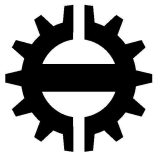


TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology

Matti Vihola

Random Sets for Multitarget Tracking and Data Fusion

Licentiate Thesis

Examiners: Prof. Keijo Ruohonen (TUT)
Prof. Ari Visa (TUT)
Dr. Tech. Petri Salmela (Datactica Ltd)

Preface

First of all, I would like to thank Datactica Ltd for providing me the opportunity to write this thesis. It has been a highly therapeutic experience to dig deeper into the theoretical aspects of some problems that I have previously faced. Althgough this thesis is primarily a partial requirement for an academic degree, I hope that other people find the thesis useful as well.

I am deeply grateful to all the people that have read versions of this thesis, commented, and proposed changes. Specifically, I would like acknowledge Prof. Keijo Ruohonen, Dr. Tech. Petri Salmela, MSc Kari Heine, and MSc Tuomo Valkonen for valuable comments. This thesis is prepared with LaTeX, written using the Jed text editor, on Linux platform. Many of the diagrams are created with the Dia drawing utility. I would like to thank the people that have contributed in the development of these free software projects, and more generally, all the people in the free software community. In addition, I would like to acknowledge the authors that have made their publications freely available in the Web.

Tampere, 13th August 2004

Matti Vihola
Opiskelijankatu 4 E 289
33720 Tampere
Tel: (+358) 50 3045553

Contents

Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology
Institute of Mathematics

VIHOLA, MATTI: Random Sets for Multitarget Tracking and Data Fusion
Licentiate thesis, 115 pages, 19 enclosure pages.
Examiners: Prof. Keijo Ruohonen, Prof. Ari Visa, and Dr. Tech. Petri Salmela.
Funding: Datactica Ltd.
August 2004
Keywords: Bayesian estimation, data fusion, finite random set, Monte Carlo, particle filtering, target tracking, track initiation.

This thesis is a background study of the theory of finite random sets, and their application to target tracking. The main purpose is to provide a solid theoretical basis for further research and development of random set tracking methods. The random set approach to multitarget tracking is a theoretically sound framework, that covers the joint estimation of the number of targets, and the state of the targets. In addition to the theory of random sets, the thesis provides a review of related subjects, such as general probability theory and recursive Bayesian estimation. Sequential Monte Carlo (SMC) is selected as an approximative computational strategy, so a summary of SMC methods is included. In addition, a literature review on Bayesian target tracking methods is given.

The random set tracking model that is considered in this thesis includes a model for such sensors that produce at most one measurement per report. A SMC implementation for the random set tracking model is derived. The results that were obtained during the limited tests of the SMC algorithm serve as a "proof-of-concept". Based on the results, the practical usability of the algorithm cannot be assessed. Merely, the results in this thesis show that the random set framework has potential for challenging tracking situations. The tests consisted of synthetic bearings-only tracking scenarios with up to four targets, in the presence of false alarms, and missed measurements. The algorithm was capable to track up to two targets quite reliably. In addition, the algorithm showed robustness against relatively high false alarm rates. Moreover, the performance of the algorithm degraded gradually with respect to an increasing mismatch in the false alarm rate parameter value and the parameter value in the tracking environment.

List of Symbols

General

Symbol	Description
$a, b, c \ldots$	Elements of some set.
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \ldots$	Random elements.
$\underline{a}, \underline{b}, \underline{c} \ldots$	Vectors in a Euclidean space $\mathbb{R}^d$.
$A, B, C \ldots$	Sets.
$\mathrm{A}, \mathrm{B}, \mathrm{C} \ldots$	Matrices.
$\mathcal{A}, \mathcal{B}, \mathcal{C} \ldots$	Collections of sets or collections.
$[a, b]$	A closed real interval, $a \leq x \leq b$.
(a, b)	An open real interval, $a < x < b$.
$[a, b)$	A half-open real interval, $a \leq x < b$.
(a, b, c)	An ordered set with three elements.
$[a, b, c]^T$	A column vector.
$\underline{a}^T, \mathrm{A}^T$	The transpose of a vector $\underline{a}$, and a matrix A, respectively.
$a \in A$	a is a member of A.
$A \subset B$	The set A is a subset (not necessarily proper) of the set B.
$A \supset B$	The set A is a superset (not necessarily proper) of the set B.
$A \setminus B$	The set difference, $A \cap \complement B$.
$\complement A$	The complement of a set A.
$\lvert A \rvert$	The cardinality of (number of elements in) the set A.
$A \times B$	The Cartesian product of sets A and B.
$\times_{i=1}^{n} A_i$	The Cartesian product of sets $A_1, \ldots, A_n$.
A^n	The Cartesian product space $\times_{i=1}^{n} A$.
$\mathcal{P}(A)$	The power set of A, i.e. the collection of subsets of A.
$a \triangleq b$	An equality defining a as b.
$a \leftarrow b$	The value b is substituted into a.
$\operatorname{atan2}(y, x)$	The four quadrant inverse tangent.
$B_x(r)$	An open ball with radius r centred at x.
$\overline{B}_x(r)$	A closed ball with radius r centred at x.
$\mathcal{B}(X)$	The collection of the Borel sets of a topological space X.

Symbol	Description
f^+, f^-	The decomposition of a function $f : X \to \overline{\mathbb{R}}$, so that $f = f^+ - f^-$, where $f^+ \geq 0$ and $f^- \geq 0$.
$n!$	The n factorial, $\prod_{i=1}^{n} i$.
$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$	The sets of natural numbers (≥ 0), integers, rational numbers, real numbers, and complex numbers, respectively.
$\operatorname{rng}(f)$	The range of function f.
$\overline{\mathbb{R}}$	The set of extended real numbers, $\mathbb{R} \cup \{\infty\} \cup \{-\infty\}$.
$\boldsymbol{x} \sim D$	The random variable $\boldsymbol{x}$ has a distribution D.
$\overleftarrow{\boldsymbol{x}}(S)$	The preimage of a set S with respect to a function $\boldsymbol{x}$.
$\{x \in X : \operatorname{cond}(x)\}$	A set consisting of such elements x of X that the condition $\operatorname{cond}(x)$ is satisfied.
$\mathbb{Z}_+$	The strictly positive integers.
$\delta_x(y)$	The Dirac delta function: $\delta_x(y) = 1$, when $y = x$, and 0 otherwise.
λ, λ^d	The Lebesgue measure in $\mathbb{R}$, and $\mathbb{R}^d$, respectively.
μ, ν, ξ	General measures.
$\mu \ll \nu$	The measure (possibly signed) μ is absolutely continuous with respect to the measure ν ($\mu(N) = 0 \implies \nu(N) = 0$).
ν^d	The d-fold product measure of ν.
$\sigma(\mathcal{G}), \sigma(\boldsymbol{y})$	The σ-algebra generated by a collection $\mathcal{G}$, or a random variable $\boldsymbol{y}$, respectively.
$\chi_A(x)$	The characteristic function (indicator function) of a set A.
$\int f(x)\mathrm{d}\mu(x)$	The integral of a function f with respect to a measure μ.
$\int_A f(x)\mathrm{d}\mu(x)$	The integral over a set A.
$\int f(x)\mathrm{d}x$	The Riemann integral in $\mathbb{R}$.

Random Variables

Symbol	Description
$c(a, b)$	The Bayes cost function of estimating b as a.
$\mathsf{corr}(\boldsymbol{x}, \boldsymbol{y})$	The correlation of random variables $\boldsymbol{x}$ and $\boldsymbol{y}$.
$\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y})$	The covariance of random variables $\boldsymbol{x}$ and $\boldsymbol{y}$.
$\mathsf{E}[\boldsymbol{x}]$	The expectation of a random variable $\boldsymbol{x}$.
$\mathsf{E}[\boldsymbol{x} \mid \mathcal{G}]$	The conditional expectation of a random variable $\boldsymbol{x}$ given a σ-algebra $\mathcal{G}$.
$\mathsf{E}[\boldsymbol{x} \mid \boldsymbol{y}]$	The Conditional expectation of a random variable $\boldsymbol{x}$ given the σ-algebra induced by a random variable $\boldsymbol{y}$.

Symbol	Description
$\mathsf{E}[\boldsymbol{x} \mid \boldsymbol{y} = y]$	The conditional expectation of a random variable $\boldsymbol{x}$ given $\boldsymbol{y} = y$.
$f_{\boldsymbol{x}}$	The Radon-Nikodym derivative of a measure $P_{\boldsymbol{x}}$ (with respect to some measure).
$h : \Omega \to X$	A function h maps values from a set Ω to a set X.
$h : (\Omega, \mathcal{M}) \to (X, \mathcal{N})$	A measurable function h maps values from a measurable space $(\Omega, \mathcal{M})$ to a measurable space $(X, \mathcal{N})$.
$N(\,\cdot\,; \underline{m}, \mathrm{R})$	The measure of a Gaussian distribution with a mean vector $\underline{m}$ and a covariance matrix R.
$p_{m\|n}$	The posterior density of $\boldsymbol{x}_m$ after $\boldsymbol{y}_{1:n} = y_{1:n}$ have been observed.
P	The probability measure in $(\Omega, \mathcal{M})$.
$P_{\boldsymbol{x}}$	The probability measure of a random variable $\boldsymbol{x}$.
$P(\boldsymbol{x} \in B \mid \boldsymbol{y})$	The conditional probability of $\boldsymbol{x}$ given $\boldsymbol{y}$.
$P(\boldsymbol{x} \in B \mid \boldsymbol{y} = y)$	The conditional probability of $\boldsymbol{x}$ given $\boldsymbol{y} = y$.
$\mathrm{R}_{\boldsymbol{x}}$	The covariance matrix of a random vector $\boldsymbol{x}$.
$\mathsf{V}[\boldsymbol{x}]$	The variance of a random variable $\boldsymbol{x}$.
$\boldsymbol{x}_k \xrightarrow{\text{a.s.}} \boldsymbol{x}$	Almost sure convergence.
$\boldsymbol{x}_k \xrightarrow{d} \boldsymbol{x}$	Convergence in distribution.
$\boldsymbol{x}_k \xrightarrow{L^p} \boldsymbol{x}$	Convergence in mean of order p.
$\boldsymbol{x}_k \xrightarrow{P} \boldsymbol{x}$	Convergence in probability.
$(\Omega, \mathcal{M}, P)$	A probability space, where Ω is a sample space, $\mathcal{M}$ are events and P is a probability measure.
$\pi_{m\|n}$	The posterior measure of $\boldsymbol{x}_m$ after $\boldsymbol{y}_{1:n} = y_{1:n}$ have been observed.

Random Sets

Symbol	Description
$D(s)$	The probability hypothesis density.
$f\|_A$	The restriction of a function $f : X \to Y$ into a set $A \subset X$.
$\mathcal{F}, \mathcal{G}, \mathcal{K}$	The collections of the finite, the open, and the compact sets of $\mathbb{S}$, respectively.
$\mathcal{F}_A, \mathcal{F}^A$	The collections of the closed sets missing and hitting a set $A \subset \mathbb{S}$, respectively.
$\mathcal{F}(k)$	The collections of the k-element sets of $\mathbb{S}$.
$\mathcal{F}_*$	The collection of the finite sets of $\mathbb{S}$.
$\mathcal{O}(C)$	The collection of the closed subsets of C.
$\mathbb{S}$	The hybrid space $\mathbb{R}^d \times D$, where D is a finite space.
$\beta_{\boldsymbol{X}}$	The belief measure of a random set $\boldsymbol{X}$.

Symbol	Description
$\delta\beta/\delta Z$	The set derivative of a belief measure β with respect to a finite set Z.
$\overline{\lambda}, \overline{\lambda}^k$	The extended Lebesgue measure in $\mathbb{S}$, and $\mathbb{S}^k$, respectively.
$\int f(Z)\delta Z$	The set integral.

Target Tracking

Symbol	Description
b_j	The probability that j targets are born.
$\boldsymbol{B}_k$	The multitarget state of the targets that are born during time interval $(t_{k-1}, t_k]$.
$\boldsymbol{c}_k$	The contact (association) indicator variable at time t_k.
d_j	The probability that j false measurements occur.
$D_{m\|n}$	The posterior PHD of $\boldsymbol{X}_m$ after $\boldsymbol{Y}_{1:n} = Y_{1:n}$ have been observed.
f_b	The birth density.
f_f	The false alarm density.
$p_d(x)$	The probability of detection for a target in state x.
p_f	The probability of false alarm.
$p_s(x)$	The probability of survival for a target in state x.
P_f	The probability measure of one false alarm measurement.
P_j	The probability measure of a born target.
$\boldsymbol{S}_k$	The multitarget state of the targets surviving from time instant t_{k-1} to t_k.
$\boldsymbol{w}_k^{(i)}$	The weight of the i'th sample (particle) at the k'th time instant.
$\boldsymbol{x}_k$	The single-target or the multitarget state variable at time t_k.
$\boldsymbol{x}_k^{(i)}$	The state variable of the i'th target at time t_k.
$\boldsymbol{X}_k$	The multitarget state variable at time t_k.
$\boldsymbol{y}_k$	The measurement variable at time t_k.
$\boldsymbol{Y}_k$	The random set measurement variable at time t_k.
$\boldsymbol{Z}_k^{(i)}$	The state of the i'th sample (particle) at time instant k.
η	The birth intensity parameter.
γ	The false alarm intensity parameter.
τ_k	The time difference between two consecutive updates, $\tau_k = t_k - t_{k-1}$.

Acronyms and Abbreviations

Abbreviation	Description
a.s, a.e.	Almost surely, almost every
BN	Bayesian network
CSM	Countably simple and measurable
DAP	Data association problem
DOA	Direction of arrival
DS	Dempster-Shafer
EAP	Expected *a posteriori*
EKF	Extended Kalman filter
FAR	False alarm rate
GPB	Generalised pseudo-Bayesian
GRND	Generalised Radon-Nikodym derivative
HMM	Hidden Markov model
IID	Independent and identically distributed
IMM	Interacting multiple model
IPDA	Integrated probabilistic data association
IS	Importance sampling
JIPDA	Joint integrated probabilistic data association
JoME	Joint multitarget estimator
JPDA	Joint probabilistic data association
KF	Kalman filter
KFM	Kalman filter model
LCHS	Locally compact, Hausdorff, and separable
MaME	Marginal multitarget estimator
MAP	Maximum *a posteriori*
MCDA	Monte Carlo data association
MHT	Multiple hypothesis tracking
PDA	Probabilistic data association
PDF	Probability density function
PHD	Probability hypothesis density
PHS	Probability hypothesis surface
RFS	Random finite set
RND	Radon-Nikodym derivative
SDE	Stochastic differential equation

Abbreviation	Description
SIS	Sequential importance sampling
SISR	Sequential importance sampling-resampling
SKFM	Switching Kalman filter model
SMC	Sequential Monte Carlo
SNR	Signal-to-noise ratio
SSM	State-space model
UKF	Unscented Kalman filter

Chapter 1

Introduction

Mathematical modelling is always a compromise. Naylor and Sell [1982] suggests that when a model includes many aspects of a real system, the model becomes complex, and mathematically intractable. On the other hand, when a model is made simple enough to be mathematically tractable, the modelling accuracy of the model may become poor. One philosophy for constructing a model is to use all available information, and create as accurate model as possible, but then seek for a tractable computational strategy [Stone et al. 1999]. Modern computer resources provide means to approximately handle complex models. There is a whole variety of simulation based algorithms available, which have gained attention in the last decades [Doucet et al. 2001; Liu 2003; Robert and Casella 1999].

This thesis addresses the problem of tracking an unknown number of targets. Such a problem arises naturally in a surveillance application, where a sensor, or a suite of sensors produce noisy measurements [Goodman et al. 1997]. In such an application, the targets can be, for example, aircrafts whose geokinematic state is tracked [Bergman 1999]. Similar problem can be considered to arise also in such different application areas as robotics, or automatic transcription of polyphonic music. In robotics, the problem can arise in determining the positions of humans surrounding the robot [Schulz et al. 2001]. In the latter problem, "targets" are, e.g., instruments, and their state is the note or chord that is played at current time [Klapuri 2004].

1.1 Goal

The traditional approach for building tracking systems separates the following subproblems to be solved [Blackman and Popoli 1999; Goodman et al. 1997].

1. Kinematic filtering of single targets, i.e. estimation of the geokinematic position of a target given noisy or indirect measurements.
2. Attribute fusion of single targets, i.e. estimation of some other characteristics, e.g. identification and type information, given noisy or indirect measurements.
3. Data-association, i.e. estimation of the measurement-target correspondence.

4. Initiation and track maintenance, i.e. estimation of the number of targets currently in the surveillance region.

This decomposition has been more or less the *de facto* standard for solving tracking problems. The popularity of the separation is probably based on the existing algorithms for "solving" the subproblems[1]. However, the combination of the algorithms "solving" the subproblems has been rather difficult.

The separation of the problem into the above mentioned subproblems has been found to work rather well in practice in several situations. However, the implementation of such functionalities need to be specifically tailored for the purpose of each particular tracking situation. In addition, since the subproblems are considered as separate, one needs to make *hard decisions* on intermediate estimation problems. The author's opinion is that making such hard decisions can be considered justified, if the uncertainty of the decisions is rather low. That is, the data that is used in estimation is rather accurate. If the estimation problem is very challenging, i.e. the data rate or accuracy is poor, making such decisions is hard, and wrong decisions are easily made. And what follows is that the tracking result is unreliable.

Goodman et al. [1997] refer such an estimation framework as "indirect", in which an intermediate estimation problem is solved before the final objective. This thesis aims to providing a "direct" estimation framework, in which all the above estimation problems are considered as one. The main purpose of the thesis is to bind together the recent development in Bayesian target tracking, that concerns the joint estimation of target count, and their state, which may consist of geokinematic and attribute parts. This thesis is primarily based on the random set formalism presented for the purposes of data fusion in [Goodman et al. 1997]. The random set framework is the only rigorously formulated approach, that has been proposed for general multitarget tracking in the literature[2]. The application of the random set approach requires an approximative computational strategy. In this thesis, the selected strategy is the sequential Monte Carlo simulation.

In addition to reviewing the random set approach to tracking, this thesis contains basics of general probability theory as well as estimation theory, to be as self-contained as possible. The sequential Monte Carlo techniques are reviewed as well, in the depth required by the application. Few proofs are presented—the results are merely listed. As a theoretical perspective, the purpose of this thesis is to uncover the random set approach, providing a solid basis for further development of the algorithms presented in the literature, and means for bringing the algorithms into practical applications.

---

1. For example, the (extended) Kalman filter solves the kinematic filtering, the hidden Markov model filter solves the attribute fusion (if attributes are assumed discrete-valued), the joint probabilistic data-association solves the data-association, and the track maintenance is solved by expert systems. [Blackman and Popoli 1999; Korpisaari 2001]

2. There are approaches presented in the literature, that have been formulated differently. Most of them can essentially be considered as special cases of the random set approach. For more discussion, see, e.g., [Mahler 2003a].

1.2 Structure

Unfortunately, this thesis contains some mathematics, that may be rather laborious to follow by an engineering oriented person. For example, the probability theory is kept quite abstract, and cannot be considered a good tutorial to the subject. The references that are given in the text lead into better tutorial sources. The abstract probability theory is considered necessary, since the main goal of the thesis is to study random sets, which cannot be considered as any kind of random variables or vectors, which are covered in elementary probability theory. Most of the contents of the thesis, however, can be read by anyone with elementary mathematical background—most importantly by any engineering oriented person. One should keep in mind when reading, that the author is not a mathematician, but merely an engineer.

The thesis starts with a review of probability theory in Chapter 2. Recursive Bayesian estimation is discussed in brief in Chapter 3. The sequential Monte Carlo approach as a computational strategy for recursive Bayesian estimation is covered in Chapter 4. The theory related to random sets is summarised in Chapter 5. Chapter 6 provides a literature review of target tracking, and contains some of the basic models that are commonly used in tracking. The random set approach to target tracking is discussed in Chapter 7. The limited experiments are described in Chapter 8, and conclusions are drawn finally in Chapter 9. The Appendices are included for the thesis to be self-contained. Appendix A contains some results of analysis, measure theory, and integration. Appendix B contains a brief description of the notion of Bayesian network graphs of probabilistic models.

Chapter 2

Probability Theory

This chapter covers some probability theory that is needed for the applications presented in the following chapters. The reader is referred to Appendix A for definitions and results of analysis. Readers without elementary probability theory background are advised to look through first an introductory book, e.g. [Shiryaev 1996, Chapter I], or selected parts of [Feller 1971].

Section 2.1 starts the chapter with the definition of a probability space. Section 2.2 covers the definition of a random element, and some special cases, such as random variables, vectors, and processes. The density function of a random element is defined in Section 2.3, and independence of random elements in Section 2.4. The expectation of a random variable and some related concepts are given in Section 2.5, while Section 2.6 continues with the definitions of conditional expectations, and conditional probability. Convergence of random sequences of random variables is reviewed in Section 2.7. The definitions of some common random distributions and processes are given in Section 2.8. Finally, Section 2.9 summarises some concepts of estimation theory in terms of general probability theory.

2.1 Probability Space

In this thesis, a probability space is defined directly as a measure space, with the property that the measure of the whole space is one. No explicit "(Kolmogorov) axioms of probability" are given, since they are included in the definition of a general measure[1]. In many books of probability, the definition is given without reference to measure spaces. However, since measure spaces are needed in this thesis anyway, it is convenient to give the definition in this form.

Definition 2.1 *The measure space* $(\Omega, \mathcal{M}, P)$ *is a* ***probability space****, if* $P(\Omega) = 1$*. The set* Ω *is called the* ***sample space****, the* σ*-algebra* $\mathcal{M}$ *on* Ω *constitutes of the possible* ***events****, and* P *is the* ***probability measure****. The probability measure* P *defines probability* $P(E)$ *for all events* $E \in \mathcal{M}$*.*

1. For a historical review of the development of probability theory, see e.g. [Doob 1996].

An example of a probability space is given next. The example can be considered, indeed, a useful one for the purposes of intuition.

Example 2.2 Consider the case where the sample space is the real hypercube $\Omega = [0,1]^d$, $\mathcal{L}$ is the collection of all Lebesgue-measurable subsets in Ω, and $\lambda^d : \mathcal{L} \to [0,1]$ is the Lebesgue measure in $\mathbb{R}^d$. Then, the measure space $(\Omega, \mathcal{L}, \lambda^d)$ is a probability space. $\Diamond$

2.2 Random Elements

The probability space introduced in Section 2.1 is an abstraction that is hidden in many applications. The problems are often convenient to formulate using random elements[2] and their distributions.

Definition 2.3 *A* ***random element*** $\boldsymbol{x} : \Omega \to X$ *is a measurable mapping from the probability space* $(\Omega, \mathcal{M}, P)$ *to some measurable space* $(X, \mathcal{N})$*. That is, for all* $B \in \mathcal{N}$*,*

$$\overleftarrow{\boldsymbol{x}}(B) = \{\omega \in \Omega : \boldsymbol{x}(\omega) \in B\} \in \mathcal{M}$$

where the notation $\overleftarrow{\boldsymbol{x}}(B)$ *denotes the preimage of the set* B*.*

Definition 2.3 above ensures that the probability measure P defines a probability measure $P_{\boldsymbol{x}}$ for all measurable $B \in \mathcal{N}$ as follows

$$P_{\boldsymbol{x}}(B) \triangleq P(\overleftarrow{\boldsymbol{x}}(B)) \tag{2.1}$$

It is obvious that $(X, \mathcal{N}, P_{\boldsymbol{x}})$ is a probability space.

Quite often, the sample space X of the random element is a topological space, and the selected σ-algebra is the Borel sets, $\mathcal{N} = \mathcal{B}(X)$. In particular, when $X = \mathbb{R}$ and $\mathcal{N} = \mathcal{B}(\mathbb{R})$, the random element is referred to as a **random variable**. If $X = \overline{\mathbb{R}}$ (the extended real numbers) and $\mathcal{N} = \mathcal{B}(\overline{\mathbb{R}})$, the random element is referred to as an **extended random variable**. In applications, the most common random element is a random vector, which is introduced next.

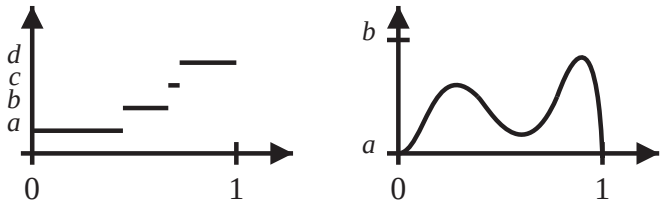
Definition 2.4 *A random element* $\underline{\boldsymbol{x}}(\omega) = [\boldsymbol{x}_1(\omega), \ldots, \boldsymbol{x}_d(\omega)]^T$ *is a* ***random vector*** *in* $\mathbb{R}^d$*, if* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ *are random variables.*

One could ask, whether the direct definition of a random vector as a random element $\underline{\boldsymbol{x}} : \Omega \to \mathbb{R}^d$ would be more general than the definition given above. It turns out, that every random element in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a random vector in the sense of Definition 2.4, and every d-dimensional random vector is a random element in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ [Shiryaev 1996, p. 177].

Definition 2.5 *We call the random element* $\boldsymbol{x} = (\boldsymbol{x}_t)_{t \in T}$ *a* ***stochastic process****, where* T *is an arbitrary index set, and each* $\boldsymbol{x}_t : \Omega \to X$ *is a random element.*

---

2. Sometimes referred to as random variable, or random object.



(a) (b)

Figure 2.1: (a) Example of a discrete random variable, when $S = \{a, b, c, d\}$. (b) Example of continuous random variable, when $S = [a, b] \subset \mathbb{R}$.

In this thesis, the index set T is some subset of $[0, \infty)$, or $\mathbb{N}$, and stands for time, or time index. Generally, if T is countable, the stochastic process is referred to as a random sequence, or a discrete-time process. If T is uncountable, then the process is sometimes referred to as a continuous-time process.

Using the probability space defined in Example 2.2, in the case $\Omega = [0, 1]$, two simple examples of random elements are given.

Example 2.6 A **discrete random variable** is a random element with a countable range. For example, in the case of a finite range, one can characterise the random variable as a map to a set S with finite number of elements, $S = \{s_1, \ldots, s_n\}$. The space S is endowed with discrete topology, and the random variable $\boldsymbol{x} : \Omega \to S$ is a measurable simple function. The mapping $\boldsymbol{x}$ can be defined indirectly, by assigning a probability $0 \leq P^*(s) \leq 1$ for each individual realisation $s \in S$, so that $\sum_{s \in S} P^*(s) = 1$. Then, one can define

$$\boldsymbol{x}(\omega) = \begin{cases} s_1, & \omega \leq P^*(s_1) \\ s_i, & \sum_{j=1}^{i-1} P^*(s_j) < \omega \leq \sum_{j=1}^{i} P^*(s_j) \qquad \text{for } i = 2, \ldots, n \end{cases} \tag{2.2}$$

The probability measure $P_{\boldsymbol{x}}$ can be given as follows.

$$P_{\boldsymbol{x}}(B) = \sum_{s \in B} P^*(s)$$

where $B \subset S$. This construction is straightforward to extend to cover random variables with a countable range, in an obvious manner. ◊

Figure 2.1 (a) shows an example of a discrete random variable when S has four elements, and $\boldsymbol{x} : \Omega \to \mathbb{R}$ is constructed as given in Equation (2.2). Another example of a random variable can be constructed so that one defines any other measurable mapping, e.g. continuous one, from the interval $[0, 1]$ to some subset of $\mathbb{R}$. The mapping of such a random variable is exemplified in Figure 2.1 (b).

Remark 2.7 The concept of a random element can be illustrated also through actual generation of pseudo-random numbers with a computer. Consider the examples given above. Many random number generation schemes are based on a

uniform $[0,1]$ random number generator (giving ω), followed by transformation methods (the random element $\boldsymbol{x}(\omega)$) [Robert and Casella 1999, p. 35].

2.3 Probability Density Function

We begin with a general definition of the probability density function, or density of a measure [Shiryaev 1996, p. 196].

Definition 2.8 *Suppose that a random element $\boldsymbol{x} : \Omega \to X$ has a probability measure $P_{\boldsymbol{x}}$ that is absolutely continuous with respect to a σ-finite measure μ on $(X, \mathcal{N})$. The **probability density function** (PDF) of $\boldsymbol{x}$ is such function $f_{\boldsymbol{x}} : X \to [0, \infty)$, that for all $E \in \mathcal{M}$,*

$$P_{\boldsymbol{x}}(E) = \int_E f_{\boldsymbol{x}}(x) \mathrm{d}\mu(x)$$

That is, $f_{\boldsymbol{x}} = \mathrm{d}P_{\boldsymbol{x}}/\mathrm{d}\mu$ is the Radon-Nikodym derivative (RND) of $P_{\boldsymbol{x}}$ with respect to μ.

The Radon-Nikodym Theorem (A.42) ensures that $f_{\boldsymbol{x}}$ exists, and is unique μ-a.s. Notice, however, that the existence of $f_{\boldsymbol{x}}$ requires $P_{\boldsymbol{x}}$ to be absolutely continuous with respect to μ.

Many times, probability measures are constructed directly using probability density functions. Consider a situation where $(X, \mathcal{N})$ is a measurable space, and μ a σ-finite measure. Then, suppose $f : X \to [0, \infty)$ is a nonnegative measurable function with the property that $\int f \mathrm{d}\mu = 1$. Then, a set function $P : \mathcal{N} \to [0, 1]$ can be constructed as follows.

$$P(E) = \int_E f(x) \mathrm{d}\mu(x) \tag{2.3}$$

Clearly, P defined above is a probability measure on $(X, \mathcal{N})$, and P is absolutely continuous with respect to μ. So, $f = \mathrm{d}P/\mathrm{d}\mu$, and $(X, \mathcal{N}, P)$ is a probability space.

2.4 Independence

The notion of independence of events, σ-algebras and random elements is often encountered. The definitions in this section are given according to [Shiryaev 1996, pp. 28–29; 179].

Definition 2.9 *Let $\mathcal{A} \subset \mathcal{M}$ be a finite collection of events. The events in $\mathcal{A}$ are **independent**, if the following condition is satisfied.*

$$P\Big(\bigcap_{M \in \mathcal{A}} M\Big) = \prod_{M \in \mathcal{A}} P(M)$$

The σ-algebras $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are independent, if the events $M_1, \ldots, M_n$ are independent, for each selection $M_i \in \mathcal{M}_i$.

Definition 2.10 *Let* $\{\boldsymbol{x}_i : i \in \mathcal{I}\}$ *be a collection of random elements* $\boldsymbol{x}_i : (\Omega, \mathcal{M}) \to (X_i, \mathcal{N}_i)$*. The random elements are (collectively)* ***independent****, if for every finite collection* $\{i_k\}_{k=1}^n \subset \mathcal{I}$ *the following condition is satisfied*

$$P(\boldsymbol{x}_{i_1} \in B_{i_1}, \ldots, \boldsymbol{x}_{i_n} \in B_{i_n}) = \prod_{k=1}^{n} P(\boldsymbol{x}_{i_k} \in B_{i_k})$$

for each selection of $B_{i_k} \in \mathcal{N}_{i_k}$*.*

Since it often occurs, that one manipulates functions of independent random variables, the following proposition is worth stating.

Proposition 2.11 *Let* $\{\boldsymbol{x}_i : i \in \mathcal{I}\}$ *be a collection of independent random elements,* $\boldsymbol{x}_i : (\Omega, \mathcal{M}) \to (X, \mathcal{N})$*, and let* $h : (X, \mathcal{N}) \to (Y, \mathcal{Q})$ *be a measurable function. Then the random elements* $h(\boldsymbol{x}_i)$ *are independent.*

Proof. Let $\{i_k\}_{k=1}^n \subset \mathcal{I}$ be an arbitrary finite subcollection of indices. Suppose $C_{i_1}, \ldots, C_{i_n} \in \mathcal{Q}$ are arbitrary measurable sets.

$$\begin{aligned} P\left(h(\boldsymbol{x}_{i_1}) \in C_{i_1}, \ldots, h(\boldsymbol{x}_{i_n}) \in C_{i_n}\right) &= P(\boldsymbol{x}_{i_1} \in A_{i_1}, \ldots, \boldsymbol{x}_{i_n} \in A_{i_n}) \\ &\overset{(*)}{=} \prod_{k=1}^{n} P(\boldsymbol{x}_{i_k} \in A_{i_k}) = \prod_{k=1}^{n} P(h(\boldsymbol{x}_{i_k}) \in C_{i_k}) \end{aligned}$$

where the equality $(*)$ is due to independence of $\boldsymbol{x}_i$, since the sets $A_i = \overleftarrow{h}(C_i) \in \mathcal{N}$, by assumption of measurability of h. □

2.5 Expectation

This section introduces the expectation of a random variable, and some related concepts. We begin with the definition [Shiryaev 1996, p. 182].

Definition 2.12 *The* ***expectation*** *of a (possibly extended) random variable* $\boldsymbol{x} : \Omega \to \overline{\mathbb{R}}$*, denoted by* $\mathsf{E}[\boldsymbol{x}]$*, is defined as*

$$\mathsf{E}[\boldsymbol{x}] \triangleq \int \boldsymbol{x} \mathrm{d}P$$

if the integral exists.

Notice, that $\mathsf{E}[\boldsymbol{x}]$ may be finite, or equal to ∞ or $-\infty$, or be undefined. Taking expectation of a random variable is integration, so the properties of general integrals, some of which are listed in Section A.5.2, apply also to the expectation operator.

The expectation is a basis for some other concepts. The following definition lists perhaps the most popular ones. Notice, that some of the expectations may be infinite, or undefined in the definitions [Shiryaev 1996, pp. 182; 234].

Definition 2.13 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two random variables, with finite expectations $\mathsf{E}[\boldsymbol{x}] = m_{\boldsymbol{x}}$ and $\mathsf{E}[\boldsymbol{y}] = m_{\boldsymbol{y}}$. Their* ***covariance*** *is defined as follows.*

$$\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \mathsf{E}[(\boldsymbol{x} - m_{\boldsymbol{x}})(\boldsymbol{y} - m_{\boldsymbol{y}})]$$

If $\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) = 0$, the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be ***uncorrelated****. The expectation $\mathsf{E}[\boldsymbol{x}^r]$, where $r > 0$, is called the* ***r'th moment*** *of $\boldsymbol{x}$. The* ***variance*** *of $\boldsymbol{x}$, denoted by $\mathsf{V}[\boldsymbol{x}]$ is defined as follows.*

$$\mathsf{V}[\boldsymbol{x}] \triangleq \mathsf{E}\left[(\boldsymbol{x} - m_{\boldsymbol{x}})^2\right] = \mathsf{cov}(\boldsymbol{x}, \boldsymbol{x}) = \mathsf{E}\left[\boldsymbol{x}^2\right] - m_{\boldsymbol{x}}^2$$

If $0 < \mathsf{V}[\boldsymbol{x}], \mathsf{V}[\boldsymbol{y}] < \infty$, the ***correlation coefficient*** *for $\boldsymbol{x}$ and $\boldsymbol{y}$ is*

$$\mathsf{corr}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \frac{\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathsf{V}[\boldsymbol{x}]\,\mathsf{V}[\boldsymbol{y}]}}$$

The next theorem gives a general method for assessing an upper bound for absolute variation of a random variable in terms of variance [Shiryaev 1996, p. 192].

Theorem 2.14 (Chebyshev) *Let $\boldsymbol{x}$ be a random variable with a finite expectation and variance. Then, the following inequality is satisfied.*

$$P(|\boldsymbol{x} - \mathsf{E}[\boldsymbol{x}]| \geq \varepsilon) \leq \frac{\mathsf{V}[\boldsymbol{x}^2]}{\varepsilon^2}$$

Since the tracking application discussed in Chapter 6 requires random vectors, the concepts of a covariance matrix and a mean vector are required [Shiryaev 1996, p. 235].

Definition 2.15 *Let $\underline{\boldsymbol{x}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d]^T$ be a random vector in $\mathbb{R}^d$. The* ***covariance matrix*** *of $\underline{\boldsymbol{x}}$ is a $d \times d$-matrix $\mathrm{R}_{\boldsymbol{x}}$ defined elementwise as follows.*

$$[\mathrm{R}_{\boldsymbol{x}}]_{ij} \triangleq \mathsf{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

The covariance matrix can also be given in a matrix algebra form

$$\mathrm{R}_{\boldsymbol{x}} = \mathsf{E}\left[(\boldsymbol{x} - \mathsf{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathsf{E}[\boldsymbol{x}])^T\right]$$

where the expectations are taken elementwise, e.g. $\mathsf{E}[\underline{\boldsymbol{x}}] = [\mathsf{E}[\boldsymbol{x}_1], \ldots, \mathsf{E}[\boldsymbol{x}_d]]^T = \underline{m}_{\boldsymbol{x}}$. The vector $\underline{m}_{\boldsymbol{x}}$ is called the ***mean vector*** *of $\underline{\boldsymbol{x}}$.*

The covariance matrix $\mathrm{R}_{\boldsymbol{x}}$ is always symmetric and positive semi-definite.

2.6 Conditional Expectation and Conditional Probability

The definition of the conditional expectation given in this section follows the presentation in [Shiryaev 1996, p. 213]. We first define the conditional expectation of a nonnegative random variable, and then give the general definition for the conditional expectation.

Definition 2.16 *Let $\mathcal{G} \subset \mathcal{M}$ be a sub-σ-algebra of $\mathcal{M}$. The conditional expectation of a nonnegative random variable $\boldsymbol{x}^+$ given a σ-algebra $\mathcal{G}$, denoted by $\mathsf{E}[\boldsymbol{x}^+ \mid \mathcal{G}]$, is a $\mathcal{G}$-measurable random variable that satisfies*

$$\int_A \boldsymbol{x}^+ \mathrm{d}P = \int_A \mathsf{E}\left[\boldsymbol{x}^+ \mid \mathcal{G}\right] \mathrm{d}P$$

for all $A \in \mathcal{G}$.

The conditional expectation $\mathsf{E}[\boldsymbol{x}^+ \mid \mathcal{G}]$ exists and is almost surely unique by the Radon-Nikodym theorem, for let $Q(B) = \int_B \boldsymbol{x}^+ \mathrm{d}P$, which is absolutely continuous with respect to P. Then, by the Radon-Nikodym theorem, Q has a RND with respect to P, which is $\mathcal{G}$-measurable. That is, $\mathsf{E}[\boldsymbol{x}^+ \mid \mathcal{G}] = \mathrm{d}Q/\mathrm{d}P$.

Definition 2.17 *The **conditional expectation** of an extended random variable **$\boldsymbol{x}$ given a σ-algebra** $\mathcal{G}$ is considered to be defined, if at least one of $\mathsf{E}[\boldsymbol{x}^+ \mid \mathcal{G}]$ and $\mathsf{E}[\boldsymbol{x}^- \mid \mathcal{G}]$ is finite P-a.s. Then, it is given by*

$$\mathsf{E}[\boldsymbol{x} \mid \mathcal{G}] \triangleq \mathsf{E}\left[\boldsymbol{x}^+ \mid \mathcal{G}\right] - \mathsf{E}\left[\boldsymbol{x}^- \mid \mathcal{G}\right]$$

The conditional probability of an event $A \in \mathcal{M}$ can be defined in terms of the conditional expectation, if one takes the characteristic function of A,

$$\chi_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

as a random variable.

Definition 2.18 *The **conditional probability** of an event $A \in \mathcal{M}$ with respect to a σ-algebra $\mathcal{G}$ is defined as follows.*

$$P(A \mid \mathcal{G}) \triangleq \mathsf{E}[\chi_A \mid \mathcal{G}]$$

where $\chi_A : \Omega \to \mathbb{R}$ is the characteristic function of the set A.

More often, in applications, one encounters conditional expectations or probabilities with respect to a random element [Shiryaev 1996, p. 214].

Definition 2.19 *Let $\boldsymbol{x} : \Omega \to \overline{\mathbb{R}}$ be an extended random variable and $\boldsymbol{y} : \Omega \to Y$ be a random element. Denote the σ-algebra generated by $\boldsymbol{y}$ as $\mathcal{G}_{\boldsymbol{y}} = \sigma(\boldsymbol{y})$. Then, the **conditional expectation of $\boldsymbol{x}$ with respect to $\boldsymbol{y}$** is defined as follows.*

$$\mathsf{E}[\boldsymbol{x} \mid \boldsymbol{y}] \triangleq \mathsf{E}[\boldsymbol{x} \mid \mathcal{G}_{\boldsymbol{y}}]$$

if the conditional expectation of $\boldsymbol{x}$ with respect to $\mathcal{G}_{\boldsymbol{y}}$ is defined.

Next definition gives a more intuitively appealing characterisation of the conditional expectation of a random variable given a random element [Shiryaev 1996, p. 220].

Definition 2.20 *Let $\boldsymbol{x} : \Omega \to \overline{\mathbb{R}}$ be a random variable with $\mathsf{E}[\boldsymbol{x}]$ defined, and let $\boldsymbol{y} : (\Omega, \mathcal{M}) \to (Y, \mathcal{N})$ be a random element. The conditional expectation of $\boldsymbol{x}$ with respect to the condition $\boldsymbol{y} = y$ is a measurable function $m(y)$, for which the following is satisfied*

$$\int_{\overleftarrow{\boldsymbol{y}}(B)} \boldsymbol{x} \mathrm{d}P = \int_B m(y) \mathrm{d}P_{\boldsymbol{y}}(y)$$

for all measurable $B \in \mathcal{N}$.

The function m always exists, since if one defines $Q(B) = \int_{\overleftarrow{\boldsymbol{y}}(B)} \boldsymbol{x} \mathrm{d}P$, then clearly Q is a signed measure that is absolutely continuous with respect to $P_{\boldsymbol{y}}$, so the Radon-Nikodym theorem ensures that m exists. In addition, $m(\boldsymbol{y})$ is $\mathcal{G}_{\boldsymbol{y}}$ measurable. Hence, Definitions 2.19 and 2.20 are dual, since one can reconstruct $\mathsf{E}[\boldsymbol{x} \mid \boldsymbol{y}]$ from $m(y) = \mathsf{E}[\boldsymbol{x} \mid \boldsymbol{y} = y]$ and vice versa.

The following theorem lists some basic properties of the conditional expectation [Shiryaev 1996, pp. 216–221].

Theorem 2.21 *Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent random variables, $\{\boldsymbol{z}_i\}_{i=1}^n$ are nonnegative random variables, f is a measurable function, and $\mathcal{H} \subset \mathcal{G}$ are σ-algebras. Then,*

1. $\mathsf{E}[\mathsf{E}[\boldsymbol{x} \mid \mathcal{G}] \mid \mathcal{H}] = \mathsf{E}[\boldsymbol{x} \mid \mathcal{H}]$ *(P-a.s.)*
2. $\mathsf{E}[f(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} = y] = \mathsf{E}[f(\boldsymbol{x}, y)]$ *($P_{\boldsymbol{y}}$-a.s.)*
3. $\mathsf{E}[\sum_{i=1}^n \boldsymbol{z}_n \mid \mathcal{G}] = \sum_{i=1}^n \mathsf{E}[\boldsymbol{z}_i \mid \mathcal{G}]$ *(P-a.s.)*

The conditional probability with respect to a random element can be given in an obvious manner, using Definitions 2.19 and 2.20.

Definition 2.22 *Let $\boldsymbol{y} : \Omega \to Y$ be a random element. The conditional probability of event A given the condition $\boldsymbol{y} = y$ is defined as follows.*

$$P(A \mid \boldsymbol{y} = y) \triangleq \mathsf{E}[\chi_A \mid \boldsymbol{y} = y]$$

Equivalently, $P(A \mid \boldsymbol{y} = y)$ can be defined as a measurable function $Y \to \mathbb{R}$, such that

$$P(A \cap \overleftarrow{\boldsymbol{y}}(B)) = \int_B P(A \mid \boldsymbol{y} = y) \mathrm{d}P_{\boldsymbol{y}}(y) \tag{2.4}$$

for every measurable $B \subset Y$.

The next example is from [Shiryaev 1996, p. 222], but it is modified to cover general product spaces instead of $\mathbb{R} \times \mathbb{R}$.

Example 2.23 Consider two random elements $\boldsymbol{x}$ and $\boldsymbol{y}$ with images on measure spaces $(X, \mathcal{N}_X, \mu)$ and $(Y, \mathcal{N}_Y, \nu)$, respectively. Suppose, that $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$ is a probability density function of $\boldsymbol{x}$ and $\boldsymbol{y}$ with respect to the σ-finite product measure $(\mu \times \nu)$,

$$P((\boldsymbol{x}, \boldsymbol{y}) \in B) = \int_B f_{\boldsymbol{x},\boldsymbol{y}}(x, y) \mathrm{d}(\mu \times \nu)(x, y), \quad \text{for all} \quad B \in \mathcal{N}_{X \times Y} \tag{2.5}$$

Then, $f_{\boldsymbol{x}} = \int f_{\boldsymbol{x},\boldsymbol{y}}(x,y)\mathrm{d}\nu(y)$ is the RND $\mathrm{d}P_{\boldsymbol{x}}/\mathrm{d}\mu$, and similarly $f_{\boldsymbol{y}} = \mathrm{d}P_{\boldsymbol{y}}/\mathrm{d}\nu$ according to Fubini's theorem. Define the function $f_{\boldsymbol{x}|\boldsymbol{y}}$ as follows.

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y) = \begin{cases} \dfrac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)}, & f_{\boldsymbol{y}}(y) > 0 \\ 0, & f_{\boldsymbol{y}}(y) = 0 \end{cases} \tag{2.6}$$

Then, it follows that $f_{\boldsymbol{x}|\boldsymbol{y}} = \mathrm{d}P(\boldsymbol{x} \mid \boldsymbol{y})/\mathrm{d}\mu$ is the density of the conditional probability distribution. That is,

$$P(\boldsymbol{x} \in C \mid \boldsymbol{y} = y) = \int_C f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y)\mathrm{d}\mu(x) \tag{2.7}$$

Clearly $f_{\boldsymbol{x}|\boldsymbol{y}}$ as defined in Equation (2.6) is μ-measurable for all y. To verify Equation (2.7), it is sufficient to verify Equation (2.4) in the case $A = \overleftarrow{\boldsymbol{x}}(C)$.

$$\begin{aligned} \int_B \left[\int_C f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y)\mathrm{d}\mu(x)\right] \mathrm{d}P_{\boldsymbol{y}}(y) &= \int_B \left[\int_C f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y)\mathrm{d}\mu(x)\right] f_{\boldsymbol{y}}(y)\mathrm{d}\nu(y) \\ = \int_{C\times B} f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y) f_{\boldsymbol{y}}(y)\mathrm{d}(\mu\times\nu)(x,y) &= \int_{C\times B} f_{\boldsymbol{x},\boldsymbol{y}}(x,y)\mathrm{d}(\mu\times\nu)(x,y) \\ &= P(\overleftarrow{\boldsymbol{x}}(C) \cap \overleftarrow{\boldsymbol{y}}(B)) \end{aligned}$$

according to Fubini's theorem. ◊

The conditional probability $P(A \mid \mathcal{G})(\omega)$ satisfies the following properties for a.e. $\omega \in \Omega$.[3]

1. $P(\Omega \mid \mathcal{G})(\omega) = 1$
2. $P(\emptyset \mid \mathcal{G})(\omega) = 0$
3. $P(\bigcup_{k=1}^{\infty} A_k \mid \mathcal{G}) = \sum_{k=1}^{\infty} P(A_k \mid \mathcal{G})$, for disjoint $A_k \in \mathcal{M}$.

That is, the conditional probability $P(\cdot \mid \mathcal{G})$ determines a probability measure for a.e. $\omega \in \Omega$. It is desired, that the conditional probability satisfies the conditions 1–3 for every $\omega \in \Omega$. The following definition gives a characterisation of such conditional probabilities [Shiryaev 1996, p. 226–227].

Definition 2.24 *Let $\mathcal{G} \subset \mathcal{M}$ be a sub-σ-algebra of $\mathcal{M}$. A function $P(\omega, B)$ defined for all $\omega \in \Omega$ and $B \in \mathcal{M}$ is a* ***regular conditional probability*** *with respect to $\mathcal{G}$, if*

1. *For each $\omega \in \Omega$, the function $P(\omega, \cdot)$ is a probability measure on $\mathcal{M}$.*
2. *For each $B \in \mathcal{M}$, the function $P(\cdot, B)$ is a variant of the conditional probability, i.e. $P(\omega, B) \overset{\text{a.s.}}{=} P(B \mid \mathcal{G})(\omega)$.*

Let $\boldsymbol{x} : \Omega \to X$ be a random element with values in a measurable space $(X, \mathcal{N})$. A function $Q(\omega, B)$ defined for $\omega \in \Omega$ and $B \in \mathcal{N}$ is a ***regular conditional distribution of $\boldsymbol{x}$*** *with respect to $\mathcal{G}$ if the following conditions are satisfied.*

1. *For each $\omega \in \Omega$, the function $Q(\omega, \cdot)$ is a probability measure on $\mathcal{N}$.*

3. The two first are obvious, while the third follows from property 3 in Theorem 2.21.

2. *For each* $B \in \mathcal{N}$*, the function* $Q(\cdot, B)$ *is a variant of the conditional probability with respect to* $\mathcal{G}$*, i.e.* $Q(\omega, B) \overset{\text{a.s.}}{=} P(\boldsymbol{x} \in B \mid \mathcal{G})(\omega)$.

