1

Graphical models and probabilistic reasoning

1.1 Introduction

This text considers the subject of *graphical models*, which is an interaction between probability theory and graph theory. The topic provides a natural tool for dealing with a large class of problems containing uncertainty and complexity. These features occur throughout applied mathematics and engineering and therefore the material has diverse applications in the engineering sciences. A complex model is built by combining simpler parts, an idea known as *modularity*. The *uncertainty* in the system is modelled using probability theory; the graph helps to indicate independence structures that enable the probability distribution to be decomposed into smaller pieces.

Bayesian networks represent joint probability models among given variables. Each variable is represented by a node in a graph. The direct dependencies between the variables are represented by *directed edges* between the corresponding nodes and the conditional probabilities for each variable (that is the probabilities conditioned on the various possible combinations of values for the immediate predecessors in the network) are stored in potentials (or tables) attached to the dependent nodes. Information about the observed value of a variable is propagated through the network to *update the probability distributions* over other variables that are not observed directly. Using *Bayes' rule*, these influences may also be identified in a 'backwards' direction, from dependent variables to their predecessors.

The *Bayesian* approach to uncertainty ensures that the system as a whole remains consistent and provides a way to apply the model to data. Graph theory helps to illustrate

Bayesian Networks: An Introduction T. Koski, J. Noble

and utilize independence structures within interacting sets of variables, hence facilitating the design of efficient algorithms.

In many situations, the directed edges between variables in a Bayesian network can have a simple and natural interpretation as graphical representations of *causal* relationships. This occurs when a graph is used to model a situation where the values of the immediate *predecessors* of a variable in a network are to be interpreted as the immediate *causes* of the values taken by that variable. This representation of causal relationships is probabilistic; the relation between the value taken by a variable and the values taken by its predecessors is specified by a conditional probability distribution. When a graph structure is given and the modelling assumptions permit a causal interpretation, then the estimates of the conditional probability tables obtained from data may be used to infer a system of causation from a set of conditional probability distributions. A Bayesian network is essentially a directed acyclic graph, together with the associated conditional probability distributions. When a Bayesian network represents a causal structure between the variables, it may be used to assess the effects of an intervention, where the manipulation of a cause will influence the effect.

The ability to infer causal relationships forms the basis for learning and acting in an intelligent manner in the external world. Statistical and probabilistic techniques may be used to assess direct associations between variables; some additional common sense and modelling assumptions, when these are appropriate, enable these direct associations to be understood as direct *causal* relations. It is the knowledge of causal relations, rather than simply statistical associations, that gives a sense of genuine understanding, together with a sense of potential control resulting from the ability to predict the consequences of actions that have not been performed, as J. Pearl writes in [1].

K. Pearson, an early, pre-eminent statistician, argued that the only proper goal of scientific investigation was to provide descriptions of experience in a mathematical form (see [2] written by Pearson in 1892); for example, a coefficient of correlation. Any effort to advance beyond a description of associations and to deduce *causal* relations meant, according to this view, to evoke hidden or metaphysical ideas such as causes; he did not consider such modelling assumptions to be scientific.

R.A. Fisher, possibly the most influential statistician, considered that causation could be inferred from experimental data only when controlled, or randomized experiments were employed. A majority of statistical studies follow the approach of Fisher and only infer ‘correlation’ or ‘association’ unless randomized experimental trials have been performed.

It is not within the scope of this treatment of Bayesian networks to review the sophisticated attempts at characterizing causality and the ensuing controversies (starting with David Hume [3]) amongst scholars; the reader is referred to the enlightening treatment by J. Williamson [4]. This text attempts to take what seems to be a ‘common sense’ point of view. The human mind is capable of detecting and approving causes of events in an intuitive manner. For example, the nineteenth century physician Ignaz Semmelweis in Vienna investigated, without knowing about germs, the causes of child bed fever. He instituted a policy that doctors should use a solution of chlorinated lime to wash their hands between autopsy work and the examination of patients, with the effect that the mortality rate at the maternity wards of hospitals dropped substantially. Such reasoning about causes and effects is not as straightforward for computers and it is not clear that it is valid in terms of philosophical analysis.

Causality statements are often expressed in terms of large and often complicated objects or populations: for example, ‘smoking causes lung cancer’. Causal connections at finer levels of detail will have a different context: for example, the processes at the cell level that cause lung cancer. Causal connections are also contingent on many other conditions and causal laws than those explicitly under consideration. This introduces a level of uncertainty and the causal connections are therefore probabilistic.

Correlations or statistical associations between two variables often imply causation, even if there is not a direct causal relation between the two variables in question; one need not be the cause of the other. A correlation may indicate the presence of hidden variables, that are common causes for both the observed variables, so that the two observed variables are statistically associated.

When there is a possibility that there may be such unknown hidden variables, it is necessary to separate the ‘cause’ from the extraneous factors that may influence it in order to conclude that there is a causal relationship and a randomized (or controlled) experiment achieves this. For a randomized experiment, the groups for each level of treatment, and a control group to which no treatment is applied, are chosen at random so that the allocation of members to treatment groups is not affected by any hidden variables. Unfortunately, there are situations where it may be unethical to carry out a randomized experiment. For example, to prove that smoking causes lung cancer, it is perhaps inappropriate to force a non-smoker to start smoking.

In the example of smoking and lung cancer, the model is unknown and has to be inferred from data. The statistical analyst would like to establish whether smoking causes lung cancer, or whether there are additional hidden variables that are causes for both variables. Note that common sense plays a role here; the possibility that lung cancer may cause smoking is not considered. In terms of graphical models, where the direction of the pointed arrow indicates cause to effect, the analyst wants to determine which of the models in Figure 1.1 are appropriate.

When he carries out a controlled experiment, randomly assigning people to ‘smokers’ and ‘non-smokers’ respectively, the association between the hidden variables and smoking for this experiment are broken and the causal diagram is therefore given by Figure 1.2.

By carrying out a ‘controlled experiment’, an intervention is made whereby the causal path from the hidden variable is removed, thus ensuring that only the causal connection of interest is responsible for an observed association.

In many situations, a controlled experiment seems the only satisfactory way to demonstrate conclusively that an association between variables is due to a causal link. Without a controlled experiment, there may remain some doubt. For example, levels of smoking dropped when the first announcements were made that smoking caused lung cancer

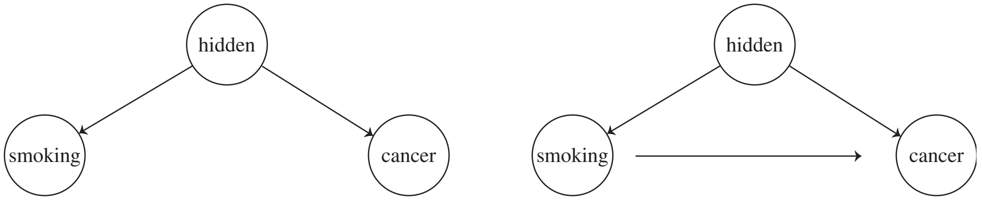

Figure 1.1 Smoking and lung cancer.

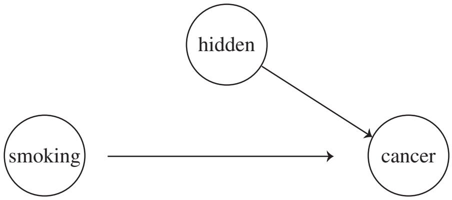

Figure 1.2 Controlled experiment: the control has removed the hidden causes of smoking from consideration.

and levels of lung cancer also dropped substantially. A controlled experiment would have demonstrated conclusively that the drop in lung cancer was not due to other environmental factors, such as a decline in heavy polluting industry that occurred at the same time.

Bayesian networks provide a straightforward mathematical language to express relations between variables in a clear way. In many engineering examples, the variables that should be present in the model are well defined. From an appropriate model that contains the hidden (or non-observable) variables and the observable variables, and where it is clear which variables may be intervened on, it will be possible to verify whether certain 'identifiability' conditions hold and hence to conclude whether or not there is a causal relation from the data, without a controlled experiment.

Many of the classical probabilistic systems studied in fields such as systems engineering, information theory, pattern recognition and statistical mechanics are problems that may be expressed as graphical models. Hidden Markov models may be considered as graphical models. Engineers are also accustomed to using circuit diagrams, signal flow graphs, and trellises, which may be treated using the framework of graphical models.

Examples of applied fields where Bayesian networks have recently been found to provide a natural mathematical framework are reliability engineering [5], software testing [6], cellular networks [7], and intrusion detection in computer systems [8]. In 1996, William H. Gates III, a co-founder of Microsoft, stated that expertise in Bayesian networks had enhanced the competitive advantage of Microsoft. Bayesian networks are used in software for electrical, financial and medical engineering, artificial intelligence and many other applications.

1.2 Axioms of probability and basic notations

The basic set notations that will be used will now be established.

Definition 1.1 (Notations) The following notations will be used throughout:

- The *universal set* will be denoted by Ω. This is the *context* of the experiment. In Bayesian analysis, the *unknown parameters* are considered to be random. Therefore, the context Ω consists of the set of all possible outcomes of an experiment, *together with all possible values of the unknown parameters.*
- The notation $\mathcal{X}$ will be used to denote the sample space; the set of all possible outcomes of the experiment.
- $\tilde{\Theta}$ will be used to denote the parameter space, so that $\Omega = \mathcal{X} \times \tilde{\Theta}$.

The following notations will be used when considering sets:

- If A and B are two sets, then $A \cup B$ or $A \vee B$ denotes their union. If $A_1, \dots, A_n$ are a finite collection of sets, then $\cup_{j=1}^n A_j$ or $\bigvee_{j=1}^n A_j$ denotes their union. Also, $A_1 \cup \dots \cup A_n$ or $A_1 \vee \dots \vee A_n$ may be used to denote their union.
- If A and B are two sets, then $A \cap B$ or AB or $A \wedge B$ may be used to denote their intersection. If $A_1, \dots, A_n$ are a finite collection of sets, then $A_1 \dots A_n$, $A_1 \cap \dots \cap A_n$, $\cap_{j=1}^n A_j$ or $A_1 \wedge \dots \wedge A_n$ all denote their intersection.
- $A \subset B$ denotes that A is a strict subset of B. $A \subseteq B$ denotes that A is a subset of B, possibly equal to B.
- The empty set will be denoted by ϕ.
- A^c denotes the complement of A; namely, $\Omega \backslash A$, where Ω denotes the universal set. The symbol \ denotes exclusion.
- Together with the *universal set* Ω, an *event space* $\mathcal{F}$ is required. This is a collection of subsets of Ω. The *event space* $\mathcal{F}$ is an *algebra* of *constructable sets*. That is, $\mathcal{F}$ satisfies the following:
  1. $\phi \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.
  2. Each $A \in \mathcal{F}$ may be constructed. That is, for each $A \in \mathcal{F}$, let $i_A : \Omega \to \{0, 1\}$ denote the mapping such that $i_A(\omega) = 1$ if $\omega \in A$ and $i_A(\omega) = 0$ if $\omega \in A^c$. Then there is a procedure to determine the value of $i_A(\omega)$ for each $\omega \in \Omega$.[1]
  3. If for a finite collection of events $(A_j)_{j=1}^n$ each A_j satisfies $A_j \in \mathcal{F}$, then $\cup_{j=1}^n A_j \in \mathcal{F}$.
  4. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \Omega \backslash A$ denotes the complement of A.
  5. Each $A \in \mathcal{F}$ satisfies $A \subseteq \Omega$.

For Ω, $\mathcal{F}$ satisfying the above conditions, a *probability distribution* p *over* $(\Omega, \mathcal{F})$ (or, in short, a probability distribution over Ω, when $\mathcal{F}$ is understood) is a function $p : \mathcal{F} \to [0, 1]$ satisfying the following version of the *Kolmogorov axioms*:

1. $p(\phi) = 0$ and $p(\Omega) = 1$.
2. If $(A_j)_{j=1}^n$ is a finite collection such that each $A_j \in \mathcal{F}$ and the events satisfy $A_j \cap A_k = \phi$ for all $j \neq k$, then
$$p\left(\cup_{j=1}^n A_j\right) = \sum_{j=1}^n p(A_j).$$
3. $0 \leq p(A) \leq 1$ for all $A \in \mathcal{F}$.

