

# 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 expectations 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-dimensional, 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 concludaturque. 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 11
1.1.4	Change of variables 13
1.1.5	Expectations 16
1.1.6	Conditional expectations and densities 17
1.2	Graphical models 18
1.2.1	Directed and undirected graphical models 18
1.2.2	Factor graphs 18
1.2.3	Stochastic computation graphs 18
1.3	Inference 18
1.3.1	Posterior expectations 19
1.3.2	Model evidence 19
2	APPROXIMATE INFERENCE 21
2.1	Deterministic approaches 21
2.1.1	Laplace's method 21
2.1.2	Variational inference 21
2.1.3	Expectation propagation 21
2.2	Stochastic approaches 21
2.2.1	Monte Carlo method 21
2.2.2	Rejection sampling 21
2.2.3	Importance sampling 21
2.2.4	Markov chain Monte Carlo 21

3	MARKOV CHAIN MONTE CARLO	23
3.1	Metropolis–Hastings	23
3.2	Gibbs sampling	23
3.3	Slice sampling	23
3.4	Hamiltonian Monte Carlo	23
4	THERMODYNAMIC METHODS	25
4.1	Simulated tempering	25
4.2	Parallel tempering	25
4.3	Tempered transitions	25
4.4	Annealed importance sampling	25
4.5	Path sampling	25
4.6	Adiabatic Monte Carlo	25



## LIST OF FIGURES

## 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.

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.

### 1.1 PROBABILITY THEORY

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

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

*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

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

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

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  $P(E)$ .

The measure-theoretic approach has the advantage 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  
 $f : X \rightarrow Y$  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 observed and unobserved quantities in the model of interest. A random variable  $x : S \rightarrow X$  is defined as a measurable function from the sample space to a measurable space  $(X, \mathcal{F})$ .

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

Often  $X$  is the reals,  $\mathbb{R}$ , and  $\mathcal{F}$  is the Borel  $\sigma$ -algebra on the reals,  $\mathcal{B}(\mathbb{R})$ , 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  $\mathcal{F} = \mathcal{B}(\mathbb{R}^D)$  - in this case we will term the resulting random variable a *random vector* and use the notation  $\mathbf{x} : S \rightarrow X$ . A final special case is when  $X$  is countable and  $\mathcal{F}$  is the power set  $\mathcal{P}(X)$  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,  $\mu$  a measure  
 on these spaces and  
 $f : X \rightarrow Y$  a  
 measurable function,  
 the pushforward  
 measure  $\mu_f$  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  $x$

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

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  $P_x(A)$ .

### 1.1.2 Joint and conditional probability

Often we will jointly define multiple random variables on the same probability space. Let  $(S, \mathcal{E}, P)$  be a probability space and  $x : S \rightarrow X$ ,

$y : S \rightarrow Y$  be two random variables with corresponding  $\sigma$ -algebras  $\mathcal{F}$  and  $\mathcal{G}$ . Then the *joint probability* of  $x$  and  $y$  is defined as

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

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

$$P_{x,y}(A, Y) = P_x(A), P_{x,y}(X, B) = P_y(B) \quad \forall A \in \mathcal{F}, B \in \mathcal{G}. \quad (1.3)$$

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

Also useful is the definition of *conditional probability*

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad \forall A \in \mathcal{E}, B \in \mathcal{E} : P(B) \neq 0. \quad (1.4)$$

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

$$P_{x,y}(A, B) = P_{x|y}(A | B) P_y(B) = P_{y|x}(B | A) P_x(A) \quad (1.5)$$

$$\forall A \in \mathcal{F}, B \in \mathcal{G} : P_{x,y}(A, B) \neq 0,$$

*In Kolmogorov's probability theory, (1.4) 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.*

which is sometimes referred to as the product rule.

An implication of (1.5) is what is often termed *Bayes' theorem*

$$P_{x|y}(A | B) = \frac{P_{y|x}(B | A) P_x(A)}{P_y(B)} \quad \forall A \in \mathcal{F}, B \in \mathcal{G} : P_y(B) \neq 0, \quad (1.6)$$

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 arbitrarily 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 (1.5) 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  $\sigma$ -finite if  $X$  is a countable union of finite measure sets.