Regular conditional probabilities are convenient in practice, since for example the conditional expectations can be given in terms of the conditional probability, as stated in the following theorem [Shiryaev 1996, p. 227].

Theorem 2.25 *If* $P(A \mid \mathcal{G})$ *is a regular conditional probability, then*

$$\mathsf{E}\left[\boldsymbol{x} \mid \mathcal{G}\right](\omega) \overset{\text{a.s.}}{=} \int \boldsymbol{x}(\omega')\mathrm{d}P(\omega' \mid \mathcal{G})(\omega)$$

The following theorem ensures, that conditional probability distributions exist for quite broad class of random elements [Shiryaev 1996, p. 229].

Theorem 2.26 *Let* $\boldsymbol{x} : \Omega \to X$ *be a random element with values in a Borel space* $(X, \mathcal{N})$*. Then there is a regular conditional distribution of* $\boldsymbol{x}$ *with respect to a* σ*-algebra* $\mathcal{G}$*.*

Especially, $\mathbb{R}^d$ is a Borel space, as well as any complete, separable metric space. See Definition A.35 for the exact definition of a Borel space.

Remark 2.27 The conditional expectation with respect to several random elements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, where $\boldsymbol{x}_i : (\Omega, \mathcal{M}) \to (X_i, \mathcal{N}_i)$, is defined as follows.

$$\mathsf{E}\left[A \mid \boldsymbol{x}, \ldots, \boldsymbol{x}_k\right] \triangleq \mathsf{E}\left[A \mid \sigma(\mathcal{G}_{\boldsymbol{x}_1}, \ldots, \mathcal{G}_{\boldsymbol{x}_k})\right]$$

where $\mathcal{G}_{\boldsymbol{x}_i} = \sigma(\boldsymbol{x}_i)$ is the σ-algebra generated by the random variable $\boldsymbol{x}_i$. The conditional probability is denoted similarly.

2.7 Convergence of Sequences of Random Elements

Consider a sequence of real-valued outcomes of an random experiment. When the average of increasing number of such outcomes is computed, it tends to converge to a certain value, namely the expectation. So, suppose in general a sequence $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ of random elements. The sequence can converge to a random element $\boldsymbol{x}$ in many different ways. This section outlines the basic types of convergence of random elements [Shiryaev 1996, pp. 252–261].

Definition 2.28 *Let* $\boldsymbol{x}$ *be a random variable, and* $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *a sequence of random variables. The sequence* $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *is said to converge to the random variable* $\boldsymbol{x}$

1. ***with probability one****, or* ***a.s.****, denoted by* $\boldsymbol{x}_k \xrightarrow{\text{a.s.}} \boldsymbol{x}$*, if*

$$P(\{\omega : \lim_{k\to\infty} \boldsymbol{x}_k(\omega) \neq \boldsymbol{x}(\omega)\}) = 0$$

2. ***in mean of order*** p, *denoted by* $\boldsymbol{x}_k \xrightarrow{L^p} \boldsymbol{x}$, *if*

$$\lim_{k\to\infty} \mathsf{E}\left[|\boldsymbol{x}_k - \boldsymbol{x}|^p\right] = 0$$

where $p > 0$. *In the special case when* $p = 2$, *the convergence is referred to as* ***mean square convergence***, *and denoted*[4] *by* $\text{l.i.m.}\, \boldsymbol{x}_k = \boldsymbol{x}$.

3. ***in probability***, *denoted by* $\boldsymbol{x}_k \xrightarrow{P} \boldsymbol{x}$, *if for all* $\epsilon > 0$,

$$\lim_{k\to\infty} P(|\boldsymbol{x}_k - \boldsymbol{x}| > \epsilon) = 0$$

4. ***in distribution***, *denoted by* $\boldsymbol{x}_k \xrightarrow{d} \boldsymbol{x}$, *if*

$$\lim_{k\to\infty} \mathsf{E}\left[f(\boldsymbol{x}_k)\right] = \mathsf{E}\left[f(\boldsymbol{x})\right]$$

for each bounded and continuous function f.

The almost sure convergence (1.) is meaningful to consider for random elements in any topological space. The convergence modes in mean of order p and in probability (2. and 3.) can be generalised into random elements $\boldsymbol{x}$, $\boldsymbol{x}_k$ in a metric space (X, d), by replacing $|\boldsymbol{x}_k - \boldsymbol{x}|$ with $d(\boldsymbol{x}_k, \boldsymbol{x})$. The convergence in distribution (4.) can be considered even for random elements defined in different probability spaces.

The convergence modes in Definition 2.28 are ordered according to their "strength", since the following implications can be shown valid [Shiryaev 1996, p. 256].

Theorem 2.29

$$\begin{aligned}
\boldsymbol{x}_k \xrightarrow{\text{a.s.}} \boldsymbol{x} &\implies \boldsymbol{x}_k \xrightarrow{P} \boldsymbol{x} \\
\boldsymbol{x}_k \xrightarrow{L^p} \boldsymbol{x} &\implies \boldsymbol{x}_k \xrightarrow{P} \boldsymbol{x} \\
\boldsymbol{x}_k \xrightarrow{P} \boldsymbol{x} &\implies \boldsymbol{x}_k \xrightarrow{d} \boldsymbol{x}
\end{aligned}$$

Finally, the two most popular limit theorems for independent and identically distributed (IID) random variables are stated [Shiryaev 1996, pp. 326; 389–391].

Theorem 2.30 (Central Limit Theorem) *Let* $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *be a sequence of IID nondegenerate, i.e.* $\mathsf{V}[\boldsymbol{x}_1] > 0$, *random variables with finite second moments. Let* $\boldsymbol{s}_n = \sum_{k=1}^{n} \boldsymbol{x}_k$. *Then,*

$$\frac{\boldsymbol{s}_n - \mathsf{E}[\boldsymbol{s}_n]}{\sqrt{\mathsf{V}[\boldsymbol{s}_n]}} = \sqrt{\frac{n}{\mathsf{V}[\boldsymbol{x}_1]}} \left(\frac{\boldsymbol{s}_n}{n} - \frac{\mathsf{E}[\boldsymbol{s}_n]}{n} \right) \xrightarrow{d} N(0, 1)$$

4. "l.i.m." stands for "limit in the mean".

Theorem 2.31 (Strong Law of Large Numbers) *Suppose that* $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ *is a sequence of IID random variables with* $\mathsf{E}[|\boldsymbol{x}_1|] < \infty$. *Denote* $\boldsymbol{s}_n = \sum_{k=1}^{n} \boldsymbol{x}_k$. *Then,*

$$\frac{\boldsymbol{s}_n}{n} \xrightarrow{\text{a.s.}} m$$

where $m = \mathsf{E}[\boldsymbol{x}_1]$.

These results, while being merely theoretical, are often referred to as, when the Monte Carlo methods are considered. The Monte Carlo methods are discussed in Chapter 4.

2.8 Common Distributions and Processes

This section covers the definitions of some common random distributions and stochastic processes. The emphasis is on presenting the distributions that are needed in Chapters 6 and 7.

2.8.1 Gaussian Distribution

The Gaussian distribution is by far the most frequently used random distribution in applications. The Gaussian distribution is most often used for practical reasons, since it has many "user-friendly" properties[5]. Most commonly, the Gaussian distribution is defined directly through a probability density function. We introduce the Gaussian distribution via the characteristic function, to include also the degenerate case [Shiryaev 1996, p. 275].

Definition 2.32 *The* ***characteristic function*** $\phi_{\underline{\boldsymbol{x}}} : \mathbb{R}^d \to \mathbb{C}$ *of a random vector* $\underline{\boldsymbol{x}}$ *is defined as follows.*

$$\phi_{\underline{\boldsymbol{x}}}(\underline{t}) = \mathsf{E}\left[\exp(i\underline{t}^T\underline{\boldsymbol{x}})\right] = \int_{\mathbb{R}^d} \exp(i\underline{t}^T\underline{x}) \mathrm{d}P_{\underline{\boldsymbol{x}}}(\underline{x})$$

where i *is the imaginary unit, and the expectation (integral) of the complex-valued random variable is taken separately for the real and the imaginary parts.*

The characteristic function defines uniquely a random distribution. That is, if $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$ have the same characteristic function, their probability measures are also the same, $P_{\underline{\boldsymbol{x}}} \equiv P_{\underline{\boldsymbol{y}}}$.

The multivariate Gaussian distribution on $\mathbb{R}^d$ is completely defined by the mean vector and the covariance matrix [Shiryaev 1996, p. 299].

Definition 2.33 *A random vector* $\underline{\boldsymbol{x}}$ *is* ***Gaussian****, or* ***normally distributed****, if the characteristic function of* $\underline{\boldsymbol{x}}$ *has the form*

$$\phi_{\underline{\boldsymbol{x}}}(\underline{t}) = \exp\left[i\underline{t}^T\underline{m} - (1/2)\underline{t}^T(\mathrm{R}\underline{t})\right]$$

5. For example, sum of two independent Gaussian random variables is Gaussian. In recursive Bayesian estimation (Chapter 3), the linear-Gaussian case is one of the few special cases, that admit closed form solution.

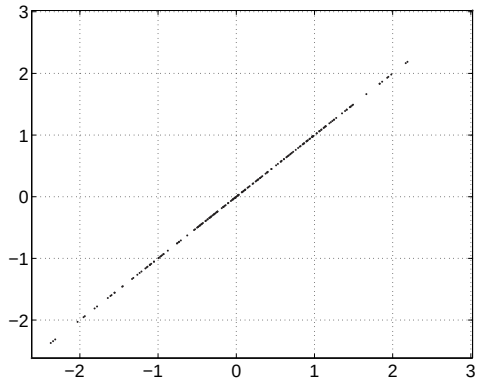


Figure 2.2: Generated random samples from a Gaussian distribution in $\mathbb{R}^2$ with a non-invertible covariance matrix.

where $\underline{m} \in \mathbb{R}^d$ *is the mean vector and* $\mathrm{R} \in \mathbb{R}^d \times \mathbb{R}^d$ *is a symmetric and positive semi-definite matrix. The probability measure of a Gaussian random vector with parameters* $\underline{m}$ *and* R *is denoted as* $N(B; \underline{m}, \mathrm{R}) \triangleq P_{\underline{\boldsymbol{x}}}(B)$.

If R is positive definite, hence invertible, the random measure $N(\cdot; \underline{m}, \mathrm{R})$ is absolutely continuous with respect to the Lebesgue measure λ^d in $\mathbb{R}^d$, and can be given as follows.

$$N(B; \underline{m}, \mathrm{R}) = \int_B f_{\underline{\boldsymbol{x}}}(\underline{x}; \underline{m}, \mathrm{R}) \mathrm{d}\lambda^d(\underline{x}) \tag{2.8}$$

with the probability density function given as

$$f_{\underline{\boldsymbol{x}}}(\underline{x}; \underline{m}, \mathrm{R}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathrm{R})}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{m})^T \mathrm{R}^{-1}(\underline{x} - \underline{m})\right] \tag{2.9}$$

where $\det(\mathrm{R})$ denotes the determinant of the covariance matrix, and R^{-1} is the inverse matrix.

In one-dimensional case, the Gaussian distribution with parameters $m \in \mathbb{R}$ and $r = \mathsf{V}[\boldsymbol{x}] = 0$ is a degenerate random variable, i.e. $P(\boldsymbol{x} = m) = 1$. In $\mathbb{R}^2$, an example of a Gaussian distribution with a non-invertible covariance matrix $[\mathrm{R}]_{ij} = 1$ is depicted in Figure 2.2.

2.8.2 Markov Chain

Markov chains are quite simple processes, that admit computationally convenient properties. They are, however, sufficiently general to characterise properties of many practical processes, and thus appear frequently in applications[6]. The following definition is given according to [Shiryaev 1996, p. 564].

6. For example, the recursive Bayesian estimation discussed in Chapter 3, is based on a Markov chain model of the estimated quantity.

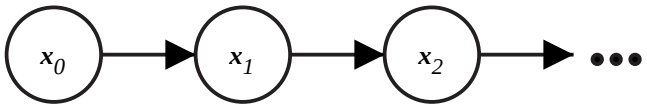


Figure 2.3: A Bayesian network graph representing a Markov chain.

Definition 2.34 *Let $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \cdots \subset \mathcal{M}$ be a sequence of σ-algebras, and let $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ be a sequence of random elements $\boldsymbol{x}_k : \Omega \to X$, such that $\sigma(\boldsymbol{x}_k) \subset \mathcal{M}_k$, i.e., each $\boldsymbol{x}_k$ is $\mathcal{M}_k$-measurable. The sequence $\boldsymbol{x}_k$ is a **Markov chain** (with respect to P), if*

$$P(\boldsymbol{x}_k \in B \mid \mathcal{M}_n) \overset{\text{a.s.}}{=} P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_n)$$

for all $k \geq n \geq 0$ and all measurable B.

If all singletons are measurable, i.e. $\{x\} \in \mathcal{N}$, the space $(X, \mathcal{N})$ is called a **phase space**. Then, the Markov chain is characterised by the initial distribution $\pi(B)$, and the one-step transition probabilities $P_k(x; B)$ such that

$$\pi(B) = P(\boldsymbol{x}_0 \in B) \qquad P_k(x; B) \overset{\text{a.s.}}{=} P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1} = x)$$

A regular conditional probability $P_k(x_{k-1}; B)$ is often called a **transition kernel** [Robert and Casella 1999, p. 141]. In this thesis, only such Markov chains are considered, that admit a transition kernel. A Markov chain expressed as a Bayesian network is depicted in Figure 2.3.

Example 2.35 If X is a finite set with n elements, the conditional probability distributions $P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1})$ can be characterised by the $n \times n$ **transition matrices** $[\mathrm{P}_k]_{ij} = P(\boldsymbol{x}_k = x_j \mid \boldsymbol{x}_{k-1} = x_i)$. ◊

2.8.3 Poisson Process

A continuous-time process $(\boldsymbol{n}_t)_{t\geq 0}$ is a counting process, if $\boldsymbol{n}_t : \Omega \to \mathbb{N}$ represents the total number of events that have occurred up to time t [Ross 1983, p. 31]. Consequently, if $s < t$, then $\boldsymbol{n}_t - \boldsymbol{n}_s$ is the number of events that have occurred in interval $(s, t]$. A counting process has independent increments, if the numbers of events occurring in disjoint time intervals are independent.

Definition 2.36 *A counting process $(\boldsymbol{n}_t)_{t\geq 0}$ is a **Poisson process** with rate $\eta \geq 0$, if the following properties are satisfied.*

1. *$\boldsymbol{n}_0 \equiv 0$.*
2. *The process has independent increments.*
3. *The number of events in any interval of length t is Poisson-distributed. That is, for all $s, t \geq 0$,*

$$P(\boldsymbol{n}_{t+s} - \boldsymbol{n}_s = k) = \frac{(\eta t)^k}{k!} \exp(-\eta t), \quad k \in \mathbb{N} \tag{2.10}$$

where the parameter η determines the mean number of events occurring within one time unit[7].

The Poisson process is quite often used as a model for arrival times in queueing applications.

2.9 Estimation Theory

This section covers some basic definitions and concepts of estimation theory. As a whole, the field of estimation theory is wide, and this section is just a tip of the iceberg. The reader with no prior knowledge on estimation theory can find a good elementary introduction, for example, in [Kay 1993].

There are two schools in estimation theory: the Bayesian and the classical[8]. The classical estimation is based on the assumption that the estimated parameter ϕ is nonrandom—fixed but unknown. In Bayesian estimation, the estimated parameter $\boldsymbol{\phi}$ is considered to be a random element in the same probability space as the data.

2.9.1 Classical Estimation

Suppose that there is a parametrised family of random elements of the form $\boldsymbol{x}^\phi : \Omega \to X$ where $\phi \in \Phi$. Similarly, there is a parametrised family of random measures corresponding the random elements $P_\phi \triangleq P_{\boldsymbol{x}^\phi}$.

Definition 2.37 *Suppose $\boldsymbol{x}_1^\phi, \ldots, \boldsymbol{x}_k^\phi$ are random elements with a fixed ϕ. A function $\varphi_k : X^k \to \Phi$ is an **estimator** of the parameter ϕ.*

It is clear, that the above definition is very general, and does not assume any "good" properties for the estimator. The rest of this section is dedicated to summarising the definitions of some desired properties of estimators [Kay 1993, p. 160; van Trees 1968, pp. 64; 70].

Definition 2.38 *Suppose $\Phi = \mathbb{R}^d$. That is, $\underline{\phi}$ is a vector parameter. Denote $\widehat{\underline{\boldsymbol{\phi}}}_k = \varphi_k(\boldsymbol{x}_1^\phi, \ldots, \boldsymbol{x}_k^\phi)$. The estimator φ_k is **unbiased**, if*

$$\mathsf{E}\left[\widehat{\underline{\boldsymbol{\phi}}}_k\right] = \underline{\phi}, \qquad \forall \underline{\phi} \in \Phi$$

*where the expectation is taken componentwise. The estimators $(\varphi_k)_{k=1}^\infty$ are said to be **asymptotically unbiased**, if*

$$\mathsf{E}\left[\widehat{\underline{\boldsymbol{\phi}}}_k\right] \xrightarrow{k\to\infty} \underline{\phi}, \qquad \forall \underline{\phi} \in \Phi$$

The following property can be considered for more generally, in any metric space.

---

7. $\mathsf{E}\left[\boldsymbol{n}_{t+s} - \boldsymbol{n}_s\right] = \sum_{k=0}^\infty k \frac{(\eta t)^k}{k!} e^{-\eta t} = e^{-\eta t} \eta t \sum_{k=0}^\infty \frac{(\eta t)^k}{k!} = e^{-\eta t} e^{\eta t} \eta t = \eta t$
8. also referred to as non-Bayesian, orthodox, and Fisher.

Definition 2.39 *Suppose that (Φ, d) is a metric space. The estimators φ_k are **consistent**, if the estimators converge in probability to the true value. That is, denote $\widehat{\boldsymbol{\phi}}_k = \varphi_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, then for all $\epsilon > 0$*

$$P(d(\widehat{\boldsymbol{\phi}}_k, \phi)) > \epsilon) \xrightarrow{k\to\infty} 0$$

2.9.2 Bayesian Estimation

As mentioned, In Bayesian estimation it is assumed that the estimated parameter $\boldsymbol{\phi}$ is a random element. So, suppose that $\boldsymbol{\phi} : \Omega \to \Phi$ and $(\boldsymbol{x}_k : \Omega \to X_k)_{k\in\mathbb{N}}$ are random elements. Based on these assumptions, the definition of a general Bayes estimator can be given [Kay 1993, p. 342].

Definition 2.40 *Denote $\widehat{\boldsymbol{\phi}}_k = \varphi_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. The **Bayes risk** of an estimator φ_k is the expected cost,*

$$R(\varphi_k) = \mathsf{E}\left[c(\widehat{\boldsymbol{\phi}}_k, \boldsymbol{\phi})\right]$$

*where $c : \Phi^2 \to [0, \infty]$ is the **Bayes cost** function. The **Bayes estimator** φ_k is the one minimising the Bayes risk. That is, φ_k is the Bayes estimator, if*

$$R(\varphi_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)) \leq R(\varphi'_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k))$$

for any other estimator φ'_k.

The definition of the Bayes estimator is very general. The choice of the cost function c affects the nature of the estimator greatly. In the following, some of the most common choices of the cost function are listed. The cost functions are given in the case of a vector parameter, $\Phi = \mathbb{R}^d$ [Kay 1993, pp. 342–344].

The **mean square error** cost: $$c(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\|^2 \tag{2.11}$$

The **absolute error** cost: $$c(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\| \tag{2.12}$$

The **hit-or-miss** cost: $$c(\underline{x}, \underline{y}) = \begin{cases} 0, & \|\underline{x} - \underline{y}\| \leq \Delta/2 \\ 1, & \|\underline{x} - \underline{y}\| > \Delta/2 \end{cases} \tag{2.13}$$

where Δ is a "small" positive number. The mean square error cost leads into the expected *a posteriori* (EAP) estimator[9]. The EAP estimator can be given as follows

$$\varphi_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = \mathsf{E}\left[\underline{\boldsymbol{\phi}} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\right]$$

The absolute error cost leads into the median estimator. The hit-or-miss cost function, as Δ is decreased, leads into the maximum *a posteriori* (MAP) estimator. If the joint distribution of $\boldsymbol{\phi}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ has a continuous density function with respect to the Lebesgue measure, then the MAP estimator can be given as follows.

$$\varphi_k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = \arg\max_{\underline{\phi}} f_{\underline{\boldsymbol{\phi}}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k}(\underline{\phi} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$$

9. Often also referred to as the mimimum mean square error (MMSE) estimator.

if the maximum exists. The given cost functions can be extended to any metric space, by replacing the norm $\|\underline{x} - \underline{y}\|$ by $d(x, y)$ in Equations (2.11)–(2.13). This way, one can obtain Bayes estimators in general metric spaces.

Remark 2.41 The comparison of Bayes estimators and classical estimators can be considered cumbersome. This is due to the different philosophy behind the methods. The Bayes estimators are *Bayes optimal*, given the cost function they minimise. However, it is natural that consistency (when applicable) is a desirable property of any estimator, also a Bayesian one. It is rather obvious, that the Bayes estimates (of random variables or random vectors) are not, in general, unbiased, since the prior distribution of the estimated parameter affects the estimator. Asymptotical unbiasedness is a weaker condition, that should be fulfilled also by the Bayes estimators (of random variables or random vectors).

2.9.3 Computation of Bayes Estimates

One may ask how to compute Bayes estimates in practice? This section provides a partial answer, by providing theoretical background for computing the posterior distribution. The practical computation or approximation of the posterior distribution is discussed further in Chapters 3 and 4.

Suppose that there is a regular conditional probability of $\boldsymbol{\phi}$ with respect to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. Then, one may define the following conditional probability.

$$P_{\boldsymbol{\phi}|x_1,\ldots,x_k}(B) \triangleq P(\boldsymbol{\phi} \in B \mid \boldsymbol{x}_1 = x_1, \ldots, \boldsymbol{x}_k = x_k) \tag{2.14}$$

This conditional distribution is referred to as the **posterior distribution** (given the data $\boldsymbol{x}_1 = x_1, \ldots, \boldsymbol{x}_k = x_k$). So, suppose that the posterior distribution is given. Then, one can obtain the Bayes estimate $\varphi_k(x_1, \ldots, x_k)$, since the Bayes estimator is the function φ_k that minimises

$$\int_{\Phi} c(\phi, \varphi_k(x_1, \ldots, x_k)) \mathrm{d}P_{\boldsymbol{\phi}|x_1,\ldots,x_k}(\phi)$$

separately for all $x_1, \ldots, x_k \in X$. This follows from the fact that the Bayes risk can be written as follows.

$$\begin{aligned} R(\varphi_k) &= \int_{X^n \times \Phi} c(\varphi_k(x_1, \ldots, x_k), \phi) \mathrm{d}P_{\boldsymbol{\phi}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k}(\phi, x_1, \ldots, x_k) \\ &= \int_{X^n} \left[\int_{\Phi} c(\varphi_k(x_1, \ldots, x_k), \phi) \mathrm{d}P_{\boldsymbol{\phi}|x_1,\ldots,x_k}(\phi) \right] \mathrm{d}P_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k}(x_1, \ldots, x_k) \end{aligned}$$

This justifies the idea of Bayesian estimation: given the data $\boldsymbol{x}_1 = x_1, \ldots, \boldsymbol{x}_k = x_k$, the posterior distribution of the estimated parameter, $P_{\boldsymbol{\phi}|x_1,\ldots,x_k}$, contains all the required information to extract any Bayes estimate. So, how to obtain the posterior distribution given in Equation (2.14)? The generalised Bayes theorem given next will provide means for that [Shiryaev 1996, p. 231].

Theorem 2.42 (Bayes) *Suppose $\boldsymbol{x} : (\Omega, \mathcal{M}) \to (X, \mathcal{N}_X)$ is a random element, and that there is a regular conditional probability $P(B \mid \boldsymbol{x} = x)$. In addition, suppose that there is a nonnegative function $\rho(\omega; x)$ that is measurable with respect to jointly both the arguments, such that*

$$P(B \mid \boldsymbol{x} = x) = \int_B \rho(\omega; x) \mathrm{d}\nu(\omega), \qquad B \in \mathcal{G}$$

where ν is a σ-finite measure on $(\Omega, \mathcal{G})$, and $\mathcal{G} \subset \mathcal{M}$ is a σ-algebra. Let $g : (X, \mathcal{N}_X) \to (\mathbb{R}, \mathcal{B}(R))$ be a measurable function, with $\mathsf{E}[|g(\boldsymbol{x})|] < \infty$. Then,

$$\mathsf{E}[g(\boldsymbol{x}) \mid \mathcal{G}] \overset{\text{a.s.}}{=} \frac{\int g(x)\rho(\omega; x)\mathrm{d}P_{\boldsymbol{x}}(x)}{\int \rho(\omega; x)\mathrm{d}P_{\boldsymbol{x}}(x)}$$

Proof. Define the following function for all $B \in \mathcal{G}$.

$$\begin{aligned} Q(B) &= \int_B g(\boldsymbol{x}(\omega))\mathrm{d}P(\omega) = \int g(x)P(B \mid \boldsymbol{x} = x)\mathrm{d}P_{\boldsymbol{x}}(x) \\ &= \int g(x) \left[\int_B \rho(\omega; x)\mathrm{d}\nu(\omega) \right] \mathrm{d}P_{\boldsymbol{x}}(x) \\ &= \int_B \left[\int g(x)\rho(\omega; x)\mathrm{d}P_{\boldsymbol{x}}(x) \right] \mathrm{d}\nu(\omega) \end{aligned}$$

where the last equality is due to Fubini's theorem. Similarly, one can write for all $B \in \mathcal{G}$

$$\begin{aligned} P(B) &= \int P(B \mid \boldsymbol{x})(\omega)\mathrm{d}P(\omega) = \int P(B \mid \boldsymbol{x} = x)\mathrm{d}P_{\boldsymbol{x}}(x) \\ &= \int_B \left[\int \rho(\omega; x)\mathrm{d}P_{\boldsymbol{x}}(x) \right] \mathrm{d}\nu(\omega) \end{aligned}$$

Then, the proposition follows from

$$\mathsf{E}[g(\boldsymbol{x}) \mid \mathcal{G}] = \frac{\mathrm{d}Q}{\mathrm{d}P} \overset{\text{a.s.}}{=} \frac{\mathrm{d}Q/\mathrm{d}\nu}{\mathrm{d}P/\mathrm{d}\nu}$$

where the last a.s.-equality is according to Theorem A.43. □

The following example gives the special case of the generalised Bayes theorem, that is particularly useful in applications [Shiryaev 1996, p. 233].

Example 2.43 Suppose $\boldsymbol{x} : (\Omega, \mathcal{M}) \to (X, \mathcal{N}_X)$ and $\boldsymbol{y} : (\Omega, \mathcal{M}) \to (Y, \mathcal{N}_Y)$ are random elements. Also, suppose that $P(\boldsymbol{y} \mid \boldsymbol{x} = x)$ is regular, and can be represented as follows.

$$P(\boldsymbol{y} \in B \mid \boldsymbol{x} = x) = \int_B f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x)\mathrm{d}\nu(y)$$

where ν is a σ-finite measure in $(Y, \mathcal{N}_Y)$, and $f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x)$ is jointly measurable with respect to both y and x. Then,

$$\mathsf{E}\left[g(\boldsymbol{x}) \mid \boldsymbol{y} = y\right] \overset{\text{a.s.}}{=} \frac{\int g(x) f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) \mathrm{d}P_{\boldsymbol{x}}(x)}{\int f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) \mathrm{d}P_{\boldsymbol{x}}(x)} \tag{2.15}$$

One can set $g = \chi_B$, where $B \in \mathcal{N}_X$, and get that

$$P(\boldsymbol{x} \in B \mid \boldsymbol{y} = y) \overset{\text{a.s.}}{=} \frac{\int_B f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) \mathrm{d}P_{\boldsymbol{x}}(x)}{\int f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) \mathrm{d}P_{\boldsymbol{x}}(x)} \tag{2.16}$$

Especially, if $(\boldsymbol{x}, \boldsymbol{y})$ have a joint density function with respect to the σ-finite product measure $(\mu \times \nu)$ in $(X \times Y)$, one can define $f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x)$ as in Equation (2.6). Then, Equations (2.15) and (2.16) can be restated by replacing "$\mathrm{d}P_{\boldsymbol{x}}(x)$" with "$f_{\boldsymbol{x}}(x)\mathrm{d}\mu(x)$". In particular, the following is true.

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x \mid y) \overset{\text{a.s.}}{=} \frac{f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) f_{\boldsymbol{x}}(x)}{\int f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x) f_{\boldsymbol{x}}(x) \mathrm{d}\mu(x)}$$

This is one of the most common forms of the Bayes theorem. ◊

Chapter 3

Recursive Bayesian Estimation

The main application in this thesis is target tracking based on the recursive Bayesian estimation framework. Recursive Bayesian estimation, often referred to as *filtering*[1], is a framework that is also used in many other applications, e.g. speech recognition, optimal control, computer vision, econometrics, and molecular biology, just to mention few [Doucet et al. 2001; Liu 2003].

This chapter outlines the theory of recursive Bayesian estimation. First, the underlying probabilistic model, the state-space model (SSM), is described in Section 3.1. Once the model is specified, Section 3.2 describes, in a conceptual level, how the inference in a SSM can be performed recursively. The inference is based on recursive computation of the posterior distribution, given the observed data. Having the recursion implemented, an estimate, or several estimates, can be extracted from the posterior distribution. Finally, Section 3.3 reviews some special cases, in which exact inference is computationally tractable. Some approximative methods are summarised as well, except for sequential Monte Carlo, for which Chapter 4 is devoted.

3.1 State-Space Model

Quite often, in applications, the probabilistic model is determined by a signal process, which is a Markov chain $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$, where $\boldsymbol{x}_k : \Omega \to X$. In addition, there is a separate observation process $(\boldsymbol{y}_k)_{k\in\mathbb{Z}_+}$, where $\boldsymbol{y}_k : \Omega \to Y_k$. The observations $\boldsymbol{y}_k$ are assumed conditionally independent given the current state $\boldsymbol{x}_k = x_k$. That is,

$$P(\boldsymbol{y}_k \in B \mid \boldsymbol{x}_{0:n}, \boldsymbol{y}_{1:n\setminus k}) \overset{\text{a.s.}}{=} P(\boldsymbol{y}_k \in B \mid \boldsymbol{x}_k) \tag{3.1}$$

$$P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{1:k-1}) \overset{\text{a.s.}}{=} P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1}) \tag{3.2}$$

for all $k \geq 1$ and $n \in \mathbb{N}$. The notation $\boldsymbol{x}_{a:b}$, where $a \leq b$, denotes the set $\{\boldsymbol{x}_a, \ldots, \boldsymbol{x}_b\}$. Similarly, $\boldsymbol{x}_{0:n} = x_{0:n}$ will denote $\boldsymbol{x}_i = x_i$ for all $0 \leq i \leq n$. In addition, $\boldsymbol{y}_{1:n\setminus k}$ means all $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, except $\boldsymbol{y}_k$.

1. Sometimes, the terms "stochastic filtering" and "optimal filtering" are used to distinguish recursive Bayesian estimation, e.g., from frequency-selective filtering in signal processing.

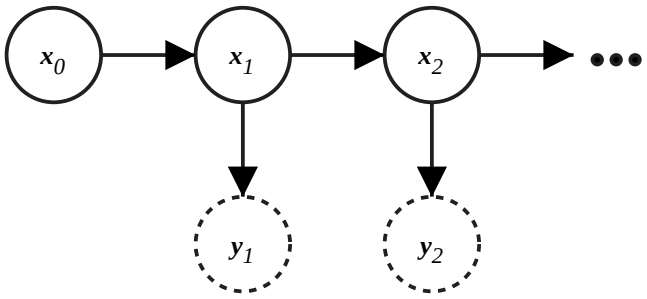


Figure 3.1: A Bayesian network graph representing a state-space model.

This general model is referred to as the **state-space model** (SSM). The model is depicted as a Bayesian network in Figure 3.1. Intuitively, SSM includes an unknown, dynamically changing parameter $\boldsymbol{x}_k$. There are noisy measurements $\boldsymbol{y}_k$, which depend on $\boldsymbol{x}_k$ in some known manner. In the following, the conditional probability distributions $P(\boldsymbol{y}_k \in B \mid \boldsymbol{x}_k)$ and $P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1})$ are referred to as the *measurement model*, and the *dynamic model*, respectively.

3.2 Bayes Recursion

Before the recursive computation of the posterior distribution can be presented, some additional regularity assumptions are needed. It is required that both the dynamic model and the measurement model are regular conditional probabilities. In addition, it is assumed that the measurement model admits the following representation

$$P(\boldsymbol{y}_k \in B \mid \boldsymbol{x}_k = x_k) = \int_B g_k(y_k \mid x_k) \mathrm{d}\nu_k(y_k)$$

where ν_k is a σ-finite measure on Y_k. That is, $g_k(y_k \mid x_k)$ is the RND of the measurement model. Since the posterior distributions are often considered, the following notation is introduced for conciseness.

$$\pi_{m|n}(B) \triangleq P(\boldsymbol{x}_m \in B \mid \boldsymbol{y}_{1:n} = y_{1:n})$$

If the above mentioned assumptions are satisfied, the propagation of the posterior distribution in a SSM can be carried out using a procedure that is referred to as the Bayes recursion.

Theorem 3.1 *The marginal probability measures for $k \geq 1$ can be obtained from the following recursion.*

$$\pi_{k|k-1}(B) \overset{\text{a.s.}}{=} \int P_k(x_{k-1}; B) \mathrm{d}\pi_{k-1|k-1}(x_{k-1}) \tag{3.3}$$

$$\pi_{k|k}(B) \overset{\text{a.s.}}{=} \frac{\int_B g_k(y_k \mid x_k) \mathrm{d}\pi_{k|k-1}(x_k)}{\int g_k(y_k \mid x_k) \mathrm{d}\pi_{k|k-1}(x_k)} \tag{3.4}$$

Proof. Equation (3.3) is due to the following equality, where one can replace $h(x) = \chi_{\overleftarrow{\boldsymbol{x}}_k(B)}(x)$ for any measurable B.

$$\begin{aligned}\mathsf{E}\left[h(\boldsymbol{x}_k) \mid \boldsymbol{y}_{1:k-1}\right] &\overset{\text{a.s.}}{=} \mathsf{E}\left[\mathsf{E}\left[h(\boldsymbol{x}_k) \mid \boldsymbol{y}_{1:k-1}, \boldsymbol{x}_{0:k-1}\right] \mid \boldsymbol{y}_{1:k-1}\right] \\ &\overset{\text{a.s.}}{=} \mathsf{E}\left[\mathsf{E}\left[h(\boldsymbol{x}_k) \mid \boldsymbol{x}_{k-1}\right] \mid \boldsymbol{y}_{1:k-1}\right]\end{aligned}$$

where the first equality follows from property 2 in Theorem 2.21, since $\sigma(\boldsymbol{y}_{1:k-1}) \subset \sigma(\boldsymbol{y}_{1:k-1}, \boldsymbol{x}_{0:k-1})$. The second equality is due to the independence assumption in Equation (3.2). The proof of Equation (3.3) is omitted for brevity. It is given, e.g., in [Doucet et al. 2001, pp. 39–41]. □

Suppose next, that the dynamic model, i.e. the Markov transition kernels, have a RND with respect to a σ-finite measure μ on X. That is,

$$P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1} = x_{k-1}) = P_k(x_{k-1}; B) = \int_B f_k(x_k \mid x_{k-1}) \mathrm{d}\mu(x_k)$$

In addition, suppose that the initial distribution $\pi_{0|0}$ has RND $p_{0|0} = \mathrm{d}\pi_{0|0}/\mathrm{d}\mu$. Then, the Bayes recursion can be given in terms of densities, as follows.

$$p_{k|k-1}(x_k) \overset{\text{a.s.}}{=} \int f_k(x_k \mid x_{k-1}) p_{k-1|k-1}(x_{k-1}) \mathrm{d}\mu(x_{k-1}) \tag{3.5}$$

$$p_{k|k}(x_k) \overset{\text{a.s.}}{=} \frac{g_k(y_k \mid x_k) p_{k|k-1}(x_k)}{\int g_k(y_k \mid x'_k) p_{k|k-1}(x'_k) \mathrm{d}\mu(x'_k)} \tag{3.6}$$

This form of the Bayes recursion can be considered sufficient for most applications. It is, however, sometimes convenient to consider the general form given in Equations (3.3) and (3.4). This is the case in Section 4.5, when sequential Monte Carlo methods are considered as sequential approximations of the posterior measure.

3.3 Inference

The computation of the posterior distribution using the Bayes recursion is a form of Bayesian inference[2]. In general, Equation (3.4) cannot be computed in a closed form. This section covers briefly methods for inference in recursive Bayesian estimation. First, the tractable special cases are covered in Section 3.3.1, and then some approximate solutions to the intractable cases are discussed in Section 3.3.2. The methods are not covered thoroughly, since the main emphasis in this thesis is on the sequential Monte Carlo approximation of the Bayes recursion, which is discussed in Chapter 4. Furthermore, approximate inference in some more application-specific models is discussed in Chapter 6.

2. In general, Bayesian inference may refer to computation of any posterior marginal distribution, or any Bayes estimate from some posterior distribution.

3.3.1 Exact Inference

There are two special cases, in which the Bayes recursion can be performed in a closed form. The first special case is when the random elements in the Markov chain are finite. Then, the SSM is referred to as a **hidden Markov model** (HMM)[3]. Inference in a HMM can be performed in a closed form, since the integrals in Equation (3.3) reduce to finite summations, and the marginal densities[4] $p_{k|k}$ can be represented as n-tuples[5]. The recursion can be given in the following form [Ristic et al. 2004, p. 9].

$$p_{k|k-1}(x_k) = \sum_{x \in X} f_k(x_k \mid x) p_{k-1|k-1}(x) \tag{3.7}$$

$$p_{k|k}(x_k) = \frac{g_k(y_k \mid x_k) p_{k|k-1}(x_k)}{\sum_{x \in X} g_k(y_k \mid x) p_{k|k-1}(x)} \tag{3.8}$$

The above HMM recursion requires $O(d^2)$ operations, since the prediction in Equation (3.7) requires d^2 summations and products. The update in Equation (3.8) requires $O(d)$ operations, since $p_{k|k}(x_k) \propto g_k(y_k \mid x_k) p_{k|k-1}(x_k)$.

The second model that admits a tractable inference is the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, and the conditional probabilities are linear-Gaussian, i.e., for all $B \in \mathcal{B}(\mathbb{R}^n)$ and $C \in \mathcal{B}(\mathbb{R}^m)$,

$$\begin{aligned} P(\underline{\boldsymbol{x}}_k \in B \mid \underline{\boldsymbol{x}}_{k-1} = x) &= N(B; \mathrm{A}_k \underline{x}, \mathrm{Q}_k) \\ P(\underline{\boldsymbol{y}}_k \in C \mid \underline{\boldsymbol{x}}_k = x) &= N(C; \mathrm{H}_k \underline{x}, \mathrm{R}_k) \end{aligned} \tag{3.9}$$

where A_k, H_k are $n \times n$ and $m \times n$ real matrices, and Q_k, R_k are positive definite and symmetric $n \times n$ and $m \times m$ matrices. In this case, the model is referred to as a **Kalman filter model** (KFM). In a KFM, the marginal densities[6] $p_{k|k}$ can be represented by their mean vectors and covariance matrices, since $\pi_{k|k}$ is always Gaussian. The update recursions for the mean vector and the covariance matrix can be given as follows [Murphy 2002, Section 3.6.1]. First, $\pi_{k|k-1}$ is a Gaussian with mean and covariance given in the following equation.

$$\underline{m}_{k|k-1} = \mathrm{A}_k \underline{m}_{k-1|k-1} \tag{3.10}$$

$$\mathrm{P}_{k|k-1} = \mathrm{A}_k \mathrm{P}_{k-1|k-1} \mathrm{A}_k^T + \mathrm{Q}_k \tag{3.11}$$

---

3. This terminology is not used by all. Some authors talk about HMMs, when they mean what is called a SSM here; some refer to SSMs, meaning the model what is introduced here to be a KFM.
4. RNDs with respect to the counting measure in the finite set X.
5. That is, the ordered sets with n elements. In this thesis, the term "vector" is designated to elements of a vector space.
6. RNDs with respect to the Lebesgue measure in $\mathbb{R}^n$.

Then, the moments of $\pi_{k|k}$ can be obtained as follows.

$$\begin{aligned} \underline{v} &= \underline{y}_k - \mathrm{H}_k \underline{m}_{k|k-1} \\ \mathrm{S} &= \mathrm{H}_k \mathrm{P}_{k|k-1} \mathrm{H}_k^T + \mathrm{R}_k \\ \mathrm{K} &= \mathrm{P}_{k|k-1} \mathrm{H}_k^T \mathrm{S}^{-1} \\ \underline{m}_{k|k} &= \underline{m}_{k|k-1} + \mathrm{K}\underline{v} \qquad (3.12) \\ \mathrm{P}_{k|k} &= \mathrm{P}_{k|k-1} - \mathrm{KSK}^T \qquad (3.13) \end{aligned}$$

where $\underline{v}$, S, and K are auxiliary variables, which are referred to as the innovation, covariance of the innovation, and the Kalman gain, respectively. That is, the KFM recursion requires only matrix multiplications and summations, and inversion of one positive definite and symmetric matrix.

The model in Equation (3.9) can be expressed also as a generative model as follows.

$$\begin{aligned} \underline{\boldsymbol{x}}_k &= \mathrm{A}_k \underline{\boldsymbol{x}}_{k-1} + \underline{\boldsymbol{u}}_k \\ \underline{\boldsymbol{y}}_k &= \mathrm{H}_k \underline{\boldsymbol{x}}_k + \underline{\boldsymbol{v}}_k \end{aligned} \qquad (3.14)$$

where $\underline{\boldsymbol{u}}_k$ and $\underline{\boldsymbol{v}}_k$ are independent zero-mean nondegenerate Gaussian random vectors with covariance matrices Q_k and R_k, respectively.

3.3.2 Approximate Inference

One straightforward method for approximating the Bayes recursion is to discretise the state space X [Arulampalam et al. 2002; Bergman 1999; Stone et al. 1999]. Then, the SSM reduces into a HMM, which admits exact inference. Often, discretisation leads to practically infeasible inference, since the number of discretisation points will become large. This is the case, e.g., when $X = \mathbb{R}^n$ where n is large. For example, in a six dimensional case, where each dimension is discretised into one hundred values leads to 10^{12} discretisation points, which can be considered infeasible in practice. This problem, sometimes referred to as the "curse of dimensionality", is stated to be overcome by the sequential Monte Carlo methods [Crisan and Doucet 2002, p. 744], which are discussed in Chapter 4. They can be considered to provide an alternative, "stochastic discretisation" method.

Many times, $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, i.e. $\underline{\boldsymbol{x}}_k$ and $\underline{\boldsymbol{y}}_k$ are random vectors, but the distributions are not exactly linear-Gaussian, but rather close to. That is, they are not strongly nonlinear, or non-Gaussian. Instead of Equation (3.14), one has a generative model that can be written as follows

$$\begin{aligned} \underline{\boldsymbol{x}}_k &= a_k(\underline{\boldsymbol{x}}_{k-1}, \underline{\boldsymbol{u}}_k) \\ \underline{\boldsymbol{y}}_k &= h_k(\underline{\boldsymbol{x}}_k, \underline{\boldsymbol{v}}_k) \end{aligned} \qquad (3.15)$$

where $a_k : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ and $h_k : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^m$ are measurable functions, and $\underline{\boldsymbol{u}}_k$ and $\underline{\boldsymbol{v}}_k$ are nondegenerate Gaussian random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. If a_k and h_k are differentiable, then a very popular method, that has also been

applied to real-life applications, is the extended Kalman filter (EKF) [Arulampalam et al. 2002; Fearnhead 1998]. EKF is based on a first-order approximation (local linearisation) of the functions a_k and h_k, in the neighbourhood of the previous posterior mean $\underline{m}_{k-1|k-1}$ and of the predicted mean $\underline{m}_{k|k-1}$, respectively. This linearisation essentially converts the model sequentially to a KFM, in which the update can be performed.

Recently, alternative methods have been proposed to replace the EKF. For example, the unscented Kalman filter (UKF) [Wan and van der Merwe 2000] and the central difference filter (CDF) [Ito and Xiong 2000; Nørgaard et al. 1998] have shown superior performance compared to the EKF, at least in some applications. All these algorithms suffer from the same drawback, though: they constrain the posterior distribution $\pi_{k:k}$ to a Gaussian. If the true posterior distribution is very far from Gaussian, e.g. multimodal, the approximation is evidently poor, which leads into unexpected behaviour of these algorithms. A more flexible approach than EKF, UKF, or CDF is to consider a sum of Gaussians approximation of the posterior distribution [Alspach and Sorenson 1972; Ito and Xiong 2000]. However, the algorithms propagating a sum of Gaussians tend to be computationally more complicated than the Gaussian approximations.

Chapter 4

Sequential Monte Carlo

Monte Carlo simulation methods have been used by statisticians for decades[1]. The increasing computing capabilities have enabled much more complicated simulation based approaches to be used in practice [Robert and Casella 1999]. The sequential Monte Carlo (SMC) methods, which are Monte Carlo approximations of the Bayes recursion, have gained much attention lately in many applications, including computer vision, target tracking, and many other fields [Doucet et al. 2001].

This chapter includes the basic concepts related to SMC. The Monte Carlo and importance sampling approaches are reviewed in Sections 4.1 and 4.2. The sequential importance sampling is introduced in Section 4.3, and the added resampling step in Section 4.4. Section 4.5 provides an alternative view of the Monte Carlo algorithms: as approximations of probability measures. Since the variety of different SMC methods is huge, this chapter can give only a coarse view. The collection edited by [Doucet et al. 2001] contains some theoretical discussion, practical applications, and a comprehensive list of references[2].

4.1 Monte Carlo

The Monte Carlo simulation method can be used to approximately compute expectation values for functions of random elements using independent samples of random elements.

Definition 4.1 *Let $h : X \to \mathbb{R}$ be a measurable function, and $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ be n independent and identically distributed random elements $\boldsymbol{x}^{(i)} : \Omega \to X$ having a distribution $P_{\boldsymbol{x}}$. The **Monte Carlo estimate** of the expected value of $h(\boldsymbol{x}^{(i)})$ is*

$$\boldsymbol{m}_n = \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}^{(i)}) \tag{4.1}$$

1. "The Monte Carlo method" was presented with this name by Metropolis and Ulam [1949], but the method dates back to as early as the 18th century [see, e.g. Liu 2003].
2. The Web page http://www-sigproc.eng.cam.ac.uk/smc/ contains also an extensive and up-to-date list of publications, demos, and an FAQ related to SMC.

Note that $\boldsymbol{m}_n : \Omega \to \mathbb{R}$ is a random variable. In a practical simulation, one often deals with one realisation of $\boldsymbol{m}_n$.

The Monte Carlo estimate converges almost surely, $\boldsymbol{m}_n \xrightarrow{\text{a.s.}} \mathsf{E}\left[h(\boldsymbol{x}^{(1)})\right]$, if the expectation is finite, due to Theorem 2.31 (Strong Law of Large Numbers). Of course, the theorem only suggests that $\boldsymbol{m}_n(\omega)$ converges for a.e. $\omega \in \Omega$, but does not give any bounds of variance for $\boldsymbol{m}_n$ with fixed n.

If $h(\boldsymbol{x}^{(1)})$ has a finite second moment, the speed of the convergence can be assessed, for denote the residual $\boldsymbol{e}_n = \boldsymbol{m}_n - \mathsf{E}\left[h(\boldsymbol{x}^{(1)})\right]$. Then, one obtains

$$\mathsf{V}\left[\boldsymbol{m}_n\right] = \mathsf{V}\left[\boldsymbol{e}_n\right] = \frac{1}{n^2}\mathsf{V}\left[\sum_{i=1}^{n} h(\boldsymbol{x}^{(i)})\right] = \frac{1}{n}\mathsf{V}\left[h(\boldsymbol{x}^{(1)})\right]$$

since $h(\boldsymbol{x}^{(i)})$ are independent. The Chebyshev's inequality (Theorem 2.14) can be used to give upper confidence bounds for absolute deviation of $\boldsymbol{m}_n$. In addition, if $h(\boldsymbol{x}^{(1)})$ has a finite second moment, the conditions of the Central Limit Theorem are satisfied, which states that

$$\sqrt{\frac{n}{\mathsf{V}\left[h(\boldsymbol{x}^{(1)})\right]}}\boldsymbol{e}_n \xrightarrow{d} N(\cdot; 0, 1)$$

Robert and Casella [1999] write that this information can be used to construct a convergence test, and give approximate confidence bounds.

Remark 4.2 Sometimes the term **Monte Carlo integration** is used in the literature. This term is due to the fact that taking expectation is, in general, integration. Thus, Monte Carlo simulation can be considered also as a numerical integration method, that applies to certain kinds of integrals, i.e. such integrals that may be considered expectations.

