AUXILIARY VARIABLE MARKOV CHAIN MONTE CARLO METHODS

MATTHEW MCKENZIE GRAHAM

Doctor of Philosophy
University of Edinburgh
2017

Matthew Mckenzie Graham: *Auxiliary variable Markov chain Monte Carlo methods,* 2017.
Submitted for the degree of Doctor of Philosophy, University of Edinburgh.

SUPERVISOR:
Dr. Amos J. Storkey

ABSTRACT

Inference, the process of drawing conclusions from evidence, is at the heart of the scientific method. Probability theory offers a consistent framework for performing inference in the presence of uncertainty, posing the inference problem as the task of computing expectations of functions of interest with respect to a probability distribution.

In complex models with a large number of unknown variables to be inferred, these expecations can be intractable to compute exactly. This has motivated the development of a large class of approximate inference methods which tradeoff a lack of exactness for greater computational tractability. Monte Carlo methods are one class of such techniques, with the integrals or summations across the whole state space approximated by summations over a finite number of randomly sampled points. This maps the inference problem to that of drawing samples from, often complex, high-dimesional, probability distributions.

LAY SUMMARY

A lay summary is intended to facilitate knowledge exchange, public engagement and outreach. It should be in simple, non-technical terms that are easily understandable by a lay audience, who may be non-professional, non-scientific and outside the research area.

Abstracts, particularly in science, engineering, medicine and veterinary medicine, may be highly technical or contain scientific language that is not easily understandable to readers outside the research area. Therefore, the lay summary is intended as supplementary to the abstract.

ACKNOWLEDGEMENTS

Put acknowledgments here.

Lorem ipsum at nusquam appellantur his, ut eos erant homero concluda turque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Errem omnium ea per, pro Markov chain Monte Carlo (MCMC) con populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, June 2017

Matthew Mckenzie Graham

CONTENTS

FRONT MATTER

Abstract 1
Lay summary 2
Acknowledgements 3
Declaration 4
List of Figures 7
List of Tables 7
List of Abbreviations 7

1 PROBABILISTIC INFERENCE 9
1.1 Probability theory 9
1.1.1 Random variables 10
1.1.2 Joint and conditional probability 10
1.1.3 Probability densities 12
1.1.4 Transforms of random variables 14
1.1.5 Expectations 17
1.1.6 Conditional expectations and densities 18
1.2 Graphical models 19
1.2.1 Directed and undirected graphical models 20
1.2.2 Factor graphs 22
1.2.3 Stochastic computation graphs 23
1.3 Inference 23
1.3.1 Posterior expectations 24
1.3.2 Model evidence 24

2 APPROXIMATE INFERENCE 25
2.1 Deterministic approaches 25
2.1.1 Laplace's method 25
2.1.2 Variational inference 25
2.1.3 Expectation propagation 25
2.2 Stochastic approaches 25
2.2.1 Monte Carlo method 25
2.2.2 Rejection sampling 25
2.2.3 Importance sampling 25
2.2.4 Markov chain Monte Carlo 25

3 MARKOV CHAIN MONTE CARLO 27
3.1 Metropolis–Hastings 27
3.2 Gibbs sampling 27
3.3 Slice sampling 27
3.4 Hamiltonian Monte Carlo 27

4 THERMODYNAMIC METHODS 29
4.1 Simulated tempering 29
4.2 Parallel tempering 29
4.3 Tempered transitions 29
4.4 Annealed importance sampling 29
4.5 Path sampling 29
4.6 Adiabatic Monte Carlo 29

LIST OF FIGURES

Figure 1.1 Directed and undirected graphical models. 20
Figure 1.2 Factor graph examples. 23
Figure 1.3 Hierarchical linear regression model factor graph. 23
Figure 1.4 Hierarchical linear regression model stochastic computation graph. 23

LIST OF TABLES

LIST OF ABBREVIATIONS

MCMC Markov chain Monte Carlo

wrt with respect to

1 PROBABILISTIC INFERENCE

Inference is the process of drawing conclusions from evidence. Much of our lives are spent making inferences about the world given our observations of it; in particular inference is a central aspect of the scientific process. Although 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 subject to uncertainty.

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

To make inferences under conditions of uncertainty, we must instead turn to probability theory. Probabilities offer a consistent framework for quantifying the uncertainty in our beliefs about the world and making inferences given these beliefs. The output of the inference process is itself probabilistic, reflecting that the conclusions we make given uncertain information will themselves be subject to uncertainty.

In this chapter we will first introduce the probability notation we will use in the rest of this work, and state some basic results which will be important in the later chapters. We will introduce graphical models as a compact way of visualising structure in probabilistic models. Finally we will give a concrete definition of the probabilistic inference tasks that the methods presented in the rest of this thesis are aimed at computing (approximate) solutions to, and motivate why such approximate computational methods are needed.

