > s$ we say that M_t is a martingale ($\mathbf{F}$-martingale). If the above has $\leqslant$ or $\geqslant$ we say that M_t is either a **submartingale** or **supermartingale** respectively.*

Naturally, this defintions may easily be extended to discrete models and we have the trivial equality:

$$E[M_t - M_s \mid \mathcal{F}_s] = 0.$$

Martingales is useful, when proofing probalistic statements as the posses tractable properties. A useful technique often include the construction of the martingale

$$M_t = E[X \mid \mathcal{F}_t].$$

Chapter 16

Stochastic calculus

16.1 Stochastic Integrals

We want to formulate financial markets in continuous time and the most elegant theory is obtained from processes that can be defined in terms of **stochastic differential equations** or in other words by their dynamics. We may call them **diffusion processes**, as they may be approximated by a stochastic difference equation:

$$X_{t+\Delta t} - X_t = \mu(t, X_t)\Delta t + \sigma(t, X_t)Z_t. \tag{4.1}$$

Above Z_t is a normally distributed random variable (a disturbance). In this formulation we say that S_t is driven by two forces: on one hand a locally deterministic velocity or drift $\mu(t, X_t)$ and on the other hand a Gaussian term amplified by the deterministic factor $\sigma(t, X_t)$.

16.1.1 Information

We consider a primary process X_t and we introduce the notion of information generated by X_t in terms of the natural filtration. The idea can be summed up in the following definition.

Definition 4.3. (Bjork) *The symbol $\mathcal{F}_t^X \subseteq \mathcal{F}$ denotes "the information generated by X_t on the interval $[0, t]$", or alternatively "what has happened to X_t over ther interval $[0, t]$".*

1. *If, based upon observations of the trajectory $\{X_s;\ 0 \leqslant s \leqslant t\}$, it is possible to decide whether a given event A has occurred or not, then we write this as*

$$A \in \mathcal{F}_t^X$$

   or say that "A is $\mathcal{F}_t^X$-measurable".
2. *If the value of a given random variable Z can be completely determined given observations of the tragectory $\{X_s;\ 0 \leqslant s \leqslant t\}$, then we also write*

$$Z \in \mathcal{F}_t^X.\ (Z \text{ is } \mathcal{F}_t^X\text{-measurable})$$

3._ If Y is a stochastic process such that we have_

$$Y_t \in \mathcal{F}_t^X$$

*for all $t \geqslant 0$ then we say that Y_t is **adapted** to the **filtration** $\{\mathcal{F}_t^X\}_{t\geqslant 0}$. For brevity of notation, we will sometimes write the filtration as $\{\mathcal{F}_t^X\}_{t\geqslant 0} = \mathbf{F}$.*

16.1.2 Stochastic Integrals

We will now formulate the theory of stochastic integrals, that is, processes written in terms of stochastic processes with stochastic integrator and/or stochastic integrant. We will consider some given standard Brownian motion W_t and another stochastic process X_t. We need som integrability condition on X_t in order to do the calculations. We therefore determine a selection of suitable stochastic processes X must be contained in.

Definition 4.4. (Bjork) *Let X_t be a stochastic process, then*

i. *We say that X_t belongs to the class $\mathcal{L}^2[a,b]$ if X_t is adapted to the filtration $\mathcal{F}_t^X$ and the following holds*

$$\int_a^b E[X_s^2]\ ds < \infty$$

ii. *We say that X_t belongs to the class $\mathcal{L}^2$ if $X_t \in \mathcal{L}^2[0,t]$* for all $t > 0$.

We now want to define what we mean by

$$\int_a^b X_t\ dW_s$$

for some process $X_t \in \mathcal{L}^2$. We see that a way to go about this problem is to start by defining the concept for a simple stochastic process X_t. By *simple* we mean a process X_t that is constant on between some deterministic points in time $a = t_0 < t_1 < \cdots < t_n = b$. In that case we may define the integral as

$$\int_a^b X_s\ dW_s = \sum_{k=0}^{n-1} X_{t_k}[W_{t_{k+1}} - W_{t_k}]. \tag{4.8}$$

In the more general setting we may follow the following approach:

1. Approximate X with a sequence $\{X^n\}_{n\in\mathbb{N}}$ of simple processes such that the following convergence criteria hold

$$\int_a^b E[(X_s^n - X_s)^2]\ ds \to 0,\ n \to \infty$$

2. For each n the integral $\int_a^b X_s^n\ dW_s := Z^n$ is well defined and it is possible to prove, using DCT, that a variable Z exists such that $Z^n \to Z$ that is in L^2.
3. We now define the stochastic integral by the limit

$$\int_a^b X_s\ dW_s = \lim_{n\to\infty} \int_a^b X_s^n\ dW_s. \tag{4.9}$$

Obviously the hardest step is finding the processes X^n. This stochastic har some properties we will use.

Proposition 4.5. (Bjork) *Let $X_t \in \mathcal{L}^2$, then*

$$E\left[\int_a^b X_s\ dW_s\right] = 0. \tag{4.12}$$

$$E\left[\left(\int_a^b X_s\ dW_s\right)^2\right] = \int_a^b E[X_s^2]\ dW_s. \tag{4.13}$$

$$\int_a^b X_s\ dW_s \ \text{ is } \mathcal{F}_b^W\text{-measurable.} \tag{4.14}$$

16.1.3 Martingales

Definition 4.7. (Bjork) *Let M_t be a stochastic process defined on a background space $(\Omega, \mathcal{F}, P)$. Let $(\mathcal{F}_t)_{t\geqslant 0}$ be a filtration. If M_t is adapted to the filtration $\mathcal{F}_t$, $E|M_t| < \infty$ and*

$$E[M_t|\mathcal{F}_s] = M_s, \qquad P - \text{a.s.}$$

*holds for any $t > s$ we say that M_t is a martingale ($\mathbf{F}$-martingale). If the above has $\geqslant$ or $\leqslant$ we say that M_t is either a **submartingale** or **supermartingale** respectively.*

