1 INTRODUCTION

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of probabilities
—James Clerk Maxwell

Inference is the process of drawing conclusions from evidence. While deductive logic offers a framework for inferring conclusions from absolute statements of truth, it does not apply to the more typical real-world setting where the information we receive is uncertain. To make inferences under conditions of uncertainty, we must instead turn to probability theory, which both offers a consistent framework for quantifying our beliefs and making inferences given these beliefs.

The key computational task in inference is computing integrals with respect to probability distributions on the variables in a proposed model. Typically these integrals will not have analytic solutions and the large number of variables being integrated over mean numerical quadrature methods are impractically costly. In these cases we must resort to approximate methods which tradeoff an introduction of error for an increase in computational tractability. *Markov chain Monte Carlo* (MCMC) methods are a very generally applicable class of approximate inference techniques which estimate the integrals of interest by computing averages over the states of a Markov chain.

The topic of this thesis is the development of MCMC methods. In particular we introduce several novel methods which exploit reparameterisations and augmentations of the state of a Markov chain to improve upon the computational efficiency, ease of use or degree of approximation error of existing approaches.

In this chapter we discuss the basic concepts of probabilistic modelling which underpin the inference methods discussed in later chapters. In particular we review the terminology and basic concepts of the measure-theoretic description of probability as some of the later results in the thesis are most clearly described within this framework. We also introduce graphical models as a compact way of visualising structure in probabilistic models. Finally we conclude with a discussion of the specific inference problems that the methods presented in the rest of this thesis are intended to help tackle.

1.1 PROBABILITY THEORY

A σ-algebra, $\mathcal{E}$, on a set S is set of subsets of S with $S \in \mathcal{E}$, $\emptyset \in \mathcal{E}$ and which is closed under complement and countable unions and intersections.

A *probability space* is defined as a triplet $(S, \mathcal{E}, \mathsf{P})$ where

- S is the *sample space*, the set of all possible outcomes,
- $\mathcal{E}$ is the *event space*, a σ-algebra on S, defining all possible events (measurable subsets of S),
- P is the *probability measure*, a finite measure satisfying $\mathsf{P}(S) = 1$, which specifies the probabilities of events in $\mathcal{E}$.

Kolmogorov's axioms:

1. *Non-negativity:* $\mathsf{P}(E) \geq 0 \ \forall E \in \mathcal{E}$,
2. *Normalisation:* $\mathsf{P}(S) = 1$,
3. *Countable additivity: for any countable set of disjoint events* $\{E_i\}_i : E_i \in \mathcal{F} \ \forall i$, $E_i \cap E_j = \emptyset \ \forall i \neq j$, $\mathsf{P}(\cup_i E_i) = \sum_i \mathsf{P}(E_i)$.

Given this definition of a probability space, Kolmogorov's axioms [15] can be used to derive a measure-theoretic formulation of probability theory. The probability of an event $E \in \mathcal{E}$ is defined as its measure $\mathsf{P}(E)$. Two events $A, B \in \mathcal{E}$ are *independent* if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$.

The key advantage of the measure-theoretic approach to probability is that it provides a consistent definition of the probability of an event in any space we can define a measure on. This allows a unified treatment of the common cases of probability distributions of discrete and continuous random variables but also makes it possible to consider distributions on more general spaces. In Chapter 4 we will consider problems which involve distributions on implicitly-defined manifolds where this generality will be key to understanding the proposed methods.

1.1.1 Random variables

If $(X, \mathcal{F})$ and $(Y, \mathcal{G})$ are two measurable spaces, a function $f : X \to Y$ is measurable if $f^{-1}(E) \in \mathcal{F} \ \forall E \in \mathcal{G}$.

When modelling real-world processes, rather than considering events as subsets of an abstract sample space, it is usually more helpful to consider *random variables* which represent quantities in the model of interest. A random variable $\mathrm{x} : S \to X$ is defined as a measurable function from the sample space to a measurable space $(X, \mathcal{F})$.

The Borel σ-algebra $\mathscr{B}(\mathbb{R})$ is the smallest σ-algebra on $\mathbb{R}$ which contains all open real intervals.

Often X is the reals, $\mathbb{R}$, and $\mathcal{F}$ is the Borel σ-algebra on the reals, $\mathscr{B}(\mathbb{R})$, in which case we refer to a *real random variable*. It is also common to consider cases where X is a real vector space, $\mathbb{R}^D$, and $\mathcal{F} = \mathscr{B}(\mathbb{R}^D)$ - in this case refer to a *real random vector* and use the notation $\mathbf{x} : S \to X$. A final specific case is when X is countable and $\mathcal{F}$ is the power set $\mathscr{P}(X)$ in which case we refer to x as a *discrete random variable*. As we are most often concerned with real-valued random variables and vectors in this thesis, when it is unambiguous to do so we drop the 'real' qualifier and simply refer to *random variables* and *random vectors*.

Due to the definition of a random variable as a measurable function, we can define a pushforward measure on a random variable x

$$\mathsf{P}_{\mathsf{x}}(A) = \mathsf{P} \circ \mathsf{x}^{-1}(A) = \mathsf{P}(\{s \in S : \mathsf{x}(s) \in A\}) \quad \forall A \in \mathcal{F}. \tag{1.1}$$

If $(X, \mathcal{F})$ and $(Y, \mathcal{G})$ are two measurable spaces, μ a measure on these spaces and $f : X \to Y$ a measurable function, the pushforward measure μ_f satisfies $\mu_f(A) = \mu \circ f^{-1}(A)$ $\forall A \in \mathcal{G}$.

The measure P_{x} specifies that the probability of the event that the random variable x takes a value in a measurable set $A \in \mathcal{F}$ is $\mathsf{P}_{\mathsf{x}}(A)$. We typically describe P_{x} as the *distribution* of x.

1.1.2 Joint and conditional probability

Typically we will jointly define multiple random variables on the same probability space. Let $(S, \mathcal{E}, \mathsf{P})$ be a probability space and $\mathsf{x} : S \to X$, $\mathsf{y} : S \to Y$ be two random variables with corresponding σ-algebras $\mathcal{F}$ and $\mathcal{G}$. Then the *joint probability* of x and y is defined as

$$\mathsf{P}_{\mathsf{x,y}}(A, B) = \mathsf{P}\left(\mathsf{x}^{-1}(A) \cap \mathsf{y}^{-1}(B)\right) \quad \forall A \in \mathcal{F}, B \in \mathcal{G}. \tag{1.2}$$

The joint probability is related to P_{x} and P_{y} by

$$\mathsf{P}_{\mathsf{x,y}}(A, Y) = \mathsf{P}_{\mathsf{x}}(A), \; \mathsf{P}_{\mathsf{x,y}}(X, B) = \mathsf{P}_{\mathsf{y}}(B) \quad \forall A \in \mathcal{F}, B \in \mathcal{G}. \tag{1.3}$$

In this context P_{x} and P_{y} are referred to as *marginal distributions* of the joint distribution. Two random variables x and y are said to be independent if and only if

$$\mathsf{P}_{\mathsf{x,\,y}}(A, B) = \mathsf{P}_{\mathsf{x}}(A)\,\mathsf{P}_{\mathsf{y}}(B) \quad \forall A \in \mathcal{F}, B \in \mathcal{G}. \tag{1.4}$$

A particularly key concept for inference is the definition of *conditional probability*. The conditional probability of an event $A \in \mathcal{E}$ occuring given another event $B \in \mathcal{E}$ has occured is denoted $\mathsf{P}(A \,|\, B)$ and we have the definition

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} \quad \forall A \in \mathcal{E}, B \in \mathcal{E} : \mathsf{P}(B) \neq 0. \tag{1.5}$$

In Kolmogorov's probability theory, (1.5) is given as an additional definition distinct from the basic axioms. In alternatives such as the work of Cox [5, 6] and de Finetti [9], conditional probabilities are instead viewed as a primitive.

Correspondingly we denote the conditional probability of the event of the random variable x taking a value in $A \in \mathcal{F}$ given the event that the random variable y takes a value in $B \in \mathcal{G}$ as $\mathsf{P}_{\mathsf{x|y}}(A \,|\, B)$. Using (1.5) and (1.2), $\mathsf{P}_{\mathsf{x|y}}$ and $\mathsf{P}_{\mathsf{y|x}}$ can be shown to satisfy

$$\begin{aligned} \mathsf{P}_{\mathsf{x,y}}(A, B) = \mathsf{P}_{\mathsf{x|y}}(A \,|\, B)\,\mathsf{P}_{\mathsf{y}}(B) = \mathsf{P}_{\mathsf{y|x}}(B \,|\, A)\,\mathsf{P}_{\mathsf{x}}(A) \\ \forall A \in \mathcal{F}, B \in \mathcal{G}. \end{aligned} \tag{1.6}$$

This decomposition of a joint probability into a product of a conditional and marginal is sometimes referred to as the product rule. An implication of (1.6) is what is often termed *Bayes' theorem*

$$\mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \mid B) = \frac{\mathsf{P}_{\mathsf{y}|\mathsf{x}}(B \mid A)\, \mathsf{P}_{\mathsf{x}}(A)}{\mathsf{P}_{\mathsf{y}}(B)} \quad \forall A \in \mathcal{F},\, B \in \mathcal{G} : \mathsf{P}_{\mathsf{y}}(B) \neq 0, \qquad (1.7)$$

which will be of key importance in the later discussion of inference.

The definition in (1.2) of the joint probability of a pair of random variables can be extended to arbitarily large collections of random variables. Similarly conditional probabilities can be defined for collections of multiple jointly dependent random variables, with the product rule given in (1.6) generalising to a combinatorial number of possible factorisations of the joint probability. Graphical models offer a convenient way of representing the dependencies between large collections of random variables and any resulting factorisation structure in their joint probability, and are discussed in Section 1.2.