Probability theory is nothing but common sense reduced to calculation.
— Pierre-Simon Laplace

1.1 PROBABILITY THEORY

A *probability space* is defined as a triplet $(S, \mathcal{E}, \mathrm{P})$ 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 $\mathrm{P}(S) = 1$, which specifies the probabilities of events in $\mathcal{E}$.

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.

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 [] can be used to derive a measure-theoretic formulation of probability theory. The probability of an event $E \in \mathcal{E}$ is defined as the measure of that event $\mathsf{P}(E)$. Two events $A, B \in \mathcal{E}$ are said to be *independent* if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$.

The measure-theoretic approach has the advanatage of providing a unified treatment for describing probabilities on both finite and infinite sample spaces. Although alternative derivations of the laws of probability from different premises such as Cox's theorem [] have been proposed, modern extensions of this work result in a calculus of probabilities that is equivalent to Kolmogorov's [], with the differences mainly being in the philosophical interpretations of probabilities.

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 $\mathsf{x} : S \to X$ 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 will 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 we will term the resulting random variable a *random vector* and use the notation $\mathbf{x} : S \to X$. A final special case is when X is countable and $\mathcal{F}$ is the power set $\mathscr{P}(X)$ in which case we will refer to x as a *discrete random variable*.

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 $\mu_f(A) = \mu \circ f^{-1}(A) \; \forall A \in \mathcal{G}$.

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

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

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 $\mathsf{P}_\mathsf{x}(A)$.

1.1.2 Joint and conditional probability

Often 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

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

The joint probability is related to the probabilities P_x and P_y by

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

In this context P_x and P_y are referred to as *marginals* of the joint.

The two random variables are said to be independent if and only if

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

Also useful is the definition of *conditional probability*

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

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 [] and de Finetti [], conditional probabilities are instead viewed as a primitive.

Correspondingly, the conditional probabilities of random variables $\mathsf{P}_{\mathsf{x}|\mathsf{y}}$ and $\mathsf{P}_{\mathsf{x}|\mathsf{y}}$ can likewise be defined as satisfying

$$\begin{aligned}\mathsf{P}_{\mathsf{x},\mathsf{y}}(A,\, B) = \mathsf{P}_{\mathsf{x}|\mathsf{y}}(A \mid B)\, \mathsf{P}_\mathsf{y}(B) = \mathsf{P}_{\mathsf{y}|\mathsf{x}}(B \mid A)\, \mathsf{P}_\mathsf{x}(A) \\ \forall A \in \mathcal{F},\, B \in \mathcal{G} : \mathsf{P}_{\mathsf{x},\mathsf{y}}(A,\, B) \neq 0,\end{aligned} \tag{1.6}$$

which is sometimes referred to as the product rule.

An implication of (1.6) is what is often termed *Bayes' theorem*

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

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 will be discussed later in this chapter in section 1.2 .

1.1.3 Probability densities

So far we have ignored how the probability measure P is defined and by consequence the probability of a random variable.

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

The Radon–Nikodyn theorem [] guarantees that for a pair of $\sigma-$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

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

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

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 measure with respect to an appropriate base measure. It can also be shown that if $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ and g is a measurable function that

$$\int_X g \, \mathrm{d}\nu = \int_X g \, f \, \mathrm{d}\mu, \tag{1.9}$$

which we will use later when discussing calculation of expectations.

For real random variables, an appropriate base measure is usually the *Lebesgue measure*, λ, on $\mathbb{R}$. The probability P_x of a real random variable x can then be defined via a *probability density* $\mathsf{p}_\mathsf{x} : \mathbb{R} \to [0, \infty)$ by

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

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

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

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

For discrete random variables, an appropriate base measure is instead the *counting measure*, $\#$. The probability of a discrete random variable

is then defined via a probability density $\mathsf{p}_\mathsf{x} : X \to [0, 1]$ by

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

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 with respect to a base measure, this avoiding difficulties when defininining 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 $\mathsf{p}_{\mathsf{x},\mathsf{y}} : X \times Y \to [0, \infty)$ by

$$\mathsf{P}_{\mathsf{x},\mathsf{y}}(A, B) = \int_{A \times B} \mathsf{p}_{\mathsf{x},\mathsf{y}} \, \mathrm{d}(\mu_\mathsf{x} \times \mu_\mathsf{y}) \quad \forall A \in \mathcal{F}, \, B \in \mathcal{G}, \tag{1.13}$$

where $\mu_\mathsf{x} \times \mu_\mathsf{y}$ represents the product measure of two appropriate base measures μ_x and μ_y, e.g. $\mu_\mathsf{x} = \lambda$ and $\mu_\mathsf{y} = \#$ if x is a real random variable and y is a discrete random variable.

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 $(\mu_1 \times \mu_2)(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ $\forall A_1 \in \mathcal{F}_1, \, A_2 \in \mathcal{F}_2$.

When dealing with random variables, we will usually 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 $(S, \mathcal{E}, \mathsf{P})$ which the random variable(s) map from. In this case we will typically use the loose notation $\mathsf{x} \in X$ to mean a random variable x with co-domain X.