4.2 Importance Sampling

There are many situations in which drawing samples straight from the distribution $P_{\boldsymbol{x}}$ is either infeasible, or gives a high-variance estimator. In that case, one method that can be used is the importance sampling (IS) [Robert and Casella 1999]. In IS, there is another distribution $P_{\boldsymbol{z}}$, from which independent samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$ can be drawn. In the following, we refer $P_{\boldsymbol{x}}$ to as the **target distribution**, whereas $P_{\boldsymbol{z}}$ is referred to as the **importance distribution**.

Definition 4.3 *Let $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$, $\boldsymbol{z}^{(i)} : \Omega \to X$ be independent and identically distributed samples form the importance density $f_{\boldsymbol{z}}$. In addition, let the target density be $f_{\boldsymbol{x}}$, and let $h : X \to \mathbb{R}$ be a measurable function. The **importance sampling** estimate of $\mathsf{E}\left[h(\boldsymbol{x})\right]$ is*

$$\boldsymbol{m}_n = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{z}^{(i)})\frac{f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})}{f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})}, \quad \textit{where} \quad f^*_{\boldsymbol{z}}(z) = \begin{cases} f_{\boldsymbol{z}}(z), & f_{\boldsymbol{z}}(z) > 0 \\ 1, & f_{\boldsymbol{z}}(z) = 0 \end{cases} \tag{4.2}$$

The IS estimate is often written in the following form

$$\boldsymbol{m}_n = \sum_{i=1}^{n} \boldsymbol{w}^{(i)} h(\boldsymbol{z}^{(i)}) \quad \text{where} \quad \boldsymbol{w}^{(i)} = \frac{f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})}{n f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})},$$

where $\boldsymbol{w}^{(i)}$ are referred to as *importance weights*. When IS is applied, the collection of pairs $\{(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i)})\}_{i=1}^n$ is often referred to as the set of *weighted samples* or *particles*, especially if the function h is not fixed beforehand, or expectations of several functions are considered [Liu 2003].

Proposition 4.4 *Sufficient condition for a.s. convergence of the IS estimate,* $\boldsymbol{m}_n \xrightarrow{\text{a.s.}} \mathsf{E}\left[h(\boldsymbol{x})\right]$ *is that* $P_{\boldsymbol{x}} \ll P_{\boldsymbol{z}}$*, i.e.* $P_{\boldsymbol{x}}(A) = 0$ *whenever* $P_{\boldsymbol{z}}(A) = 0$*. Especially, the estimate converges a.s., if* $\{x : f_{\boldsymbol{x}}(x) > 0\} \subset \{x : f_{\boldsymbol{z}}(x) > 0\}$.[3]

Proof. Let $N = \{z : f_{\boldsymbol{z}}(z) = 0\}$. Then, since $f_{\boldsymbol{z}}$ must be measurable, N is measurable. The measure of N is zero, since

$$P_{\boldsymbol{z}}(N) = \int_N \mathrm{d}P_{\boldsymbol{z}} = \int_N f_{\boldsymbol{z}}(x)\mathrm{d}\mu(x) = 0$$

By assumption, then $P_{\boldsymbol{x}}(N) = 0$, and it follows that

$$\begin{aligned}\mathsf{E}\left[h(\boldsymbol{x})\right] &= \int_S h(x) f_{\boldsymbol{x}}(x)\mathrm{d}\mu(x) = \int_{S\setminus N} h(x) f_{\boldsymbol{x}}(x)\mathrm{d}\mu(x) \\ &= \int_{S\setminus N} h(x)\frac{f_{\boldsymbol{x}}(x)}{f^*_{\boldsymbol{z}}(x)} f_{\boldsymbol{z}}(x)\mathrm{d}\mu(x) = \mathsf{E}\left[h(\boldsymbol{z})\frac{f_{\boldsymbol{x}}(\boldsymbol{z})}{f^*_{\boldsymbol{z}}(\boldsymbol{z})}\right]\end{aligned}$$

The functions $f_{\boldsymbol{x}}$, $f_{\boldsymbol{z}}$, $f^*_{\boldsymbol{z}}$, and h are measurable, so their products and reciprocals are measurable as well. Now, it holds that

$$\boldsymbol{m}_n \xrightarrow{\text{a.s.}} \mathsf{E}\left[h(\boldsymbol{z})\frac{f_{\boldsymbol{x}}(\boldsymbol{z})}{f^*_{\boldsymbol{z}}(\boldsymbol{z})}\right] = \mathsf{E}\left[h(\boldsymbol{x})\right]$$

That is, if the expectations are finite, the IS estimate converges as a regular Monte Carlo estimate. Clearly, if $\{x : f_{\boldsymbol{x}}(x) > 0\} \subset \{x : f_{\boldsymbol{z}}(x) > 0\}$, then $P_{\boldsymbol{x}} \ll P_{\boldsymbol{z}}$. □

It is worth noticing, that if $P_{\boldsymbol{x}} \ll P_{\boldsymbol{z}}$, then there are densities $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{z}}$ with respect to $P_{\boldsymbol{z}}$, given as follows $f_{\boldsymbol{x}} = \mathrm{d}P_{\boldsymbol{x}}/\mathrm{d}P_{\boldsymbol{z}}$ and $f_{\boldsymbol{z}} = \mathrm{d}P_{\boldsymbol{z}}/\mathrm{d}P_{\boldsymbol{z}} \equiv 1$.

This sufficient condition of the convergence of the IS estimate is rather easy to fulfil. One should notice, that this does not mean it is irrelevant how

3. The book of Robert and Casella [1999] describes the condition $\mathrm{supp}(f_{\boldsymbol{z}}) \supset \mathrm{supp}(f_{\boldsymbol{x}})$ sufficient for convergence, but do no specify what they mean by support (or density). Common definition of support, the *closure* of the set $\{x : f(x) \neq 0\}$, (see, e.g. [Råde and Westergren 1998]), does not guarantee convergence. For example, suppose $f_{\boldsymbol{x}}(x) = 1$, whenever $x \in [0, 1]$, and zero otherwise, and $f_{\boldsymbol{z}}(x) = 1$ whenever $x \in [1, 2] \cup (\mathbb{Q} \cap [0, 1])$, and zero otherwise. Then, the both are density functions (with respect to the Lebesgue measure), specify probability measures, and $\mathrm{supp}(f_{\boldsymbol{z}}) = [0, 2] \supset [0, 1] = \mathrm{supp}(f_{\boldsymbol{x}})$, but $\boldsymbol{m}_n \overset{\text{a.s.}}{=} 0$.

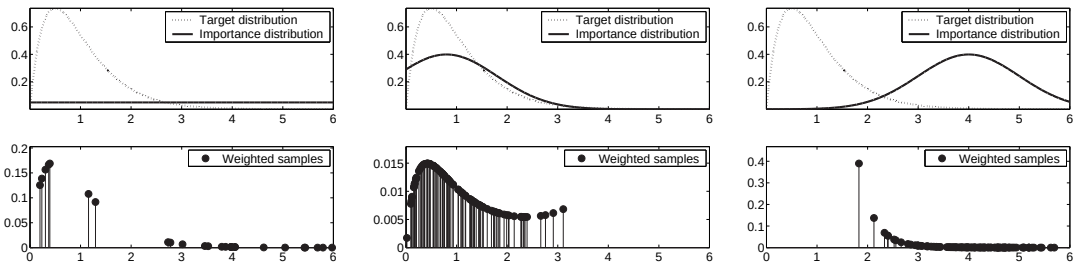


Figure 4.1: Examples of weight distributions in importance sampling with 200 samples and different choices of the importance distribution. The height of the stems corresponds to the weight w_i.

the importance distribution $f_{\boldsymbol{z}}$ is chosen. On the contrary, a proper selection of $f_{\boldsymbol{z}}$ plays a significant role in the practical usability of the IS estimate. Firstly, if $f_{\boldsymbol{z}}$ is selected so that the ratio $f_{\boldsymbol{x}}/f_{\boldsymbol{z}}$ is unbounded, the variance of the estimate may well be infinite. Secondly, even if the ratio is bounded, the selection of $f_{\boldsymbol{z}}$ may change the magnitude of the variance of the estimate. Next, an illustrative example of different choices of the importance distribution are given, which should appeal to reader's intuition.

Example 4.5 Figure 4.1 shows examples of the weighted sample sets, when the target distribution $f_{\boldsymbol{x}}$ is the Erlang distribution given by the density function

$$f_{\boldsymbol{x}}(x) = \frac{(\lambda x)^{n-1}}{(n-1)!}\lambda e^{-\lambda x}$$

with parameters $n = 2$ and $\lambda = 2$, which results in $\mathsf{E}[\boldsymbol{x}] = n/\lambda = 1$, $\mathsf{V}[\boldsymbol{x}] = n/\lambda^2 = 0.5$. The chosen importance distributions $f_{\boldsymbol{z}}$ are shown in Figures 4.1 (a)–(c). In figure (a), $f_{\boldsymbol{z}}$ is the uniform distribution in the interval $[0, 20]$. In figures (b) and (c), $f_{\boldsymbol{z}}$ is a Gaussian distribution with parameters $\mathsf{E}[\boldsymbol{z}] = 0.8$, and $\mathsf{V}[\boldsymbol{z}] = 1$; and $\mathsf{E}[\boldsymbol{z}] = 4$, and $\mathsf{V}[\boldsymbol{z}] = 1$, respectively.

Clearly, a large mismatch of the importance distribution $f_{\boldsymbol{z}}$ and the target distribution $f_{\boldsymbol{x}}$ in Figure 4.1 (c) results in a large variation in the importance weights. It is obvious, that e.g. the estimate of $\mathsf{E}[\boldsymbol{x}]$ would have a much higher variance in the case of Figure 4.1 (c) than in Figure 4.1 (b), where the importance distribution resembles more the target distribution. ◊

In some applications, a different importance sampling scheme is convenient to use. The alternative IS estimate can be given in the following form [Doucet et al. 2001; Geweke 1989; Liu 2003].

$$\boldsymbol{m}'_n = \sum_{i=1}^{n} \boldsymbol{w}^{(i)} h(\boldsymbol{z}^{(i)}), \quad \text{where} \quad \boldsymbol{w}^{(i)} = \frac{f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})/f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})}{\sum_{i=1}^{n} f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})/f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})} \tag{4.3}$$

where $\boldsymbol{w}_i$ are the *normalised* importance weights. The advantage of this method is that one needs to know the density functions $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{z}}$ only proportionally,

since the weights $\boldsymbol{w}^{(i)}$ are always normalised to sum to unity. The connection between the estimates in Equations (4.3) and (4.2) is

$$\boldsymbol{m}'_n = \frac{1}{\sum_{i=1}^{n} f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})/f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})} \sum_{i=1}^{n} h(\boldsymbol{z}^{(i)}) \frac{f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})}{f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})} = \frac{n}{\sum_{i=1}^{n} f_{\boldsymbol{x}}(\boldsymbol{z}^{(i)})/f^*_{\boldsymbol{z}}(\boldsymbol{z}^{(i)})} \boldsymbol{m}_n$$

The estimate given in Equation (4.3), is biased unlike the one given in Equation (4.2) [Liu 2003]. The estimate is, however, asymptotically unbiased [Doucet et al. 2001, p. 8].

Remark 4.6 In the scope of this thesis, IS is applied because samples cannot be drawn, or are hard to draw directly from the target distribution. In general, however, IS may also be applied to accelerate the convergence of a Monte Carlo estimate. For example, consider a situation in which the function $h(x)$ is very "spiky". That is, the function has value zero, or almost zero, with most probable values of x. Then, the importance distribution can be chosen so that less samples are drawn in the regions where $h(x) \approx 0$. In fact, [Robert and Casella 1999, p. 84] prove, that the importance distribution, that minimises the variance of the estimate, has the following density.

$$g^*(x) = \frac{|h(x)| f_{\boldsymbol{x}}(x)}{\int |h(x)| f_{\boldsymbol{x}}(x) \mathrm{d}\mu(x)}$$

This result is impractical, since the evaluation of the integral in the denominator is almost the expectation that was to be approximated with IS in the first place. However, if one uses the biased estimate given in Equation (4.3), the denominator need not be known.

4.3 Sequential Importance Sampling

Suppose that the model at hand is the state-space model introduced in Chapter 3. Since the Bayes recursion given in Section 3.2 is generally intractable to compute, one may consider applying Monte Carlo methods. The computation of the Bayes estimates involves conditional expectations. In general, one needs to obtain the following expectation for each $k \in \mathbb{N}$

$$\mathsf{E}\left[h(\boldsymbol{x}_{0:k}) \mid \boldsymbol{y}_{1:k} = y_{1:k}\right] = \int h(x_{0:k}) \mathrm{d}\pi_{0:k|k}(x_k) \tag{4.4}$$

for some measurable function $h : X^{k+1} \to \mathbb{R}$. Most often, the function of interest depends only on the latest state element, $\boldsymbol{x}_k$. Sequential importance sampling (SIS) can be formulated as a regular importance sampling estimation of Equation (4.4).

The importance distribution must admit certain properties, to allow the estimate to be computed sequentially. Suppose that there is another process

$(\boldsymbol{z}_k)_{k\in\mathbb{N}}$, for which the conditional probability $P(\boldsymbol{z}_{0:k} \mid \boldsymbol{y}_{1:k} = y_{1:k})$ has the density $q_{0:k}(z_{0:k} \mid y_{1:k})$ which admits the following factorisation [Doucet et al. 2000].

$$q_{0:k}(z_{0:k} \mid y_{1:k}) = q_0(z_0) \prod_{j=1}^{k} q_j(z_j \mid z_{0:j-1}, y_{1:j}) \tag{4.5}$$

where q_0 is a density, and for all $1 \leq j \leq k$ and all $z_{0:j-1}$ and $y_{1:j}$, the functions $q_j(\cdot \mid z_{0:j-1}, y_{1:j})$ are density functions. Furthermore, assume that the transition kernel of the Markov chain $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ has the density $f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}$, and the prior distribution $P_{\boldsymbol{x}_0}$ has the density $p_{\boldsymbol{x}_0}$. Then, the prior distribution of $\boldsymbol{x}_{0:k}$ has a density that can be represented in terms of the prior and the kernel densities as follows.

$$f_{\boldsymbol{x}_{0:k}}(x_{0:k}) = p_{\boldsymbol{x}_0}(x_0) \prod_{j=1}^{k} f_{\boldsymbol{x}_j|\boldsymbol{x}_{j-1}}(x_j \mid x_{j-1})$$

Suppose that $\boldsymbol{z}_0^{(1)}, \ldots, \boldsymbol{z}_0^{(n)}$ are IID samples drawn from the distribution with the density $q_{0|0}$. It is obvious, that the importance weights can be computed as follows.

$$\boldsymbol{w}_0^{(i)} \propto \frac{p_{\boldsymbol{x}_0}(\boldsymbol{z}_0^{(i)})}{q_0(\boldsymbol{z}_0^{(i)})}$$

Then, suppose that $\boldsymbol{z}_{0:k}^{(1)}, \ldots, \boldsymbol{z}_{0:k}^{(n)}$ are IID samples from the distribution with the density $q_{0:k|k}(z_{0:k} \mid y_{1:k})$. In this case, the importance weights can be computed as follows.

$$\begin{aligned}
\boldsymbol{w}_k^{(i)} &\propto \frac{p_{0:k|k}}{q_{0:k}} \\
&\propto \frac{p_0(\boldsymbol{z}_0^{(i)}) \prod_{j=1}^{k} g_j(y_j \mid \boldsymbol{z}_j^{(i)}) f_{\boldsymbol{z}_j|\boldsymbol{z}_{j-1}}(\boldsymbol{z}_j^{(i)} \mid \boldsymbol{z}_{j-1}^{(i)})}{q_0(\boldsymbol{z}_0^{(i)}) \prod_{j=1}^{k} q_j(\boldsymbol{z}_j^{(i)} \mid \boldsymbol{z}_{0:j-1}^{(i)}, y_{1:j})} \\
&\propto \frac{p_{0:k-1|k-1}}{q_{0:k-1|k-1}} \frac{g_k(y_k \mid \boldsymbol{z}_k^{(i)}) f_{\boldsymbol{z}_k|\boldsymbol{z}_{k-1}}(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{k-1}^{(i)})}{q_k(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{0:k-1}^{(i)}, y_{1:k})} \\
&\propto \boldsymbol{w}_{k-1}^{(i)} \frac{g_k(y_k \mid \boldsymbol{z}_k^{(i)}) f_{\boldsymbol{z}_k|\boldsymbol{z}_{k-1}}(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{k-1}^{(i)})}{q_k(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{0:k}^{(i)}, y_{1:k})}
\end{aligned} \tag{4.6}$$

where the proportionalities follow from the Bayes recursion, and the assumed factorisation of $q_{0:k}$ given in Equation (4.5). The derivation above shows that the importance weights can be computed recursively. Furthermore, it is easy to see, that the samples $\boldsymbol{z}_{0:k}^{(i)}$ can be obtained recursively as well, by sampling $\boldsymbol{z}_j^{(i)}$ from the distribution determined by the density $q_j(\cdot \mid \boldsymbol{z}_{0:j-1}^{(i)}, y_{1:j})$. The SIS algorithm is summarised in Algorithm 4.1.

Consider next the special case that the importance density is chosen to be the unconditional distribution, i.e., $q_0 = p_{\boldsymbol{x}_0}$ and $q_k = f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}$. Consequently,

$\boldsymbol{z}_0^{(i)} \sim q_0$

$\boldsymbol{w}_0^{(i)} \leftarrow \dfrac{p_{\boldsymbol{x}_0}(\boldsymbol{z}_0^{(i)})}{q_0(\boldsymbol{z}_0^{(i)})}$

for $k = 1, 2, \ldots$ **do**

$\boldsymbol{z}_k^{(i)} \sim q_k(\cdot \mid \boldsymbol{z}_{0:k-1}^{(i)}, y_{1:k})$

$\widehat{\boldsymbol{w}}_k^{(i)} \leftarrow \boldsymbol{w}_{k-1}^{(i)} \dfrac{g_k(y_k \mid \boldsymbol{z}_k^{(i)}) f_{\boldsymbol{z}_k|\boldsymbol{z}_{k-1}}(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{k-1}^{(i)})}{q_k(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{0:k}^{(i)}, y_{1:k})}$

$\boldsymbol{w}_k^{(i)} \leftarrow \dfrac{\widehat{\boldsymbol{w}}_k^{(i)}}{\sum_{i=1}^n \widehat{\boldsymbol{w}}_k^{(i)}}$

$\mathsf{E}\left[h(\boldsymbol{x}_{0:k}) \mid \boldsymbol{y}_{1:k} = y_{1:k}\right] \approx \sum_{i=1}^n \boldsymbol{w}_k^{(i)} h(\boldsymbol{z}_{0:k}^{(i)})$

end for

Algorithm 4.1: The SIS algorithm. The symbol "$\sim$" means that the random element is distributed according to the given density, and "$\leftarrow$" denotes substitution.

the samples $\boldsymbol{z}_{0:k}^{(i)}$ are distributed according to the unconditional distribution $P_{\boldsymbol{x}_{0:k}}$. Theoretically, this is a sensible decision, since the posterior distribution is absolutely continuous with respect to the unconditional distribution, $\pi_{0:k|k} \ll P_{\boldsymbol{x}_{0:k}}$. In this special case, the weight update formula given in Equation (4.6) reduces into the following.

$$\boldsymbol{w}_k^{(i)} \propto \boldsymbol{w}_{k-1}^{(i)} g_k(y_k \mid \boldsymbol{z}_k^{(i)})$$

The next example illustrates how SIS works in this situation.

Example 4.7 Suppose that the Markov chain is homogeneous, i.e. the transition kernel is independent time. Suppose that the prior and the transition kernel have the form.

$$P(\boldsymbol{x}_0 \in B) = N(B; 0, 0.1)$$
$$P_k(x; B) = N(B; x, 1)$$

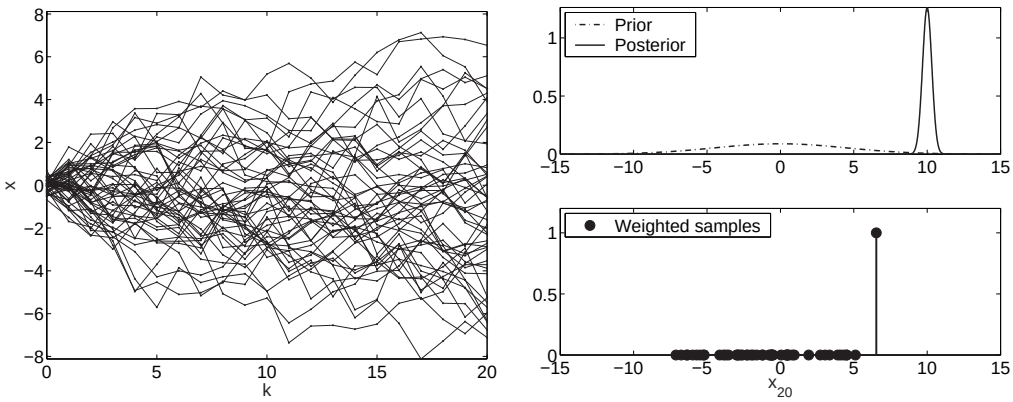
Figure 4.2 (a) shows 50 simulated IID sample paths of $\boldsymbol{x}_{0:20}$. The measurements $\boldsymbol{y}_k$ are assumed distributed as follows.

$$P(\boldsymbol{y}_k \in B \mid {}_k = x_k) = N(B; x, 0.11)$$

and the measurements $y_{0:20}$ can be given as $y_k = 10(k+1)/21 + 0.05$. It is easy to show[4], that the prior and the posterior distribution of $\boldsymbol{x}_{20}$ can be given as follows.

$$P(\boldsymbol{x}_{20} \in B) = N(B; 0, 20.1)$$
$$P(\boldsymbol{x}_{20} \in B \mid \boldsymbol{y}_{1:20} = y_{1:20}) = N(B; 10, 0.1)$$

4. In fact, this model is linear-Gaussian, so Equations (3.10)–(3.13) provide the result.



(a) (b)

Figure 4.2: (a) Sample paths of a homogeneous Markov chain. (b) The prior density and the posterior density of $\boldsymbol{x}_{20}$, and the weighted samples when prior distribution is used as importance distribution.

In this case, the prior and the posterior distribution have a rather large mismatch, with the property that the prior is much wider than the posterior. The prior and posterior distributions are depicted in Figure 4.2 (b). ◊

In practice, it often occurs that the posterior distribution $\pi_{k|k}$ is far from the prior distribution $P_{\boldsymbol{x}_k}$, as it is in Figure 4.2 (b). This being the case, most of the SIS samples $\boldsymbol{x}_k^{(i)}$ have very small weight values, which in turn means that the samples have very little effect on the estimate[5]

$$\mathsf{E}\left[h(\boldsymbol{x}_k) \mid \boldsymbol{y}_{1:k} = y_{1:k}\right] \approx \sum_{i=1}^{n} \boldsymbol{w}_k^{(i)} h(\boldsymbol{x}_k^{(i)})$$

Consequently, the variance of the estimate is high. This phenomenon, that only few samples have a nonnegligible weight, is known as *sample degeneracy* [Doucet et al. 2001]. It may occur, that an importance distribution better than the prior is hard to find. Fortunately, other methods exist for overcoming the problem of degeneracy. The next section describes resampling, which addresses the issue.

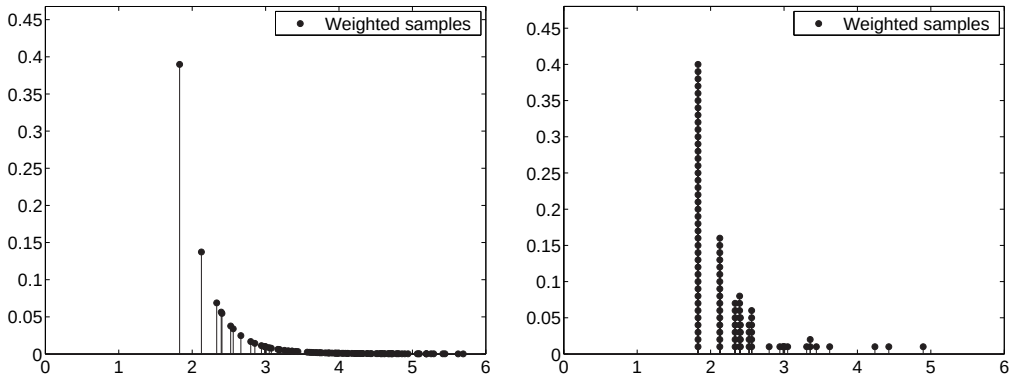
4.4 Sequential Importance Sampling with Resampling

Since the basic SIS framework does not work in practice due to sample degeneracy, one needs methods for overcoming the problem. One, rather straightforward method is resampling [Gordon et al. 1993]. Resampling can be considered intuitively as stochastic pruning, in which the least likely particles are discarded, while the most promising ones are duplicated. The resampling algorithm shown in Algorithm 4.2 was proposed by Gordon et al. [1993]. In brief, the algorithm

5. Provided that h is assumed regular enough, e.g., continuous and bounded.

$\boldsymbol{s}_0 \leftarrow 0$
$\boldsymbol{s}_j \leftarrow \sum_{i=1}^{j} \boldsymbol{w}^{(i)}$ for all $1 \leq j \leq n$
for $i = 1, \ldots, n$ **do**
  $\boldsymbol{u} \sim U(0, 1)$.
  Find such j, that $\boldsymbol{s}_{j-1} < u \leq \boldsymbol{s}_j$
  $\widehat{\boldsymbol{z}}^{(i)} \leftarrow \boldsymbol{z}^{(j)}$
  $\widehat{\boldsymbol{w}}^{(i)} \leftarrow 1/n$
end for

Algorithm 4.2: Resampling algorithm. The input to the algorithm is a set of weighted samples $(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i)})_{i=1}^{n}$, and the output is the set of resampled samples $(\widehat{\boldsymbol{z}}^{(i)}, \widehat{\boldsymbol{w}}^{(i)})_{i=1}^{n}$.



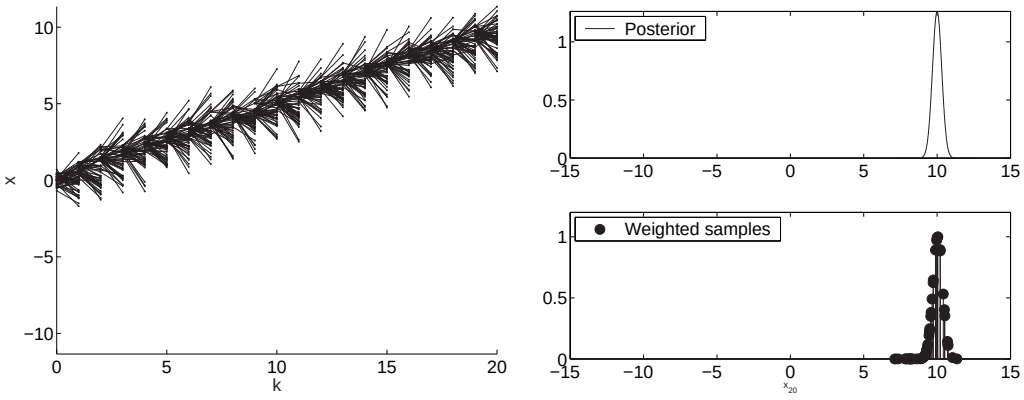
(a) (b)

Figure 4.3: (a) The set of weighted samples (b) The resampled samples.

can be described just to pick n samples from the discrete distribution determined by weighted samples $(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i)})_{i=1}^{n}$. The core idea of resampling can be best introduced in terms of an example.

Example 4.8 Consider the set of weighted samples shown in Figure 4.3 (a) (which is reproduction of Figure 4.1 (c)). Resampling, Algorithm 4.2, is applied to the set of weighted samples, and the resulting samples are shown in Figure 4.3 (b). The duplicated particles are shown as stacked for visualisation. The resampling can be considered, in fact, some kind of stochastic discretisation of the set of weighted samples. ◊

Of course, the duplications in the resampling algorithm are random, since the n resampled samples are drawn at random. It is rather intuitive, that the expected number of the resampled duplicates of a sample $\boldsymbol{z}^{(i)}$ is exactly proportional to the weight of the sample $\boldsymbol{w}^{(i)}$. Furthermore, the expectation of the IS approximation induced by a set of weighted samples is the same as the IS approximation



(a) (b)

Figure 4.4: (a) Sample paths of the bootstrap filter (b) The posterior density and the weighted sample.

induced by the set of resampled samples. That is,

$$\mathsf{E}\left[\boldsymbol{m}_n\right] = \mathsf{E}\left[\widehat{\boldsymbol{m}}_n\right] \quad \text{where} \quad \boldsymbol{m}_n = \sum_{i=1}^{n} \boldsymbol{w}^{(i)} h(\boldsymbol{z}^{(i)}) \;\text{ and }\; \widehat{\boldsymbol{m}}_n = \sum_{i=1}^{n} \widehat{\boldsymbol{w}}^{(i)} h(\widehat{\boldsymbol{z}}^{(i)})$$

However, resampling adds extra variance to the estimator [Carpenter et al. 1999]. Since resampling discards some samples and duplicates others, there is a loss of diversity of particles. This phenomenon is referred to as *sample impoverishment*.

Since the variance is increased, it does not make sense to perform resampling in the regular IS. For SIS, the resampling step provides a "trick" to reduce the sample degeneracy. Such a SIS algorithm that includes resampling is referred here to as a SISR algorithm[6]. The SISR algorithm proposed by Gordon et al. [1993], which is referred to as the *bootstrap filter*, assumes that the importance distribution is the prior density, and resampling is performed at each step.

Example 4.9 Consider the problem given in Example 4.7. The bootstrap filter was applied to the same problem, and Figure 4.4 (a) shows the sample paths of the bootstrap filter. The sample paths are "pruned" in the resampling step. Figure 4.4 (b) shows the weighted samples corresponding to the posterior $\pi_{20|20}$. Comparing Figures 4.2 and 4.4 shows how the bootstrap filter concentrates the samples sequentially, whereas in the case of a plain SIS, the samples are scattered uncontrolled to much wider area. ◊

The bootstrap filter is only one specific SISR algorithm. There are many other SISR algorithms, with different choices of the importance distribution and the resampling algorithm. The resampling algorithm given in Algorithm 4.2 is perhaps conceptually the simplest, but not the most efficient. There are algorithms that perform resampling in $O(n)$ time, and admit smaller variance [Carpenter

6. Sequential Importance Sampling with Resampling.

$\boldsymbol{z}_0^{(i)} \sim p_{\boldsymbol{x}_0}$
$\boldsymbol{w}_0^{(i)} \leftarrow 1/n$
for $k = 1, 2, \ldots$ **do**
  $\boldsymbol{z}_k^{(i)} \sim f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(\cdot \mid \boldsymbol{z}_{k-1}^{(i)})$
  $\widehat{\boldsymbol{w}}_k^{(i)} \leftarrow g_k(y_k \mid \boldsymbol{z}_k^{(i)})$
  $\boldsymbol{w}_k^{(i)} \leftarrow \dfrac{\widehat{\boldsymbol{w}}_k^{(i)}}{\sum_{i=1}^n \widehat{\boldsymbol{w}}_k^{(i)}}$
  $\mathsf{E}\left[h(\boldsymbol{x}_{0:k}) \mid \boldsymbol{y}_{1:k} = y_{1:k}\right] \approx \sum_{i=1}^{n} \boldsymbol{w}_k^{(i)} h(\boldsymbol{z}_{0:k}^{(i)})$
  if $n_{\text{eff}}(\{\boldsymbol{w}_k^{(i)}\}_{i=1}^n\}) < n_{\text{th}}$ **then**
    $(\boldsymbol{z}_k^{(i)}, \boldsymbol{w}_k^{(i)})_{i=1}^n \leftarrow \text{resample}\left[(\boldsymbol{z}_k^{(i)}, \boldsymbol{w}_k^{(i)})_{i=1}^n\right]$
  end if
end for

Algorithm 4.3: SISR algorithm with prior importance distribution and adaptive resampling. The "resample" function refers to Algorithm 4.2.

et al. 1999; Kitagawa 1996]. There are also other methods proposed for enhancing the resampling [see, e.g, Doucet et al. 2001, pp. 230–234]. Furthermore, Doucet [1998] suggests that resampling should not be performed in each iteration, but adaptively. This method, *adaptive resampling*, is based on the estimated *effective sample size*,

$$n_{\text{eff}}(\{\boldsymbol{w}_k^{(i)}\}_{i=1}^n) = \frac{1}{\sum_{i=1}^n (\boldsymbol{w}_k^{(i)})^2} \tag{4.7}$$

Note that $0 < n_{\text{eff}} \leq n$. The resampling procedure is carried out only if n_{eff} drops below a predefined threshold $0 < n_{\text{th}} \leq n$. Otherwise, $\boldsymbol{w}_k$ and $\boldsymbol{x}_k$ remain intact. The SISR algorithm with prior importance distribution and adaptive resampling is given in Algorithm 4.3.

Remark 4.10 Russell and Norvig [2003] outline the history of Sequential Monte Carlo, describing that resampling dates back to 1975. The resampling approach has been reinvented and renamed again since, with such names as sequential importance-sampling resampling, particle filtering [Gordon et al. 1993], survival of the fittest [Kanazawa et al. 1995], and Conditional Density Propagation (Condensation) algorithm [Blake and Isard 1997].

4.5 SMC as Measure Approximation

All the Monte Carlo methods considered in this thesis can be thought as approximations of a probability measure. This approximation can be written as a finite

sum of weighted delta measures, as follows

$$\boldsymbol{P}_n(A) = \sum_{i=1}^{n} \boldsymbol{w}^{(i)} \delta_{\boldsymbol{z}^{(i)}}(A)$$

where $(\boldsymbol{w}^{(i)}, \boldsymbol{z}^{(i)})_{i=1}^n$ constitute the set of weighted samples. The expectation integral can be considered as integration over this (random) measure.

$$\mathsf{E}\left[h(\boldsymbol{x})\right] = \int h \mathrm{d}P_{\boldsymbol{x}} \approx \int h \mathrm{d}\boldsymbol{P}_n = \sum_{i=1}^{n} \boldsymbol{w}_i h(\boldsymbol{x}_i)$$

One may notice, that the measure $\boldsymbol{P}_n(A)$ is not necessarily a *probability* measure. For example, if $\boldsymbol{P}_n(A)$ corresponds to the IS method given in Definition 4.3, the weights $\boldsymbol{w}_i$ may not sum to unity. However, $\boldsymbol{P}_n(A)$ is a measure, so integration is meaningful, and results in the IS estimate.

If the random measures $\boldsymbol{P}_n$ converge in distribution to $P_{\boldsymbol{x}}$, then by definition, the following holds for any bounded and continuous function h

$$\lim_{n\to\infty} \mathsf{E}\left[h(\boldsymbol{z}_n)\right] = \mathsf{E}\left[h(\boldsymbol{x})\right]$$

where the random elements $\boldsymbol{z}_n$ are distributed according to $\boldsymbol{P}_n$. Equivalently, this can be expressed as follows.

$$\lim_{n\to\infty} \int h(x) \mathrm{d}\boldsymbol{P}_n(x) = \int h(x) \mathrm{d}P_{\boldsymbol{x}}(x)$$

Since almost sure convergence implies convergence in distribution (Theorem 2.29), the random measure corresponding to the MC estimate given in Definition 4.1 converges in distribution to the true distribution. Similarly, if $P_{\boldsymbol{x}} \ll P_{\boldsymbol{y}}$, random measure corresponding to the IS estimate given in Definition 4.3 converges in distribution. The SIS estimate converges as well, if the posterior distribution is absolutely continuous with respect to the importance distribution.

In light of the random measure perspective to the Monte Carlo approach, the SISR algorithm can be considered as a sequential approximation of the posterior measure $\pi_{0:k|k}$. Let us go through an informal construction of the measure approximation update. For detailed description, see e.g. [Crisan and Doucet 2002]. Suppose there is a set of weighted samples $(\boldsymbol{z}_{0:k-1}^{(i)}, \boldsymbol{w}_{k-1}^{(i)})_{i=1}^n$ from the distribution $\pi_{0:k-1|k-1}$. That is, the random measure $\boldsymbol{\pi}_{k-1|k-1}$ approximates the true posterior,

$$\boldsymbol{\pi}_{0:k-1|k-1} = \sum_{i=1}^{n} \boldsymbol{w}_{k-1}^{(i)} \delta_{\boldsymbol{z}_{0:k-1}^{(i)}} \approx \pi_{0:k-1|k-1}$$

One can obtain a set of weighted samples from $(\boldsymbol{z}_{0:k}^{(i)}, \widetilde{\boldsymbol{w}}_k^{(i)})$ by sampling each $\boldsymbol{z}_k^{(i)}$ from $q_k(\cdot \mid \boldsymbol{z}_{0:k-1}, y_{1:k})$, and computing the weights as follows.

$$\widetilde{\boldsymbol{w}}_k^{(i)} \propto \frac{f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{k-1}^{(i)})}{q_k(\boldsymbol{z}_k^{(i)} \mid \boldsymbol{z}_{0:k-1}, y_{1:k})}$$

Now, the predictive density is approximated as follows.

$$\boldsymbol{\pi}_{0:k|k-1} = \sum_{i=1}^{n} \widetilde{\boldsymbol{w}}_k^{(i)} \delta_{\boldsymbol{z}_{0:k}^{(i)}} \approx \pi_{0:k|k-1}$$

The approximation of the posterior distribution can then be obtained from the Bayes recursion, Equation (3.4).

$$\boldsymbol{\pi}_{0:k|k}(B) \overset{\text{a.s.}}{=} \frac{\int_B g_k(y_k \mid x_k) \mathrm{d}\boldsymbol{\pi}_{0:k|k-1}(x_k)}{\int g_k(y_k \mid x_k) \mathrm{d}\boldsymbol{\pi}_{0:k|k-1}(x_k)} = \frac{\sum_{i=1}^{n} g_k(y_k \mid \boldsymbol{z}_k^{(i)}) \widetilde{\boldsymbol{w}}_k^{(i)} \delta_{\boldsymbol{z}_{0:k}^{(i)}}(B)}{\sum_{i=1}^{n} g_k(y_k \mid \boldsymbol{z}_k^{(i)}) \widetilde{\boldsymbol{w}}_k^{(i)}}$$

That is, if one obtains the weights $\boldsymbol{w}_k^{(i)}$ according to Equation (4.6), the weighted samples $(\boldsymbol{z}_{0:k}^{(i)}, \boldsymbol{w}_k^{(i)})_{i=1}^n$ are from the posterior distribution, and they constitute an approximation of the posterior distribution,

$$\boldsymbol{\pi}_{0:k|k} = \sum_{i=1}^{n} \boldsymbol{w}_k^{(i)} \delta_{\boldsymbol{z}_{0:k}^{(i)}} \approx \pi_{0:k|k}$$

It is not obvious, that $\boldsymbol{\pi}_{0:k|k}$ will converge in distribution to $\pi_{0:k|k}$, since the random samples $\boldsymbol{z}_{0:k}^{(i)}$ are not independent, due to resampling. The reader is advised to take a look on the recent survey paper of the convergence results obtained for SISR [Crisan and Doucet 2002].

Chapter 5

Random Sets

This chapter presents the concept of random sets which is one of the main topics in this thesis. The chapter starts with Section 5.1 outlining the concept of random set in a finite, discrete universe. The discrete universe case contains many simplifications compared with the general case, providing intuitive view on the subject. Section 5.2 covers briefly the general concept of random closed sets. After that, the concept of random finite sets is presented in Section 5.3. Random finite sets are the main topic of this chapter, since they are applied to target tracking in Chapter 7.

5.1 Random Sets of Finite Universe

To get a feel of what a "random set" means, we examine first the case of random sets of a finite universe. This introductory exploration of the finite universe case serves us also as a very brief introduction to the theory of evidence (also known as Dempster-Shafer reasoning) from the perspective of random sets.

5.1.1 Probability Measure and Belief Measure

Let U be a set with n elements. Enumerate the elements as $U = \{u_1, u_2, \ldots, u_n\}$. One can define a random set $\boldsymbol{X} : \Omega \to \mathcal{U}$, where $\mathcal{U} = \mathcal{P}(U)$, the collection of all subsets of U. A measure can be defined by assigning probabilities $P(\boldsymbol{X} = A) = P_{\boldsymbol{X}}(\{A\})$ directly to each $A \in \mathcal{U}$, since $\mathcal{U}$ is finite[1]. This probability assignment is denoted in this section as $m_{\boldsymbol{X}}(A) \triangleq P_{\boldsymbol{X}}(\{A\})$.

Sometimes, it is more convenient to determine a distribution using the belief function

$$\beta_{\boldsymbol{X}}(A) \triangleq P(\boldsymbol{X} \subset A) = \sum_{B \subset A} m_{\boldsymbol{X}}(B) \tag{5.1}$$

The belief function can be considered to be the counterpart of the cumulative distribution function $P(\boldsymbol{x} \leq x)$ of a random variable $\boldsymbol{x}$ in $\mathbb{R}$. Notice, however, that the collection of subsets $\mathcal{U}$ is only partially ordered with the inclusion relation "$\subset$".

1. The set $\mathcal{P}(U)$ contains 2^n elements.

The belief function $\beta_{\boldsymbol{X}}$ is dual to the probability assignment $m_{\boldsymbol{X}}$. One can obtain the probabilities from the belief function using the so called Möbius inversion formula [Goodman et al. 1997]

$$m_{\boldsymbol{X}}(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} \beta_{\boldsymbol{X}}(B) \tag{5.2}$$

Next, a simple example of the definition of a finite random set is given.

Example 5.1 Take a two-element set $U = \{a, b\}$. Then, the collection of the subsets of U is $\mathcal{U} = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$. Define a probability for each event in $\mathcal{U}$ so that

$$m_{\boldsymbol{X}}(\emptyset) = 0.1,\ m_{\boldsymbol{X}}(\{a\}) = 0.4,\ m_{\boldsymbol{X}}(\{b\}) = 0.3,\ m_{\boldsymbol{X}}(U) = 0.2$$

From this probability assignment, one can compute the values of the belief function

$$\beta_{\boldsymbol{X}}(\emptyset) = 0.1,\ \beta_{\boldsymbol{X}}(\{a\}) = 0.5,\ \beta_{\boldsymbol{X}}(\{b\}) = 0.4,\ \beta_{\boldsymbol{X}}(U) = 1.$$

It is easy to verify that the Möbius inversion formula in Equation (5.2) works. ◊

5.1.2 Connection with Dempster-Shafer Reasoning

The Dempster-Shafer (DS) reasoning, or the theory of evidence, deals with imprecise evidence [Shafer 1976]. The theory can be considered to be a special case of random sets in a finite universe, such that the probability of the empty set is assumed zero. In the following, the terminology of DS reasoning is explained from the perspective of random sets. Note, that this section is not intended to provide introduction to the DS philosophy, but only lists some of the terms in the DS reasoning as probability assignments of random sets.

The belief of a set $V \subset U$ is the total belief that is consistent with V, i.e. the probability that $\boldsymbol{X}$ is contained in V,

$$\mathrm{Bel}_{\boldsymbol{X}}(V) = \sum_{W \subset V} m_{\boldsymbol{X}}(W) = \beta_{\boldsymbol{X}}(V)$$

The plausibility is the total weight of all events that do not contradict V, i.e. the probability that $\boldsymbol{X}$ intersects V.

$$\mathrm{Pl}_{\boldsymbol{X}}(V) = \sum_{W \cap V \neq \emptyset} m_{\boldsymbol{X}}(W) = \sum_{W \not\subseteq \complement V} m_{\boldsymbol{X}}(W) = 1 - \beta_{\boldsymbol{X}}(\complement V)$$

It is easy to see that $\mathrm{Bel}_m(V) \leq \mathrm{Pl}_m(V)$ for all $V \subset U$. The conflict between two random sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ with distributions $m_{\boldsymbol{X}}$ and $m_{\boldsymbol{Y}}$ is defined as follows.

$$K(m_{\boldsymbol{X}}, m_{\boldsymbol{Y}}) = \sum_{W \cap V = \emptyset} m_{\boldsymbol{X}}(V) m_{\boldsymbol{Y}}(W) = 1 - \sum_{W \cap V \neq \emptyset} m_{\boldsymbol{X}}(V) m_{\boldsymbol{Y}}(W)$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed independent, the conflict is the probability that $\boldsymbol{X}$ and $\boldsymbol{Y}$ do not intersect, $K(m_{\boldsymbol{X}}, m_{\boldsymbol{Y}}) = P(\boldsymbol{X} \cap \boldsymbol{Y} = \emptyset)$.

The orthogonal sum of two distributions $m_{\boldsymbol{X}}$ and $m_{\boldsymbol{Y}}$, also known as the Dempster's rule of combination is [Shafer 1976].

$$m_{\boldsymbol{X}} \oplus m_{\boldsymbol{Y}}(V) = \frac{1}{1 - K(m_{\boldsymbol{X}}, m_{\boldsymbol{Y}})} \sum_{W \cap T = V} m_{\boldsymbol{X}}(W) m_{\boldsymbol{Y}}(T)$$

Again, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed independent, one can write

$$\begin{aligned} m_{\boldsymbol{X}} \oplus m_{\boldsymbol{Y}}(V) &= \frac{\sum_{W \cap T = V} m_{\boldsymbol{X}}(W) m_{\boldsymbol{Y}}(T)}{\sum_{W \cap T \neq \emptyset} m_{\boldsymbol{X}}(W) m_{\boldsymbol{Y}}(T)} = \frac{P(\boldsymbol{X} \cap \boldsymbol{Y} = V)}{P(\boldsymbol{X} \cap \boldsymbol{Y} \neq \emptyset)} \\ &= P(\boldsymbol{X} \cap \boldsymbol{Y} = V \mid \boldsymbol{X} \cap \boldsymbol{Y} \neq \emptyset) \end{aligned}$$

So, the Dempster's rule of combination can be regarded as computation of the probability distribution for the random set $\boldsymbol{Z} = \boldsymbol{X} \cap \boldsymbol{Y}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random sets, given the condition that $\boldsymbol{Z}$ is nonempty.

5.2 Random Closed Sets

The next topic that is covered concerns the random closed sets. The presentation in this section is quite general. The space Θ of the random sets is not fixed, but only assumed to be a locally compact, Hausdorff and separable (LCHS) (see Appendix A for description of these terms). It is worth mentioning that the Euclidean space $\mathbb{R}^d$ is such a space, since replacing the abstract space Θ with $\mathbb{R}^d$ (or even $\mathbb{R}$) in mind, one can get a better idea of the nature of the presented concepts.

The collections of open, closed and compact sets of Θ are denoted by $\mathcal{G}$, $\mathcal{F}$ and $\mathcal{K}$, respectively. For any $A \subset \Theta$, we define the collections of closed sets "missing" and "hitting" A

$$\mathcal{F}^A = \{F \in \mathcal{F} : F \cap A = \emptyset\} \qquad \mathcal{F}_A = \{F \in \mathcal{F} : F \cap A \neq \emptyset\}$$

respectively. Figure 5.1 depicts an example of these collections of sets. Using the "hit" and "miss" collections, a topology on $\mathcal{F}$ can be defined. The topology is generated by the subbase of the collections $\mathcal{F}^K$ and $\mathcal{F}_G$ where $K \in \mathcal{K}$ and $G \in \mathcal{G}$. The topology has the base [Matheron 1975]

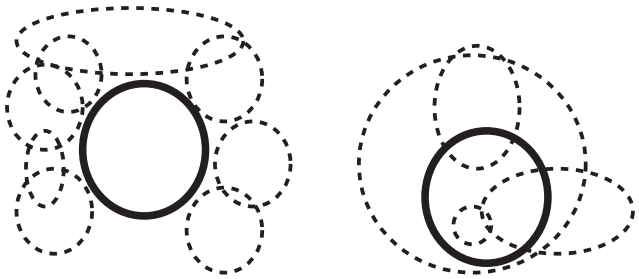
$$\mathcal{D}_{H/M} = \left\{\mathcal{F}^K \cap \mathcal{F}_{G_1} \cap \mathcal{F}_{G_2} \cap \cdots \cap \mathcal{F}_{G_n} : n \geq 0, K \in \mathcal{K}, G_k \in \mathcal{G}\right\} \tag{5.3}$$

This topology is called the *hit-or-miss topology*. The obtained topological space $(\mathcal{F}, \mathcal{D}_{H/M})$ is LCHS [Matheron 1975].

Let us denote the Borel sets of the topological space by $\mathcal{B}(\mathcal{F})$. Then, we may define a random closed set by a map $\boldsymbol{X} : \Omega \to \mathcal{F}$ which is measurable (with respect to $\mathcal{B}(\mathcal{F})$). The probability of an event $\mathcal{A} \in \mathcal{B}(\mathcal{F})$ can be obtained as usual

$$P_{\boldsymbol{X}}(\mathcal{A}) = P(\overleftarrow{\boldsymbol{X}}(\mathcal{A})) \tag{5.4}$$

Since the probability measure $P_{\boldsymbol{X}}$ is impractical to determine straightforwardly, Matheron [1975] suggests that the measure is defined indirectly, using a set function $T : \mathcal{K} \to [0, 1]$ having the following properties.