Proposition 4.8. (Bjork) *For any process $X_t \in \mathcal{L}^2[s,t]$ the following hold:*

$$E\left[\int_s^t X_s \, dW_s \middle| \mathcal{F}_s^W\right] = 0$$

Corollary 4.9. (Bjork) *For any process $X_t \in \mathcal{L}^2$ the process*

$$M_t = \int_s^t X_s \, dW_s,$$

*is an $(\mathcal{F}_t^W)$-martingale. In other words, modulo an integrability condition, **every stochastic integral is a martingale**.*

Lemma 4.10. (Bjork) *Let M_t be a stochastic process with stochastic differential, then M_t is a martingale if and only if the stochastic differential has the form $dM_t = X_t \, dW_t$ i.e. M_t as no dt-term.*

16.1.4 Stochastic Calculus and the Ito Formula

Given this breif introduction to stochastic integrals we may formulate som simple calculus revolving around Ito's formula. We consider the stochastic process X_t and we suppose that there exist a real number X_0 and adapted processes μ and σ wrt. $\mathcal{F}_t^W$ such that for all $t \geqslant 0$ we have

$$X_t = X_0 + \int_0^t \mu_s \, ds + \int_0^t \sigma_s \, dW_s, \tag{4.16}$$

where W_t is a standard Brownian motion. We know from earlier courses that the above may be written in terms of the dynamics (pure notation):

$$\begin{cases} dX_t = \mu_t \, dt + \sigma_t \, dW_t, \\ X_0 = X_0. \end{cases} \tag{4.17/18}$$

Here we intepret the above as X_t has boundary condition X_0 and evolves with a drift $\mu_t \, dt$ amplified and distorted by the drift $\sigma_t \, dW_t$. We say that X_t has **stochastic differential** dX_t and initial condition X_0.

We want to understand how transformation of such an integral behaves and therefore we introduce some calculus which will tell how for instance $f(t, X_t)$ (for some $C^{1,2}$-function) behaves. This insight is given by the important Ito's formula.

Thoerem 4.11. (Bjork) (Ito's formula, one-dimensional) *Assume that the process X has a stochastic differential form given by*

$$dX_t = \mu_t \, dt + \sigma_t \, dW_t, \tag{4.28}$$

where μ and σ are adapted processes, and let $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ be a $C^{1,2}$-function. Define the process Z by $Z_t = f(t, X_t)$. Then Z has stochastic differential given by

$$df(t, X_t) = \left(\frac{\partial f}{\partial t}(t, X_t) + \mu_t \frac{\partial f}{\partial x}(t, X_t) + \frac{1}{2}\sigma_t^2 \frac{\partial^2 f}{\partial x^2}(t, X_t)\right) \, dt + \sigma_t \frac{\partial f}{\partial x}(t, X_t) \, dW_t. \tag{4.29}$$

Proposition 4.12. (Bjork) (Ito's formula, one-dimensional) *With assumptions as in theorem 4.11, df is given by*

$$df = \frac{\partial f}{\partial t} \, dt + \frac{\partial f}{\partial x} \, dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2} \, (dX_t)^2, \tag{4.31}$$

where we use the following table

$$\begin{cases} (dt)^2 = 0, \\ dt \cdot dW_t = 0, \\ (dW_t)^2 = dt. \end{cases}$$

Lemma 4.18. (Bjork) *Let $\sigma(t)$ be deterministic function of time and define the process X by*

$$X_t = \int_0^t \sigma(s)\ dW_s. \tag{4.37}$$

Then

$$X_t \sim \mathcal{N}\left(0, \int_0^t \sigma^2(s)\ ds\right).$$

16.1.5 The multidimensional Ito Formula

Consider a vector process $X = (X^1, ..., X^n)^\top$ where each component X^i has stochastic differential

$$dX_t^i = \mu_t^i\ dt + \sum_{j=1}^d \sigma_t^{ij}\ dW_t^j$$

where $W^1, ..., W^d$ is independent Brownian motions. Then we have respectively the drift vector process μ_t in n dimensions, the vector Brownian motion in d dimensions and a $n \times d$-dimensional **diffusion matrix** σ_t given as below

$$\mu_t = \begin{bmatrix} \mu_t^1 \\ \vdots \\ \mu_t^n \end{bmatrix}, \quad W_t = \begin{bmatrix} W_t^1 \\ \vdots \\ W_t^d \end{bmatrix}, \quad \sigma_t = \begin{bmatrix} \sigma_t^{11} & \cdots & \sigma_t^{1d} \\ \vdots & \ddots & \vdots \\ \sigma_t^{n1} & \cdots & \sigma_t^{nd} \end{bmatrix}.$$

Given this we may write the dynamics of X as

$$dX_t = \mu_t\ dt + \sigma_t\ dW_t \in \mathbb{R}^n.$$

Consider now a function $f : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ which is a $C^{1,2}$-mapping. We want to study the dynamics of the process

$$Z_t = f(t, X_t).$$

The dynamics is given in the multidimensional version of Ito's formula.