This less explicit but more succinct probability notation in terms of random variables and densities is common in the machine learning and computational statitistics literature and will generally be preferred to improve readability. The underlying measure-theoretic basis of these concepts will however be important for some of the upcoming definitions in this chapter and some of the derivations later in the thesis.

1.1.4 Transforms of random variables

It is common to define a random variable via a transform of another. 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 $\mathsf{y} = \phi \circ \mathsf{x}$ 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

$$\mathsf{P}_\mathsf{y}(A) = \mathsf{P}_\mathsf{x} \circ \phi^{-1}(A) = \mathsf{P}_\mathsf{x}(\{x \in X : \phi(x) \in A\}) \quad \forall A \in \mathcal{G}, \tag{1.14}$$

i.e. the probability of the event $\mathsf{y} \in A$ is equal to the probability of x being 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 probability P_x is defined by a probability density p_x with respect to a measure μ_x, we can also in some cases find a density p_y on the transformed variable $\mathsf{y} = \phi(\mathsf{x})$ with respect to a (potentially different) measure μ_y which can be used to calculate the probability P_y,

$$\mathsf{P}_\mathsf{y}(A) = \int_{\phi^{-1}(A)} \mathsf{p}_\mathsf{x} \, \mathrm{d}\mu_\mathsf{x} = \int_A \mathsf{p}_\mathsf{y} \, \mathrm{d}\mu_\mathsf{y} \quad \forall A \in \mathcal{G}. \tag{1.15}$$

For random variables with countable co-domains where the integral in (1.15) corresponds to a sum, a p_y satisfying (1.15) is simple to identify. If x is a discrete random variable with probability density p_x with respect to the counting measure, then $\mathsf{y} = \phi(\mathsf{x})$ will necessarily also be a discrete random variable. Applying (1.15) for $\mathsf{p}_\mathsf{x} = \frac{\mathrm{d}\mathsf{P}_\mathsf{x}}{\mathrm{d}\#}$ we have that

$$\begin{aligned}\int_{\phi^{-1}(A)} \mathsf{p}_\mathsf{x}(x) \, \mathrm{d}\#(x) &= \sum_{x \in \phi^{-1}(A)} \mathsf{p}_\mathsf{x}(x) = \sum_{y \in A} \sum_{x \in \phi^{-1}(y)} \mathsf{p}_\mathsf{x}(x) \\ &= \int_A \sum_{x \in \phi^{-1}(y)} \mathsf{p}_\mathsf{x}(x) \, \mathrm{d}\#(y) \quad \forall A \in \mathcal{G}. \end{aligned} \tag{1.16}$$

We can therefore define $\mathsf{p}_\mathsf{y} = \frac{\mathrm{d}\mathsf{P}_\mathsf{y}}{\mathrm{d}\#}$ in terms of p_x as

$$\mathsf{p}_\mathsf{y}(y) = \sum_{x \in \phi^{-1}(y)} \mathsf{p}_\mathsf{x}(x) \quad \forall y \in Y. \tag{1.17}$$

In the special case that ϕ is bijective we have that

$$\mathrm{p}_{\mathrm{y}}(y) = \mathrm{p}_{\mathrm{x}} \circ \phi^{-1}(y) \quad \forall y \in Y. \tag{1.18}$$

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 map ϕ. Let $X = \mathbb{R}^M$ and $Y = \mathbb{R}^N$ with $N \leq M$, $N, M \in \mathbb{N}$. We will need a result from geometric measure theory, the *co-area formula* []. Let g be an L^1 integrable function and $\boldsymbol{\phi} : X \to Y$ a Lipschitz map. Then the co-area formula states that

$$\int_X g(\boldsymbol{x})\, J_{\boldsymbol{\phi}}(\boldsymbol{x})\, \mathrm{d}\lambda^M(\boldsymbol{x}) = \int_Y \int_{\boldsymbol{\phi}^{-1}(\boldsymbol{y})} g(\boldsymbol{x})\, \mathrm{d}\mathcal{H}^{M-N}(\boldsymbol{x})\, \mathrm{d}\lambda^N(\boldsymbol{y}) \tag{1.19}$$

where $\mathcal{H}^D$ is the D-dimensional *Hausdorff measure* and $J_{\boldsymbol{\phi}} : X \to [0, \infty)$ is the *Jacobian determinant* defined as

$$J_{\boldsymbol{\phi}}(\boldsymbol{x}) = \left| \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}}^{\mathsf{T}} \right|^{\frac{1}{2}} \quad \forall \boldsymbol{x} \in X. \tag{1.20}$$

The D-dimensional Hausdorff measure $\mathcal{H}^D$ on $\mathbb{R}^N$ for $D \in \mathbb{N}$, $0 < D < N$ formalises a measure of the 'volume' of D-dimensional submanifolds of $\mathbb{R}^N$ - e.g. for $D = 1$ it corresponds to the length of a curve in $\mathbb{R}^N$. Additionally $\mathcal{H}^N = \lambda^N$ and $\mathcal{H}^0 = \#$.

