Bayesian probability

Herman Bruyninckx
Dept. of Mechanical Engineering, K.U.Leuven, Belgium
http://people.mech.kuleuven.ac.be/~bruyninc

November 2002

Abstract

This document introduces the foundations of Bayesian probability theory. The emphasis is on understanding why Bayesian probability theory works, and on realizing that the theory relies, on the one hand, on a very limited number of fundamental properties for information processing, and, on the other hand, on a number of application-dependent and arbitrary choices for decision making.

1 What is Bayesian probability theory?

Bayesian probability, is one of the major theoretical and practical frameworks for reasoning and decision making under uncertainty. The historical roots of this theory lie in the late 18th, early 19th century, with Thomas Bayes [2] and Pierre-Simon de Laplace [6]. It was "forgotten" for a long time, and began to be re-appreciated in different application domains, during various periods of the 20th century. Hence, Bayesian probability was never developed as one single, homogeneous piece of scientific activity. So, it should come as no surprise that its concepts, methods and solution practices became known under various names:

- the Bayesian approach to uncertainty reasoning.
- Bayesianism.
- the Bayesian framework.
- the Bayesian paradigm.
- plausible inference.
- Bayesian reasoning.
- ...

This course speaks most often of *Bayesian probability (theory)*, mainly because its terminology borrows most from statistical probability theory. The "theory" is much more than just a theory: it's a *systematic way* of approaching an application problem in which one is confronted with incomplete knowledge about

the problem. When reading papers that use Bayesian probability one gets the impression that the core of the systematic approach is that one uses all the time the well-known laws of probability: sum rule, product rule, and Bayes' rule. However, this is only part of the whole picture, because applying Bayesian probability to an application problem also means that one takes great care to unambiguously define and motivate the *system models* and *decision criteria* that are used in the problem solution.

Message 1 (Modelling, information processing, decision making.) *This document does a big effort to separate explicitly the discussion on Bayesian probability theory into these three sub-problems:* modelling*,* information processing*, and* decision making*. And to give a clear definition of what* information *means in the context of Bayesian probability.*

1.1 Examples

This section gives some examples of applications of Bayesian probability theory:

(Parameter) estimation.

- *Police speed radar.*
- *Prediction of celestial body motions.*
- *Steering space ships to Jupiter.*

Pattern matching/Hypothesis testing.

- *Detection of tumor in a scan image.*
- *Detection of gene sequences.*
- *Speech recognition.*

Model building.

- *Reverse engineering.*
- *Autonomously navigating robot.*

Inference.

- *Reasoning in court decisions.*
- *Weather forecasting.*

1.2 Information processing and decision making

The examples above all illustrate the typical use of Bayesian probability as a *input–output* information processing activity:

Input:

- a given "world," "system," or "context," in which names, phenomena and values have a clear meaning to practitioners in the domain.
- facts, data, measurements, relationships, constrainsts, laws of nature, ..., that provide information about the "system."

Output:

- an *analysis* of the "system": the understanding and interpretation of what is going on in the system. The explication of what where the *causes* of the observed *effects* in the "system."
- *predictions*: about how the "system" will evolve in the near future, and how it will react to specific inputs.
- *decisions*: does the "system" satisfies its requirements? What actions should be taken on the basis of the available information?

All the reasoning about the "system" is done under *uncertainty*: numerical conclusions about the "system" are (usually) not given as logical, binary numbers, which are either *true* or *false*. On the contrary, each conclusion is accompanied with a measure of its uncertainty. How to represent information about the "system" in a numerical manner is explained in a following Section.

1.3 The essence of Bayesian probability

The essence of Bayesian probability is that it gives *precise* answers to the following questions:

1. What is *information*?
2. How is information *mathematically modelled*?
3. Where does information come from? How does it change? And how is it processed?
4. How does one draw *conclusions* or *make decisions* on the basis of available information?

Message 2 (Probability as information processing tool) *This course explains why one can have good faith in using the mathematics of probability theory as a* consistent*,* unique *and* plausible *tool for dealing with uncertainty in real-world systems.*

(Note that this message doesn't talk about the whole of Bayesian probability theory, but just about the *information processing* part of it.)

Definition 1 (Consistency) *It does not matter in what form or order the available information is processed by the Bayesian probability tools and algorithms, because the result will always be the same. That is, the Bayesian framework is free from* paradoxes *and* internal contradictions*.*

A later Section will show how this internal consistency is derived from an *axiomatic basis.*

Definition 2 (Unique) *Bayesian probability theory provides an unique way to process information from "input" to "output."*

However, this unique way can be too complex to describe, and/or too computationally intensive to calculate. Hence, many approximate processing algorithms have been developed (and will be introduced in a later Section).

Definition 3 (Plausible) *In accordance with what seems logical to a human being.*

Hence, Bayesian probability has become quite popular in much of the modern research and products in *Artificial Intelligence*. One of the major achievements in the 20th century development of Bayesian probability is that this "vague" definition of plausibility can be nicely formalised in mathematical form, and that the axioms of probability theory (sum, product and Bayes' rule) can be derived from it. It is also interesting to learn that this development was almost exclusively driven by physicists, and that almost no mathematicians or statisticians were involved.

Fact 1 (Determinism) *Bayesian probability is a fully deterministic theory to deal with undeterministic systems and data.*

Definition 4 (Uncertainty, Belief, Evidence) *Uncertainty, Belief and Evidence are used as synonyms of information.*

The terminology "information" will most often be used in this text.

1.4 Modelling: representation of information

Every practically useful mathematical theory about information processing in a given "system" must have a way to numerically represent and quantify information. Bayesian probability uses the following representation:

1. *Variables.* The "system" is fully described by a number of variables, which can be given numerical values. These variables can be *discrete* or *continuous*; they can be *scalar* (i.e., consisting of only one single number) or *composite* (i.e., consisting of more than one single number).

2. *Information.* The information about the N variables in the "system" is a *single-valued* function over the N-dimensional base space of variables.

3. *Relationships.* The variables is the "system" can, in general, not take aribitrary values independent of each other. Hence, there are functional relationships that indicate the inter-dependence of the variables. These relationship functions need *not* be single-valued.

4. *State.* The "system" is, at a particular instant in time, in a particular *state*, i.e., the information functions on the system's variables have particular values. The set of all possible values that the state variables can have is called the *state space*.

It is important to observe that the information function is *single-valued*. Such functions are, in the context of statistics, commonly known as *probability density functions* (PDF), Figure 1, and this text will use this terminology. Note also that the absolute value of the function is not important, but only the relative values at different points.

The function value at each point of the base space indicates how "likely" the particular combination of numerical values of the different variables is at that point. In fact, this value of the PDF at a single point is without much meaning, at least for the case of a continuous PDF; what is important is the *integral* of the PDF over an *area* of the variables space.

Message 3 (PDF = function + measure) *A PDF is not fully characterized by the function over the variables space: also the* (density) measure *at each point of this space is important.*

So, only expressions like

$$\int_D p(x)\,dx \tag{1}$$

have real meaning (i.e., the amount of "probability mass" contained in the domain of the integral). The dx is called the *measure* of the variables space: the total probability mass is the product of the value of the PDF function, times the "density" of variables at this particular place. This density or measure is in general not equal in different places of the state space. For example, on the surface of the earth, there is a diminishing amount of ground between two longitude and latitude increments (i.e., the "square" formed by moving one unit in either direction), when coming closer to the poles.

Measures are not just used for notational purposes (i.e., to denote the variables over which to integrate), but they are examples of so-called *differential forms*. An n-dimensional differential form maps n tangent vectors to a real number. The tangent vectors form the edges of a "parallellepipedum" of volume in the parameter space; the result of applying the differential form on this parallellepipedum is the amount of volume enclosed in it. Measures have their own transformation rules when changing the representation (change of coordinates, change of physical units, etc.): changing the representation changes the coordinates of the tangent vectors, and hence also the mathematical representation of the differential forms should change, in order to keep the enclosed volume *invariant*. An invariant measure puts *structure* on the parameter space. It can also be interpreted as the generalized form of a *uniform* probability distribution: with this distribution, each unit of volume in parameter space is equally probable.

The concept of an invariant measure is important for parameter spaces that are fundamentally different from $\mathbb{R}^n$, i.e., that have a different *structure* than $\mathbb{R}^n$, even though they may have n real numbers as coordinates. The best-known example is probably the measure used to calculate the surface integral over a sphere: depending on whether one uses Cartesian x, y, and z coordinates, or spherical r and θ coordinates, the measure changes, [36]:

$$\int_A \{f(x,y,z)\}\ \{dx\ dy\ dz\} = \int_A \{f(r,\theta,\phi)\}\ \{r\ dr\ \sin(\theta) d\theta\ d\phi\}\,. \tag{2}$$

Fact 2 (Properties of PDF) *Classical statistics often characterizes a PDF by saying that its integral over its whole domain is equal to 1. However, this "1" doesn't have any intrinsic meaning, and could be replaced by any other positive number.*

Definition 5 (Model) *The combination of a set of variables, their information and relationship function(s) is called a* model.

Choice 1 (Choice of model) *Any "system" can be given many possible mathematical models; which one to choose is a rather arbitrary choice of the practitioners. They can be guided in their choice by various criteria: level of detail of the model; computational complexity; tradition; etc.*

Fact 3 (Importance of model selection) *Choosing an appropriate model for a particular application is not a simple job. And (the mathematical representation chosen in) the model determines to a very large extent the efficiency and accuracy of the information processing that will take place in the system.*

Definition 6 (Constraint) *A constraint is used as a synonym of a relationship between variables.*

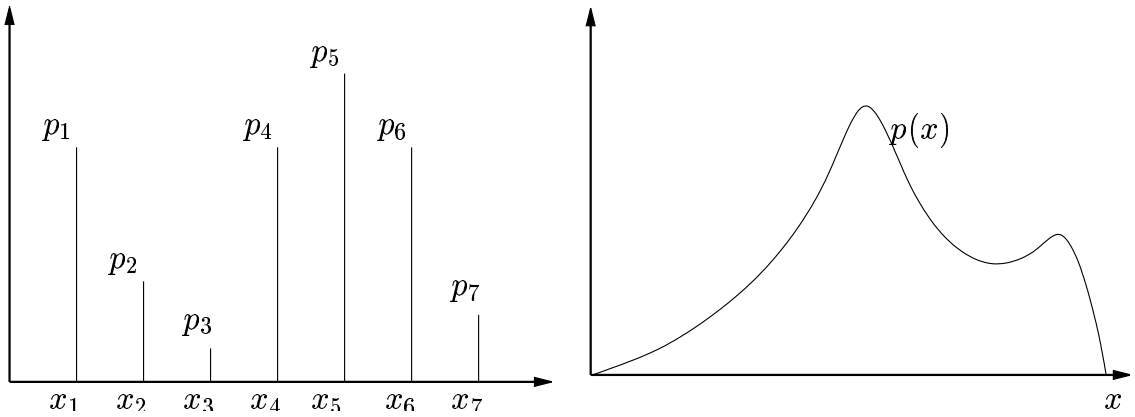

Figure 1: Examples of discrete and continuous information functions, i.e., probability distributions.

Learning the syntax and semantics of the model (state, information functions, relationships) in a particular application domain is usually also the hardest part of building a reasoning system for that application. Once the model has been built and understood, all information processing should be straigthforward, as will be shown in later sections.

Definition 7 ((Joint) probability density function) *The multi-dimensional, single-valued information function $p(X, Y, Z, \dots)$ that describes the dependencies between the variables $X, Y, Z, \dots$ in a model is called a* (joint) probability density function.