Thoerem 4.19. (Bjork) (Ito's formula, multi-dimensional) *Let X be given as above. Then the following holds:*

- *The process $f(t, X_t)$ has the stochastisc differential given by*

$$df(t, X_t) = \left(\frac{\partial f}{\partial t}(t, X_t) + \sum_{i=1}^n \mu_t^i \frac{\partial f}{\partial x^i}(t, X_t) + \frac{1}{2}\sum_{i,j=1}^n C_t^{ij} \frac{\partial^2 f}{\partial x^i \partial x^j}(t, X_t)\right)\ dt + \sum_{i=1}^n \sigma_t^i \frac{\partial f}{\partial x^i}(t, X_t)\ dW_t.$$

  Here the row vector σ_t^i is the i'th row of the matrix σ_t and the matrix C is defined by $C = \sigma\sigma^\top$.
- *Alternatively, the differential is given by the formula*

$$df(t, X_t) = \frac{\partial f}{\partial t}(t, X_t)\ dt + \sum_{i=1}^n \frac{\partial f}{\partial x^i}(t, X_t)\ dX_t^i + \frac{1}{2}\sum_{i,j=1}^n \frac{\partial^2 f}{\partial x^i \partial x^j}(t, X_t)\ dX_t^i dX_t^j,$$

  with the formal multiplication table

$$\begin{cases} (dt)^2 = 0, & \\ dt \cdot dW_t^i = 0, & i = 1, ..., d, \\ (dW_t^i)^2 = dt, & i = 1, ..., d, \\ dW_t^i \cdot dW_t^i = 0, & i \neq j. \end{cases}$$

Obviously, one can write the differential in Ito's formula in many other ways including a matrix-wise version using the Hessian matrix $H_{ij} = \frac{\partial^2 f}{\partial x^i \partial x^j}$.

16.1.6 Correlated Brownian motions

In the previous section the d-dimensional Brownian was assumed to have independent Brownian motions. However we may instead consider a variation where we have some dependence between the Brownian motions.

This section has not been finished.

16.2 Discrete Stochastic Integrals

This section has not been finished.

16.3 Stochastic Differential Equations

We start the chapter by formalising some used objects. We consider the following objects.

- $M(n,d)$ denotes the class of $n \times d$-matrices.
- W is a d-dimensional Brownian motion
- μ is a $\mathbb{R}^n$-valued function with arguments (t, X_t) with X_t being a n-dimensional stochastic process.
- σ a $M(n,d)$-valued function with arguments as in μ.
- x_0 a $\mathbb{R}^n$-valued vector.

We want then to understand when the following has a solution

$$dX_t = \mu(t, X_t)\ dt + \sigma(t, X_t)\ dW_t, \quad X_0 = x_0. \tag{5.1/2}$$

We call such an equation the **stochastic differential equation** or simply SDE. We know that the above is loosely notation for the integral form as

$$X_t = x_0 + \int_0^t \mu(s, X_s)\ ds + \int_0^t \sigma(s, X_s)\ dW_s, \tag{5.3}$$

for all $t \geqslant 0$. The following proposition tells us when an solution exist to the problem above. In the below $\|\cdot\|$ is usual euclidian norm

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Proposition 5.1. (Bjork) *Suppose that there existis a constant K such that the following conditions are satisfied for all x, y and t.*

$$\|\mu(t,x) - \mu(t,y)\| \leqslant K\|x - y\|, \tag{5.6}$$
$$\|\sigma(t,x) - \sigma(t,y)\| \leqslant K\|x - y\|, \tag{5.7}$$
$$\|\mu(t,x)\| + \|\sigma(t,x)\| \leqslant K(1 + \|x\|). \tag{5.8}$$

Then there exists a unique solution to the SDE above. Furthermore, the solution has the properties

1. *X is $\mathcal{F}_t^W$-adapted.*
2. *X has continuous trajectories.*
3. *X is a Markov process.*
4. *There exists a constant C such that*

$$E[\|X_t\|^2] \leqslant Ce^{Ct}(1 + \|x_0\|^2). \tag{5.9}$$

In genereal the solution to an SDE is so complicated, that it in practical terms is unsolvable and may only be approximated on a finely subdividet grid as jumps. There does however exist som nontrivial cases where we may infer a analytical solution. One is the rather important **Geometric Brownian motion**.

Proposition 5.2. (Bjork) *Consider the SDE*

$$dX_t = \alpha X_t\ dt + \sigma X_t\ dW_t, \tag{5.13}$$

with $X_0 = x_0$. Then the solution is given as

$$X_t = x_0 \cdot \exp\left\{\left(\alpha - \frac{\sigma^2}{2}\right)t + \sigma W_t\right\}. \tag{5.15}$$

The expected value of X is given as $E[X_t] = x_0 e^{\alpha t}$ (eq. 5.16).

One other generalisation that is analytically solvable is the Linear SDE.

Proposition 5.3. (Bjork) *Consider the SDE*

$$dX_t = (AX_t + b_t)\ dt + \sigma_t\ dW_t, \tag{5.19}$$

with $X_0 = x_0$ *and* $A \in M(n,n)$ *and* b_t *being a real-valued function. Then the solution is given as*

$$X_t = e^{At}x_0 + \int_0^t e^{A(t-s)}b_s\ ds + \int_0^t e^{A(t-s)}\sigma_s\ dW_s. \tag{5.20}$$

Where we define the exponential of a matrix as below

$$e^{At} = \sum_{k=0}^{\infty} A^k \frac{1}{k!}t^k.$$

In general with the SDE we have a partial differential operator $\mathcal{A}$ called the **infinitesimal operator** of X which has some interesting analytical properties regarding X.