Now let $\mathbf{x}$ be a random vector with co-domain the measurable space $(X, \mathscr{B}(\mathbb{R}^M))$ and define $\mathbf{y} = \boldsymbol{\phi} \circ \mathbf{x}$ as a random vector with co-domain the measurable space $(Y, \mathscr{B}(\mathbb{R}^N))$ with $\boldsymbol{\phi} : X \to Y$ a Lipschitz map as above. Let $Z = \left\{ \boldsymbol{x} \in X : J_{\boldsymbol{\phi}}(\boldsymbol{x}) = 0 \right\}$ and require that $\mathrm{P}_{\mathrm{x}}(Z) = 0$. Then for $A \in \mathscr{B}(\mathbb{R}^N)$ define an L^1 integrable function g as

$$g(\boldsymbol{x}) = \begin{cases} \mathbb{1}_A \circ \boldsymbol{\phi}(\boldsymbol{x})\, \mathrm{p}_{\mathbf{x}}(\boldsymbol{x})\, J_{\boldsymbol{\phi}}(\boldsymbol{x})^{-1} & \forall \boldsymbol{x} \in X \setminus Z \\ 0 & \forall \boldsymbol{x} \in Z \end{cases}. \tag{1.21}$$

Integrating $g(\boldsymbol{x})\, J_{\boldsymbol{\phi}}(\boldsymbol{x})$ over $\boldsymbol{x} \in X$ we have that

$$\begin{aligned} \int_X g(\boldsymbol{x})\, J_{\boldsymbol{\phi}}(\boldsymbol{x})\, \mathrm{d}\lambda^M(\boldsymbol{x}) &= \int_{X \setminus Z} \mathbb{1}_A \circ \boldsymbol{\phi}(\boldsymbol{x})\, \mathrm{p}_{\mathbf{x}}(\boldsymbol{x})\, \mathrm{d}\lambda^M(\boldsymbol{x}) && (1.22) \\ &= \int_X \mathbb{1}_A \circ \boldsymbol{\phi}(\boldsymbol{x})\, \mathrm{dP}_{\mathrm{x}}(\boldsymbol{x}) && (1.23) \\ &= \int_{\phi^{-1}(A)} \mathrm{dP}_{\mathrm{x}}(\boldsymbol{x}) = \mathrm{P}_{\mathrm{y}}(A). && (1.24) \end{aligned}$$

The equality between first and second lines comes from the requirement $\mathrm{P}_{\mathrm{x}}(Z) = 0$, with the Lebesgue integrals of a function over two

sets which differ by only a zero-measure set equal. Now applying the co-area formula (1.19) to the left-hand side gives

$$\int_Y \int_{\boldsymbol{\phi}^{-1}(\boldsymbol{y})} g(\boldsymbol{x})\, \mathrm{d}\mathcal{H}^{M-N}(\boldsymbol{x})\, \mathrm{d}\lambda^N(\boldsymbol{y}) = \mathrm{P}_{\mathsf{y}}(A). \tag{1.25}$$

Therefore we can define a density $\mathrm{p}_{\mathsf{y}} = \frac{\mathrm{dP}_{\mathsf{y}}}{\mathrm{d}\lambda^N}$ satisfying (1.15) as

$$\mathrm{p}_{\mathsf{y}}(\boldsymbol{y}) = \int_{\boldsymbol{\phi}^{-1}(\boldsymbol{y})} \mathrm{p}_{\mathsf{x}}(\boldsymbol{x})\, J_{\boldsymbol{\phi}}(\boldsymbol{x})^{-1}\, \mathrm{d}\mathcal{H}^{M-N}(\boldsymbol{x}) \quad \forall \boldsymbol{y} \notin \boldsymbol{\phi}(Z). \tag{1.26}$$

For the special case of a dimension-preserving map $\boldsymbol{\phi}$ with $N = M$ the integral in (1.26) is with respect to $\mathcal{H}^0$ which is equivalent to the counting measure #. In this case $J_{\boldsymbol{\phi}}(\boldsymbol{x}) = \left|\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}}\right|$ and we therefore get

$$\mathrm{p}_{\mathsf{y}}(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in \boldsymbol{\phi}^{-1}(\boldsymbol{y})} \mathrm{p}_{\mathsf{x}}(\boldsymbol{x}) \left|\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}}\right|^{-1} \quad \forall \boldsymbol{y} \notin \boldsymbol{\phi}(Z). \tag{1.27}$$

Under the further restriction that $\boldsymbol{\phi}$ is bi-Lipschitz, i.e. it is bijective and Lipschitz in both directions, we recover the more commonly presented multidimensional change of variables formula

$$\mathrm{p}_{\mathsf{y}}(\boldsymbol{y}) = \mathrm{p}_{\mathsf{x}} \circ \boldsymbol{\phi}^{-1}(\boldsymbol{y}) \left|\frac{\partial \boldsymbol{\phi}^{-1}}{\partial \boldsymbol{y}}\right| \quad \forall \boldsymbol{y} \in Y. \tag{1.28}$$

In both of the cases considered, we have seen that if the function ϕ the random variable x is mapped through is bijective, the resulting expression for the density on the mapped random variable y is simpler in the sense that the pre-image $\phi^{-1}(y)$ of a point $y \in Y$ is itself a point and so we do not need to integrate or sum over points in the pre-image which will often be difficult to do analytically.