Definition 8 (Dependent variables) *Two variables X and Y are (statistically) dependent if a change in the value of X is correlated with a change in the value of Y.*

This does not necessarily mean that there is a physical *causal connection* between X and Y. In the example of the alarm, X could be *JohnCalls*, and Y is *MaryCalls*; there is no physical connection between John and Mary making calles; but in this case, both are sometimes influenced by a common cause, i.e., the alarm that sounds.

Definition 9 (Conditional PDF) *The information function $p(X, Y, Z, \dots)$ between the variables $X, Y, Z, \dots$ can depend on the given values of some other variables $A, B, C, \dots$. This sort of information function is called a* conditional PDF*, and the dependence is denoted by a vertical bar: $p(X, Y, Z, \dots | A, B, C, \dots)$.*

Figure 1 shows a discrete and a continuous PDF in one single dimension. Figure 2 shows a set of four PDFs from a very common family of one-dimensional PDFs, called *normal distribution* or *Gaussian distribution*. These latter distributions are very popular, because (i) the information about many "systems" can be described by Gaussians, and (ii) their mathematical properties are very attractive. That is, one needs only two parameters to represent the normal distribution, denoted by $\mathcal{N}(\mu, \sigma)$: its *mean* μ and its *standard deviation* σ:

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{3}$$

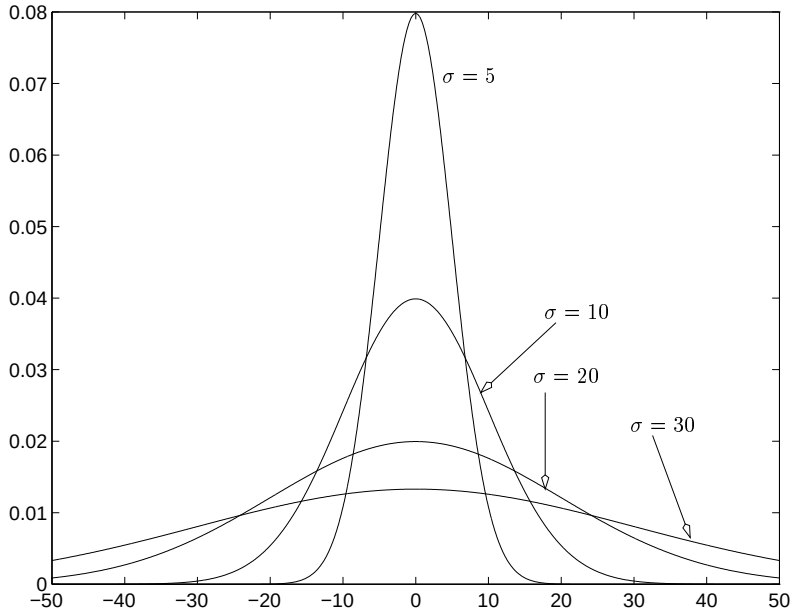

Figure 2: Gaussian probability distributions, with mean 0 and different standard deviations σ.

σ^2 is called the *covariance* of the distribution. The Gaussian distribution above is *univariate distribution*, i.e., it is a function of one single parameter x. Its *multivariate* generalization has the form

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P}) = \frac{1}{\sqrt{(2\pi)^n \ \det(\boldsymbol{P})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{4}$$

$\boldsymbol{x}$ is an n-dimensional vector, $\boldsymbol{\mu}$ is the vector of the mean (*first moment*, or "*expected value*") of $\boldsymbol{x}$:

$$\boldsymbol{\mu} = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \left\{\boldsymbol{x}\ p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{P})\right\}. \tag{5}$$

$||\boldsymbol{P}||$ is the *two-norm* (i.e., the square root of the largest eigenvalue of $\boldsymbol{P}^T\boldsymbol{P}$, [11]) of the *second moment*, or *covariance matrix* $\boldsymbol{P}$:

$$\boldsymbol{P} = \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \ \left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\ p(\boldsymbol{a}|\boldsymbol{\mu}, \boldsymbol{P})\right\}. \tag{6}$$

Note that the term $(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$ in the above equation is a *matrix*, while the term $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ in Eq. (4) is a *number*. It can be shown that this number has all properties of a *distance* on the parameter space of $\boldsymbol{x}$; in other words, the inverse of the covariance matrix $\boldsymbol{P}$ is a *metric* on that space, [27], and hence determines *structure* on the space: points in the space are "ordered" relative to each other by their distances.

1.5 Marginalization and Bayes network

Definition 10 (Marginalization) *Given a PDF $p(X, Y, Z)$ that represents the information about in what ways the values of the variables X, Y and Z can occur together. Marginalization is then the process to derive the information about X and Y, given* all possible values *of Z.*

The mathematical instantiation of marginalization is conceptually simple: integrate $p(X, Y, Z)$ over all possible values of Z:

$$p(X, Y) = \int_Z p(X, Y, Z)\, dz. \tag{7}$$

Fact 4 (Importance of marginalization.) *Marginalization is of utmost importance for all* inference *in Bayesian probability: the information about a subset of the system's variables is derived by "integrating out" all "superfluous" variables.*

These superfluous variables are not superfluous at all: they are needed in the system because it's usually through them that information enters the model, and can be used to infer information about the variables that the application user is interested in.

Calculating the integrals needed in marginalization can, in practice, be *very* time consuming. This difficulty has motivated two complementary research developments: (i) more efficient integration algorithms, and (ii) modeling of uncertaint systems with PDF functions that are not too computational complex to work with. Note that both developments are a trade-off between efficiency and accuracy. One successful result of this quest for efficient representation (and hence also processing) of information are Bayesian networks.

Definition 11 (Bayes network) *A Bayes network is a directed graph structure to* represent *the (probabilistic) dependencies between various variables in a system, and that is used to reduce the computational complexity with respect to working with the full joint PDF.*

Later sections will give more details, but Figure 3 illustrates the basic idea: the variables X, Y and Z are connected by a PDF $p(X, Y, Z)$; the network represents the fact that X and Y are independent, *given* the information about the variable Z. This structure in the dependencies can be exploited to simplify marginalization operations on $p(X, Y, Z)$.

1.6 Information is subjective

Fact 5 (Information is subjective) *The PDF that represents the information about a given system does not represent the "real state" of that system, but it represents all information that we, as human observers of the system, (think we) have collected about the system.*

The collection of information comes from various sources: prior knowledge about the system (such as physical laws it has to obey, or known initial conditions); measurements from sensors that can observe (parts of) the system; data from the past; etc.

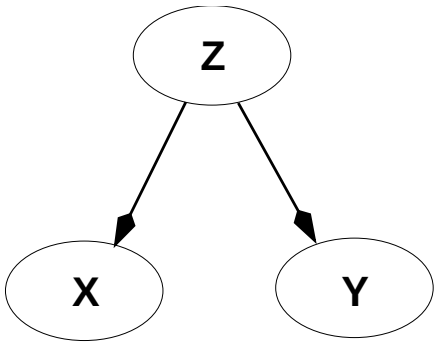

Figure 3: Smallest element in a Bayesian network: X and Z are conditionally independent, given Z.

1.7 Information processing is objective

The information that is available about a given system is often not in the most appropriate form to be really useful for a particular application. Hence, this information has to be *processed*, to result in a more appropriate form. Basically, there are two forms of processing:

1. *Transformation.* The information is available as a PDF $p(X, Y, Z, A, B, \dots)$, but the user prefers a functional dependence on the variables C and D. *If* a relationship exists between, for example, A and B, and C and D, the PDF $p(X, Y, Z, A, B, \dots)$ can be transformed into a PDF $p(X, Y, Z, C, D, \dots)$.

   Very often, the transformation consists of *data reduction*: the N-dimensional PDF is transformed into a much lower-dimensional PDF. In the case of decision making (Section 1.8), the reduced "PDF" is just one single number. Hence, a significant part of the information is thrown out in the reduction process.

2. *Combination.* From time to time, new information about the system is coming in. For example, via sensors that make measurements of some variables of the system. This new information should be taken into account, and combined with the already available information.

   The unique mathematical tool to perform combination of information is through *Bayes' rule*. A later Section gives more details about where this rule comes from, and how it should be used.

Figure 4 gives examples of information transformation and combination: a mobile robot drives around in an office corridor, which has three doors. When its power is switched on, it doesn't know where in the corridor it is. This lack of information is represented by a *uniform PDF*. The robot takes a first sensor measurement, and "sees" a door. This leads to a non-uniform PDF, with (low) peaks at the three possible doors. After having moved a bit further, the robot sees another door, which raises the probability of being close to the second door.

In this example, *transformation* of information could be that, from its position in the map, the robot derives the direction in which it expects a certain visual clue against the walls.

Updating the information about its position is an example of *combination* of information, i.e., the previously collected information about its position is combined with the new sensor information.

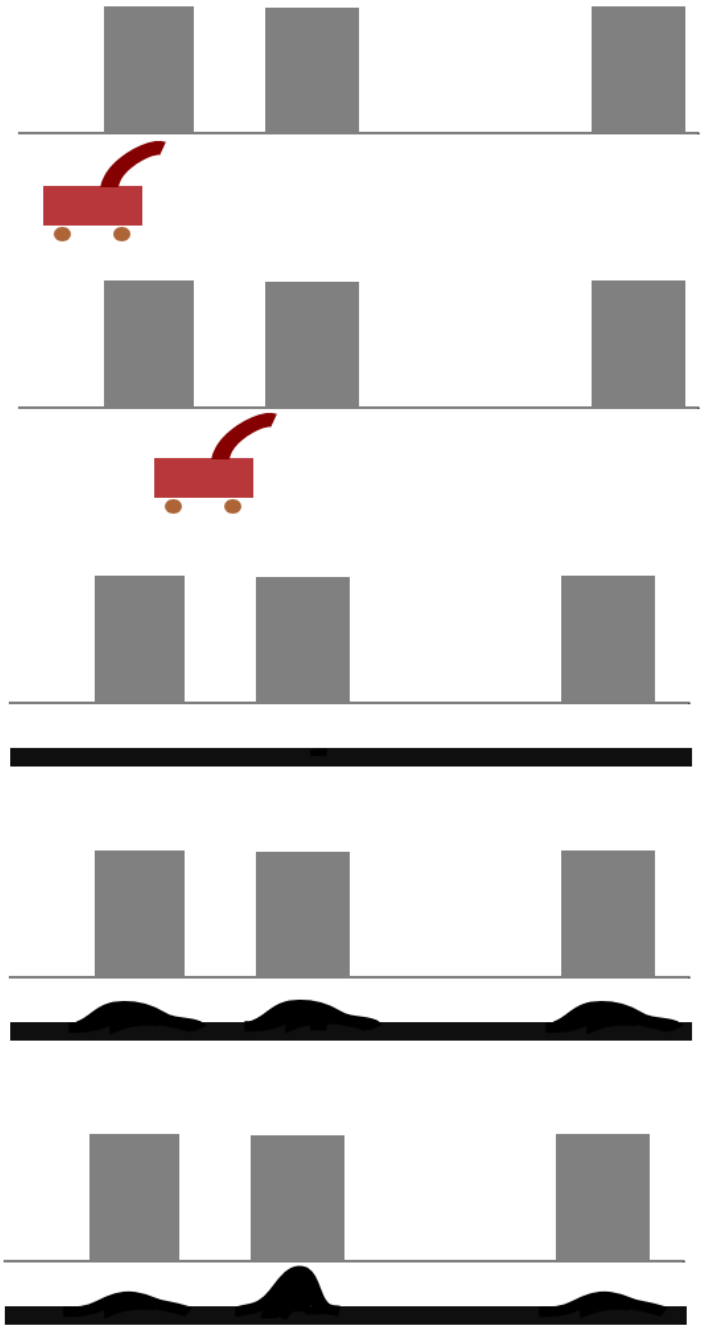

Figure 4: A mobile robot drives in an office corridor, and builds up information about where it is positioned with respect to the doors in the corridor.