[1] **A non-constructable set** The following example illustrates what is intended by the term *constructable* set. Let $A_{n,k} = (\frac{k}{2^n} - \frac{1}{2^{3(n+1)}}, \frac{k}{2^n} + \frac{1}{2^{3(n+1)}})$ and let $A = \cup_{n=1}^\infty \cup_{k=0}^{2^n+1} A_{n,k}$. Then A is not constructable in the sense given above. For any number $x \in A \cap [0, 1]$, there is a well defined algorithm that will show that it is in $A \cap [0, 1]$ within a finite number of steps. Consider a number $x \in [0, 1]$ and take its dyadic expansion. Then, for each $A_{n,k}$ it is clear whether or not $x \in A_{n,k}$. Therefore, if $x \in A$, this may be determined within a finite number of steps. But if $x \in A^c$, there is no algorithm to determine this within a finite number of steps.

This is a reduced version of the Kolmogorov axioms. The Kolmogorov axioms require countable additivity rather than finite additivity. They invoke axiomatic set theory and do not therefore require the constructive hypothesis. For the Kolmogorov axioms, the event space $\mathcal{F}$ is taken to be a *sigma*-algebra and the second axiom requires *countable* additivity.

- Let $(\theta, x) \in \Omega = \tilde{\Theta} \times \mathcal{X}$. For each $A \in \mathcal{F}$, let $A_\Theta = \{\theta \in \tilde{\Theta} | (\theta, x) \in A\}$ and let $A_\mathcal{X} = \{\mathbf{x} \in \mathcal{X} | (\theta, x) \in A\}$. Then $\mathcal{F}_\Theta$ will be used to denote the algebra $\{A_\Theta | A \in \mathcal{F}\}$ and $\mathcal{F}_\mathcal{X} = \{A_\mathcal{X} | A \in \mathcal{F}\}$.

Definition 1.2 (Probability Distribution over $\mathcal{X}$) *If $\mathcal{X}$ contains a finite number of elements, then $\mathcal{F}_\mathcal{X}$ contains a finite number of elements. In this setting, a probability distribution over $\mathcal{X}$ satisfies:*

- $p(\{x\}) \geq 0$ *for all* $x \in \mathcal{X}$,
- *For any* $A \in \mathcal{F}_\mathcal{X}$, $\sum_{x \in A} p(\{x\}) = p(A)$.
- *In particular,* $\sum_{x \in \mathcal{X}} p(\{x\}) = 1$

Definition 1.3 (Notation) *Let $\mathcal{X}$ be a finite state space and let $\mathcal{A}_X$ denote the set of all subsets of $\mathcal{X}$ (including ϕ the empty set and $\mathcal{X}$). For probability distribution $p : \mathcal{A}_X \to [0, 1]$ defined above (Definition 1.2), p will also be used to denote the function $p : \mathcal{X} \to [0, 1]$ such that $p(x) = p(\{x\})$. The meaning of p will be clear from the context.*

The definitions and notations will now be established for random variables.

Definition 1.4 (Random Variables and Random Vectors) *Discrete random variables, continuous random variables and random vectors satisfy the following properties:*

- For this text, a *discrete random variable* Y is a function $Y : \Omega \to \mathcal{C}$ where $\mathcal{C}$ is a countable space, that satisfies the following conditions: for each $x \in \mathcal{C}$, $\{\omega | Y(\omega) = x\} \in \mathcal{F}$, and there is a function $p_Y : \mathcal{C} \to \mathbf{R}_+$, known as the probability function of Y, such that for each $x \in \mathcal{C}$,

  $$p_Y(x) = p(\{\omega | Y(\omega) = x\}).$$

  This is said to be the ‘probability that Y is *instantiated* at x’. Therefore, for any subset $C \subseteq \mathcal{C}$, p_Y satisfies

  $$\sum_{x \in C} p_Y(x) = p(\{\omega | Y(\omega) \in C\}).$$

  In particular, taking $C = \mathcal{C}$,

  $$\sum_{x \in \mathcal{C}} p_Y(x) = 1.$$

- A *continuous random variable* Ξ is defined as a function $\Xi : \Omega \rightarrow \mathbf{R}$ such that for any set $A \subseteq \mathbf{R}$ such that the function $\mathbf{1}_{A\cap[-N,N]}$ is Riemann integrable for all $N \in \mathbf{Z}_+$, the set $\{\omega|\Xi(\omega) \in A\} \in \mathcal{F}$, and for which there is a function π_Ξ, known as the probability density function, or simply density function, such that for any $A \subset \mathbf{R}$ with Riemann integrable indicator function,

  $$p(\{\omega|\Xi(\omega) \in A\}) = \int_A \pi_\Xi(x)dx.$$

  In particular,

  $$p(\{\omega|\Xi(\omega) \in \mathbf{R}\}) = p(\Omega) = \int_{\mathbf{R}} \pi_\Xi(x)dx = 1.$$

- A *random vector* $\underline{Y}$ is a vector of random variables. It will be taken as a *row* vector if it represents different characteristics of a single observation; it will be taken as a *column* vector if it represents a collection of independent identically distributed random variables.
  This convention is motivated by the way that data is presented in a data matrix. Each column of a data matrix represents a different attribute, each row represents a different observation.
  A random row vector $\underline{Y}$ is a collection of random variables that satisfies the following requirements: suppose $\underline{Y} = (Y_1, \ldots, Y_m)$ and for each $j = 1, \ldots, m$, Y_j is a discrete random variable that takes values in a countable space $\mathcal{C}_j$ and let $\mathcal{C} = \mathcal{C}_1 \times \ldots \times \mathcal{C}_m$, then $\underline{Y}$ is a random vector if for each $(y_1, \ldots, y_m) \in \mathcal{C}$, $\{\omega|(Y_1(\omega), \ldots, Y_m(\omega))\} \in \mathcal{F}$, and there is a joint probability function $p_{Y_1,\ldots,Y_m} : \mathcal{C} \rightarrow [0, 1]$ such that for any set $C \subseteq \mathcal{C}$,

  $$\sum_{(y_1,\ldots,y_m)\in C} p_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m) = p(\{\omega|(Y_1(\omega), \ldots, Y_m(\omega)) \in C\}).$$

  In particular,

  $$\sum_{(y_1,\ldots,y_m)\in\mathcal{C}} p_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m) = 1.$$

  If $\underline{\Xi} := (\Xi_1, \ldots, \Xi_n)$ is a collection of n random variables where for each $j = 1, \ldots, n$ Ξ_j is a continuous random variable, then $\underline{\Xi}$ is a random vector if for each set $A \subset \mathbf{R}^n$ such that $\mathbf{1}_{A\cap[N,N]^n}$ is Riemann integrable for each $N \in \mathbf{Z}_+$,

  $$\{\omega|(\Xi_1(\omega), \ldots, \Xi_n(\omega)) \in A\} \in \mathcal{F},$$

  and there is a Riemann integrable function $\pi_{\Xi_1,\ldots,\Xi_n} : \mathbf{R}^n \rightarrow \mathbf{R}_+$, where $\mathbf{R}_+$ denotes the non-negative real numbers such that for each set $A \subset \mathbf{R}^n$ such that $\mathbf{1}_{A\cap[N,N]^n}$ is Riemann integrable for each $N \in \mathbf{Z}_+$,

  $$p(\{\omega|(\Xi_1(\omega), \ldots, \Xi_n(\omega)) \in A\}) = \int_A \pi_{\Xi_1,\ldots,\Xi_n}(x_1, \ldots, x_n)dx_1 \ldots dx_n.$$

In particular,

$$\int_{\mathbf{R}^n} \pi_{\Xi_1,\ldots,\Xi_n}(x_1, \ldots, x_n)dx_1 \ldots dx_n = 1.$$

A collection of random variables $\underline{Y}$ of length $m+n$, containing m discrete variables and n continuous variables, is a random vector if it satisfies the following: there is an ordering σ of $1, \ldots, m+n$ such that $(Y_{\sigma(1)}, \ldots, Y_{\sigma(m)})$ is a discrete random vector and $(Y_{\sigma(m+1)}, \ldots, Y_{\sigma(m+n)})$ is a continuous random vector. Furthermore, let $\tilde{\mathcal{C}}$ denote the state space for $(Y_{\sigma(1)}, \ldots, Y_{\sigma(m)})$, then for each $(y_1, \ldots, y_m) \in \tilde{\mathcal{C}}$, there is a Riemann integrable function $\pi^{(\sigma)}_{Y_{\sigma(m+1)},\ldots Y_{\sigma(m+n)}|y_1,\ldots,y_m} : \mathbf{R}^n \to \mathbf{R}_+$ such that for any set $A \in \mathbf{R}^n$ such that $\mathbf{1}_{A\cap[N,N]^n}$ is Riemann integrable for each $N \in \mathbf{Z}_+$,

$$\begin{aligned}&p(\{\omega|(Y_{\sigma(1)}, \ldots, Y_{\sigma(m)}) = (y_1, \ldots, y_m), (Y_{\sigma(m+1)}, \ldots, Y_{\sigma(m+n)}) \in A\})\\ &= p_{Y_{\sigma(1)},\ldots,Y_{\sigma(m)}}(y_1, \ldots, y_m) \int_A \pi^{(\sigma)}_{Y_{\sigma(m+1)},\ldots Y_{\sigma(m+n)}|y_1,\ldots,y_m}(x_1, \ldots, x_n)dx_1 \ldots dx_n.\end{aligned}$$

and such that

$$\int_{\mathbf{R}^n} \pi^{(\sigma)}_{Y_{\sigma(m+1)},\ldots Y_{\sigma(m+n)}|y_1,\ldots,y_m}(x_1, \ldots, x_n)dx_1 \ldots dx_n = 1.$$

Definition 1.5 (Marginal Distribution) Let $X = (X_1, \ldots, X_n)$ be a discrete random vector, with joint probability function $p_{X_1,\ldots,X_n}$. The probability distribution for $(X_{j_1}, \ldots, X_{j_m})$, where $m \leq n$ and $1 \leq j_1 < \ldots < j_m \leq n$ is known as the *marginal distribution*, and the marginal probability function is defined as

$$p_{X_{j_1},\ldots,X_{j_m}}(x_1, \ldots, x_m) = \sum_{(y_1,\ldots,y_n)|(y_{j_1},\ldots,y_{j_m})=(x_{j_1},\ldots,x_{j_m})} p_{X_1,\ldots,X_n}(y_1, \ldots, y_n).$$

In particular, for two discrete variables X and Y taking values in the spaces $\mathcal{C}_X$ and $\mathcal{C}_Y$ respectively, with joint probability function $p_{X,Y}$, the marginal probability function for the random variable X is defined as

$$p_X(x) = \sum_{y\in\mathcal{C}_Y} p_{X,Y}(x, y)$$

and the marginal probability function for the random variable Y is defined as

$$p_Y(y) = \sum_{x\in\mathcal{C}_X} p_{X,Y}(x, y).$$

If $\Xi = (\Xi_1, \ldots, \Xi_n)$ is a continuous random vector, with joint probability density function $\pi_{\Xi_1,\ldots,\Xi_n}$, then the *marginal density function* for $\Xi_{j_1}, \ldots, \Xi_{j_m}$, where $\{j_1, \ldots, j_m\} \subset \{1, \ldots, n\}$ is defined as

$$\begin{aligned}\pi_{\Xi_{j_1},\ldots,\Xi_{j_m}}(x_1, \ldots, x_m) &= \int_{\mathbf{R}^{n-m}} \pi_{\Xi_1,\ldots,\Xi_n}(y_1, \ldots, y_n|(y_{j_1}, \ldots, y_{j_m})\\ &= (x_1, \ldots, x_m)) \prod_{k\notin(j_1,\ldots j_m)} dy_k.\end{aligned}$$

□

Categorical random variables In this text, the sample space $\mathcal{X}$ will contain a finite, or countably infinite, number of outcomes, while usually the parameter space $\tilde{\Theta} \subseteq \mathbf{R}^n$, where n is the number of parameters. Most of the discrete variables arising will be *categorical*, in the sense that the outcomes are classified in several categories. For example, suppose an urn contains 16 balls, of which four are white, six are green, five are blue and one is red. Pick a ball at random and let C denote the colour of the ball, then C is an example of a discrete random variable. The probability distribution of a categorical variable C is denoted by p_C. Here, for example, $p(\{C = \text{green}\}) = p_C(\text{green}) = \frac{6}{16} = \frac{3}{8}$;

$$p_C = \begin{array}{cccc} \text{white} & \text{green} & \text{blue} & \text{red} \\ \hline \frac{1}{4} & \frac{3}{8} & \frac{5}{16} & \frac{1}{16} \end{array}$$

The set-up described above is adequate for this text, where only two types of random variables are considered; continuous (which, by definition, have a Riemann integrable density function) and discrete variables. Statements of uncertainty made within this framework are consistent and coherent. Any probability function p over $\mathcal{F}$ that is to provide a quantitative assessment of the probability of an event, which is also to be mathematically coherent over a constructive algebra of events, must satisfy the axioms listed above. Any set function over an algebra of sets that satisfies these axioms will provide a mathematically coherent measure of uncertainty.