1.1.3 Probability densities

So far we have not specified how the probability measure P is defined and by consequence the probability (distribution) of a random variable. The Radon–Nikodym theorem guarantees that for a pair of σ−finite measures μ and ν on a measurable space $(X, \mathcal{F})$ where ν is absolutely continuous with respect to μ, then there is a unique (up to μ-null sets) measurable function $f : X \to [0, \infty)$ termed a *density* such that

A measure on X is σ-finite if X is a countable union of finite measure sets.

$$\nu(A) = \int_A f \,\mathrm{d}\mu = \int_A f(x)\,\mu(\mathrm{d}x) \quad \forall A \in \mathcal{F}. \qquad (1.8)$$

The two Lebesgue integral notations above are equivalent and we will use them interchangeably. The density function f is also termed the *Radon-Nikodym derivative* of ν with respect to μ, denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. Density functions therefore represent a convenient way to define a probability distribution with respect to a reference measure we will term the *base measure*. The key requirement defining what is an appropriate base measure to use it that the probability measure of interest is absolutely continuous with respect to it.

If μ and ν are measures on a measurable space $(X, \mathcal{F})$ then ν has absolute continuity WRT to μ if $\forall A \in \mathcal{F}$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$.

It can also be shown that if $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ and g is a measurable function

$$\int_X g(x)\,\nu(\mathrm{d}x) = \int_X g(x)\, f(x)\,\mu(\mathrm{d}x), \qquad (1.9)$$

which we will use later when discussing calculation of expectations.

Real random variables will typically have a distribution P_{x} defined by a probability density $\mathsf{p}_{\mathsf{x}} : \mathbb{R} \to [0, \infty)$ with respect to the *Lebesgue measure*, λ, on $\mathbb{R}$,

$$\mathsf{P}_{\mathsf{x}}(A) = \int_A \mathsf{p}_{\mathsf{x}}(x)\,\lambda(\mathrm{d}x) = \int_A \mathsf{p}_{\mathsf{x}}(x)\,\mathrm{d}x \qquad \forall A \in \mathscr{B}(\mathbb{R}). \tag{1.10}$$

The Lebesgue measure assigns a measure to subsets of a Euclidean space, and for $\mathbb{R}$, $\mathbb{R}^2$ and $\mathbb{R}^3$ formalises the intuitive concepts of length, area and volume of subsets respectively.

Analagously for a random vector $\mathbf{x}$ with density $\mathsf{p}_{\mathbf{x}} : \mathbb{R}^D \to [0, \infty)$ with respect to the D-dimensional Lebesgue measure λ^D we have that

$$\mathsf{P}_{\mathbf{x}}(A) = \int_A \mathsf{p}_{\mathbf{x}}(\boldsymbol{x})\,\lambda^D(\mathrm{d}\boldsymbol{x}) = \int_A \mathsf{p}_{\mathbf{x}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall A \in \mathscr{B}(\mathbb{R}^D). \tag{1.11}$$

The notation in the second equalities in (1.10) and (1.11) uses a convention that will be used throughout this thesis that integrals without an explicit measure are with respect to the Lebesgue measure.

The probability distribution of a discrete random variable can be defined via probability density $\mathsf{p}_{\mathsf{x}} : X \to [0, 1]$ with respect to the *counting measure* $\#$,

$$\mathsf{P}_{\mathsf{x}}(A) = \int_A \mathsf{p}_{\mathsf{x}}(x)\,\#(\mathrm{d}x) = \sum_{x \in A} \mathsf{p}_{\mathsf{x}}(x) \qquad \forall A \in \mathscr{P}(X). \tag{1.12}$$

The counting measure $\#$ is defined as $\#(A) = |A|$ for all finite A and $\#(A) = +\infty$ otherwise.

The co-domain of a probability density p_{x} for a discrete random variable is restricted to $[0, 1]$ due to the non-negativity and normalisation requirements for the probability measure P_{x}, with $\sum_{x \in X} \mathsf{p}_{\mathsf{x}}(x) = 1$. Commonly for the case of a discrete random variable, the density p_{x} is instead referred to as a *probability mass function*, with density reserved for real random variables. We will however use *probability density* in both cases in keeping with the earlier definition of a density, this avoiding difficulties with terminology and notation when definining joint probabilities on a mixture of real and discrete random variables.

The joint probability $\mathsf{P}_{\mathsf{x},\mathsf{y}}$ of a pair of random variables x and y with co-domains the measurable spaces $(X, \mathcal{F})$ and $(Y, \mathcal{G})$ respectively, can be defined via a joint probability density $\mathsf{p}_{\mathsf{x},\mathsf{y}} : X \times Y \to [0, \infty)$ with respect to a product measure $\mu_{\mathsf{x},\mathsf{y}} = \mu_{\mathsf{x}} \times \mu_{\mathsf{y}}$ by

$$\mathsf{P}_{\mathsf{x},\mathsf{y}}(A, B) = \int_{A \times B} \mathsf{p}_{\mathsf{x},\mathsf{y}}(x, y)\,\mu_{\mathsf{x},\mathsf{y}}(\mathrm{d}x, \mathrm{d}y) \tag{1.13}$$

If $(X_1, \mathcal{F}_1, \mu_1)$ and $(X_2, \mathcal{F}_2, \mu_2)$ are two measure spaces, the product measure $\mu_1 \times \mu_2$ on a measurable space $(X_1 \times X_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ is defined as satisfying $(\mu_1 \times \mu_2)(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ $\forall A_1 \in \mathcal{F}_1,\ A_2 \in \mathcal{F}_2$.

As a consequence of Fubini's theorem, the integral over $\mu_{\mathrm{x,y}}$ can be expressed as iterated integrals over μ_{x} and μ_{y}

$$\begin{aligned}\mathrm{P}_{\mathrm{x,y}}(A, B) &= \int_A \int_B \mathrm{p}_{\mathrm{x,y}}(x, y)\, \mu_{\mathrm{x}}(\mathrm{d}x)\, \mu_{\mathrm{y}}(\mathrm{d}y) \\ &= \int_B \int_A \mathrm{p}_{\mathrm{x,y}}(x, y)\, \mu_{\mathrm{y}}(\mathrm{d}y)\, \mu_{\mathrm{x}}(\mathrm{d}x) \quad \forall A \in \mathcal{F},\, B \in \mathcal{G}.\end{aligned} \tag{1.14}$$

The two measures μ_{x} and μ_{y} can differ for example $\mu_{\mathrm{x}} = \lambda$ and $\mu_{\mathrm{y}} = \#$ if x is a real random variable and y is a discrete random variable.

When dealing with random variables, we will often only specify the co-domain of the random variable(s) and a (joint) probability density, with the base measure being implicitly defined as the Lebesgue measure for real random variables (or vectors), counting measure for discrete random variables and an appropriate product measure for a mix of random variables. Similarly we will usually neglect to explicitly define the probability space $(S, \mathcal{E}, \mathrm{P})$ which the random variable(s) map from. In this case we will typically use the loose notation $\mathrm{x} \in X$ to mean a random variable x with co-domain X.

Tables A.1, A.2 and A.3 in Appendix A give definitions of the densities and shorthand notation for some common parametric probability distributions that we use in this thesis.

1.1.4 Transforms of random variables

A key concept we make use of in this thesis is defining transformations of random variables. Let x be a random variable with co-domain the measurable space $(X, \mathcal{F})$. Further let $(Y, \mathcal{G})$ be a second measurable space and $\phi : X \to Y$ a measurable function between the two spaces. If we define $\mathrm{y} = \phi \circ \mathrm{x}$ then analagously to our original definition of P_{x} as the pushforward measure of P under the measurable function defining x, we can define P_{y} in terms of P_{x} as

$$\mathrm{P}_{\mathrm{y}}(A) = \mathrm{P}_{\mathrm{x}} \circ \phi^{-1}(A) = \mathrm{P}_{\mathrm{x}}(\{x \in X : \phi(x) \in A\}) \quad \forall A \in \mathcal{G}, \tag{1.15}$$

i.e. the probability of the event $\mathrm{y} \in A$ is equal to the probability of x taking a value in the pre-image under ϕ of A. To calculate probabilities of transformed random variables therefore we will therefore need to be able to find the pre-images of values of the transformed variable.