Message 4 (Conservation of information) *The processing of information should not change the information content; i.e., it should not delete, nor add information.*

This seems a trivial statement to make, but is not in practice. One of the fundamental questions to answer first before one can check whether information is conserved, is how to *measure* information. This will be dealt with in Section 2.

1.8 Decision making

All practical applications of information processing spent time on this processing for one single purpose only: collect as much information as possible on a system, in order to be able to *draw conclusions*, or *make a decision* to perform specific actions. Decision making necessarily reduces higher-dimensional information representations into one single number (or low-dimensional vector of numbers).

Definition 12 (Decision and utility functions) *Both terms are used as synonyms in this text. A utility function is a mapping from the state space of the system to the real line. It represents the value of the information acquired thus far.*

Fact 6 (Choice of decision function) *The choice of a decision function is most often very arbitrary. There is no equally attractive axiomatic theory on decision making as for the axiomatic foundations of Bayes' rule (see later).*

In the example of Figure 4, the robot must make a decision to enter a particular door or not, based in the collected information thus far. Hence, from all the *possible* positions, it *has* to choose only one single position, and plan its motion to move through the "door" as if this chosen position was the real position.

Message 5 *In every application, always make the explicit distinction between modelling, information processing, and decision making.*

Fact 7 (Process first, decide later) *In Bayesian probability, information processing is performed every time the system transforms, and/or gets new information. Decisions can be taken at any moment, using the information gathered until that moment. The decision doesn't transform the available information.*

Fact 8 (Conflicting beliefs) *Many reasoning systems feel the need to introduce the concept of conflicting beliefs. In Bayesian probability, this concept does not exist on the level of information processing. However, it* is *possible that two different people define decision functions whose evalution contradicts each other.*

2 Information measures

Any uncertainty reasoning system should be able to tell its users *how much* information it has gathered about a particular part of the system under study. This means the system must have a procedure to reduce the available information (stored in the form of possibly high-dimensional PDFs) to *one single scalar* number.

Fact 9 (Ambiguity of information measures) *Calculating the information content in a PDF is an* irreversible *operation: it is impossible to reconstruct the original PDF from knowing only a scalar that quantifies its information.*

Of course, an infinite number of ways exist to reduce a PDF to one single number. But history has provided us with a (small) number of choices that have interesting and "natural" properties. *Entropy* is one of the most "pure" measures, because, as this Section will show, it is derived from a minimal and very plausible set of desirable information measure properties.

2.1 Good's information function

Let's formally denote by $I(M{:}E|C)$ the information about model M that can be derived from the information E, and given the "background" ("context") information C. Good [12] proposed the following structural properties for $I(M{:}E|C)$:

1. $I(M{:}E \text{ AND } F|C) = f\{I(M{:}E|C), I(M{:}F|E \text{ AND } C)\}$.

   The information about M derived from both E and F is a function of the information about M derived from E only, and the extra information given by F, interpreted in the context that E is already taken for granted.

   f simply represents a *functional relationship* between the different terms, without saying anything about the exact form of this relationship.

2. $I(M{:}E \text{ AND } M|C) = I(M|C)$.

   If one knows already everything about M, other information cannot add anything anymore.

3. $I(M{:}E|C)$ is a strictly increasing function of its arguments.

   If the information content of one of the parameters of I increases, the information I increases too.

4. $I(M_1 \text{ AND } M_2{:}M_1|C) = I(M_1{:}M_1|C)$ if M_1 and M_2 are mutually irrelevant pieces of information.

5. $I(M_1 \text{ AND } M_2|M_1 \text{ AND } C) = I(M_2|C)$.

   Considering the information contained in M_1 doesn't increase the total information if this information was already incorporated.

The details of the Good's derivation are skipped in this text, but Good found in a rather straightforward way that these specifications lead to *many alternatives* for the representation of information. And he also proved that, *if* information $I(M|C)$ is represented by a *measurable* function[1] $p(M|C)$, *then* composition of

[1] Loosely speaking, this means that calculation of information requires no involved mathematical technicalities.

information becomes *additive* if and only if I is any function of $\log(p)$, or $\ln(p)$. The simplest choice being, of course, $I = \ln(p)$, which is the rationale behind the abundance of logarithms in statistics, for example in the information measures discussed in the following Subsections. Indeed, *addition* is the *natural* operator on the space of these logarithms, and is easy to work with.

The following subsection explains how Shannon found one particular logarithm-based function as a very attractive measure of information.

2.2 Shannon entropy

Claude Elwood Shannon (1916–), [4, 31, 32], presented a *scalar* "average" (or "expected") measure to quantify the quality of communication channels, i.e., their capacity to transmit information. However, his measure also qualifies as a fully general information measure, as was first explained by Jaynes [14]. Shannon gave his measure the name of *entropy*, because it models the similar concept with the same name in thermodynamics: the higher the entropy of a thermodynamic system, the higher our uncertainty about the state of the system, or, in other words, the higher its "disorder." Note that entropy is a "subjective" feature of the system: it represents the knowledge (or uncertainty) that the *observer* has of the system, but it is not a physical property of the system itself. However, it is "objective" in the sense that each observer comes to the same conclusion when given the same information.

Assume the parameter x takes one of the values from a set $\{x_1, \ldots, x_n\}$, with $p(x) = \{p_1, \ldots, p_n\}$ the corresponding probability distribution. The following three properties are straightforward, plausible *desiderata* for any information measure $H(p)$ of the probability distribution $p(x)$:

Axioms for entropy

I	H is a *continuous* function of p.
II	If all n probabilities p_i are equal (and hence equal to $1/n$, if we choose them to have to sum to 1), the entropy $H(1/n, \ldots, 1/n)$ is a *monotonically increasing* function of n.
III	H is an *invariant*, i.e., the uncertainty should not depend on how one orders or groups the elements x_i.

The first and second specifications model our intuition that (i) small changes in probability imply only small changes in entropy, and (ii) our uncertainty about the exact value of a parameter increases when it is a member of a larger group.

The third desideratum is represented mathematicly as follows: the information measure H obeys the following *additive composition law*:

$$\begin{aligned} H(p_1, \ldots, p_n) = & H(w_1, w_2, \ldots) \\ & + w_1 H(p_1|w_1, \ldots, p_k|w_1) \\ & + w_2 H(p_{k+1}|w_2, \ldots, p_{k+m}|w_2) + \ldots, \end{aligned} \tag{8}$$

where w_1 is the probability of the set $\{x_1, \ldots, x_k\}$, w_2 is the probability of the set $\{x_{k+1}, \ldots, x_{k+m}\}$, and so on, Figure 5; $p_i|w_j$ is the probability of the alternative x_i *if* one knows that the parameter x comes from the set that has probability w_j.

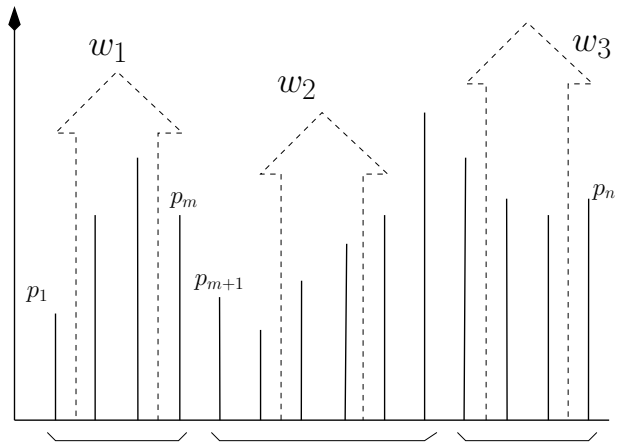

Figure 5: Grouping in a discrete probability distribution p_i, in three sub-groups. The dashed lines represent the discrete probability density function w_j of the three groups.

For example, assume that x comes from a set of three members, with the alternatives occurring with probabilities $1/2, 1/3$ and $1/6$, respectively. If one then groups the second and third alternatives together (i.e., $w_1 = p_1 = 1/2$, the probability of the set $\{x_1\}$, and $w_2 = p_2 + p_3 = 1/3 + 1/6 = 1/2$, the probability of the set $\{x_2, x_3\}$), the composition law gives $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2}H(2/3, 1/3)$ since $2/3$ and $1/3$ are the probabilities of x_2 and x_3 within the set $\{x_2, x_3\}$.

The three above-mentioned *axioms* suffice to derive an analytical expression for the information measure function $H(p)$. The first axiom implies that it is sufficient to determine $H(p)$ for *rational* values $p_i = n_i / \sum_{j=1}^{n} n_j$ (with n_j integer numbers) only; the reason is that the rational numbers are a *dense* subset of the real numbers. One then uses the composition law to find that $H(p)$ can be found from the uniform probability distribution $P = (1/N, \ldots, 1/N)$ over $N = \sum_{i=1}^{n} n_i$ alternatives. Indeed, the composition law says that the entropy $H(p)$ is equal to the entropy $H(P)$, because in P one can group the first n_1 alternatives, the following n_2 alternatives, and so on, which reduces to the original distribution. For example, let $n = 3$ and $(n_1, n_2, n_3) = (3, 4, 2)$ such that $N = 3 + 4 + 2 = 9$; denoting $H(1/N, \ldots, 1/N)$ by $H(N)$ yields

$$H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}H(3) + \frac{4}{9}H(4) + \frac{2}{9}H(2) = H(9).$$

In general, this could be written as

$$H(p_1, \ldots, p_n) + \sum p_i H(n_i) = H\left(\sum n_i\right). \tag{9}$$

The special case of all n_i equal to the same integer m gives

$$H(n) + H(m) = H(mn).$$

A solution to this equation is given by $H(n) = K \ln(n)$, with $K > 0$ because of the monotonicity rule. All

this yields the following expression for the entropy:

$$H(p_1, \ldots, p_n) = K \ln\left(\sum n_i\right) - K \sum p_i \ln(n_i), \tag{10}$$

$$= -K \sum p_i \ln\left(\frac{n_i}{\sum n_i}\right). \tag{11}$$

$$= -K \sum p_i \ln(p_i). \tag{12}$$

Fact 10 (Entropy of a discrete probability distribution)

$$\boxed{H(p_1, \ldots, p_n) = -K \sum p_i \ln(p_i).} \tag{13}$$

Note that $\ln(p_i) < 0$, because $(n_i / \sum n_i) < 1$. The minus sign in the entropy (or information) makes the entropy measure positive, and increasing when *uncertainty* increases; this is the same interpretation as in statistical mechanics, the science that originally defined the concept of entropy. The constant K has no influence: it is nothing but a factor that sets the scale of the entropy. The entropy need not be a monotonically decreasing function of the amount of information received: entropy can increase with new information ("evidence") coming in, if this new information contradicts the previous assumptions. Note also that the uncertainty in many *dynamic* systems increases naturally over the time period that no new information is received from the system: the probability distributions "flatten out" and hence the entropy increases.

Figure 6 gives examples of entropy functions for various simple PDFs. It shows that entropy corresponds, to some extent, to our intuition about information: a PDF with a sharp peak has more information than one with a broad peak. A PDF with much variation has a lower information than one without much variation. A PDF with many alternating peaks and valleys has the same information as one which as all peaks and valleys assembled together.

Note also that the information measure can never reach "absolute zero," because this would require $p_i \ln(n_i / \sum n_i)$ to vanish, which can only occur for trivial PDFs with one single element.