1.3 The Bayes update of probability

The *prior* probability is the probability distribution p over $\mathcal{F}$ before any relevant data is obtained. The prior probability is related to background information, modelling assumptions, or simply a function introduced for mathematical convenience, which is intended to express a degree of vagueness. It could be written $p^{(K)}(.)$, where K designates what could be called the *context* or a *frame of knowledge* [9]. The notation $p(.|K)$ is often employed, although this is a little misleading since, formally, conditional probability is only defined in terms of an initial distribution. Here, the *initial* distribution is based on K. Since the notation is now standard, it will be employed in this text, although caution should be observed.

Consider a prior distribution, denoted by $p(.)$. Suppose the experimenter, or agent, has the information that $B \in \mathcal{F}$ holds and desires to update the probabilities based on this piece of information. This update involves introducing a new probability distribution p^* on $\mathcal{F}$. Suppose also that $p(B) > 0$. Since it is now known that B is a certainty, the update requires

$$p^*(B) = 1$$

so that $p^*(B^c) = 0$, where $B^c = \Omega \backslash B$ is the complement of B in Ω. The updated probability is constructed so that the *ratio* of probabilities for any $B_1 \subset B$ and $B_2 \subset B$ does not change. That is, for $B_i \subset B$, $i = 1, 2$,

$$\frac{p^*(B_1)}{p^*(B_2)} = \frac{p(B_1)}{p(B_2)}.$$

For arbitrary $A \in \mathcal{F}$, the axioms yield

$$p^*(A) = p^*(A \cap B) + p^*(A \cap B^c).$$

But since $p^*(B^c) = 0$, it follows that $p^*(A \cap B^c) \leq p^*(B^c) = 0$. In other words, it follows that for arbitrary $A \in \mathcal{F}$, since $p^*(B) = 1$,

$$p^*(A) = \frac{p^*(A \cap B) + p^*(A \cap B^c)}{p^*(B)} = \frac{p^*(A \cap B)}{p^*(B)} = \frac{p(A \cap B)}{p(B)}.$$

The transformation $p \to p^*$ is known as the *Bayes update of probability*. When the evidence is obtained is precisely that an event $B \in \mathcal{F}$ *within the algebra* has happened, Bayes' rule may be used to update the probability distribution. It is customary to use the notation $p(A|B)$ to denote $p^*(A)$. Hence

$$p(A|B) \overset{def}{=} p^*(A) = \frac{p(A \cap B)}{p(B)}. \tag{1.1}$$

This is called the conditional probability of A given B. This characterization of the conditional probability $p(A|B)$ follows [10]. Further discussion may be found in [11]. This is the update used when the evidence is precisely that an event $B \in \mathcal{F}$ has occurred. Different updates are required to incorporate knowledge that cannot be expressed in this way. This is discussed in [12].

From the definition of $p(A|B)$, the trivial but important identity

$$p(A|B)p(B) = p(A \cap B) \tag{1.2}$$

follows for any $A, B \in \mathcal{F}$.

The Bayes factor Bayes' rule simply states that for any two events A and C,

$$p(A|C) = \frac{p(C|A)p(A)}{p(C)}.$$

If event C represents new evidence, and $p^*(.) = p(.|C)$ represents the updated probability distribution, then for any two events A and B, Bayes' rule yields:

$$\frac{p^*(A)}{p^*(B)} = \frac{p(C|A)}{p(C|B)}\frac{p(A)}{p(B)}.$$

The factor $\frac{p(C|A)}{p(C|B)}$ therefore updates the ratio $\frac{p(A)}{p(B)}$ to $\frac{p^*(A)}{p^*(B)}$. Note that

$$\frac{p(C|A)}{p(C|B)} = \frac{p^*(A)/p^*(B)}{p(A)/p(B)}.$$

This leads to the following definition, which will be used later.

Definition 1.6 *Let p and q denote two probability distributions over an algebra $\mathcal{A}$. For any two events $A, B \in \mathcal{A}$, the* Bayes factor *$F_{q,p}(A; B)$ is defined as*

$$F_{q,p}(A, B) := \frac{q(A)/q(B)}{p(A)/p(B)}. \tag{1.3}$$

Here p plays the role of a probability before updating and q plays the role of an updated probability. The Bayes factor indicates whether or not the new information has increased the *odds* of an event A relative to B.

Bayes' rule applied to random variables Let X and Y be two discrete random variables. Let p_X, p_Y, $p_{X|Y}$ and $p_{Y|X}$ denote the probability mass functions of X, Y, X given Y and Y given X respectively. It follows directly from Bayes' rule that for all x, y

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}. \tag{1.4}$$

If X and Y are continuous random variables with density functions π_X and π_Y, then the conditional probability density function of X, given $Y = y$ is

$$\pi_{X|Y}(x|y) = \frac{\pi_{Y|X}(y|x)\pi_X(x)}{\pi_Y(y)}.$$

If X is a discrete random variable and Θ is a continuous random variable with state space $\tilde{\Theta}$, where $p_{X|\Theta}(.|\theta)$ is the conditional probability function for X given $\Theta = \theta$ and Θ has density function π_Θ, then Bayes' rule in this context gives

$$\pi_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)\pi_\Theta(\theta)}{\int_{\tilde{\Theta}} p_{X|\Theta}(x|\theta)\pi_\Theta(\theta)d\theta} = \frac{p_{X|\Theta}(x|\theta)\pi_\Theta(\theta)}{p_X(x)}. \tag{1.5}$$

1.4 Inductive learning

The task of *inductive learning* is, roughly stated, to find a *general* law, based of a *finite* number of particular examples. Without further information, a law established in this way cannot be a certainty, but necessarily has a level of uncertainty attached to it. The assessment uses the idea that future events similar to past events will cause similar outcomes, so that outcomes from past events may be used to predict outcomes from future events. There is a *subjective* component in the assessment of the uncertainty, which enters through the *prior* distribution. This is the assessment made before any particular examples are taken into consideration.

In any real situation within the engineering sciences, a mathematical model is only ever at best a *model* and can never present a *full* description of the situation that is being modelled. In many situations, the information is not presented in terms of absolute certainties to which deductive logic may be applied and 'inductive learning', as described above, is the only way to proceed.

For machine learning of situations arising in the engineering sciences, the machine learns inductively from past examples, from which it makes predictions of future behaviour and takes appropriate decisions; for example, deciding whether a paper mill is running abnormally and, if it is, shutting it down and locating the error.

In the following discussion, the second person singular pronoun, 'You', will be used to denote the person, or machine,[2] with well defined instructions, analysing some uncertain statements and making predictions based on this analysis.

[2] A more precise description of the machine is the hypothetical artificial intelligence robot 'Robbie' in [13].

The Bayes update rule is now applied to learning from experience. The update is based on the definition of the conditional probability $p(E|A)$ of an event E given an event A. The rule for calculating the conditional probability of A given E (or, given that E 'occurs') is known as *Bayes' rule* and is given by

$$p(A|E) = \frac{p(E|A)p(A)}{p(E)}. \tag{1.6}$$

Equation (1.6) follows immediately from the definitions introduced in Equations (1.1) and (1.2). Here E is a mnemonic notation for *evidence*.

The formula (1.6) was for a long time known more widely as 'inverse probability' rather than as 'Bayes' rule'. The quantity $p(A|E)$ does not always need to be computed using this formula; sometimes it is arrived at by other means. For example, the Jeffrey's update rule, as will be seen later, uses a particular set of conditional probabilities as *fundamental* and derives the other probabilities from it.

1.4.1 Bayes' rule

A more instructive form of Bayes' rule is obtained by considering a finite exhaustive set of mutually exclusive hypotheses $\{H_i\}_{i=1}^m$. The law of total probability gives for any E

$$p(E) = \sum_{i=1}^{m} p(H_i)p(E|H_i) \tag{1.7}$$

and Equation (1.6) yields for any H_i

$$p(H_i|E) = \frac{p(E|H_i)p(H_i)}{\sum_{i=1}^{m} p(H_i)p(E|H_i)}. \tag{1.8}$$

Here $p(H_i)$ is the initial or *prior* probability for a hypothesis, before any evidence E is obtained. The prior probability will have a subjective element, depending on background information and how You interpret this information. As discussed earlier, this background information is often denoted by the letter K. In computer science, this is often referred to as *domain knowledge* [14]. The prior is therefore often denoted $p(H_i|K)$. As discussed earlier, this is misleading, because no application of the formula given in Equation (1.1) has been made. The notation is simplified by dropping K.

The quantity in $p(H_i|E)$ in Equation (1.8) denotes how $p(H_i)$ is *updated* to a new probability, called the *posterior probability*, based on this new evidence. This update is the key concept of Bayesian learning from experience.

There is a basic question: why should probability calculus be used when performing the inductive logic in the presence of uncertainty described above? This is connected with the notion of *coherence*. The probability calculus outlined above ensures that the uncertainty statements 'fit together' or 'cohere' and it is the only way to ensure that mathematical statements of uncertainty are consistent. To ensure coherence, inductive logic should therefore be expressed through probability. Inference about uncertain events H_i from an observed event E will be mathematically coherent if and only if they are made by computing $p(H_i|E)$ in the way described above. All calculations, in order to achieve coherence, must be made within this probability calculus. Hence, to make coherent statements about different events, Your calculations for learning by experience have to satisfy Bayes' rule.

The introduction of an arbitrary prior distribution may appear, at first sight, to be ad hoc. The important point here is that all modelling contains some ad hoc element in the choice of the model. The strength of the Bayesian approach is that once the prior distribution is declared, this contains the modelling assumptions and the ad hoc element becomes transparent. The ad hoc element is always present and is stated clearly in the Bayesian approach.

When an event E occurs that belongs to the well defined algebra of events, over which there are well defined probabilities, the Bayes rule for updating the probability of an event A from $p(A)$ to $p(A|E)$ is applied. There are situations where the evidence E may be given in terms of observations that are less precise (that is, E is not an event that clearly belongs to the algebra), but nevertheless an update of the the probability function $p(\cdot)$ is required. Jeffrey's rule and Pearl's method of virtual evidence can be useful in these situations.

1.4.2 Jeffrey's rule

Suppose that the new evidence implies that You form an exhaustive set of r mutually exclusive hypotheses $(G_i)_{i=1}^r$ which, following the soft evidence have probabilities $p^*(G_i)$. The question is how to update the probability of any event $A \in \mathcal{F}$. Note that You cannot use Bayes' rule (Equation (1.6)), since the evidence has not been expressed in terms of a well defined event E for which the prior probability value is known. Jeffrey's rule may be applied to the situation where it may be assumed that for all events $A \in \mathcal{F}$, the probabilities $p(A|G_i)$ remain unchanged. It is only the assessment of the the mutually exclusive hypotheses $(G_i)_{i=1}^r$ that changes; no new information is given about the the relevance of $(G_i)_{i=1}^r$ to *other* events.

Definition 1.7 (Jeffrey's Update) *The* Jeffrey's rule *for computing the update of the probability for any event A, is given by*

$$p^*(A) = \sum_{i=1}^{r} p^*(G_i)p(A|G_i). \tag{1.9}$$

This is discussed in [15]. Jeffrey's rule provides a consistent probability function, such that $p^*(A|G_i) = p(A|G_i)$ for all i. Equation (1.9) is therefore an expression of the rule of total probability (Definition 1.7). □

1.4.3 Pearl's method of virtual evidence

In the situation considered by Pearl, new evidence gives information on a set of mutually exclusive and exhaustive events $G_1, \ldots, G_n$, but is not specified as a set of new probabilities for these events. Instead, for each of the events $G_1, \ldots, G_n$, the ratio $\lambda_j = \frac{p(A|G_j)}{p(A|G_1)}$, for $j = 1, \ldots, n$ is given for an event A. That is, λ_j represents the likelihood ratio that the event A occurs given that G_j occurs, compared with G_1. Note that $\lambda_1 = 1$.