If the distribution $\mathrm{P_x}$ is defined by a density $\mathrm{p_x}$ with respect to a measure μ_x, we can also in some cases find a density $\mathrm{p_y}$ on the transformed variable $\mathrm{y} = \phi(\mathrm{x})$ with respect to a (potentially different) measure μ_y

$$\mathrm{P_y}(A) = \int_{\phi^{-1}(A)} \mathrm{p_x}(x)\,\mu_\mathrm{x}(\mathrm{d}x) = \int_A \mathrm{p_y}(y)\,\mu_\mathrm{y}(\mathrm{d}y) \quad \forall A \in \mathcal{G}. \tag{1.16}$$

For random variables with countable co-domains where the integral in (1.16) corresponds to a sum, a $\mathrm{p_y}$ satisfying (1.16) is simple to identify. If x is a discrete random variable with probability density $\mathrm{p_x}$ with respect to the counting measure, then $\mathrm{y} = \phi(\mathrm{x})$ will necessarily also be a discrete random variable. Applying (1.16) for $\mathrm{p_x} = \frac{\mathrm{dP_x}}{\mathrm{d}\#}$ we have that[1]

$$\begin{aligned}\int_{\phi^{-1}(A)} \mathrm{p_x}(x)\,\#(\mathrm{d}x) &= \sum_{x\in\phi^{-1}(A)} \mathrm{p_x}(x) = \sum_{y\in A}\sum_{x\in\phi^{-1}(y)} \mathrm{p_x}(x)\\ &= \int_A \sum_{x\in\phi^{-1}(y)} \mathrm{p_x}(x)\,\#(\mathrm{d}y) \quad \forall A \in \mathcal{G}.\end{aligned} \tag{1.17}$$

We can therefore define $\mathrm{p_y} = \frac{\mathrm{dP_y}}{\mathrm{d}\#}$ in terms of $\mathrm{p_x}$ as

$$\mathrm{p_y}(y) = \sum_{x\in\phi^{-1}(y)} \mathrm{p_x}(x) \quad \forall y \in Y. \tag{1.18}$$

In the special case that ϕ is bijective with an inverse ϕ^{-1} we have that

$$\mathrm{p_y}(y) = \mathrm{p_x} \circ \phi^{-1}(y) \quad \forall y \in Y. \tag{1.19}$$

For transformations of real random variables and vectors, the situation is more complicated as we need to account for any local contraction or expansion of space by the transformation. We will consider here the special case where the transformation is a *diffeomorphism*: a differentiable bijection which has an inverse which is also differentiable. In Chapter 4 we will consider how this can be generalised to non-bijective differentiable functions, including the case where the dimensionalities of the domain and co-domain of the function differ.

[1] We use $\phi^{-1}(y)$ as a shorthand here for $\phi^{-1}(\{y\})$ i.e. the pre-image of a singleton set. In cases where an inverse function exists we will also use the same notation, which of the three meanings is intended should be clear from the context.

Let $X \subseteq \mathbb{R}^N$ and $Y \subseteq \mathbb{R}^N$ and $\boldsymbol{\phi} : X \to Y$ be a diffeomorphism. Then the *Jacobian* of $\boldsymbol{\phi}$ is defined as

$$\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{x}) = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_N}{\partial x_1} & \cdots & \frac{\partial \phi_N}{\partial x_N} \end{bmatrix}. \tag{1.20}$$

The Jacobian matrix describes the local transformation of an infinitesimal volume element $\mathrm{d}\boldsymbol{x}$ in X under the map $\boldsymbol{\phi}$. In particular the corresponding volume element in Y under the map will be an infinitesimal parallelotope spanned by the columns of the Jacobian $\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{x})$. The Jacobian matrix determinant $\left|\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{x})\right|$ which corresponds to the volume of this parallelotope therefore determines hows the volume elements scales unders the map - a value more than one corresponds to a local expansion and less than one to a contraction. Informally we therefore have that $\mathrm{d}\boldsymbol{y} = \left|\mathbf{J}_{\boldsymbol{\phi}}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x}$ and applying the same arguments to the inverse map $\boldsymbol{\phi}^{-1}$, $\mathrm{d}\boldsymbol{x} = \left|\mathbf{J}_{\boldsymbol{\phi}^{-1}}(\boldsymbol{y})\right| \mathrm{d}\boldsymbol{y}$.

Let $\mathbf{x}$ be a random vector taking values in the measurable space $(X, \mathscr{B}(X))$ and define $\mathbf{y} = \boldsymbol{\phi} \circ \mathbf{x}$ as a random vector taking values in $(Y, \mathscr{B}(Y))$. If $\mathsf{P}_{\mathbf{x}}$ has a density $\mathsf{p}_{\mathbf{x}}$ with respect to the Lebesgue measure

$$\begin{aligned} \mathsf{P}_{\mathbf{y}}(A) = \mathsf{P}_{\mathbf{x}} \circ \boldsymbol{\phi}^{-1}(A) &= \int_{\boldsymbol{\phi}^{-1}(A)} \mathsf{p}_{\mathbf{x}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &= \int_A \mathsf{p}_{\mathbf{x}} \circ \boldsymbol{\phi}^{-1}(\boldsymbol{y}) \left|\mathbf{J}_{\boldsymbol{\phi}^{-1}}(\boldsymbol{y})\right| \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{1.21}$$

Therefore $\mathsf{P}_{\mathbf{y}}$ has a density $\mathsf{p}_{\mathbf{y}}$ with respect to the Lebesgue measure

$$\mathsf{p}_{\mathbf{y}}(\boldsymbol{y}) = \mathsf{p}_{\mathbf{x}} \circ \boldsymbol{\phi}^{-1}(\boldsymbol{y}) \left|\mathbf{J}_{\boldsymbol{\phi}^{-1}}(\boldsymbol{y})\right| \quad \forall \boldsymbol{y} \in Y. \tag{1.22}$$

In both the cases considered, we have seen that if the function ϕ the random variable x is mapped through is bijective, it is tractable to compute a density on the mapped random variable y as the pre-image $\phi^{-1}(y)$ of a point $y \in Y$ is itself a point. Bijectivity is a very limiting condition however, with many models involving non-bijective transformations of random variables. In Chapter 4 we will discuss methods for performing inference in generative models defined by complex, non-dimension preserving and non-bijective transformations of random variables.

1.1.5 Expectations

A key operation when working with probabilistic models is computing expectations. Let $(S, \mathcal{E}, \mathsf{P})$ be a probability space, and $\mathsf{x} : S \to X$ a random variable on this space. The *expected value of* x is defined as

$$\mathbb{E}[\mathsf{x}] = \int_S \mathsf{x}\,\mathrm{d}\mathsf{P}. \tag{1.23}$$

Usually it will be more convenient to express expectations in terms of the probability P_x. If $g : X \to \mathbb{R}$ is an integrable function then

$$\int_X g(x)\,\mathsf{P}_\mathsf{x}(\mathrm{d}x) = \int_S g \circ \mathsf{x}(s)\,\mathsf{P}(\mathrm{d}s). \tag{1.24}$$

If we take g as the identity map we therefore have that

$$\mathbb{E}[\mathsf{x}] = \int_X x\,\mathsf{P}_\mathsf{x}(\mathrm{d}x). \tag{1.25}$$

If P_x is given by a density $\mathsf{p}_\mathsf{x} = \frac{\mathrm{d}\mathsf{P}_\mathsf{x}}{\mathrm{d}\mu}$ then using (1.9) we also have

$$\mathbb{E}[\mathsf{x}] = \int_X x\,\mathsf{p}_\mathsf{x}(x)\,\mu(\mathrm{d}x). \tag{1.26}$$

A further useful implication of (1.24) is what is sometimes termed the *Law of the unconscious statistician*. Let $\mathsf{x} : S \to X$ be a random variable, $\phi : X \to Y$ a measurable function and define $\mathsf{y} = \phi \circ \mathsf{x}$. Then

$$\mathbb{E}[\mathsf{y}] = \int_S \mathsf{y}(s)\,\mathsf{P}(\mathrm{d}s) = \int_S \phi \circ \mathsf{x}(s)\,\mathsf{P}(\mathrm{d}s) = \int_X \phi(x)\,\mathsf{P}_\mathsf{x}(\mathrm{d}x), \tag{1.27}$$

i.e. it can be calculated by integrating ϕ with respect to P_x. This means we can calculate the expected value of a transformed random variable $\mathsf{y} = \phi(\mathsf{x})$ without needing to use the change of variables formulae from Section 1.1.4 to explicitly calculate the probability P_y (or density p_y) and with a relatively weak condition of measurability on ϕ.

Closely related to the expected value are the *variance* and *covariance* of a random variable. The variance of a random variable x is defined

$$\mathbb{V}[\mathsf{x}] = \mathbb{E}\left[(\mathsf{x} - \mathbb{E}[\mathsf{x}])^2\right] = \mathbb{E}\left[\mathsf{x}^2\right] - \mathbb{E}[\mathsf{x}]^2. \tag{1.28}$$

For a pair of random variables x and y their covariance is defined

$$\mathbb{C}[\mathsf{x}, \mathsf{y}] = \mathbb{E}[(\mathsf{x} - \mathbb{E}[\mathsf{x}])(\mathsf{y} - \mathbb{E}[\mathsf{y}])]. \tag{1.29}$$

1.1.6 Conditional expectations and densities

A related concept, and one which will be key to the discussion of inference, is conditional expectation. Let $(S, \mathcal{E}, \mathsf{P})$ be a probability space, $(X, \mathcal{F})$ and $(Y, \mathcal{G})$ two measurable spaces and $\mathsf{x} : S \to X$ and $\mathsf{y} : S \to Y$ two random variables. Then the *conditional expectation of* x *given* y, is defined as a measurable function $\mathbb{E}[\mathsf{x} \,|\, \mathsf{y}] : Y \to X$ satisfying

$$\int_{\mathsf{y}^{-1}(B)} \mathsf{x}(s)\, \mathsf{P}(\mathrm{d}s) = \int_B \mathbb{E}[\mathsf{x} \,|\, \mathsf{y}](y)\, \mathsf{P}_{\mathsf{y}}(\mathrm{d}y) \quad \forall B \in \mathcal{G}. \tag{1.30}$$

$\mathbb{E}[\mathsf{x} \,|\, \mathsf{y}]$ is guaranteed to be uniquely defined almost everywhere in Y by (1.30), i.e. up to P_{y}-null sets. As a particular case where $B = Y$ we recover what is sometimes termed the *Law of total expectation*

$$\int_S \mathsf{x}\, \mathrm{d}\mathsf{P} = \int_S \mathbb{E}[\mathsf{x} \,|\, \mathsf{y}] \circ \mathsf{y}\, \mathrm{d}\mathsf{P} \implies \mathbb{E}[\mathsf{x}] = \mathbb{E}[\mathbb{E}[\mathsf{x} \,|\, \mathsf{y}] \circ \mathsf{y}]. \tag{1.31}$$

We will also use a more standard notation for the conditional expectation evaluated at point $\mathbb{E}[\mathsf{x} \,|\, \mathsf{y} = y] \equiv \mathbb{E}[\mathsf{x} \,|\, \mathsf{y}](y)$ but use the latter in this section to stress its definition as a measurable function.