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

The Radon–Nikodym theorem [1] guarantees that for a pair of  $\sigma$ -finite measures  $\mu$  and  $\nu$  on a measurable space  $(X, \mathcal{F})$  where  $\nu$  is absolutely continuous with respect to  $\mu$ , then there is a unique (up to  $\mu$ -null sets) measurable function  $f : X \rightarrow [0, \infty)$  termed a *density* such that

$$\nu(A) = \int_A f \, d\mu \quad \forall A \in \mathcal{F}. \quad (1.7)$$

The density function  $f$  is also termed the *Radon-Nikodym derivative* of  $\nu$  with respect to  $\mu$ , denoted  $\frac{d\nu}{d\mu}$ . 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  $f = \frac{d\nu}{d\mu}$  and  $g$  is a measurable function that

$$\int_X g \, d\nu = \int_X g f \, d\mu, \quad (1.8)$$

which we will use later when discussing calculation of expectations.

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

$$P_x(A) = \int_A p_x \, d\lambda = \int_A p_x(x) \, dx \quad \forall A \in \mathcal{B}(\mathbb{R}). \quad (1.9)$$

Analagously for a random vector  $\mathbf{x}$  with density  $p_{\mathbf{x}} : \mathbb{R}^D \rightarrow [0, \infty)$  with respect to the  $D$ -dimensional Lebesgue measure  $\lambda^D$

$$P_{\mathbf{x}}(A) = \int_A p_{\mathbf{x}} \, d\lambda^D = \int_A p_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x} \quad \forall A \in \mathcal{B}(\mathbb{R}^D). \quad (1.10)$$

The notation in the second equalities in (1.9) and (1.10) use a convention that will be used throughout this thesis that integrals without an explicit measure (but with an explicit variable of integration) are assumed to be 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  $p_x : X \rightarrow [0, 1]$  by

$$P_x(A) = \int_A p_x \, d\# = \sum_{x \in A} p_x(x) \quad \forall A \in \mathcal{P}(X). \quad (1.11)$$

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  $\sum_{x \in X} p_x(x) = 1$ . 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 defining joint probabilities on a mixture of real and discrete random variables.

The joint probability  $P_{x,y}$  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  $p_{x,y} : X \times Y \rightarrow [0, \infty)$  by

$$P_{x,y}(A, B) = \int_{A \times B} p_{x,y} d(\mu_x \times \mu_y) \quad \forall A \in \mathcal{F}, B \in \mathcal{G}, \quad (1.12)$$

where  $\mu_x \times \mu_y$  represents the product measure of two appropriate base measures  $\mu_x$  and  $\mu_y$  e.g.  $\mu_x = \lambda$  and  $\mu_y = \#$  if  $x$  is a real random variable and  $y$  is a discrete random variable.

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  $(S, \mathcal{E}, P)$  which the random variable(s) map from. In this case we will typically use the loose notation  $x \in X$  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 statistics 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 Change of variables

It is common to define a random variable via a transform of another. Let  $x \in X$  be a random variable with co-domain the measurable space

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

$(X, \mathcal{F})$ . Further let  $(Y, \mathcal{G})$  be a second measurable space and  $\phi : X \rightarrow Y$  a measurable function between the two spaces. Then if we define  $y = \phi(x)$  then analogously to our original definition of  $P_x$  as the push-forward measure of  $P$  under the measurable function defining  $x$ , we can define  $P_y$  in terms of  $P_x$  as

$$P_y(A) = P_x \circ \phi^{-1}(A) = P_x(\{x \in X : \phi(x) \in A\}) \quad \forall A \in \mathcal{G}, \quad (1.13)$$

i.e. the probability of the event  $y \in A$  is equal to the probability of  $x$  being in the pre-image under  $\phi$  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  $\mu_x$ , we can also in some cases find a density  $p_y$  on the transformed variable  $y = \phi(x)$  with respect to a (potentially different) measure  $\mu_y$  which can be used to calculate the probability  $P_y$ ,

$$P_y(A) = \int_A p_y d\mu_y = \int_{\phi^{-1}(A)} p_x d\mu_x \quad \forall A \in \mathcal{G}. \quad (1.14)$$

For random variables with countable co-domains where the integral in (1.14) corresponds to a sum, a  $p_y$  satisfying (1.14) is simple to identify. If  $x$  is a discrete random variable with probability density  $p_x$  with respect to the counting measure, then  $y = \phi(x)$  will necessarily also be a discrete random variable and we can define a probability density  $p_y$  with respect to the counting measure in terms of  $p_x$  as

$$p_y(y) = \sum_{x \in \phi^{-1}(y)} p_x(x) \quad \forall y \in Y. \quad (1.15)$$