Definition 1.8 *Let p denote a probability distribution over a countable space $\mathcal{X}$ (Definition 1.2) and let $G_1, \ldots, G_n \in \mathcal{F}$ be a mutually exclusive (that is $G_i \cap G_j = \phi$ for all*

$i \neq j$) and exhaustive (that is $\cup_{j=1}^{n} G_j = \Omega$) events, where $p(G_j) = p_j$. Set $\lambda_j = \frac{p(A|G_j)}{p(A|G_1)}$ for $j = 1, \ldots, n$. Then, for each $x \in \mathcal{X}$, the Pearl update *$\tilde{p}$ is defined as*

$$\tilde{p}(\{x\}) = p(\{x\}) \frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j p_j} \qquad x \in G_j, \qquad j = 1, \ldots, n. \tag{1.10}$$

It is clear that this provides a well defined probability distribution.

Jeffrey's rule and Pearl's method for virtual evidence will be discussed further in Section 3.2. They are methods that, under some circumstances, enable evidence to be incorporated that does not fit directly into the framework of the chosen statistical model.

1.5 Interpretations of probability and Bayesian networks

Loosely speaking, 'classical statistics' proposes a probability distribution over the event space, where the probability distribution is a member of a parametric family, where the value of the parameters are unknown. The parameters are *estimated*, and the estimates are used to obtain an *approximation* to the 'true' probability, which is unknown. In *Bayesian* probability, the lack of knowledge of the parameter is expressed though a probability distribution over the parameter space, to feature Your personal assessment of the probability of where the parameter may lie. For this reason, Bayesian probability is also referred to as *personal* probability, or *epistemological* probability. Built into a Bayesian probability is Your a priori state of knowledge, understanding and assessment concerning the model and the source of data. So, loosely speaking, classic statistics yields an approximation to an objectively 'true' probability. The 'true' probability is fixed, but unknown, and the *estimate* of this probability may differ between researchers. Bayesian statistics yields an *exact* computation of a *subjective* probability. It is the *probability itself*, rather than the estimate of the probability, that may differ between researchers.

Bayesian networks are frequently implemented as information processing components of Expert Systems (in artificial intelligence) [14], where personal and epistemological probability provides a natural framework for the machine to learn from its experience. P. Cheeseman in [13] and [16] argues that personal probability as the calculus of plausible reasoning is the natural paradigm for artificial intelligence.

J. Williamson [4] and other authors distinguish between *subjective* and *objective* Bayesian probability. Consider two different learning agents, called (for convenience) Robbie$_\alpha$ and Robbie$_\beta$. These are two hardware copies of Robbie in [13], with some different internal representations (i.e. different ways of assessing prior information). A Bayesian probability is said to be *objective* if, with the same background information, the two agents will agree on the probabilities. Bayesian probabilities are *subjective* if the two different learning agents, Robbie$_\alpha$ and Robbie$_\beta$ may disagree about the probabilities, even though they share the same background knowledge, but without either of then being provably wrong.

It seems rational to require that *subjective* probabilities should be calibrated with the external world. This does not follow from requirements of coherence; it is rather a principle of inference that is imposed in any serious statistical study. For this, one often cites *Lewis's principal principle*, stated in [17]. This is the assertion that Your subjective probability for an event, conditional upon the knowledge of a *physical* probability of that event, should equal the physical probability. In terms of the formula (where *ch* denotes the physical chance),

$$p(A \mid ch\,(A) = x) = x.$$

There are several theoretical foundations for reconciling subjective opinion to objective probability. One of the more prominent was given by R.J. Aumann.[3] He proved the following fact. If two agents have same priors, and if their posteriors for an event A are *common knowledge* (i.e. Robbie$_\alpha$ and Robbie$_\beta$ know the posterior, and Robbie$_\alpha$ knows that Robbie$_\beta$ knows and that Robbie$_\beta$ knows that Robbie$_\alpha$ knows that Robbie$_\beta$ knows and so on, ad infinitum), then the two posterior probability masses will be identical [18].

This relies on the rational idea that if someone presents an opinion different from Yours, then this is an important piece of information which should induce You to revise Your opinion. The result by Aumann quoted above implies that there will be a process of revision, which will continue until objective probability (an equilibrium of consensus) is reached. It can be shown that this will happen in a finite number of steps.

In probabilistic theories of causation there is a set of variables V, over which there is a probability distribution p. The causal relationships between the variables in V are the object of study. In [19], the variables are indexed by a time parameter and causality is reduced to probability. The Bayesian network approach is different; in addition to the probability distribution p, the variables are nodes of a directed acyclic graph $\mathcal{G}$, where the edges represent direct causal relationships between the variables in V. This requires, as pointed out by D. Freedman and P. Humphreys [20], that we already know the causal structure obtained, for example, by exercise of common sense and knowledge of the variables V. Both p and $\mathcal{G}$ are required [21]. Additional assumptions are therefore required to infer the direction of cause to effect relationships between variables; there is no form of Bayesian coherence from which they may be inferred. The role of a directed graph is to represent information about dependence between variables V. In particular, the graph may be used to indicate what would happen to values of some variables under changes to other variables that are called interventions; namely, the variable is forced to take a particular value irrespective of the state of the rest of the system. The graph will also indicate how the various different distributions for subsets of V are consistently connected to each other to yield p.

1.6 Learning as inference about parameters

Consider a random row vector $\underline{X} = (X_1, \ldots, X_d)$, denoting d attributes. Suppose that n independent observations are made on $\underline{X}$ and the values obtained are recorded in a matrix $\mathbf{x}$, where x_{jk} denotes the jth recorded value of attribute k. The $n \times d$ matrix $\mathbf{x}$ may be regarded as an instantiation of the $n \times d$ random matrix $\mathbf{X}$, where each row of $\mathbf{X}$ is an independent copy of $\underline{X}$. Suppose that the *evidence* $\mathbf{x}$ is to be used to update the probability function for another collection of m variables, $(Y_1, \ldots, Y_m)$ from $p_{Y_1,\ldots,Y_m}(.)$ to $p_{Y_1,\ldots,Y_m|\mathbf{X}}(.|\mathbf{x})$.

A fundamental special case, discussed in [22], is that of computing the predictive probability for the next observation in the univariate setting. That is, $d = 1$. Here, the

[3] 2005 Laureate of the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel.

matrix $\mathbf{X}$ is an $n \times 1$ matrix and may therefore be considered as a column vector $\underline{X}_{(n)} = (X_1, \ldots, X_n)^t$, where X_j, $j = 1, \ldots, n$ are independent identically distributed. Here, $\mathbf{x} = \underline{x}_{(n)} = (x_1, \ldots, x_n)^t$, a vector of n observed values. The random vector $\underline{Y}$ is simply X_{n+1}. The problem is to compute the conditional probability distribution of Y given $\underline{X}_{(n)} = \underline{x}_{(n)}$. The connection between $\underline{x}_{(n)}$ and y is described by a probability $p_{Y|\underline{X}_{(n)}}(y|\underline{x}_{(n)})$ of Y given $\underline{X}_{(n)} = \underline{x}_{(n)}$. Let $\underline{X}_{(n+1)} = (X_1, \ldots, X_{n+1})^t$ and $\underline{x}_{(n+1)} = (x_1, \ldots x_{n+1})^t$. You compute

$$p_{X_{n+1}|\underline{X}_{(n)}}(x_{n+1}|\underline{x}_{(n)}) = \frac{p_{\underline{X}_{(n+1)}}(\underline{x}_{(n+1)})}{p_{\underline{X}_{(n)}}(\underline{x}_{(n)})}. \tag{1.11}$$

The progression from here requires a mathematical model containing *parameters* denoted θ such that, given θ, the random variables $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.). That is, with notation $(.|\theta)$ to denote the parameter fixed as θ, there is a decomposition

$$p_{\underline{X}_{(n)}}(\underline{x}_{(n)}|\theta) = \prod_1^n q_{X_i}(x_i|\theta).$$

The family of probability functions $q_{X_i}(.|\theta)$, $\theta \in \tilde{\Theta}$ (where $\tilde{\Theta}$ is the space of all permissible values of the unknown parameter θ) and the parameter θ need to be specified. Since the value of θ is unknown, the Bayesian approach is to consider a *probability distribution* over $\tilde{\Theta}$. Thus, θ may be regarded as a realisation of a random variable Θ. You need to specify a prior distribution over $\tilde{\Theta}$. This discussion confines itself to the case where Θ is considered to be a *continuous* random variable and hence the prior distribution is described by a probability density function $\pi_\Theta : \tilde{\Theta} \to \mathbf{R}_+$. The prior predictive distribution may then be written as

$$p_{\underline{X}_{(n)}}(\underline{x}_{(n)}) = \int_{\tilde{\Theta}} \prod_1^n q_{X_i}(x_i|\theta)\pi(\theta)d\theta. \tag{1.12}$$

Definition 1.9 (Prior Predictive Probability Distribution) *The prior distribution $p_{\underline{X}_{(n)}}$ for the collection of random variables $\underline{X}_{(n)}$ (for which $\underline{x}_{(n)}$ is an observation) is known as the* Prior Predictive Probability Distribution.

B. De Finetti showed in [23] that if the x_js are infinitely exchangeable (and lie in a reasonable space), then the structure for $p_{\underline{X}_{(n)}}(.)$ given by Equation (1.12) is the only possible one.

Inserting (1.12) in the right hand side of (1.11) yields

$$\begin{aligned} p_{X_{n+1}}(x_{n+1}|\underline{x}_{(n)}) &= \frac{\int_{\tilde{\Theta}} \prod_1^{n+1} q_{X_i}(x_i|\theta)\pi_\Theta(\theta)d\theta}{\int_{\tilde{\Theta}} \prod_1^n q_{X_i}(x_i|\theta)\pi_\Theta(\theta)d\theta} \\ &= \int_{\tilde{\Theta}} q_{X_{n+1}}(x_{n+1}|\theta)\frac{\prod_1^n q_{X_i}(x_i|\theta)\pi_\Theta(\theta)}{\int_{\tilde{\Theta}} \prod_1^n q_{X_i}(x_i|\theta)\pi_\Theta(\theta)d\theta}d\theta. \end{aligned}$$

The conditional probability density of Θ given $\underline{X}_{(n)} = \underline{x}_{(n)}$ may be obtained by Bayes' rule:

$$\pi_{\Theta|\underline{X}_{(n)}}\left(\theta \mid \underline{x}_{(n)}\right) = \frac{\prod_1^n q_{X_i}(x_i|\theta)\pi_\Theta(\theta)}{\int_{\tilde{\Theta}} \prod_1^n q_{X_i}(x_i|\theta)\pi_\Theta(\theta)d\theta}. \tag{1.13}$$

It follows directly that

$$p_{X_{n+1}}(x_{n+1}|\underline{x}_{(n)}) = \int_{\tilde{\Theta}} q_{X_{n+1}}(x_{n+1}|\theta)\pi_{\Theta|\underline{X}_{(n)}}\left(\theta \mid \underline{x}_{(n)}\right) d\theta. \tag{1.14}$$

In Section 1.9, explicit examples of evaluations of Equation (1.14) are given.

The probability density function π_Θ, placed over $\tilde{\Theta}$ before any data is observed is known as the prior density; the probability density function $\pi_{\Theta|\underline{X}_{(n)}}$ defined by Equation (1.13) is known as the *posterior* probability density. Equation (1.14) shows how Bayesian learning about X_{n+1} is based on learning about θ from $\underline{x}_{(n)}$. *Bayesian statistical inference* is the term used to denote Bayesian learning of the posterior distribution of a set of parameters.

The meaning of causality for K. Pearson, see Chapter 4 in [2], seems to be expressible by Equation (1.14), as he writes 'that a certain sequence has occurred and recurred in the past is a matter of experience to which we give expression in the concept of causation, that it will recur in the future is a matter of belief to which we give expression in the concept of probability.'

1.7 Bayesian statistical inference

The aim of learning is to predict the nature of future data based on past experience [22]. One constructs a probabilistic model for a situation where the model contains unknown parameters. The parameters are only a *mechanism* to help estimate *future* behaviour; they are not an end in themselves.