Message 6 (Comparison of PDFs) *The number of samples as well as the (arbitrary) scaling constant K make comparisons of the absolute values of the entropies of two PDFs quite useless.*

2.3 Entropy for continuous PDF

At first sight, extending Eq. (13) from discrete to continuous PDFs seems straigthforward, but it isn't. This can be seen from the reasoning that led to that formula, because there the value $H(1/N, \ldots, 1/N)$ of the entropy of a *uniform* distribution is used. And this uniform distribution is not well defined for a continuous PDF running over all real values.

A second look at Eq. (11) suggests another interpretation:

1. the fraction $n_i / \sum n_i$ can be taken in each interval of the PDF parameter(s).
2. the *ratio* of two of these ratios makes sense locally, in terms of the PDFs' *densities*:

$$\frac{n_i / \sum n_i}{m_j / \sum m_j} \to \frac{dx}{dy}, \tag{14}$$

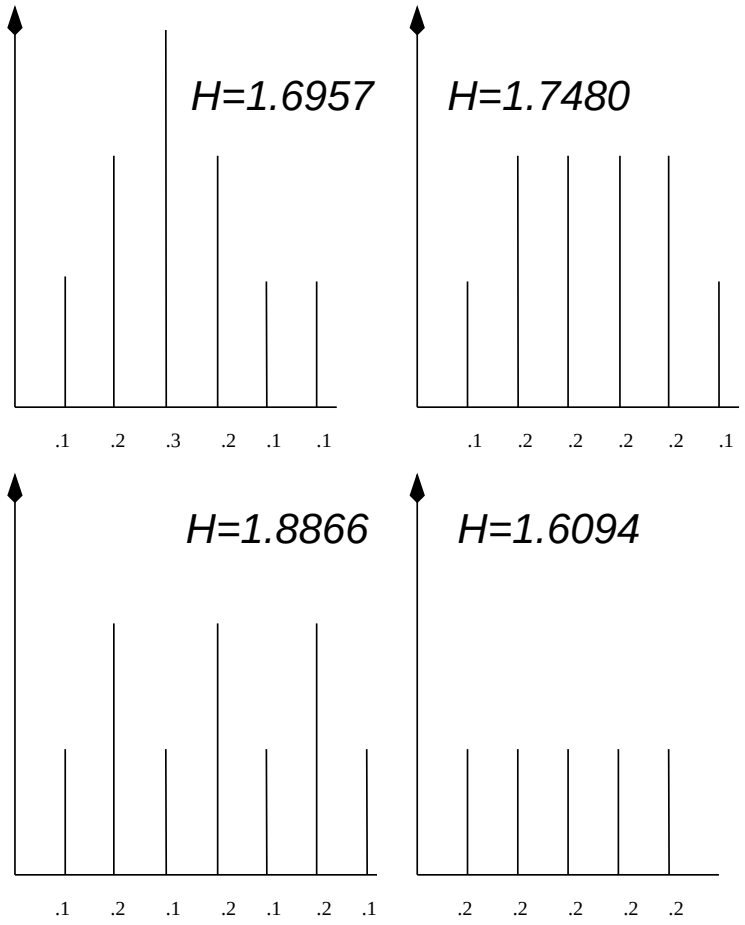

Figure 6: Examples of entropies H of discrete PDFs.

where dx is the density of the continuous PDF that we obtain from "taking the limit" of the discrete PDF n_i, and simularly for dY and m_j.

Hence, only *relative* information measures are possible. This is consistent with the fact that there is no *absolute zero* information, to which the "distance" of a PDF $p(x)\,dx$ could be taken.

Fact 11 (Mutual information) *The relative information measure (often called "*mutual information*") $H(p,q)$ of two continuous PDFs $p(x)$ and $q(x)$ is defined as:*

$$\boxed{H(p,q) = -\int p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx.} \tag{15}$$

This scalar is also called the *Kullback-Leibler divergence*, (after the duo that first presented it, [22], [23]), or also *mutual entropy*, or *cross entropy*, of both probability measures $p(x)$ and $m(x)$, [22, 28].

It is a (coordinate-independent) measure for how much information one needs to go from the probability distribution $m(x)$ to the probability distribution $p(x)$. As Shannon's entropy, $H\big(p(x){:}m(x)\big)$ is a *global* measure, since all information contributions $\log(p(x)/m(x))$ are weighted by $p(x)dx$, and then added together.

2.4 Fisher Information

The relative entropy $H\big(p(x), m(x)\big)$ of Eq. (15) is *not* a *distance function* (or "metric") on the space of all PDFs, since it is not symmetric in its arguments:

$$H\big(p(x), m(x)\big) \neq H\big(m(x), p(x)\big). \tag{16}$$

Rao [27] was the first to come up with a real distance function on the *manifold* $\mathcal{M}_\Sigma$ of probability distributions $p(x, \Sigma)$ over the state space x and described by a parameter vector $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$.

Message 7 (PDF manifold) *$\mathcal{M}_\Sigma$ is a parameterized space of PDFs, which is* not *the same space as the state space of the system on which the PDF is defined.*

Define *tangent vectors* $v = (v^1, \ldots, v^n)$ to the manifold $\mathcal{M}_\Sigma$ as follows:

$$v(x, \Sigma) = \sum_{i=1}^{n} v^i \frac{\partial l(x, \Sigma)}{\partial \sigma^i}, \qquad \text{with} \quad l(x, \Sigma) = \log\big(p(x, \Sigma)\big). \tag{17}$$

The v^i are the coordinates of the tangent vector in the basis formed by the tangent vectors of the *logarithms* of the σ-coordinates. A metric $\mathcal{M}$ at the point p (which is a probability distribution) is a bilinear mapping that gives a real number when applied to two (logarithmic) tangent vectors v and w attached to p. Rao showed that the *covariance* of both vectors satisfies all properties of a metric. Hence, the elements g_{ij} of the matrix representing the metric are found from the covariance of the coordinate tangent vectors:

$$g_{ij}(\Sigma) = \int \big(\partial_i l(x, \Sigma)\big)\big(\partial_j l(x, \Sigma)\big) p(x, \Sigma)\, dx. \tag{18}$$

The matrix g_{ij} got the name *Fisher information matrix*. The covariance "integrates out" the dependency on the state space coordinates x, hence the metric is only a function of the statistical coordinates Σ. This metric is defined on the tangent space to the manifold $\mathcal{M}_\Sigma$ of the σ-parameterized family of probability distributions over the x-parameterized state space $\mathcal{X}$. Kullback and Leibler already proved the following relationship between the relative entropy of two "infinitely separated" probability distributions Σ and $\Sigma + \epsilon v$ on the one hand, and the Fisher Information matrix $g_{ij}(\Sigma)$ on the other hand:

$$H(\Sigma{:}\Sigma + \epsilon v) = \frac{1}{2} \sum_{i,j} g_{ij}(\Sigma) v^i v^j + \mathcal{O}(\epsilon^2). \tag{19}$$

Hence, Fisher Information represents the *local* behaviour of the relative entropy: it indicates the rate of change in information in a given direction of the probability manifold (*not* in a given direction of the state space!).

Fact 12 (Fisher Information of a Gaussian PDF) *It can be proven that the Fisher Information of a Gaussian distribution gives the distribution's covariance matrix.*

This covariance matrix P is a function of the variables on the PDF parameter space, i.e., mean μ and variance σ.

2.5 Measures for Gaussian PDFs

The Gaussian PDFs (Section 1.4) are among the most widely used PDFs, because of their mathematical simplicity. They also offer simple *information measures*, based on the covariance matrix $P(\mu, \sigma)$. Equation (4) is repeated here for convenience:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P}) = \frac{1}{\sqrt{(2\pi)^n}\,||\boldsymbol{P}||^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

The argument of the exponential function is a *real number*. Hence,

$$\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{20}$$

can be considered as the *magnitude* of the state space "vector" $\boldsymbol{x}$, and so, $\boldsymbol{P}$ (or, rather, its inverse) is a *metric* on the state space.

Fact 13 (Generalized least-squares) *Equation (20) generalizes the well-known* least-squares *criterion of measuring the deviation between a state space vector $\boldsymbol{x}$ and another state space vector $\boldsymbol{\mu}$.*

In addition, the covariance matrix itself is often used to derive measures for the information contained in the corresponding Gaussian PDF. This is a list of often used measures:

- *Trace.* That is, the *sum* of the eigenvalues of the matrix. This is one of the so-called *linear invariants* of the covariance matrix.
- *Determinant.* Another invariant, being the *product* of the eigenvalues.
- *Ratio of singular values.* Most often, the condition number (ratio of smalles to largest singular value) is taken as measure.

Fact 14 (Arbitrariness of Gaussian measures) *None of the above-mentioned measures derived from the covariance matrix of a Gaussian PDF has the status of a* natural *or* absolute *information measure.*

2.6 No information—Ignorance

Every piece of software, hence also every set of software agents that together implement an "intelligent" robot, must start from a certain initial state. In the context of plausible inference, one is often tempted to describe this initial state as the state in which the robot knows "nothing" yet. This raises the two questions as to (i) what such a total ignorance really means, and (ii) how to represent it. Since many researchers didn't find satisfactory answers to these two questions, they jumped to the conclusion that

probability theory is not a valid framework for A.I. However, the fact that they didn't find satisfactory answers says much about their own state of ignorance, since ignorance can be dealt with in a very clean and formal way. (But, it is true, not always in a *simple* way.) The French scientist Pierre Simon de Laplace (1749–1827) was the first to propose the *uniform distribution* of a parameter as the state of ignorance about its exact value. This approach of assigning an a priori distribution this way later got the name of *Laplace's principle of indifference.* Harold Jeffreys [19] generalized this, and presented the *invariant volume form* as the *non-informative* prior distribution. However, Jeffreys was not accepted by the then active community of statisticians, so his ideas were not widespread. A similar fate befell Edwin Thompson Jaynes (1922–1998), in the 50s, 60s and 70s, although he did add mathematical rigour to the somewhat intuitive ideas of Jeffreys, [14, 29], and, especially, did a lot to spread the Bayesian ideas.

Let's first discuss the question about what total ignorance means. In fact, it doesn't mean much: one always knows *something* about the system one is interested in; or, at least, one could come up with some models, even though one would have no idea about the values of the parameters in it. Or, in the words of Jaynes: "merely knowing the physical meaning of our parameters [in a model], *already constitutes highly relevant prior information* which our intuition is able to use at once" (emphasis is Jaynes').

The question about which "ignorance priors" (or "*noninformative priors*") to choose has still not been answered completely satisfactorily: Jeffreys' non-informative prior distribution works only for *location parameters*, such as the mean value of a parameter. For other properties, such as e.g. the standard deviation, other ignorance prior distributions are needed. Jaynes's approaches to find ignorance priors are [15, 16, 17, 18]:

1. *Invariance under transformations.* If the only thing one knows about the system is a model or an hypothesis, this ignorance should not change if the mathematical representation of the model is transformed into an equivalent representation. For example, a uniform distribution on x does not necessarily lead to a uniform distribution on $y = x^2$.

2. *Maximum Entropy (MaxEnt) principle.* If all one knows about a system is a number of constraints it has to obey, the prior distribution of the parameters describing this system is given my maximizing the entropy of the distribution taking into account the constraints. For example, if one knows the mean and variance of a parameter, the probability distribution that adds no extra a priori information (and hence has the largest entropy) turns out to be the Gaussian with given mean and variance, e.g., [3, 18, 24, 25].

3. *Extra "I don't know" hypothesis.* If the robot has a set of hypotheses for the system under consideration, and it doesn't know at all which one to prefer, or even whether one of these hypotheses is valid, it can add a new hypothesis that just says "I don't know what hypothesis is valid." Total ignorance is then represented by giving all the probability to this last hypothesis.