Conditional expectation can be used to define the *regular conditional distribution* of a random variable conditioned on another random variable taking a particular value

$$\mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \,|\, y) = \mathbb{E}[\mathbb{1}_A \circ \mathsf{x} \,|\, \mathsf{y}](y) \quad \forall y \in Y,\, A \in \mathcal{F}. \tag{1.32}$$

We have reused our notation for conditional probability of random variables from Section 1.1.2 here, however it should be clear from whether the value conditioned on is a point or a set which is being referred to. A regular conditional distribution $\mathsf{P}_{\mathsf{x}|\mathsf{y}}(\cdot \,|\, y)$ defines a valid probability measure on $(X, \mathcal{F})$ for P_{y}-almost all y and using (1.30) we have

$$\mathsf{P}_{\mathsf{x},\mathsf{y}}(A, B) = \int_B \mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \,|\, y)\, \mathsf{P}_{\mathsf{y}}(\mathrm{d}y) \qquad \forall A \in \mathcal{F},\, B \in \mathcal{G}. \tag{1.33}$$

We can use this relationship to also motivate a definition of conditional density. We require that a joint density $\mathsf{p}_{\mathsf{x},\mathsf{y}} = \frac{\mathrm{d}\mathsf{P}_{\mathsf{x},\mathsf{y}}}{\mathrm{d}(\mu_{\mathsf{x}} \times \mu_{\mathsf{y}})}$ exists and has marginal density $\mathsf{p}_{\mathsf{y}} = \frac{\mathrm{d}\mathsf{P}_{\mathsf{y}}}{\mathrm{d}\mu_{\mathsf{y}}}$. Then for all $A \in \mathcal{F}, B \in \mathcal{G}$

$$\int_B \mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \,|\, y)\, \mathsf{P}_{\mathsf{y}}(\mathrm{d}y) = \int_B \int_A \mathsf{p}_{\mathsf{x},\mathsf{y}}(x, y)\, \mu_{\mathsf{x}}(\mathrm{d}x)\, \mu_{\mathsf{y}}(\mathrm{d}y) \tag{1.34}$$

If we define the *conditional density* $\mathsf{p}_{\mathsf{x}|\mathsf{y}}$ as

$$\mathsf{p}_{\mathsf{x}|\mathsf{y}}(x \mid y) = \begin{cases} \frac{\mathsf{p}_{\mathsf{x},\mathsf{y}}(x,y)}{\mathsf{p}_{\mathsf{y}}(y)} & \forall x \in X,\ y \in Y : \mathsf{p}_{\mathsf{y}}(y) > 0 \\ 0 & \forall x \in X,\ y \in Y : \mathsf{p}_{\mathsf{y}}(y) = 0 \end{cases} \tag{1.35}$$

then subsituting this definition in to (1.34) we have

$$\int_B \mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \mid y)\, \mathsf{P}_{\mathsf{y}}(\mathrm{d}y) = \int_B \int_A \mathsf{p}_{\mathsf{x}|\mathsf{y}}(x \mid y)\, \mu_{\mathsf{x}}(\mathrm{d}x)\, \mathsf{P}_{\mathsf{y}}(\mathrm{d}y). \tag{1.36}$$

Therefore $\mathsf{p}_{\mathsf{x}|\mathsf{y}}$ is the density of the regular conditional distribution $\mathsf{P}_{\mathsf{x}|\mathsf{y}}$. We also have that if $\mathsf{p}_{\mathsf{x},\mathsf{y}}$ and p_{y} can be defined that

$$\mathbb{E}[\mathsf{x} \mid \mathsf{y}](y) = \int_X x\, \mathsf{p}_{\mathsf{x}|\mathsf{y}}(x \mid y)\, \mu_{\mathsf{x}}(\mathrm{d}x) \quad \forall y \in Y : \mathsf{p}_{\mathsf{y}}(y) > 0. \tag{1.37}$$

Note that the initial definition of conditional expectation in (1.30) was not dependent on a joint density $\mathsf{p}_{\mathsf{x},\mathsf{y}}$ being defined.

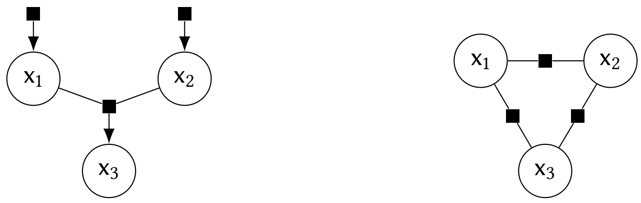
1.2 GRAPHICAL MODELS

Graphical models = statistics × graph theory × computer science
—Zoubin Ghahramani

When working with probabilistic models involving large numbers of random variables, it will often be the case that not all the variables are jointly dependent on each other but that instead there are more local relationships between them. Graphical models, which use graphs to describe the dependencies between random variables, are a useful framework for visualising the structure of complex models.

Several graphical formalisms for representing dependency structure in probabilistic models have been proposed, with *directed graphical models* [18] (also known as *Bayesian networks*) and *undirected graphical models* [14] (also known as *Markov random fields*) both common in practice and each offering a more natural representation for certain model classes. In this thesis we will instead use *factor graphs* [10, 11] which combine the representational abilities of both directed and undirected graphical models while also offering a richer framework for representing fine-grained information about model structure.

Factor graphs are bipartite graphs consisting of two node types: *variable nodes*, displayed as labelled circles and representing individual (potentially non-scalar) random variables in a probabilistic model, and *factor nodes*, displayed as filled squares and representing individual



(a) Example directed factor graph.

(b) Example undirected factor graph.

Figure 1.1: Examples of simple directed and undirected factor graphs. Square black nodes correspond to individual factors depending on the connected variables represented by circular nodes in the joint density.

factors in the joint density across the random variables in the model. Edges between nodes in a factor graph are always between nodes of disparate types i.e. between factor and variable nodes, but never between factor and factor or variable and variable nodes.

Factors may be either directed or undirected. Undirected factors, denoted by factor nodes in which all edges connecting to variable nodes are undirected, correspond to a factor in the joint density which depends on all of the variables with nodes connected to the factor, but without any requirement that the factor corresponds to a conditional or marginal probability density.

Directed factors, factor nodes in which at least one edge from the factor node to a variable node is directed, correspond to a conditional density on the variables pointed to by directed edges given the values of the variables connected to the the factor node by undirected edges. If there are no undirected edges then the factor instead corresponds to a marginal density. Graphs with directed factors must not contain any directed cycles, i.e. connected loops of edges in which one of every pair of edges connected to a factor on the loop is directed and all of the directed edges point in the same sense around the loop.

Figure 1.1a shows a simple example of fully directed factor graph for three random variables. The graph implies that the joint density for the model can be factorised as

$$\mathrm{p}_{\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3}(x_1, x_2, x_3) = \mathrm{p}_{\mathrm{x}_3|\mathrm{x}_1,\mathrm{x}_2}(x_3 \mid x_1, x_2)\, \mathrm{p}_{\mathrm{x}_1}(x_1)\, \mathrm{p}_{\mathrm{x}_2}(x_2). \qquad (1.38)$$

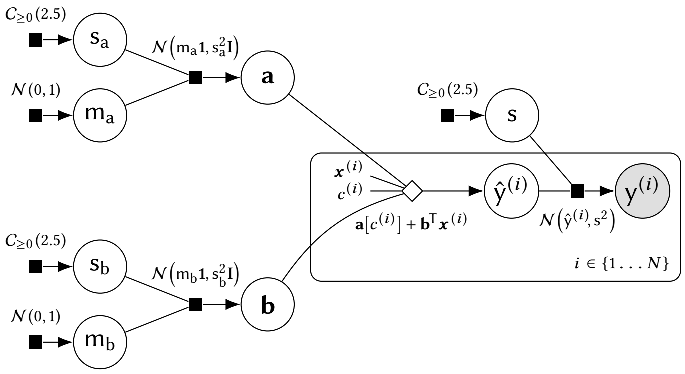


Figure 1.2: Hierarchical linear regression model factor graph showing examples of extended factor graph notation.

Figure 1.1b shows a fully undirected factor graph on three random variables. If $\psi_{i,j}$ denotes the unnormalised density factor on the pair $(\mathrm{x}_i, \mathrm{x}_j)$ then the graph implies the joint density can be factorised as

$$\mathrm{p}_{\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3}(x_1, x_2, x_3) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{1,3}(x_1, x_3)\psi_{2,3}(x_2, x_3) \tag{1.39}$$

with Z a normalising constant such that the density integrates to one.

Figure 1.2 shows examples of some additional useful factor graph notation we will use in this thesis. We use as an example a factor graph corresponding to a hierarchical linear regression model which will be discussed in Chapter 5.

It will often be useful to be able to explicitly represent deterministic functions applied to the random variables in a factor graph. For this purpose we introduce an additional node type denoted by an unfilled diamond (◇). The semantics of this node type are similar to standard directed factor nodes. Variables acting as inputs to the function are connected to the node by undirected edges and the variable corresponding to the function output indicated by a directed edge from the node to the relevant variable. Like standard factor nodes, the deterministic factor nodes only ever connect to variable nodes. The operations performed by the function on the inputs will usually be included as a label adjacent to the node as illustrated by the example in Figure 1.2. A deterministic factor node can informally[2] be considered equivalent to a directed factor node with a degenerate Dirac delta conditional density on the

[2] A Dirac delta is not strictly a density as it is not the Radon–Nikodyn derivative of an absolutely continuous measure, however informally we treat is as the density of a singular Dirac measure with respect to the Lebesgue measure $\int f(x)\,\delta(\mathrm{d}x) \simeq \int f(x)\delta(x)\,\mathrm{d}x$.

output variable which concentrates all the probability mass at the output of the function applied to the inputs variables.