(a) (b)

Figure 5.1: Some example sets in collections (a) $\mathcal{F}^A$ and (b) $\mathcal{F}_A$. The set A is the one with solid border, and the sets in the respective collections are those with dotted border.

1. $T(\emptyset) = 0$.
2. If $\{K_i\}$ is a decreasing sequence[2] of compact sets and $\cap_{i\to\infty} K_i = K$, then $\lim_{i\to\infty} T(K_i) = T(K)$.
3. For any $n \geq 1$ and compact sets $K, K_1, \ldots, K_n$, the functions S_n determined by the recursion

$$\begin{aligned} S_1(K; K_1) &= T(K \cup K_1) - T(K) \\ S_n(K; K_1, \ldots, K_n) &= S_{n-1}(K; K_1, \ldots, K_{n-1}) \\ &\quad - S_{n-1}(K \cup K_n; K_1, \ldots, K_{n-1}) \end{aligned}$$

have a nonnegative value.

Theorem 5.2 (Choquet) *If the conditions 1–3 are satisfied, the function T determines a* unique *probability measure on $\mathcal{B}(\mathcal{F})$ satisfying the following condition.*

$$P_{\boldsymbol{X}}(\mathcal{F}_K) = T(K), \qquad \forall K \in \mathcal{K} \tag{5.5}$$

The function T is called the Choquet capacity functional [cf. Matheron 1975]. The reader may notice an analogy between the Choquet capacity functional and the belief measure given in Equation (5.1), since $P_{\boldsymbol{X}}(\mathcal{F}_K) = 1 - P_{\boldsymbol{X}}(\mathcal{F}^K) = 1 - P(\boldsymbol{X} \subset \complement K)$.

In the following, a simple example of a Choquet capacity functional is given. The example is a special case of the example in [Goodman et al. 1997, p. 95].

Example 5.3 Let $\Theta = \mathbb{R}$, and $f : \mathbb{R} \to [0, 1]$ be an upper-semicontinuous function[3]. Then, the function $T : \mathcal{K} \to [0, 1]$ defined by

$$T(K) = \sup_{x \in K} f(x)$$

2. The sequence of sets is decreasing, if $K_i \supset K_j$ for all $i < j$.
3. The function f is upper-semicontinuous, if for all $x \in \mathbb{R}$ and $\epsilon > 0$ there exists such $r > 0$ that $f(t) < f(x) + \epsilon$ whenever $\|t - x\| < r$.

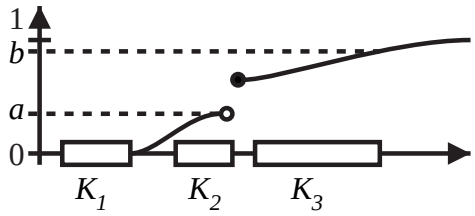


Figure 5.2: An example of a capacity functional defined using an upper-semicontinuous function $f : \mathbb{R} \to [0, 1]$. In this example, $T(K_1) = 0$, $T(K_2) = a$, and $T(K_3) = b$.

is a capacity functional. Figure 5.2 exemplifies this definition. ◊

5.3 Random Finite Sets

The presentation of random closed sets in Section 5.2 was kept general, without fixing the space Θ. The space was only required to be LCHS. This section covers a more specific class of random closed sets, the random finite sets (RFS). As the name suggests, we allow the sets to contain only a *finite* number of elements. The general definition of a RFS can be given as follows.

Definition 5.4 *The random element $\boldsymbol{X} : \Omega \to \mathcal{F}$ is a **random finite set** (RFS), if for all $\omega \in \Omega$, the set is finite, i.e. $|\boldsymbol{X}(\omega)| < \infty$. The random finite set $\boldsymbol{X}$ is **bounded**[4], if there is an integer M, so that $|\boldsymbol{X}(\omega)| \leq M$, for all $\omega \in \Omega$.*

In the following, it is assumed that the space Θ is the hybrid space, introduced in Section 5.3.1. The rest of this section is divided into seven subsections whose content is briefly following. Section 5.3.2 outlines the connection between the k-fold product spaces and the finite sets with exactly k elements. This connection is used to derive measures for k-element finite sets. Section 5.3.3 continues by defining a measure to the finite set space, which is base of the set integral. Section 5.3.4 introduces the belief measure of a RFS. The concept of the set derivative is introduced in Section 5.3.5, and some general properties of the set integral and the set derivative are given in Section 5.3.6. The last two sections are dedicated for analysis of a RFS distribution. Section 5.3.7 discusses of the probability hypothesis density, which can be considered as a "projection" of the random set distribution into a density in the hybrid space. Section 5.3.8 outlines two estimators, that can be used to extract an estimate from a random set distribution.

5.3.1 The Hybrid Space

In this thesis, the space Θ of finite random sets is assumed to be a so called *hybrid space* [Goodman et al. 1997, p. 135]. The hybrid space $\mathbb{S} = \mathbb{R}^d \times U$ is the

4. The definition of RFS given in [Goodman et al. 1997, p. 152] restricts RFS to be bounded.

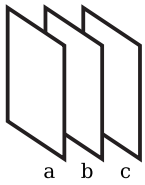


Figure 5.3: Example of a hybrid space $\mathbb{R}^2 \times \{a, b, c\}$.

Cartesian product of a d-dimensional Euclidean space $\mathbb{R}^d$, and a finite discrete space U. An element in this space $(\underline{v}, u) = s \in \mathbb{S}$ consists of the Euclidean part (vector) $\underline{v} \in \mathbb{R}^d$, and the discrete part $u \in U$. Figure 5.3 illustrates an example of a hybrid space.

The hybrid space is endowed with the product topology of the Euclidean topology in $\mathbb{R}^d$, and the discrete topology in U. This means that open sets in $\mathbb{S}$ are the sets of form $W \subset \mathbb{S}$, where $W_u = \{\underline{v} : (\underline{v}, u) \in W\} \subset \mathbb{R}^d$ is open for all "slices" $u \in U$. The closed and compact sets are determined similarly.[5] The space $\mathbb{S}$ is endowed with the product measure of the Lebesgue measure λ in $\mathbb{R}^d$ and the counting measure c in U.

Definition 5.5 *The product measure $\overline{\lambda} = \lambda \times c$ on space $\mathbb{S} = \mathbb{R}^d \times U$ is referred to as the* ***hybrid Lebesgue measure****.*

The measurable sets in $\mathbb{S}$ are all such $W \subset \mathbb{S}$, for which each "slice" W_u is Lebesgue-measurable [Goodman et al. 1997, p. 136].

It is quite obvious, that the space $\mathbb{S}$ endowed with the measure $\overline{\lambda}$ is σ-finite, since we may set $E_i = (B_{\underline{0}}(i), U)$, for which $\overline{\lambda}(E_i) = |U|\lambda(B_{\underline{0}}(i)) < \infty$ for all $i \in \mathbb{N}$, and clearly $\mathbb{S} = \cup_i E_i$. Therefore, all subsets of $\mathbb{S}$ are σ-finite. According to Fubini's theorem, for all measurable sets $W \subset \mathbb{S}$, it then holds that

$$\overline{\lambda}(W) = \int_U \lambda(W_u) \mathrm{d}c(u) = \sum_{u \in U} \lambda(W_u) \tag{5.6}$$

and for each integrable $f : \mathbb{S} \to \overline{\mathbb{R}}$ it holds that

$$\int f \mathrm{d}\overline{\lambda} = \int_U \left[\int_{\mathbb{R}^d} f(\underline{v}, u) \mathrm{d}\lambda(\underline{v}) \right] \mathrm{d}c(u) = \sum_{u \in U} \int_{\mathbb{R}^d} f(\underline{v}, u) \mathrm{d}\lambda(\underline{v}) \tag{5.7}$$

The hybrid space $\mathbb{S}$ is endowed with the following metric.

$$d(s, s') = \|\underline{v} - \underline{v}'\| + \delta_u(u'), \qquad \text{where} \quad \delta_u(u') = \begin{cases} 1, & u = u' \\ 0, & u \neq u' \end{cases} \tag{5.8}$$

where $s = (\underline{v}, u)$ and $s' = (\underline{v}', u')$, and $\| \cdot \|$ denotes the Euclidean norm.

5. In the discrete topology, all sets are open, thus also closed. In a finite space, all sets are compact.

Theorem 5.6 *The space $\mathbb{S}$ is locally compact, Hausdorff and separable (LCHS).*

Proof. Suppose $(u, v) = s \in \mathbb{S}$. One can pick open set $B_s(1)$, for which $\overline{B}_s(1)$ is closed and bounded in $\mathbb{R}^d$, hence compact in $\mathbb{S}$. Thus, $\mathbb{S}$ is locally compact. It is obvious, that $\mathbb{S}$ is Hausdorff, for suppose $(u', v') = s' \in \mathbb{S}$, so that $s \neq s'$. Then we can choose non-intersecting open neighbourhoods $B_s(\epsilon)$ and $B_{s'}(\epsilon)$, by setting $\epsilon = d(s, s')/2$. The space $\mathbb{R}^d$ is known to be separable, with dense set $\mathbb{Q}^d$ (the rational vectors). The space $\mathbb{S}$ is then obviously separable, since each slice $\mathbb{S}_u = \mathbb{R}^d$ is separable, and there is a finite number of slices. □

5.3.2 Product Spaces and Finite Sets

Throughout this section, the notations like $\mathcal{F}(k)$ and $\mathcal{F}(\leq k)$ are used, corresponding to the closed sets with exactly k elements and the closed sets with no more than k elements, respectively. Since the goal is to construct measures on $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$, it is important to figure out that such collections of sets are Borel in the hit-or-miss topology [Goodman et al. 1997, p. 132].

Proposition 5.7 *The collections $\mathcal{F}(\geq k)$, $\mathcal{F}(\leq k)$ and $\mathcal{F}(k)$ are open, closed, and Borel, respectively, in $\mathcal{F}$.*

Proof. Define the collection

$$\mathcal{A}(k) = \bigcup \left\{ \bigcap_{i=1}^{k} \mathcal{F}_{G_i} : G_i \in \mathcal{G} \text{ and } G_i \cap G_j = \emptyset \text{ for all } i \neq j \right\}$$

The set $\mathcal{A}(k)$ is open in $\mathcal{F}$, since it is union of open sets in $\mathcal{F}$. Below, we show that $\mathcal{A}(k) = \mathcal{F}(\geq k)$, by first showing that $\mathcal{A}(k) \subset \mathcal{F}(\geq k)$, and then the inverse.

Suppose $C \in \mathcal{A}(k)$. Then, there are such $G_1, \ldots, G_k$, that C intersects each G_i, i.e. $C \cap G_i \neq \emptyset$. Since G_i are disjoint, C must contain at least k elements, i.e., $C \in \mathcal{F}(\geq k)$. Conversely, suppose $C \in \mathcal{F}(\geq k)$. Then, there are unique $s_1, \ldots, s_k \in C$. Since $\mathbb{S}$ is Hausdorff, there are disjoint open sets $G_1, \ldots, G_k$, for which $s_i \in G_i$. It follows, that $C \cap G_i \neq \emptyset$, and then $C \in \cap_{i=1}^{k} \mathcal{F}_{G_i}$. Concluding, $C \in \mathcal{A}(k)$.

Now, since $\mathcal{F}(\geq k)$ is open in $\mathcal{F}$, a direct consequence is that $\mathcal{F}(\leq k) = \complement\mathcal{F}(\geq k+1)$ is closed. Then, $\mathcal{F}(k) = \mathcal{F}(\leq k) \cap \mathcal{F}(\geq k)$ is intersection of a closed and an open set, so $\mathcal{F}(k) \in \mathcal{B}(\mathcal{F})$. □

Next, we start to build a correspondence between the product spaces $\mathbb{S}^k = \times_{i=1}^{k} \mathbb{S}$ and the k-element finite sets on $\mathbb{S}$. Consider the map $\eta : \mathbb{S}^k \to \mathcal{F}(k)$ defined as follows.

$$\eta(s_1, \ldots, s_k) = \{s_1, \ldots, s_k\} \tag{5.9}$$

Clearly, η is surjective. It is not injective, since $\eta(s_1, \ldots, s_k) = \eta(s_{\pi 1}, \ldots, s_{\pi k})$ for any permutation π of integers from 1 to k. In addition, η is not defined at all in the case any two of the arguments are same, $s_i = s_j$ for $i \neq j$.

To make η injective, a lexicographic ordering $s \prec s'$ is introduced. Suppose $s = (\underline{v}, u) \in \mathbb{S}$ and $s' = (\underline{v}', u') \in \mathbb{S}$. Define that $s \prec s'$, if one of the following statements is true.

- $u < u'$
- $u = u'$ and $\underline{v}(1) < \underline{v}'(1)$
- $u = u'$, $\underline{v}(1) = \underline{v}'(1)$ and $\underline{v}(2) < \underline{v}'(2)$

  $\vdots$
- $u = u'$, $\underline{v}(i) = \underline{v}'(i)$, for all $i < d$, and $\underline{v}(d) < \underline{v}'(d)$

where $\underline{v}(i)$ denotes the i'th element of the vector $\underline{v}$. The order in the finite set U can be arbitrary total ordering, i.e. for any $u \neq u'$ either $u < u'$ or $u > u'$. Now, consider the map η restricted to the following set.

$$[\mathbb{S}]^k = \{(s_1, \ldots, s_k) : s_1, \ldots, s_k \in \mathbb{S} \text{ and } s_i \prec s_j \text{ for all } i < j\}$$

This restriction, denoted as $\eta_* \triangleq \eta|_{[\mathbb{S}]^k}$, is a homeomorphism (see Definition A.24) between $[\mathbb{S}]^k$ and $\mathcal{F}(k)$. Thus, a symmetric function f defined on $\mathbb{S}^k$ corresponds to exactly one set function f^* with domain $\mathcal{F}(k)$,

$$f^*(\{s_1, \ldots, s_k\}) = f(s_1, \ldots, s_k)$$

and vice versa. In addition, the measurable and the continuous functions on $(\mathbb{S}^k, \mathcal{B}(\mathbb{S}^k))$ are exactly those measurable and continuous in $(\mathcal{F}(k), \mathcal{B}(\mathcal{F}(k)))$, respectively [Goodman et al. 1997, pp. 131–135].

Since $\mathcal{F}(k)$ is homeomorphic to $[\mathbb{S}]^k$, it makes sense to construct a measure to $\mathcal{B}(\mathcal{F})$ using a measure in $\mathbb{S}^k$. Let $\mathcal{O} \in \mathcal{B}(\mathcal{F})$, and define set functions $\mu_k : \mathcal{B}(\mathcal{F}) \to [0, \infty]$ as follows.

$$\mu_0(\mathcal{O}) \triangleq \delta_\emptyset(\mathcal{O}) = \begin{cases} 1, & \emptyset \in \mathcal{O} \\ 0, & \emptyset \notin \mathcal{O} \end{cases} \tag{5.10}$$

$$\mu_k(\mathcal{O}) \triangleq \overline{\lambda}^k(\overleftarrow{\eta_*}(\mathcal{O} \cap \mathcal{F}(k))), \quad k \geq 1 \tag{5.11}$$

Proposition 5.8 *The set functions μ_k are measures in $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$.*

Proof. Clearly, μ_k are nonnegative, and $\mu_k(\emptyset) = 0$ for all $k \geq 0$. Countable additivity of μ_0 is trivial, so assume $k \geq 1$. Let $\{\mathcal{O}^{(i)}\}_{i=1}^\infty$ be a sequence of disjoint and measurable collections, $\mathcal{O}^{(i)} \in \mathcal{B}(\mathcal{F})$. Denote $\mathcal{O}_k^{(i)} = \mathcal{O}^{(i)} \cap \mathcal{F}(k)$. Then,

$$\begin{aligned} \mu_k(\bigcup_{i=1}^\infty \mathcal{O}^{(i)}) &= \overline{\lambda}^k(\overleftarrow{\eta_*}(\bigcup_{i=1}^\infty \mathcal{O}_k^{(i)})) = \overline{\lambda}^k(\bigcup_{i=1}^\infty \overleftarrow{\eta_*}(\mathcal{O}_k^{(i)})) \\ &= \sum_{i=1}^\infty \overline{\lambda}^k(\overleftarrow{\eta_*}(\mathcal{O}_k^{(i)})) = \sum_{i=1}^\infty \mu_k(\mathcal{O}^{(i)}) \end{aligned}$$

In addition, since $\mathcal{F}(k)$ and $\mathbb{S}^k$ are homeomorphic, for every $\mathcal{O} \in \mathcal{B}(\mathcal{F})$, the preimage $\overleftarrow{\eta_*}(\mathcal{O}_k) \in \mathcal{B}(\mathbb{S})$, and thus is $\overline{\lambda}$-measurable. □

For practical reasons, it is convenient to work with the original map η rather than η_*, so the following equality is worth noticing.

Proposition 5.9 *For $k \geq 1$, the following holds.*

$$\mu_k(\mathcal{O}) = \frac{1}{k!}\overline{\lambda}^k(\overleftarrow{\eta}(\mathcal{O} \cap \mathcal{F}(k)))$$

Proof. The equality follows from the fact, that the choice of η_* is not unique. One may choose any other η_{*i}, that is similar restriction than η_*, but with a different choice of the basis. That is, instead of restricting into $[\mathbb{S}]^k$, such that $s_1 \prec s_2 \prec \cdots \prec s_k$ one may choose any permutation π_i and restrict to $[\mathbb{S}]^k_{\pi_i}$ determined so that $s_{\pi_i 1} \prec s_{\pi_i 2} \prec \cdots \prec s_{\pi_i k}$.

Suppose that $\mathcal{O} \in \mathcal{B}(\mathcal{F})$ and denote $\mathcal{O}_k = \mathcal{O} \cap \mathcal{F}(k)$. Clearly,

$$\overleftarrow{\eta}(\mathcal{O}_k) = \bigcup_{i=1}^{k!} \overleftarrow{\eta_{*i}}(\mathcal{O}_k)$$

where η_{*i} are all unique restrictions into $[\mathbb{S}]^k_{\pi_i}$, which follow from different permutations π_i. Clearly also $\overleftarrow{\eta_{*i}}(\mathcal{O}_k) \cap \overleftarrow{\eta_{*j}}(\mathcal{O}_k) = \emptyset$ for all $i \neq j$. Then, it follows that

$$\overline{\lambda}^k(\overleftarrow{\eta}(\mathcal{O}_k)) = \overline{\lambda}^k(\bigcup_{i=1}^{k!} \overleftarrow{\eta_{*i}}(\mathcal{O}_k)) = \sum_{i=1}^{k!} \overline{\lambda}^k(\overleftarrow{\eta_{*i}}(\mathcal{O}_k)) = k! \cdot \overline{\lambda}^k(\overleftarrow{\eta_*}(\mathcal{O}_k))$$

where the last equality is due to invariance of the product hybrid Lebesgue measure with respect to the order of the basis. □

Remark 5.10 One can choose any other measure on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ than the hybrid Lebesgue measure $\overline{\lambda}$. For example, suppose that ξ is a measure on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$. Then, one can define measure ν_k on $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ as follows.

$$\nu_k(\mathcal{O}) = \xi^k(\overleftarrow{\eta}(\mathcal{O} \cap \mathcal{F}(k)))$$

Remark 5.11 The subcollections of $\complement\mathcal{F}(k) = \mathcal{F} \setminus \mathcal{F}(k)$ are null sets with respect to measure μ_k. Especially, collections of infinite sets are null sets, with respect to all μ_k. That is,

$$\mu_k(\mathcal{F} \setminus \mathcal{F}_*) = 0$$

where $\mathcal{F}_* = \bigcup_{k=0}^{\infty} \mathcal{F}(k)$ is the collection of all finite sets.

5.3.3 Set Measure and Integral

We introduce a measure to $\mathcal{F}$ using the measures μ_k given in Equation (5.10) and Equation (5.11) as follows.[6]

$$\mu(\mathcal{O}) = \sum_{k=0}^{\infty} \mu_k(\mathcal{O}) = \delta_{\{\emptyset\}}(\mathcal{O}) + \sum_{k=1}^{\infty} \frac{1}{k!}\overline{\lambda}^k(\overleftarrow{\eta}(\mathcal{O}_k)) \tag{5.12}$$

6. This measure was given by Valkonen [2002], but without explicit development of μ_k.

Proposition 5.12 *The set function μ is a measure in $\mathcal{B}(\mathcal{F})$.*

Proof. Clearly, μ is nonnegative, and $\mu(\emptyset) = 0$. In addition, μ is countably additive, for suppose $\{\mathcal{O}^{(i)}\}_{i=1}^{\infty}$ is a sequence of disjoint collections. Then

$$\mu(\bigcup_{i=1}^{\infty}\mathcal{O}^{(i)}) = \sum_{k=0}^{\infty}\mu_k(\bigcup_{i=1}^{\infty}\mathcal{O}^{(i)}) = \sum_{k=0}^{\infty}\sum_{i=1}^{\infty}\mu_k(\mathcal{O}^{(i)}) = \sum_{i=1}^{\infty}\mu(\mathcal{O}^{(i)})$$

where the last equality follows, since the summation order can be changed in a positive term series. □

The measure μ defined in Equation (5.12) introduces, of course, an integral to the real-valued functions defined in $\mathcal{F}$. The integral is identical to the set integral defined in [Goodman et al. 1997, p. 142].

Definition 5.13 *The **set integral** is defined as an integral with respect to the measure μ for all $\mathcal{O} \in \mathcal{B}(\mathcal{F})$ as follows.*

$$\begin{aligned}\int_{\mathcal{O}} f(Z)\delta Z \triangleq \int_{\mathcal{O}} f(Z)\mathrm{d}\mu(Z) &= \int_{\cup_{k=0}^{\infty}\mathcal{O}_k} f(Z)\mathrm{d}\mu(Z) = \sum_{k=0}^{\infty}\int_{\mathcal{O}_k} f(Z)\mathrm{d}\mu_k(Z) \\ &= f(\{\emptyset\})\delta_{\{\emptyset\}}(\mathcal{O}) + \sum_{k=1}^{\infty}\frac{1}{k!}\int_{\overleftarrow{\eta}(\mathcal{O}_k)} f(\{s_1,\ldots,s_k\})\mathrm{d}\overline{\lambda}^k(s_1,\ldots,s_k)\end{aligned}$$

where the shorthand notation $\mathcal{O}_k = \mathcal{O} \cap \mathcal{F}(k)$ is used as usual.

Suppose $\mathcal{O}(C) = \complement(\mathcal{F}_{\complement C})$, for some closed, nonempty $C \subset \mathbb{S}$. That is, $\mathcal{O}(C)$ is the collection of closed subsets of C. Then, it is clear that $\overleftarrow{\eta}(\mathcal{O}(C) \cap \mathcal{F}(k)) = C^k$. In this case, the set integral can be written in the following form.

$$\begin{aligned}\int_{\mathcal{O}(C)} f(Z)\delta Z &= f(\{\emptyset\}) + \sum_{k=1}^{\infty}\frac{1}{k!}\int_{C^k} f(\{s_1,\ldots,s_k\})\mathrm{d}\overline{\lambda}^k(s_1,\ldots,s_k) \\ &= f(\{\emptyset\}) + \sum_{k=1}^{\infty}\frac{1}{k!}\int_{C}\cdots\int_{C} f(\{s_1,\ldots,s_k\})\mathrm{d}\overline{\lambda}(s_1)\cdots\mathrm{d}\overline{\lambda}(s_k)\end{aligned}$$

where the last form follows from Fubini's theorem.

Remark 5.14 The set integral over the collection $\mathcal{O}(C)$ above is often denoted briefly as follows

$$\int_{C} f(Z)\delta Z \triangleq \int_{\mathcal{O}(C)} f(Z)\delta Z$$

when no confusion should arise. In addition, the set integral over $\mathcal{F}$ is denoted, as

$$\int f(Z)\delta(Z) \triangleq \int_{\mathcal{F}} f(Z)\delta(Z) = \int_{\mathbb{S}} f(Z)\delta(Z)$$

Remark 5.15 Since the set integrals are general integrals with respect to a certain measure, they admit the properties of a general integral, for example the properties listed in Section A.5.2. Goodman et al. [1997] (p. 141) *define* such functions f set integrable, for which each k-element function $f(\{s_1, \ldots, s_k\})$ is $\overline{\lambda}^k$-integrable, and for sufficiently large k, the function is identically zero. It is true, that functions of this type are μ-integrable, but here, the restriction requiring an upper bound for k is not made. The only restriction is the general integrability.

5.3.4 Belief Measure

This section aims to provide means to construct RFS distributions. Direct construction of a probability measure $P_{\boldsymbol{X}}$ may be considered impractical for most purposes. Therefore, other methods are needed. The first is the belief measure, introduced in Section 5.1.1 in finite universe case.

Definition 5.16 *The **belief measure** of a random finite set $\boldsymbol{X}$ is defined as*

$$\beta_{\boldsymbol{X}}(C) \triangleq P(\boldsymbol{X} \subset C) = P(\boldsymbol{X} \in \mathcal{O}(C))$$

where $\mathcal{O}(C) = \complement(\mathcal{F}_{\complement C})$. The belief measure is defined for all $\overline{\lambda}$-measurable $C \subset \mathbb{S}$.

One should notice, that the belief measure $\beta_{\boldsymbol{X}}(C)$ is not a measure, since $\beta_{\boldsymbol{X}}(\emptyset) \neq 0$ in general. The belief measure characterises the probability measure P uniquely according to Choquet Theorem (5.2).[7]

The next result is important, when one wishes to construct a belief measure in practice. The theorem is given in a form that is not restricted to bounded RFS, which is the case in [Goodman et al. 1997, p. 155].

Theorem 5.17 *Let $\boldsymbol{X}$ be a RFS, with belief measure $\beta_{\boldsymbol{X}}(C) = P(\boldsymbol{X} \subset C)$. The belief measure can be written in a factored form as follows.*

$$\beta_{\boldsymbol{X}}(C) = a_0 + \sum_{k=1}^{\infty} a_k q_k(C^k)$$

where the constants $a_k \geq 0$ sum to unity, and q_k are probability measures in $\mathbb{S}^k$. In addition,

$$a_k = P(|\boldsymbol{X}| = k), \qquad q_k(C^k) = P\left(\boldsymbol{X} \subset C \;\middle|\; |\boldsymbol{X}| = k\right) \tag{5.13}$$

Conversely, any choice of nonnegative a_k that sum to unity, and probability measures q_k in $\mathbb{S}^k$ specify a belief measure of a RFS.

Proof. Suppose $\boldsymbol{X}$ is a RFS. Then, $|\boldsymbol{X}(\omega)|$ is a random integer, since the collections $\mathcal{F}(k)$ are measurable. One can define $a_k = P(\boldsymbol{X} \in \mathcal{F}(k)) = P(|\boldsymbol{X}| = k)$ for all $k \geq 0$. Obviously, $\sum_{k=1}^{\infty} a_k = 1$, since $\overleftarrow{\boldsymbol{X}}(\cup_{k=1}^{\infty} \mathcal{F}(k)) = \Omega$. Then, since $\mathcal{O}(C)$

7. Set $T(K) = 1 - \beta_{\boldsymbol{X}}(\complement K)$ for all compact K.

and $\mathcal{F}(k)$ are measurable, so is their intersection. Therefore, for $k \geq 1$, one can define

$$q_k(C^k) = P(\boldsymbol{X} \in \mathcal{O}(C) \mid \boldsymbol{X} \in \mathcal{F}(k)) = \begin{cases} \dfrac{P_{\boldsymbol{X}}(\mathcal{O}(C) \cap \mathcal{F}(k))}{a_k}, & a_k > 0 \\ M_k(C^k), & a_k = 0 \end{cases}$$

where $M_k(C)$ is any probability measure in $\mathbb{S}^k$. Since a_k and q_k are constructed as described, Equation (5.13) is satisfied.

Suppose then that a_k and q_k satisfy the given conditions. Define $b(C) = a_0 + \sum_{k=1}^{\infty} a_k q_k(C^k)$ and $T(K) = 1 - b(\complement K)$. Next, it is proved that $T(K)$ satisfies the conditions of a Choquet capacity functional (Theorem 5.2).

1. $T(\emptyset) = 1 - b(\complement\emptyset) = 1 - b(\mathbb{S}) = 1 - \sum_{k=0}^{\infty} a_k q_k(\mathbb{S}^k) = 1 - \sum_{k=0}^{\infty} a_k = 0$.
2. Suppose $\{K_i\}_{i=1}^{\infty}$ is a decreasing sequence of compact sets, such that $\cap_i K_i = K$. Then, $G_i = \complement K_i$ is an increasing sequence of open sets, such that $\cup_i G_i = G = \complement K$. Then,

$$\begin{aligned} \lim_{i\to\infty} T(K_i) &= 1 - \lim_{i\to\infty} b(G_i) = 1 - \lim_{i\to\infty} \left[a_0 + \sum_{k=1}^{\infty} a_k q_k(G_i^k) \right] \\ &= 1 - \left[a_0 + \sum_{k=1}^{\infty} a_k q_k(G^k) \right] = 1 - b(G) = T(K) \end{aligned}$$

3. Suppose $K, K_1, \ldots, K_n \in \mathcal{K}$ are arbitrary compact sets, and the open sets $G, G_1, \ldots, G_n \in \mathcal{G}$ are their complements, respectively. Let us show that $S_n(K; K_1, \ldots, K_n) = b\left(G \setminus \bigcup_{i=1}^{n} G_i\right)$, which is nonnegative for all n. Clearly, the proposition holds for $n = 1$, since

$$\begin{aligned} S_1(K; K_1) &= T(K \cup K_1) - T(K) = (1 - b(G \cap G_1)) - (1 - b(G)) \\ &= b(G) - b(G \cap G_1) = b(G \setminus G_1) \geq 0 \end{aligned}$$

where the last equality follows from additivity of b. Next, suppose that $S_{n-1}(K; K_1, \ldots, K_{n-1}) = b\left(G \setminus \bigcup_{i=1}^{n-1} G_i\right)$. Then,

$$\begin{aligned} S_n(K; \cdots) &= S_{n-1}(K; K_1, \ldots, K_{n-1}) - S_{n-1}(K \cup K_n; K_1, \ldots, K_{n-1}) \\ &= b\left(G \setminus \bigcup_{i=1}^{n-1} G_i\right) - b\left(G_n \cap G \setminus \bigcup_{i=1}^{n-1} G_i\right) = b\left(G \setminus \bigcup_{i=1}^{n} G_i\right) \end{aligned}$$

Since T is a Choquet capacity functional, it specifies a unique probability measure $P_{\boldsymbol{X}}$. It is clear, that the belief measure of this probability measure is b. □

Theorem 5.17 states that the belief measure, hence the probability distribution, of a RFS $\boldsymbol{X}$ is completely characterised by a discrete probability distribution of number of elements on $\mathbb{N}$, and probability measures in $\mathbb{S}^k$. This form of the belief measure is encountered in Chapter 7.

Example 5.18 Suppose $\mathbb{S} = \mathbb{R}^d$. Let $a \in (0, 1)$, and $q(C) = N(C; \underline{m}, \mathrm{R})$. Then, a belief measure can be constructed as follows.

$$\beta_{\boldsymbol{X}}(C) = 1 - a + aq(C)$$

Obviously, $P(|\boldsymbol{X}| = 0) = 1 - a$, $P(|\boldsymbol{X}| = 1) = a$, and for $k \geq 2$, $P(|\boldsymbol{X}| = k) = 0$. That is, this model allows only singletons and the empty set.

If the situation $a = 1$ is allowed, the random set model above reduces into a singleton, $P(|\boldsymbol{X}| = 1) = 1$. A singleton random set can be considered similar to a random element in $\mathbb{S}$. Obviously, if $a = 0$, the random set model is degenerate, $P(\boldsymbol{X} = \emptyset) = 1$. ♢

Definition 5.19 *Suppose $\beta_{\boldsymbol{X}}(T) > 0$. Then, the conditional belief measure is defined as*

$$\beta_{\boldsymbol{X}}(C \mid T) \triangleq P_{\boldsymbol{X}}(\mathcal{O}(C) \mid \mathcal{O}(T)) = \frac{\beta_{\boldsymbol{X}}(C \cap T)}{\beta_{\boldsymbol{X}}(T)}$$

5.3.5 Set Density and Set Derivative

The next practical issue is to consider the density function of a random finite set distribution. The definition is based on the set integral developed above in Section 5.3.3.

Definition 5.20 *Suppose f is a nonnegative integrable finite set function, $f : \mathcal{F}_* \to [0, \infty]$, such that*

$$\int f(Z)\delta Z = 1$$

*Then, the function f is a **set density function**.*

A set density $f_{\boldsymbol{X}}$ defines uniquely a probability measure to $\mathcal{F}$, as follows.

$$P_{\boldsymbol{X}}(\mathcal{O}) = \int_{\mathcal{O}} f_{\boldsymbol{X}}(Z)\delta Z$$

The measure $P_{\boldsymbol{X}}$ is absolutely continuous with respect to μ, and $f_{\boldsymbol{X}}$ is the probability density function of $\boldsymbol{X}$ with respect to μ.

In this section, a construction of the density function from a given belief measure is considered. The construction is based on the set derivative, that is defined using the generalised Radon-Nikodym derivative [Goodman et al. 1997, p. 145].

Definition 5.21 *Let $\Phi : \mathcal{F} \to \mathbb{R}^d$ be a set function, and $s = (\underline{v}, u) \in \mathbb{S}$. The **generalised Radon-Nikodym derivative** (GRND) is defined as a function*

$$\frac{\delta\Phi}{\delta s}(T) = \lim_{j\to\infty} \lim_{i\to\infty} \frac{\Phi\big[(T \setminus B_s(1/j)) \cup \overline{B}_s(1/i)\big] - \Phi\big[T \setminus B_s(1/j))\big]}{\overline{\lambda}(\overline{B}_s(1/i))}$$

if the limit exists.

In the case $s \notin T$, and noticing that $\overline{B}_s(1/i) = \overline{B}_{\underline{v}}(1/i) \times \{u\}$ for all $i > 1$, the GRND simplifies to the following form.

$$\frac{\delta\Phi}{\delta s}(T) = \lim_{i\to\infty} \frac{\Phi\big[T \cup \overline{B}_s(1/i)\big] - \Phi\big[T\big]}{\lambda(\overline{B}_{\underline{v}}(1/i))}$$

Iterated GRNDs are defined as follows.

$$\frac{\delta^k\Phi}{\delta s_1 \cdots \delta s_k}(T) \triangleq \frac{\delta}{\delta s_1}\left[\frac{\delta^{k-1}\Phi}{\delta s_2 \cdots \delta s_k}\right](T)$$

The set derivative is defined using the GRND as follows.

Definition 5.22 *Let $\Phi : \mathcal{F} \to \mathbb{R}^d$ be a set function, and $Z = \{s_1, \ldots, s_k\} \subset \mathbb{S}$ a finite set. The set derivative is defined as*

$$\frac{\delta\Phi}{\delta Z}(T) \triangleq \frac{\delta^k\Phi}{\delta s_1 \cdots \delta s_k}(T)$$
$$\frac{\delta\Phi}{\delta\emptyset}(T) \triangleq \Phi(T)$$

if all the iterated generalised RNDs exist.

It is rather laborious to obtain set derivatives starting directly from Definition 5.22. The next section provides a set calculus "toolbox", containing results that make the derivation much more straightforward.

5.3.6 Some Properties of Set Integral and Set Derivative

This section gives a summary of some basic properties of the set derivative and the set integral. Assume that $\beta(C)$ is a belief measure, and $f(Z)$ is a set density function. The set derivative and the set integral are related to each other as follows [Mahler 2000, p. 30].

$$\beta(C) = \int_C \frac{\delta\beta}{\delta Z}(\emptyset)\delta Z \qquad f(Z) = \frac{\delta}{\delta Z}\left[\int_C f(Y)\delta Y\right](C = \emptyset) \tag{5.14}$$

Suppose that $\beta, \beta_1, \ldots, \beta_n$ are belief measures, $a_1, a_2 \in \mathbb{R}$ are constants, and $g(x_1, \ldots, x_n) : [0,1]^n \to \mathbb{R}$ is a function. The set derivative has the following properties [Mahler 2000, pp. 31–32].

Sum rule

$$\frac{\delta}{\delta Z}\left[a_1\beta_1(C) + a_2\beta_2(C)\right] = a_1\frac{\delta\beta_1}{\delta Z}(C) + a_2\frac{\delta\beta_2}{\delta Z}(C) \tag{5.15}$$

Product rules

$$\frac{\delta}{\delta z}\left[\beta_1(C)\beta_2(C)\right] = \frac{\delta\beta_1}{\delta z}(C)\beta_2(C) + \frac{\delta\beta_2}{\delta z}(C)\beta_1(C) \tag{5.16}$$

$$\frac{\delta}{\delta Z}\left[\beta_1(C)\beta_2(C)\right] = \sum_{W\subset Z}\left[\frac{\delta\beta_1}{\delta W}(C)\frac{\delta\beta_2}{\delta(Z\setminus W)}(C)\right] \tag{5.17}$$

Constant rule

$$\frac{\delta}{\delta Z}a_1 = 0, \quad \text{if } Z \neq \emptyset \tag{5.18}$$

Chain rules

$$\frac{\delta}{\delta z}g(\beta(C)) = \frac{\mathrm{d}g}{\mathrm{d}x}(\beta(C))\frac{\delta\beta}{\delta z}(C) \tag{5.19}$$

$$\frac{\delta}{\delta z}g(\beta_1(C),\dots,\beta_n(C)) = \sum_{i=1}^{n}\left[\frac{\partial g}{\partial x_i}(\beta_1(C),\dots,\beta_n(C))\frac{\delta\beta_i}{\delta z}(C)\right] \tag{5.20}$$

Power rule

$$\frac{\delta}{\delta Z}p(C)^n = \begin{cases} \dfrac{n!}{(n-|Z|)!}p(C)^{n-|Z|}\displaystyle\prod_{z\in Z} f_p(z), & |Z| \leq n \\ 0, & |Z| > n \end{cases} \tag{5.21}$$

where p is a probability measure of a random vector, and f_p its density. In Equation (5.21), the conventions $0^0 = 1$ and $0! = 1$ are used.

The following result is needed in Chapter 7, but it is included here due to its general nature.

Proposition 5.23 *Suppose $\beta_1,\dots,\beta_n$ are belief measures. Then*

$$\left[\frac{\delta}{\delta z}\prod_{i=1}^{n}\beta_i\right](C) = \sum_{i=1}^{n}\left[\frac{\delta\beta_i}{\delta z}(C)\prod_{\substack{j=1\\ j\neq i}}^{n}\beta_j(C)\right] \tag{5.22}$$

Proof. The proof proceeds by induction. The case $n = 1$ is trivial. Then, suppose Equation (5.22) holds for $n = k-1$. Omitting the argument C for convenience,

$$\begin{aligned}\frac{\delta}{\delta z}\prod_{i=1}^{k}\beta_i &= \frac{\delta}{\delta z}\beta_k\prod_{i=1}^{k-1}\beta_i \overset{(*)}{=} \beta_k\frac{\delta}{\delta z}\prod_{i=1}^{k-1}\beta_i + \prod_{i=1}^{k-1}\beta_i\frac{\delta\beta_k}{\delta z} \\ &= \beta_k\sum_{i=1}^{k-1}\left[\frac{\delta\beta_i}{\delta z}\prod_{\substack{j=1\\ j\neq i}}^{k-1}\beta_j\right] + \prod_{i=1}^{k-1}\beta_i\frac{\delta\beta_k}{\delta z} = \sum_{i=1}^{k}\left[\frac{\delta\beta_i}{\delta z}\prod_{\substack{j=1\\ j\neq i}}^{k}\beta_j\right]\end{aligned}$$

where the equality $(*)$ follows from product rule in Equation (5.16). □

5.3.7 Probability Hypothesis Density

The random set density function, introduced in Section 5.3.5, is defined piecewise for different number of elements in the set. The density is hard to visualise, even in the case of a simple low-dimensional space, such as $\mathbb{R}$. On the other hand, if only one extracted estimate $\{x_1,\dots,x_n\} \subset \mathbb{S}$ of the density is shown, the information can be considered insufficient, or almost misleading. This is because the estimate does not contain any information of the uncertainty.

In the author's opinion, one possible visualisation of the set density could be the so called Probability Hypothesis Density (PHD). We give first the definition

of a more general density, the Probability Hypothesis Surface (PHS) [Goodman et al. 1997, p. 169].

Definition 5.24 *The **probability hypothesis surface** (PHS) function for random set $\boldsymbol{X}$, denoted by $D_{\boldsymbol{X}}(Z)$, is defined by*

$$D_{\boldsymbol{X}}(Z) \triangleq \int f_{\boldsymbol{X}}(Z \cup Y)\delta Y$$

*In the case $|Z| = 1$, the PHS is referred to as **probability hypothesis density**, and denoted as $D_{\boldsymbol{X}}(s) \triangleq D_{\boldsymbol{X}}(\{s\})$.*

The PHS can be obtained from the belief measure as follows

$$D_{\boldsymbol{X}}(Z) = \frac{\delta \beta_{\boldsymbol{X}}}{\delta Z}(\mathbb{S})$$

The probability hypothesis density satisfies

$$\mathsf{E}\left[|\boldsymbol{X} \cap A|\right] = \int_A D_{\boldsymbol{X}}(s)\mathrm{d}\overline{\lambda}(s)$$

for all measurable $A \subset \mathbb{S}$. That is, $D_{\boldsymbol{X}}$ is the RND of the measure $\mathsf{E}\left[|\boldsymbol{X} \cap \cdot|\right]$ with respect to $\overline{\lambda}$.

5.3.8 Finite Set Estimators

Since the space of finite sets does not constitute a vector space, taking expectations does not make sense. This means, that the most popular Bayes estimator, the EAP, cannot be used. So, alternative estimators are required. This section covers two proposed finite set estimators. Recollect that, in general, Bayes estimators are obtained from the posterior distribution. In this section, it is assumed that there are densities of the posterior distributions. The notations in this section are similar to notation in Chapter 3. That is, the unknown and estimated parameter process is $(\boldsymbol{X}_k)_{k\in\mathbb{N}}$, while the observation process is $(\boldsymbol{Y}_k)_{k\in\mathbb{Z}_+}$.

Two Bayes estimators, "GMAP-I" and "GMAP-II"[8] were presented in [Goodman et al. 1997, pp. 191–192]. Mahler [2000] stated new, more descriptive names for the estimators[9]: The Marginal Multitarget Estimator (MaME) and the Joint Multitarget Estimator (JoME), respectively. MaME is obtained so that one first finds $\widehat{n}$, the maximum of the cardinality distribution. Then, one finds a MAP estimate of the conditional distribution given that the number of elements is $\widehat{n}$. That is,

$$\widehat{n} = \arg\max_n P(|\boldsymbol{X}_k| = n \mid \boldsymbol{Y}_{1:k} = Y_{1:k})$$
$$\widehat{X}_k^{\mathrm{MaME}} = \arg\max_{|X|=\widehat{n}} f_{\boldsymbol{X}_k|Y_{1:k}}(X)$$

8. Global MAP of the first and second kind.
9. These names are also rather application-specific.

JoME differs from MaME in that it does not involve two-step maximisation procedure. JoME is formed directly by maximising the following quantity.

$$\widehat{X}_k^{\text{JoME}} = \arg\max_X f_{\boldsymbol{X}_k|Y_{1:k}}(X)\frac{c^{|X|}}{|X|}$$

where c is a "small" constant. Mahler [2000] states that both the estimators are Bayes-optimal (given an appropriate cost function), and that JoME has been proved to be a consistent estimator [Goodman et al. 1997, pp. 200–205].

Chapter 6

Target Tracking

In this thesis, "target tracking" is considered as a rather general state estimation problem. This chapter covers the specifics of a multitarget multisensor data fusion system, but the general problem is common in many other fields. The approach presented in this chapter is based on the Bayesian target tracking. The presentation is influenced by the references [Goodman et al. 1997; Stone 2001; Stone et al. 1999]. More engineering and implementation oriented approaches to target tracking can be found, e.g., in the books [Bar-Shalom and Fortmann 1988; Bar-Shalom and Li 1995; Blackman and Popoli 1999; Hall and McMullen 2004].

This chapter starts in a traditional manner, by first considering single target tracking in Section 6.1. This approach is fruitful in the sense that the single target tracking problem is rather easy to formulate rigorously, and the formulation is rather standard in the tracking community. Moreover, the models and ideas of single target tracking are needed also in the case of multiple targets. Next, Section 6.2 gives a summary of the most common methods for tracking multiple known number of targets. At last, the most general problem of tracking multiple unknown number of targets, is considered in Section 6.3.

6.1 Single Target Tracking

Modern single target tracking systems are usually based on recursive Bayesian estimation, covered in general in Chapter 3. The state-space model is built so that the Markov chain $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ (the signal process) represents the target state. Most commonly, the target state consists of the geokinematic part (usually position and velocity components in $\mathbb{R}^3$). Often, the target state is augmented to contain also some other types of elements, which are referred to as *attributes*. A common formulation for the state space is to consider a hybrid space $\mathbb{S} = \mathbb{R}^d \times U$ where $\mathbb{R}^d$ is Euclidean and U a discrete space. The hybrid space is discussed in detail in Section 5.3.1. The construction of the Markov chain $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ can be considered as *dynamic modelling*, since it indeed characterises the dynamic behaviour of the state variable $\boldsymbol{x}_k$. Some common methods of dynamic modelling are discussed in Section 6.1.1.

The observation process $(\boldsymbol{y}_k)_{k\in\mathbb{Z}_+}$ consists of the sensor reports. Each observation $\boldsymbol{y}_k$ is assumed to contain a measurement, usually noisy and indirect, of the target state at the same time $\boldsymbol{x}_k$. Each measurement $\boldsymbol{y}_k$ is assumed to have values in hybrid space $\mathbb{S}_k$. Specifically, it can occur that the measurement space and the target space are different, $\mathbb{S}_k \neq \mathbb{S}$. In addition, it is allowed that different measurements are defined in different spaces, so in general for $k \neq k'$ it may occur that $\mathbb{S}_k \neq \mathbb{S}_{k'}$. Section 6.1.2 covers some common issues of sensor modelling.

6.1.1 Dynamic Modelling

Often, in tracking applications, the geokinematic dynamics of the target are modelled most naturally as continuous-time processes. This is especially the case, if the estimation process is to be carried out in irregular sample intervals. In this case, the target state is considered to be a stochastic process $(\boldsymbol{x}_t)_{t\in[0,\infty)}$, where $\boldsymbol{x}_t : \Omega \to \mathbb{R}^d$. The stochastic process is often constructed using a stochastic differential equation. For computations, the continuous-time model is discretised.

Perhaps the simplest physically attractive dynamic (kinematic) model, the *constant velocity model*, is introduced next. The model is defined in the case $\mathbb{S} = \mathbb{R}^{2d}$, where d is the dimension of the position space. The state $\underline{\boldsymbol{x}}_t = (\underline{\boldsymbol{z}}_t, \underline{\boldsymbol{z}}'_t)$ consists of the position and velocity in $\mathbb{R}^d$, respectively. The constant velocity model can be given as follows [Bar-Shalom and Fortmann 1988, p. 83–84][1]

$$\frac{\mathrm{d}\underline{\boldsymbol{x}}_t}{\mathrm{d}t} = \begin{bmatrix} 0_d & \mathrm{I}_d \\ 0_d & 0_d \end{bmatrix} \underline{\boldsymbol{x}}_t + \rho \begin{bmatrix} 0_d \\ \mathrm{I}_d \end{bmatrix} \underline{\boldsymbol{w}}_t \tag{6.1}$$

where I_d and 0_d are a $d \times d$ identity and zero matrix, respectively. In addition, $\underline{\boldsymbol{w}}_t$ is a zero-mean "white noise process[2]" in $\mathbb{R}^d$, and the positive constant ρ is the process noise deviation. The model is, in fact, an "almost constant velocity model", since the velocity components are perturbed by the *process noise* $\rho\underline{\boldsymbol{w}}_t$. The process $\underline{\boldsymbol{x}}_t$ has independent increments, so one can create a discrete-time Markov chain corresponding the continuous-time version given above.

Suppose $(t_k)_{k\in\mathbb{N}} \subset [0,\infty)$ is a countable set of increasing times, i.e. $t_k \leq t_{k+1}$ for all $k \in \mathbb{N}$. Consider the sequence of random vectors $(\underline{\boldsymbol{x}}_k)_{k=1}^{\infty}$ where $\underline{\boldsymbol{x}}_k = \underline{\boldsymbol{x}}_{t_k}$. Then, the discrete-time model corresponding to Equation (6.1) can be represented as follows.

$$\boldsymbol{x}_{k+1} = \mathrm{A}_k \underline{\boldsymbol{x}}_k + \underline{\boldsymbol{v}}_k \tag{6.2}$$

where $\underline{\boldsymbol{v}}_k$ is a zero-mean Gaussian distributed random vector with a covariance matrix Q_k. The matrices A_k and Q_k can be given as follows.

$$\mathrm{A}_k = \begin{bmatrix} \mathrm{I}_d & \tau_k \mathrm{I}_d \\ 0_d & \mathrm{I}_d \end{bmatrix}, \qquad \mathrm{Q}_k = \rho^2 \begin{bmatrix} (\tau_k^3/3)\mathrm{I}_d & (\tau_k^2/2)\mathrm{I}_d \\ (\tau_k^2/2)\mathrm{I}_d & \tau_k \mathrm{I}_d \end{bmatrix}$$

---

1. The model was presented in the reference in the case $d = 1$, but here it is given in vector case explicitly, assuming that unique coordinate components are independent of each other.
2. Intuitively, $\underline{\boldsymbol{w}}_t$ can be considered to be such a process for which each $\underline{\boldsymbol{w}}_t$ is Gaussian, and for all $t \neq t'$ the random variables $\underline{\boldsymbol{w}}_t$ and $\underline{\boldsymbol{w}}_{t'}$ are independent and identically distributed. Unfortunately, such process does not exist. The theory of stochastic differential equations is beyond the scope of this thesis, but an interested reader may read a introduction e.g. in [Øksendahl 1985].

where $\tau_k = t_{k+1} - t_k$ is the time interval between $\underline{\boldsymbol{x}}_{k+1}$ and $\underline{\boldsymbol{x}}_k$. In a similar manner, *constant acceleration* and *coordinated turn* models can be formed [Bar-Shalom and Fortmann 1988; Karlsson 2002].