Bijectivity is a very limiting condition however, with many models involving non-bijective transformations of random variables. Later in this thesis we will see that methods used for defining the more general forms for calculating the density of a transformed variable are key to proposed 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 fundamental operation when working with probabilistic models is computing expectations of random variables. Let $(S, \mathcal{E}, \mathsf{P})$ be a probability space, and $\mathsf{x} : S \to X$ a random variable on this space. Then the *expected value of* x is defined as

$$\mathbb{E}[\mathsf{x}] = \int_S \mathsf{x}(s)\, \mathrm{dP}(s). \tag{1.29}$$

Often it will be more convenient to express expectations in terms of the probability P_x instead. If $f : S \to X$ is a measurable function and μ a measure on S then the integral with respect to the pushforward measure μ_f of an integrable function g satisfies

$$\int_X g(x)\, \mathrm{d}\mu_f(x) = \int_S g \circ f(s)\, \mathrm{d}\mu(s). \tag{1.30}$$

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

$$\mathbb{E}[\mathsf{x}] = \int_X x\, \mathrm{dP}_\mathsf{x}(x). \tag{1.31}$$

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

$$\mathbb{E}[\mathsf{x}] = \int_X x\, \mathsf{p}_\mathsf{x}(x)\, \mathrm{d}\mu(x), \tag{1.32}$$

which is often the form used for computation.

A further useful implication of (1.30) 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 the expected value of y is

$$\mathbb{E}[\mathsf{y}] = \int_S \mathsf{y}(s)\, \mathrm{dP}(s) = \int_S \phi \circ \mathsf{x}(s)\, \mathrm{dP}(s) = \int_X \phi(x)\, \mathrm{dP}_\mathsf{x}(x), \tag{1.33}$$

i.e. it can be calculated by integrating ϕ with respect to P_x. This means we can calculate expectations 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.

1.1.6 Conditional expectations and densities

A related concept, and one which will be key in our 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 $\mathbb{E}[\mathsf{x} \mid \mathsf{y}] : Y \to X$ satisfying

$$\int_{\mathsf{y}^{-1}(A)} \mathsf{x}(s) \, \mathrm{dP}(s) = \int_A \mathbb{E}[\mathsf{x} \mid \mathsf{y}](y) \, \mathrm{dP}_{\mathsf{y}}(y) \quad \forall A \in \mathcal{G}. \tag{1.34}$$

$\mathbb{E}[\mathsf{x} \mid \mathsf{y}]$ is guaranteed to be uniquely defined almost everywhere in Y by (1.34), i.e. up to P_{y}-null sets []. As a particular case where $A = Y$ we recover what is sometimes termed the *Law of total expectation*

$$\int_S \mathsf{x} \, \mathrm{dP} = \int_S \mathbb{E}[\mathsf{x} \mid \mathsf{y}] \circ \mathsf{y} \, \mathrm{dP} \implies \mathbb{E}[\mathsf{x}] = \mathbb{E}[\mathbb{E}[\mathsf{x} \mid \mathsf{y}] \circ \mathsf{y}]. \tag{1.35}$$

We can also motivate a definition of conditional density in terms of conditional expectation. Assume a joint density $\mathsf{p}_{\mathsf{x},\mathsf{y}} = \frac{\mathrm{dP}_{\mathsf{x},\mathsf{y}}}{\mathrm{d}(\mu_{\mathsf{x}} \times \mu_{\mathsf{y}})}$ exists and has marginal density $\mathsf{p}_{\mathsf{y}} = \frac{\mathrm{dP}_{\mathsf{y}}}{\mathrm{d}\mu_{\mathsf{y}}}$. Then for all $A \in \mathcal{G}$

$$\begin{aligned}
\int_{\mathsf{y}^{-1}(A)} \mathsf{x}(s) \, \mathrm{dP}(s) &= \int_S \mathsf{x}(s) \, \mathbb{1}_A \circ \mathsf{y}(s) \, \mathrm{dP}(s) && (1.36)\\
&= \int_{X \times Y} x \, \mathbb{1}_A(y) \, \mathrm{dP}_{\mathsf{x},\mathsf{y}}(x, y) && (1.37)\\
&= \int_A \int_X x \, \mathsf{p}_{\mathsf{x},\mathsf{y}}(x, y) \, \mathrm{d}\mu_{\mathsf{x}}(x) \, \mathrm{d}\mu_{\mathsf{y}}(y). && (1.38)
\end{aligned}$$