Definition 5.4. (Bjork) *Consider the SDE*

$$dX_t = \mu(t, X_t)\ dt + \sigma(t, X_t)\ dW_t. \tag{5.21}$$

The partial differential operator $\mathcal{A}$ *is defined, for any function* $h \in C^2(\mathbb{R}^n)$*, by*

$$\mathcal{A}h(t,x) = \sum_{i=1}^{n} \mu_i(t,x)\frac{\partial h}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^{n}(\sigma(t,x)\sigma(t,x)^\top)_{ij}\frac{\partial^2 h}{\partial x_i \partial x_j}(x).$$

We see that in terms of Ito's formula the operator is included as such

$$df(t,X_t) = \left\{\frac{\partial f}{\partial t}(t,X_t) + \mathcal{A}f(t,x)\right\}\ dt + [\nabla_x f](t,X_t)\sigma(t,X_t)\ dW_t,$$

where ∇_x is the gradient for function $h \in C^1(\mathbb{R}^n)$ as

$$\nabla_x h(x) = \left[\frac{\partial h}{\partial x_1}(x), ..., \frac{\partial h}{\partial x_n}(x)\right].$$

16.4 Partial differential equations

Proposition 5.5. (Bjork) (Feynmann-Kac) *Assume that* F *is a solution to the boundary value problem*

$$\frac{\partial F}{\partial t}(t,x) + \mu(t,x)\frac{\partial F}{\partial x}(x,t) + \frac{1}{2}\sigma^2(t,x)\frac{\partial^2 F}{\partial x^2}(t,x) = 0,$$

with boundary condition $F(T,x) = \Phi(x)$*. Assume furthermore that the process*

$$\sigma(s, X_s)\frac{\partial F}{\partial x}(s, X_s) \in \mathcal{L}^2$$

as per definition 4.4, where X *is defined below. Then* F *has the representation*

$$F(t,x) = E_{t,x}[\Phi(X_T)] = E[\Phi(X_T) \mid X_t = x], \tag{5.29}$$

where X *satisfies the SDE*

$$dX_s = \mu(s, X_s)\ ds + \sigma(s, X_s)\ dW_s, \tag{5.30}$$

with boundary condition $X_t = x$*.*

Proposition 5.6. (Bjork) (Feynmann-Kac) *Assume that* F *is a solution to the boundary value problem*

$$\frac{\partial F}{\partial t}(t,x) + \mu(t,x)\frac{\partial F}{\partial x}(x,t) + \frac{1}{2}\sigma^2(t,x)\frac{\partial^2 F}{\partial x^2}(t,x) - rF(t,x) = 0, \tag{5.34}$$

with boundary condition $F(T,x) = \Phi(x)$*. Assume furthermore that the process*

$$e^{-rs}\sigma(s, X_s)\frac{\partial F}{\partial x}(s, X_s) \in \mathcal{L}^2$$

as per definition 4.4, where X is defined below. Then F has the representation

$$F(t,x) = e^{-r(T-t)}E_{t,x}[\Phi(X_T)] = e^{-r(T-t)}E[\Phi(X_T) \mid X_t = x], \tag{5.36}$$

where X satisfies the SDE

$$dX_s = \mu(s,X_s)\,ds + \sigma(s,X_s)\,dW_s, \tag{5.37}$$

with boundary condition $X_t = x$.

Proposition 5.8. (Bjork) (Feynmann-Kac) *Assume that F is a solution to the boundary value problem*

$$\frac{\partial F}{\partial t}(t,x) + \sum_{i=1}^{n}\mu_i(t,x)\frac{\partial F}{\partial x}(x,t) + \frac{1}{2}\sum_{i,j=1}^{n}C_{ij}(t,x)\frac{\partial^2 F}{\partial x^2}(t,x) - rF(t,x) = 0,$$

with boundary condition $F(T,x) = \Phi(x)$ and $C_{ij} = \sigma\sigma^\top$. Assume furthermore that the process

$$e^{-rs}\sum_{i=1}^{n}\sigma_i(s,X_s)\frac{\partial F}{\partial x}(s,X_s) \in \mathcal{L}^2$$

as per definition 4.4, where X is defined below. Then F has the representation

$$F(t,x) = e^{-r(T-t)}E_{t,x}[\Phi(X_T)], \tag{5.39}$$

where X satisfies the SDE

$$dX_s = \mu(s,X_s)\,ds + \sigma(s,X_s)\,dW_s, \tag{5.40}$$

with boundary condition $X_t = x$.

Proposition 5.9. (Bjork) *Consider as given a vector process X with generator $\mathcal{A}$, and a function $F(t,x)$. Then, modulo some integrability condition, the following hold:*

- *The process $F(t,X_t)$ is a martingale relative to the filtration $\mathcal{F}^X$ if and only if F satisfies the PDE*

$$\frac{\partial F}{\partial t} + \mathcal{A}F = 0.$$

- *The process $F(t,X_t)$ is a martingale relative to the filtration $\mathcal{F}^X$ if and only if, for every (t,x) and $T \geqslant t$, we have*

$$F(t,x) = E_{t,x}[F(T,X_T)].$$

16.5 The Product Integral

Consider the (stochastic) function $Y : \mathbb{R} \to \mathbb{R}^{n\times n}$, where $\mathbb{R}^{n\times n}$ is the set of all $n \times n$ real-valued matrices, that is Y has the matrix-representation

$$Y(t) = \begin{bmatrix} Y_{11}(t) & \cdots & Y_{1n}(t) \\ Y_{21}(t) & \cdots & Y_{2n}(t) \\ \vdots & \ddots & \vdots \\ Y_{n1}(t) & \cdots & Y_{nn}(t) \end{bmatrix}.$$

Assume furthermore that for each coordinate function $Y_{ij}(t)$ the equation

$$\frac{dY_{ij}}{dt}(t) = Y_{i1}(t)A_{1j}(t) + \cdots + Y_{in}(t)A_{nj}(t), \qquad Y_{ij}(s) = C_{ij},$$

is satisfied. That is on matrix form the linear system of differential equation

$$\frac{dY}{dt}(t) = Y(t)A(t), \qquad Y(s) = C,$$

for some function $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ and initial condition $C \in \mathbb{R}^{n\times n}$. If A is continuous we know that Y is well-defined and absolutely continuous. We may then state the following theorem regarding uniqueness and existence of such a function above.