As stated, the 'classical' approach regards a parameter as fixed. It is unknown and has to be estimated, but it is not considered to be a random variable. One therefore computes approximations to the unknown parameters, and uses these to compute an approximation to the probability density. The parameter is considered to be fixed and unknown, because there is usually a basic assumption that in ideal circumstances, the experiment could be repeated infinitely often and the estimating procedure would return a precise value for the parameter. That is, if one increases the number of replications indefinitely, the *estimate* of the unknown parameter converges, with probability one, to the true value. This is known as the 'frequentist' interpretation of probability.

Basic to the 'frequentist' interpretation is the assumption that an experiment may be repeated, in an identical manner, an indefinite number of times. With the classical approach, the sample space Ω and the event space $\mathcal{A}$ of subsets of Ω have to be defined in advance. This makes incorporation of 'soft evidence' or 'virtual evidence,' that will be considered later in the text, harder; these are situations where the information obtained cannot be expressed as one of the well defined events of the event space. Then, the probability distribution is interpreted as follows: for each $A \in \mathcal{A}$, $p(A)$ is interpreted as the limit, *from observed data* that would be obtained if the experiment could be repeated independently, under identical circumstance, arbitrarily often. That is,

$$p(A) = \lim_{n \to +\infty} \frac{N(n, A)}{n},$$

where $N(n, A)$ denotes the number of times event A has been observed from n replications of the experiment.

This interpretation is intuitively appealing, but there is room for caution, since the infinite independent replications are *imagined*, and therefore the convergence of relative frequencies to a limit is hypothetical; the imagined infinite sequence of replications under *exactly* the same conditions is a textbook construction and abstraction. In concrete terms, it is supposed that there are many sources, each with large numbers of data so that the 'empirical' distribution can approximate the limit with arbitrary precision. Despite the hypothetical element in the formulation, the 'frequentist' interpretation of probability follows basic human common sense; the probability distribution is interpreted as the long run average. The 'long run average' interpretation assumes prior knowledge; when an agent like Robbie is to compute probability for its actions, it cannot be instructed to wait for the result in an infinite outcome of experiments and, indeed, it cannot run in 'real time' if it is expected to wait for a large number of outcomes.

Once the existence of a probability measure p over $(\Omega, \mathcal{A})$ has been established, which may be interpreted in the sense of 'long run averages', it is then a matter of computation to prove that the parameter estimates $\hat{\theta}_n$ based on n observations converge with probability 1, provided a sensible estimating procedure is used, to a parameter value θ.

As discussed earlier, the *Bayesian* approach takes the view that since the parameter is unknown, it is a random variable as far as You are concerned. A probability distribution, known as the *prior distribution*, is put over the *parameter space*, based on a prior assessment of where the parameter may lie. One then carries out the experiment and, using the data available, which is necessarily a *finite* number of data, one uses the Bayes rule to compute the *posterior distribution* in Equation (1.13), which is the updated probability distribution over the parameter space.

The posterior distribution over the parameter space is then used to compute the probability distribution for future events, based on past experience. Unlike the classical approach, this is an exact distribution, but it contains a subjective element. The subjective element is described by the prior distribution.

In Bayesian statistics, the computation of the posterior distribution usually requires numerical methods, and Markov chain Monte Carlo methods seem to be the most efficient. This technique is 'frequentist', in the sense that it relies upon an arbitrarily large supply of independent random numbers to obtain the desired precision. From an engineering point of view, there are efficient pseudo-random number generators that supply arbitrarily large sequences of 'random' numbers of very good quality. That is, there are tests available to show whether a sequence 'behaves' like an observation of a sequence of suitable independent random numbers.

Both approaches to statistical inference have an arbitrary element. For the classical approach, one sees this in the choice of sample space. The sample space is, to use H. Jeffreys' [24] vivid description, 'the class of observations that might have been obtained, but weren't'. For some experiments, the sample space is a clear and well defined object, but for others, there is an arbitrary element in the choice of the sample space. For example, an experiment may be set up with n plants, but some of the plants may die before the results are established.

Alternative hypotheses There is *no distinction* within the Bayesian approach between the various values of the parameter except in the prior $\pi(\theta)$. The view is one of contrast between various values of θ. Consider the case where the parameter space consists of just two values, (θ_0, θ_1). Dropping subscripts where they are clearly implied, Bayes' rule for data x gives

$$\pi(\theta_0|x) = \frac{p(x|\theta_0)\pi(\theta_0)}{p(x)}$$

and

$$\pi(\theta_1|x) = \frac{p(x|\theta_1)\pi(\theta_1)}{p(x)}.$$

It follows that

$$\frac{\pi(\theta_0|x)}{\pi(\theta_1|x)} = \frac{p(x|\theta_0)\pi(\theta_0)}{p(x|\theta_1)\pi(\theta_1)}. \tag{1.15}$$

The *likelihood ratio* for two different parameter values is the ratio of the likelihood functions for these parameter values; denoting the likelihood ratio by LR,

$$LR(\theta_0, \theta_1; x) = \frac{p(x|\theta_0)}{p(x|\theta_1)}.$$

The *prior odds ratio* is simply the ratio $\pi(\theta_0)/\pi(\theta_1)$ and the *posterior odds ratio* is simply the ratio $\pi(\theta_0|x)/\pi(\theta_1|x)$. An odds ratio of greater than 1 indicates support for the parameter value in the numerator.

Equation (1.15) may be rewritten as

$$\text{posterior odds} \quad = \quad LR \times \text{prior odds}.$$

The data affect the change of assessment of probabilities through the likelihood ratio, comparing the probabilities of data on θ_0 and θ_1. This is in contrast with a sampling theory, or tail area significance test, where only the null hypothesis (say θ_0) is considered by the user of the test.

In 'classical' statistics, statements about parameters may be made through *confidence intervals*. It is important to note that a confidence interval for θ is *not* a probability statement about θ, because in classical statistics θ is *not* a random variable. It is a fixed, though unknown, value. The confidence interval is derived from probability statements about x the observation, namely from $p(x|\theta)$.

There is no axiomatic system that leads to confidence *measures*, while the axioms of probability are well defined. Operations strictly in accord with the calculus of probability give coherent conclusions. Ideas outside the probability calculus may give anomalies.

The next two sections give a detailed examination of two probability distributions that are often central to the analysis of Bayesian networks. Section 1.8 discusses binary variables, while Section 1.9 discusses multinomial variables. The distributions discussed in Section 1.8 are a useful special case of those discussed in Section 1.9.

1.8 Tossing a thumb-tack

The discussion of the thumb-tack is taken from D. Heckerman [25].

If a thumb-tack is thrown in the air, it will come to rest either on its point (0) or on its head (1). Suppose the thumb-tack is flipped n times, making sure that the physical properties of the thumb-tack and the conditions under which it is flipped remain stable over time. Let $\underline{x}_{(n)}$ denote the sequence of outcomes

$$\underline{x}_{(n)} = (x_1, \ldots, x_n)^t.$$

Each trial is a *Bernoulli trial* with probability θ of success (obtaining a 1). This is denoted by

$$X_i \sim Be(\theta), \qquad i = 1, \ldots, n.$$

Using the Bayesian approach, the parameter θ is be regarded as the outcome of a random variable, which is denoted by Θ. The outcomes are *conditionally independent, given* θ. This is denoted by

$$X_i \perp X_j | \Theta, \qquad i \neq j.$$

When $\Theta = \theta$ is given, the random variables $X_1, \ldots, X_n$ are independent, so that

$$p_{\underline{X}_{(n)}}(\underline{x}_{(n)}|\theta) = \prod_{l=1}^{n} \theta^{x_l}(1-\theta)^{1-x_l} = \theta^k(1-\theta)^{n-k}$$

where $k = \sum_{l=1}^{n} x_l$.

The problem is to estimate θ, finding the value that is best for $\underline{x}_{(n)}$. The Bayesian approach is, starting with a prior density $\pi_\Theta(.)$ over the parameter space $\tilde{\Theta} = [0, 1]$, to find the posterior density $\pi_{\Theta|\underline{X}_{(n)}}(.|\underline{x}_{(n)})$.

$$\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)}) = \frac{p_{\underline{X}_{(n)}|\Theta}(\underline{x}_{(n)}|\theta)\pi_\Theta(\theta)}{p_{\underline{X}_{(n)}}(\underline{x}_{(n)})} = \frac{p_{\underline{X}_{(n)}|\Theta}(\underline{x}_{(n)}|\theta)\pi_\Theta(\theta)}{\int p_{\underline{X}_{(n)}|\Theta}(\underline{x}_{(n)}|\phi)\pi_\Theta(\phi)d\phi}.$$

Let π_Θ be the uniform density on [0, 1]. This represents that initially You have no preference concerning θ; all values are equally plausible.[4] The choice of prior may seem arbitrary, but following the computations below, it should be clear that, from a large class of priors, the final answer does not depend much on the choice of prior if the thumb-tack is thrown a large number of times.

[4] As previously stated, the prior distribution contains the ‘ad hoc’ element. The results obtained from any statistical analysis are only reliable if there is sufficient data so that any inference will be robust under a rather general choice of prior.

There are well known difficulties with the statement that a uniform prior represents no preference concerning the value of θ. If the prior density for Θ is uniform, then the prior density of Θ^2 will *not* be uniform, so ‘no preference’ for values of Θ indicates that there is a *distinct* preference among possible initial values of Θ^2. If $\pi_1(x) = 1$ for $0 < x < 1$ is the density function for Θ and π_2 is the density function for Θ^2, then $\pi_2(x) = \frac{1}{2x^{1/2}}$ for $0 < x < 1$.

With the uniform prior,

$$\int_0^1 p_{\underline{X}_{(n)}|\Theta}(\underline{x}_{(n)}|\theta)\pi_\Theta(\theta)d\theta = \int_0^1 \theta^k(1-\theta)^{n-k}d\theta = \frac{k!(n-k)!}{(n+1)!}. \tag{1.16}$$

This may be computed using integration by parts, as follows. Set

$$I_{n,k} = \int_0^1 \theta^k(1-\theta)^{n-k}d\theta,$$

then

$$I_{n,0} = \int_0^1 (1-\theta)^n d\theta = \frac{1}{n+1}.$$

Using integration by parts,

$$I_{n,k} = \left[-\frac{\theta^k(1-\theta)^{n-k+1}}{n-k+1}\right]_{\theta=0}^1 + \frac{k}{n-k+1}I_{n,k-1} = \frac{k}{n-k+1}I_{n,k-1}.$$

From this,

$$I_{n,k} = \frac{k!}{n(n-1)\dots(n-k+1)}\frac{1}{(n+1)} = \frac{k!(n-k)!}{(n+1)!}.$$

This is an example of the *Beta integral*. The *posterior distribution* is therefore a *Beta density*

$$\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}^{(n)}) = \begin{cases} \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} & 0 \le \theta \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1.17}$$

It should be apparent that, in this case, there would have been tremendous difficulties carrying out the integral if the prior had been anything other than the uniform, or a member of the Beta family. The computational aspects are, or were, prior to the development of Markov chain Monte Carlo (McMC) methods [26], the main drawback to the Bayesian approach.

The Beta distribution is not restricted to integer values; the *Euler gamma function* is necessary to extend the definition to positive real numbers.

Definition 1.10 (Euler Gamma function) *The* Euler Gamma function $\Gamma(\alpha) : (0, +\infty) \to (0, +\infty)$ *is defined as*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx. \tag{1.18}$$

The Euler Gamma function satisfies the following properties.

Lemma 1.1 *For all* $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. *If* n *is an integer satisfying* $n \ge 1$, *then*

$$\Gamma(n) = (n-1)!$$

Proof of Lemma 1.1 Note that $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$. For all $\alpha > 0$, integration by parts gives

$$\Gamma(\alpha + 1) = \int_0^\infty x^\alpha e^{-x} dx = \alpha\Gamma(\alpha). \tag{1.19}$$

The result follows directly. □

Definition 1.11 (Beta Density) *The Beta density Beta*(α, β) *with parameters* $\alpha > 0$ *and* $\beta > 0$ *is defined as the function*

$$\psi(t) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1} & t \in [0, 1] \\ 0 & t \notin [0, 1] \end{cases} \tag{1.20}$$

The following results show that the Beta density is a probability density function for all real $\alpha > 0$ and $\beta > 0$.