Define $g : Y \to X$ as

$$g(y) = \begin{cases} \int_X x \, \frac{\mathsf{p}_{\mathsf{x},\mathsf{y}}(x,y)}{\mathsf{p}_{\mathsf{y}}(y)} \, \mathrm{d}\mu_{\mathsf{x}}(x) & \forall y \in Y : \mathsf{p}_{\mathsf{y}}(y) > 0 \\ 0 & \forall y \in Y : \mathsf{p}_{\mathsf{y}}(y) = 0. \end{cases} \tag{1.39}$$

Then from (1.38) we have that for all $A \in \mathcal{G}$

$$\int_{\mathsf{y}^{-1}(A)} \mathsf{x}(s) \, \mathrm{dP}(s) = \int_A g(y) \, \mathsf{p}_{\mathsf{y}}(y) \, \mathrm{d}\mu_{\mathsf{y}}(y) = \int_A g(y) \, \mathrm{dP}_{\mathsf{y}}(y). \tag{1.40}$$

The definition of g in (1.39) therefore satisfies the definition of conditional expectation in (1.34) and is uniquely defined up to a P_y-null set. Therefore if $\mathsf{p}_{\mathsf{x},\mathsf{y}}$ and p_y can be defined we have that

$$\mathbb{E}[\mathsf{x} \mid \mathsf{y}](y) = \int_X x \, \mathsf{p}_{\mathsf{x}|\mathsf{y}}(x \mid y) \, \mathrm{d}\mu_\mathsf{x}(x) \quad \forall y \in Y : \mathsf{p}_\mathsf{y}(y) > 0 \tag{1.41}$$

where the *conditional density of* x *given* y, $\mathsf{p}_{\mathsf{x}|\mathsf{y}}$, is defined as

$$\mathsf{p}_{\mathsf{x}|\mathsf{y}}(x \mid y) = \frac{\mathsf{p}_{\mathsf{x},\mathsf{y}}(x, y)}{\mathsf{p}_\mathsf{y}(y)} \quad \forall x \in X, \, y \in Y : \mathsf{p}_\mathsf{y}(y) > 0 \tag{1.42}$$

which can be seen to be analagous to the definition of conditional probability in (1.5). Note the definition of conditional expectation in (1.34) was not dependent on a joint density $\mathsf{p}_{\mathsf{x},\mathsf{y}}$ being defined and so is more general than (1.41).

1.2 GRAPHICAL MODELS

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

When working with probabilistic models defining large collections 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 conditional relationships between them. Graphical models, which use graphs to describe the relationship between random variables, are a useful framework for visualising the structure in complex probabilistic models and for giving a graph-theoretic basis for establishing the dependence between sets of random variables.

Central to all graphical models in the concept of conditional independence. Let $(S, \mathcal{E}, \mathsf{P})$ be a probability space and $\mathsf{x} : S \to X$, $\mathsf{y} : S \to Y$ and $\mathsf{z} : S \to Z$ be three random variables with corresponding σ-algebras, $\mathcal{F}_\mathsf{x}$, $\mathcal{F}_\mathsf{y}$ and $\mathcal{F}_\mathsf{z}$ respectively. Analagously to the earlier definition of (unconditional) independence of random variables in (1.4), we say that x *and* y *are conditionally independent given* z, denoted $\mathsf{x} \perp \mathsf{y} \mid \mathsf{z}$, if

$$\begin{aligned} \mathsf{x} \perp \mathsf{y} \mid \mathsf{z} \iff \mathsf{P}_{\mathsf{x},\mathsf{y}|\mathsf{z}}(A, B \mid C) = \mathsf{P}_{\mathsf{x}|\mathsf{z}}(A \mid C)\mathsf{P}_{\mathsf{y}|\mathsf{z}}(B \mid C) \\ \forall A \in \mathcal{F}_\mathsf{x}, \, B \in \mathcal{F}_\mathsf{y}, \, C \in \mathcal{F}_\mathsf{z}. \end{aligned} \tag{1.43}$$

An equivalent property can also be defined for conditional densities

$$\begin{aligned} \mathsf{x} \perp \mathsf{y} \mid \mathsf{z} \iff \mathsf{p}_{\mathsf{x},\mathsf{y}|\mathsf{z}}(x, y \mid z) = \mathsf{p}_{\mathsf{x}|\mathsf{z}}(x \mid z)\mathsf{p}_{\mathsf{y}|\mathsf{z}}(y \mid z) \\ \forall x \in X, \, y \in Y, \, z \in Z. \end{aligned} \tag{1.44}$$

These definitions can be naturally extended to conditional independence of more than two random variables and when conditioning on more than one random variable, for example if $\mathsf{u} \perp \mathsf{v} \perp \mathsf{x} \mid \mathsf{y}, \mathsf{z}$

$$\mathsf{p}_{\mathsf{u},\mathsf{v},\mathsf{x}|\mathsf{y},\mathsf{z}}(u, v, x \mid y, z) = \mathsf{p}_{\mathsf{u}|\mathsf{y},\mathsf{z}}(u \mid y, z)\mathsf{p}_{\mathsf{v}|\mathsf{y},\mathsf{z}}(v \mid y, z)\mathsf{p}_{\mathsf{x}|\mathsf{y},\mathsf{z}}(x \mid y, z). \quad (1.45)$$

1.2.1 Directed and undirected graphical models

(a) Directed graphical model. (b) Undirected graphical model.