Theorem 1.1. (Bladt) (Uniqueness) *Consider the homogeneous system of linear differential equations*

$$\mathbf{Y}'(t) = \mathbf{Y}(t)\mathbf{A}(t), \qquad \mathbf{Y}(s) = \mathbf{C}, \tag{1}$$

where $\mathbf{Y}(t)$, $\mathbf{A}(t)$ and $\mathbf{C}$ are $n \times n$-matrices and $\mathbf{A}(t)$ is continuous on $[s,t]$. Then (1) has at most one solution.

Theorem 1.2. (Bladt) (Existence) *The matrix function*

$$\mathbf{Y}(t) = \sum_{k=0}^{\infty} \mathbf{Y}_k(t), \tag{3}$$

converges uniformly and absolutely on finite intervals, and solves the differential equation

$$\mathbf{Y}'(t) = \mathbf{Y}(t)\mathbf{A}(t), \qquad \mathbf{Y}(s) = \mathbf{C}.$$

Now, we know that for any system as in (1) with a continuous function $\mathbf{A}(t)$ we can always construct the solution as some converging series as per theorem 1.2 then theorem 1.1 gives that the solution is unique. We can then with a piece of mind define a symbol for such a solution, without care for the *exact* solution.

Definition 1.3. (Bladt) (The Product Integral) *For any continuous matrix function $\mathbf{A}(t)$ we define the product integral as*

$$\prod_s^t (\mathbf{I} + \mathbf{A}(x)\ dx)$$

as the unique solution $\mathbf{Y}(t)$ to

$$\mathbf{Y}'(t) = \mathbf{Y}(t)\mathbf{A}(t), \qquad \mathbf{Y}(s) = \mathbf{I}.$$

From a simple integral argument we may construct *a* converging series not containing $\mathbf{Y}$ itself. We see that

$$\begin{aligned}
\prod_s^t (\mathbf{I} + \mathbf{A}(x)\ dx) &= \mathbf{I} + \int_s^t \mathbf{Y}'(x)\ dx = \mathbf{I} + \int_s^t \mathbf{Y}(x)\mathbf{A}(x)\ dx \\
&= \mathbf{I} + \int_s^t \left[\mathbf{I} + \int_s^{x_1} \mathbf{Y}'(x_2)\ dx_2\right] \mathbf{A}(x_1)\ dx_1 \\
&= \mathbf{I} + \int_s^t \left[\mathbf{I} + \int_s^{x_1} \mathbf{Y}(x_2)\mathbf{A}(x_2)\ dx_2\right] \mathbf{A}(x_1)\ dx_1 \\
&= \mathbf{I} + \int_s^t \mathbf{A}(x_1)\ dx_1 + \int_s^t \int_s^{x_1} \mathbf{Y}(x_2)\mathbf{A}(x_2)\mathbf{A}(x_1)\ dx_2\ dx_1.
\end{aligned}$$

One can continue indefinitely and see that we have the following representation.

Corollary 1.4. (Bladt) (Peano Representation) *The prduct integral has series representation given by*

$$\begin{aligned}
\prod_s^t (\mathbf{I} + \mathbf{A}(x)\ dx) &= \mathbf{I} + \sum_{i=1}^{\infty} \int_s^t \int_s^{x_1} \cdots \int_s^{x_n} A(x_n)\cdots\mathbf{A}(x_2)\mathbf{A}(x_1)\ dx_n \cdots dx_2\ dx_1 \\
&= \mathbf{I} + \int_s^t \mathbf{A}(x_1)\ dx_1 + \sum_{i=2}^{\infty} \int_s^t \int_s^{x_1} \cdots \int_s^{x_n} A(x_n)\cdots\mathbf{A}(x_2)\mathbf{A}(x_1)\ dx_n \cdots dx_2\ dx_1.
\end{aligned}$$

16.5.1 Properties of the Product Integral

The product integral is the fundamental solution in the sense that if

$$\mathbf{Y}'(t) = \mathbf{Y}(t)\mathbf{A}(t), \qquad \mathbf{Y}(s) = \mathbf{C},$$

then

$$\mathbf{Y}(t) = \mathbf{C}\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big).$$

We furthermore have for any s, t, u have

$$\mathbf{Y}(t) = \mathbf{Y}(s)\prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\prod_u^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big).$$

From uniqueness we then get the following theorem.

Theorem 1.5. (Bladt) *For any s, t, u we have that*

$$\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\prod_u^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big).$$

By choosing $u = t$ in the above we get the following corollary.

Corollary 1.6. (Bladt) *The inverse of the product integral is*

$$\left[\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\right]^{-1} = \prod_t^s \big(\mathbf{I} + \mathbf{A}(x)\ dx\big).$$

Theorem 1.7. (Bladt) *If $\mathbf{A}(x)$ commutes withall $\mathbf{B}(x =$ i.e. $\mathbf{A}(x)\mathbf{B}(x) = \mathbf{B}(x)\mathbf{A}(x)$, then*

$$\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\prod_s^t \big(\mathbf{I} + \mathbf{B}(x)\ dx\big) = \prod_s^t \big(\mathbf{I} + (\mathbf{A}(x) + \mathbf{B}(x))\ dx\big).$$

Corollary 1.8. (Bladt) *Let r be a real-valued function, then*

$$\exp\left(-\int_s^t r(x)\ dx\right)\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = \prod_s^t \big(\mathbf{I} + (\mathbf{A}(x) - r(x)\mathbf{I})\ dx\big).$$

Theorem 1.9. (Bladt) *If $\mathbf{A}(x)$ commutes for all x i.e. $\mathbf{A}(y)\mathbf{A}(x) = \mathbf{A}(x)\mathbf{A}(y)$, then*

$$\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = e^{\int_s^t \mathbf{A}(x)\ dx}.$$

In particular, if $\mathbf{A}(x) = \mathbf{A}\lambda(x)$ for some real-valued function λ and a $n \times n$ matrix $\mathbf{A}$ then

$$\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = e^{\mathbf{A}\int_s^t \lambda(x)\ dx}.$$

Taking the function $\lambda = 1$ we have that $\int_s^t \lambda(x)\ dx = (t - s)$ and so we get the result.