The deterministic node notation allows generative models consisting of complex compositions of deterministic functions and probabilistic sampling operations to be represented in a unified framework. Subgraphs of a directed factor graph consisting entirely of determinstic nodes can be viewed as *computation graphs*, a graphical formalism typically used in numerical computing frameworks to support efficient *automatic differentiation* algorithms. We exploit this idea in later in the thesis to allow propagation of derivatives through complex probabilistic models and make extensive use of automatic differentiation implementations in frameworks such as *Theano* [19] in numerical experiments. In Appendix B we provide a short review of the basic concepts of computation graphs and automatic differentation and a discussion of their links to directed factor graphs.

In some cases constant values used in a model will be included in a factor graph as plain nodes indicated only by a label. The $\boldsymbol{x}^{(i)}$ and $c^{(i)}$ nodes in Figure 1.2 are an example of this notation.

A commonly used convention in factor graphs (and other graphical models) is *plate notation* [4], with an example of a plate shown by the rounded rectangle bounding some of the nodes in Figure 1.2. Plates are used to indicate a subgraph in the model which is replicated multiple times (with the replications being indexed over a set which is typically indicated in the lower right corner of the plate as in Figure 1.2). The subgraph entirely contained on the plate is assumed to be replicated the relevant number of times, with any edges crossing into the plate from variable nodes outside of the plate being repeated once for each subgraph replication.

Each of the factors in Figure 1.2 is labelled with a shorthand for a probability density function corresponding to the conditional or marginal density factor associated with the node. Definitions for the shorthand notations that are used for densities in this thesis are given in Appendix A. The dependence of the factors on the value of the random variable the density is defined on is omitted in the labels for brevity.

A final additional notation used in Figure 1.2 is the use of a shaded variable node (corresponding to $\mathrm{y}^{(i)}$) to indicate a random variable corresponding to an observed quantity in the model.

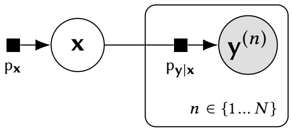

$\mathrm{p}_{\mathbf{x}}$
$\mathbf{x}$
$\mathrm{p}_{\mathbf{y}|\mathbf{x}}$
$\mathbf{y}^{(n)}$
$n \in \{1 \dots N\}$

Figure 1.3: Factor graph of N observations $\mathbf{y}^{(n)}$ independently and identically distributed according to a distribution with parameters $\mathbf{x}$.

1.3 INFERENCE

Having now introduced the tools and notation we use to define probabilistic models, we will now describe the inference problems we consider approximate approaches to solving in this thesis. We begin with a overview of *Bayesian inference*.

The starting point for any inference problem is to define a model specifying proposed relationships between the observed data and unknown quantities to be inferred. The model codifies the assumptions we make about the problem and any prior beliefs we have. In virtually all real inference problems the model will be a simplified description of a much more complex underlying process, usually motivated by prior empirical observations that the behaviour proposed by the model is a reasonable description of reality. For now we will consider the model as a singular fixed object we will perform inference with. We consider probabilistic model comparison in a subsequent subsection.

You cannot do inference without making assumptions —David Mackay

Amongst the simplest, but also most common, modelling assumptions made is that the observed data values are *independently and identically distributed* (IID) according to a parametric probability distribution. If we denote the collection of N observed variables $\{\mathbf{y}^{(n)}\}_{n=1}^{N}$ then we assume that each is independently generated from a distribution $\mathsf{P}_{\mathbf{y}^{(n)}|\mathbf{x}} = \mathsf{P}_{\mathbf{y}|\mathbf{x}}\ \forall n \in \{1 \dots N\}$ with density $\mathsf{p}_{\mathbf{y}|\mathbf{x}} = \frac{\partial \mathsf{P}_{\mathbf{y}|\mathbf{x}}}{\partial \mu_{\mathbf{y}}}$ and governed by a set of unknown parameters $\mathbf{x} \in X$.

Any beliefs we have about the plausible values for the parameters prior to observing data are integrated into the model by choosing an appropriate, typically parametric, marginal distribution $\mathsf{P}_{\mathbf{x}}$, with this distribution, and the corresponding density $\mathsf{p}_{\mathbf{x}} = \frac{\partial \mathsf{P}_{\mathbf{x}}}{\partial \mu_{\mathbf{x}}}$, referred to as the *prior*. The joint distribution on the model variables then factorises as

$$\mathsf{p}_{\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)},\mathbf{x}}(\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)}, \boldsymbol{x}) = \prod_{n=1}^{N} \mathsf{p}_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}^{(n)} \mid \boldsymbol{x})\, \mathsf{p}_{\mathbf{x}}(\boldsymbol{x}) \qquad (1.40)$$

with this structure illustrated as a directed factor graph in Figure 1.3. In analogy to the naming of the prior, the conditional distribution on the unknown model parameters after conditioning on observed data values is termed the *posterior* and from the definition of conditional density (1.35) we can express its density as

$$\mathsf{p}_{\mathbf{x}\,|\,\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}(\boldsymbol{x} \mid \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)}) = \frac{\prod_{n=1}^{N} \mathsf{p}_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}^{(n)} \mid \boldsymbol{x})\, \mathsf{p}_{\mathbf{x}}(\boldsymbol{x})}{\mathsf{p}_{\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}(\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)})}. \tag{1.41}$$

Bayesian inference is named after Thomas Bayes, an 18th century Presbyterian minister, who proved a special case of what is now termed Bayes' theorem. Pierre-Simon Laplace later independently derived a more general statement of Bayes' theorem closer to the modern form.

This expression relating the posterior on the unknown parameters to the prior distribution and model of the observations is an example of *Bayes' theorem*. Typically the product of the conditional densities $\mathsf{p}_{\mathbf{y}|\mathbf{x}}$ is termed the *likelihood* and considered a function of the value $\boldsymbol{x}$ of the unknown parameters $\mathbf{x}$, with the observed data values $\{\boldsymbol{y}^{(n)}\}_{n=1}^{N}$ fixed. The denominator of the right-hand side (1.41), the marginal density on the observed variables, can be written as a integral marginalising out the parameters from the joint density

$$\mathsf{p}_{\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}(\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)}) = \int_X \prod_{n=1}^{N} \mathsf{p}_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}^{(n)} \mid \boldsymbol{x})\, \mathsf{p}_{\mathbf{x}}(\boldsymbol{x})\, \mu_{\mathbf{x}}(\mathrm{d}\boldsymbol{x}). \tag{1.42}$$

A conditional density $\mathsf{p}_{\mathbf{u}|\mathbf{v}}$ *is from the exponential family if it can be written as* $\mathsf{p}_{\mathbf{u}|\mathbf{v}}(\boldsymbol{u} \mid \boldsymbol{v}) = \frac{h(\boldsymbol{u})\exp(\boldsymbol{\eta}(\boldsymbol{v})^{\mathsf{T}}\boldsymbol{t}(\boldsymbol{u}))}{z(\boldsymbol{v})}$, *with* $\boldsymbol{\eta}(\boldsymbol{v})$ *termed the* natural parameters *and* $\boldsymbol{t}(\boldsymbol{u})$ *termed the* sufficient statistics.

This term is often described as the *marginal likelihood* or the *model evidence*. Generally this integral will not have an analytic solution though there are exceptions to this in a few special cases. For example if the densities $\mathsf{p}_{\mathbf{y}|\mathbf{x}}$ and $\mathsf{p}_{\mathbf{x}}$ are both of *exponential family distributions* and form a conjugate pair such that the posterior density is in the same family as the prior density then (1.42) will have a closed-form solution. For models in which the parameters are discrete the integral in (1.42) corresponds to a summation and so is in theory exactly solvable, though if the total number of possible configurations of the parameters is very large this summation can be infeasible to compute in practice. If the parameters are instead real-valued but of a low-dimensionality it may be possible to use numerical quadrature methods [7] to compute the integral in (1.42) to a reasonable accuracy. Quadrature methods involve evaluating the integrand across a grid of points and then computing a weighted sum of these values. For a fixed grid resolution however the cost of quadrature scales exponentially with the dimension of the space being integrated over - if G points are used per dimension, for a D dimensional parameter space evaluating (1.42) would require summing the joint density over G^D parameter values.

For real-valued parameter spaces of a more than ~ 10 dimensions[3] evaluating the model evidence term (1.42) is therefore typically computationally intractable. We can therefore often only evalulate the posterior density (1.41) up to an unknown constant. The posterior density itself is usually not of direct interest as it is only a proxy for describing the posterior distribution and is dependent on the particular model parameterisation chosen. However most quantities of interest from an inference perspective involve integrating functions against the posterior distribution and as with the model evidence these integrals will typically be intractable to compute exactly.