In the special case that  $\phi$  is bijective we have that

$$p_y(y) = p_x \circ \phi^{-1}(y) \quad \forall y \in Y. \quad (1.16)$$

Now consider the case where  $x$  and  $y$  are both real random variables. Here the situation is more complicated as we need to account for any local contraction / expansion of space by the map  $\phi$ . For In one dimension this is described by  $\left| \frac{\partial \phi}{\partial x} \right|$  the absolute value of the derivative of  $\phi$ . If  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  is an absolutely continuous function, then let  $Z \subseteq \mathbb{R}$  be the (possibly empty) set of points where the derivative  $\frac{\partial \phi}{\partial x}$  does not exist



or is equal to zero. If  $P_x(Z) = 0$  and  $\mathbf{x}$  has probability density  $p_x$  with respect to the Lebesgue measure then

$$p_y(y) = \sum_{x \in \phi^{-1}(y)} p_x(x) \left| \frac{\partial \phi}{\partial x} \right|^{-1} \quad \forall y \notin \phi(Z), \quad (1.17)$$

defines a probability density for  $y$ . The density can take arbitrary finite values for  $y \in \phi(Z)$  as  $\lambda(\phi(Z)) = 0$  due to the assumed absolute continuity of  $\phi$  and the absolute continuity of  $P_x$  with respect to  $\lambda$  so that  $P_x(Z) = 0 \Rightarrow \lambda(Z) = 0$ . For bijective  $\phi$  such that  $\phi^{-1}$  is uniquely defined, then we recover the more commonly quoted special case

$$p_y(y) = p_x \circ \phi^{-1}(y) \left| \frac{\partial \phi^{-1}}{\partial y} \right| \quad \forall y \in Y. \quad (1.18)$$

We can also generalise to functions of random vectors. Let  $\mathbf{x} \in \mathbb{R}^M$  and  $\mathbf{y} \in \mathbb{R}^N$  be two random vectors with  $M \geq N$ . Further let  $\phi : \mathbb{R}^M \rightarrow \mathbb{R}^N$  be a Lipschitz map and  $Z \subseteq \mathbb{R}^M$  be the (possibly empty) set of points where the Jacobian ( $N \times M$  matrix of partial derivatives)  $\frac{\partial \phi}{\partial \mathbf{x}}$  does not exist or has determinant zero. If  $P_x(Z) = 0$  and  $\mathbf{x}$  has probability density  $p_x$  with respect to the Lebesgue measure  $\lambda^M$  then

$$p_y(\mathbf{y}) = \int_{\phi^{-1}(\mathbf{y})} p_x(\mathbf{x}) \left| \frac{\partial \phi}{\partial \mathbf{x}} \frac{\partial \phi^\top}{\partial \mathbf{x}} \right|^{-\frac{1}{2}} d\mathcal{H}^{M-N}(\mathbf{x}) \quad \forall \mathbf{y} \notin \phi(Z), \quad (1.19)$$

where  $\mathcal{H}^D$  is the  $D$ -dimensional *Hausdorff measure*. Again  $p_y$  can take arbitrary finite values on  $\phi(Z)$ . This result can be derived from *Federer's Co-Area Formula* []. As the Hausdorff measure of dimension zero is equivalent to the counting measure i.e.  $\mathcal{H}^0 = \#$ , then for the case where  $M = N$  i.e.  $\phi$  is dimension preserving then we have that

$$p_y(\mathbf{y}) = \sum_{x \in \phi^{-1}(\mathbf{y})} p_x(x) \left| \frac{\partial \phi}{\partial \mathbf{x}} \right|^{-1} \quad \forall \mathbf{y} \notin \phi(Z), \quad (1.20)$$

which (1.17) can be seen to be a special case of for  $M = N = 1$ . If  $\phi$  is further constrained to be a diffeomorphism, i.e. it is differentiable and bijective and its inverse is also differentiable, then the standard multidimensional change of variables formula is recovered

$$p_y(\mathbf{y}) = p_x \circ \phi^{-1}(\mathbf{y}) \left| \frac{\partial \phi^{-1}}{\partial \mathbf{y}} \right| \quad \forall \mathbf{y} \in Y. \quad (1.21)$$

In all of the cases considered, we have seen that if the function  $\phi$  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.