Quite often, a good model for target kinematics can be created from multiple simple kinematic models, such as the one given above. The multiple-model approach assumes, that the target can switch the kinematic model between the n possible models. The switch is assumed to happen instantaneously. The state variable $\boldsymbol{x}_k = (\underline{\boldsymbol{z}}_k, \boldsymbol{r}_k)$ is decomposed into a random vector $\underline{\boldsymbol{z}}_k$, and an integer-valued random variable $\boldsymbol{r}_k$. The vector $\underline{\boldsymbol{z}}_k$ corresponds to the geokinematic state, and is referred to as the base state. The integer $\boldsymbol{r}_k$ is the index of the kinematic model that has been active during the time interval $(t_{k-1}, t_k]$, and is referred to as the modal state [Bar-Shalom and Li 1993, p. 450]. The model can be given as follows.

$$\begin{aligned} P(\boldsymbol{r}_k = j \mid \boldsymbol{r}_{k-1} = i) &= [\mathrm{M}]_{ij} \\ P(\underline{\boldsymbol{z}}_k \in B \mid \underline{\boldsymbol{z}}_{k-1} = \underline{z}, \boldsymbol{r}_k = r) &= N(B; \mathrm{A}_k^{(r)} \underline{z}, \mathrm{Q}_k^{(r)}) \end{aligned} \tag{6.3}$$

where the matrix M is the Markov transition kernel for $\boldsymbol{r}_k$, and the matrices $\mathrm{A}_k^{(r)}$ and $\mathrm{Q}_k^{(r)}$ are the parameter matrices of the r'th kinematic model. An interested reader may see the PhD thesis of Herman [2002] for deeper discussion of multiple model kinematic modelling of an aeroplane.

If the state contains attributes, i.e. variables that cannot be directly associated with geokinematics, there are no general approaches to modelling the dynamic behaviour, since the attributes can be virtually anything. If the attributes can be considered independent of the geokinematics, the dynamic behaviour of the attributes and the geokinematics can be modelled as follows.

$$P(\underline{\boldsymbol{z}}_k \in B, \boldsymbol{a}_k \in A \mid \underline{\boldsymbol{z}}_{k-1}, \boldsymbol{a}_{k-1}) = P(\underline{\boldsymbol{z}}_k \in B \mid \underline{\boldsymbol{z}}_{k-1}) P(\boldsymbol{a}_k \in A \mid \boldsymbol{a}_{k-1})$$

where the state $\boldsymbol{x}_k = (\underline{\boldsymbol{z}}_k, \boldsymbol{a}_k) \in \mathbb{S}$ consists of the geokinematic part $\underline{\boldsymbol{z}}_k$, and the attribute part $\boldsymbol{a}_k$.

6.1.2 Measurement Modelling

While the dynamic state variable $\boldsymbol{x}_t$ is usually modelled as a continuous process, the measurements are assumed to depend on the state variable at one instantaneous time. That is, a measurement collected at time t_k is conditionally independent of all the other state variables given the current state $\boldsymbol{x}_k$. The measurement model is characterised by the conditional probability $P(\boldsymbol{y}_k \in C \mid \boldsymbol{x}_k)$ which is defined for any measurable $C \subset \mathbb{S}_k$. In this thesis, it is assumed that this probability measure is absolutely continuous with respect to the hybrid measure $\overline{\lambda}$ in $\mathbb{S}_k$, so that there is a density function $f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k)$ that satisfies the following equation.

$$P(\boldsymbol{y}_k \in C \mid \boldsymbol{x}_k = x_k) = \int_C f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k) \mathrm{d}\overline{\lambda}(y_k)$$

Often, a measurement model is constructed according to the following additive noise model.

$$\boldsymbol{y}_k = h(x_k) + \boldsymbol{v}_k$$

where $\boldsymbol{v}_k$ is a noise term, having a density $f_{\boldsymbol{v}}$ with respect to $\overline{\lambda}$, and $h : \mathbb{S} \to \mathbb{S}_k$ is a measurable function. Then, the density of $\boldsymbol{y}_k$ can be given as follows.

$$f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k) = f_{\boldsymbol{v}}(y_k - h(x_k))$$

Since there are many different types of sensors, no more specific models are listed here. The sensor models that are used in the experiments are described in Chapter 8.

Many sensors produce also *false alarms*, i.e. measurements that are due to uninteresting phenomena, e.g. environment conditions or sensor malfunction. It is clear, that false alarms must be included in the measurement model. If it is reasonable to assume, that a certain percentage of the measurements produced by the sensor are false alarms, one can create a simple measurement model as follows.

$$f'_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k) = (1 - p_f) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k) + p_f f_f(y_k)$$

where $p_f \in [0, 1]$ is the probability corresponding the percentage of false alarms, and $f_f(y_k)$ is the probability density function of the false measurements. Typically, $f_f(y_k)$ is quite uninformative, e.g. a uniform distribution.

In some applications, it is reasonable to assume that one can obtain information from a failed measurement attempt [Mori et al. 1986; Stone et al. 1999]. This is the case, if it is reasonable to model the *probability of detection*. A reasonable model can be constructed by allowing $\boldsymbol{y}_k$ to take the value $\ominus$ denoting a "failed measurement". That is, the measurement space is augmented with the new element, $\mathbb{S}'_k = \mathbb{S}_k \cup \{\ominus\}$. The measurement model density is altered as follows.

$$f'_{\boldsymbol{y}|\boldsymbol{x}}(y_k \mid x_k) = \begin{cases} 1 - p_d(x_k), & y_k = \ominus \\ p_d(x_k) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y_k \mid x_k), & y_k \neq \ominus \end{cases} \tag{6.4}$$

where $f_{\boldsymbol{y}_k|\boldsymbol{x}_k}$ is a density in $\mathbb{S}_k$. The measurable function $p_d(x) : \mathbb{S} \to [0, 1]$ determines the probability of detection of a target in location $x \in \mathbb{S}$. This framework is described in detail and exemplified in [Stone et al. 1999, pp. 46–49].

6.1.3 Filtering

Given the above described kinematic model and measurement model, tracking is just estimation of the unknown state variable $\boldsymbol{x}_k$. Of course, what is still missing, is the prior distribution $\pi_{0|0}$. The prior distribution is, most often, the parameter that is more or less heuristically given. In that case, the prior distribution is best to be as uninformative as possible.

Once all the models are fixed, the Bayesian estimation needs to be addressed. Chapter 3 contained the basic methods how one can perform this inference in practice. If one is lucky, the state and measurement spaces are Euclidean,

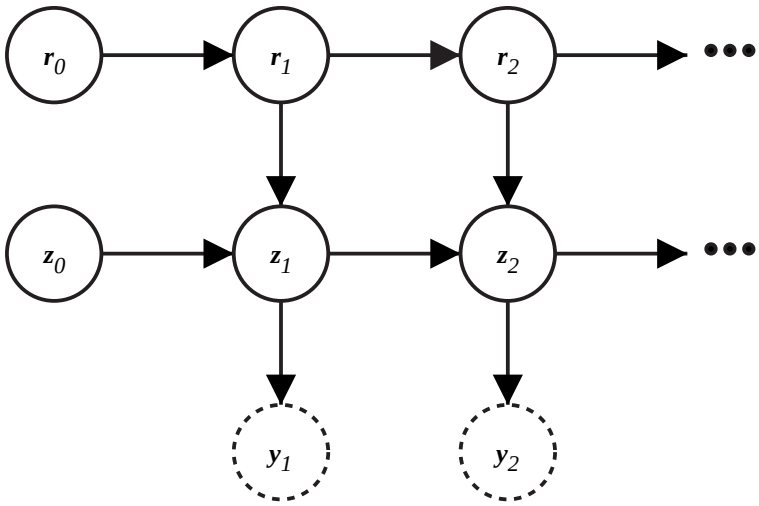


Figure 6.1: A switching Kalman filter model as a dynamic Bayesian network.

the prior density is Gaussian, and the conditional densities are linear Gaussian, as given in Equation (3.9). In this case, the Kalman filter can be used to obtain the exact posterior distribution $p_{k|k}$, from which any Bayes estimate can be computed. Most often, the EAP estimator is used, which is the posterior mean $\underline{m}_{k|k}$.

Obviously, it happens quite rarely in real life, that all the assumptions of the Kalman filter are satisfied. Most importantly, it often occurs that the linearity or the Gaussianity assumptions fail. The EKF algorithm can be considered perhaps the "standard" tracking algorithm, which has been applied to a variety of tracking problems. As mentioned in Section 3.3.2, the recently proposed UKF algorithm serves as an alternative for EKF.

If the multiple model approach is used in kinematic modelling, and the measurement model is linear-Gaussian, the model is referred to as a switching Kalman filter model[3] (SKFM). For convenience, the SKFM model is summarised below.

$$\begin{aligned} P(\boldsymbol{r}_k = j \mid \boldsymbol{r}_{k-1} = i) &= [\mathrm{M}]_{ij} \\ P(\underline{\boldsymbol{z}}_k \in B \mid \underline{\boldsymbol{z}}_{k-1} = \underline{z}, \boldsymbol{r}_k = r) &= N(B; \mathrm{A}_k^{(r)} \underline{z}, \mathrm{Q}_k^{(r)}) \\ P(\underline{\boldsymbol{y}}_k \in C \mid \underline{\boldsymbol{z}}_k = \underline{z}) &= N(C; \mathrm{H}_k \underline{z}, \mathrm{R}_k) \end{aligned} \tag{6.5}$$

Figure 6.1 shows SKFM as a dynamic Bayesian network. In principle, inference in a SKFM can be performed in an exact manner. This is due to the fact that if $\boldsymbol{r}_1 = r_1, \ldots$ are given, the model reduces to a KFM. In practice, the computational complexity of an exact inference algorithm will grow exponentially, due

3. Also referred to as a switching SSM, a switching linear dynamical system (LDS), a jump-Markov model, a jump-linear system, a conditional dynamic linear model (DLM), etc. [Murphy 2002, p. 41].

to an increasing number of Gaussian mixture components[4] [Murphy 2002, Section B.5.2]. Therefore, approximate recursive inference methods must be used in practice.

There are a couple of straightforward approximate inference algorithms that may be used for approximate recursive inference of a SKFM. The idea of the algorithms is to keep the number of mixture components constant. The generalised second order pseudo-Bayesian algorithm (GPB2) algorithm is based on the idea, that given a mixture distribution of q Gaussians on round $k-1$, the algorithm performs q^2 KF updates to obtain the exact posterior with q^2 Gaussians. Then, the q^2 Gaussians are "collapsed[5]" into q Gaussians. The interacting multiple model (IMM) algorithm [Blom and Bar-Shalom 1988], which seems to be more popular in tracking applications, is very similar to GPB2. The IMM algorithm reduces the number of Gaussians to q after the prediction, but before performing the KF update, so that only q KF updates are needed [Murphy 2002, Section 4.3]. The IMM algorithm is found to perform almost as well as GPB2, but with less computational load [cf. Bar-Shalom and Li 1993, pp. 482–483].

As the reader may guess, the above mentioned approximate inference algorithms may be applied also in other cases, if the state variable can be factored into Euclidean and discrete parts, and if the conditional densities are conditionally linear-Gaussian, as given in Equation (6.5). Furthermore, if the dependencies are again "not exactly, but close to linear-Gaussian", the first-order linearisation may be applied as well, to obtain, for example, an "EKF-IMM" algorithm. The PhD thesis of Fearnhead [1998] contains more discussion on EKF, IMM and GPB, as well as some other approaches.

There are two classes of methods that remain applicable for cases in which the densities are strongly multimodal and/or non-Gaussian: the sample based Monte Carlo methods, and the grid methods based on a (deterministic) discretisation. The discretisation methods, in general, are rather tricky to implement in a real-life application. This is due to the often high-dimensional state space. Therefore, one must use some kind of dynamic discretisation method, that concentrates the grid points to the high-probability regions of the state space. The book of Stone et al. [1999] and the PhD thesis of Bergman [1999] consider this, certainly not trivial, approach.

Recently, there has been a "Monte Carlo gold rush" in nonlinear and non-Gaussian tracking. The first tracking application of sequential Monte Carlo (SMC) seems to be the bearings-only tracking in [Gordon et al. 1993]. Ever since, SMC have been applied at least to positioning and navigation [Bergman 1999; Gustafsson et al. 2002; Karlsson et al. 2003], bearings-only tracking [Carpenter et al. 1999; Gilks and Berzuini 2001], visual tracking [Blake and Isard 1997], passive radar

4. Starting from a Gaussian prior, the number of mixture components in the posterior at time instant k is q^k, where q is the number of kinematic models, i.e. the number of values $\boldsymbol{r}_k$ can take.

5. In fact, the mixture of Gaussians distribution is replaced with a Gaussian distribution with the mean and the covariance of the mixture distribution. This procedure is sometimes referred to as a "weak marginalisation" in the Bayesian network literature [Murphy 2002, Section B.5.2].

tracking [Herman 2002], and ground moving target indicator (GMTI) tracking [Mallick et al. 2002]. Most of the publications are applications of the bootstrap filter, that was introduced by Gordon et al. [1993] (possibly with minor modifications). The SMC approach has been compared with the performance of the EKF on some nonlinear tracking problems, and SMC has been found to overcome EKF. Of course, the computational requirements of the SMC methods are usually much higher than the requirements of EKF, so comparison may be considered somehow "unfair". The specifics of these Monte Carlo tracking algorithms are not covered here. An interested reader is advised to take a look of the recent book devoted to SMC in single-target tracking [Ristic et al. 2004].

6.2 Tracking of Multiple Known Number of Targets

Tracking multiple known number of targets does not differ conceptually from single-target tracking. Suppose there are n targets to track. It suffices to concatenate the state variables of the targets together, and obtain a joint state variable $\boldsymbol{x}_k = (\boldsymbol{x}_k^{(1)}, \ldots, \boldsymbol{x}_k^{(n)})$, where $\boldsymbol{x}_k^{(i)}$ is the state variable of the i'th target. That is, the n-target state space is the Cartesian product space of n single-target state spaces.

In practice, the multitarget tracking problem is, of course, trickier than the problem of single-target tracking. The only multitarget case, in which exact inference is tractable, is the HMM, i.e. the case when the state space $\mathbb{S}$ is finite[6]. That is, excluding the HMM case, every multitarget tracking problem is intractable. The problem has been studied for decades, and practical methods have been developed to obtain good suboptimal solutions. Below, in Sections 6.2.1 and 6.2.2 some common methods for building a multitarget model are discussed. In Section 6.2.3, a summary of computational solutions for multitarget tracking are given.

6.2.1 Dynamic Modelling

Usually, the construction of a multitarget dynamic model is carried out so that one assumes the target dynamics independent of each other. That is, the multitarget state evolution model is defined as follows.

$$\begin{aligned} P(\boldsymbol{x}_k \in B \mid \boldsymbol{x}_{k-1}) &= P(\boldsymbol{x}_k^{(1)} \in B^{(1)}, \ldots, \boldsymbol{x}_k^{(n)} \in B^{(n)} \mid \boldsymbol{x}_{k-1}^{(1)}, \ldots, \boldsymbol{x}_{k-1}^{(n)}) \\ &= \prod_{i=1}^{n} P(\boldsymbol{x}_k^{(i)} \in B^{(i)} \mid \boldsymbol{x}_{k-1}^{(i)}) \end{aligned} \tag{6.6}$$

where $B = \times_{i=1}^{n} B^{(i)}$. Figure 6.2 shows the independent motion multitarget model as a Bayesian network. In the figure, the prior state variable $\boldsymbol{x}_0$ is assumed to have independent components as well, $P(\boldsymbol{x}_0 \in B) = \prod_{i=1}^{n} P(\boldsymbol{x}_0^{(i)} \in B^{(i)})$.

6. And even in this case, the number of elements in $\mathbb{S}$ can only be moderate.

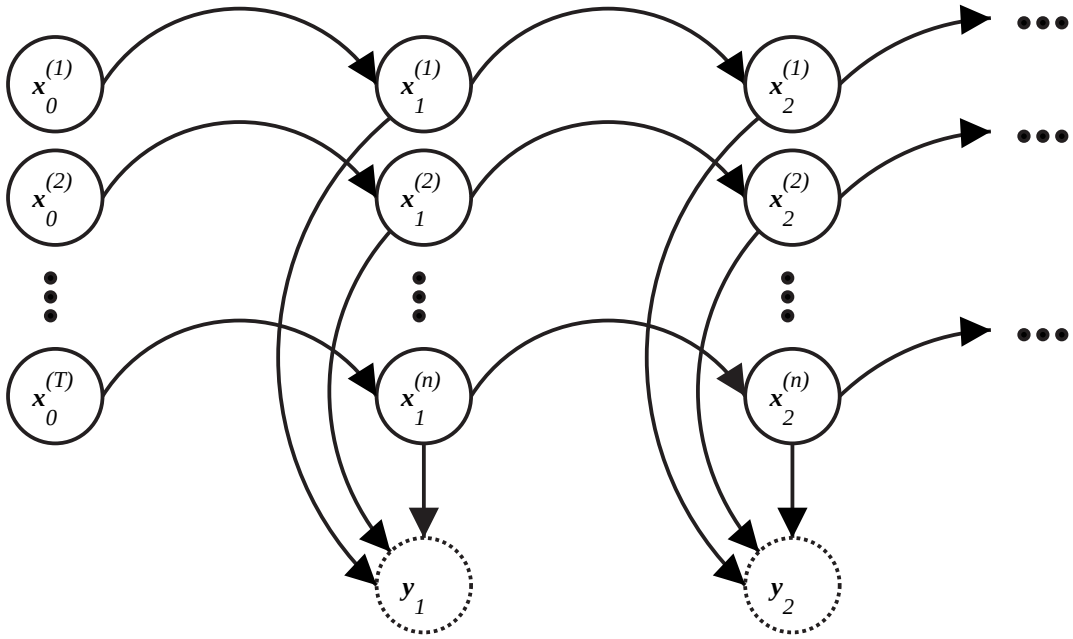


Figure 6.2: The special case of a multitarget tracking model, when the motion models of the targets are assumed independent of each other.

Mahler [2001b] (p. 14-5) suggests that the assumption of independent motions is unrealistic. The assumption is, of course, not necessary from a theoretical perspective. In practice, however, it may be quite hard to construct a multitarget state evolution model, in which the inter-target dependence is modelled in a sufficient accuracy. In addition, the independence assumption admits simplifications in computations.

6.2.2 Measurement Modelling

Consider first that there are single measurements, that arrive sequentially. Often, it is reasonable to assume that a single measurement depends on at most one target. In that case, one may include a so called association[7] variable $\boldsymbol{c}_k$, which tells which target (if any) originated the measurement. That is, one assumes that the measurement model can be given as follows.

$$P(\boldsymbol{y}_k \in C \mid \boldsymbol{x}_k, \boldsymbol{c}_k = i) = \begin{cases} P(\boldsymbol{y}_k \in C \mid \boldsymbol{x}_k^{(i)}), & i > 1 \\ P_f(\boldsymbol{y}_k \in C), & i = 0 \end{cases} \tag{6.7}$$

where P_f is the distribution of false measurements. The association variable "switches" the dependency of the measurement of targets, as exemplified in Figure 6.3. Of course, the value of the association is generally unknown. The associ-

7. Also referred to as contact.

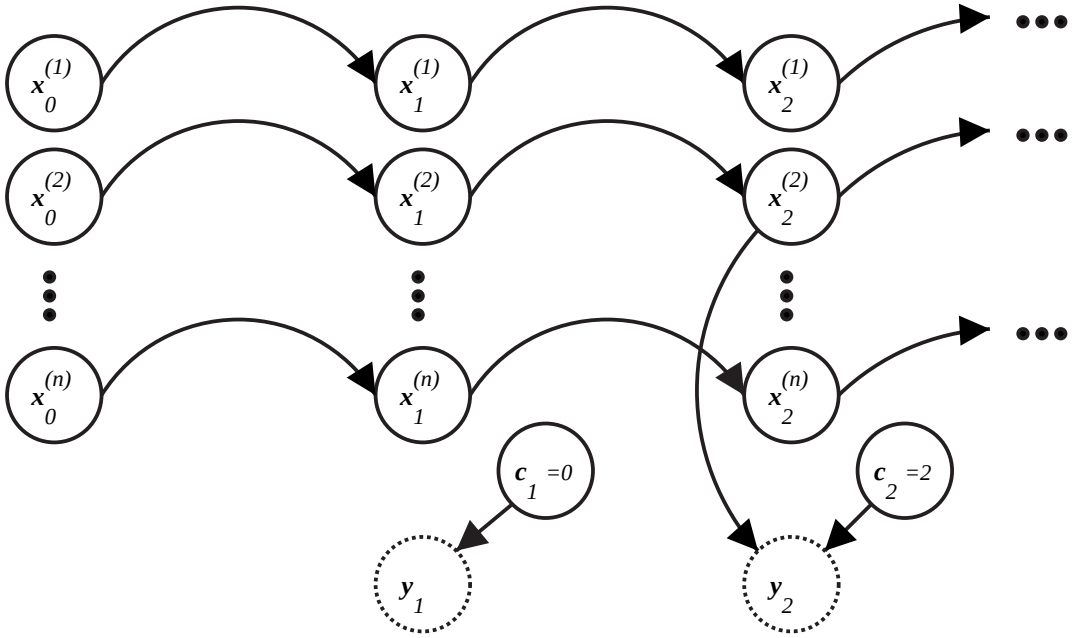


Figure 6.3: The special case of a multitarget tracking model, when the motion models are assumed independent of each other, and the measurement model is the one given in Equation (6.7).

ation variable $\boldsymbol{c}_{1:k}$ can be considered to be an unimportant[8] nuisance parameter. The association indicator can, in general, depend on the past associations. In the most simple form, depicted also in Figure 6.3, the association variable $\boldsymbol{c}_k$ is assumed independent of the past given the current measurement,

$$P(\boldsymbol{c}_k = c \mid \boldsymbol{y}_{1:k} = y_{1:k}, \boldsymbol{c}_{1:k-1} = c_{1:k-1}) = P(\boldsymbol{c}_k = c \mid \boldsymbol{y}_{1:k} = y_{1:k})$$

i.e. $\boldsymbol{c}_k$ is assumed to have only the prior distribution $P(\boldsymbol{c}_k)$.

Often, a set of measurements is assumed to arrive at once, or almost at the same time. Such a set of measurements is referred to as data, scan, frame, or return set. The association of these measurements into targets is often characterised by an *assignment function* $\iota : \{1, \ldots, m\} \to \{0, \ldots, n\}$, which assigns each of the m measurements to one of the n targets, or to false alarm.

$$P(\boldsymbol{y}_k \in A \mid \boldsymbol{x}_k, \boldsymbol{c}_k = \iota) = \prod_{j=1}^{m} P(\boldsymbol{y}_k^{(j)} \in A^{(j)} \mid \boldsymbol{x}_k^{(\iota(j))})$$

where $A = \times_{j=1}^{m} A^{(j)}$, and the notation $P(\boldsymbol{y}_k^{(j)} \in A^{(j)} \mid \boldsymbol{x}_k^{(0)}) = P_f(\boldsymbol{y}_k^{(j)} \in A^{(j)})$ is used for convenience. That is, the measurements are independent given the

8. It is assumed, that one is interested only in the state variable $\boldsymbol{x}_{1:k}$

value of the association function ι, and the multitarget state $\boldsymbol{x}_k$. The association function ι is often constrained in some manner, e.g. such that at most one measurement is allowed to be assigned to one target.

A more general assumption is to consider the case in which the set of measurements is allowed to have a random number of elements [Goodman et al. 1997; Mori et al. 1986]. Although this type of model could be sensible to consider in the case of known number of targets, the description of the random set approach is postponed to Chapter 7. In addition, the constraint that a measurement depends on at most one target can be relaxed. This is the case, when an "association" in the sense of Equation (6.7) is not meaningful[9]. Such a case is not considered here, but an interested reader may find information on that, e.g., in [Stone et al. 1999, Chapter 5].

6.2.3 Filtering

As mentioned above, excluding the case of a finite state-space, the multitarget tracking problem is intractable. This means, that suboptimal methods must be introduced to implement practical systems. The field of multiple target tracking has been under rather intense research for decades. The basic methods for multiple target tracking has been developed in the 70's [Bar-Shalom 1978]. Consequently, to date, a range of algorithms have been proposed for the purpose.

Often, a multiple target tracking algorithm is based on the assumptions of independent target motions and a unique measurement-to-target association, which were covered in Sections 6.2.1 and 6.2.2. If, in addition, the single-target dynamic models and measurement models are assumed linear-Gaussian[10], i.e.

$$\begin{aligned} P(\underline{\boldsymbol{x}}_k^{(i)} \in B \mid \underline{\boldsymbol{x}}_{k-1} = \underline{x}_{k-1}) &= N(B; \mathrm{A}_k \underline{x}_{k-1}^{(i)}, \mathrm{Q}_k) \\ P(\underline{\boldsymbol{y}}_k \in C \mid \underline{\boldsymbol{x}}_k = \underline{x}_k, \boldsymbol{c}_k = i) &= N(C; \mathrm{H}_k \underline{x}_k^{(i)}, \mathrm{R}_k) \end{aligned} \tag{6.8}$$

then the problem is rather similar to the case of SKFM discussed in Section 6.1.3. In principle, one can compute the exact posterior distribution, but the computational complexity increases exponentially with respect to k (the number of Bayes recursions). The framework of computing the posterior distribution by means of exploring all the association possibilities, is referred here to as multiple hypothesis tracking (MHT). The MHT framework has been formulated theoretically in the 70's by Reid [1979], and has been reformulated again, e.g., by Stone et al. [1999].

Since the theoretical MHT framework cannot be used in practice, suboptimal MHT *algorithms* have been proposed. The core for many modern MHT algorithms, e.g. Blackman [2004], is the original algorithm proposed of Reid [1979]. Stone et al. [1999] suggest an extension of MHT, "unified tracking", allowing merged measurements[11]. The association problem has been formulated also as an optimisation problem, in which the most likely[12] association is found for the last s

9. For example, the measured signal is a superposition of two or more signals.
10. Or almost linear-Gaussian, in which case EKF can be employed.
11. Merged measurements are such measurements that depend on more than one target.
12. In general, any other cost function can be used as well.

processed measurements. That is, one tries to find the most likely value for $\boldsymbol{c}_{k-s:k}$ for some fixed s [e.g. Deb et al. 1997; Pattipati et al. 1992].

The joint probabilistic data association (JPDA) framework provides a somewhat different philosophy for the data association problem [Bar-Shalom and Fortmann 1988; Fortmann et al. 1983]. In JPDA, no explicit "data-association hypotheses" are formed, but the posterior distribution is computed approximately. The core idea of the JPDA approach is similar to the idea behind the IMM algorithm. In JPDA, it is assumed, that all the single target posterior distributions in the recursion $k-1$ are Gaussian. In the recursion k, one performs the KF update for each possible association, and then replaces the obtained mixture of Gaussians distribution with a single Gaussian. That is, the posterior distribution of each target is always approximated by a Gaussian distribution. Somewhat similar approach to JPDA is the "event averaged maximum likelihood estimator" (EAMLE) that was proposed by Kastella [1995]. The EAMLE differs from JPDA in that inter-target correlation is modelled. In addition Kastella [1995] proposes an approximative approach, based on mean-field theory, that provides reduced computational complexity.

Of course, sequential Monte Carlo approach has been proposed for the multiple target tracking problem. Indeed, the multiple-target problem is quite natural to be solved with SMC, since it is inherently non-linear and non-Gaussian inference problem. In addition, if the target dynamic model, or the measurement model are very nonlinear or non-Gaussian, the traditional KF based methods cannot be used. Avitzour [1995] proposed the basic bootstrap filter to be used in multitarget tracking. Gordon [1997] considers a hybrid bootstrap filter for multiple target tracking, in which the resampling step is replaced by a step that fits a finite mixture distribution[13] to the posterior samples, and draws samples from the mixture density. Hue et al. [2000, 2002] consider data-association as a missing data problem, and use Gibbs sampling in order to obtain samples of the unknown association variable. Karlsson and Gustafsson [2001] and Schulz et al. [2001] propose a Monte Carlo JPDA approach, which combines the bootstrap filter and the JPDA association method. Särkkä et al. [2004] propose Rao-Blackwellised Monte Carlo data association (RBMCDA), which works in the restricted case of conditionally linear-Gaussian model given in Equation (6.8). The RBMCDA method can be characterised as a "stochastic MHT" implementation, since it forms and maintains hypotheses like MHT, but performs hypothesis creation and pruning in a stochastic manner.

6.3 Tracking Unknown Number of Targets

Often, in real-life applications, there are unknown number of targets to be tracked. In addition, the number of targets can change from time instant to another. In such a case, the assumption of known number of targets, that was made in

13. In the article, mixture of Gaussians were used.

Section 6.2, turns out to be invalid. Consequently, the framework presented in Section 6.2 turns out to be insufficient. Basically, there are two approaches that can be used to enable unknown and random change in the number of targets. The first is to develop directly an *algorithm* that somehow allows the number to change. The second is to formulate the estimation problem differently, so that the number of targets is a dynamic and unknown random variable. After a rigorous probabilistic formulation, an algorithm that computes the (approximate) posterior distribution of the multitarget state can be developed.

Typically, the "algorithmic" approach is based on some kind of a heuristic that makes decisions on the number of targets. The multitarget tracking system developed in Section 6.2 can then be used conditioned on the number of targets determined by the heuristic. The task of finding new targets is referred to as a track initiation problem. Of course, in addition to initiation of tracks, there is certainly a need for a heuristic that decides when a target disappears from the surveillance region. Such a heuristic is substantially easier to implement, since all that is required is the decision that certain target disappears.

Hu et al. [1997] analyse some approaches for track initiation. The methods that have been developed for initiation purposes have been rather application-dependent. For example, the approaches based on the (modified) Hough transform require that the state space is a subset of $\mathbb{R}^2$, and that the measurements are points in the plane [Hu et al. 1997]. Similarly, the rule-based and the logic-based approaches are developed specifically for a certain type of sensor and target model. For example, Yeddanapudi et al. [1995] introduce a track initiation method for ballistic missiles. Lately, an initiation approach based on minimum description length (MDL) has been proposed [Chen et al. 2003]. Unfortunately, even that approach needs to be tailored to a specific problem[14]. As a conclusion, one can state that the initiation methods that have been developed are not too general. The problem-specific approaches may well be suitable for the application for which they were developed, but they cannot be considered suitable for a heterogeneous multi-sensor systems, with a variety of sensor types.

Alternative approaches, that are based on a development of a rigorous theory for changing number of targets, have been developed by several authors. True multitarget estimation with unknown number of targets is joint estimation of target count, individual target states, and measurement-to-target association[15]. The random set approach, which is the method selected in this thesis, is covered in Chapter 7. Mahler [2003a] gives a rather extensive summary of the alternative approaches to the multitarget problem, and references to the approaches. A short summary and references to selected frameworks is given next.

The MHT approach of Reid [1979] includes also the case of changing

14. The application that is considered in the article is "detection and initiation of tracks of incoming tactical ballistic missiles in the exo-atmospheric phase using a surface-based electronically scanned array (ESA) radar".

15. The association is often considered as a nuisance parameter, and the objective is to estimate the *multitarget state*, i.e. target count and states.

number of targets. However, the article does not provide a probabilistic model that would account also for the change in the number of targets. The article of Mori et al. [1986] presents a model that closely resembles the random set approach, but instead of random set formalism, the problem is specified in a disjoint union space $X = \bigcup_{n=0}^{\infty} \mathbb{S}^n \times \{n\}$, where n denotes the number of targets. Mori et al. [1986] suggest the inference to be carried out in the spirit of Reid's MHT approach.

The integrated probabilistic data association (IPDA) approach considers the case of zero or one targets [Mušicki et al. 1994]. The target appearance and disappearance are included to the model by means of a two-state Markov chain. The idea of the IPDA approach is similar to the one proposed in [Stone et al. 1999, p. 181] and in [Valkonen 2002, Chapter 5]. There, the target appearance and disappearance is achieved by an additional element, say "$\oslash$", in the state space. The state $\oslash$ of a target means that the target is not present. Recently, an extension of IPDA, the joint IPDA (JIPDA) approach has been proposed [Mušicki and Evans 2002]. JIPDA considers the case of multiple targets, with the same basic idea. In both IPDA and JIPDA, the posterior distribution of each target is approximated by a Gaussian distribution, like in PDA and JPDA, respectively. Challa et al. [2002] have found that the IPDA algorithm can be derived from the random set formalism.

Chapter 7

Random Set Tracking

In this thesis, construction of a model for tracking unknown number of targets is carried out in the random set framework, whose theoretical background was covered in Chapter 5. The multiple target tracking problem can be modelled in a very flexible and natural fashion with random sets. The unknown estimated multitarget state $\boldsymbol{X}_k$ is a random set that consists of unknown and varying number of targets. Similarly, for many types of sensors, the number of measurements that are obtained at one time instant[1] is varying, and can be considered random. The model that is derived in this chapter assumes numerous independencies, and a rather specific sensor model.

Since this chapter contains some specific derivations, it is suggested that a reader with no prior knowledge of finite random sets in target tracking looks through some introduction. There are many sources, where one can find tutorial material. Perhaps the most thorough survey is given in the book of Goodman et al. [1997]. More engineering-oriented introductions can be found, e.g., in the following references [Mahler 2000, 2001a,b, 2004].

This chapter is ordered as follows. First, a simple dynamic model is developed in Section 7.1. Section 7.2 covers development of measurement models for three different types of sensors. Section 7.3 discusses of practical filtering problems that arise, and previously proposed computational strategies tackling the problems. Section 7.3.3 proposes a SMC implementation for the model that was developed in Sections 7.1 and 7.2. Finally, the visualisation of the multitarget posterior distribution is considered in Section 7.4.

7.1 Dynamic Model

Conceptually, the dynamic model for unknown and varying number of targets is constructed by defining the conditional probabilities $P(\boldsymbol{X}_k \mid \boldsymbol{X}_{k-1})$, i.e. the Markov transition kernels for $(\boldsymbol{X}_k)_{k\in\mathbb{N}}$. These conditional probabilities contain a number of *submodels*. The submodels need not be independent, but the submodels are listed separately below as a reminder of what is needed to construct a sensible multitarget dynamic model.

1. Or during one period of time

1. A *single-target* dynamic model, which is discussed in Section 6.1.1.
2. A *birth model*, which characterises how new targets can appear in the surveillance region.
3. A *death model*, which characterises how the existing targets can disappear from the surveillance region.

In practice, the target motions are usually assumed independent, as given in Equation (6.6). Consider, in addition, that new target appearance and old target disappearance are independent both with respect to each other and with respect to target movements[2]. Then one can obtain the model that is introduced in this section.

Since the births are assumed independent, a dynamic model can be constructed in parts, as follows.

$$\boldsymbol{X}_k = \boldsymbol{S}_k \cup \boldsymbol{B}_k \tag{7.1}$$

where $\boldsymbol{B}_k$ is the set of the appeared new targets, and $\boldsymbol{S}_k$ is the set of the targets survived from the previous round $k-1$. The dynamic models of sets $\boldsymbol{S}_k$ and $\boldsymbol{B}_k$ are considered next in Sections 7.1.1 and 7.1.2, respectively. The density of the dynamic model $f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}$ is given in Section 7.1.3.

7.1.1 Individual Movements and Deaths

This section is devoted to construction of the dynamic model $P(\boldsymbol{S}_k \mid \boldsymbol{X}_{k-1})$ of the survived targets. As mentioned in Section 5.3.4, this can be achieved by defining a belief measure $\beta_{\boldsymbol{S}_k}(C \mid X_{k-1}) = P(\boldsymbol{S}_k \subset C \mid \boldsymbol{X}_{k-1} = X_{k-1})$. Suppose that there are n targets on the set $X_{k-1} = \{x_{k-1}^{(1)}, \ldots, x_{k-1}^{(n)}\}$. Then the set of survived targets can be represented as follows.

$$\boldsymbol{S}_k = \bigcup_{i=1}^{n} \boldsymbol{X}_k^{(i)} \qquad \text{where} \qquad \boldsymbol{X}_k^{(i)} = \{\boldsymbol{x}_k^{(i)}\} \cap \boldsymbol{D}_k^{(i)} \tag{7.2}$$

where the old-target sets $\boldsymbol{X}_k^{(i)}$ are either empty sets, or singletons. The random sets $\boldsymbol{D}_k^{(i)}$ characterise target disappearance, and can be given as follows[3].

$$\boldsymbol{D}_k^{(i)} = \begin{cases} \mathbb{S}, & \text{with probability } p_s(x_{k-1}^{(i)}) \\ \emptyset, & \text{with probability } 1 - p_s(x_{k-1}^{(i)}) \end{cases} \tag{7.3}$$

where $p_s : \mathbb{S} \to [0,1]$ characterises the probability of target survival in a given position of the state space.

---

2. Mahler [2000] suggests a *spawn model*, that accounts for targets spawning from the existing targets. This means, that the birth model includes a dependency on the previous state. Spawning is not considered in this thesis. However, such dependencies can be included in the tracking model without a theoretical difficulty.
3. $\boldsymbol{D}_k^{(i)}$ is a random closed set, but not finite! Since $\{\boldsymbol{x}_k^{(i)}\}$ are finite, $\boldsymbol{X}_k^{(i)}$ is finite, though.

Since $\boldsymbol{D}_k^{(i)}$ and $\{\boldsymbol{x}_k^{(i)}\}$ were assumed independent, the conditional belief measure of $\boldsymbol{X}_k^{(i)}$ can be given in the form of Example 5.18.

$$\begin{aligned}\beta_{\boldsymbol{X}_k^{(i)}}(C \mid X_{k-1}) &= P(\{\boldsymbol{x}_k^{(i)}\} \cap \boldsymbol{D}_k^{(i)} \subset C \mid \boldsymbol{x}_{k-1}^{(i)} = x_{k-1}) \\ &= P(\boldsymbol{D}_k^{(i)} = \emptyset \mid \boldsymbol{x}_{k-1}^{(i)} = x_{k-1}) + P(\boldsymbol{D}_k^{(i)} = \mathbb{S}, \{\boldsymbol{x}_k^{(i)}\} \subset C \mid \boldsymbol{x}_{k-1}^{(i)} = x_{k-1}) \\ &= [1 - p_s(x_{k-1})] + p_s(x_{k-1}) P(\boldsymbol{x}_k^{(i)} \in C \mid \boldsymbol{x}_{k-1}^{(i)} = x_{k-1})\end{aligned}$$

where $1 - p_s(x_{k-1})$ is the probability that a target in position x_{k-1} will disappear, and $P(\boldsymbol{x}_k^{(i)} \in C \mid \boldsymbol{x}_{k-1}^{(i)} = x_{k-1})$ is a single-target dynamic model, discussed in Section 6.1.1. If it is reasonable to assume, that target disappearance does not depend on the position x_{k-1}, then p_s is constant.

It is rather trivial[4] to obtain the belief density corresponding to the belief measure $\beta_{\boldsymbol{X}_k^{(i)}}(C \mid X_{k-1})$.

$$f_{\boldsymbol{X}_k^{(i)}}(Z) = \begin{cases} 1 - p_s(x_{k-1}^{(i)}), & Z = \emptyset \\ p_s(x_{k-1}^{(i)}) f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(z \mid x_{k-1}^{(i)}), & Z = \{z\} \\ 0, & |Z| > 1 \end{cases} \tag{7.4}$$

where $f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}$ is the density of the single-target dynamic model $P(\boldsymbol{x}_k^{(i)} \mid \boldsymbol{x}_{k-1}^{(i)})$. Then, one can ask how to obtain the belief density corresponding $\beta_{\boldsymbol{S}_k}$. It is obvious, that $\beta_{\boldsymbol{S}_k} = \prod_{i=1}^n \beta_{\boldsymbol{X}_k^{(i)}}$, since $\boldsymbol{X}_k^{(i)}$ were assumed independent, and thus one obtains

$$\beta_{\boldsymbol{S}_k}(C \mid X) = P\left(\bigcup_{i=1}^n \boldsymbol{X}_k^{(i)} \subset C \;\middle|\; \boldsymbol{X}_{k-1} = X \right) = \prod_{i=1}^n P(\boldsymbol{X}_k^{(i)} \subset C \mid \boldsymbol{X}_{k-1} = X)$$

$$\prod_{i=1}^n P(\boldsymbol{X}_k^{(i)} \subset C \mid \boldsymbol{x}_{k-1}^{(i)} = x^{(i)}) = \prod_{i=1}^n \beta_{\boldsymbol{X}_k^{(i)}}(C \mid x^{(i)})$$

The following proposition provides means for obtaining the belief density of a product of belief measures.

Proposition 7.1 *Let $\beta_1, \ldots, \beta_n$ be belief functions of "singleton-or-empty" random sets. That is, $\delta\beta_i / \delta Z = 0$, whenever $|Z| > 1$. Then, the following holds.*

$$\frac{\delta}{\delta\{z_1, \ldots, z_m\}} \prod_{i=1}^n \beta_i = \begin{cases} \dfrac{1}{(n-m)!} \displaystyle\sum_{\pi} \left[\prod_{i=1}^m \frac{\delta \beta_{\pi i}}{\delta z_i} \prod_{i=m+1}^n \beta_{\pi i} \right], & m \leq n \\ 0, & m > n \end{cases}$$

where π goes over every permutation of the integers $1, \ldots, n$.

4. Most importantly, the power rule in Equation (5.21) is needed.

Proof. Let us first convince ourselves of the case $m > n$. First, observe that if $n = 1$, then the condition holds by assumption. Next, suppose that the condition holds for $n - 1$. Then, by the product rule in Equation (5.17),

$$\begin{aligned}\frac{\delta}{\delta Z}\prod_{i=1}^{n}\beta_i &= \sum_{W\subset Z}\left[\overbrace{\frac{\delta\beta_n}{\delta W}}^{|W|\leq 1}\frac{\delta}{\delta(Z\setminus W)}\prod_{i=1}^{n-1}\beta_i\right] \\ &= \overbrace{\beta_n\frac{\delta}{\delta Z}\prod_{i=1}^{n-1}\beta_i}^{|Z|\leq n-1} + \sum_{z\in Z}\left[\frac{\delta\beta_n}{\delta z}\overbrace{\frac{\delta}{\delta(Z\setminus\{z\})}\prod_{i=1}^{n-1}\beta_i}^{|Z|\leq n}\right]\end{aligned}$$

where the conditions above the terms indicate when the term may be nonzero.

Suppose next that $m \leq n$. The case $m = 0$ is trivial. Assume that the proposition holds for $m - 1$. Then, one obtains

$$\begin{aligned}\frac{\delta}{\delta\{z_1,\ldots,z_m\}}\prod_{i=1}^{n}\beta_i &= \frac{\delta}{\delta z_m}\left[\frac{\delta}{\delta\{z_1,\ldots,z_{m-1}\}}\prod_{i=1}^{n}\beta_i\right] \\ &= \frac{1}{(n-m+1)!}\sum_{\pi}\left[\overbrace{\frac{\delta}{\delta z_m}\left(\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\prod_{i=m}^{n}\beta_{\pi i}\right)}^{(*)}\right]\end{aligned}$$

Now, observe that $(*)$ decomposes as follows according to the product rule in Equation (5.17), and Proposition 5.23, respectively.

$$\begin{aligned}\frac{\delta}{\delta z_m}\left[\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\prod_{i=m}^{n}\beta_{\pi i}\right] &= \prod_{i=m}^{n}\beta_{\pi i}\frac{\delta}{\delta z_m}\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i} + \prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\frac{\delta}{\delta z_m}\prod_{i=m}^{n}\beta_{\pi i} \\ &= \prod_{i=m}^{n}\beta_{\pi i}\sum_{i=1}^{m-1}\left[\overbrace{\frac{\delta\beta_{\pi i}}{\delta z_i\delta z_m}}^{=0}\prod_{\substack{j=1\\ j\neq i}}^{m-1}\beta_{\pi j}\right] + \overbrace{\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\sum_{i=m}^{n}\left[\frac{\delta\beta_{\pi i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi j}\right]}^{(**)}\end{aligned}$$

Applying $(**)$ back into $(*)$, one gets

$$\frac{\delta}{\delta\{z_1,\ldots,z_m\}}\prod_{i=1}^{n}\beta_i = \frac{1}{(n-m+1)!}\overbrace{\sum_{\pi}\left[\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\sum_{i=m}^{n}\left(\frac{\delta\beta_{\pi i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi j}\right)\right]}^{(\dagger)} \tag{7.5}$$

where the sum over permutations (†) can be written in parts, so that one gets

$$\sum_{\pi}\left[\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi i}}{\delta z_i}\sum_{i=m}^{n}\left(\frac{\delta\beta_{\pi i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi j}\right)\right]=\sum_{\pi^*}\sum_{\pi_*}\left[\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi^* i}}{\delta z_i}\sum_{i=m}^{n}\left(\frac{\delta\beta_{\pi_* i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi_* j}\right)\right]$$

$$=\sum_{\pi^*}\left[\prod_{i=1}^{m-1}\frac{\delta\beta_{\pi^* i}}{\delta z_i}\sum_{i=m}^{n}\overbrace{\sum_{\pi_*}\left(\frac{\delta\beta_{\pi_* i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi j}\right)}^{(\ddagger)}\right] \tag{7.6}$$

where the permutations π are split into two parts, as follows

$$\pi i=\begin{cases}\pi^* i, & 1\leq i<m\\ \pi_* i, & m\leq i\leq n\end{cases}$$

Consequently, the sum over permutations π is split into the two sums over π^* and π_*. The map π^* goes through all one-to-one maps $\pi^*:\{1,\ldots,m-1\}\to\{1,\ldots,n\}$. For each π^*, the map π_* goes through all one-to-one maps $\pi_*:\{m,\ldots,n\}\to\{1,\ldots,n\}\setminus\mathrm{rng}(\pi^*)$.

Since the sum in (‡) covers all one-to-one maps of numbers $m,\ldots,n$ to $\{1,\ldots,n\}\setminus\mathrm{rng}(\pi^*)$, it is independent of i, as long as $m\leq i\leq n$. Thus, one can write

$$\sum_{i=m}^{n}\sum_{\pi_*}\left(\frac{\delta\beta_{\pi_* i}}{\delta z_m}\prod_{\substack{j=m\\ j\neq i}}^{n}\beta_{\pi j}\right)=(n-m+1)\sum_{\pi_*}\left(\frac{\delta\beta_{\pi_* m}}{\delta z_m}\prod_{j=m+1}^{n}\beta_{\pi j}\right) \tag{7.7}$$

Combining Equations (7.5)–(7.7), one gets

$$\frac{\delta}{\delta\{z_1,\ldots,z_m\}}\prod_{i=1}^{n}\beta_i=\frac{1}{(n-m)!}\sum_{\pi}\left[\prod_{i=1}^{m}\frac{\delta\beta_{\pi i}}{\delta z_i}\prod_{i=m+1}^{n}\beta_{\pi i}\right] \tag{7.8}$$

which is the proposed form. □

According to Proposition 7.1, the belief density of $\boldsymbol{S}_k$ conditioned on $\boldsymbol{X}_{k-1}$ can now be given as follows.

$$f_{\boldsymbol{S}_k|\boldsymbol{X}_k}(Z\mid X)=\begin{cases}\dfrac{1}{(|X|-|Z|)!}\displaystyle\sum_{\pi}\left[\prod_{i=1}^{|Z|}\Big(p_s(x_{\pi i})f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(z_i\mid x_{\pi i})\Big)\right.\\ \qquad\qquad\left.\displaystyle\prod_{i=|Z|+1}^{|X|}\Big(1-p_s(x_{\pi i})\Big)\right], & |Z|\leq|X|\\ 0, & |Z|>|X|\end{cases} \tag{7.9}$$

where π goes over every permutation of integers $1,\ldots,|X|$.