Figure 1.1: Examples of directed and undirected graphical models.

Several different graphical frameworks have been proposed for representing conditional independency relationships (and other information) in probabilistic models.

Directed graphical models[], also known as *Bayesian networks*, represent probabilistic models as *directed acyclic graphs* (i.e. a directed graph in which there are no directed cycles), with the nodes in the graph representing random variables in the model and the edges of the graph defining a factorisation of the joint density over these variables into a product of conditional and marginal densities. In particular a conditional density factor is included for each node with parents (on the node random variable value given the parent variable values) and a marginal density factor for each root node without any parents.

An example directed graphical model for three random variables, x_1, x_2 and x_3, is shown in Figure 1.1a. The graph implies that the joint density can be factorised as

$$\mathsf{p}_{\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3}(x_1, x_2, x_3) = \mathsf{p}_{\mathsf{x}_3|\mathsf{x}_1,\mathsf{x}_2}(x_3 \mid x_1, x_2)\, \mathsf{p}_{\mathsf{x}_1}(x_1)\, \mathsf{p}_{\mathsf{x}_2}(x_2). \quad (1.46)$$

Note that this factorisation would not be valid for all joint densities on the three variables; in particular we have that x_1 and x_2 are (unconditionally) independent and so that the joint density $\mathsf{p}_{\mathsf{x}_1,\mathsf{x}_2}$ can be written as the product of the two marginals $\mathsf{p}_{\mathsf{x}_1}$ and $\mathsf{p}_{\mathsf{x}_2}$.

Directed graphical models are a natural way of specifying *generative models* - i.e. probabilistic models which can be used to generate simulated observable quantities. Typically the factorisation specified by a directed graphical model gives a natural way to generate values from the joint density, via *ancestral sampling*.

Ancestral sampling in a directed graphical model corresponds to first sampling values from all the root nodes from their marginal densities, then iteratively sampling from the conditional densities on each node for which all the parents nodes already have sampled values to condition on.

An alternative formalism for graphically representing probabilistic models is that of *undirected graphical models*[], which are also known as *Markov networks*. As with directed graphical models, each node in the graph represents a random variable, but here the edges connecting nodes are undirected. Rather than describing a factorisation of a joint density into conditional and marginal densities, an undirected graphical model indicates the factorisation of a joint density into a product of *clique potentials* on each of the *maximal cliques* in the graph.

A *clique* is a fully connected component of the graph - i.e. a subset of nodes in the graph such that all pairs of nodes in the subset are connected by an edge. A *maximal clique* is a clique which is not a strict subset of any other clique. A *clique potential* is a non-negative function of the values of the random variables in the clique; it does necessarily correspond to any conditional or marginal probabilty density.

An example undirected graphical model on four random variables, x_1, x_2, x_3 and x_4, is shown in Figure 1.1b. Here the (maximal) cliques correspond to all the connected pairs of nodes. If $\psi_{a,b}$ denotes the clique potential on the pair (a, b) then the graphical model implies the joint density can be factorised as

$$\begin{aligned} \mathrm{p}_{\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3,\mathrm{x}_4}(x_1, x_2, x_3, x_4) = \\ \frac{1}{Z}\psi_{\mathrm{x}_1,\mathrm{x}_2}(x_1, x_2)\psi_{\mathrm{x}_1,\mathrm{x}_3}(x_1, x_3)\psi_{\mathrm{x}_2,\mathrm{x}_4}(x_2, x_4)\psi_{\mathrm{x}_3,\mathrm{x}_4}(x_3, x_4), \end{aligned} \tag{1.47}$$

with Z a normalising constant such that the density integrates to 1 and so defines a valid probability measure.

Undirected graphical models are a natural representation for models of systems of mutually interacting components. For example they are commonly used in models of images to represent dependencies between pixel values and models of ferromagnetism to represent interactions between lattices of particles.

Unlike directed models, generating joint configurations of the random variables in an undirected graphical model from the implied joint dis-

tribution is typically a non-trivial task, with no general equivalent to ancestral sampling. Further the joint density can typically only be evaluated upto an unknown normalising constant, with the integral needed to evaluate this constant often intractable for models involving a large number of variables or complex potentials. These properties mean that inference in distributions defined by undirected graphical models is often particularly challenging.