There still exists discussion about the appropriateness of these approaches; much of this controversy is caused by the fact that most researchers do not thoroughly understand the importance of *structure*... Also for the MaxEnt criterion, one has found a small set of very intuitive desiderata, [33, 34], that lead to an axiomatic treatment of, and hence fundamental *structure* on, the space of probability distributions of a system:

Axioms for MaxEnt	
I	*Subset independence*: information about one domain should not affect the information description in another domain, provided that there are no constraints linking both domains.
II	*Coordinate invariance*.
III	*System independence*: if the principle is valid in general, it should also work on specific individual systems.
IV	*Scaling*: the entropy of the probability distribution of a system should not change if no new information is added to the system.

This text does not give the proof that these intuitive desiderata indeed lead to the MaxEnt principle, but the approach is very similar to the approach applied on the desiderata for entropy; see the cited references for more details.

3 Axiomatic foundations for Bayesian calculus

This Section explains the axiomatic approach behind the calculus of Bayesian probability, i.e., the sum, product and Bayes' rule. The discussion is done in the "plausible reasoning" context of Jeffreys, Jaynes, Cox and others, [5, **?**, 13].

Good's approach (Section 2.1) does not say how to "calculate" with information, i.e., how to combine the information from different sources, or from the same source but collected at different time instants. This is the area of *probability theory*. This theory has many historical roots, and many approaches exist to explain its fundamental mathematical properties. The approach developed by Cox [5, **?**], and later refined by Jaynes [14], [18], [29] fits nicely into the information-based approach to inference presented in previous Sections. Cox and Jaynes looked at probability theory as the theory of how to deal with uncertain statements and hypotheses, and *not* as the more classical theory describing frequencies of random events. The former approach is currently known as *Bayesian* probability theory (e.g., [21]), in contrast to the latter "orthodox statistics" à la Fisher, [9]. Cox [5] starts from only three functional relationships to describe the *structure* that operations in uncertain reasoning should obey:

1. What we know about two statements A and B should be a smooth function of (i) what we know about A, and (ii) what we know about B *given* that A is taken for granted. In functional form this becomes:
$$p(A,B) = f\big\{p(A), p(B|A)\big\}, \tag{21}$$
with $p(A)$ the probability of proposition A, $p(B|A)$ the conditional proposition of B given A, and f an as yet unspecified function. (This is the same starting point as Good's reasoning that led to the definition of information.) $p(A)$ can be a continuous probability distribution *function* (instead of a single number), for example if it represents the value of a parameter in the system under study. Cox then proves that relation (21) leads to the following form for f:
$$f\big\{p(A), p(B|A)\big\} = f\big\{p(A)\big\}^m f\big\{p(B|A)\big\}^m, \tag{22}$$
for an arbitrary m and an arbitrary f. Hence, it turns out that the historical literature in statistics had already fixed the choices $m = 1$ and $f(u) = u$, not out of necessity but most probably just for computational convenience.

2. The negation of the negation of a proposition A is equal to the proposition A. This means that a function g must exist such that:
$$g\Big(g\big(p(A)\big)\Big) = p(A). \tag{23}$$
3. The same function g should also satisfy the following law of logic:
$$g\big(p(A \text{ OR } B)\big) = g\big(p(A)\big) \text{ AND } g\big(p(B)\big). \tag{24}$$

These structural prescriptions are sufficient to derive the product rule $p(A \text{ AND } B|M) = p(A|M)p(B|A \text{ AND } M)$, and the additivity to one, $p(A) + p(\text{NOT } A) = 1$. ($M$ represents all available knowledge ("models") used in the inference procedure.)

Jaynes builds further on the approach by Cox, and states some assumptions (e.g., invariance) a bit more explicitly. In his unfinished manuscript [18], the basic rules of plausible reasoning are formulated as follows:

Axioms for plausible Bayesian inference

I	Degrees of plausibility are represented by real numbers.
II	Qualitative correspondence with common sense.
III	If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.
IV	Always take into account all of the evidence one has.
V	Always represent equivalent states of knowledge by equivalent plausibility assignments.

These "axioms" are not accepted without discussion, e.g., [10, p. 241]:

- Assigning equivalent probabilities for equivalent states seems to assume that the modeller has "absolute knowledge," since one often doesn't know that states are equivalent.
- Many people state that representing probability by one single real number is not always possible or desirable.

It can be proved, [5, 18], in a way again very similar to the proof given for the entropy desiderata, that the above-mentioned axioms lead to the following well-known probability rules:

Bayesian calculus

Sum rule
$$\boxed{p(x + y|H) = p(x|H) + p(y|H).} \tag{25}$$

Product rule
$$\boxed{p(xy|H) = p(x|yH)p(y|H).} \tag{26}$$

Bayes' rule
$$\boxed{p(\text{Model}|\text{Data}, H) = \frac{p(\text{Data}|\text{Model}, H)}{p(\text{Data}|H)} p(\text{Model}|H).} \tag{27}$$

Equation (27) suggestively uses the names "Data" and "Model," since this is the contents of these variables in many robotics inference problems.

Fact 15 (Bayes' rule and PDFs) *Bayes' rule is defined to work with probabilities, not densities, but the differentials dx "cancel out" of the equations, [25, p. 104].*

Fact 16 (Bayes' rule and the product rule) *Bayes' rule follows straightforwardly from applying the product rule twice, developing $p(Model, Data|H$ once for Model and once for Data.*

The denominator in Bayes' rule is usually considered to be nothing more than just a normalization constant, [25, p. 105]; it is independent of the Model, and predicts the Data given only the prior information.

The term $p(\text{Model}|\text{Data}, H)$ is called the *posterior (probability)*; $p(\text{Data}|\text{Model}, H)/p(\text{Data}|H)$ is the *likelihood*; and $p(\text{Model}|H)$ is the *prior probability* of the hypothesis. The likelihood is *not* a probability distribution in itself; as the ration of two probability distributions with values between 0 and 1 it can have any positive real value.

Fact 17 (Bayes' rule represents model-based learning) *It expresses the probability of the Model, given the Data (and the background information H), as a function of (i) the probability of the Data when the Model is assumed to be known, and (ii) the probability of that Model given the background information.*

Let's take a look at what each term in Bayes' rule means. The left-hand side is what the reasoning system wants to calculate: the PDF representing the information on the Model, taking into account all Data; this PDF is a function over the *Model's* parameter space. The right-most term is also a PDF over the same parameter space, but before the Data was taken into account. The other term is written down as the ratio of two PDF functions, *seemingly* having both the same domain, i.e., the parameter space of the *Data*. However, the likelihood is also a function over the *Model* space. The Data is "measured" in its own parameter space, but in Bayes' rule it is used after *transformation* to the Model domain, through the mathematical relationship between both. This relationship is sometimes called the *measurement equation*, and makes up an indispensable part of the modelling step in any Bayesian reasoning system.

The transformation of a PDF through a functional relationship between two domains is quite straightforward. Assume that a PDF $p(x)$ is given over the x domain, and that x is transformed to the y domain through a functional relationship $x = f(y)$. Then the PDFs are related as follows:

$$p(x)\,dx = p(f(y))\,\frac{\partial f}{\partial y}\,dy. \tag{28}$$

Figure 7 shows Bayes' rule in action, working on a Gaussian PDF as prior, and a likelihood which is proportional to a Gaussian PDF. The resulting posterior is again a Gaussian PDF, [24, p. 7]. Recall, however, that the prior and the likelihood in general do not have the same function domain. So, what is depicted in the Figure is the Data PDF after transformation through the measurement equation.

Bayesian probability theory gives an axiomatically founded treatment of *all* structures and concepts needed in reasoning with uncertainty; that's why it is presented in this text. Classical ("orthodox") statistics gives some shortcuts for particular problems; this is convenient for specific implementations (such as many parameter identification applications) but it is seldom easy and intuitive to know what shortcuts are used in a given robotics task.

3.1 Optimality of information processing

The reasoning in Section 2.1 also allows to interpret Bayes' rule as a procedure to combine information from two sources *without loss of information*: the first source is the prior information already contained in the current state, and the second source is the new information added by the current measurement. This relationship is straightforward: take Bayes' rule for two models M_1 and M_2 that receive the same new $Data$:

$$p(M_1|Data) = \frac{p(Data|M_1)}{p(Data)}\ p(M_1),$$
$$p(M_2|Data) = \frac{p(Data|M_2)}{p(Data)}\ p(M_2).$$

Taking the logarithms of the ratio of both relationships yields

$$\log \frac{p(M_1|Data)}{p(M_2|Data)} = \log \frac{p(Data|M_1)}{p(Data|M_2)} + \log \frac{p(M_1)}{p(M_2)}. \tag{29}$$

The left-hand side is the information *after* the measurement; the right-hand side represents the information contributions of both sources. Hence:

Fact 18 (Bayes' rule is optimal information processor) *Bayes' rule has equal information "before" its application as "after," and hence is optimal in the sense that it doesn't add nor delete information, [37].*

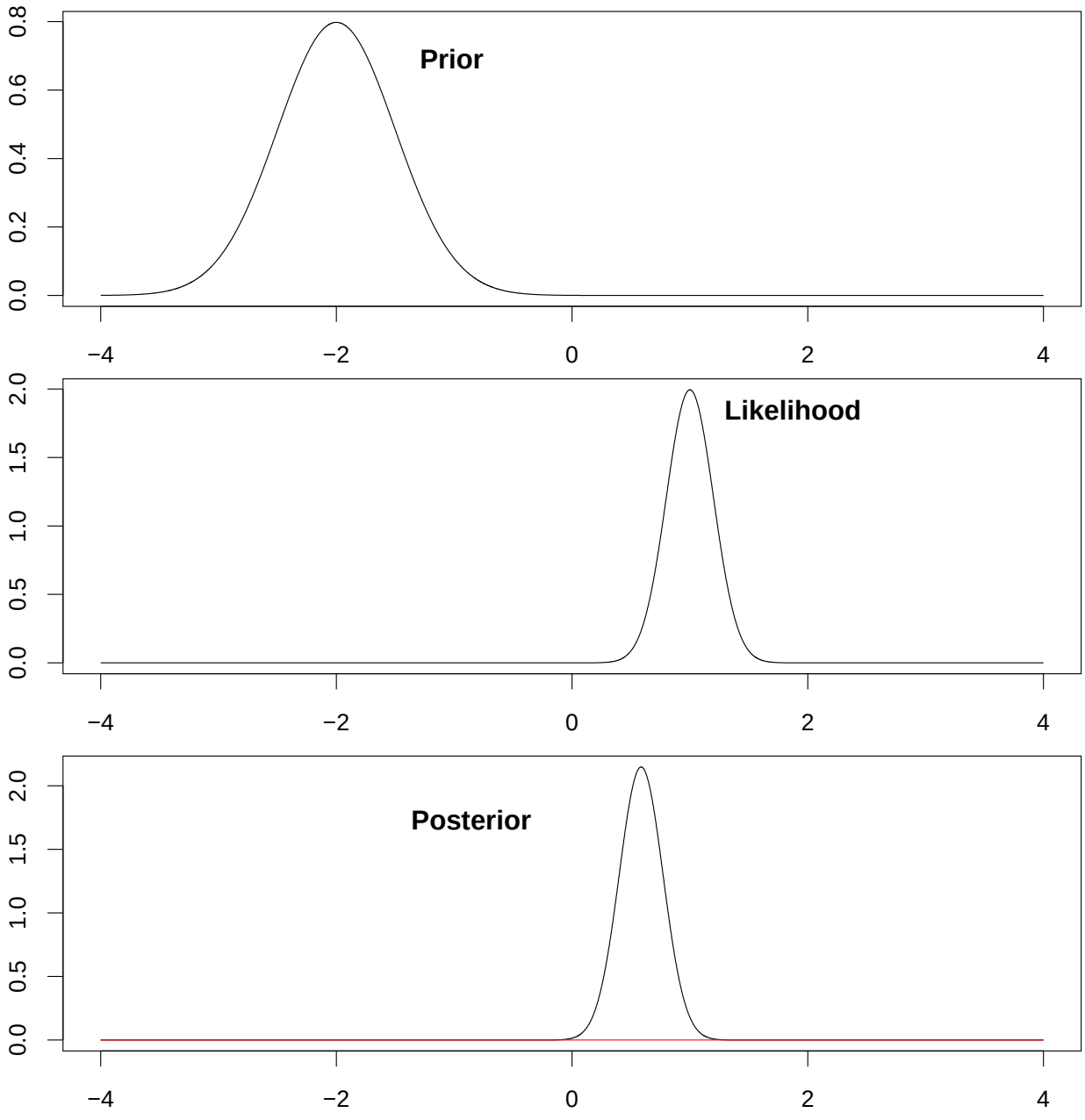

Figure 7: Example of Bayes' rule, applied to two Gaussian PDFs.