7.1.2 Poisson Birth

The set of new targets that have appeared can have, in general, any distribution. A rather uninformative model can be constructed so that one assumes that targets appear independently in different regions of the state space. That is, given that k targets are born, the elements $\boldsymbol{b}_k^{(i)}$ of the set $\boldsymbol{B}_k$ are independent. This kind of model can be constructed according to Theorem 5.17 as follows

$$\beta_{\boldsymbol{B}_k}(C \mid X_{k-1}) = \beta_{\boldsymbol{B}_k}(C) = \sum_{j=0}^{\infty} b_j P_b(C)^j \tag{7.10}$$

where $P_b(C) = P(\boldsymbol{b}_k^{(i)} \in C)$ is a probability measure for single target appearance. The terms $b_j = P(|\boldsymbol{B}_k| = j)$ denote the probability distribution that j targets are born. In this thesis, it is assumed that the birth events form a Poisson process. That is, b_j follow the Poisson distribution,

$$b_j = \frac{(\eta\tau_k)^j}{j!} e^{-\eta\tau_k}, \quad j \in \mathbb{N} \tag{7.11}$$

where $\tau_k = t_k - t_{k-1}$ is the time difference between $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k-1}$, and η is the birth intensity parameter, that describes the expectation of the number of born targets per time unit.

The density of the belief measure given in Equation (7.10) can be written as follows, according to the power rule in Equation (5.21).

$$\frac{\delta\beta_{\boldsymbol{B}_k}}{\delta Z}(\emptyset) = \sum_{j=0}^{\infty} b_j \frac{\delta}{\delta Z} P_b(\emptyset)^j = \sum_{j=m}^{\infty} b_j \frac{j!}{(j-m)!} P_b(\emptyset)^{j-m} \prod_{i=1}^{m} f_b(z_i)$$

where $Z = \{z_1, \ldots, z_m\}$. But $P_b(\emptyset) = 0$, and thus $P_b(\emptyset)^{j-m} = 0$, whenever $m \neq j$, so

$$f_{\boldsymbol{B}_k}(Z) = \frac{\delta\beta_{\boldsymbol{B}_k}}{\delta Z}(\emptyset) = |Z|! \cdot b_{|Z|} \prod_{z \in Z} f_b(z) \tag{7.12}$$

where f_b is the density of the single-target birth measure P_b, product over empty set is defined one, and $0! = 1$.

7.1.3 Individual Movements, Births, Deaths, and Poisson Birth

Due to the independence of individual motion models, disappearance, and appearance of new targets, one can write the conditional belief as follows.

$$\begin{aligned}\beta_{\boldsymbol{X}_k}(C \mid X_{k-1}) &= P\left(\boldsymbol{S}_k \cup \boldsymbol{B}_k \subset C \mid \boldsymbol{X}_{k-1} = X_{k-1}\right) \\ &= P\left(\boldsymbol{B}_k \subset C\right) P\left(\boldsymbol{S}_k \subset C \mid \boldsymbol{X}_{k-1} = X_{k-1}\right)) = \beta_{\boldsymbol{B}_k}(C)\beta_{\boldsymbol{S}_k}(C \mid X_{k-1})\end{aligned}$$

So, the belief density $f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(Z \mid X_{k-1})$ can be obtained according to the product rule in Equation (5.17) from the belief densities of $\boldsymbol{B}_k$ and $\boldsymbol{S}_k$ as follows.

$$
\begin{aligned}
f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(Z \mid X) &= \sum_{W \subset Z} \left[f_{\boldsymbol{S}_k|\boldsymbol{X}_{k-1}}(W \mid X) f_{\boldsymbol{B}_k}(Z \setminus W) \right] \\
&= \sum_{\substack{W \subset Z \\ |W| \leq |X|}} \left[\frac{|Z \setminus W|! \cdot b_{|Z \setminus W|}}{(|X| - |W|)!} \prod_{u \in Z \setminus W} f_b(u) \right. \\
&\qquad \left. \sum_{\pi} \left(\prod_{i=1}^{|W|} \left(p_s(x_{\pi i}) f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(w_i \mid x_{\pi i}) \right) \prod_{i=|W|+1}^{n} \left(1 - p_s(x_{\pi i}) \right) \right) \right]
\end{aligned} \tag{7.13}
$$

where π goes over every permutation of integers $1, \ldots, |X|$.

7.2 Measurement Model

In the case of unknown number of targets, the measurement model has quite an important role. In addition to giving information on the position of a target, the measurements give information of the number of targets in the surveillance region. Conceptually, the construction of the multitarget measurement model reduces to definition of the conditional probability $P(\boldsymbol{Y}_k \mid \boldsymbol{X}_k)$. In contrast to Goodman et al. [1997], the whole sensor suite is not reconceptualised as one "meta-sensor" in this thesis. Instead, it is assumed that the sensors operate rather independently, and their measurements are processed separately. In addition, the reports from different sensors are allowed to arrive irregularly, even without any prior knowledge of sensors' reporting frequencies.

This section covers specifications of three different measurement models. The categorisation of sensors is near to the categorisation into "type 1" and "type 2" sensors due to Reid [1979]. The first sensor type, covered in Section 7.2.1, is a sensor that produces a varying number of measurements in each report. A practical example of this type of sensor can be considered to be a microwave radar, that reports the detections obtained during one scan that covers the surveillance region. This sensor type admits meaningful concepts of a "measurement attempt", and a "probability of detection". The second type of sensor, discussed in Section 7.2.2, is one that produces reports with exactly one detection. Section 7.2.3 presents the third type of sensor, that produces at most one detection in a report. Such a sensor is considered to be a kind of "mixture" of the first and second types of sensors.

7.2.1 Varying Number of Measurements

Let us first consider a measurement model that allows the number of measurements to change. If the target measurements, the missed measurements, and the false alarms are considered independent, the model reduces analogous to the random set dynamic model given in Equation (7.1). That is, the measurement set

$\boldsymbol{Y}_k$ can be characterised by the model

$$\boldsymbol{Y}_k = \boldsymbol{T}_k \cup \boldsymbol{F}_k \tag{7.14}$$

where $\boldsymbol{T}_k = \bigcup_{i=1}^{n} \boldsymbol{Y}_k^{(i)}$ are the measurements obtained from the n targets, and $\boldsymbol{F}_k$ are the false measurements. If it is assumed, that there is at most one measurement of each target, the random sets $\boldsymbol{Y}_k^{(i)}$ are either singleton or empty, similar to $\boldsymbol{X}_k^{(i)}$ given in Equation (7.2),

$$\boldsymbol{T}_k = \bigcup_{i=1}^{n} \boldsymbol{Y}_k^{(i)} \qquad \text{where} \qquad \boldsymbol{Y}_k^{(i)} = \{\boldsymbol{y}_k^{(i)}\} \cap \boldsymbol{M}_k^{(i)} \tag{7.15}$$

where $\boldsymbol{M}_k^{(i)}$ are random sets that characterise the target measurability similar to Equation (7.3),

$$\boldsymbol{M}_k^{(i)} = \begin{cases} \mathbb{S}, & \text{with probability } p_d(x_{k-1}^{(i)}) \\ \emptyset, & \text{with probability } 1 - p_d(x_{k-1}^{(i)}) \end{cases}$$

where $p_d : \mathbb{S} \to [0, 1]$ determines the probability of detection in each position of the state space. If the detection and the measurement processes ($\boldsymbol{M}_k^{(i)}$ and $\boldsymbol{y}_k^{(i)}$, respectively) are assumed independent, the belief function can be given as follows.

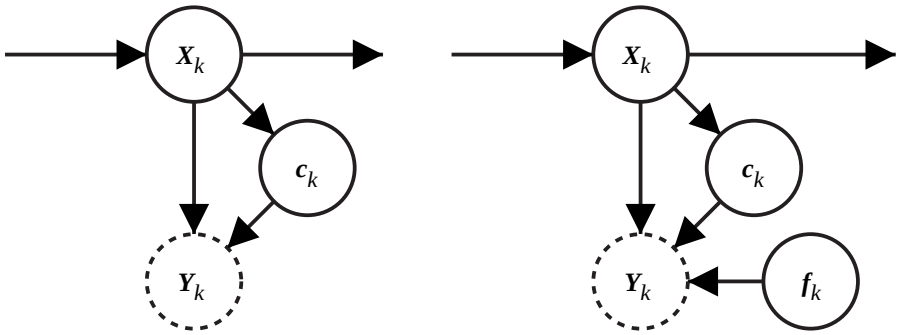
$$\beta_{\boldsymbol{Y}_k^{(i)}}(C \mid X_k) = \left[1 - p_d(x_k^{(i)})\right] + p_d(x_k^{(i)}) P(\boldsymbol{y}_k^{(i)} \in C \mid \boldsymbol{x}_k^{(i)} = x_k^{(i)})$$

where $p_d(x)$ is the *probability of detection* of a target in state space point $x \in \mathbb{S}$. The probability of detection model includes, e.g., the sensor field-of-view (FOV), and obviously $p_d(x) = 0$ for such state-space points x where it is impossible for the sensor to obtain measurements of the target.

A Poisson model has been used as a model for false alarms in the literature, e.g. in derivation of the PDA algorithm [Bar-Shalom and Fortmann 1988, p. 168]. In the case of a Poisson false alarm model, derivation of the measurement model is similar to derivation of the dynamic model in Section 7.1. The resulting density $f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(Y \mid X)$ is similar to the dynamic model given in Equation (7.13),

$$\begin{aligned} f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(Y \mid X) &= \sum_{W \subset Y} \left[f_{\boldsymbol{T}_k|\boldsymbol{X}_k}(W \mid X) f_{\boldsymbol{F}_k}(Y \setminus W) \right] \\ &= \sum_{\substack{W \subset Y \\ |W| \leq |X|}} \left[\frac{|Y \setminus W|! \cdot d_{|Y \setminus W|}}{(|X| - |W|)!} \prod_{u \in Y \setminus W} f_f(u) \right. \\ &\qquad \left. \sum_{\pi} \left(\prod_{i=1}^{|W|} \left(p_d(x_{\pi i}) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(w_i \mid x_{\pi i}) \right) \prod_{i=|W|+1}^{n} \left(1 - p_d(x_{\pi i}) \right) \right) \right] \end{aligned} \tag{7.16}$$

where π goes over all the permutations of the integers $1, \ldots, |X|$, p_d is the probability of detection, $f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x)$ is the single-target measurement model density,



(a) (b)

Figure 7.1: A Bayesian network representation of a random set tracking model with an explicit association indicator. (a) The model in the single-target case. (b) The model in the case of zero or one measurements, with constant false alarm rate.

and f_f is the false alarm density. The constants d_j correspond to the Poisson density of the false alarms, which can be given as follows

$$d_j = \frac{(\gamma\tau_k)^j}{j!}e^{-\gamma\tau_k}, \quad j \in \mathbb{N}$$

where γ is the false alarm intensity, i.e. the expected number of false alarms per time unit, and τ_k is the time difference from the previous report of the current sensor.

7.2.2 Exactly One Measurement

In the case of a sensor that produces exactly one measurement per report, the cardinality of the set $\boldsymbol{Y}_k$ must be considered fixed. That is, it is assumed that the measurement is a singleton $\boldsymbol{Y}_k = \{\boldsymbol{y}_k\}$. This means, that the cardinality distribution is assumed to be

$$P\left(|\boldsymbol{Y}_k| = n \;\middle|\; \boldsymbol{X}_k = X\right) = P(|\boldsymbol{Y}_k| = n) = \begin{cases} 1, & n = 1 \\ 0, & n \neq 1 \end{cases}$$

It is assumed, that the measurement can either be false alarm, or originate equally likely from any of the targets. This kind of model can be formulated using an association indicator introduced in Section 6.2.2. The measurement model can be presented more intuitively as a Bayesian network such as given in Figure 7.1 (a). The distribution of the association indicator $\boldsymbol{c}_k$ depends only on the cardinality of $\boldsymbol{X}_k$,

$$P(\boldsymbol{c}_k = i \mid \boldsymbol{X}_k) = P\left(\boldsymbol{c}_k = i \;\middle|\; |\boldsymbol{X}_k|\right)$$

The model can be given either by defining a fixed false alarm percentage in the presence of targets,

$$P\left(\boldsymbol{c}_k = i \;\middle|\; |\boldsymbol{X}_k| = n\right) = \begin{cases} p_f, & i = 0 \\ n^{-1}(1 - p_f), & 1 \leq i \leq n \\ 0, & i > n \end{cases} \tag{7.17}$$

or by defining just a uniform distribution from 0 to n,

$$P\left(\boldsymbol{c}_k = i \;\middle|\; |\boldsymbol{X}_k| = n\right) = \begin{cases} (n+1)^{-1} & 0 \leq i \leq n \\ 0, & i > n \end{cases} \tag{7.18}$$

Now, suppose that the elements in X are enumerated arbitrarily, $X = \{x^{(1)}, \ldots, x^{(n)}\}$. The measurement model can be given as follows similar to the model given in Equation (6.7).

$$P(\{\boldsymbol{y}_k\} \subset C \mid \boldsymbol{X}_k = X, \boldsymbol{c}_k = i) = \begin{cases} P(\boldsymbol{y}_k \in C \mid x^{(i)}), & i > 0 \\ P_f(C), & i = 0 \end{cases}$$

where P_f is the probability measure of the false alarms. Since the association variable $\boldsymbol{c}_k$ is unknown, and considered unimportant, we need to obtain the measurement model without $\boldsymbol{c}_k$.

$$\begin{aligned} &\beta_{\{\boldsymbol{y}_k\}}(C \mid \boldsymbol{X}_k = X) = \sum_{i=0}^{\infty} P(\boldsymbol{y}_k \in C, \boldsymbol{c}_k = i \mid \boldsymbol{X}_k = X) \\ &= \sum_{i=0}^{\infty} \left[P(\boldsymbol{y}_k \in C \mid \boldsymbol{X}_k = X, \boldsymbol{c}_k = i) P(\boldsymbol{c}_k = i \mid \boldsymbol{X}_k = X) \right] \\ &= P\left(\boldsymbol{c}_k = 0 \;\middle|\; |\boldsymbol{X}_k| = n\right) P_f(C) + \sum_{i=1}^{n} \left[P(\boldsymbol{y}_k \in C \mid x^{(i)}) P\left(\boldsymbol{c} = i \;\middle|\; |\boldsymbol{X}_k| = n\right) \right] \end{aligned}$$

Derivation of the corresponding belief densities is rather straightforward, due to the linearity of the set derivative. The density corresponding to the uniform association distribution in Equation (7.18) can be given as follows

$$f_{\boldsymbol{y}_k|\boldsymbol{X}_k}(y \mid X) = \frac{1}{|X| + 1} \left[f_f(y) + \sum_{x \in X} f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x) \right]$$

where f_f is the density corresponding P_f, the false alarm distribution. In the case of a constant false alarm rate, i.e. when the association variable has the distribution given in Equation (7.17), the density can be given as follows

$$f_{\boldsymbol{y}_k|\boldsymbol{X}_k}(y \mid X) = \begin{cases} p_f \cdot f_f(y) + \frac{1 - p_f}{|X|} \sum_{x \in X} f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x), & X \neq \emptyset \\ f_f(y), & X = \emptyset \end{cases}$$

7.2.3 Zero or One Measurement

Suppose that there is a sensor, which includes either one measurement, or no measurements in each report. This section introduces a model for this type of sensor. The sensor model includes independent false alarms, that occur with probability p_f. In addition, the targets are assumed to be measured equiprobably. However, when a target is measured, it may occur that the sensor is not able to produce a measurement, thus there is also a model for probability of detection. This measurement model can be factored into a Bayesian network shown in Figure 7.1 (b). The binary random variable $\boldsymbol{f}_k$ can take values 0 (no false alarm) or 1 (false alarm). The distribution of $\boldsymbol{f}_k$ can be given as follows.

$$P(\boldsymbol{f}_k = f) = \begin{cases} 1 - p_f, & f = 0 \\ p_f, & f = 1 \end{cases} \tag{7.19}$$

The association variable $\boldsymbol{c}_k$ is integer-valued, and has a distribution that depends only on the current number of targets, i.e. the cardinality of $\boldsymbol{X}_k$.

$$P\left(\boldsymbol{c}_k = i \;\middle|\; |\boldsymbol{X}_k| = n\right) = \begin{cases} 1/n, & 1 \leq i \leq n \\ 0, & \text{otherwise} \end{cases} \tag{7.20}$$

where it is assumed that $n \geq 1$. As we shortly see, the distribution of $\boldsymbol{c}_k$ is unimportant[5], if $n = 0$. The measurement $\boldsymbol{Y}_k$ is assumed to depend on $\boldsymbol{X}_k$, $\boldsymbol{c}_k$, and $\boldsymbol{f}_k$, so that

$$\begin{aligned} \beta_{\boldsymbol{Y}_k}(C \mid \boldsymbol{X}_k, \boldsymbol{c}_k = i, \boldsymbol{f}_k = 1) &= P_f(C) \\ \beta_{\boldsymbol{Y}_k}(C \mid \boldsymbol{X}_k = \emptyset, \boldsymbol{c}_k = i, \boldsymbol{f}_k = 0) &= 1 \\ \beta_{\boldsymbol{Y}_k}(C \mid \boldsymbol{X}_k = X, \boldsymbol{c}_k = i, \boldsymbol{f}_k = 0) &= 1 - p_d(x^{(i)}) + p_d(x^{(i)}) P(\boldsymbol{y} \in C \mid x^{(i)}) \end{aligned} \tag{7.21}$$

Now, the measurement model $P(\boldsymbol{Y}_k \mid \boldsymbol{X}_k)$ can be derived, since one can write

$$\begin{aligned} \beta_{\boldsymbol{Y}_k}(C \mid X) &= \sum_{f=0}^{1} \sum_{i=1}^{\infty} P(\boldsymbol{Y}_k \subset C, \boldsymbol{c}_k = i, \boldsymbol{f}_k = f \mid \boldsymbol{X}_k = X) \\ &= \sum_{f=0}^{1} \sum_{i=1}^{\infty} \left[\beta_{\boldsymbol{Y}_k}(C \mid \boldsymbol{X}_k = X, \boldsymbol{c}_k = i, \boldsymbol{f}_k = f) P(\boldsymbol{c}_k = i, \boldsymbol{f}_k = f \mid \boldsymbol{X}_k = X) \right] \\ &= \sum_{f=0}^{1} \sum_{i=1}^{\infty} \left[\beta_{\boldsymbol{Y}_k}(C \mid \boldsymbol{X}_k = X, \boldsymbol{c}_k = i, \boldsymbol{f}_k = f) P(\boldsymbol{c}_k = i \mid \boldsymbol{X}_k) P(\boldsymbol{f}_k = f) \right] \end{aligned} \tag{7.22}$$

Substituting Equations (7.19)–(7.21) into Equation (7.22), one obtains the following belief measure $\beta_{\boldsymbol{Y}_k}(C \mid X) = P(\boldsymbol{Y}_k \subset C \mid \boldsymbol{X}_k = X)$.

$$\begin{aligned} \beta_{\boldsymbol{Y}_k}(C \mid \emptyset) &= (1 - p_f) + p_f P_f(C) \\ \beta_{\boldsymbol{Y}_k}(C \mid X) &= \frac{1 - p_f}{|X|} \left[\sum_{x \in X} \Big(1 - p_d(x)\Big) + \sum_{x \in X} \Big(p_d(x) P(\boldsymbol{y}_k \in C \mid x)\Big) \right] + p_f P_f(C) \end{aligned}$$

5. One can assume, for example, that $P(\boldsymbol{c}_k = 1 \mid |\boldsymbol{X}_k| = 0) = 1$.

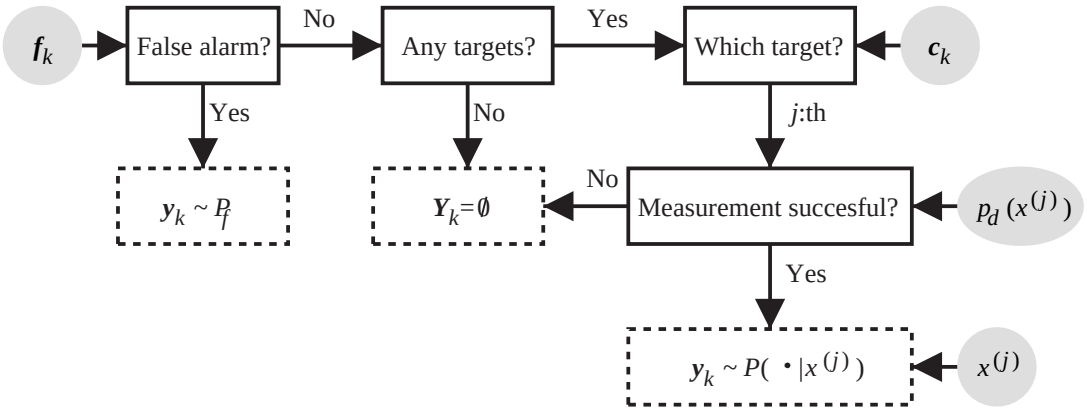


Figure 7.2: The data generation procedure for the model of zero or one measurements. The shaded ellipses denote the "inputs", while the dotted boxes contain the different possible outputs.

where $X \neq \emptyset$. The density corresponding to the belief measure above can be given as follows

$$\begin{aligned} f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(\emptyset \mid \emptyset) &= 1 - p_f \\ f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(\{y\} \mid \emptyset) &= p_f f_f(y) \\ f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(\emptyset \mid X) &= \frac{1-p_f}{|X|} \sum_{x \in X} \Big[1 - p_d(x)\Big] \\ f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(\{y\} \mid X) &= \frac{1-p_f}{|X|} \sum_{x \in X} \Big[p_d(x) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x)\Big] + p_f f_f(y), \end{aligned} \tag{7.23}$$

where X denotes a nonempty set.

The derivation of the density in Equation (7.23) was technical. The model can be considered more intuitively through a data generation procedure, which is outlined in Figure 7.2. The generation of a datum starts from the random variable $\boldsymbol{f}_k$, which determines whether there is a false alarm. If there is a false alarm, the measurement is distributed according to the false alarm distribution P_f. If the measurement is not a false alarm, i.e $\boldsymbol{f}_k = 0$, then the procedure continues with a query of any targets. If there are no targets, the measurement is empty. Otherwise, the i'th target is picked according to $\boldsymbol{c}_k = i$. The following query determines whether the sensor's attempt to obtain a measurement of the i'th target is a success. If the measurement attempt fails, the measurement is again empty. Otherwise, the measurement is generated according to the single-target measurement model $P(\boldsymbol{y}_k \in C \mid \boldsymbol{x}_k = x^{(i)})$.

7.3 Filtering

Before starting this section, one may recall, that tracking multiple known number of targets was in general an intractable estimation problem. Since the case of a

known number of targets is a special case of the random set tracking framework, it is expected that no closed form solutions can be found. On the contrary, since there are more unknown variables that need to be estimated, one can assume, that there will be severe problems in practical implementation of a random set tracking system.

Goodman et al. [1997] presented the theoretical foundations of the Bayesian random set tracking. However, such algorithms that can be considered computationally practical have appeared only recently. This is probably due to the very nonlinear nature of the multitarget tracking problem. The traditional approximative methods, such as linearisation, tend not to suffice. Conceptually, the multitarget tracking problem reduces into recursive Bayesian estimation, discussed in Chapter 3. More specifically, the core problem is to find a tractable approximation of the Bayes recursion, which can be given in the case of random finite set densities as follows

$$p_{k|k-1}(X_k) = \int f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(X_k \mid X_{k-1}) p_{k-1|k-1}(X_{k-1}) \delta X_{k-1} \tag{7.24}$$

$$p_{k|k}(X_k) = \frac{f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(Y_k \mid X_k) p_{k|k-1}(X_k)}{\int f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(Y_k \mid X'_k) p_{k|k-1}(X'_k) \delta X'_k} \tag{7.25}$$

where the notation $p_{a|b}(X) \triangleq f_{\boldsymbol{X}_a|\boldsymbol{Y}_{1:b}}(X \mid Y_{1:b})$ from Chapter 3 is used, and the integral in Equation (7.24) is the set integral, introduced in Section 5.3.3. It is easy to believe, that a closed form solution to Equations (7.24) and (7.25) cannot be obtained in general. Even the pointwise evaluation of $f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}$ and $f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}$ can be computationally heavy, for large cardinality Y_k or X_k. This depends, of course, on the specific choice of the dynamic model and the measurement model.

There are some approaches that have been proposed to bring the recursion into a computationally feasible form. The first approach that is discussed in Section 7.3.1 is the approach based on the probability hypothesis density (PHD). The PHD approach does not provide an explicit algorithm for computing the Bayes recursion, but transfers the problem into a different form. The assumptions that are needed for the PHD approach to be applicable are not satisfied by the sensor models discussed in Sections 7.2.2 and 7.2.3. So, other methods that have been proposed for random set estimation are reviewed in Section 7.3.2. Finally, Section 7.3.3 presents the sequential Monte Carlo algorithm that was developed in this thesis.

7.3.1 The PHD Approach

The concept of PHD was introduced in Section 5.3.7, but merely as a statistic or visualisation that can be obtained from the multitarget density. PHD can be considered also as the "first moment" of the multitarget distribution [Mahler 2003a,b]. Mahler has proposed, that PHD can be used directly in the estimation [see, e.g. Mahler 2003a]. The PHD estimation algorithms are based on the idea that one computes the PHD of the posterior density from the PHD of the previous posterior. Let us denote the PHDs so that $D_{a|b}(x) \triangleq \int p_{a|b}(\{x\} \cup W \mid Y_{1:k}) \delta W$. Then,

the PHD filter can be characterised so that one seeks an algorithm that performs the Bayes update of the PHDs [Mahler 2003a].

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & p_{k-1|k-1} & \xrightarrow{\text{Eq. (7.24)}} & p_{k|k-1} & \xrightarrow{\text{Eq. (7.25)}} & p_{k|k} & \longrightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \cdots & \longrightarrow & D_{k-1|k-1} & \xrightarrow{\text{PHD-predict}} & D_{k|k-1} & \xrightarrow{\text{PHD-update}} & D_{k|k} & \longrightarrow & \cdots \end{array}$$

That is, one needs to develop the "PHD-predict" and the "PHD-update" steps to the above diagram, so that the multitarget densities $p_{a|b}$ need not be considered explicitly.

Mahler [2003a] has derived the approximate update equations for the PHD, given that the following assumptions are satisfied [Mahler 2003a, pp. 1166–1168][6].

1. The motions of the targets are independent, with a transition kernel density $f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}$.
2. The probability of each target's survival is independent of the other targets, and is characterised by the function $p_s : \mathbb{S} \to [0, 1]$.
3. Appearance of new targets is independent of the states of the existent targets. New target appearance has the density $f_{B_k}(X)$. The PHD of this density is denoted as $D'_k(x) \triangleq \int f_{B_k}(\{x\} \cup W)\delta W$.
4. The target measurements produced by the sensor are characterised by the density $f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x)$.
5. The probability of each target's detection is independent, and characterised by the function $p_d : \mathbb{S} \to [0, 1]$.
6. The false alarms are Poisson, with an average number $\gamma\tau_k$ of independent false alarms with a spatial distribution $f_f(y)$.

The assumptions 1–3 are fulfilled by the dynamic model that was presented in Section 7.1, and the assumptions 4–6 are fulfilled by the measurement model in Section 7.2.1. The sensor models introduced in Sections 7.2.2 and 7.2.3 do not satisfy these assumptions.

In the case that the above listed assumptions are valid, the PHD predictor equation can be written as follows [Mahler 2003a, p. 1167].

$$D_{k|k-1}(x) = D'_k(x) + \int p_s(w) f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(x \mid w) D_{k-1|k-1}(w) \mathrm{d}w \tag{7.26}$$

However, the PHD measurement update equation needs to be approximated. The update equation requires that the predicted density is approximately Poisson[7]. That is, for all finite $X \subset \mathbb{S}$, one can write

$$p_{k|k-1}(X) \approx e^{-\mu}\mu^{|X|} \prod_{x \in X} s(x) \tag{7.27}$$

6. Mahler [2003a] allows also spawning, which is not included here.
7. "Poisson" means here a Poisson process in a more general sense than what was given in Definition 2.36. The general definition of a Poisson process is related to the theory of *point processes*, from which one can find information, e.g., in [Karr 1991].

for some $\mu \geq 0$ and probability density $s(x)$. Then, the PHD of the predicted density can be written as $D_{k|k-1}(x) = \mu \cdot s(x)$. Assuming that the prior is approximately Poisson, then one can write an approximate PHD measurement update as follows [Mahler 2003a, p. 1168].

$$D_{k|k}(x) \approx F_k(Y_k \mid x) D_{k|k-1}(x) \tag{7.28}$$

where the "PHD-likelihood" term F_k can be written as follows

$$F_k(Y_k \mid x) = \sum_{y \in Y_k} \frac{p_d(x) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x)}{\eta_k c(y) + \int D_{k|k-1} p_d(x) f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(y \mid x) D_{k|k-1}(x) \mathrm{d}x} + 1 - p_d(x)$$

These update formulae turn out much less complicated than the direct inference according to Equations (7.24) and (7.25). The combinatorial explosion of Equations (7.13) and (7.16) is reduced drastically. But what is the cost of this approach? Since this approximation propagates only one density $D_{k|k}$ instead of the full multitarget distribution, information is evidently lost. However, since PHD can be considered sufficient for many purposes, the most important question is that how drastic the approximation given in Equation (7.27) is? This question can be stated to be under rather intense research, but here is what the developer has written [Mahler 2003a, p. 1155].

> *... But if both sensor covariances and sensor false alarm densities are small then observations will be tightly clustered around target states, confusion due to false alarms will be small, and so the time-evolving multitarget posteriors will be roughly characterized by their first-order moments.*

In other words, the PHD approximation inherently assumes that the sensor has a rather high probability of detection, and the measurement errors are quite small. In addition, the number of false alarms should be moderate.

There is intense research going on with the practical implementation of the PHD approximation. Vo et al. [2003a,b] and Sidenbladh [2003] suggest sequential Monte Carlo implementation of Equations (7.26) and (7.28). Sidenbladh [2003] compared the performance of the PHD filter with respect to direct implementation of the random set Bayes recursions. They concluded that the performance of the PHD filter falls quickly with respect to the sensor signal-to-noise ratio (SNR). Vo et al. [2003a,b] reported simulations in a two-dimensional state space[8], in which the algorithm was able to detect tracks quite reliably from a cluttered background.

7.3.2 Other Proposed Methods

Mahler has proposed a para-Gaussian approximation of the multitarget estimation [Mahler 2000, pp. 50–52]. The approach is, however, considered too restrictive in the scope of this thesis. The article of Morelande and Challa [2003] outlined a random set tracking algorithm, but only on a conceptual level. The article did not provide a concrete algorithm, but postponed the development into the future.

---

8. The state space consisted of position and velocity components.

Since the multitarget distributions tend to be nonlinear and non-Gaussian, it is natural to consider a sequential Monte Carlo implementation. Sidenbladh and Wirkander [2003, 2004] have suggested a SMC implementation of the random set Bayes recursion. Their sensor model and the dynamic model of Sidenbladh and Wirkander [2004] are similar[9] to the models in Sections 7.1 and 7.2.1. The tracking model of Sidenbladh and Wirkander [2004] includes a flavour that may be considered contradictory: the birth model allows a dependency on the measurements[10]. The SMC algorithm Sidenbladh and Wirkander [2004] propose is based on the idea that one keeps a fixed amount of particles at each cardinality level. Consequently, this approach requires an upper bound for the number of targets. In addition, the computational complexity of the algorithm increases drastically with respect to this upper bound. Since the computational complexity of the algorithm is high, Sidenbladh and Wirkander [2004] propose some methods for bringing the complexity down. Most importantly, Sidenbladh and Wirkander [2004] suggest the PHD approach discussed in Section 7.3.1.

7.3.3 Implemented Sequential Monte Carlo Filter

This section outlines an SMC implementation of the random set tracking, that is based on the general SISR algorithm given in Algorithm 4.3. The implementation covers the random set tracking model that was introduced in Sections 7.1 and 7.2. The algorithm is best suited for sensors, that admit at most one measurement per report, such as the sensors that were discussed in Sections 7.2.2 and 7.2.3. In the case of varying number of measurements, i.e. the model given in Section 7.2.1, this algorithm has an exponentially increasing computational complexity with respect to the number of measurements and targets.

The implementation is somewhat different to what Sidenbladh and Wirkander [2003, 2004] suggested. The implementation admits a computational complexity that increases approximately linearly with respect to the expected number of targets in the scene. No prior limit for the number of targets is required, unlike in the implementation of Sidenbladh and Wirkander [2004]. Vo et al. [2003a,b] proposed this kind of an algorithm in a conceptual level, but did not present any tests. They suggested that the algorithm with the prior importance distribution could not be practical.

The implementation that is proposed in this section is, in fact, exactly the bootstrap filter, with adaptive resampling, given in a general form in Algorithm 4.3. The algorithm is theoretically not too complicated, but includes all sorts

9. In their model, the birth and false alarms events were not modelled as Poisson processes. There was a fixed amount of target objects, which were "alive" (in the surveillance region), or "dead" (outside the surveillance region). The objects were allowed to enter from one state to another with a certain probability. Similarly, there was a fixed amount of "clutter" objects, which each became measured with a certain probability.

10. They justify their choice with the notion that this model allows more efficient exploration of the state space, i.e. more efficient allocation of the random samples. Usually, in SMC algorithms, this is achieved by a proper choice of the importance distribution, as mentioned in Section 4.3, while the actual model is left untouched.

$\boldsymbol{Z}_0^{(i)} \sim P_{\boldsymbol{X}_0}$
$\boldsymbol{w}_0^{(i)} \leftarrow 1/n$
for $k = 1, 2, \ldots$ **do**
  $\boldsymbol{Z}_k^{(i)} \sim f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\cdot \mid \boldsymbol{Z}_{k-1}^{(i)})$
  $\widehat{\boldsymbol{w}}_k^{(i)} \leftarrow f_{\boldsymbol{Y}_k|\boldsymbol{X}_k}(Y_k \mid \boldsymbol{Z}_k^{(i)})$
  $\boldsymbol{w}_k^{(i)} \leftarrow \dfrac{\widehat{\boldsymbol{w}}_k^{(i)}}{\sum_{i=1}^n \widehat{\boldsymbol{w}}_k^{(i)}}$
  $\mathsf{E}\left[h(\boldsymbol{X}_{0:k}) \mid \boldsymbol{Y}_{1:k} = Y_{1:k}\right] \approx \sum_{i=1}^{n} \boldsymbol{w}_k^{(i)} h(\boldsymbol{Z}_{0:k}^{(i)})$
  if $n_{\text{eff}}(\{\boldsymbol{w}_k^{(i)}\}_{i=1}^n\}) < n_{\text{th}}$ **then**
    $(\boldsymbol{Z}_k^{(i)}, \boldsymbol{w}_k^{(i)})_{i=1}^n \leftarrow \text{resample}\left[(\boldsymbol{Z}_k^{(i)}, \boldsymbol{w}_k^{(i)})_{i=1}^n\right]$
  end if
end for

Algorithm 7.1: SISR implementation of random set estimation framework.

of implementation issues, regarding the choice of the components. The proposed algorithm is outlined in Algorithm 7.1. The symbols in the algorithm are in accordance with the symbols in Sections 7.1 and 7.2. Notice that the expectation that is estimated has to be real or vector valued, i.e. the function $h : \mathcal{F}_* \to \mathbb{R}^d$ maps finite sets into vectors.

The line of Algorithm 7.1 that states "$\boldsymbol{Z}_k^{(i)} \sim f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\cdot \mid \boldsymbol{Z}_{k-1}^{(i)})$" calls for some comments. The generation of random samples $\boldsymbol{Z}_k^{(i)}$ given the previous state $\boldsymbol{Z}_{k-1}^{(i)}$ is summarised in Algorithm 7.2. The generation of $\boldsymbol{Z}_k^{(i)}$ follows directly the presentation in Section 7.1. At first, the number of born targets $\boldsymbol{m}$ is drawn from the discrete Poisson distribution given in Equation (7.11). Then, the state of each of the born targets is drawn independently according to the birth density f_b. For each target in $\boldsymbol{Z}_{k-1}^{(i)}$, it is determined whether the target will survive. For the surviving targets, samples are drawn from the single-target motion model $f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}$.

7.4 Visualisation and Output

The multitarget posterior distribution $\pi_{k|k}(\mathcal{O}) \triangleq P(\boldsymbol{X}_k \in \mathcal{O} \mid \boldsymbol{Y}_{1:k} = Y_{1:k})$ contains all information that is relevant in view of estimation. An important question is that how one can extract a reasonable estimate or estimates from the posterior distribution $\pi_{k|k}$ when approximated with particles,

$$\pi_{k|k}(\mathcal{O}) \approx \boldsymbol{P}_{k|k}(\mathcal{O}) \triangleq \sum_{i=1}^{N} \boldsymbol{w}_k^{(i)} \delta_{\boldsymbol{Z}_k^{(i)}}(\mathcal{O}) \tag{7.29}$$

$\boldsymbol{m} \sim \text{Poisson}(\eta\tau_k)$
$\boldsymbol{B}_k \leftarrow \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{\boldsymbol{m}_k}\}$, where $\boldsymbol{b}_i \sim f_b(\cdot)$
$\boldsymbol{S}_k \leftarrow \emptyset$
for all $\boldsymbol{z} \in \boldsymbol{Z}_{k-1}$ **do**
  $\boldsymbol{u} \leftarrow U(0,1)$
  if $\boldsymbol{u} < p_s(\boldsymbol{z})$ **then**
    $\boldsymbol{S}_k \leftarrow \boldsymbol{S}_k \cup \{\boldsymbol{z}'\}$, where $\boldsymbol{z}' \sim f_{\boldsymbol{x}_k|\boldsymbol{x}_{k-1}}(\cdot \mid \boldsymbol{z})$
  end if
end for
$\boldsymbol{Z}_k \leftarrow \boldsymbol{S}_k \cup \boldsymbol{B}_k$

Algorithm 7.2: Algorithm that draws an sample $\boldsymbol{Z}_k$ from the predictive density $f_{\boldsymbol{X}_k|\boldsymbol{X}_{k-1}}(\cdot \mid \boldsymbol{Z}_{k-1})$ given in Equation (7.13).

where $\boldsymbol{w}_k^{(i)}$ and $\boldsymbol{Z}_k^{(i)}$ are the weights and the samples, respectively, that correspond to the symbols in Algorithm 7.1. The JoME and MaME estimates considered in Section 5.3.8 cannot be extracted from this representation straightforwardly, since both JoME and MaME are "MAP-like" estimates, which require maximisation of the posterior density[11].

A reasonable estimator for the whole multitarget state is difficult to construct from the particle representation. Simple estimators can, however, be obtained for the number of targets. The first estimator is the expected *a posteriori* (EAP) number of targets, given as follows

$$\widehat{n}_{k|k}^{\text{EAP}} = \mathsf{E}\left[|\boldsymbol{X}_k| \;\middle|\; \boldsymbol{Y}_{1:k} = Y_{1:k}\right] = \sum_{n=0}^{\infty} \left[n \cdot \pi_{k|k}(\mathcal{F}(n))\right]$$

where $\mathcal{F}(n) = \{X \subset \mathbb{S} : |X| = n\}$, i.e. the collection of all subsets of $\mathbb{S}$ with cardinality n. This estimate is also referred to as the mean estimated number of targets. The second estimator for target count can be obtained by maximising the posterior cardinality distribution,

$$\widehat{n}_{k|k}^{\text{MAP}} = \arg\max_{n \in \mathbb{N}} \pi_{k|k}(\mathcal{F}(n))$$

This is the maximum *a posteriori* (MAP) estimate of number of targets. These estimates are quite easy to obtain from the particle representation of Equa-

11. However, if one considers the hit-or-miss Bayes cost function given in Equation (2.13) with some macroscopic Δ, such estimate can be found that minimises the cost. The MAP estimate was given in the first place as a limit when Δ is decreased, and thus the MAP estimate cannot be considered a "proper" Bayes estimate. This estimate obtained from the particle representation cannot be considered very practical. The estimate has evidently quite large variance, if Δ is very small. On the other hand, if Δ is large, the estimate is coarse.

tion (7.29), since

$$\pi_{k|k}(\mathcal{F}(n)) = \sum_{i=1}^{N} \boldsymbol{w}_k^{(i)} \delta_{\boldsymbol{Z}_k^{(i)}}(\mathcal{F}(n)) = \sum_{i \in I(n)} \boldsymbol{w}_k^{(i)}$$

where $I(n) = \{i : |\boldsymbol{Z}_k^{(i)}| = n\}$, i.e. the set that contains the indices of the particles of cardinality n. Now, the EAP and MAP number of targets estimates can be extracted as follows

$$\widehat{n}_{k|k}^{\mathrm{EAP}} = \sum_{n=0}^{\infty} \left[n \sum_{i \in I(n)} \boldsymbol{w}_k^{(i)} \right] \tag{7.30}$$

$$\widehat{n}_{k|k}^{\mathrm{MAP}} = \arg\max_{n \in \mathbb{N}} \sum_{i \in I(n)} \boldsymbol{w}_k^{(i)} \tag{7.31}$$

The infinite sum in Equation (7.30) reduces naturally to a finite sum, since each particle contains a finite number of elements. Similarly, maximum has to be searched only from a finite set of possibilities in Equation (7.31). The EAP and MAP estimators for number of targets differ also in the sense that $\widehat{n}_{k|k}^{\mathrm{MAP}}$ is integer-valued, but $\widehat{n}_{k|k}^{\mathrm{EAP}}$ is real-valued.

Since the position distribution of the targets is even more important than the cardinality distribution, it is clear that the position distribution has to be illustrated in some manner—preferably in a manner that is easy to interpret by a human. In the case of single-target tracking, a rather straightforward and intuitive visualisation of the single-target position distribution would be to consider a two-dimensional histogram corresponding the (x, y) coordinates. This type of histogram could be illustrated as an intensity image. In the multiple target case, similar information can be shown in the form of PHD. Recall that PHD is the density of the target count measure,

$$\mathsf{E}\left[|\boldsymbol{X}_k \cap C| \,\middle|\, \boldsymbol{Y}_{1:k} = Y_{1:k} \right] = \int_C D_{k|k}(x) \mathrm{d}x$$

where $D_{k|k}$ is the posterior PHD. The PHD estimator that is considered in this thesis is defined so that the surveillance region is discretised into finite resolution cells. Then, for each finite resolution cell, the expected number of targets in the resolution cell is determined. This can be achieved as follows. Suppose A is a set corresponding to a resolution cell. Then, one determines

$$\widehat{n}_{k|k}^{\mathrm{EAP}}(A) = \mathsf{E}\left[|\boldsymbol{X}_k \cap A| \mid \boldsymbol{Y}_{1:k} = Y_{1:k} \right] = \int |Z \cap A| \mathrm{d}\pi_{k|k}(Z)$$

From this definition, one obtains for the particle representation

$$\widehat{n}_{k|k}^{\mathrm{EAP}}(A) = \sum_{i=1}^{N} \boldsymbol{w}_k^{(i)} |\boldsymbol{Z}_k^{(i)} \cap A| \tag{7.32}$$

The interpretation of the obtained PHD approximation is quite straightforward: the value $\widehat{n}_{k|k}^{\mathrm{EAP}}(A)$ of each resolution cell A corresponds to the (approximate) expectation of the number of targets within that cell.

While the PHD with an estimate of target count might be a sufficient representation of the posterior distribution for many purposes, there might still be situations where extraction of a unique estimate is desirable. Sidenbladh and Wirkander [2004] suggested fitting a mixture of Gaussians into the PHD to find peaks. The parameters of the Gaussians would then represent estimates of the target state distributions. It should be pointed out, however, that this method can be best thought as an indirect, heuristic estimator. The mixture-Gaussian representation of the PHD is an approximation, and information is lost when the PHD is extracted from the true multitarget distribution.

Chapter 8

Experimental Setup

This chapter covers the limited experimental setup, that is intended to give an idea how the algorithm described in Section 7.3.3 works in practice. The test scenarios consist of a varying number of targets in a surveillance region. There are sensors that measure the angle from the sensor towards the target based on some signal emitted by the target. This *bearings-only tracking* is a rather standard tracking scenario, that has been used in testing of several single-target and multitarget tracking algorithms [e.g. Carpenter et al. 1999; Gilks and Berzuini 2001; Gordon et al. 1993; Hue et al. 2002; Särkkä et al. 2004].

The experiments in this thesis are based on artificial tracks, and the data is generated according to the ideal measurement model, that is used also by the tracker. Since there are bearings-only measurements, which are corrupted by false alarms and missed measurements, the scenario can be considered difficult. The performance of the random set tracking algorithm was not compared with an existing tracking algorithm. This was due to the fact that no such alternative approaches were found in the literature that could be used in this kind of scenario.

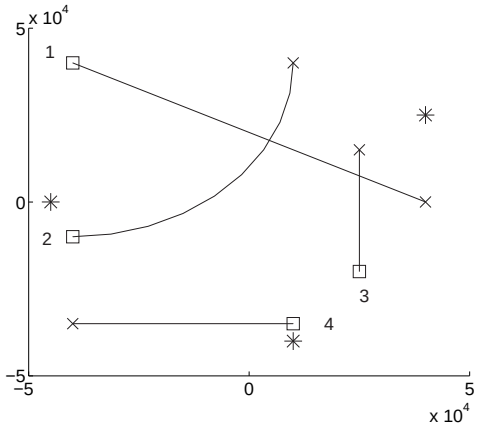
Section 8.1 describes the details of the models that were used in the tests. In addition, the general parameters of the tracking scenarios are outlined. The specifics of the parameters of each test setup are given in Section 8.2, where the results are also analysed.

8.1 Data and System Parameters

The algorithm that was tested was the SMC implementation of the random set tracking framework described in Section 7.3.3. The algorithm was written in Matlab[1] version 6.5. The tests were run on a PC having 2 GHz Intel Pentium 4 CPU and 512 MB of main memory. The number of Monte Carlo samples (particles) was fixed throughout the tests to 20000. The resample threshold n_{th} was set to one fourth of the total number of particles, 5000.

The state space $\mathbb{S}$ in the test scenarios was four-dimensional consisting of the x and y-coordinates, and the x and y velocities, denoted as x' and y',

1. See http://www.mathworks.com/products/matlab/.



ID	Start time	End time	Speed
1	5	220	416.0
2	35	270	891.1
3	70	210	339.7
4	50	280	272.7

Figure 8.1: The sensors ($*$) and the tracks (lines) in the scenarios. The tracks start from a square, and end into a cross. The table gives a summary of the temporal properties of the tracks.

respectively. The state space was bounded, so that for each $\underline{v} = [x, y, x', y']^T \in \mathbb{S}$, the following conditions were satisfied.[2]

- -50 km $\leq x, y \leq 50$ km
- -1000 m/s $\leq x', y' \leq 1000$ m/s

Figure 8.1 shows the positions of the three angular sensors, and the four constant speed tracks involved in the scenarios. All of the three sensors were used in all the tests, but the number of tracks that were selected varied. In the single-target scenario, only target 1 was selected. Correspondingly, the two-target scenario included targets 1 and 2, and the three-target scenario included, in addition, target 3.

If not otherwise mentioned, the measurement data was generated according to the measurement model involved in the corresponding test. In particular, all the tests used the zero or one measurement model described in Section 7.2.3. All the scenarios had a time span from 1 to 300 seconds. The measurements arrived sequentially at each second. The measurement generation process was such that at each second, the sensor that produced the measurement was picked randomly. Then, a measurement was generated according to the sensor model, as illustrated in Figure 7.2. Section 8.1.2 describes the default parameters of the measurement model, while the test-specific parameters are given in Section 8.2. The parameters of the dynamic model were fixed throughout the tests to the values described in Section 8.1.1.

2. Since it is convenient, for the sake of intuition, to introduce some units of measurement, we use the standard physical units of position and time. One should keep in mind, however, that the tests are purely synthetic. That is, one should not consider the tests mimicking some real world phenomena.

8.1.1 Dynamic Model

The dynamic model that was used in all the experiments is the model introduced in Section 7.1. The single-target dynamic model was the constant velocity model, given in Equation (6.2). The process noise standard deviation ρ was fixed to the value[3] 35 m/s$^{3/2}$. The value was set empirically.

The initial distribution was constrained to be "no targets", i.e. $P(\boldsymbol{X}_0 = \emptyset) = 1$. In all the scenarios, the true situation started with no targets in the surveillance region. However, this can be considered a reasonable initial distribution for any multitarget surveillance system, if there is no better prior information of the situation available. The track birth process was modelled as Poisson, as discussed in Section 7.1.2. The track birth rate[4] η was fixed to value 0.1 through all simulations. The distribution of the born targets P_b was uniform over the whole surveillance region. For practical reasons, the Poisson distribution was truncated so that the maximum number of born targets during one processing step was one.

The probability of track survival was 0.95 at maximum. To prevent the targets getting out of the surveillance region, soft boundaries were created so that the probability of track survival decreased linearly, if any of x, y is nearer than 10 km to the surveillance region border, or if any of x', y' is nearer than 50 m/s to the allowed velocity limits. When representing the target state as $\underline{v} = [x, y, x', y']^T$, the probability of track survival can be given as follows

$$p_s(\underline{v}) = \begin{cases} \min_i(0.95, \dfrac{\underline{v}(i) - \underline{m}(i)}{\underline{b}(i)}, \dfrac{\underline{M}(i) - \underline{v}(i)}{\underline{b}(i)}), & \underline{v} \in \mathbb{S} \\ 0, & \underline{v} \notin \mathbb{S} \end{cases}$$

where $\underline{m} = -[50000, 50000, 1000, 1000]^T$, $\underline{M} = [50000, 50000, 1000, 1000]^T$, and $\underline{b} = [10000, 10000, 50, 50]^T$. The case $\underline{v} \notin \mathbb{S}$, i.e. the vector is out of the surveillance region, is included due to the fact that the constant velocity single-target dynamic model allows the target to drift outside the surveillance region[5]. The probability of survival is exemplified in Figure 8.2 as a function of the (x, y)-position, with zero velocities.