Theorem 1.10. (Bladt) *If $\mathbf{A}(x) = \mathbf{A}$ is constant then*

$$\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = e^{\mathbf{A}(t-s)}.$$

Theorem 1.11. (Bladt) *The product integral satisfies*

$$\frac{\partial}{\partial s}\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) = -\mathbf{A}(s)\prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big).$$

On the other hand, the solution to the system of differential equations

$$\frac{\partial}{\partial s}\mathbf{X}(s) = -\mathbf{A}(s)\mathbf{X}(s)$$

with initial condition $\mathbf{X}(t) = \mathbf{I}$ *is the function*

$$s \mapsto \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)$$

We have a very useful result for insurance mathematics and markov chains given by the Van Loan's theorem giving the product integral of a block matrix with zero lower left block.

Theorem 1.12. (Bladt) *Let* $\mathbf{A}(x)$, $\mathbf{B}(x)$ *and* $\mathbf{C}(x)$ *be continuous matrix functions such that*

$$\mathbf{D}(x) = \begin{bmatrix} \mathbf{A}(x) & \mathbf{B}(x) \\ \mathbf{0} & \mathbf{C}(x) \end{bmatrix},$$

is a square matrix. Then the product integral of $\mathbf{D}$ *is*

$$\prod_s^t \big(\mathbf{I} + \mathbf{D}(x)\ dx\big) = \begin{bmatrix} \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) & \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \\ \mathbf{0} & \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \end{bmatrix}.$$

Proof.

Start by defining $\mathbf{X}(t)$ as below

$$\begin{aligned} \mathbf{X}(t) &= \begin{bmatrix} \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big) & \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \\ \mathbf{0} & \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \end{bmatrix} \\ &:= \begin{bmatrix} \mathbf{X}_{11}(t) & \mathbf{X}_{12}(t) \\ \mathbf{X}_{21}(t) & \mathbf{X}_{22}(t) \end{bmatrix}. \end{aligned}$$

The derivative is then

$$\frac{d}{dt}\mathbf{X}(t) = \begin{bmatrix} \frac{d}{dt}\mathbf{X}_{11}(t) & \frac{d}{dt}\mathbf{X}_{12}(t) \\ \frac{d}{dt}\mathbf{X}_{21}(t) & \frac{d}{dt}\mathbf{X}_{22}(t) \end{bmatrix}.$$

with

$$\begin{aligned} \frac{d}{dt}\mathbf{X}_{11}(t) &= \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{A}(t), \\ \frac{d}{dt}\mathbf{X}_{12}(t) &= \frac{d}{dt}\int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du, \\ \frac{d}{dt}\mathbf{X}_{21}(t) &= \mathbf{0}, \\ \frac{d}{dt}\mathbf{X}_{22}(t) &= \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\mathbf{C}(t). \end{aligned}$$

Consider that

$$\begin{aligned} \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) &= \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \underbrace{\prod_t^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)}_{=\mathbf{I}} \\ &= \prod_u^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big), \end{aligned}$$

hence using af Riemann approximation of the integral $\mathbf{X}_{12}(t)$ we may write

$$\begin{aligned}
\mathbf{X}_{12}(t) &= \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \\
&= \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \\
&= \lim_{n\to\infty} \frac{t-s}{n} \sum_{i=0}^{n} \prod_s^{s+(t-s)i/n} \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(s+(t-s)i/n) \prod_{s+(t-s)i/n}^{s} \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \\
&\prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \\
&= \left\{ \lim_{n\to\infty} \frac{t-s}{n} \sum_{i=0}^{n} \prod_s^{s+(t-s)i/n} \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(s+(t-s)i/n) \prod_{s+(t-s)i/n}^{s} \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \right\} \\
&\prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \\
&= \left\{ \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \right\} \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big).
\end{aligned}$$

We then get that

$$\begin{aligned}
\frac{d}{dt}\mathbf{X}_{12}(t) &= \frac{d}{dt} \left\{ \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \right\} \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \\
&= \left\{ \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(t) \prod_t^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \right\} \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big) \\
&+ \left\{ \int_s^t \prod_s^u \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(u) \prod_u^s \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\ du \right\} \prod_s^t \big(\mathbf{I} + \mathbf{C}(x)\ dx\big)\mathbf{C}(t) \\
&= \prod_s^t \big(\mathbf{I} + \mathbf{A}(x)\ dx\big)\mathbf{B}(t) + \mathbf{X}_{12}(t)\mathbf{C}(t).
\end{aligned}$$

We hope to show that $\frac{d}{dt}\mathbf{X}(t) = \mathbf{X}(t)\mathbf{D}(t)$. Calculating the right side we have

$$\begin{aligned}
\mathbf{X}(t)\mathbf{D}(t) &= \begin{bmatrix} \mathbf{X}_{11}(t) & \mathbf{X}_{12}(t) \\ \mathbf{0} & \mathbf{X}_{22}(t) \end{bmatrix} \begin{bmatrix} \mathbf{A}(x) & \mathbf{B}(x) \\ \mathbf{0} & \mathbf{C}(x) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}_{11}(t)\mathbf{A}(x) & \mathbf{X}_{11}(t)\mathbf{B}(x) + \mathbf{X}_{12}(t)\mathbf{C}(x) \\ \mathbf{0} & \mathbf{X}_{22}(t)\mathbf{C}(x) \end{bmatrix},
\end{aligned}$$

given that $\mathbf{X}$ satisfies the desired differential equation and so it follows that $\mathbf{X}$ is the product integral $\prod_s^t \big(\mathbf{I} + \mathbf{D}(x)\ dx\big)$ as desired.■

Chapter 17

Linear Algebra

17.1 Invertible matrices

This sections study some fundamental properties of the *invertible matrix*. We start by defining what an invertible matrix is.