4 Estimation

Estimation is the activity of taking as input the PDF that represents the information about a system, and perform a *data reduction* to one single scalar that is believed to summarize the information about this parameter. By the way we define estimation, it is clear that some arbitrary choices are involved.

In general, the reasoning system should build the joint PDF over all model and data parameters, as the complete source of information about the whole system. However, the dimension of this PDF would grow unbounded for systems that receive new information on a regular basis. Hence, one reduces these high-dimensional PDFs to lower-dimensional ones, in a way that one can hope to keep most of the information in this data reduction step.

Marginalization is one way to reduce the amount of representation data; estimators are another way. The following types of estimators are most widely used:

Maximum Likelihood (ML) The PDF over the parameters that represent the Model information is replaced by that parameter combination that has the highest value in the likelihood function.

Figure 8 depicts an ML estimate. It also shows that ML only works reliably for so-called *unimodal* PDFs: PDFs that have only one "peak." But even in that case, the ML estimate could be a poor data reduction; e.g., in the case of a very "flat" PDF, there are many parameters with almost the same function value, so the maximum is poorly conditioned numerically.

Maximum a Posteriori (MAP) Instead of looking only at the likelihood function, the MAP estimator considers the complete posterior PDF. However, one usually looks for the maximum of $p(x)$, while one should look for the *interval* I in parameter space on which the integral $\int_I p(x)\,dx$ is largest, [26].

Least Squares (LS) This is the minimum of the "squared error function," i.e., the function $(\hat{x}-x)^T(\hat{x}-x)$. It seems at first sight that this estimator involves no probability distributions on the parameter(s) x; but the squared error is proportional to the exponent of a Gaussian distribution with equal covariance in all parameters.

Mean value The integral $\int p(x)\,dx$ gives the average or mean value of the PDF $p(x)$. This is also one scalar, and hence a valid procedure for doing data reduction. Figure 8 also shows that the mean is a poor choice for a multi-modal PDF.

Some of these estimators become equivalent in special cases:

- ML = MAP if the prior is noninformative, i.e., "flat."
- LS = ML, for indentically distributed and independent ("*iid*"), symmetric and zero mean distributions.
- ML = MAP = mean for Gaussian PDFs.

5 Hypothesis tests

Many reasoning systems are asked to answer the question whether it has enough information to decide in favour of a particular "hypothesis" in the system. In the Bayesian framework, the validity of an

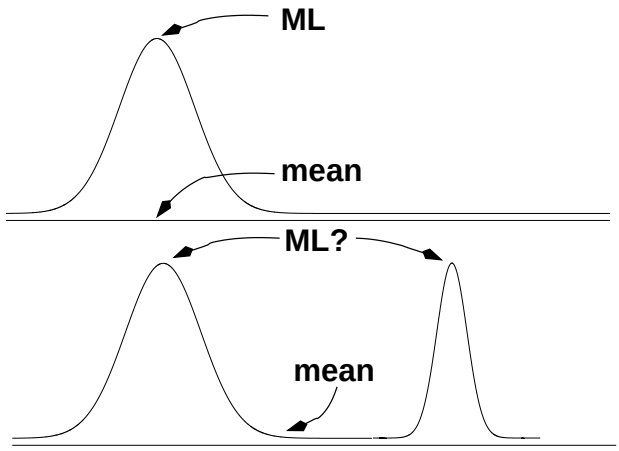

Figure 8: Example of Maximum Likelihood estimation (top). In the bottom figure, i.e., the case of a multi-modal PDF, the ML and mean are not so good estimators.

hypothesis cannot be determined *absolutely*; only the *relative* quantification of hypotheses is meaningful. Equation (29), derived from Bayes' rule in a previous Section, is the basis for *hypothesis* testing. Written without the logarithms one gets:

$$\frac{p(M_1|Data)}{p(M_2|Data)} = \frac{p(Data|M_1)}{p(Data|M_2)} \frac{p(M_1)}{p(M_2)}. \tag{30}$$

This relationship answers the question whether the information available in the reasoning system prefers model M_1 over model M_2, depending on whether the left-hand side ratio is larger or smaller than 1. Or rather, it *seems* to answer this question, because taking a closer look reveals some problems:

- The terms in Eq. (30) are not scalar numbers, but *functions*.
- In addition, the models M_1 and M_2 need not even be defined over the same parameters spaces. M_1 could have parameters x, y and z, while M_2 has parameters a and b. So, the ratios in the equation have no unambiguous meaning. (In contrast to the PDF ratio in the likelihood function of Bayes' rule, which *is* a physically meaningful ratio.)

This means that Eq. (30) has only *qualitative* value, but is quantitatively wrong: the ratio of any two PDF functions doesn't make sense, in general. However, the particular ratio involving the Data PDFs does make sense, because both PDFs are defined over the same Data parameter space.

So, a data reduction step must be performed on some of the PDFs in Eq. (30), using *information measures* to transform the PDFs into scalars that can be multiplied, divided, and compared. In this respect, hypothesis tests are, in fact, the same things as parameter estimators, [25, p. 104]: estimating a parameter to have a certain value is the same problem as testing the hypothesis that it has that value against the hypotheses that it has other values.

The relative probability of two Models not only depends on how well they predict the Data (i.e., the first term on the right-hand side), but also on how *complex* the models are (i.e., represented by the second

term). The predictive power and the complexity are always in a trade-off. More complex models have a higher probability to fit the Data better. Simpler models have less "degrees of freedom," and hence the probability mass can be spread less. So, $p(M|H)$ will be higher for a simpler model M. This principle is often called *Occam's razor*, after (the Latin name of) the Englishman William of Ockham (1288–1348), who got famous (long after his death!) for his quotations "*Frustra fit per plura, quod fieri potest per pauciora* (It is vain to do with more what can be done with less) and "*Essentia non sunt multiplicanda praeter necessitatem*" (Entities should not be multiplied unnecessarily).

6 Model building

Some of the more advanced reasoning systems have the goal of doing the *modelling* part of a Bayesian system autonomously, i.e., without supervision by a human operator. However, these so-called *model building* systems rely on estimation and hypothesis testing. Indeed, they start from a set of "modelling primitives", with which to build models. Then, model building corresponds to finding which combination of primitives fits "best" with the data. The system typically also is given some algorithms or procedures to define which combinations of primitives should be looked at.

7 Approximations and algorithms

The previous Sections have made clear that Bayesian theory is a very attractive framework to use for reasoning about uncertain knowledge-based systems. However, applying the Bayesian information processing methods is quite often not straigthforward, because of the computational complexity of working with high-dimensional PDFs. Hence, many people have been motivated to find computationally tractable ways to perform these Bayesian information processing operations. This Section gives an overview of the most general and successful approximations and algorithms.

7.1 Markov assumption

Most reasoning systems are *iterative*: they get new information at regular time intervals. Every time a new piece of information arrives, they do an information update through Bayes' rule. In general, the right-hand side part in all PDFs in the system (i.e., behind the condition bar "|") should contain all received information packages. This would make the PDFs inefficiently complex, so one often *assumes* that this history should only be kept for N steps in the past. This assumption is called the *Markov assumption*. And most often, $N = 1$. The result of such a Markov assumption is that

$$p\left(x(k)|H, x(k-1), x(k-2), \dots\right) = p(x(k)|H, x(k-1)). \tag{31}$$

7.2 Markov model

Some systems have a simple state space representation, as depicted in Figure 9: only a limited number of discrete states exist, and *transition probabilities* a_{ij} exist that represent the probability that, in the next time instant, the state will change from i to j:

$$p(x(k) = \text{State}j|x(k-1) = \text{State}i) = a_{ij}. \tag{32}$$

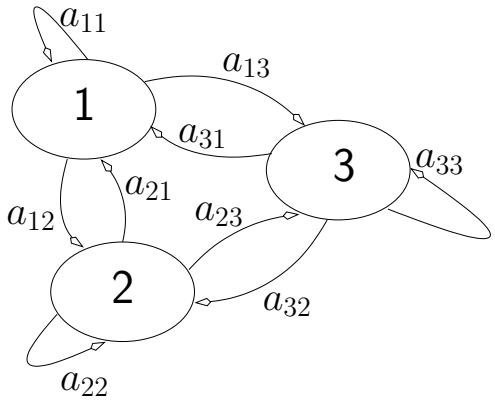

Figure 9: A Markov model system.

7.3 Hidden Markov model

A Hidden Markov Model (HMM) is a generalization of the Markov Model, where the "hidden" refers to the fact that not all states can be observed *directly*. Only some of them can, while the others have to be *estimated* from a PDF on the measurements that each of the hidden states generates. So, Figure 9 must be extended with measurement probabilities m_i: this is the probability of measuring m if it was the state i that has generated the measured signal.

HMMs are popular in signal processing tasks such as speech recognition: when a human pronounces a sentence, the speech computer can only measure the signals that are produced in the phonemes (= units of speech) uttered by the speaker, but it cannot measure the letters or words directly.

So, (Hidden) Markov Models are just that: models. Some efficient algorithms have been developed to do the estimation and information processing of systems modelled by HMMs (see below).

7.4 Kalman Filter

The Kalman Filter, [20, 35], is one of the most efficient Bayesian estimators. It was developed by Kalman around 1960, without any mention of Bayesian theory, but on the basis of least-squares estimation. However, with the background material provided by the previous Sections, it's very easy to fully describe the Kalman Filter (Fig. 10):

- The system is time sampled, where subsequent samples get the index k and $k+1$.
- *Process equation.* The state $x(k)$ of the system has an evolution over time which can be represented by a *linear* relationship: $x(k+1) = F\,x(k)$.
- *Measurement equation.* The mapping from state $x(k)$ to measurement $z(k)$ is also linear: $z(k) = H\,x(k)$.
- The *uncertainty* on the state, its evolution over time, and the measurements are represented by Gaussian PDFs, i.e., by their means and covariance matrices.

Then, the Kalman Filter is nothing else but Bayes' rule applied to this particular kind of model. The *innovation* in Fig. 10 is the difference between the *predicted* value of the measurement and the actual value; this is indeed the new information, because one learns nothing new from predictions that turn out to correspond to the reality.