8.1.2 Measurement Model

The measurement model that was used throughout the tests was the one described in Section 7.2.3. That is, the sensor included at most one measurement in each report. The sensors produced bearings-only measurements. This means, that the sensors returned only the direction of arrival (DOA) of the signal they measured. In the scenarios, the angular measurement model assumed additive noise. This noisy DOA measurement model can be given as follows.

$$\boldsymbol{y}_k = \operatorname{atan2}(\widehat{y}, \widehat{x}) + \boldsymbol{v}_k \mod 2\pi \tag{8.1}$$

3. The unit of the process noise variance ρ^2 may be more intuitive: $(\text{m/s})^2/\text{s}$.
4. That is, the expected number of born targets per second.
5. Strictly speaking, we should consider the state space as $\mathbb{S} = \mathbb{R}^4$. It is more convenient, though, to speak of the effective restriction of $\mathbb{R}^4$ as the state space.

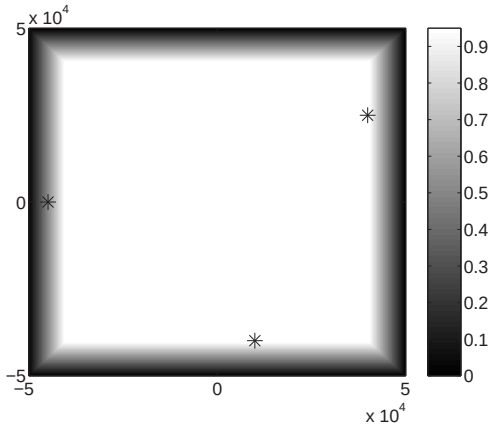


Figure 8.2: The probability of survival of a track as a function of the (x, y) position.

where $\boldsymbol{v}_k$ is a Gaussian distributed random variable, $\widehat{x} = x - s_x$ and $\widehat{y} = y - s_y$ are the target x and y coordinates with respect to the sensor position $\underline{s} = [s_x, s_y]^T$, and the function atan2 is the four quadrant inverse tangent, defined as follows.

$$\text{atan2}(y, x) = \begin{cases} \arctan(y/x), & x > 0 \\ \text{sgn}(y) \cdot \pi/2, & x = 0 \\ \arctan(y/x) + \pi, & x < 0 \end{cases} \quad \text{where} \quad \text{sgn}(y) = \begin{cases} 1, & y \geq 0 \\ -1, & y < 0 \end{cases}$$

That is, each measurement $y_k \in [0, 2\pi)$ is the true target angle perturbed by an additive Gaussian noise term[6],

$$P(\boldsymbol{y}_k \in A \mid \boldsymbol{x}_k = [x, y, x', y']^T) = N(A'; 0, \sigma_a^2)$$

where σ_a is the standard deviation of the angular measurements, having value $(2/180)\pi$. The set $A' \subset [-\pi, \pi)$ is the set $A \subset [0, 2\pi)$ mapped to residuals around the true angle, and can be given as follows

$$A' = \{a \in A : [\text{atan2}(\widehat{y}, \widehat{x}) - a + \pi \mod 2\pi] - \pi\}$$

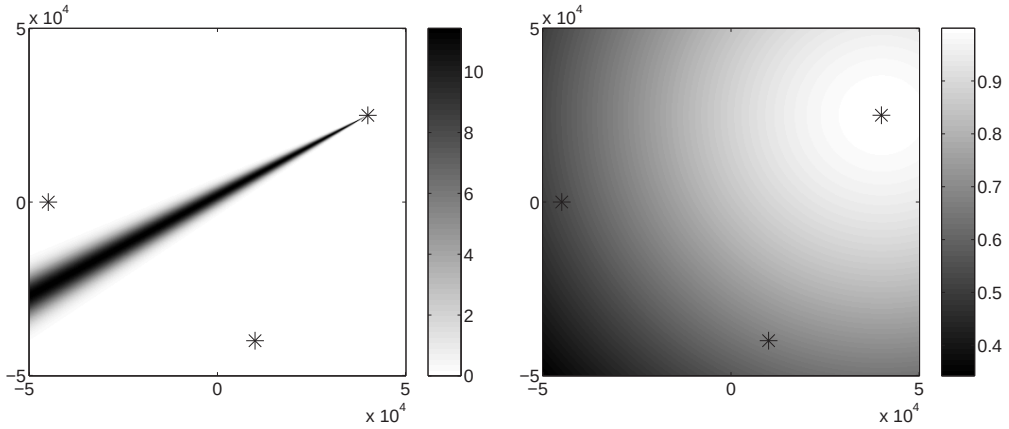
The density corresponding to this measure can be given as follows

$$f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(a \mid [x, y, x', y']^T) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left\{-\frac{1}{2\sigma_a^2} d^2\right\} \tag{8.2}$$

where

$$d = [\text{atan2}(\widehat{y}, \widehat{x}) - a + \pi \mod 2\pi] - \pi$$

6. In fact, the noise term is assumed to be *truncated Gaussian distributed*, since the tails of a true Gaussian distribution would continue beyond $[-\pi, \pi)$. However, since the standard deviation of the noise term is much less than π, the amount of probability mass beyond the limits can be considered negligible.



(a) (b)

Figure 8.3: (a) The single-target measurement likelihood function in the case of the measurement $y_k = 7\pi/6$ (b) The probability of detection for the same sensor. Both the images illustrate the values of the functions with respect to the (x, y) position.

The construction of the angular measurement model above contained many piecewise definitions, and was not too intuitive. Figure 8.3 (a) illustrates the measurement model, showing an intensity image of values of $f_{\boldsymbol{y}_k|\boldsymbol{x}_k}(a \mid [x, y, x', y']^T)$ with respect to x and y, when $a = 7\pi/6$. The measurement density $f_{\boldsymbol{y}|\boldsymbol{x}}(y \mid x)$ considered as a function of x for a fixed y is often referred to as the *likelihood function* of measurement $\boldsymbol{y} = y$.

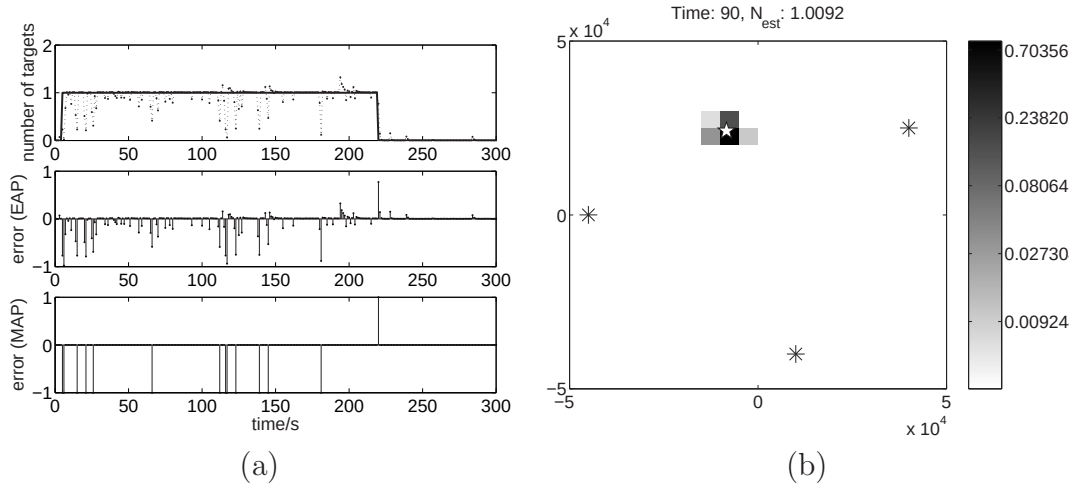
The measurement model allowed also detections to be missed. For that purpose, the function $p_d(\underline{v})$ was defined to determine the probability of detection of a target position $\underline{v}$. The model that was used in the experiments can be given as follows.

$$p_d([x, y, x', y']^T) = \exp\left(-\frac{1}{2\sigma_r^2}\|[\widehat{x}, \widehat{y}]^T\|^2\right)$$

The sensor range parameter σ_r that was used in the experiments was 80 km. Figure 8.3 (b) illustrates the values p_d gets in the different (x, y) positions of the state space. Finally, the sensor was allowed to produce false alarms. The default setting was that 5% of the measurements were false alarms, i.e. $p_f = 0.05$. False alarms were assumed to be distributed uniformly in $[0, 2\pi)$, i.e. $f_f(a) = 1/(2\pi)$ for all $0 \leq a < 2\pi$.

8.2 Tests

The outputs of the tracking algorithm that were included in the analysis consisted of the two estimates for the number of targets given in Equations (7.30) and (7.31). The PHD visualisation was carried out according to the estimator given in Equation (7.32). The surveillance region was divided into 20 equiwidth



(a) (b)

Figure 8.4: (a) The development of the estimates of target count. (b) A snapshot of the PHD after processing the 90th measurement in the single-target scenario.

intervals with respect to x and y dimensions, determining 400 equivolume (x, y) resolution cells. The velocity resolution was one, i.e. the velocity components were not distinguished in any manner.

Figure 8.4 (a) shows the development of the estimators for number of targets in the single-target scenario run with the default parameters. The uppermost graph shows the development of the EAP estimator (the dotted line), and the true number of targets (the solid line). The two graphs below show the development of errors of the EAP and MAP estimators, i.e. the differences between the values of the estimators and the true target count. Figure 8.4 (b) shows a snapshot of the PHD approximation at one time instant, after processing the 90th measurement. The star (pentagram) symbol in the PHD image corresponds to the true position of a target. The asterisks denote the positions of the three sensors. Notice the logarithmic scale of the intensities in the PHD image[7]. The numerical value of the EAP estimator is shown in the title of Figure 8.4 (b) in addition to the current time instant in seconds.

The actual tests consisted of three setups. The first setup analysed in Section 8.2.1 tests how the increase in the number of targets affects the algorithm. The effect of increased false alarm levels is considered in Section 8.2.2, while Section 8.2.3 contains a test which determines how the algorithm can handle with a mismatch between the true value of false alarm parameter and the parameter value given to the algorithm. Finally, the results that were obtained from the different setups are summarised in Section 8.2.4.

7. The absolute value of a particular bin of PHD is generally of no great importance. The general shape of the PHD image and the current estimated number of targets are more important.

8.2.1 Different Number of Targets

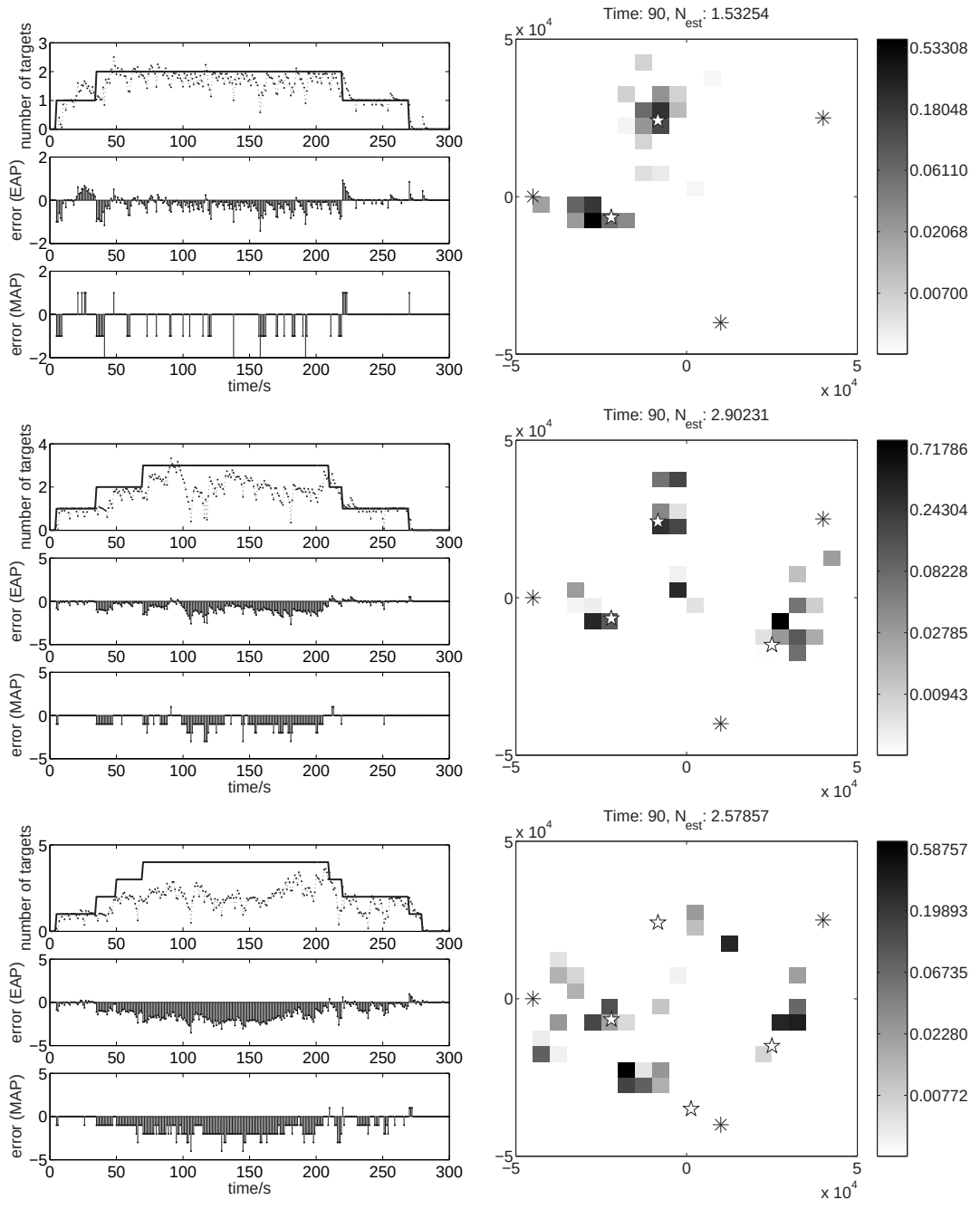
This test was intended to provide a glance on the capability of the algorithm to handle increasing number of targets. It is worth noticing, though, that the measurement model that was used does not provide too much information on the target situation. This is due the fact that the average number of measurements per target decreases linearly with respect to the total number of targets. Consequently, there is less information available for the algorithm to deal with a more complicated scenario.

The tests in this scenario were run with the default parameter values. Only the true number of targets was changed. Figure 8.5 shows the results obtained for the two, the three and the four target scenarios. The corresponding single-target scenario is shown in Figure 8.4. The results reflect what was expected: the performance decreases quite rapidly, as the number of targets increases. The results are quite reliable for up to two targets, but for a higher target count, the results can be considered poor. In the three-target case, the algorithm seems to "find" all the targets temporarily, as can be seen in the PHD snapshot of Figure 8.5 (b). In the four-target case, the algorithm never seems to truly find the targets, but oscillates mostly between single-target and three-target hypotheses.

There can be, of course, several reasons for the performance degradation. In author's opinion, however, the measurement data is insufficient, providing low number of true measurements per target, and can be considered the largest source of error. The number of particles, 20000, can also be considered low for tracking this many targets. It is obvious, that the state space that is effectively sampled becomes larger, as more targets appear in the scenario. This increases the variance in the results of the algorithm, so the number of particles would have to be increased as the number of targets is increased. Figure 8.6 shows the results of three runs of the algorithm with an identical input data, but different random seeds. Although the results for all the runs are quite good, there is noticeable variation in the results, especially in the beginning of the scenario. This can be considered as evidence of an insufficient number of particles. If the number of particles were sufficient, the Monte Carlo estimates would be relatively stable, and would not change noticeably between the runs of the algorithm. The number of particles could not, however, be increased due to the limited computational resources.

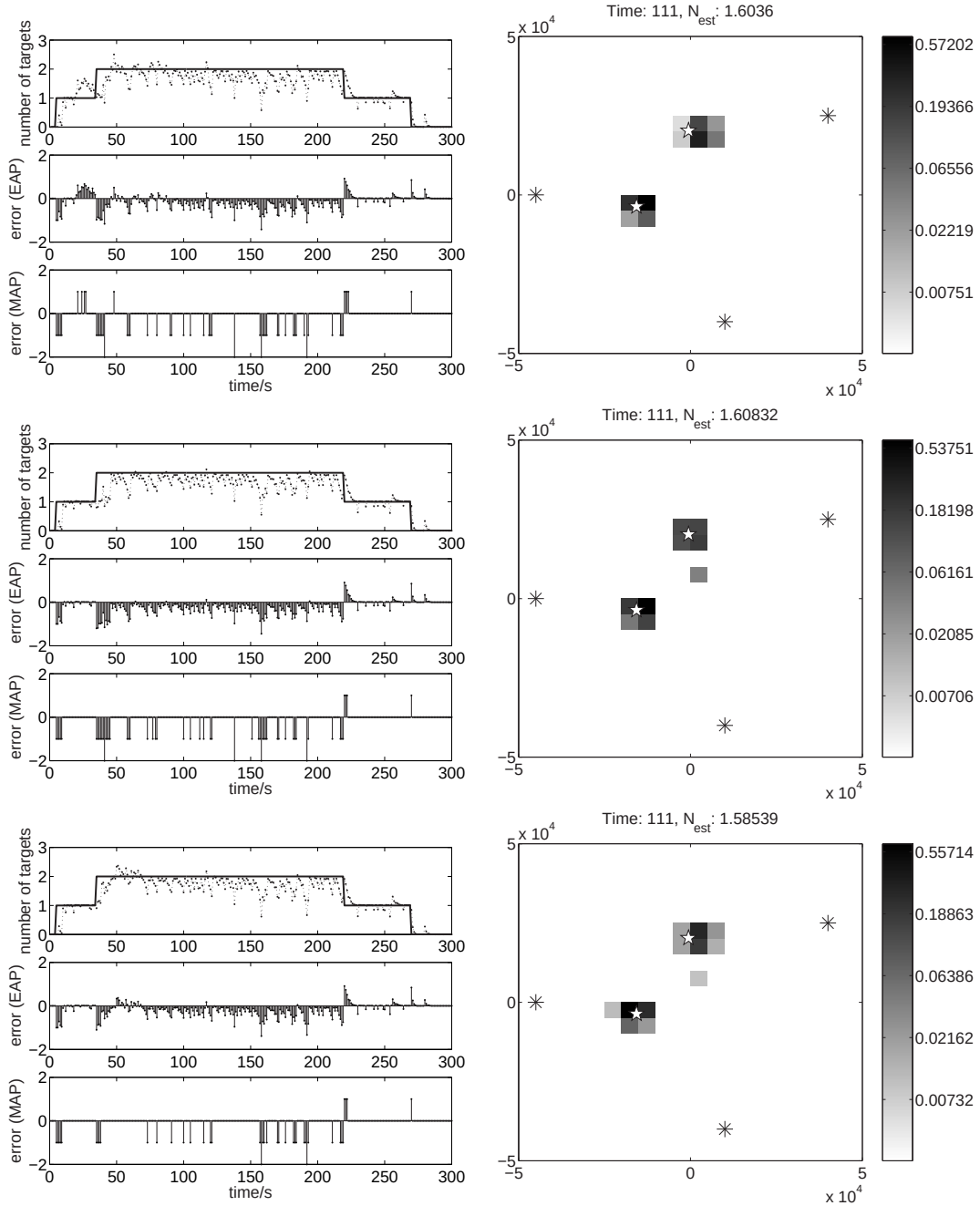
8.2.2 False Alarm Rate

The second test setup consisted of test runs with the single-target and the two-target scenario, with three different increased false alarm rates (FARs). The results for the single-target scenario with FARs 30%, 70%, and 80% are shown in Figure 8.7. As a comparison, the result with the default FAR, 5%, is shown in Figure 8.4. The PHD snapshots in Figure 8.7 (b) show how the algorithm reacts to the false alarms. There are "ghost" tracks in addition to the true one. Due to the increased FAR, there are less true target measurement, which delays the initiation of the track. This can be seen in Figure 8.7 (a), where the EAP and MAP



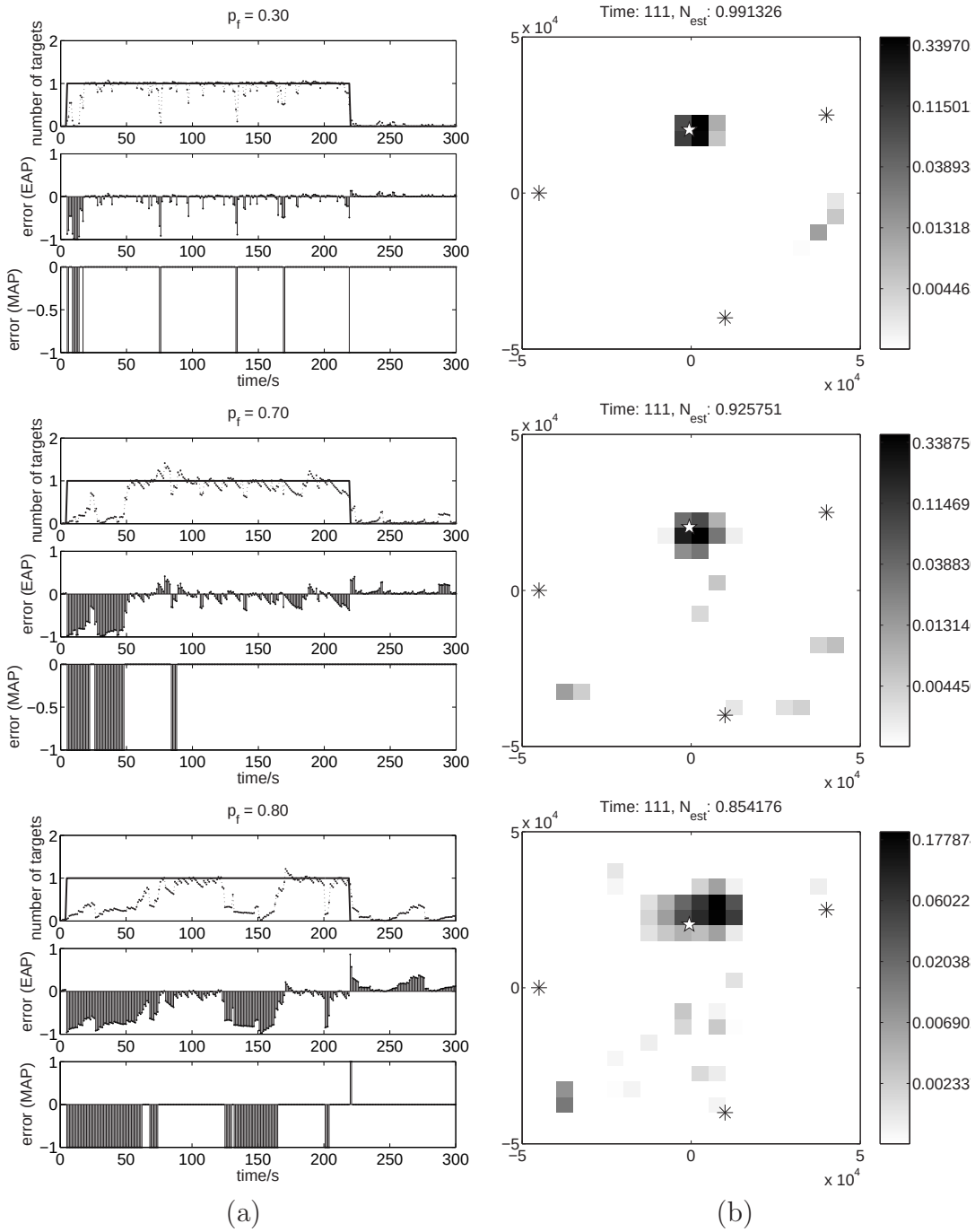
(a) (b)

Figure 8.5: (a) The development of the estimates of target count, with different number of true targets. (b) The PHD snapshots after processing the 90th measurement.



(a) (b)

Figure 8.6: (a) The development of the estimates of target count in the two-target scenario with the default parameters. (b) The snapshot of the PHD after processing the 111th measurement.



(a) (b)

Figure 8.7: (a) Development of the estimates of number of targets in the single-target scenario, with different increased probability of false alarm levels. (b) PHD snapshots after processing of the 111th measurement.

estimates rise into unity approximately after the 15th measurement in the case of 30% FAR, but only after the 50th measurement in the case of 70% FAR. According to this setup, the algorithm can handle false alarms rather well, providing quite reliable tracking results with as high FAR as 70%. As FAR is increased to 80%, the algorithm seems to lose the track temporarily.

The algorithm was tested also in the two-target scenario, with three increased FARs: 10%, 30%, and 50%. The development of the estimates of the number of targets are shown in Figure 8.8. As expected, the performance of the algorithm decreases more rapidly in the two-target scenario, when the FAR is increased. The reason for this is, most probably, the same what was considered to decrease the performance in the case of the three and the four-target scenarios: insufficiency of the data. As FAR is increased, the number of true measurements from each target decreases. Of course, as there are two targets, each of the non-false-alarms are equiprobably from either one of the targets.

This test setup showed, most importantly, that the degradation of the performance of the algorithm is gradual with respect to the increasing FAR. This property is promising in view of such low-fidelity sensors, that are prone to produce many false alarms.

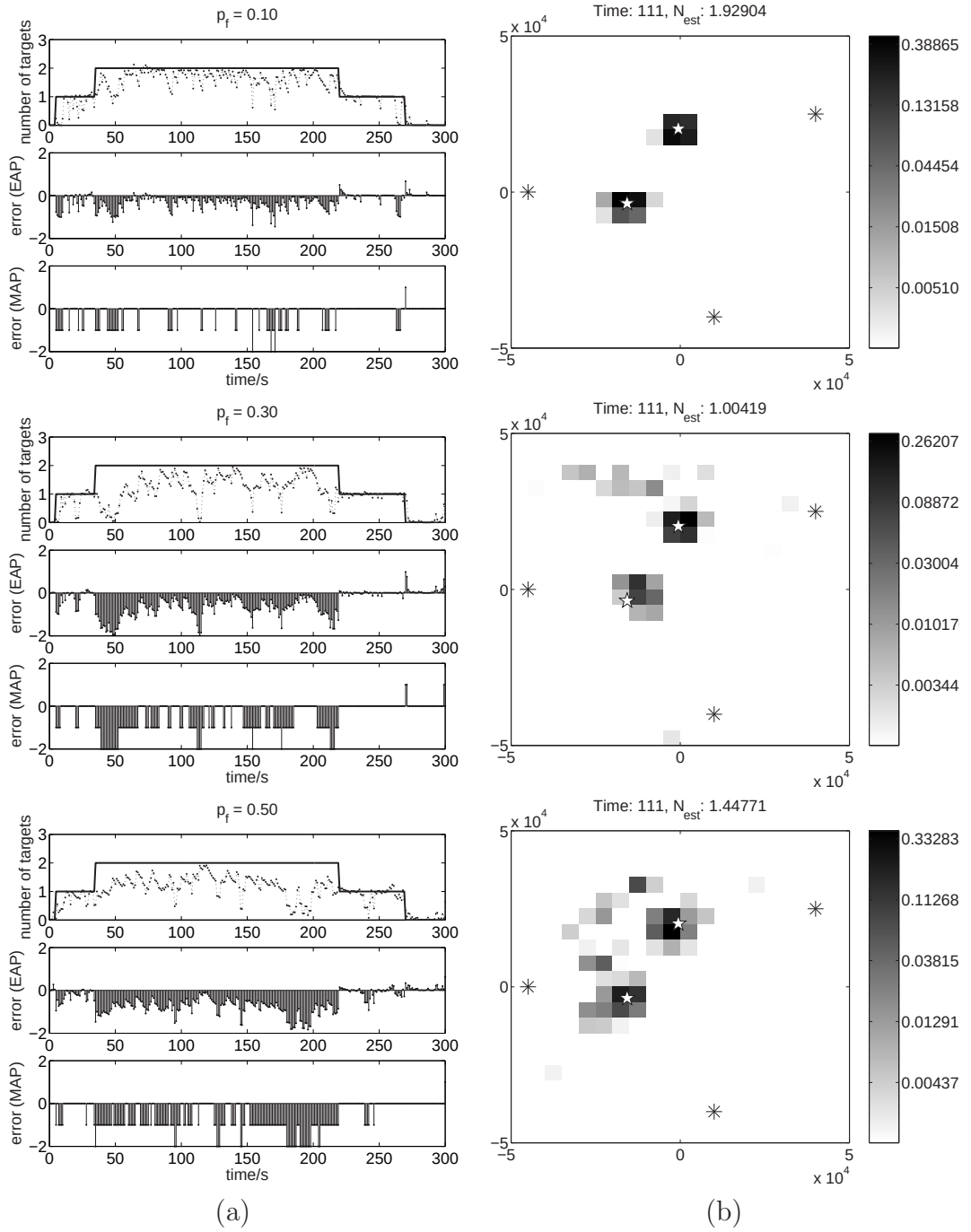
8.2.3 Parameter Mismatch

In real world, the tracking model rarely matches the characteristics of the true tracking environment. Most importantly, the parameters of the model that are adjusted manually can be expected to be misleading in some cases. So, a tracking framework should be robust against incorrect parameter values, at least in some extent[8]. Therefore, a test of robustness against a parameter mismatch was included. Since there was a test how the false alarm rate affects, the chosen target parameter was the percentage of false alarms.

The test was carried out using the single-target scenario. The only parameter value that was altered from its default value was the probability of false alarm, p_f. The FARs that were used in the experiment were 5%, 30% and 70%. The input data, i.e. the measurements, were generated separately according to each FAR. After that, for each input data, the algorithm was run with each of the three values of the probability of false alarm parameter p_f.

The mean absolute errors of the EAP and MAP estimates in the case of all the combinations of the parameter values and the true (simulated) values are shown in Table 8.1. The trend seems to be that the true parameter values produced the best results. However, for 5% FAR input data, the mean absolute error of the EAP for algorithm running on 30% setting was the smallest. Similarly, for 30% FAR input data, the algorithm running on p_f setting 5% had a smaller mean MAP error than the true setting. These "anomalies" can be explained by the Monte Carlo variation in the output of the algorithm, and the particular input data. Most importantly, though, the performance of the algorithm decreased

8. Of course, if an algorithm is made extremely robust against incorrect parameter values, then the parameters do not affect in any way.



(a) (b)

Figure 8.8: (a) The development of the estimates of target count in the two-target scenario with different increased probability of false alarm levels. (b) The corresponding PHD snapshots after processing of the 111th measurement.

gradually with respect to the mismatch in the parameter value. The worst result was obtained with input data having 70% false alarms, and parameter value set to 5%. Figure 8.9 shows snapshots of the PHD after processing the 111th measurement[9]. The gradual decrease of the performance with respect to the increasing mismatch in the parameter value can also be seen in the PHD snapshots.

8.2.4 Summary of Results

Before getting into the summary of the results that were obtained during the performed tests, there is a practical aspect, that is worth noticing. Figure 8.10 shows how the computational load of the algorithm[10] increases with respect to the mean cardinality of the particles. The processing time is not a direct function of the mean cardinality, but the tendency can be seen: the processing load increases approximately linearly with respect to the mean cardinality.

In light of the results that were obtained in the tests, the framework seems promising. Even though the scenarios that were used were purely synthetic, and can be easily judged toy examples, they provide a challenging task for any tracking algorithm. There were spurious detections (false alarms) as well as missed detections[11], and the measurements were bearings-only type. The implemented algorithm seemed to work out well in this challenging setup.

The algorithm showed robustness against relatively high false alarm levels. In addition, the performance of the algorithm degraded gradually in the case of invalid probability of false alarm parameter values. The number of particles that is required to obtain stable behaviour of the algorithm is rather high, since there was noticeable variation between the results obtained from different runs of the two-target scenario with 20000 particles. In the case of bearings-only data with zero or one measurements per report, the method was able to handle up to two targets quite reliably, and so that the computational load did not increase too much.

---

9. The colour bars are omitted due to a limited space, and since the absolute values of the PHD are considered unimportant.
10. The measure of computational load is the elapsed CPU time (Matlab command `cputime`).
11. Such detections are also referred to as *false positives* and *false negatives*, respectively.

Table 8.1: Mean absolute error of the estimates of number of targets, when the parameter value was altered from the true ones. The errors of the EAP estimate are shown in (a), and the errors of the MAP estimate in (b). The smallest error value in each row is shown in boldface.

(a)

True value	Parameter value		
	5%	30%	70%
5%	0.077	**0.075**	0.106
30%	0.104	**0.086**	0.107
70%	0.581	0.244	**0.214**

(b)

True value	Parameter value		
	5%	30%	70%
5%	**0.047**	0.050	0.087
30%	**0.040**	0.053	0.073
70%	0.480	0.180	**0.153**

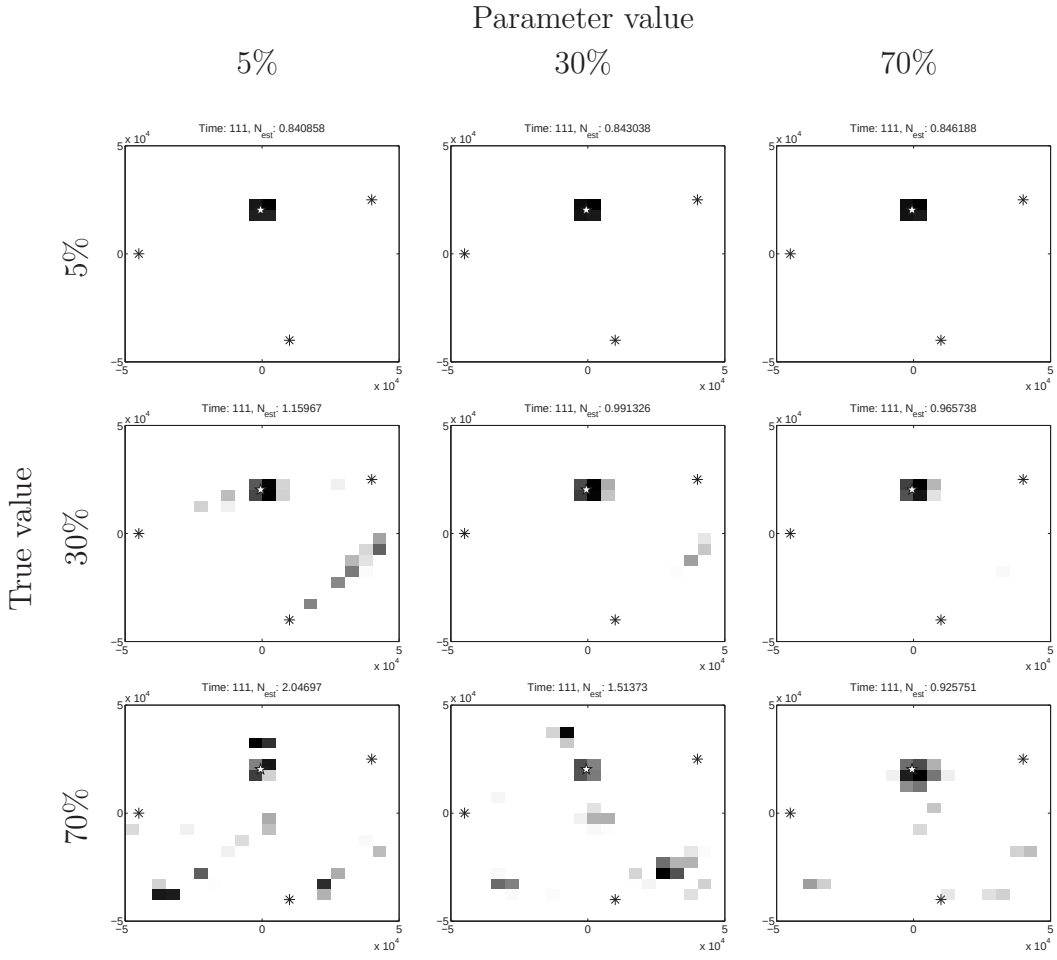


Figure 8.9: The PHD snapshots corresponding the setup in Table 8.1 after processing the 111th measurement.

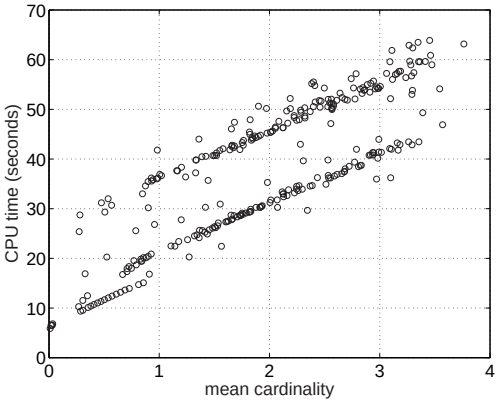


Figure 8.10: CPU time consumption for one processed measurement with respect to mean cardinality of particles.

Chapter 9

Conclusions

This thesis gathered together background theory for the random set tracking framework. The background theory included results in general probability theory and recursive Bayesian estimation. The finite random sets were considered as a special case of an abstract random element. The theory of random sets was presented in view of finite set measures and general integrals with respect to those measures. Sequential Monte Carlo (SMC) was selected as the computational strategy for practical implementation of random set estimation. The thesis included also a literature review of some common SMC methods.

The application-oriented part of the thesis covered a literature review and a summary of Bayesian target tracking methods. Even though the random set formalism of target tracking includes some theoretical aspects, that are not considered in the prevailing state of the art tracking framework, they share a lot in common. Random set models for tracking include the single-target dynamic models and the single-target measurement models, that are used by any conventional target tracking system. The random set tracking framework was presented in this thesis as a straightforward generalisation of the conventional tracking techniques.

The thesis contained derivation of a simple random set model for tracking. The model can be considered, however, to be sufficiently general for practical application. The random set dynamic model included independent target motions, an independent Poisson birth model and an independent death model. In addition to those random set sensor models that have appeared in the literature, this thesis proposed a model for such sensors that include at most one measurement in each report. After introducing a random set tracking model, a simple and straightforward SMC implementation performing recursive estimation in the model was introduced. The results that were obtained in the limited experiments are summarised in Section 9.1. The experiments and the obtained results can be considered best as a "proof-of-concept", rather than tests assessing the practical usability of the method. Section 9.2 contains the future development ideas, which were considered important.

9.1 Results

The SMC implementation was tested in rather simple yet challenging bearings-only tracking scenarios. The experimental results show that the random set framework has potential. Based on the results that were obtained, one can summarise that the proposed random set implementation can be applied to tracking of up to two targets quite reliably. The algorithm showed robustness against relatively high false alarm rates, as well as a mismatch in the true parameter value and the parameter value of the algorithm. Such situations, in which the false alarm rate is high, can be considered hard to tackle with conventional framework including a heuristics-based track initiation algorithm. At least such algorithms that process the input data in a fixed size time-window may be problematic.

The algorithm that was developed is a rather simple one. Further research and development is required for the purposes of real-life application. The computational complexity of the implementation is high. It seems, that without substantial reduction of the computational complexity, the SMC implementation of the random set Bayes recursion cannot be applied to tracking of a large number (tens or more) of targets. The computational complexity of the algorithm increases only approximately linearly with respect to the number of targets, when the number of Monte Carlo samples is kept fixed. However, to achieve a given accuracy, the number of Monte Carlo samples needs to be increased, as the number of targets increases. Hence, the computational complexity of the algorithm increases more rapidly in practice. For such purposes, when there are a lot of targets, the PHD approximation of the random set tracking framework may be considered a better alternative. For further development purposes, the developed simple implementation can serve as a good baseline performance test.

The tests that were included in this thesis considered only few aspects. There are dozens of other parameters, that were fixed throughout the experiments, and need to be considered in thorough tests. For example, the birth intensity and the probability of survival parameters should be examined. Other types of sensors models in addition to angular-only sensors might be considered as well. The state space that was considered, $\mathbb{S} \subset \mathbb{R}^4$, was quite low-dimensional considering real-life application. In addition, the state space that was used in the experiments did not include any non-geokinematic quantities, i.e. attributes. If attributes are included, they may affect the performance of the algorithm substantially.

9.2 Future Work

The theoretical part of the thesis provides a solid basis for further development of random set target tracking algorithms. During writing of this thesis, some ideas for further development have already arisen. This section presents the three branches of possible further development, that were considered the most promising, and worth stating.

The predictive importance distribution that was used in the implementa-

tion is ineffective, especially since the born target states are drawn from a sparse (uniform) prior distribution, without any dependence on the measurements. Many samples are drawn from very unlikely regions of the state space, and consequently become discarded almost immediately in the next resampling step. Therefore, a choice of a better importance distribution could be beneficial. Especially such a choice of an importance distribution that takes the measurements into account in drawing states of the born targets could be considered. An importance distribution other than the predictive one raises a practical issue, which may affect the computational complexity of the algorithm. The use of the predictive importance distribution avoided the need to compute Equation (7.13), but any other choice of the importance distribution requires the ratio of the dynamic model density and the importance density to be computed.

The state space $\mathbb{S}$ is typically at least 6-dimensional in a real-world scenario[1]. When there are multiple targets, say n, the state space that is sampled is effectively $\mathbb{S}^n$. Monte Carlo sampling in higher dimensions leads to higher variance estimates, or increased number of samples to attain a given accuracy. This problem has been proposed to be partially overcome by Rao-Blackwellisation of sampling schemes. Loosely speaking, Rao-Blackwellisation is a Monte Carlo sampling framework in which some parts of the model are integrated analytically, while the others are sampled. Rao-Blackwellisation has been proposed to tracking multiple targets with bearings-only measurements [Särkkä et al. 2004]. It could be possible to derive a similar approach in the random set framework.

The PHD approximation in the random set tracking framework has drawn much attention recently. The PHD approach is attractive because it reduces the filtering task to the state space $\mathbb{S}$, instead of the full random finite set space, which is infinite dimensional[2]. The PHD method is perhaps the single most interesting area of further research. Considering the development of a practical random set tracking system, research on an efficient SMC implementation of the PHD filter will most likely produce the best results.

In addition to these ideas of further development of the tracking algorithm, there is a practical issue, considering the output of the algorithm. It is not straightforward to obtain theoretically sound estimators for the multitarget state, when the posterior is represented with random samples. This is the case in the SMC implementation. Hence, an important aspect, that requires further research, is finding a good method to extract such an estimate.

---

1. The dimensions corresponding to the 3D position and velocity components, when a constant velocity model is used.
2. The random set space is essentially of form $\bigcup_{k=0}^{\infty} \mathbb{S}^k$.

References

D. L. Alspach and H. W. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, AC-17(4):439–448, Aug. 1972.

M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb. 2002.

D. Avitzour. Stochastic simulation Bayesian approach to multitarget tracking. *IEE Proceedings—Radar, Sonar and Navigation*, 142(2):41–44, Apr. 1995.

Y. Bar-Shalom. Tracking methods in a multitarget environment. *IEEE Transactions on Automatic Control*, AC-23(4):618–626, Aug. 1978.

Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, San Diego, 1988. ISBN 0-12-079760-7.

Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking: Principles, Techniques, and Software*. Artech House Inc., Norwood, MA, 1993. ISBN 0-89006-643-4.

Y. Bar-Shalom and X.-R. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, Storrs; CT, 1995. ISBN 0-9648312-0-1.

R. G. Bartle. Return to the Riemann integral. *The American Mathematical Monthly*, 103(8):625–632, Oct. 1996.

N. Bergman. *Recursive Bayesian Estimation: Navigation and Tracking Applications*. PhD thesis, Linköping University, 1999. URL http://www.control.isy.liu.se/research/reports/Ph.D.Thesis/PhD579.pdf.

S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18, Jan. 2004.

S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, Boston, 1999. ISBN 1-58053-006-0.

A. Blake and M. Isard. The CONDENSATION algortihm—conditional density propagation and applications to visual tracking. In *Proceedings of the 1996 Conference on Advances in Neural Information Processing Systems 9*, pages 361–367. MIT Press, 1997.

H. A. P. Blom and Y. Bar-Shalom. The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control*, 33(8), Aug. 1988.

J. Carpenter, P. Clifford, and P. Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings—Radar, Sonar and Navigation*, 146(1):2–7, Feb. 1999.

S. Challa, B.-G. Vo, and X. Wang. Bayesian approaches to track existence – IPDA and random sets. In *Proceedings of the Fifth International Conference on Information Fusion*, volume 2, pages 1228–1235, July 2002.

H. Chen, Y. Bar-Shalom, K. R. Pattipati, and T. Kirubarajan. MDL approach for multiple low observable track initiation. *IEEE Transactions on Aerospace and Electronic Systems*, 39(3):862–882, July 2003.

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2003. ISBN 0-387-98767-3.

D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3), Mar. 2002.

S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom. A generalized S-D assignment algorithm for multisensor-multitarget state estimation. *IEEE Transactions on Aerospace and Electornic Systems*, 33(2):523–538, Apr. 1997.

J. L. Doob. The development of rigor in mathematical probability (1900–1950). *The American Mathematical Monthly*, 103(7):586–595, 1996.

A. Doucet. On sequential simulation-based methods for Bayesian filtering. Technical Report CUED/F-INFENG/TR.310, Signal Processing Group, Department of Engineering, University of Cambridge, Cambridge, UK, 1998.

A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001. ISBN 0-387-95146-6.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

P. Fearnhead. *Sequential Monte Carlo methods in filter theory*. PhD thesis, University of Oxford, 1998. URL http://www.maths.lancs.ac.uk/~fearnhea/publications/.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley & Sons, New York, second edition, 1971. ISBN 0-471-257095.

T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, OE-8(3):173–184, July 1983.

R. F. Gariepy and W. P. Ziemer. *Modern Real Analysis*. PWS Publishing Company, 1995. ISBN 0-534-94404-3.

J. Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339, Nov. 1989.

W. R. Gilks and C. Berzuini. Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):127–146, 2001.

I. R. Goodman, R. P. S. Mahler, and H. T. Nguyen. *Mathematics of Data Fusion*, volume 39 of *Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers, AA Dordrecht, The Netherlands, 1997. ISBN 0-7923-4674-2.

N. Gordon. A hybrid bootstrap filter for target tracking in clutter. *IEEE Transactions on Aerospace and Electronic Systems*, 33(1):353–358, Jan. 1997.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund. Particle filters for positioning, navigation, and tracking. *IEEE Transactions on Signal Processing*, 50(2), Feb. 2002.

D. L. Hall and S. A. H. McMullen. *Mathematical Techniques in Multisensor Data*

Fusion. Artech House Publishers, Norwood, MA, second edition, 2004. ISBN 1-58053-335-3.

S. M. Herman. *A Particle Filtering Approach to Joint Passive Radar Tracking and Target Classification*. PhD thesis, University of Illinois, Urbana, Illinois, 2002. URL http://users.ece.gatech.edu/~lanterma/pcl/papers/herman_thesis_1side.pdf.

Z. Hu, H. Leung, and M. Blanchette. Statistical performance analysis of track initiation techniques. *IEEE Transactions on Signal Processing*, 45(2):445–456, Feb. 1997.

C. Hue, J.-P. Le Cadre, and P. Pérez. Tracking multiple objects with particle filtering. Technical report, Institut national de recherche en informatique et en automatique (INRIA), Oct. 2000.

C. Hue, J.-P. Le Cadre, and P. Pérez. Sequential monte carlo methods for multiple target tracking and data fusion. *IEEE Transactions on Signal Processing*, 50 (2):309–325, Feb. 2002.

K. Ito and K. Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927, May 2000.

F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001. ISBN 0-387-95259-4.

O. Kaleva. *Funktionaalianalyysi*. Tampereen Teknillinen Yliopisto, Tampere, 2003. (in Finnish).

K. Kanazawa, D. Koller, and S. Russell. Stochastic simulation algorithms for dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Eleventh Conference on*, pages 246–351, Montreal, Canada, 1995. Morgan Kaufmann.

R. Karlsson. *Simulation Based Methods for Target Tracking*. Lic. thesis, Lingköpings universitet, Lingköping, 2002.

R. Karlsson and F. Gustafsson. Monte Carlo data association for multiple target tracking. In *Proceedings of the IEE International Seminar on Target Tracking: Algorithms and Applications*, volume 1, pages 13/1–13/5, Oct. 2001.

R. Karlsson, F. Gustafsson, and T. Karlsson. Particle filtering and Cramér-Rao lower bound for underwater navigation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

A. F. Karr. *Point Processes and Their Statistical Inference*. Marcel Dekker, Inc., New York, second edition, 1991. ISBN 0-8247-8532-0.

K. Kastella. Event averaged maximum likelihood estimator and mean-field theory in multitarget tracking. *IEEE Transactions on Automatic Control*, 40(6):1070–1074, June 1995.

S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993. ISBN 0-13-042268-1.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. Dr. Tech. thesis, Tampere University of Technology, Tampere, 2004.

P. Korpisaari. *Studies on Data Fusion Techniques for Multitarget Tracking*. Dr. Tech. thesis, Tampere University of Technology, Tampere, 2001.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 2002. ISBN 0-19-852219-3.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2003. ISBN 0-387-95230-6.