For example under an IID assumption the density of the *predictive distribution* of a new data point $\mathbf{y}^*$ given the previously observed data is formed by integrating $\mathsf{p}_{\mathsf{y}|\mathsf{x}}$ against the posterior distribution

$$
\begin{aligned}
&\mathsf{p}_{\mathsf{y}^*|\mathsf{y}^{(1)},\dots,\mathsf{y}^{(N)}}(\boldsymbol{y}^* \,|\, \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)}) \\
&\quad= \int_X \mathsf{p}_{\mathsf{y}|\mathsf{x}}(\boldsymbol{y}^* \,|\, \boldsymbol{x})\, \mathsf{p}_{\mathsf{x}|\mathsf{y}^{(1)},\dots,\mathsf{y}^{(N)}}(\boldsymbol{x} \,|\, \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)})\, \mu_{\mathsf{x}}(\mathrm{d}\boldsymbol{x}) \qquad (1.43) \\
&\quad= \mathbb{E}\left[\mathsf{p}_{\mathsf{y}|\mathsf{x}}(\boldsymbol{y}^* \,|\, \mathsf{x}) \,|\, \mathsf{y}^{(1)} = \boldsymbol{y}^{(1)}, \dots, \mathsf{y}^{(N)} = \boldsymbol{y}^{(N)}\right].
\end{aligned}
$$

If we wish to for example minimise the expected prediction error under some loss function this will involve integrating against this predictive distribution and so as a sub-task integrating against the posterior distribution on the model parameters. Similarly evaluating statistics of the unknown parameters under the posterior such as their mean or covariance corresponds to computing conditional expectations. In general any inferential output which takes in to account all of the information available from the posterior distribution will involve integrating against the posterior and so the computation of integrals is the key computational task in inference.

As exact evaluation of the integrals of interest is usually intractable we must instead resort to *approximate inference* methods which tradeoff an introduction of some level of approximation for an increase in computational tractability.

[3] The `C`-based implementation by Steven G. Johnson of an adaptive multi-dimensional quadrature algorithm [2] available at `https://github.com/stevengj/cubature` recommends using the package for integrals of up to around $D = 7$. Running a provided test cases for the integral of a Gaussian density across a D-dimensional space with a target error tolerance of 10^{-5} took around 2.5 seconds for $D = 5$, 50 seconds for $D = 6$ and 17 minutes for $D = 7$ on one core of a desktop CPU. Extrapolating the ~ 20-fold increase in run time per dimension, for $D = 10$ the run-time would be around 100 days.

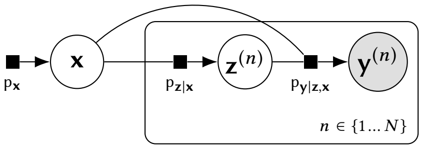


Figure 1.4: Factor graph of a simple hierarchical latent variable model with N observed variables $\mathbf{y}^{(n)}$ each associated with a local latent variable $\mathbf{z}^{(n)}$, with both observed and latent variables dependent on a set of global latent variables (parameters) $\mathbf{x}$.

The IID assumption is widely made in inference problems and although it will not be always be entirely valid in practice, it will often be a reasonable approximation. For real-valued parameter spaces X and densities $\mathsf{p_{y|x}}$ and $\mathsf{p_x}$ meeting certain regularity conditions, if an IID assumption is valid then the posterior distribution will asympotically tend to a multivariate normal distribution as the number of data points N tends to infinity [12]. For inference in models of large IID datasets where the conditions for asymptotic normality are met, while the dimensionality of the parameter space will often still require the use of approximate inference methods, the close to normal geometry of the posterior distribution will typically mean even relatively simple approximate inference methods can achieve good results.

In this thesis we will primarily be concerned with methods for performing inference in models which do not fit into this mould. In the following subsections we discuss some specific issues that can prove challenging to standard approximate inference approaches and which the methods contributed in this thesis are intended to help address.

1.3.1 Hierarchical models

In the preceding discussion of inference in a model of a IID dataset, it was assumed that the only unknown variables in the model were a set of parameters $\mathbf{x}$, the quantity of which did not depend on the number of data points N. This structure can be overly restrictive with it common that the process being modelled includes unknown quantities associated with each observed variable. Models will therefore often include local (per data point) latent variables in addition to a set of global latent variables (or parameters). This grouping structure in the observed and unobserved variables in a model can extend to multiple levels and such models often are termed *hierarchical* or *multilevel* models.

A simple example of a hierarchical model is shown as a factor graph in Figure 1.4. As in the factor graph in Figure 1.3 we assume there are N observed variables $\{\mathbf{y}^{(n)}\}_{n=1}^{N}$ and a vector of global latent variables $\mathbf{x}$. We further define N local latent variables $\{\mathbf{z}^{(n)}\}_{n=1}^{N}$ paired with each observed variable. In Figure 1.4 we assume the local latent and observed variables are conditionally independent given the global latent variables. More complex structures are also common - for example dynamical state space models for time series data assume dependencies between the latent variables corresponding to adjacent time points.

Although powerful, the introduction of local latent variables in to models can significantly increase the complexity of inference. At a basic level, as the number of unobserved variables is now dependent on the data set size, the total dimensionality of the space which needs to be integrated over when performing inference will typically be much higher than for models with a fixed number of parameters. This means the need for inference methods which scale well with dimensionality is even more essential. The growth of the the number of unobserved variables with the data set size N will typically also mean that we can no longer expect asymptotic normality of the full posterior. Typically the posterior distribution on the local and global latent variables will have a complex geometry, with strong dependencies between the global and local latent variables that can limit the performance of many standard approximate inference approaches [3].

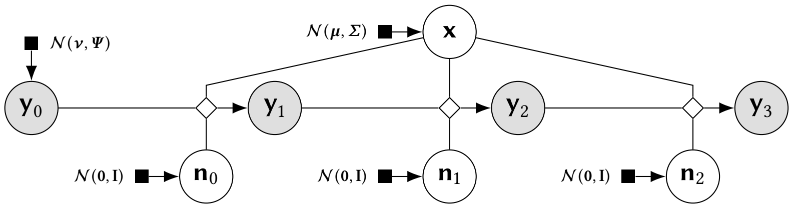
In some cases the posterior distributions of the local latent variables associated with the observed data will not be of direct interest to the downstream task. For example the conditional independence structure in Figure 1.4 means that the predictive distribution on a new unseen datapoint $\mathbf{y}^*$ given the observed data has density

$$\begin{aligned}
&\mathsf{p}_{\mathbf{y}^* | \mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}(\boldsymbol{y}^* \,|\, \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)}) \\
&\quad = \int_Z \int_X \mathsf{p}_{\mathbf{y}|\mathbf{x},\mathbf{z}}(\boldsymbol{y}^* \,|\, \boldsymbol{x}, z)\, \mathsf{p}_{\mathbf{z}|\mathbf{x}}(z \,|\, \boldsymbol{x}) \\
&\qquad\qquad \mathsf{p}_{\mathbf{x}|\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}(\boldsymbol{x} \,|\, \boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(N)})\, \mu_{\mathbf{x}}(\mathrm{d}\boldsymbol{x})\, \mu_{\mathbf{z}}(\mathrm{d}z).
\end{aligned} \tag{1.44}$$

Predictions under the model will therefore not depend on the values of the local latent variables $\{\mathbf{z}^{(n)}\}_{n=1}^{N}$, and so ideally we would marginalise out these variables from the full posterior distribution on all unobserved variables $\mathsf{P}_{\mathbf{z}^{(1)},\dots,\mathbf{z}^{(N)},\mathbf{x}|\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}$ to obtain the posterior dis-

```
$\boldsymbol{y}_0 \sim \mathcal{N}(\cdot \,|\, \boldsymbol{\nu}, \boldsymbol{\Psi})$
$\boldsymbol{x} \sim \mathcal{N}(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
for $t \in \{1 \dots T\}$ do
    $\boldsymbol{n}_{t-1} \sim \mathcal{N}(\cdot \,|\, \mathbf{0}, \mathbf{I})$
    $\boldsymbol{y}_t \leftarrow \boldsymbol{y}_{t-1} + h\boldsymbol{m}(\boldsymbol{y}_{t-1}, \boldsymbol{x}) - \frac{h}{2}\boldsymbol{s}(\boldsymbol{y}_{t-1}, \boldsymbol{x}) \odot \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{y}}(\boldsymbol{y}_{t-1}, \boldsymbol{x})$
    $\boldsymbol{y}_t \leftarrow \boldsymbol{y}_t + \sqrt{h}\,\boldsymbol{s}(\boldsymbol{y}_{t-1}, \boldsymbol{z}) \odot \boldsymbol{n}_{t-1} + \frac{h}{2}\boldsymbol{s}(\boldsymbol{y}_{t-1}, \boldsymbol{x}) \odot \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{y}}(\boldsymbol{y}_{t-1}, \boldsymbol{x}) \odot \boldsymbol{n}_{t-1}^2$
```

(a) Pseudo-code for Milstein method integration of SDE model.



(b) Directed factor graph of 3 time steps of SDE simulation.

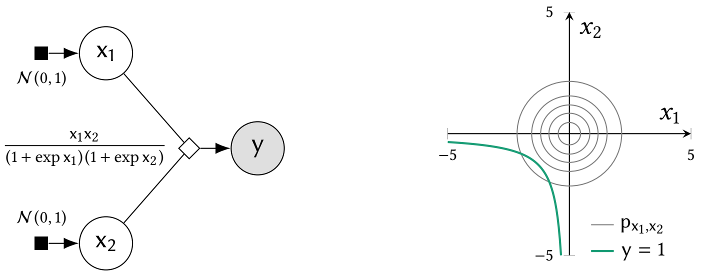
Figure 1.5: Example of a simulator model corresponding to Milstein method integration of a set of SDEs, $\mathrm{d}\mathbf{y}(t) = \boldsymbol{m}(\mathbf{y}(t), \mathbf{x})\,\mathrm{d}t + \boldsymbol{s}(\mathbf{y}(t), \mathbf{x})\,\mathrm{d}\mathbf{n}(t)$, specified as pseudo-code in (a) and a directed factor graph in (b). The dynamics of model are governed by parameters $\mathbf{x}$. In the pseudo-code the notation $\sim$ followed by a distribution shorthand represents generating a value from the associated distribution.

tribution on just the global latent variables $\mathsf{P}_{\mathbf{x}|\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}$. The distribution $\mathsf{P}_{\mathbf{x}|\mathbf{y}^{(1)},\dots,\mathbf{y}^{(N)}}$ is defined on a much lower dimensional space and will often have a simpler geometry which makes it more amenable to approximate inference methods, however generally the marginalisation over the local latent variables will not be analytically tractable. We can in some cases however approximately marginalise out the local latent variables - we discuss methods based on this idea in Chapter 3.