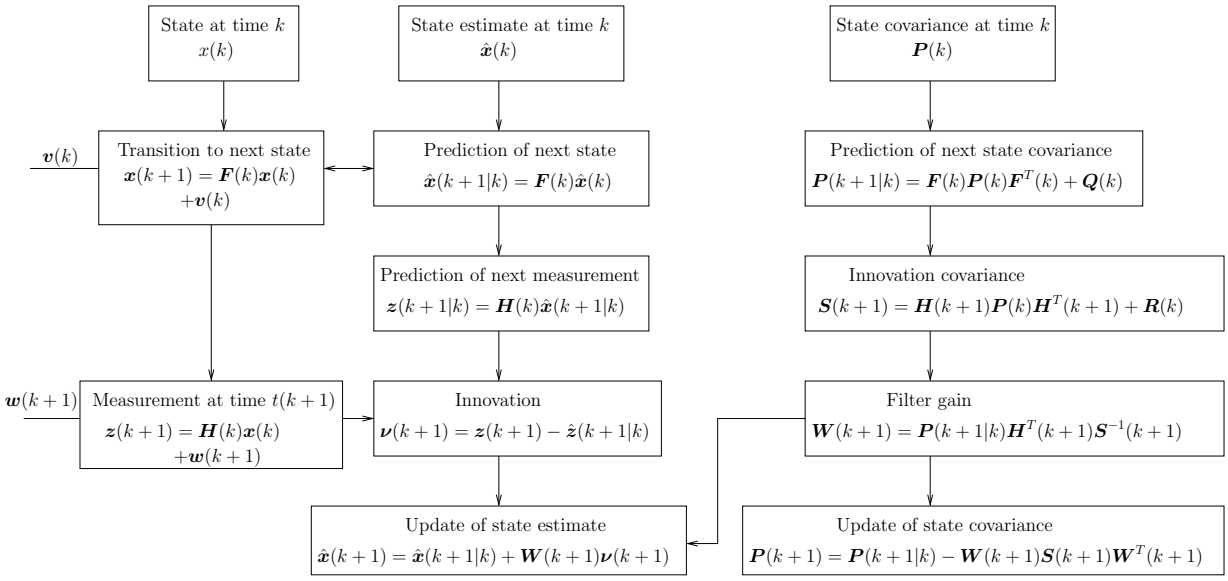

Figure 10: Computational scheme of the Kalman Filter, [1].

A Kalman Filter is an *information processor*, but also trivially useable as an estimator, because, for Gaussian PDFs, the mean of the PDF corresponds to the Maximum Likelihood estimate.

Note that the Kalman Filter works on full PDFs at all times (but these PDFs are analytically simple), and hence within the model no information is created or lost. Of course, this only holds as long as the linear process and measurement models are reliable. And quite often they are only first-order approximations.

Kalman Filters are the information processing workhorses in the aviation and aerospace applications: they are used for over fourty years to keep rockets, airplanes and satellites in their correct course. Kalman Filters are also very popular in robotics applications, such as autonomously navigating mobile robots (Fig. 11).

7.5 Viterbi algorithm

This algorithm is used as an *approximation* to the full Bayesian calculus for systems that can be modelled by an Hidden Markov Model.

Message 8 (Data reduction through Maximum Likelihood) *Many Bayesian algorithms tackle the complexity involved in the full application of Bayes' rule, by reducing the information contained in a PDF by only one single parameter, i.e., the Maximum Likelihood estimate.*

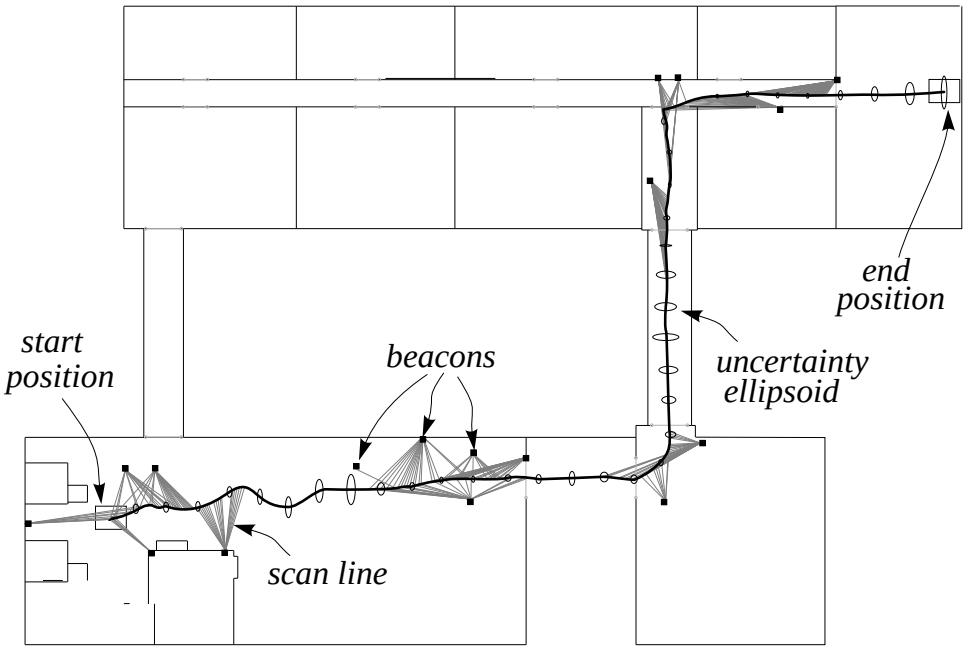

Figure 11: Mobile robot navigation based on tracking of beacons in the environment, by means of a Kalman Filter. Whenever the robot is able to see beacons in its environment (and is able to match them with its a priori map) it can reduce its position uncertainty. This is reflected by the covariance ellipses shown in the picture. (Figure courtesy of J. Vandorpe.)

Message 9 (ML and local maxima) *Every estimation or information processing algorithm that uses not the full PDFs but only a Maximum Likelihood (or any other) estimate, is prone to converge to a* local *maximum, and not the global maximum.*

The Viterbi algorithm is an example of such a ML-based approach. The underlying system is modelled by an HMM, and data correlated with the state transitions is coming in at every sample instant. The goal of the reasoning system is to estimate which state transition *sequence* has generated the measured signals. Instead of calculating a PDF over *all possible* transition paths that could have led from the past to the current measurement, the Viterbi algorithm only stores the *most probable* path.

7.6 Expectation-Maximization (EM)

The EM algorithm, [7], is another estimator that only keeps track of the Maximum Likelihood, in order to reduce the computational complexity of the full Bayesian approach, for a situation where not all states are directly observable. The "E" and "M" stand for the two steps in the algorithm that are executed sequentially:

- *Expectation.* Based on the latest ML estimates of the states, the expected measurements are predicted.
- *Maximization.* These predicted measurements are compared to the real measurements, and from the difference the state estimate is adapted, in a Maximum Likelihood sense.

Figure 12 depicts this EM loop, in two snapshots of the reasoning system inside a mobile robot that navigates in an unknown environment, building a map of that environment. The first figure shows still mapping inaccuracies, due to the inaccurate self-motion measurements of the mobile robot. The second row shows a better converged estimate at a later point in time.

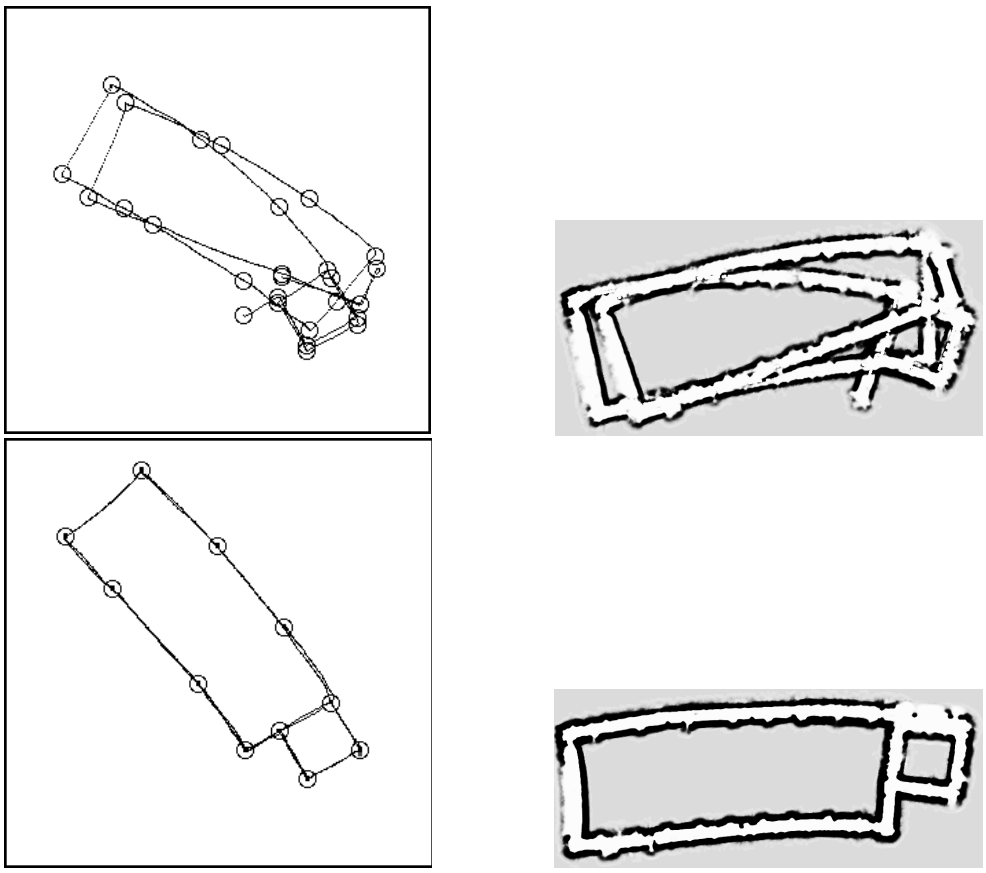

Figure 12: Intermediate and final topological maps and occupancy grids during an EM algorithm for model building by a mobile robot.

Figure 13: Sample-based information processing for a mobile robot that wants to localize itself in an office like environment of which it has a map. The differently coloured robots and the correspondingly coloured samples represent the information at subsequent instants in time.

7.7 Sample-based information

The major drawback of approximating a PDF by its Maximum Likelihood estimate is that the information processing can lead to a wrong optimum. Hence, other techniques have been developed, aiming at both computational efficiency and sufficient "coverage" of the whole parameter space of the PDFs. The sample-based methods are becoming increasingly popular in this respect.

Fact 19 (Sampled representation of PDF) *Any PDF can be approximated by a number of samples, i.e., at areas I in the parameter space where $\int_I p(x)\,dx$ is high, one puts more samples than where this integral is low. All operations on PDFs are then replaced by operations on the individual samples.*

Fact 20 (Monte Carlo) *The name* Monte Carlo *is associated with many sample-based methods.*

The animated image at `http://www.cs.washington.edu/ai/Mobile_Robotics/mcl/` and in Figure 13 show Bayes' rule in action for a mobile robot navigation and localization task.

Fact 21 (Sampling from uniform interval) *The only computationally efficient sampling algorithms sample from a uniform distribution over the interval $[0, 1]$. Sampling from other PDFs is always transformed, via various routes, to such a sampling.*

The *Cumulative Density Function (CDF)* is an important concept in sampling, Fig. 14. The CDF of any given PDF can be used to generate samples from that PDF, via the *Inverse CDF* sampling procedure, Fig. 15: one takes uniform samples from the Y axis of the CDF plot, and each of the inverse function values is a sample from the original PDF.