A. L. Madsen and F. V. Jensen. Lazy propagation: A junction tree inference algortihm based on lazy evaluation. *Artificial Intelligence*, 113:203–245, 1999.

R. Mahler. *An Introduction to Multisource-Multitarget Statistics and its Applications*. Lockheed Martin, Mar. 2000.

R. Mahler. Engineering statistics for multi-object tracking. In *Proceedings of the IEEE Workshop on Multi-Object Tracking*, pages 53–60, Vancouver, July 2001a.

R. Mahler. Random set theory for target tracking and identification. In D. L. Hall and J. Llinas, editors, *Handbook of Multisensor Data Fusion*. CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431, 2001b. ISBN 0-8493-2379-7.

R. Mahler. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152–1178, Oct. 2003a.

R. Mahler. Objective functions for Bayesian control-theoretic sensor management I: Multitarget first-moment approximation. In *Proceedings of the IEEE Aerospace Conference*, volume 4, pages 1905–1923, Mar. 2003b.

R. Mahler. "Statistics 101" for multisensor, multitarget data fusion. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):53–64, Jan. 2004.

M. Mallick, T. Kirubarajan, and S. Arulampalam. Out-of-sequence measurement processing for tracking ground target using particle filters. In *Proceedings of the IEEE Aerospace Conference*, volume 4, pages 1809–1818, Mar. 2002.

G. Matheron. *Random Sets and Integral Geometry*. John Wiley & Sons, Inc., New York, 1975. ISBN 0-471-57621-2.

N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, Sept. 1949.

M. R. Morelande and S. Challa. A multitarget tracking algorithm based on random sets. In *Proceedings of the Sixth International Conference on Information Fusion*, volume 2, pages 807–814, Cairns, Australia, July 2003.

S. Mori, C.-Y. Chong, E. Tse, and R. P. Wishner. Tracking and classifying multiple targets without *a priori* identification. *IEEE Transactions on Automatic Control*, AC-31(5):401–409, May 1986.

K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002. URL http://www.ai.mit.edu/~murphyk/papers.html.

K. P. Murphy. Dynamic Bayesian networks. In M. Jordan, editor, *Probabilistic Graphical Models*. Nov. 2003. URL http://www.ai.mit.edu/~murphyk/papers.html. (draft).

D. Mušicki and R. Evans. Joint integrated probabilistic data association - JIPDA. In *Proceedings of the Fifth International Conference on Information Fusion*, volume 2, pages 1120–1125, July 2002.

D. Mušicki, R. Evans, and S. Stankovic. Integrated probabilistic data association. *IEEE Transactions on Automatic Control*, 39(6), June 1994.

A. W. Naylor and G. R. Sell. *Linear Operator Theory in Engineering and Science*. Springer-Verlag, New York, 1982. ISBN 0-387-95001-X.

M. Nørgaard, N. K. Poulsen, and O. Ravn. Advances in derivative-free state estimation for nonlinear systems. Technical Report IMM-REP-1998-15, Technical University of Denmark, 2800 Lyngby, Denmark, 1998. URL http://www.imm.dtu.dk/documents/ftp/publtr.html. (Revised edition, April 2000).

K. R. Pattipati, S. Deb, Y. Bar-Shalom, and J. Robert B. Washburn. A new relaxation algorithm and passive sensor data association. *IEEE Transactions on Automatic Control*, 37(2):198–213, Feb. 1992.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, AC-24(6):843–854, Dec. 1979.

B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Publishers, 2004. ISBN 1-58053-631-X.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999. ISBN 0-387-98707-X.

S. M. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1983. ISBN 0-471-09942-2.

H. L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, third edition, 1989. ISBN 0-02-946620-2.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976. ISBN 0-07-Y85613-3.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Pearson Education Inc., Upper Saddle River, New Jersey, second international edition, 2003. ISBN 0-13-080302-2.

L. Råde and B. Westergren. *Beta Mathematics Handbook for Science and Engineering*. Studentlitteratur, Lund, Sweden, fourth edition, 1998. ISBN 91-44-00839-2.

D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1665–1670, Seoul, Korea, May 2001.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

A. N. Shiryaev. *Probability*. Springer-Verlag, New York, second edition, 1996. ISBN 0-387-94549-0.

H. Sidenbladh. Multi-target particle filtering for the probability hypothesis density. In *Proceedings of the Sixth International Conference on Information Fusion*, volume 2, pages 800–806, Cairns, Australia, July 2003.

H. Sidenbladh and S.-L. Wirkander. Tracking random sets of vehicles in terrain. In *Proceedings of the 2nd IEEE Workshop on Multi-Object Tracking*, Madison, WI, USA, 2003.

H. Sidenbladh and S.-L. Wirkander. Particle filtering for finite random sets. *IEEE Transactions on Aerospace and Electronic Systems*, 2004. URL http://www.nada.kth.se/~hedvig/publications_en.html. (manuscript in revision, to appear).

L. D. Stone. A Bayesian approach to multiple-target tracking. In D. L. Hall and

J. Llinas, editors, *Handbook of Multisensor Data Fusion*. CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431, 2001. ISBN 0-8493-2379-7.

L. D. Stone, C. A. Barlow, and T. L. Corwin. *Bayesian Multiple Target Tracking*. Artech House, Norwood, MA, 1999. ISBN 1-58053-024-9.

R. L. Stratonovich. *Conditional Markov Processes and Their Application to the Theory of Optimal Control*. Modern Analytic and Computational Methods in Science and Mathematics. Elsevier, New York, 1968.

S. Särkkä, A. Vehtari, and J. Lampinen. Rao-Blackwellized Monte Carlo data association for multiple target tracking. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 583–590, Stockholm, Sweden, June 2004.

T. Valkonen. Puuttuvan tiedon käsittely tilastollisessa estimoinnissa. MSc thesis, Tampere University of Technology, Tampere, Oct. 2002. (in Finnish).

H. L. van Trees. *Detection, Estimation, and Modulation Theory. Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley and Sons, New York, 1968.

B.-G. Vo, S. Singh, and A. Doucet. Sequential Monte Carlo implementation of the PHD filter for multi-target tracking. In *Proceedings of the Sixth International Conference on Information Fusion*, volume 2, pages 792–799, Cairns, Australia, July 2003a.

B.-N. Vo, S. Singh, and A. Doucet. Random finite sets and sequential Monte Carlo methods in multi-target tracking. In *Proceedings of the International Conference on Radar*, pages 486–491, Adelaide, Australia, Sept. 2003b.

E. A. Wan and R. van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of IEEE Symposium on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC)*, Lake Louise, Alberta, Canada, Oct. 2000.

M. Yeddanapudi, Y. Bar-Shalom, K. R. Pattipati, and S. Deb. Ballistic missile track initiation from satellite observations. *IEEE Transactions on Aerospace and Electronic Systems*, 31(3):1054–1071, July 1995.

B. Øksendahl. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer-Verlag, Berlin Heidelberg, 1985. ISBN 3-540-15292-X.

Appendix A

Analysis

This chapter is intended to cover some definitions and results of basic analysis, including topological spaces, abstract measures and general integration. The results are needed in the thesis, e.g., in the definitions of a probability space and a random element, and in construction of the theory of random sets. This chapter is based mostly on the book [Gariepy and Ziemer 1995]. Some of the results can also be found on [Royden 1989]. The chapter is not intended to be an "introduction to analysis", thus the reader is advised to read an introductory book of basic analysis, for example [Rudin 1976]. For deeper understanding of measure theory and integration, the book of Gariepy and Ziemer [1995] is recommended.

A.1 Topology

The definition of an abstract topological space is given first. A topology defines the open and the closed sets in a space.

Definition A.1 *A* ***topological space*** $(X, \mathcal{T})$ *consists of a nonempty set* X *and a collection* $\mathcal{T}$ *of subsets of* X*, that has the following properties*

1. $\emptyset \in \mathcal{T}$ *and* $X \in \mathcal{T}$*.*
2. *If* $\mathcal{S} \subset \mathcal{T}$*, then* $\bigcup\{U : U \in \mathcal{S}\} \in \mathcal{T}$*.*
3. *If* $\mathcal{S} \subset \mathcal{T}$ *and* $\mathcal{S}$ *is finite, then* $\bigcap\{U : U \in \mathcal{S}\} \in \mathcal{T}$*.*

The sets in $\mathcal{T}$ *are called the* ***open sets*** *and their complements are the* ***closed sets****.*

Frequently, the closure of an arbitrary set $A \subset X$ is needed.

Definition A.2 *The* ***closure*** *of a set* A*, denoted by* $\overline{A}$*, is defined as follows*

$$\overline{A} \triangleq \{x \in X : U \cap A \neq \emptyset \text{ for each open set } U \text{ containing } x\}$$

The two most "extreme" topologies that one can define are given in the following example.

Example A.3 The **discrete topology** on X is such that all subsets of X are open. That is, $\mathcal{T} = \mathcal{P}(X)$. The **mini topology** on X is $\mathcal{T} = \{\emptyset, X\}$. ◊

Most of the interesting topologies lie in between these two. However, the discrete topology is often used, when X is finite.

A.1.1 Basis and Product Topology

It is often convenient to describe a topological space using a smaller collection of sets, the basis [Kaleva 2003].

Definition A.4 *A collection of sets $\mathcal{D} \subset \mathcal{T}$ is a **basis** for the topological space, if each $A \in \mathcal{T}$ can be represented as a union of basis sets, $A = \bigcup_{i \in I} D_i$, where $D_i \in \mathcal{D}$.*

Sometimes, even a smaller collection of sets, the subbase, is considered.

Definition A.5 *A **subbase** of a topology $\mathcal{S} \subset \mathcal{T}$ is such a collection of sets, that the family of all finite intersections of the members of the subbase forms a basis for a topological space.*

Every nonempty family $\mathcal{S}$ of subsets of a space X is a subbase for a topology [Gariepy and Ziemer 1995]. The relationship between the subbase, the basis, and the topology can be summarised as follows.

$$\text{subbase} \xrightarrow{\text{finite } \bigcap} \text{basis} \xrightarrow{\text{arbitrary } \bigcup} \text{topology}$$

Sometimes, we deal with Cartesian products of two (or more) topological spaces. The definition of the product topology is given as follows [Royden 1989, p. 184].

Definition A.6 *If $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ are two topological spaces, their **product topology** $(X \times Y, \mathcal{T}_{X \times Y})$ has the base*

$$U_1 \times U_2, \quad \text{where} \quad U_1 \in \mathcal{T}_X,\ U_2 \in \mathcal{T}_Y$$

A.1.2 Borel Sets and σ-algebra

The concept of a σ-algebra[1] is needed for the definition of the Borel sets. In general, σ-algebras are necessary in measure theory, thus also in probability theory.

Definition A.7 *A **σ-algebra** $\mathcal{M}$ on some set X is such a collection of subsets of X that the following properties are satisfied.*

1. *$X \in \mathcal{M}$.*
2. *If $A \in \mathcal{M}$, then $\complement A \in \mathcal{M}$.*
3. *If $A_i \in \mathcal{M}$ for $i = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{M}$.*

1. The term σ-field is also used.

It is clear, from the definition, that a σ-algebra is a collection of sets, that is closed under countable intersection and union, and contains the whole set X and the empty set $\emptyset$. It is sometimes useful to consider a σ-algebra, that contains some (arbitrary) collection of subsets of the space.

Definition A.8 *The σ-algebra **generated** by $\mathcal{G} \subset \mathcal{P}(X)$, denoted $\sigma(\mathcal{G})$, is the smallest σ-algebra, that contains all the sets in the collection $\mathcal{G}$.*

Such a σ-algebra always exists, since $\mathcal{P}(X)$ is a σ-algebra. The smallest σ-algebra can be obtained by intersection of all the σ-algebras that contain the required sets. The next definition gives a natural σ-algebra in a topological space.

Definition A.9 *In a topological space X, the smallest σ-algebra that is generated by the collection of closed sets $\mathcal{F}$, is called the **Borel sets**, and denoted by $\mathcal{B}(X) = \sigma(\mathcal{F})$.*

Clearly, the open and the closed sets of X are Borel sets, as well as their complements, countable unions, and intersections.

In the most typical topological space in engineering, $\mathbb{R}$ endowed with the common topology, the Borel sets are all the closed and the open intervals, their complements, intersections, and unions. Therefore, an example of a *non-Borel* set in this topology is quite hard to imagine, but there exists such [Gariepy and Ziemer 1995]. It is obvious, though, that one cannot construct one using the "standard" set operations on the "standard" sets, since the Borel sets include just those.

A.1.3 Certain Types of Topological Spaces

Let us define some more concepts, which are related to topological spaces, starting with a Hausdorff space and a locally compact space [Gariepy and Ziemer 1995; Royden 1989].

Definition A.10 *A topological space $(X, \mathcal{T})$ is said to be a **Hausdorff space**, if for any $x, y \in X$, $x \neq y$ there are disjoint[2] open sets $U_x, U_y \in \mathcal{T}$ containing x and y.*

The convenience of Hausdorff spaces may become clear by noticing that in any Hausdorff space, the limit point of a sequence is unique, if it exists. In addition, any singleton[3] in a Hausdorff space is closed. [Kaleva 2003]

Next, the concept of compact sets is introduced. Compact sets are convenient in metric spaces, because they contain the limit points of the Cauchy sequences. Metric spaces and Cauchy sequences are introduced in Section A.2.

Definition A.11 *Let $(X, \mathcal{T})$ be a topological space. A collection of open sets $\mathcal{G} \subset \mathcal{T}$ is an **open cover** for a set $A \subset X$, if A is covered by $\mathcal{G}$, i.e. $A \subset \bigcup\{G :$*

2. U_x and U_y are disjoint, if $U_x \cap U_y = \emptyset$.
3. Singleton is a set with exactly one element.

$G \in \mathcal{G}\}$. *A set* $K \in \mathcal{T}$ *is* ***compact****, if every open cover of* K *has a finite subcover* $\mathcal{H} \subset \mathcal{G}$.

Definition A.12 *A topological space* X *is* ***locally compact****, if for each* $x \in X$ *there is such an open set* $U \in \mathcal{T}$ *that* $x \in U$ *and the closure* $\overline{U}$ *is compact.*

Definition A.13 *A set* Y *is* ***dense*** *in* X*, if* $\overline{Y} = X$*. The space* X *is* ***separable****, if there exists a countable dense set in* X.

A.2 Metric Spaces

Definition A.14 *A set* X *is said to be a* ***metric space****, if there is a function* $d : X \times X \to \mathbb{R}$ *that has the following properties for all* $x, y, z \in X$

1. $d(x, y) > 0$ *if* $x \neq y$*, and* $d(x, x) = 0$.
2. $d(x, y) = d(y, x)$.
3. $d(x, z) \leq d(x, y) + d(y, z)$.

A metric space is often denoted by (X, d).

The function d above is often referred to as a *metric* or a *distance* function. Using the distance function, one can define an open ball as follows [Gariepy and Ziemer 1995].

Definition A.15 *An* ***open ball*** *in a metric space* (X, d) *is defined as follows*

$$B_x(r) \triangleq \{y \in X : d(x, y) < r\}$$

The point $x \in X$ *is the centre, and* $r > 0$ *is the radius of the ball.*

Having defined an open ball, one may define a topology to any metric space. Such a topology is defined next [Gariepy and Ziemer 1995].

Definition A.16 *The topology* ***induced by a metric*** *is one having the subbase*

$$S = \{B_x(r) : x \in X, r > 0\}$$

Next, two examples of metric spaces are given. These spaces are, in fact, used in the application part of this thesis.

Example A.17 The **Euclidean space** $\mathbb{R}^d$ endowed with the Euclidean norm metric

$$d(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\| = \sqrt{(\underline{x} - \underline{y})^T(\underline{x} - \underline{y})}$$

is a metric space. ◊

The topological space $\mathbb{R}^d$ induced by the Euclidean norm is the "standard" topology in $\mathbb{R}^d$. The open sets in $\mathbb{R}^d$ are those that are considered open in elementary calculus.

Example A.18 A **finite space** D endowed with the discrete metric

$$d(x,y) = \delta_x(y) = \begin{cases} 0, & y = x \\ 1, & y \neq x \end{cases}$$

is a metric space. ◊

The topology of a finite space D induced by the discrete metric is the discrete topology, given in Example A.3.

In metric spaces, convergence of sequences can be characterised conveniently using the metric d. There are some concepts, that are related to convergence in metric spaces.

Definition A.19 *Let $\{x_i\}_{i=1}^{\infty}$ be a sequence in X. The sequence **converges** to $x \in X$, denoted as $\lim_{i\to\infty} x_i = x$, if for each $\epsilon > 0$, there exists a $m \in \mathbb{N}$, such that $d(x, x_i) < \epsilon$ for all $i > m$.*

*The sequence is a **Cauchy sequence**, if for each $\epsilon > 0$, there is a $m \in \mathbb{N}$ such that for all $i, j > m$, $d(x_i, x_j) < \epsilon$. The metric space (X, d) is **complete**, if every Cauchy sequence converges in X.*

It is rather easy to prove that $\mathbb{R}$ equipped with the Euclidean norm metric is a complete metric space. On the other hand, $\mathbb{R} \setminus \{0\}$ is not complete, since the Cauchy sequence $\{1/i\}_{i=1}^{\infty}$ does not converge in $\mathbb{R} \setminus \{0\}$.

A.3 Functions

In the following definitions, it is assumed that $f : X \to Y$ is a function, i.e. for each $x \in X$ the function relates a unique value $y \in Y$, which is denoted as $y = f(x)$.

Definition A.20 *The **preimage** of a function f is defined for all sets $A \subset Y$ as follows*

$$\overleftarrow{f}(A) \triangleq \{x \in X : f(x) \in A\}$$

Definition A.21 *The **range** of a function, denoted as* $\mathrm{rng}(f)$*, is defined as*

$$\mathrm{rng}(f) \triangleq \{y \in Y : f(x) = y, \text{ for some } x \in X\}$$

There are some special types of functions, which are introduced next

Definition A.22 *The **characteristic function** of a set $A \subset X$ is defined as follows*

$$\chi_A(x) \triangleq \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

If the range of a function $X \to \overline{\mathbb{R}}$ is finite, then the function is called **simple**. A simple function can be represented as a weighted finite sum of characteristic functions. The next definition covers also the simple functions.

Definition A.23 *If the range of a function $f : X \to \overline{\mathbb{R}}$ is countable, then the function is called **countably simple**. A countably simple function can be given using characteristic functions as follows.*

$$f(x) = \sum_{a \in \mathrm{rng}(f)} a\chi_{\overleftarrow{f}(\{a\})}(x)$$

Sometimes, it is necessary to work separately with the positive and the negative parts of an extended real valued function $f : X \to \overline{\mathbb{R}}$. The negative and positive parts are denoted as follows

$$f^+(x) = \begin{cases} f(x), & f(x) > 0 \\ 0, & f(x) \leq 0 \end{cases} \qquad f^-(x) = \begin{cases} -f(x), & f(x) < 0 \\ 0, & f(x) \geq 0 \end{cases}$$

The original function can be reconstructed as $f(x) = f^+(x) - f^-(x)$.

Finally, a function that relates two topological spaces with each other is introduced.

Definition A.24 *Consider two topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$. A function $f : X \to Y$ is **continuous**, if $\overleftarrow{f}(V)$ is open in X for each open set V in Y. A bijective function $f : X \to Y$ is a **homeomorphism**, if both the function f and its inverse f^{-1} are continuous..*

Clearly, a homeomorphism relates the topological structure of X and Y to each other. That is, if there is a homeomorphism $f : X \to Y$, then X and Y can be considered topologically invariant, and are said to be **homeomorphic**.

A.4 Measure Theory

This section introduces very briefly the concepts of a measure, a measure space, and a measurable function. All the definitions are given in a general form.

Definition A.25 *The triple $(X, \mathcal{M}, \mu)$ is a **measure space**, if $\mathcal{M}$ is a σ-algebra on a set X, and $\mu : \mathcal{M} \to [0, \infty]$ is a **measure** on $\mathcal{M}$, which has the following properties.*

1. *$\mu(\emptyset) = 0$.*
2. *If $\{E_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets, $E_i \cap E_j = \emptyset$ whenever $i \neq j$, then*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

The sets $E \in \mathcal{M}$ are referred to as the ***measurable sets****.*

The property 2 above is referred to as *countable additivity*. Sometimes, the measure is omitted, and one refers to a **measurable space** $(X, \mathcal{M})$. In probability theory, restricted classes of measures are often required.

Definition A.26 *A measure μ on a space X is σ-**finite**, if there exists such sets $E_i \in \mathcal{M}$ that $\bigcup_{i=1}^{\infty} E_i = X$ and $\mu(E_i) < \infty$ for all i. The measure is **finite**, if $\mu(X) < \infty$.*

A.4.1 Properties of Measures

If there are two measures μ_1 and μ_2 defined on a same σ-algebra $\mathcal{M}$, then one may define absolute continuity of the measures as follows.

Definition A.27 *A measure μ_1 is **absolutely continuous** with respect to a measure μ_2, denoted $\mu_1 \ll \mu_2$, if for all $E \in \mathcal{M}$ we have $\mu_2(E) = 0 \implies \mu_1(E) = 0$.*

Often, the concept of a null set[4], i.e. a set of measure zero, is encountered. This is due to the fact that null sets do not affect the value of an integral, which is defined in Section A.5. The following definition clarifies the concepts.

Definition A.28 *A set $N \in \mathcal{M}$ is a μ-**null** set, if $\mu(N) = 0$. A condition is said to hold μ-**almost surely**, (μ-a.s.), or μ-**almost everywhere** (μ-a.e.), if the condition holds everywhere except in a μ-null set.*

Often, when the measure is clear from the context, a set may be referred to as a null set, and some condition is just said to hold a.s. or a.e.

The definition of a measure restricted the range to be $[0, \infty]$. Sometimes, the concept of a signed measure comes handy. We give the definition of a signed measure as follows [Shiryaev 1996, p. 196].

Definition A.29 *The set function $\nu = \mu_1 - \mu_2$ is a **signed measure**, if at least one of the measures μ_1 and μ_2 is finite.*

Signed measures have similar properties to (positive) measures. In fact, the book of Gariepy and Ziemer [1995] defines a signed measure through the following properties.

1. The range of ν does not contain both $-\infty$ and ∞.
2. $\nu(\emptyset) = 0$.
3. If $\{E_i\}_{i=1}^{\infty}$ is a disjoint sequence of measurable sets, then

$$\nu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \nu(E_i)$$

4. In this thesis, the measure spaces are required to be complete, i.e. every subset of a null set is assumed measurable.

A.4.2 Common Measures

The most common measure one encounters is the Lebesgue measure, defined in $\mathbb{R}^d$ (endowed with the usual topology). The measure is defined indirectly via the volume of a hypercube [Gariepy and Ziemer 1995].

Definition A.30 *The **volume** of a closed interval (hypercube) $I = \times_{i=1}^{d}[a_i, b_i]$ in $\mathbb{R}^d$ is defined as*

$$v(I) = \prod_{i=1}^{d}(b_i - a_i)$$

*The **Lebesgue measure** of a Lebesgue measurable set $E \in \mathcal{M}$ is defined as*

$$\lambda(E) = \inf_{\mathcal{S}} \sum_{I \in \mathcal{S}} v(I), \quad \text{where } \mathcal{S} \text{ is countable and } \quad E \subset \bigcup_{I \in \mathcal{S}} I$$

In this thesis, one needs to know that all the Borel sets in $\mathbb{R}^d$ are Lebesgue measurable. There exists, however, non-Borel sets that are Lebesgue measurable. The intuition telling that the Lebesgue measure is the "volume" of a set in $\mathbb{R}^d$ is the most important.

Another example of a common measure is the so called counting measure [Royden 1989, p. 55]. The counting measure can be defined to any space, but it is most natural to define in a finite or a countable space.

Definition A.31 *The **counting measure** $c : \mathcal{P}(X) \to [0, \infty]$ on a set X is defined as the number of elements, i.e. the cardinality, of a set.*

$$c(A) \triangleq |A|$$

If A is infinite (countable or uncountable), the value of c is infinity.

A.4.3 Measurable Functions

A measurable function can be considered a "well-behaved" function in the sense of measures. The definition of a measurable function can be given as follows

Definition A.32 *Let $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$ be two measure spaces. A function $f : X \to Y$ is **measurable** with respect to $\mathcal{M}$, or $\mathcal{M}$-measurable, if $\overleftarrow{f}(E) \in \mathcal{M}$ whenever $E \in \mathcal{N}$.*

One sees, that a measurable function transfers a measure μ on $\mathcal{M}$ to a measure on $\mathcal{N}$. This is essential in the definition of a random element.

A useful notation that is related to functions in general, but most importantly to measurable functions, is the σ-algebra induced by a function.

Definition A.33 *Let $(X, \mathcal{M})$ and $(Y, \mathcal{N})$ be measurable spaces. The σ-algebra **induced** by a function $f : X \to Y$, is the smallest σ-algebra containing all the preimages of measurable sets in Y.*

$$\sigma(f) \triangleq \sigma\left(\{\overleftarrow{f}(N) : N \in \mathcal{N}\}\right)$$

Clearly $\sigma(f) \subset \mathcal{M}$, if f is $\mathcal{M}$-measurable.

A.4.4 Some Measure Spaces

The above definitions of a measure and a measurable function are very general. A common special case of a measure space is the extended real numbers $Y = \overline{\mathbb{R}}$.

Definition A.34 *The space* $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ *is the* ***extended real numbers****. In this space, the addition and product operations are defined as follows*

$$\begin{array}{ccc} a \pm \infty = \pm\infty & a/(\pm\infty) = 0 & b \cdot (\pm\infty) = \pm\infty \\ (\pm\infty) + (\pm\infty) = \pm\infty & (\pm\infty)/(\pm\infty) = \textit{undef.} & (-b) \cdot (\pm\infty) = \mp\infty \\ (\pm\infty) + (\mp\infty) = \textit{undef.} & (\pm\infty)/(\mp\infty) = \textit{undef.} & 0 \cdot \pm\infty = 0 \end{array}$$

for all $a \in \mathbb{R}$*, and* $b \in (0, \infty]$*.*

The order topology having a subbase consisting of the sets $[-\infty, a)$ and $(b, \infty]$ where $a, b \in \mathbb{R}$ is always used with $\overline{\mathbb{R}}$. The Borel sets in this topology are

$$\mathcal{B}(\overline{\mathbb{R}}) = \{B, B \cup \{\infty\}, B \cup \{-\infty\}, B \cup \{\infty\} \cup \{-\infty\} : B \in \mathcal{B}(\mathbb{R})\}$$

where $\mathcal{B}(\mathbb{R})$ are the Borel sets in $\mathbb{R}$ according to the Euclidean topology. In the case $Y = \overline{\mathbb{R}}$, a measurable function $f : X \to Y$ is called **Borel-measurable**, and if also $X = \overline{\mathbb{R}}$, the function is a **Borel function**.

A more general class of measurable spaces are the Borel spaces [Shiryaev 1996, p. 229].

Definition A.35 *A measurable space* $(X, \mathcal{N})$ *is a* ***Borel space****, if it is Borel equivalent to a Borel subset of the real line. That is, there exists an injective function* $\varphi : X \to \mathbb{R}$ *such that*

1. $\varphi(X) \equiv \{\varphi(x) : x \in X\} \in \mathcal{B}(\mathbb{R})$
2. φ *is* $\mathcal{N}$*-measurable, i.e.* $\overleftarrow{\varphi}(A) \in \mathcal{N}$ *whenever* $A \in \mathcal{B}(\mathbb{R})$*.*
3. $\overleftarrow{\varphi}$ *is* $\mathcal{B}(\mathbb{R})$*-measurable, i.e.* $\varphi(B) \in \mathcal{B}(\mathbb{R})$ *whenever* $B \in \mathcal{N}$*.*

At last we state without proof that complete and separable metric spaces, and in particular $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, and $(R^\infty, \mathcal{B}(\mathbb{R}^\infty))$ are Borel spaces [Shiryaev 1996, p. 230].

A.5 Integration

This section defines the concept of an integral in a general sense. The Riemann integral that is considered most of the time in engineering is sufficient for most purposes. There are some applications, however, in which more general integrals are needed. In addition, general integration covering both the discrete-valued integration (summation), and the continuous-valued cases is theoretically handy.

A.5.1 Definitions

A countably simple function can be considered a "piecewise constant" function. The integral of a nonnegative countably simple measurable (CSM) function is defined quite naturally, as follows.

Definition A.36 *If $f : X \to [0, \infty]$ is a CSM nonnegative function, the integral of the function is defined as follows.*

$$\int f(x)\mathrm{d}\mu(x) \triangleq \sum_{i=1}^{\infty} a_i \mu \overleftarrow{f}(\{a_i\})$$

Notice, that the value of the sum may be infinite. Often, the variable that is integrated is omitted in notation, for example $\int f(x)\mathrm{d}\mu(x)$ is denoted frequently as $\int f\mathrm{d}\mu$. Next, the definition is extended to cover any CSM function.

Definition A.37 *Let $f : X \to \overline{\mathbb{R}}$ be a CSM function. Then, if either $\int f^+\mathrm{d}\mu$ or $\int f^-\mathrm{d}\mu$ is finite, the integral of f is defined to be*

$$\int f\mathrm{d}\mu \triangleq \int f^+\mathrm{d}\mu - \int f^-\mathrm{d}\mu$$

The above definitions can be considered intuitive, and straightforward. The integral of a measurable function, if it exists, is defined using the integrals of countably simple functions.

Definition A.38 *Let $f : X \to \overline{\mathbb{R}}$ be a measurable function. Then, the upper integral of f is defined as follows*

$$\int^* f\mathrm{d}\mu \triangleq \inf\left\{\int g\mathrm{d}\mu : g \text{ is CSM, and } g \overset{a.s.}{\geq} f\right\}$$

and the lower integral as follows

$$\int_* f\mathrm{d}\mu \triangleq \sup\left\{\int g\mathrm{d}\mu : g \text{ is CSM, and } g \overset{a.s.}{\leq} f\right\}$$

*If the values of the upper and lower integrals agree, and the common value is finite, the function $f(x)$ is **integrable** (with respect to μ), and the **integral** of the function is*

$$\int f\mathrm{d}\mu \triangleq \int^* f\mathrm{d}\mu = \int_* f\mathrm{d}\mu$$

A shorthand notation, that is commonly used, is the integral over a measurable set $A \subset X$, and it is defined as follows

$$\int_A f\mathrm{d}\mu \triangleq \int \chi_A f\mathrm{d}\mu \tag{A.1}$$

A.5.2 General Properties

Here, we list some useful properties of a general integral. The results are given without proof. Consult, e.g., Gariepy and Ziemer [1995] for proofs and further discussion.

Theorem A.39 *At first, we notice that if f is a measurable function, and $|f|$ is integrable, then f is integrable. Suppose that f and g are integrable functions. Then, the following statements are true.*

1. *If $a, b \in \mathbb{R}$ are constants, then $\int af + bg\mathrm{d}\mu = a\int f\mathrm{d}\mu + b\int g\mathrm{d}\mu$.*
2. *If $f(x) \leq g(x)$ μ-a.e., then $\int f\mathrm{d}\mu \leq \int g\mathrm{d}\mu$.*
3. *If E is measurable, then $f\chi_E$ is integrable.*
4. *$\left|\int f\mathrm{d}\mu\right| \leq \int |f|\mathrm{d}\mu$.*

Theorem A.40 (Monotone Convergence Theorem) *Let $\{f_k\}_{k=1}^{\infty}$ be a sequence of nonnegative measurable functions, such that $f_k \leq f_{k+1}$ for all k. Then,*

$$\lim_{k\to\infty}\int f_k\mathrm{d}\mu = \int \lim_{k\to\infty} f_k\mathrm{d}\mu$$

Suppose $(X, \mathcal{M}_X, \mu)$ and $(Y, \mathcal{M}_Y, \nu)$ are two measure spaces. An important question may arise, that how one can construct a measure to $X \times Y$ from the measures μ and ν. Again, we do given the construction of the measure, but refer to [Gariepy and Ziemer 1995] that such a measure can be constructed. The key properties of the product measure space can be found within the following theorem.

Theorem A.41 (Fubini) *Suppose that $(X, \mathcal{M}_X, \mu)$ and $(Y, \mathcal{M}_Y, \nu)$ are measure spaces, and that $(X \times Y, \mathcal{M}_{X\times Y}, \mu \times \nu)$ is their **product measure space**. The product measure $(\mu \times \nu)$ has the following properties:*

1. *If $A \in \mathcal{M}_X$ and $B \in \mathcal{M}_Y$, then $A \times B \in \mathcal{M}_{X\times Y}$.*
2. *$(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$.*
3. *If $S \in \mathcal{M}_{X\times Y}$ is σ-finite with respect to $\mu \times \nu$, then the following holds*
   $S_y = \{x : (x, y) \in S\} \in \mathcal{M}_X$, for ν-a.e. $y \in Y$.
   $S_x = \{y : (x, y) \in S\} \in \mathcal{M}_Y$, for μ-a.e. $x \in X$.

$$(\mu \times \nu)(S) = \int_X \nu(S_x)\mathrm{d}\mu(x) = \int_Y \mu(S_y)\mathrm{d}\nu(y)$$

4. *Let f be a $\mu \times \nu$-integrable function. Then,*

$$\begin{aligned}\int_{X\times Y} f(x, y)\mathrm{d}(\mu \times \nu)(x, y) &= \int_X \left[\int_Y f(x, y)\mathrm{d}\nu(y)\right]\mathrm{d}\mu(x)\\ &= \int_Y \left[\int_X f(x, y)\mathrm{d}\mu(x)\right]\mathrm{d}\nu(y)\end{aligned}$$

Fubini's theorem states that the integration order can be changed in an iterated integral. In addition, of course, it states that the "partial results" are defined, and integrable. For example, $g(y) = f(x, y)$ is ν-integrable for μ-a.e. x, and $h(x) = \int_Y f(x, y)\mathrm{d}\nu(y)$ is μ-integrable.

A.5.3 Relationship with Riemann Integral and Sum

The Lebesgue integral agrees with the elementary Riemann integral, whenever a function is Riemann-integrable. That is, if $f(x)$ is Riemann integrable, then it is also Lebesgue integrable with the same integral value,

$$\int f(x)\mathrm{d}\lambda(x) = \int f(x)\mathrm{d}x \tag{A.2}$$

Notice, however, that for some function there may exist an *improper* Riemann integral, but not the Lebesgue integral. For additional information of "Riemann-like" integrals, and the Lebesgue integral, see the article [Bartle 1996].

In addition, the integral over a finite space endowed with the counting measure reduces to an ordinary sum over the elements in the space. This means that if X is finite, then

$$\int f(x)\mathrm{d}c = \sum_{x\in X} f(x) \tag{A.3}$$

Obviously, if X is countable, the finite sum above is replaced by an infinite one.

A.6 Differentiation

This section introduces the concept of a Radon-Nikodym derivative (RND). RND is a generalisation of the elementary derivative, similarly as the general integral can be considered a generalisation of the Riemann-integral. RND allows to represent probability measures with respect to some other measures using their densities. The definition is given according to [Shiryaev 1996, p. 196].

Theorem A.42 (Radon-Nikodym) *Let $(X, \mathcal{M})$ be a measurable space, with a σ-finite measure μ. Let ν be a signed measure on $(X, \mathcal{M})$, such that $\nu \ll \mu$. Then, there exists such a $\mathcal{M}$-measurable function $f : X \to \overline{\mathbb{R}}$ that*

$$\nu(A) = \int_A f\mathrm{d}\mu$$

for all $A \in \mathcal{M}$.

The function f, denoted by $\mathrm{d}\nu/\mathrm{d}\mu$, is the **Radon-Nikodym derivative** of ν with respect to μ. If ν is a positive measure, then the function f can be selected so that $f(x) \geq 0$ for all $x \in X$.

The next theorem gives some convenient properties of RND [Shiryaev 1996, p. 231].

Theorem A.43 *Let μ and ν be σ-finite measures, $\mu \ll \nu$, and f a measurable function. Then,*

$$\int f\mathrm{d}\mu = \int f\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\nu \tag{A.4}$$

If, in addition, $\xi \ll \mu$ is a signed measure, then

$$\frac{\mathrm{d}\xi}{\mathrm{d}\nu} = \frac{\mathrm{d}\xi}{\mathrm{d}\mu} \cdot \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \qquad (\nu\text{-a.s.}) \tag{A.5}$$

$$\frac{\mathrm{d}\xi}{\mathrm{d}\mu} = \frac{\mathrm{d}\xi/\mathrm{d}\nu}{\mathrm{d}\mu/\mathrm{d}\nu} \qquad (\mu\text{-a.s.}) \tag{A.6}$$

Proof. By definition of RND, Equation (A.4) holds for $f = \chi_A$, thus also for nonnegative countable simple functions. The general case follows from the decomposition $f = f^+ - f^-$, the fact that any nonnegative function can be approximated by an increasing sequence of countably simple functions, and the Monotone Convergence Theorem (A.40).

Equation (A.5) follows from Equation (A.4) by letting $f = \mathrm{d}\xi/\mathrm{d}\mu$, and the observation that $\xi \ll \nu$. Then,

$$\int_A \frac{\mathrm{d}\xi}{\mathrm{d}\mu} \cdot \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\nu = \nu(A) = \int_A \frac{\mathrm{d}\xi}{\mathrm{d}\nu}\mathrm{d}\nu$$

for arbitrary measurable set A. Finally, Equation (A.6) follows directly from Equation (A.5), by the observation that the set $\{\mathrm{d}\mu/\mathrm{d}\nu = 0\}$ is μ-null. □

Remark A.44 If one constructs a measure on $\mathbb{R}$ using a function $f : \mathbb{R} \to [0, \infty)$ with the property $\int f\mathrm{d}\lambda = 1$, so that $P(A) = \int_A f\mathrm{d}\lambda$, then the RND of the measure agrees with the elementary derivative $\mathrm{d}P/\mathrm{d}\lambda = \mathrm{d}F/\mathrm{d}x = f$, where $F(x) = P((-\infty, x])$ is the cumulative distribution function.

Appendix B

Bayesian Networks

Bayesian networks (BNs) provide a method for representing probabilistic models graphically. Most importantly, a BN graph contains information on the assumed dependencies (or equivalently, the independence assumptions) between the random elements in a model. In many applications, this is the most interesting information one needs to know about a model.

Often the concept of Bayesian networks is restricted into certain special cases. The most common is to consider only discrete valued random variables [e.g. Jensen 2001], while some authors assume conditional Gaussian (CG) distributions, allowing both discrete and conditionally multivariate Gaussian distributed random variables [Cowell et al. 2003; Lauritzen 2002]. The restriction introduces some simplifications to the inference theory. In this thesis, the primary purpose of the Bayesian networks is to provide a concise graphical representation, so inference need not be considered explicitly. An interested reader can find more information on Bayesian networks, for example, in the following references:

- Jensen [2001] is an introductory book, that may be considered as a "first course" in BNs. The book is restricted into discrete-valued random variables.
- Murphy [2003] provides a good tutorial to dynamic Bayesian networks.
- Murphy [2002] provides an extensive survey of theory of dynamic BNs, and a comprehensive list of references.

B.1 Definitions

Before the definition of a Bayesian network (in the sense of this thesis) is given, some observations are needed. Suppose $(\Omega, \mathcal{M}, P)$ is a probability space. Assume that there is a finite number of random variables $\{\boldsymbol{x}_i\}_{k=1}^n$ in measurable spaces $(S_i, \mathcal{N}_i)$. One can define $\mathcal{G}_k = \sigma(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ for $k = 1, \ldots, n$. Then, it is obvious that $\mathcal{G}_1 \subset \cdots \subset \mathcal{G}_n \subset \mathcal{M}$. The following conditional probability can be defined for all $B_k \in \mathcal{G}_k$.

$$P(B_k \mid \mathcal{G}_{k-1})(\omega) = \mathsf{E}\left[\chi_{B_k} \mid \mathcal{G}_{k-1}\right](\omega) = \frac{\mathrm{d}Q}{\mathrm{d}P}(\omega)$$

where $Q(A) = P(A \cap B_k)$. The RNDs exists, since Q is absolutely continuous with respect to P (restricted to the σ-algebra $\mathcal{G}_{k-1}$). Assuming all the conditional probabilities regular, one sees that [Stratonovich 1968, pp. 306–307].

$$P(B_k) = \int \cdots \int P(B_k \mid \mathcal{G}_{k-1})(\omega_k) \mathrm{d}P(\omega_k \mid \mathcal{G}_{k-2})(\omega_{k-1}) \cdots \mathrm{d}P(\omega_1)$$

Alternatively, one can see that the above implies the following form.

$$\begin{aligned} &P\left(\boldsymbol{x}_1 \in A_1, \ldots, \boldsymbol{x}_n \in A_n\right) \\ &= \int_{A_1} \cdots \int_{A_{n-1}} P(A_n \mid \boldsymbol{x}_{1:n-1} = x_{1:n-1}) \mathrm{d}P(x_{n-1} \mid \boldsymbol{x}_{1:n-2} = x_{1:n-2}) \cdots \mathrm{d}P(x_1) \end{aligned} \tag{B.1}$$

If the random elements have a joint density function $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(x_1, \ldots, x_n)$ with respect to a σ-finite product measure in space $\prod_{k=1}^{\infty} S_k$, then the decomposition can be given in terms of density functions.

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(x_1, \ldots, x_n) = f_{\boldsymbol{x}_1}(x_1) \prod_{k=2}^{n} f_{\boldsymbol{x}_k \mid \boldsymbol{x}_{1:k-1}}(x_k \mid x_{1:k-1}) \tag{B.2}$$

where the conditional densities $f_{\boldsymbol{x}_k \mid \boldsymbol{x}_{1:k-1}}$ are given analogously to Equation (2.6),

$$f_{\boldsymbol{x}_k \mid \boldsymbol{x}_{1:k-1}}(x_k \mid x_{1:k-1}) = \begin{cases} \dfrac{f_{\boldsymbol{x}_{1:k}}(x_{1:k})}{f_{\boldsymbol{x}_{1:k-1}}(x_{1:k-1})}, & f_{\boldsymbol{x}_{1:k-1}}(x_{1:k-1}) > 0 \\ 0, & f_{\boldsymbol{x}_{1:k-1}}(x_{1:k-1}) = 0 \end{cases}$$

What was pointed out is that the joint behaviour of the random elements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ can be characterised by a set of regular conditional probabilities. Now we are ready to proceed to the definition of a Bayesian network.

Definition B.1 (Bayesian network) *The definition is given in two parts: first the graph theoretic definition, and then the interpretation of the graph.*

1. *A Bayesian network (BN) is a directed acyclic*[1] *graph (V, E), consisting of a set of vertices V and set of directed edges E. Each directed edge $(u, v) \in E$ connects one vertex $u \in V$ to another $v \in V$.*
2. *Each vertex $v \in V$ in a BN graph corresponds to one random element. The edges correspond to the dependencies between the random variables, so that according to the conditional distributions given in Equation (B.1), there is an edge from each $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}$ to $\boldsymbol{x}_k$. If there is a (conditional) independence of random elements, the corresponding edges are omitted from the BN graph.*

The above definition is rather informal. The idea of the BN representation will perhaps become clear via a couple of examples.

1. A graph is cyclic, if one can move along the directed edges (in the direction they are pointing) infinitely. An acyclic graph is one that is not cyclic.

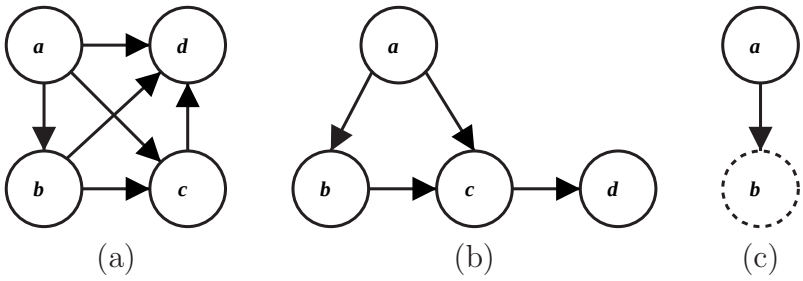


Figure B.1: Examples of Bayesian networks. In (a) and (b), $V = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}\}$. (a) A BN with no independence assumptions. (b) A BN with one independence assumption. (c) A BN with emphasised (dotted border) observation variable $\boldsymbol{b}$.

Example B.2 Consider first the general case with no independence assumptions. The BN graph is shown in such a case for four random elements in Figure B.1 (a). In the situation when there are no independence assumptions, the BN graph is "full", i.e. there is an edge between every two nodes. It is worth noticing that the Bayesian network graph is *not unique*. This is due to the fact that one can perform the factorisation in Equation (B.1) in an arbitrary order, each of which results in a different BN graph.

The second example in Figure B.1 (b) is modified from the first example so that there is one independence assumption included, which leads into omitting two edges. The assumption is that

$$P(\boldsymbol{d} \in D \mid \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = P(\boldsymbol{d} \in D \mid \boldsymbol{c})$$

Usually, the BN graphs are even more sparse than the one in this example.

Finally, one extreme case is that all the random elements in the Bayesian network are assumed independent. Then, by definition, their conditional probabilities can be given without dependencies, and the BN graph contains no edges at all. In the case of the four random elements in Figure B.1, and assuming that the joint distribution admits a density, the density can be decomposed as follows

$$f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d}}(a, b, c, d) = f_{\boldsymbol{a}}(a) f_{\boldsymbol{b}}(b) f_{\boldsymbol{c}}(c) f_{\boldsymbol{d}}(d)$$

Of course, such a BN graph with no edges is not too interesting either. The usual case is that the BN is connected, i.e. there are no unconditionally independent random elements. ◊

Sometimes, it is convenient to emphasise such a random element in a BN, that is observed in estimation. That is, consider the Bayesian estimation problem, where $\boldsymbol{a}$ is the unknown, unobserved parameter, and $\boldsymbol{b}$ is the noisy, indirect measurement of $\boldsymbol{a}$. It is convenient to emphasise this setting already in the BN graph. In Figure B.1 (c), this situation is shown, and the observation variable is denoted so that the node border is dotted. This is the notation that is used in this thesis to emphasise observed variables.

Remark B.3 Definition B.1 restricted the BN to contain a *finite* number of random elements. Frequently, in this thesis, discrete-time stochastic processes with countably infinite number of random elements are represented as Bayesian networks. The approach is sensible, since the processes are considered in an recursive manner. That is, one "constructs on the fly" an infinite Bayesian network, so that after a finite number of recursions, the number of variables (nodes) in the Bayesian network is finite. Such Bayesian networks are called in general *dynamic Bayesian networks* [Murphy 2002].

B.2 Intuitive Interpretation

The BN graphs were stated to provide an intuitive manner of considering probabilistic models. This section gives an intuitive interpretation, which many times helps to grasp the ideas behind a model. The interpretation is to consider a BN as a *generative model*.

Consider the case of discrete-valued random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. To obtain a sample $x_1 = \boldsymbol{x}_1(\omega), \ldots, x_n = \boldsymbol{x}_n(\omega)$, one can proceed as follows.

1. Find a random variable $\boldsymbol{x}_i$ that is either a root (i.e. has no parent nodes[2]), or whose parents have already been sampled.
2. Draw a sample from the conditional distribution of $\boldsymbol{x}_i$ given the sampled values of the parent nodes.
3. If all variables have not been sampled, go to step 1.

Again, an example will perhaps clarify this rather informally given procedure. Consider the case in Figure B.1 (b). First a sample a is drawn corresponding variable $\boldsymbol{a}$. Then, a sample b corresponding $\boldsymbol{b}$ can be drawn according to $P(\boldsymbol{b} \mid \boldsymbol{a} = a)$, since the parent of $\boldsymbol{b}$, namely $\boldsymbol{a}$ has been sampled. After that, sample c corresponding $\boldsymbol{c}$ can be obtained according to $P(\boldsymbol{c} \mid \boldsymbol{a} = a, \boldsymbol{b} = b)$. Finally, $\boldsymbol{d}$ is sampled according to $P(\boldsymbol{d} \mid \boldsymbol{c} = c)$. The sample (a, b, c, d) obtained in this manner is distributed according to the joint distribution $P(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d})$.

B.3 Inference

When talking about BNs, the computation of the posterior marginal distribution for an element, or a set of elements, is referred to as inference. In brief, the inference methods for BNs are all developed for random variables with finite range. The inference algorithms are based on the idea that they find such parts in the Bayesian networks, in which computations can be performed locally. Inference in tree-based networks can be carried out using the algorithm due to Pearl [1988]. For other types of networks, the algorithms are based on a secondary graph called the *junction tree*. For information on junction tree algorithms, see e.g. [Jensen 2001; Lauritzen and Spiegelhalter 1988; Madsen and Jensen 1999].

---

2. Such nodes, from which there is a directed edge towards a node, are the parent nodes.

The Bayesian network inference algorithms can be implemented straightforwardly, if the BN consists of random variables with finite range. This is due to the fact that the inference algorithms use a meta-operation "marginalisation", which is in general integration. In the case of random variables with finite range, integration reduces into finite summation. On the other hand, analytical integration is often infeasible in the case of continuous distributions. Thus, exact inference in non-discrete BNs is in general infeasible. There exists several algorithms for performing an approximate Bayesian network inference. Murphy [2002] gives an excellent survey of the existing inference methods, and formulates a variety of probabilistic models as dynamic Bayesian networks.