Lemma 1.2 *Set*

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt.$$

Then

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Proof of Lemma 1.2 Directly from the definition of the Gamma function, using the substitutions $u = a^2$ and $v = b^2$ and, at the end of the argument $\cos^2\theta = t$ so that $\frac{dt}{d\theta} = -2\cos\theta\sin\theta$,

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty \int_0^\infty e^{-u} u^{\alpha-1} e^{-v} v^{\beta-1} du dv \\
&= 4\int_0^\infty \int_0^\infty e^{-(a^2+b^2)} a^{2(\alpha-1)} b^{2(\beta-1)} ab da db \\
&= \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-(a^2+b^2)} |a|^{2\alpha-1} |b|^{2\beta-1} da db \\
&= \int_0^{2\pi} \int_0^\infty e^{-r^2} r^{2(\alpha+\beta)-2} |\cos\theta|^{2\alpha-1} |\sin\theta|^{2\beta-1} r dr d\theta \\
&= \frac{1}{2}\left(\int_0^{2\pi} |\cos\theta|^{2\alpha-1} |\sin\theta|^{2\beta-1} d\theta\right) \int_0^\infty e^{-u} u^{(\alpha+\beta)-1} du \\
&= \left(2\int_0^{\pi/2} (\cos\theta)^{2(\alpha-1)} (\sin\theta)^{2(\beta-1)} \cos\theta \sin\theta d\theta\right) \Gamma(\alpha + \beta) \\
&= \left(\int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt\right) \Gamma(\alpha + \beta) \\
&= B(\alpha, \beta)\Gamma(\alpha + \beta).
\end{aligned}$$

The result follows directly. □

Corollary 1.1 *Let ψ denote the Beta density, defined in Equation (1.20), then* $\int_0^1 \psi(\theta)d\theta = 1$.

Proof of Corollary 1.1 This is a direct consequence of Lemma 1.2. □

It follows that, for binomial sampling, updating may be carried out very easily for any prior distribution within the Beta family. Suppose the prior distribution π_0 is the $B(\alpha, \beta)$ density function, n trials are observed, with k taking the value 1 and $n-k$ taking the value 0. Then

$$\begin{aligned}\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)}) &= \frac{p_{\underline{X}_{(n)}|\Theta}(\underline{x}_{(n)}|\theta)\pi_\Theta(\theta)}{p_{\underline{X}_{(n)}}(\underline{x}_{(n)})}\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)p_{\underline{X}_{(n)}}(\underline{x}_{(n)})}\theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1} = c\theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}.\end{aligned}$$

Since $\int_0^1 \pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)})d\theta = 1$, it follows from Lemma 1.2, that

$$\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)}) = \begin{cases} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+k)\Gamma(\beta+n-k)}\theta^{\alpha+k}(1-\theta)^{\beta+n-k} & \theta \in (0,1)\\ 0 & \theta \notin (0,1).\end{cases}$$

so that $\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)})$ is a $B(\alpha+k, \beta+n-k)$ density. □

Recall the definition of the maximum likelihood estimate: it is the value of θ that maximizes $p(\underline{x}_{(n)}|\theta) = \theta^k(1-\theta)^{n-k}$. It is well known that

$$\hat{\theta}_{MLE}\left(\underline{x}_{(n)}\right) = \frac{k}{n}.$$

The same pattern of thought can be applied to maximize the posterior density.

Definition 1.12 (Maximum Posterior Estimate) *The* maximum posterior estimate*, $\hat{\theta}_{MAP}$, is the value of θ which maximizes the posterior density $\pi_{\Theta|\underline{x}_{(n)}}(\theta|\underline{x}_{(n)})$.*

When the posterior density is $B(k+\alpha, n-k+\beta)$, an easy computation gives

$$\hat{\theta}_{MAP} = \frac{\alpha+k}{\alpha+\beta+n}.$$

Note that when the prior density is uniform, as in the case above, the MAP and MLE are exactly the same. The parameter, of course, is not an end in itself. The parameter ought to be regarded as a means to computing the predictive probability. The posterior is used to compute this; c.f. (1.14) above.

The predictive probability for the next toss Recall that the 'parameter' is, in general, an artificial introduction, to help compute $p_{X_{n+1}|\underline{X}_{(n)}}(x_{n+1}|\underline{x}_{(n)})$. Suppose that $\pi(\theta|\underline{x}_{(n)})$ has a $B(\alpha+k, \beta+n-k)$ distribution. The *predictive probability for the next toss*, for $a = 0$ or 1, is given by

$$p_{X_{n+1}|\underline{X}_{(n)}}(a|\underline{x}_{(n)}) = \int_0^1 p_{X_{n+1}}(a|\theta)\pi_{\Theta|\underline{X}_{(n)}}(\theta|\underline{x}_{(n)})d\theta.$$

Since $p_{X_{n+1}}(1|\theta) = \theta$, it follows (using equation (1.19)) that

$$\begin{aligned} p_{X_{n+1}|\underline{X}_{(n)}}(1|\underline{x}_{(n)}) &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+k)\Gamma(\beta+n-k)} \int_0^1 \theta^{(\alpha+k+1)}(1-\theta)^{\beta+n-k} d\theta \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+k)\Gamma(\beta+n-k)} \frac{\Gamma(\alpha+k+1)\Gamma(\beta+n-k)}{\Gamma(\alpha+\beta+n+1)} \\ &= \frac{\alpha+k}{\alpha+\beta+n}. \end{aligned}$$

In particular, note that the *uniform* prior, $\pi_0(\theta) = 1$ for $\theta \in (0, 1)$, is the $B(1, 1)$ density function, so that for binomial sampling with a uniform prior, the predictive probability is

$$\begin{aligned} p_{X_{n+1}|\underline{X}_{(n)}}(1|\underline{x}_{(n)}) &= \frac{k+1}{n+2}; \\ p_{X_{n+1}|\underline{X}_{(n)}}(0|\underline{x}_{(n)}) &= \frac{n+1-k}{n+2}. \end{aligned} \tag{1.21}$$

This distribution, or more precisely $\frac{k+1}{n+2}$, is known as the Laplace rule of succession. A combinatorial derivation for it is given in [27].

Reconciling subjective predictive probabilities The example of *agreeing to disagree*, referred to in the preceding, is due to R.J. Aumann, in [18]. Suppose two agents, Robbie$_\alpha$ and Robbie$_\beta$, both toss a thumb-tack once without communicating the outcome to each other. Both Robbie$_\alpha$ and Robbie$_\beta$ have the same uniform prior on θ. Suppose Robbie$_\alpha$ and Robbie$_\beta$ communicate the value of their respective predictive (posterior) probabilities as

$$p(\{X_{n+1} = 1\}|\text{Robbie}_\alpha) = \frac{2}{3}; \qquad p(\{X_{n+1} = 1\}|\text{Robbie}_\beta) = \frac{1}{3}.$$

Note that in the conditional probabilities above Robbie$_\alpha$ and Robbie$_\beta$ actually refer to respective states of knowledge. Now, since both the number of tosses by each agent and the predictive probabilities held by the two agents is their common knowledge, they can revise their opinions by (1.21) to

$$p(\{X_{n+1} = 1\}|\text{Robbie}_\alpha, \text{Robbie}_\beta) = \frac{1+1}{2+2} = \frac{1}{2}.$$

This holds as Robbie$_\alpha$ and Robbie$_\beta$ deduce by (1.21) that exactly one outcome of the two tosses was 1 (and the other was 0). The revision would not hold if the number of tosses was not common knowledge.

1.9 Multinomial sampling and the Dirichlet integral

Consider the case of multinomial sampling, where an experiment can take one of k outcomes, labelled $C_1, \ldots, C_k$. Suppose that $p(X = C_j) = \theta_j$, so that $\theta_1 + \ldots + \theta_k = 1$.

Consider n independent trials, $X_1, \ldots, X_n$. The notation $\mathbf{1}_A$, to denote the indicator function of a set A, will be used; that is

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

Let $\mathbf{1}_{C_i}(x) = 1$ if $x = C_i$ and 0 otherwise. Set

$$Y_i = \sum_{j=1}^{n} \mathbf{1}_{C_i}(X_j).$$

Then Y_i denotes the number of trials that result in outcome C_i. Note that

$$Y_1 + \ldots + Y_k = n.$$

Then $(Y_1, \ldots, Y_k)$ is said to have a *multinomial* distribution and

$$p_{Y_1,\ldots,Y_k}(x_1, \ldots, x_k) = \frac{n!}{x_1!x_2!\ldots x_{k-1}!x_k!}\theta_1^{x_1}\ldots\theta_k^{x_k},$$

where the expression in front of the $\theta_1^{x_1}\ldots\theta_k^{x_k}$ is the multinomial coefficient.

In the *Bayesian* approach, a *prior distribution* is put over $\theta_1, \ldots, \theta_k$. Then, using the observations, this is updated using Bayes' rule to a posterior probability distribution over $\theta_1, \ldots, \theta_k$.

A particularly convenient family of distributions to use is the *Dirichlet* family, defined as follows.

Definition 1.13 (Dirichlet Density) *The Dirichlet density Dir*$(a_1, \ldots, a_k)$ *is the function*

$$\pi(\theta_1, \ldots, \theta_k) = \begin{cases} \frac{\Gamma(a_1+\ldots+a_k)}{\prod_{j=1}^k \Gamma(a_k)}(\prod_{j=1}^k \theta_j^{a_j-1}) & \theta_j \geq 0, \sum_{j=1}^k \theta_j = 1, \\ 0 & \textit{otherwise}, \end{cases} \tag{1.22}$$

where Γ denotes the Euler Gamma function, given in Definition 1.10. The parameters $(a_1, \ldots, a_k)$ are all strictly positive and are known as *hyper parameters*.

This density, and integration with respect to this density function, are to be understood in the following sense. Since $\theta_k = 1 - \sum_{j=1}^{k-1}\theta_j$, it follows that π may be written as $\pi(\theta_1, \ldots, \theta_k) = \tilde{\pi}(\theta_1, \ldots, \theta_{k-1})$, where

$$\tilde{\pi}(\theta_1, \ldots, \theta_{k-1}) = \begin{cases} \frac{\Gamma(a_1+\ldots+a_k)}{\prod_{j=1}^k \Gamma(a_k)}\left(\prod_{j=1}^{k-1} \theta_j^{a_j-1}\right)\left(1 - \sum_{j=1}^{k-1}\theta_j\right)^{a_k-1} & \theta_j \geq 0, \sum_{j=1}^{k-1}\theta_j \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.23}$$

Clearly, when $k = 2$, this reduces to the *Beta density*. The following results show that the Dirichlet density is a probability density function.

Lemma 1.3 *Set*

$$D(a_1, \ldots, a_k) = \int_0^1 \int_0^{1-x_1} \int_0^{1-(x_1+x_2)} \cdots \int_0^{1-\sum_{j=1}^{k-2} x_j} \left(\prod_{j=1}^{k-1} x_j^{a_j-1}\right) \left(1 - \sum_{j=1}^{k-1} x_j\right)^{a_k-1} dx_{k-1} \ldots dx_1.$$

Then

$$D(a_1, \ldots, a_k) = \frac{\prod_{j=1}^n \Gamma(a_j)}{\Gamma\left(\sum_{j=1}^k a_j\right)}.$$

Proof of Lemma 1.3 Straight from the definition of the Euler Gamma function, using the substitutions $x_j^2 = u_j$,

$$\begin{aligned}\prod_{j=1}^n \Gamma(a_j) &= \int_0^\infty \cdots \int_0^\infty e^{-\sum_{j=1}^k u_j} \prod_{j=1}^k u_j^{a_j-1} du_1 \ldots du_k \\ &= 2^k \int_0^\infty \cdots \int_0^\infty e^{-\sum_{j=1}^k x_j^2} \prod_{j=1}^k x_j^{2a_j-1} dx_1 \ldots dx_k \\ &= \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty e^{-\sum_{j=1}^k x_j^2} \prod_{j=1}^k |x_j|^{2a_j-1} dx_1 \ldots dx_k.\end{aligned}$$

Now let $r = \sqrt{\sum_{j=1}^k x_j^2}$ and $z_j = \frac{x_j}{r}$ for $1 \le j \le k-1$. Using $x_j = rz_j$ for $j = 1, \ldots, k-1$ and $x_k = r\sqrt{1 - \sum_{j=1}^{k-1} z_j^2}$, the computation of the Jacobian easy and is left as an exercise:

$$J((x_1, \ldots, x_k) \to (r, z_1, \ldots, z_{k-1})) = \frac{r^{k-1}}{\sqrt{1 - \sum_{j=1}^{k-1} z_j^2}}.$$