1.3.2 Simulator models

The probabilistic models considered so far have been defined by explicitly specifying a density over the all the variables in the model, for example via a factor graph. Rather than defining the density on the variables in a model an alternative approach is for a process for generating values for the variables in a model to be specified procedurally in code, with the resulting joint density on the model variables then only implicitly defined. Such models are sometimes termed *simulator* or *implicit* models [8].

A common setting in which such models occur is the simulation of a mechanistic model of a physical process for example described by a set



(a) Directed factor graph for model with two latent variables (x_1, x_2), and an observed variable y.

(b) Plot of marginal density on latent variables (contours) and set of values for which $y = 1$ (green curve).

Figure 1.6: Simple example of an implicit probabilistic model where the observed variable is a non-bijective function of two latent variables.

of *stochastic differential equations* (SDEs). In implementations of such simulator models, the stochasticity in the model will be introduced via draws from a pseudo-random number generator. Given these random inputs, the output of the simulator is then calculated as a series of deterministic operations and so can be described by a computation graph. The overall composition of directed factor nodes specifying the generation of random inputs from known densities by the random number generator and computation graph describing the operations performed by the simulator code together therefore define a directed factor graph. An example of a simulator model corresponding to approximate integration of a set of SDEs using the Milstein method [17] is shown as both pseudo-code and a directed factor graph in Figure 1.5.

The main complicating factor in performing inference in simulator models is the unavailability of an explicit density function on the model variables which is a prerequisite for most approximate inference methods. Computing a density function on the unobserved variables to be inferred (for example parameters of the dynamics of a SDE model) and simulated observed variables that are conditioned on requires that all other random variables used in the model are marginalised over. In some cases this marginalisation may technically be possible to exactly solve and so a density function possible to compute in theory but the complexity of the model structure means that the density is unavailable in practice.

In many cases however the density function may not be exactly evaluable even in theory. A key difference of simulator models from the probabilistic models considered previously is that the observed variables in

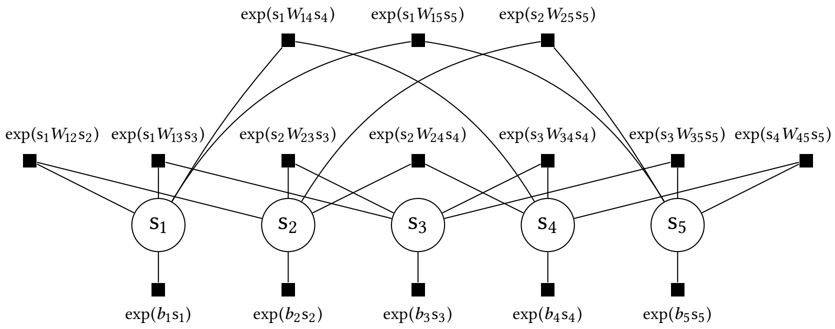


Figure 1.7: Five unit Boltzmann machine factor graph.

the model are defined via deterministic transformations of other random variables. Using our above intuition that any simulator model can be expressed as a directed factor graph with deterministic factor nodes, this means that the observed variables in the graph correspond to the outputs of deterministic factors rather than the more usual case of the observed variables being connected to probabilistic factors.

An illustration of such a case for a simple three variable model is shown in Figure 1.6. Here the observed variable y is a deterministic function of two latent (unobserved) variables x_1 and x_2. There is no analytic solution in terms of elementary functions for x_1 as a function of y and x_2 or for x_2 as a function of x_1 and y. This means the Dirac delta term corresponding to the deterministic factor cannot be integrated out. Due to the presence of the Dirac delta the joint density p_{y,x_1,x_2} is not well defined (the joint distribution P_{y,x_1,x_2} is not absolutely continuous with respect to the Lebesgue measure) which complicates evaluations of conditional expectations such as $\mathbb{E}[f(x_1, x_2) \,|\, y = 1]$. In particular the set of x_1 and x_2 values corresponding to solutions to $y = y$ for an particular y (illustrated for $y = 1$ as the green curve in Figure 1.6b) is an implicitly defined manifold (here a one-dimensional curve) in the x_1–x_2 space with zero Lebesgue measure, and the conditional distribution $P_{x_1,x_2 \,|\, y}$ has support only on this manifold. We explore methods for performing inference in implicit models in Chapter 4.

1.3.3 Undirected models

When introducing factor graphs we stated that factors can be both directed and undirected. In the preceding discussion we concentrated on directed models, both in the form of model explicitly specified via directed factor graphs as in the examples in Figure 1.3 and 1.4, and simulator models which as we argued in the previous subsection can be considered as implicitly defining a directed factor graph. A key defining feature of models corresponding to directed factor graphs is that they are natural descriptions of generative processes, with independent sampling from the joint distribution across model variables typically simple to perform via ancestral sampling (in the case of simulator models this being their defining feature).

Undirected models (which we will use here to mean models specified by factor graphs consisting solely of undirected factors) offer a complementary approach for defining a probabilistic model. Each undirected factor node is associated with a non-negative function defining a factor in the joint density across all model variables. Unlike a directed factor, this function does not correspond to a conditional or marginal density. Instead it describes a more general notion of 'compatibility' between the values of sets of variables in the model, defining a series of soft constraints as to which joint configurations are plausible (corresponding to a high value for the factor) or implausible (corresponding to a low model). This makes undirected models a natural representation for models of systems of mutually interacting components without a specific directivity in those interactions. For example they are commonly used in models of images to represent dependencies between pixel values, to model networks of stochastically spiking neurons in the brain and models of magnetic interactions in particle lattices. Unlike directed models, generating samples from the joint distribution on variables in an undirected model is typically a non-trivial task, with no general equivalent to ancestral sampling.

A particularly common form of undirected model is the *Boltzmann machine* [1] also known as a *pairwise binary Markov random field* [14] or in statistical physics settings an *Ising spin model* [13]. A Boltzmann machine consists of a set of binary random variables $\mathbf{s} = [\mathrm{s}_1 \ \cdots \ \mathrm{s}_D]^\mathsf{T}$; these are typically chosen to take values in $U = \{0, 1\}^D$ or $S = \{-1, +1\}^D$ - we will favour $S = \{-1, +1\}^D$. The joint distribution across the variables is parameterised by a symmetric weight matrix $\boldsymbol{W} \in \mathbb{R}^{D \times D}$ and a bias vector $\boldsymbol{b} \in \mathbb{R}^D$ and defined as

$$\mathrm{p}_{\mathbf{s}}(\boldsymbol{s}) = \frac{1}{Z} \exp\left(\frac{1}{2}\boldsymbol{s}^\mathsf{T}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^\mathsf{T}\boldsymbol{b}\right), \quad Z = \sum_{\boldsymbol{s} \in S} \exp\left(\frac{1}{2}\boldsymbol{s}^\mathsf{T}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^\mathsf{T}\boldsymbol{b}\right). \quad (1.45)$$

Evaluation of the normalising constant Z involves a summation over 2^D states and so for large D quickly become intractable to compute exactly. Evaluation of expectations with respect to the Boltzmann machine distribution also involves an exhaustive summation across S and so will also be intractable for high D.

If $\mathbf{s}_1$ and $\mathbf{s}_2$ are an arbitary partition of the variables in $\mathbf{s}$ then importantly the conditional distribution $\mathsf{P}_{\mathbf{s}_1|\mathbf{s}_2}$ will also be Boltzmann machine distributions. However unless the dimensionality of $\mathbf{s}_1$ is small enough that exhaustive summation over its possible states is feasible, then evaluating normalising constants of this conditional distribution and expectations with respect to it will also be intractable. Therefore inference in Boltzmann machines conditioned on observations of part of the state can be considered as a special case of computing expectations and the normalising constants of (non-conditioned) Boltzmann machine distributions, with the same challenges applying to both.

Figure 1.7 shows the factor graph for a Boltzmann machine distribution on five binary random variables $\{\mathrm{s}_i\}_{i=1}^{5}$. Each of the weights W_{ij} defines an undirected factor between a pair of variables $\mathrm{s}_i W_{ij} \mathrm{s}_j$. As the variables take on signed binary values, this factor is equal to W_{ij} when the variables are equal and so take the same sign and equal to $-W_{ij}$ when the variables take differing values. If W_{ij} is positive this factor therefore favours states where s_i and s_j are in the same configuration, while if W_{ij} is negative states with s_i and s_j in opposing configurations are preferred.

Boltzmann machine systems with a mixture of positive and negative weights will often be *frustrated* with no one global configuration which satisfies the preferences specified by each weight, and instead there being multiple states which each locally satisfy a subset of the soft constraints specified by the weights. This typically leads to a highly multi-modal distribution on the states of the system, with collections of nearby[4] states of high-probability separated sets of states with very low probability.