Definition. *Let A be an $n \times m$ matrix and B be an $m \times n$ matrix. We say that A is invertible if*

$$AB = I_{\min(m,n)}.$$

In general, we only consider square matrices. If the above holds we say that B is A's inverse and we write $B = A^{-1}$.

We can consider some equivalence statements regarding invertible matrices.

Theorem. (The Invertible Matrix Theorem) *Let A be a $n \times n$ matrix over a field K ($\mathbb{R}^n$), then the following statements are equivalent*

1. *There exists an $n \times n$ matrix B such that $AB = I_n = BA$.*
2. *There exist either a left inverse B or a right invers C i.e. $BA = I_n = AC$. In this case, $B = C$.*
3. *A has an inverse and is nonsingular and is nondegenerate.*
4. *A is row-equivalent to I_n.*
5. *A is column-equivalent to I_n.*
6. *A has n pivot positions.*
7. *A has full rank i.e. $rank(A) = n$ (spans K).*
8. *The equation $Ax = 0$ ($x \in K$) has only the trivial solution $x = 0$.*
9. *The equation $Ax = b$ has only one solution x.*
10. *The kernal of A is trivial i.e. $ker(A) = \{0\}$.*
11. *The columns of A are linearly independent.*
12. *The columns of A span K.*
13. *$span(A) = K$.*
14. *The columns of A form a basis of K.*
15. *The linear transformation Ax is a bijection from K to K.*
16. *A has non-zero determinant i.e. $det(A) \neq 0$.*
17. *A has not 0 as an eigenvalue.*
18. *The transpose of A is invertible.*
19. *A can be expressed as a finite product of elemtary matrices.*

We futhermore have some properties.

Proposition. (Properties) *Let A be an $n \times n$ invertible matrix. Then*

1. $(A^{-1})^{-1} = A$

2. $(kA)^{-1} = k^{-1}A^{-1}$ *with* $k \neq 0$.
3. $(Ax)^{+} = x^{+}A^{-1}$ *if* A *has orthonormal columns.* $(\cdot)^{+}$ *denotes the Moore-Penrose inverse and* x *is a vector.*
4. *If* B *is an* $n \times n$ *invertible matrix then* $(AB)^{-1} = B^{-1}A^{-1}$.
5. $\det(A^{-1}) = (\det(A))^{-1}$

The property 2 is especially useful in some settings. Consider for instance

$$A = \begin{bmatrix} \sigma & 0 \\ \sigma & \sigma \end{bmatrix} = \sigma \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \sigma \tilde{A}.$$

Then we simply find the inverse of $\tilde{A}$ and multiply by σ^{-1}. That is,

$$A^{-1} = \frac{1}{\sigma}\tilde{A}^{-1} = \frac{1}{\sigma}\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1/\sigma & 0 \\ -1/\sigma & 1/\sigma \end{bmatrix}.$$

We have an easy propositions regarding diagonal matrices.

Proposition. *If* A *is an diagonal matrix, then* A *is invertible. In particular,*

$$A^{-1} = \operatorname{diag}(A_{11}^{-1}, ..., A_{nn}^{-1}).$$

Chapter 18

Coding

18.1 R-Packages

18.1.1 mlr3

This short introduction to the `mlr3` package is based on the official documentation and introduction book and exercises done in the course Machine Learning in Non-Life Insurance.

We start by installing the package `mlr3verse` containing all of the importatant packages.

Index

T-bond, 45, 59
α-pruned tree, 95

absolutely continuous, 130
Adapted process, 171
admissible portfolio, 56
almost surely convergence, 143
arbitrage, 32
arbitrage portfolio, 39, 46
asymptotic normal distribution, 155

Backfitting Algorithm, 93
Bagging, 96
bank account, 14, 45
BART, 100
Basic CLT, 156
Bayes estimator, 76
Bayes risk, 76
Bayes' Theorem, 131
Binary loss function, 76
Binomial algorithm, 36
Birkhoff's ergodic theorem, 149
Black-Scholes equation, 46, 65
Black-Scholes formula, 47
Black-Scholes model, 45
Borel-Cantelli, 2nd half, 139
Brownian motion, 169
budget constraint, 42

CART, 94
Cauchy-Schwarz inequality, 135
Chapman-Kolmogorov's, 9
characteristic function, 152
Chebyshev's inequality, 134
Chebyshev-Cantelli's inequality, 134
complete market, 33, 36, 50
conditional generalized error, 77
Conditional risk, 77
consumption process, 43
consumption stream, 42
contingent claim, 33, 35, 41
Continuity mapping theorem, 154
Continuity theorem, 154
continuously compounded forward rate, 15
contract function, 33, 35
convergence almost surely, 143
convergence in distribution, 151
convergence in L-p, 144
convergence in probability, 144
convolution, 153
covariance, 135
Cramer-Wold's device, 154

decision rule, 76
Decision trees, 94
Delta method, 156
density, 133
diagonal matrix, 188
diffusion processes, 173
discrete distribution, 133
distribution, 133
distribution function, 133
Doleans exponential process, 62
Dynkin class, 138