Then

$$\begin{aligned}\prod_{j=1}^n \Gamma(a_j) = &\int_0^\infty e^{-r^2} r^{2(\sum_{j=1}^k a_j)-k} r^{k-1} dr \\ &\times \int_{-1}^1 \int_{-(1-z_1^2)}^{1-z_1^2} \cdots \int_{-(1-\sum_{j=1}^{k-2} z_j^2)}^{1-\sum_{j=1}^{k-2} z_j^2} \left(\prod_{j=1}^{k-1} z_j^{2a_j-1}\right) \left(1 - \sum_{j=1}^{k-1} z_j^2\right)^{a_k-1/2} \\ &\times \frac{1}{\sqrt{1 - \sum_{j=1}^{k-1} z_j^2}} \prod_{j=1}^{k-1} dz_j\end{aligned}$$

$$
\begin{aligned}
&= \Gamma\left(\sum_{j=1}^{k} a_j\right) \\
&\quad \times \int_0^1 \int_0^{1-z_1^2} \cdots \int_0^{1-\sum_{j=1}^{k-2} z_j^2} \left(\prod_{j=1}^{k-1} z_j^{2(a_j-1)}\right)\left(1-\sum_{j=1}^{k-1} z_j^2\right)^{a_k-1} \prod_{j=1}^{k-1} 2z_j dz_j \\
&= \Gamma\left(\sum_{j=1}^{k} a_j\right) D(a_1, \ldots, a_k)
\end{aligned}
$$

and the result follows. □

Theorem 1.1 *The function $\tilde{\pi}(\theta_1, \ldots, \theta_{k-1})$ defined by Equation (1.23) satisfies*

$$\int_0^1 \int_0^{1-\theta_1} \cdots \int_0^{1-\sum_{j=1}^{k-2}\theta_j} \tilde{\pi}(\theta_1, \ldots, \theta_{k-1}) d\theta_{k-1} \ldots d\theta_1 = 1,$$

hence the Dirichlet density (Definition 1.13) is a well defined probability density function.

Proof This follows directly from the lemma. □

Properties of the Dirichlet density Theorem 1.1 shows that the Dirichlet density is a probability density function.

Another very important property is that the Dirichlet densities $\text{Dir}(a_1, \ldots, a_k) : a_1 > 0, \ldots, a_k > 0$ form a family of distributions that is *closed under sampling*. Consider a prior distribution $\pi_\Theta \sim \text{Dir}(a_1, \ldots, a_k)$ and suppose that an observation $\underline{x} := (x_1, \ldots, x_k)$ is made on $\underline{Y} := (Y_1, \ldots, Y_k)$ based on n independent trials (i.e. $x_1 + \ldots + x_k = n$). Let $\pi_{\Theta|\underline{Y}}$ denote the posterior distribution. Then, using Bayes' rule,

$$\pi_{\Theta|\underline{Y}}(\theta_1, \ldots, \theta_k) = \frac{\pi_\Theta(\theta_1, \ldots, \theta_{k-1}) p_{\underline{Y}}(\underline{x}|\theta_1, \ldots, \theta_k)}{p_{\underline{Y}}(\underline{x})}.$$

It follows that

$$\pi_{\Theta|\underline{Y}}(\theta_1, \ldots, \theta_k) = \frac{1}{p_{\underline{Y}}(\underline{x})} \frac{n!}{x_1!x_2!\ldots x_{k-1}!x_k!} \theta_1^{a_1+x_1-1} \ldots \theta_k^{a_k+x_k-1},$$

where $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$.

Since the posterior density is a *probability* density, belonging to the Dirichlet family, it follows that the constant

$$\frac{1}{p_{\underline{Y}}(\underline{x})} \frac{n!}{x_1!x_2!\ldots x_{k-1}!x_k!} = \frac{\Gamma(a_1 + \cdots + a_k + x_1 + \cdots + x_k)}{\prod_{j=1}^k \Gamma(a_j + x_j)}$$

and hence that

$$\pi_{\Theta|\underline{Y}}(\theta_1, \ldots, \theta_k) \sim \text{Dir}(a_1 + x_1, \ldots, a_k + x_k).$$

The results in this section were perhaps first found by G. Lidstone [28].

Later in the text, the Dirichlet density will be written exclusively as a function of k variables, $\pi_\Theta(\theta_1, \ldots, \theta_k)$, where there are $k-1$ independent variables and $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$.

A question is how to select the hyper parameters $a_1, \ldots, a_k$ for the prior distribution. The choice of $a_1 = \ldots = a_k = \frac{1}{k}$ was suggested by W. Perks in [29].

Definition 1.14 (Conjugate Prior) *A prior distribution from a family that is closed under sampling is known as a* conjugate prior.

In [30], I.J. Good proved that exchangeability and *sufficientness* of samples implied that the prior is necessarily Dirichlet, if $k > 2$. The notion of sufficientness was originally defined by W.E. Johnson and I.J. Good. Loosely speaking, it means that the conditional probability of seeing case i of k possible in the next sample given n past samples, depends only on n, the number of times you have seen i in the past *and NOT on the other cases*.

Notes Full accounts of the coherence argument may be found, for example, in [4], [31] and [32]. An introduction to inductive logic is given in [33].

The monograph [34] includes a thorough presentation of the topics of statistical inference and Bayesian machine learning. The papers [13] and [16] argue for subjective probability as the appropriate inference and language procedures for artificial intelligence agents, see also [14]. The book [35] provides a clear introduction to the application of Bayesian methods in artificial intelligence.

The work [36] by Thomas Bayes (1702–1761) and Richard Price was published posthumously in 1763. This paper makes difficult reading for a modern mathematician. Consequently, there is a considerable literature investigating the question of what Bayes actually proved, see, e.g. [22, 37–39] and the references therein. There is, however, a wide consensus that [36] does contain Equation (1.17).

For this, Bayes deals with billiard balls. Suppose You throw one billiard ball o (orange) on a square table (e.g. a billiard table without pockets) and measure the shortest distance from the side of the table, when the side of the table is scaled in size to 1. Let this value be denoted by p. Then You throw n balls W (white) on the table and note the number of white balls, say k, to the left of the orange ball. Then it is understood that Bayes computed the distribution of p given k given by Equation (1.17).

In this setting the uniform prior distribution on p is based on a physical understanding that is verifiable by repeated experimentation.

There is even the question of whether Bayes was the *first* to discover the results attributed to him. This is discussed in [40]. Another up-to-date report on the life and thinking of the Reverend Thomas Bayes, by D.R. Bellhouse [41], also discusses the question of whether he was the first to prove these results. The author has discovered some previously unknown documents. The paper points out that the canonical picture of Bayes is not proved to be an image of him.[5]

An alternative procedure on the billiard table is that $n+1$ balls W are thrown on the table. One of them is then selected at random to play the role of the orange ball, and k,

[5] see http://www-history.mcs.st-andrews.ac.uk/PictDisplay/Bayes.html.

the number of balls to the left of the orange ball, is counted. Then You have a uniform distribution $\frac{1}{n+1}$ on the values of k.

It has been argued that Bayes demonstrated that a prior density $\pi(\theta)$ satisfying the equality

$$\frac{n!}{k!(n-k)!}\int_0^1 \theta^k(1-\theta)^{n-k}\pi(\theta)d\theta = \frac{1}{n+1} \tag{1.24}$$

for all $0 \leq k \leq n$ and all n must be the uniform density. It may be checked rather easily using Equation (1.16) that the uniform density indeed satisfies this equality. F.H. Murray in [42] observed that Equation (1.24) implies for $k = n$ that

$$\int_0^1 \theta^n \pi(\theta)d\theta = \frac{1}{n+1}, \tag{1.25}$$

which means that all the moments of $\pi(\theta)$ are given. Murray then went on to show that these moments determine a *unique* distribution, which is in fact the uniform distribution.[6]

The probability in Equation (1.24) is a uniform distribution on the number of successes in n Bernoulli trials with an unknown parameter. Hence Bayes (or Murray) has shown that *the uniform distribution on the number of successes is equivalent to the uniform density on the probability of success*. But this probability on the number of successes is a predictive probability on observables. This understanding of the Bayesian inference due to Thomas Bayes is different from many standard recapitulations of it, as pointed out in [39].

The ultimate question raised by reading of [36] is, ‘what is it all about?’. In other words, what was the problem that Bayes was actually dealing with?

It is hardly credible that Bayes, a clergyman, should have studied this as a mere curious speculation, and even less that scoring at a billiard room should have been at the forefront of his mind. Richard Price writes in [36],

> . . . the problem . . . mentioned [is] necessary to be solved in order to provide a sure foundation for all our reasoning concerning past facts, and what is likely to be hereafter . . .

For a layman in the history of philosophy the argument in [37] and [43] may carry a convincing power: Bayes and Price developed an inductive logic as a response to the critical and, in particular, anti-clerical objections to induction, causation and miracles advanced by David Hume [3] in his book of 1748; the famous philosopher and scholar was a contemporary of Bayes and Price.

Further evidence that this consideration may have prompted Bayes to develop a mathematical framework for inductive logic is seen from his theological interests. In 1731, he published the following paper: ‘Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures’.

[6] The *moment problem* is a classic problem; whether or not the moments of a distribution uniquely characterize the distribution. The technique usually employed is to check whether the *Carlemann conditions* are satisfied. For this problem, Murray showed *directly* that the moments uniquely determined the distribution.

The models that were later to be called Bayesian networks were introduced into artificial intelligence by J. Pearl, in the article [44]. Within the artificial intelligence literature, this is a seminal article, which concludes with the following statement: *The paper demonstrates that the centuries-old Bayes formula still retains its potency for serving as the basic belief revising rule in large, multi hypotheses, inference systems*.

1.10 Exercises: Probabilistic theories of causality, Bayes' rule, multinomial sampling and the Dirichlet density

1. This exercise considers the statistical notion of association due to G.U. Yule, who used it in a classical statistical study to demonstrate the positive effect of innoculation against cholera.

   Here the *association* between two events A and B, denoted by $\alpha(A, B)$, is defined as

   $$\alpha(A, B) = p(A \cap B) - p(A) \cdot p(B).$$

   (a) Show that

   $$\alpha(A, B) = -\alpha\left(A, B^c\right),$$

   where B^c is the complement of B.

   (b) Show that

   $$\alpha(A, B) = \alpha\left(A^c, B^c\right).$$

   Comment: Association is clearly symmetric. That is, for any two events A and B, $\alpha(A, B) = \alpha(B, A)$. It does not seem reasonable to claim that a decrease in cholera causes an increase in the number of innoculations. In this case it is common sense to conclude that there is an underlying causal relation, where innoculation (say B) causes a decreased prevalence of cholera (A), although without a controlled experiment, it is not possible to conclude that there is not a hidden factor C that both causes cholera and makes innoculation less likely.

2. **On a probabilistic theory of causality** Following the theory of causality due to P. Suppes [19] an event B_s is *defined* as a *prima facie cause*[7] of the event A_t if and only if the following three statements hold:

   - $s < t$,
   - $p(B_s) > 0$
   - $p(A_t \mid B_s) > p(A_t)$.

   Here the parameter is considered as a time parameter, and $s < t$ means that B_s occurs prior to A_t; a cause occurs before an effect.

   An event B_s is defined as a *prima facie negative cause* of an event A_t [19] if and only if the following three statements hold:

   - $s < t$,
   - $p(B_s) > 0$

[7] **Prima facie** is a Latin expression meaning 'on its first appearance', or 'by first instance'. Literally the phrase translates as first face, 'prima' first, 'facie' face. It is used in modern legal English to signify that on first examination, a matter appears to be self-evident from the facts. In common law jurisdictions, 'prima facie' denotes evidence that (unless rebutted) would be sufficient to prove a particular proposition or fact.

- $p(A_t \mid B_s) < p(A_t)$.

Intuitively, a negative cause is an event that prevents another event from happening. For example, the theory and practice of preventive medicine focuses on certain types of negative causation. In the problems the indices s, t are dropped for ease of writing.

(a) Show that if B^c is a prima facie negative cause of A, then B is a prima facie cause of A.

(b) Show that if B is a prima facie cause of A, then B^c is a prima facie cause of A^c. Also, show that if B is a prima facie *negative* cause of A, then B^c is a prima facie negative cause of A^c.