This multi-modality typically makes frustrated Boltzmann machines very challenging distributions to perform approximate inference with. In particular methods based on constructing Markov chains which explore the state of the system tend to converge very slowly as they will

[4] Nearby here being in terms of the Hamming distance between the binary states.

typically remain confined to a particular high-probability region of the state space for many iterations. In Chapter 5 we will consider methods for constructing Markov chains with improved exploration of challenging multi-modal target distributions, including methods for estimating expectations and normalising constants of frustrated Boltzmann machine distributions.

1.3.4 Model comparison

So far we have discussed inferring the unobserved variables in a single fixed model. An important second level of inference is comparing competing models for the same observed data. This can be treated consistently within the probabilistic framework we have discussed.

Given observed data, we would like to be able to make a judgement as to which of two (or more) proposed models better describes the data. To be useful this comparison must take into account the relative complexity of the models; a model with more free variables will generally be able to fit observed data more closely, however *Ockham's Razor* (and corresponding empirical evidence of the loss of predictive power of overly complex models) suggests we should prefer simpler models where possible. By marginalising over the free, unobserved variables in a model, probabilistic model comparison automatically embodies Ockham's Razor [16].

Ockham's Razor is a philoshopical principle, commonly attributed to the 14th century Franciscan friar William of Ockham, that states if there exist multiple explanations for observations, all else being equal we should prefer the simplest.

A concrete structure for model comparison is to assume that there are a finite set of M models, indexed by an *indicator* variable $\mathsf{m} \in \{1 \dots M\}$. All models share the same observed variables[5] $\mathbf{y}$, and there are a set of per model vectors of unobserved variables $\{\mathbf{x}_m\}_{m=1}^{M}$ which are assumed to be independent (before conditioning on observations). More complex structures could be assumed such as the models sharing a set of common unobserved variables, however we only consider the case of independent models here. The joint density on the observations, model indicator and latent variables is then assumed to factorise as

$$
\begin{aligned}
&\mathsf{p}_{\mathbf{y},\mathsf{m},\mathbf{x}_1,\dots,\mathbf{x}_M}(\boldsymbol{y}, m, \boldsymbol{x}_1, \dots, \boldsymbol{x}_M) = \\
&\qquad \mathsf{p}_{\mathbf{x}|\mathsf{m},\mathbf{z}_m}(\boldsymbol{y} \mid m, \boldsymbol{x}_m)\, \mathsf{p}_{\mathsf{m}}(m) \prod_{n=1}^{M} \mathsf{p}_{\mathbf{x}_n}(\boldsymbol{x}_n).
\end{aligned}
\tag{1.46}
$$

[5] For notational simplicity here we assume all observed variables have been concatenated in to one vector and similarly for the unobserved variables, with any internal model factorisation structure such as discussed in the preceding sections omitted.

The marginal density on the model indicator p_{m} represents our prior beliefs about the relative probabilities of the models before observing data. Importantly the value of the model indicator variable m selects the relevant per model conditional density on the observed variables given latent variables $\mathsf{p}_{\mathbf{y}|\mathsf{m},\mathbf{x}_m}$; this represents the assumption that conditioned on the model indicator assuming a particular model index m the observed variables are conditionally independent of the latent variables of all other models $\mathbf{y} \perp \{\mathbf{x}_n\}_{n \neq m} \mid \mathsf{m} = m$.

Given this computational set up, the task in model comparison is then to compute the relative probabilities of each of the models given observed data. These probabilities are given by

$$\mathsf{p}_{\mathsf{m}|\mathbf{y}}(m \mid \boldsymbol{y}) = \frac{\mathsf{p}_{\mathbf{y}|\mathsf{m}}(\boldsymbol{y} \mid m)\, \mathsf{p}_{\mathsf{m}}(m)}{\sum_{n=1}^{M} \left(\mathsf{p}_{\mathbf{y}|\mathsf{m}}(\boldsymbol{y} \mid n)\, \mathsf{p}_{\mathsf{m}}(n)\right)}, \tag{1.47}$$

which can be seen as a direct analogue to Bayes' theorem for the posterior density on unobserved random variables for a single model. The key quantities needed to evaluate the model posterior probabilities are the marginal densities $\mathsf{p}_{\mathbf{y}|\mathsf{m}}(\boldsymbol{y} \mid m)$ evaluated at the observed data. Computing these values requires marginalising out the unobserved variables from the per model joint densities $\mathsf{p}_{\mathbf{y},\mathbf{x}_m|\mathsf{m}}$

$$\mathsf{p}_{\mathbf{y}|\mathsf{m}}(\boldsymbol{y} \mid m) = \int_{X_m} \mathsf{p}_{\mathbf{y}|\mathsf{m},\mathbf{x}_m}(\boldsymbol{y} \mid m, \boldsymbol{x}) \mathsf{p}_{\mathbf{x}_m}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{1.48}$$

This value is equivalent to (1.42) for a single model, this explaining the naming of this term as the *model evidence*.

As described previously, evaluating the model evidence requires integrating across the space of all unobserved variables. The key computational challenge in being able to perform probabilistic model comparison with complex high dimensional models is therefore again being able to efficiently to compute integrals in high dimensional spaces. Unlike the integrals required for making predictions using a single model however, the model evidence integral cannot be naturally expressed as an expectation of a function with respect to the posterior distribution. This can complicate approximate computation of model evidence terms compared to other quantities involved in inference. In Chapter 5 we will consider extensions to a class of methods proposed for estimating model evidence terms.

1.4 SUMMARY

Probabilistic modelling offers a natural way to formalise our beliefs and assumptions about a problem and make inferences given those beliefs. Once a model has been defined the theoretical basis of the inference process is elegantly simple. Underlying this simplicity however are some very significant implementation challenges. The key computational task is the evaluation of integrals across high-dimensional spaces, which typically do not have closed form solutions and are intractable to compute using standard numerical integration approaches.

This intractability necessitates the use of approximate inference methods, the focus of this thesis. In particular we propose several novel extensions to MCMC methods, a class of approaches for drawing dependent samples from high-dimensional target distributions. In the next chapter we review the basic Monte Carlo method for integration and associated methods for generating and using independent pseudo-random variates to estimate the integrals. We then introduce the key Markov chain theory underlying MCMC methods and review some key existing MCMC algorithms. We will then conclude with an outline of the remainder of the thesis, in particular giving a a summary of the novel contributions made and how these relate to the specific inference problems discussed in the last section of this chapter.

BIBLIOGRAPHY

[1] David H Ackley, Geoffrey E Hinton and Terrence J Sejnowski. ‘A learning algorithm for Boltzmann machines’. In: *Cognitive science* 9.1 (1985), pp. 147–169.

[2] Jarle Berntsen, Terje O Espelid and Alan Genz. ‘An adaptive algorithm for the approximate calculation of multiple integrals’. In: *ACM Transactions on Mathematical Software (TOMS)* 17.4 (1991), pp. 437–451.

[3] Michael Betancourt and Mark Girolami. ‘Hamiltonian Monte Carlo for hierarchical models’. In: *Current trends in Bayesian methodology with applications* 79 (2015), p. 30.

[4] Wray L Buntine. ‘Operations for learning with graphical models’. In: *Journal of artificial intelligence research* (1994).

[5] Richard T Cox. ‘Probability, frequency and reasonable expectation’. In: *American Journal of Physics* 14.1 (1946), pp. 1–13. URL: http://dx.doi.org/10.1119/1.1990764.

[6] Richard T Cox. ‘The algebra of probable inference’. In: *American Journal of Physics* 31.1 (1963), pp. 66–67. URL: http://dx.doi.org/10.1119/1.1969248.

[7] Philip J. Davis and Philip Rabinowitz. *Numerical Integration*. Blaisdell Publishing Company, 1967.

[8] Peter J Diggle and Richard J Gratton. ‘Monte Carlo methods of inference for implicit statistical models’. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), pp. 193–227.

[9] Bruno de Finetti. ‘Foresight: its logical laws, its subjective sources’. In: *Studies in Subjective Probability*. Ed. by H. E. Kyburg. English translation of original 1937 French article *La Prévision: ses lois logiques, ses sources subjectives*. Springer, 1992, pp. 134–174.

[10] Brendan J Frey. ‘Extending factor graphs so as to unify directed and undirected graphical models’. In: *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. 2002, pp. 257–264. URL: https://arxiv.org/abs/1212.2486.

[11] Brendan J Frey, Frank R Kschischang, Hans-Andrea Loeliger and Niclas Wiberg. 'Factor graphs and algorithms'. In: *Proceedings of the 35th Annual Allerton Conference on Communication Control and Computing*. 1997.

[12] J.A. Hartigan. *Bayes theory*. Springer series in statistics. Springer-Verlag, 1983. ISBN: 9783540908838.

[13] Ernst Ising. 'Beitrag zur theorie des ferromagnetismus'. In: *Zeitschrift für Physik A Hadrons and Nuclei* 31.1 (1925), pp. 253–258.

[14] Ross Kindermann and Laurie Snell. *Markov random fields and their applications*. American Mathematical Society, 1980.

[15] Andreĭ Nikolaevich Kolmogorov. *Foundations of the Theory of Probability*. Ed. by Nathan Morrison. 2nd English Edition. English translation of original 1933 German monograph, *Grundbegriffe der Wahrscheinlichkeitrechnung*. Chelsea Publishing Company, 1956. URL: https://pdfs.semanticscholar.org/c3e1/51f71168a5f348bdebfde11752ca603fa6d0.pdf.

[16] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[17] GN Mil'shtejn. 'Approximate integration of stochastic differential equations'. In: *Theory of Probability & Its Applications* 19.3 (1975), pp. 557–562.

[18] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

[19] Theano development team. 'Theano: A Python framework for fast computation of mathematical expressions'. In: *arXiv e-prints* abs/1605.02688 (2016). URL: http://arxiv.org/abs/1605.02688.