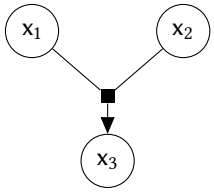
As suggested at the start of this section, both directed and undirected graphical models encode conditional independence properties of probabilistic models. In particular the rules of *D-separation* for directed graphical models and *U-separation* for undirected model give algorithmic descriptions of how to determine whether a pair of random variables are conditionally independent for a given conditioning set of random variables in terms of graph based operations.

For example the directed graphical model in Figure 1.1a encodes the (un)conditional independence property $x_1 \perp x_2 \mid \emptyset = x_1 \perp x_2$ i.e. that x_1 and x_2 are independent if the value of x_3 is *not* conditioned on. The undirected graphical model in Figure 1.1b encodes the conditional independence properties $x_1 \perp x_4 \mid x_2, x_3$ and $x_2 \perp x_3 \mid x_1, x_4$.

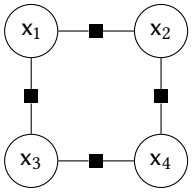
Although there are method to convert a directed graphical model to an undirected one and vice versa, in general these transformations are lossy - not all of the conditional independence relationships encoded in the original graph will necessarily be maintained in the transformed graph. For example there is no undirected graphical model which will represent the exact set of conditional independence properties represented by the directed graphical model in Figure 1.1a. Likewise there is no directed graphical model which will represent the exact set of conditional independence properties represented by the undirected graphical model in Figure 1.1b. Further there are distributions with dependency structures and factorisations which cannot be well represented by either directed or undirected graphical models [].

1.2.2 Factor graphs

An alternative graphical model format which overcomes some of the just mentioned limitations of traditional directed and undirected graphical models is factor graphs [].



(a) A factor graph equivalent of the directed model in Figure 1.1a.



(b) A factor graph equivalent to the undirected model in Figure 1.1b.

Figure 1.2: Examples of factor graphs corresponding to the directed and undirected graphical models in Figure 1.1

.

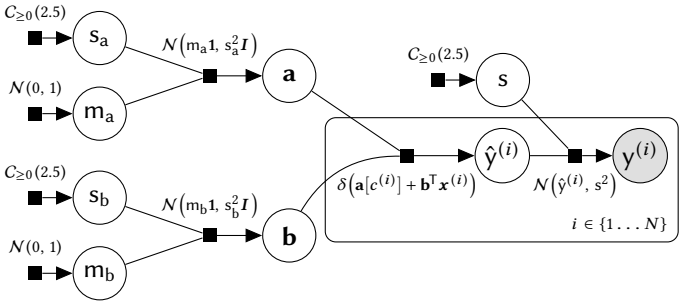


Figure 1.3: Hierarchical linear regression model factor graph.

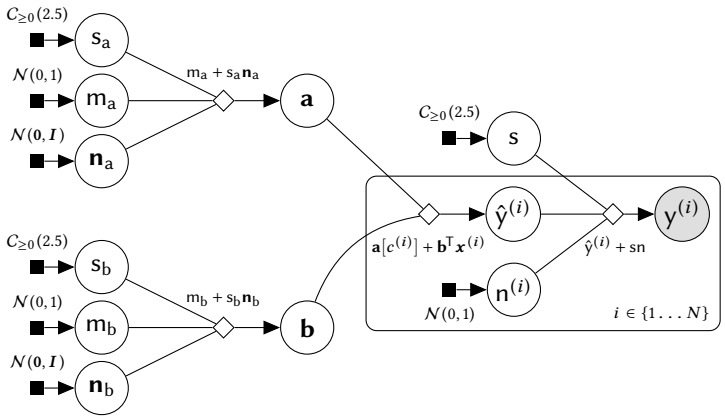


Figure 1.4: Hierarchical linear regression model stochastic computation graph.

1.2.3 Stochastic computation graphs

1.3 INFERENCE

You cannot do inference without making assumptions
—David Mackay

1.3.1 Posterior expectations

1.3.2 Model evidence

2 APPROXIMATE INFERENCE

2.1 DETERMINISTIC APPROACHES

2.1.1 Laplace's method

2.1.2 Variational inference

2.1.3 Expectation propagation

2.2 STOCHASTIC APPROACHES

2.2.1 Monte Carlo method

2.2.2 Rejection sampling

2.2.3 Importance sampling

2.2.4 Markov chain Monte Carlo

3 MARKOV CHAIN MONTE CARLO

3.1 METROPOLIS–HASTINGS

3.2 GIBBS SAMPLING

3.3 SLICE SAMPLING

3.4 HAMILTONIAN MONTE CARLO

4 THERMODYNAMIC METHODS

4.1 SIMULATED TEMPERING

4.2 PARALLEL TEMPERING

4.3 TEMPERED TRANSITIONS

4.4 ANNEALED IMPORTANCE SAMPLING

4.5 PATH SAMPLING

4.6 ADIABATIC MONTE CARLO