(c) Recall the definition of association from Exercise 1. Show that if B is a prima facie cause of A, then $\alpha(A, B) > 0$ and that if B is a prima facie negative cause of A, then $\alpha(A, B) < 0$.

3. **On odds and the weight of evidence** Let p be a probability distribution over a space $\mathcal{X}$. The *odds* of an event $A \subseteq \mathcal{X}$ given $B \subseteq \mathcal{X}$ under p, denoted by $O_p(A \mid B)$, is defined as

$$O_p(A \mid B) = \frac{p(A \mid B)}{p(A^c \mid B)}. \tag{1.26}$$

The odds ration will play an important role in Chapter 7, which considers sensitivity analysis. Next, the *weight of evidence* E in favour of an event A given B, denoted by $W(A : E \mid B)$, is defined as

$$W(A : E \mid B) = \log \frac{O_p(A \mid B \cap E)}{O_p(A \mid B)}. \tag{1.27}$$

Show that if $p(E \cap A^c \cap B) > 0$, then

$$W(A : E \mid B) = \log \frac{p(E \mid A \cap B)}{p(E \mid A^c \cap B)}. \tag{1.28}$$

4. **On a generalized odds and the weight of evidence** Let p denote a probability distribution over a space $\mathcal{X}$ and let $H_1 \subseteq \mathcal{X}$, $H_2 \subseteq \mathcal{X}$, $G \subseteq \mathcal{X}$ and $E \subseteq \mathcal{X}$. The *odds of H_1 compared to H_2* given G, denoted by $O_p(H_1/H_2 \mid G)$, is defined as

$$O_p(H_1/H_2 \mid G) = \frac{p(H_1 \mid G)}{p(H_2 \mid G)}. \tag{1.29}$$

The *generalized weight of evidence* is defined by

$$W(H_1/H_2 : E \mid G) = \log \frac{O_p(H_1/H_2 \mid G \cap E)}{O_p(H_1/H_2 \mid G)}. \tag{1.30}$$

Show that if $p(H_1 \cap G \cap E) > 0$ and $p(H_2 \cap G \cap E) > 0$ then

$$W(H_1/H_2 : E \mid B) = \log \frac{p(E \mid H_1 \cap G)}{p(E \mid H_2 \cap G)}. \tag{1.31}$$

Clearly this is just a log likelihood ratio and these notions are another expression for posterior odds = likelihood ratio × prior odds.

5. In [45], I.J. Good discusses the causes of an event that are necessary and sufficient from probabilistic view point. For example, let E is the event of being hit by a car and F the event of going for a walk. Then F tends to be a necessary cause of E. The quantitites $Q_{\text{suf}}\,(E : F \mid U)$ and $Q_{\text{nec}}\,(E : F \mid U)$ are defined to measure the probabilistic tendency of an event F to be a sufficient and/or necessary cause, respectively, for an event E with background information U, by the weights of evidence discussed in the preceding exercise. They are defined respectively by

$$Q_{\text{suf}}\,(E : F \mid U) = W\left(F^c : E^c \mid U\right) \tag{1.32}$$

and

$$Q_{\text{nec}}\,(E : F \mid U) = W\,(F : E \mid U)\,. \tag{1.33}$$

In view of the preceding definitions, Q_{suf} may be read as the weight of evidence against F provided by non-occurrence of E. Similarly, Q_{nec} is the the weight of evidence in favour of F given by occurrence of E. Both quantities are computed, to borrow a philosophical phrase, 'given the state of universe U just before F occurred'.

(a) If $p\,(E^c \mid F \cap U) > 0$, show that

$$Q_{\text{suf}}\,(E : F \mid U) = \log \frac{p\,(E^c \mid F^c \cap U)}{p\,(E^c \mid F \cap U)}. \tag{1.34}$$

(b) If $p\,(E \mid F^c \cap U) > 0$, show that

$$Q_{\text{nec}}\,(E : F \mid U) = \log \frac{p\,(E \mid F \cap U)}{p\,(E \mid F^c \cap U)}. \tag{1.35}$$

6. This exercise considers a few more properties of Q_{suf} and Q_{nec}. Following Exercise 4 above, set

$$Q_{\text{nec}}\,(E : F_1/F_2 \mid U) = W\,(F_1/F_2 : E \mid U)$$

which is the *necessitivity* of E of F_1 against F_2 and

$$Q_{\text{suf}}\,(E : F_2/F_1 \mid U) = W\left(F_1/F_2 : E^c \mid U\right),$$

which is the *sufficientivity* of E of F_1 against F_2.

(a) Show that $Q_{\text{suf}}\,(E : F \mid U) < 0$, if and only if $Q_{\text{nec}}\,(E : F \mid U) < 0$. Compare with *prima facie negative cause* in Exercise 2.

(b) Show that

$$Q_{\text{nec}}\,(E : F_1/F_2 \mid U) = Q_{\text{suf}}\left(E^c : F_2/F_1 \mid U\right).$$

This is called a *probabilistic contraposition*.

(c) Show that

$$Q_{\text{nec}}\left(E : F \mid U\right) = Q_{\text{suf}}\left(E^c : F^c \mid U\right).$$

This may interpreted along the following lines. Going for a walk F tends to be a necessary cause for being hit by a vehicle E, whereas staying home tends to be a sufficient cause for not being hit by a vehicle. (Note that cars and aircraft are known to have crashed into houses.) Both Q_{nec} and Q_{suf} should have high values in this case.

7. Let $\underline{X} = (X_1, \ldots, X_n)^t$ be an exchangeable sample of Bernoulli trials and let $T = \sum_{j=1}^{n} X_j$. Show that there is a probability density function π such that

(a)

$$p_T(t) = \int_0^1 \binom{n}{t} \theta^t (1-\theta)^{n-t} \pi(\theta) d\theta, \quad t = 0, 1, \ldots, n$$

(b)

$$E[T] = n \int_0^1 \theta \pi(\theta) d\theta.$$

You may use the result of DeFinetti.

8. Consider a sequence of n independent, identically distributed Bernoulli trials, with unknown parameter θ, the 'success' probability. For a *uniform prior* over θ, show that the posterior density for θ, if the sequence has k successes, is

$$\pi_{\Theta|\underline{x}}\left(\theta \mid \underline{x}\right) = \begin{cases} \frac{(n+1)!}{k!(n-k)!} \cdot \theta^k \left(1-\theta\right)^{n-k} & 0 \le \theta \le 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{1.36}$$

9. Consider the thumb-tack experiment and the conditional independence model for the problem and the uniform prior density for θ. What is $P_{X_{n+1}|\underline{X}_{(n)}}\left(\text{head}|\underline{x}_{(n)}\right)$, where $\underline{x}_{(n)}$ denotes the outcome of the first n throws?

10. Consider multinomial sampling, where θ_j is the probability that category j is obtained, with prior density $\pi_\Theta(\theta_1, \ldots, \theta_L)$ is the Dirichlet prior $\text{Dir}(\alpha q_1, \ldots, \alpha q_L)$ with $\sum_{j=1}^{L} q_j = 1$, defined by

$$\pi_\Theta(\theta) = \begin{cases} \frac{\Gamma(\alpha)}{\prod_{j=1}^{L} \Gamma(\alpha q_j)} \prod_{j=1}^{L} \theta_j^{\alpha q_j - 1} & \theta_1 + \ldots + \theta_L = 1, 0 \le \theta_i \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Show that for multinomial sampling, with the Dirichlet prior, the posterior density $p_{\Theta|\underline{x}}\left(\theta|\underline{x}; \alpha\right)$ is the Dirichlet density

$$Dir\left(n_1 + \alpha q_1, \ldots, n_L + \alpha q_L\right),$$

which is shorthand for

$$\pi_{\Theta|\underline{X}_{(n)}}\left(\theta|\underline{x}_{(n)}; \alpha \underline{q}\right) = \frac{\Gamma\left(n+\alpha\right)}{\prod_{i=1}^{L} \Gamma\left(\alpha q_i + n_i\right)} \prod_{i=1}^{L} \theta_i^{n_i + \alpha q_i - 1}, \tag{1.37}$$

where $\underline{q} = (q_1, \ldots, q_L)$.

11. A useful property of the Dirichlet density is that the predictive distribution of X_{n+1} may be computed explicitly by integrating $p_{X_{n+1}|\Theta}(. \mid \theta)$ with respect to the posterior distribution containing the stored experience $\underline{x}_{(n)}$. Using the previous exercise, show that

$$p_{X_{n+1}|\underline{X}_{(n)}}\left(x_i \mid \underline{x}_{(n)}\right) = \int_{S_L} \theta_i \pi\left(\theta_1, \ldots, \theta_L | \underline{x}; \alpha \underline{q}\right) d\theta_1 \ldots d\theta_L = \frac{n_i + \alpha q_i}{n + \alpha}. \quad (1.38)$$

12. Let $\underline{\Theta} = (\Theta_1, \ldots, \Theta_L)$ be a continuous random vector with $Dir(\alpha_1, \ldots, \alpha_L)$ distribution. Compute $\operatorname{Var}(\Theta_i)$.

13. Prove the *Laplace rule of succession*. Namely, let $\{X_1, \ldots, X_{n+1}\}$ be independent, identically distributed Bernoulli random variables, where $p_{X_i}(1) = 1 - p_{X_i}(0) = \theta$ and $\theta \sim U(0, 1)$. Then the Laplace rule of succession states that

$$p(\{X_{n+1} = 1\}|\{X_1 + \ldots + X_n = s\}) = \frac{s+1}{n+2}.$$

14. Let $\underline{V} = (V_1, \ldots, V_K)$ be a continuous random vector, with

$$\underline{V} \sim \operatorname{Dir}(a_1, \ldots, a_K),$$

and set

$$U_i = \frac{V_i x_i^{-1}}{\sum_{i=1}^K V_i x_i^{-1}}, \quad , i = 1, \ldots, K,$$

where $\underline{x} = (x_1, \ldots, x_K)$ is a vector of positive real numbers; that is, $x_i > 0$ for each $i = 1, \ldots, K$. Show that $\underline{U} = (U_1, \ldots, U_K)$ has density function

$$\frac{\Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^K \Gamma(a_i)} \prod_{i=1}^K u_i^{a_i - 1} \left(\frac{1}{\sum_{i=1}^K u_i x_i}\right)^{\sum_{i=1}^K a_i} \prod_{i=1}^K x_i^{a_i}.$$

This density is denoted

$$\underline{U} \sim S\left(\underline{a}, \underline{x}\right).$$

This is due to J.L. Savage [46]. Note that the Dirichlet density is obtained as a special case when $x_i = c$ for $i = 1, \ldots, K$.

15. The next two examples illustrate how the Savage distribution of the previous exercise can arise in Bayesian analysis, for updating an objective distribution over the subjective assessments of a probability distribution by several different researchers, faced with a common set of data. Consider several researchers studying an unknown quantity X, where X can take values in $\{1, 2, \ldots, K\}$. Each researcher has his own initial assessment of the probability distribution $\underline{V} = (V_1, \ldots, V_K)$ for the value that X takes. That is, for a particular researcher,

$$V_i = p_X(i), \quad i = 1, \ldots, K.$$

It is assumed that

$$\underline{V} \sim \mathrm{Dir}\left(a_1, \ldots, a_K\right).$$

Each researcher observes the same set of data with the common likelihood function

$$l_i = p\left(\mathrm{data}|\{X = i\}\right), \quad i = 1, \ldots, K.$$

The coherent posterior probability of a researcher is

$$U_i = p\left(\{X = i\} \mid \mathrm{data}\right), \quad i = 1, 2, \ldots, K.$$

Let $\underline{U} = (U_1, \ldots, U_K)$. Prove that

$$\underline{U} \sim S\left(\underline{a}, \underline{l}^{-1}\right),$$

where $\underline{a} = (a_1, \ldots, a_K)$ and $\underline{l}^{-1} = \left(l_1^{-1}, \ldots, l_K^{-1}\right)$. This is due to J.M. Dickey [47].

16. Show that the family of distributions $S\left(\underline{a}, \underline{l}^{-1}\right)$ is closed under updating of the opinion populations. In other words, if

$$\underline{V} \sim S\left(\underline{a}, \underline{z}\right),$$

before the data is considered, then

$$\underline{U} \sim S\left(\underline{a}, \underline{z} \times \underline{l}^{-1}\right),$$

after the data update, where

$$\underline{z} \times \underline{l}^{-1} = \left(z_1 l_1^{-1}, \ldots, z_K l_K^{-1}\right).$$