Elastic net, 88
empirical risk, 78
empirical risk minimizer, 78
equivalent martingale measure, 56
equivalent measure, 40
equivalent measures, 130
equivalent probability measures, 131
ergodic, 148
Ergodic theorem, $\mathcal{L}^p$-version, 149
Ergodic transformation theorem, 150
estimation error, 78
Estimator, 75
Etemadi's maximal inequality, 146
Etemahdi's version, 148
European call option, 31, 45
Excess Risk, 77
exercise date, 45
exercise price, 45
expectation, 134

Farkas' Lemma, 39
Fatou's Lemma, 145
Feynmann-Kac, 180
Filtration, 171
financial derivative, 31, 33
finite dimensional product set, 149
First Fundamental Theorem, 40, 57
forward rate, 16

gain process, 43
General Pricing Equation, 58
generalized error, 77
generically arbitrage free, 67
Geometric Brownian motion, 179
Girsanov kernel, 62
Girsanov Theorem, 61
Girsanov Theorem, converse, 62

hedgeable, 50
hedged, 41
hedging portfolio, 33, 36, 41, 50
Holder's Inequality, 145

incomplete market, 52
independent, 138
independent, jointly, 134
inductions bias, 78
infinitesimal operator, 180
intensity matrices, 7
Interest rate process, 15
Invariance Lemma, 57
invariant measure, 148
Inversion theorem, 153
invertible matrix, 187
Ito's formula, 175
Ito's formula, multi-dimensional, 176

Jacod, 58
Jensen's inequality, 134
joint distribution, 134
jointly independent, 138

k-nearest-neighbor, 88
Kernel Smoother, 89
Khintchine's ergodic theorem, 150
Khintchine's ergodic theorem, two-sided version, 150
Khintchine-Kolmogorov, 147
Kolmogorov' Differential Equations, 8
Kolmogorov's 3-series theorem, 147
Kolmogorov's zero-one law, 139
Kreps-Yan Separation Theorem, 57
Kronecker, 147

Laplace's CLT, 156
Laplace's CLT, multivariate version, 156
Lasso regression, 84
Least squares estimator, 82
Levy Characterisation of Brownian motion, 63
Levy's maximal inequality, 146
LIBOR forward rate, 15
likelihood process, 59, 66, 131
Lindeberg's CLT, 157
Lindeberg's CLT, multivariate version, 157
Lindeberg's condition, 156
Linear SDE, 179
Linear Smoother, 88
local martingale measure, 56
loss function, 76
Lyapounov's condition, 156

M-fold Cross validation, 79
Markov jump process, 7
Markov's inequality, 134
Markov-jump representation interest model, 16
Markovian, 43
martingale, 172
martingale measure, 40
martingale meausre, 32
Martingale Property, 46
Martingale Representation Theorem, 61
Maximal Ergodic Lemma, 149
measure-preserving dynamical system, 148
mixing, 148
Mortality forward rate, 21, 22
Mortality rate, 20
multi-period model, 35
multivariate Gaussian distribution, 155

natural filtration, 171
Natural Splines, 91
Nested $M_1 - M_2$ Cross-validation, 80
no arbitrage, NA, 57
No Free Lunch with Vanishing Risk, NFLVR, 57
Normal disribution, 141
normalized economy, 57
Novikov Condition, 62
numeraie asset, 40

objective probability measure, 66
one-period binomial model, 32

PAC Least squares estimator, 83
Peano Representation, 182
Phase-Type distribution, 10
Phase-type representation of bond prices, 17
Poisson Deviance, 76
Polya class, 153
portfolio, 39
portfolio strategy, 35, 43

portfolio weights, 43
portfolio-consumption pair, 44
Portmanteau's lemma, 152
Product Integral, 182
projection sigma-algebra, 150
Put-call parity, 54

Quadratic loss function, 76

Radon-Nikodym derivative, 130
Radon-Nikodym Theorem, 130
Random forest, 97
random variable, 133
reachable, 33, 35, 50
relative portfolio, 43
replicated, 33, 41, 50
replicating portfolio, 33, 36, 41, 50
Representation of Brownian Functionals, 61
Restricted eigenvalue property (REP), 85
Ridge regression, 83
Risk, 77
risk, 76
risk free asset, 45
Risk Neutral Valuation Formula, 59
Risk Neutral Valuation formula, 46
risk-neutral valueation formula, 34, 53

Scaling and squaring argument, 14
Scheffe's, 151
Second Fundamental Theorem, 41
self-financing condition, 43
self-financing portfolio, 35, 42
short interest rate, 45
short rate, 45
simple claim, 50
singular, 130
Skorokhod, 146
Skorokhod's representation theorem, 152
SLLN, $\mathcal{L}^p$-version, 148
SLLN, strong version, 148
SLLN, weak form, 146
Slutsky's lemma, 155
solution to linear differential equation system, existence, 182
solution to linear differential equation system, uniqueness, 182
sparsistent, 87
Splines, 91
state price deflator, 41
stationary, 150
stochastic differential, 175
stochastic differential equation, SDE, 179
stochastic differential equations, SDE, 173
stochastic discount factor, 41, 59
Stopping time, 9
strike price, 45
Strong Markov property, 9
submartingale, 172
subtree, 95
supermartingale, 172
Supervised Learning, 75

tail sigma-algebra, 139
Term structure equation, 18
Term structure interest model, 18
test data, 79
the First Fundamental Theorem, 56
the greeks, 55
The Hansen-Jagannathan Bounds, 68
The Second Fundamental Theorem, 58, 67
time homogeneous Markov jump processes, 7
time of maturity, 45
training data, 75
traning data, 79
transition matrix, 7
transition probabilities, 7

Univerformization, 14

validation data, 79
value process, 32, 35, 43
vanishing variance condition, 156

weakest link algorithm, 96
weakly convergence, 151
Wiener process, 169

XGBoost, 99

Yield, 16
yield curve, 16

zero coupon bond, 45, 59
Zero-Coupon Bond, 15
zero-one law, 139

Bibliography