Fact 22 (When is sampling possible?) *The PDF must be "easy" to evaluate, i.e., the value of $p(x)$ should be fast to calculate.*

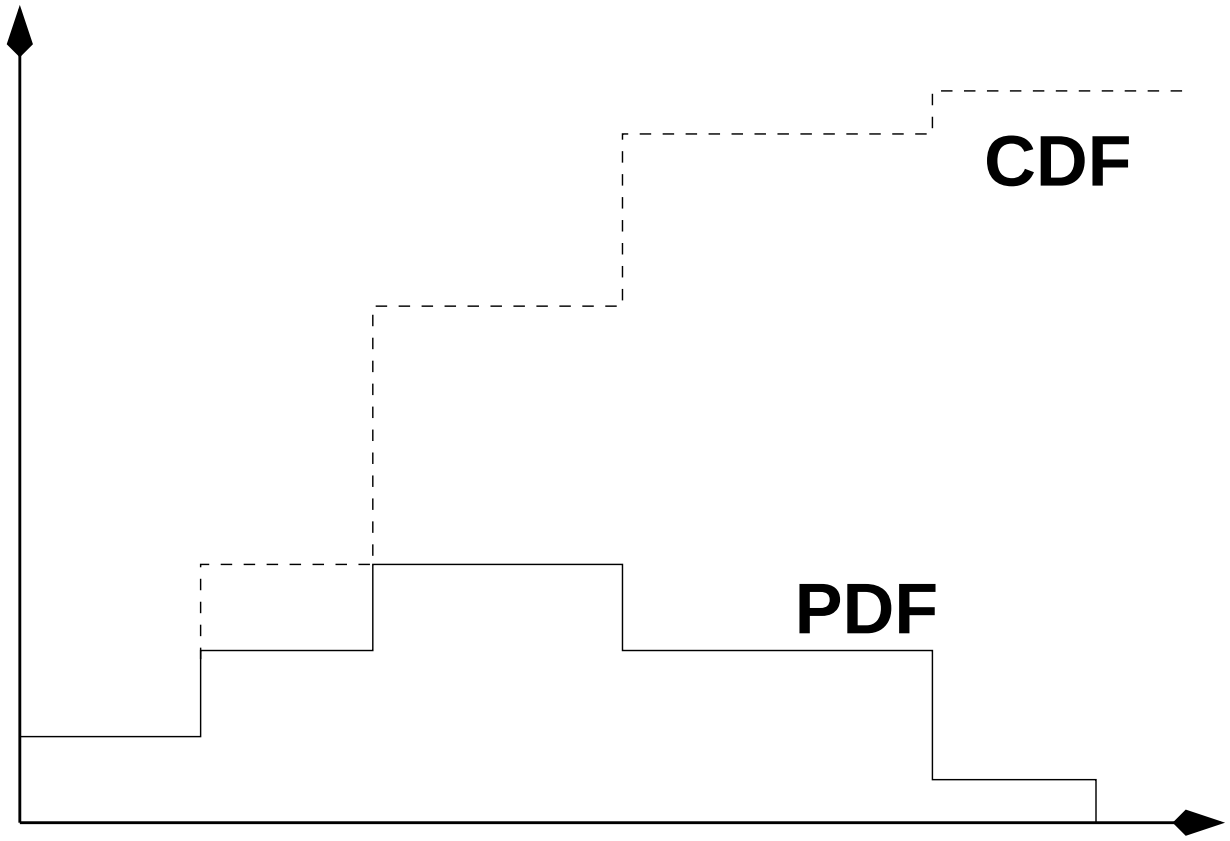

Figure 14: The Cumulative Density Function of the given PDF is the function that corresponds to the *integral* of the PDF along its domain.

The ICDF sampling is simplest for discrete or discretized PDFs; it can be quite complicated for a general analytical PDF. So, discrete *approximations* to these PDFs are most often used.

Fact 23 (Normalization of samples) *After each information processing step, it is common practice to* normalize *the samples. That is, instead of keeping information about* $p(x_i)\,dx$ *one stores a number of samples of "unit" magnitude, where the number corresponds to the integral* $\int_I p(x)\,dx$.

The "information processing step" can be: (i) transformation of the type $y = f(x)$, or (ii) the application of Bayes' rule.

Fact 24 (Monte Carlo integration) *The Bayesian framework uses a lot of integrations, for example, for doing marginalizations. Monte Carlo integration is a sample-based approximation of a real integral, formed by the sum of samples of the PDF, normalized by their "density."*

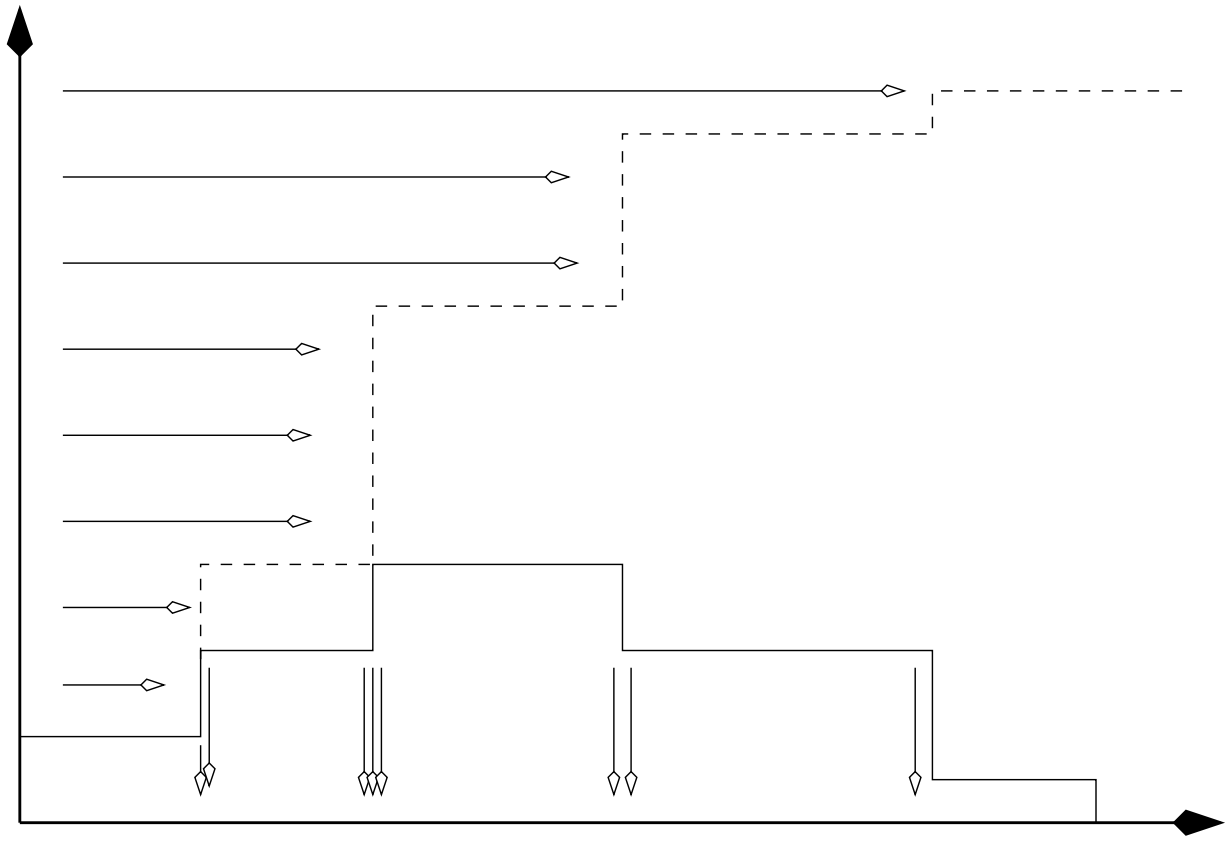

Figure 15: Sampling from any PDF can be done by "inverse" uniform sampling from its CDF.

References

[1] Y. Bar-Shalom and X.-R. Li. *Estimation and Tracking, Principles, Techniques, and Software*. Artech House, 1993.

[2] T. Bayes. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764. Reprinted in *Biometrika*, 45:293–315, 1958, and in *Fascimiles of two papers by Bayes*, W. Edwards Deming.

[3] G. L. Bretthorst. An introduction to parameter estimation using Bayesian probability theory. In P. F. Fougère, editor, *Maximum Entropy and Bayesian Methods*, pages 53–79. Kluwer, Dordrecht, The Netherlands, 1990.

[4] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, New York, NY, 1991.

[5] R. T. Cox. Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946. Reprinted in [30, p. 353].

[6] P.-S. de Laplace. *Théorie analytique des probabilités*. Courcier Imprimeur, 1812. 2nd edition, 1814; 3rd edition, 1820.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[8] G. J. Erickson and C. R. Smith, editors. *Maximum-Entropy and Bayesian Methods in Science and Engineering. Vol. 1: Foundations; Vol. 2: Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[9] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, Scotland, 13th edition, 1967.

[10] M. Ginsberg. *Essentials of Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA, 1993.

[11] G. Golub and C. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1989.

[12] I. J. Good. A derivation of the probabilistic explanation of information. *Journal of the Royal Statistical Society Ser. B*, 28:578–581, 1966.

[13] E. T. Jaynes. How does the brain do plausible reasoning? Technical Report 421, Stanford University Microwave Laboratory, 1957. Reprinted in [8, Vol. 1, p. 1–24].

[14] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957. Reprinted in [29, p. 6–16].

[15] E. T. Jaynes. Prior probabilities. *IEEE Trans. Systems Science and Cybernetics*, 4:227–241, 1968. Reprinted in [29, p. 116–130].

[16] E. T. Jaynes. Where do we stand on Maximum Entropy? In R. D. Levine and M. Tribus, editors, *The maximum entropy formalism*, pages 15–118. MIT Press, 1978. Reprinted in [29, p. 211–314].

[17] E. T. Jaynes. Bayesian methods: general background. In J. H. Justice, editor, *MaxEnt'84: Maximum Entropy and Bayesian Methods in Geophysical Inverse Problems*, pages 1–25. Cambridge University Press, 1986.

[18] E. T. Jaynes. Probability theory: The logic of science. Unfinished manuscript, `http://bayes.wustl.edu/etj`., 1996.

[19] H. Jeffreys. *Theory of Probability*. Clarendon Press, 1939. 2nd edition, 1948; 3rd edition, 1961. Reprinted by Oxford University Press, 1998.

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.*, 82:34–45, 1960.

[21] M. G. Kendall and A. O'Hagan. *Kendall's advanced theory of statistics. 2B: Bayesian inference*. Arnold, London, England, 1994.

[22] S. Kullback. *Information theory and statistics*. Wiley, New York, NY, 1959.

[23] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

[24] D. V. Lindley. *Introduction to probability and statistics from a Bayesian viewpoint. Vol. 1: Probability, Vol. 2: Inference*. Cambridge University Press, 1965.

[25] T. J. Loredo. From Laplace to supernova SN 1987a: Bayesian inference in astrophysics. In P. F. Fougère, editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer, Dordrecht, The Netherlands, 1990.

[26] D. J. C. MacKay. Information theory, inference and learning algorithms. Textbook in preparation. `http://wol.ra.phy.cam.ac.uk/mackay/itprnn/`, 1999.

[27] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematics Society*, 37:81–91, 1945.

[28] C. C. Rodríguez. The metrics induced by the Kullback number. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 415–422. Kluwer, 1989.

[29] R. D. Rosenkrantz, editor. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. D. Reidel, 1983. Second paperbound edition, Kluwer Academic Publishers, 1989.

[30] G. Shafer and J. Pearl, editors. *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Mateo, CA, 1990.

[31] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[32] C. E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, Urbana, IL, 1949.

[33] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Information Theory*, 26(1):26–37, 1980. Error corrections in Vol. 29, No. 6, pp. 942–943, 1983.

[34] J. Skilling. The axioms of Maximum Entropy. In Erickson and Smith [8], pages 173–187 (Vol. 1).

[35] H. W. Sorenson. Least-squares estimation from Gauss to Kalman. *IEEE Spectrum*, 7:63–68, 1970.

[36] G. Strang. *Calculus*. Wellesley-Cambridge Press, Wellesley, MA, 1991.

[37] A. Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42:278–284, 1988.