Bijectivity is a very limiting condition however, with many models involving non-bijective transformations of random variables. In particular later in this thesis we will see that 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  $(S, \mathcal{E}, P)$  be a probability space, and  $x : S \rightarrow X$  a random variable on this space. Then the *expected value of  $x$*  is defined as

$$\mathbb{E}[x] = \int_S x(s) dP(s). \quad (1.22)$$

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

$$\int_X g(x) d\mu_f(x) = \int_S g \circ f(s) d\mu(s). \quad (1.23)$$

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

$$\mathbb{E}[x] = \int_X x dP_x(x). \quad (1.24)$$

If  $P_x$  is given by a density  $p_x = \frac{dP_x}{d\mu}$  then using (1.8) we also have

$$\mathbb{E}[x] = \int_X x p_x(x) d\mu(x), \quad (1.25)$$

which is often the form used for computation.

A further useful implication of (1.23) is what is sometimes termed the *Law of the unconscious statistician*. Let  $x : S \rightarrow X$  be a random variable,  $\phi : X \rightarrow Y$  a measurable function and define  $y = \phi \circ x$ . Then the expected value of  $y$  is

$$\mathbb{E}[y] = \int_S y(s) dP(s) = \int_S \phi \circ x(s) dP(s) = \int_X \phi(x) dP_x(x), \quad (1.26)$$

i.e. it can be calculated by integrating  $\phi$  with respect to  $P_x$ . This means we can calculate expectations of a transformed random variable  $y = \phi(x)$  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  $\phi$ .

### 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  $(S, \mathcal{E}, P)$  be a probability space,  $(X, \mathcal{F})$  and  $(Y, \mathcal{G})$  two measurable spaces and  $x : S \rightarrow X$  and  $y : S \rightarrow Y$  two random variables. Then the *conditional expectation of  $x$  given  $y$* , is defined as a measurable function  $\mathbb{E}[x | y] : Y \rightarrow X$  satisfying

$$\int_{y^{-1}(A)} x(s) dP(s) = \int_A \mathbb{E}[x | y](y) dP_y(y) \quad \forall A \in \mathcal{G}. \quad (1.27)$$

$\mathbb{E}[x | y]$  is guaranteed to be uniquely defined almost everywhere in  $Y$  by (1.27), i.e. up to  $P_y$ -null sets  $[\cdot]$ . As a particular case where  $A = Y$  we recover what is sometimes termed the *Law of total expectation*

$$\int_S x dP = \int_S \mathbb{E}[x | y] \circ y dP \implies \mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x | y] \circ y]. \quad (1.28)$$

We can also motivate a definition of conditional density in terms of conditional expectation. Assume a joint density  $p_{x,y} = \frac{dP_{x,y}}{d(\mu_x \times \mu_y)}$  exists and has marginal density  $p_y = \frac{dP_y}{d\mu_y}$ . Then for all  $A \in \mathcal{G}$

$$\int_{y^{-1}(A)} x(s) dP(s) = \int_S x(s) \mathbb{1}_A \circ y(s) dP(s) \quad (1.29)$$

$$= \int_{X \times Y} x \mathbb{1}_A(y) dP_{x,y}(x, y) \quad (1.30)$$

$$= \int_A \int_X x p_{x,y}(x, y) d\mu_x(x) d\mu_y(y). \quad (1.31)$$

Define  $g : Y \rightarrow X$  as

$$g(y) = \begin{cases} \int_X x \frac{p_{x,y}(x,y)}{p_y(y)} d\mu_x(x) & \forall y \in Y : p_y(y) > 0 \\ 0 & \forall y \in Y : p_y(y) = 0. \end{cases} \quad (1.32)$$

Then from (1.31) we have that

$$\int_{Y^{-1}(A)} x(s) dP(s) = \int_A g(y) p_y(y) d\mu_y(y) \quad (1.33)$$

$$= \int_A g(y) dP_y(y) \quad \forall A \in \mathcal{G} \quad (1.34)$$

The definition of  $g$  in (1.32) therefore satisfies the definition of conditional expectation in (1.27) and is uniquely defined up to a  $P_y$ -null set. Therefore if  $p_{x,y}$  and  $p_y$  can be defined we have that

$$\mathbb{E}[x | y](y) = \int_X x p_{x|y}(x | y) d\mu_x(x) \quad \forall y \in Y : p_y(y) > 0 \quad (1.35)$$

where the *conditional density of  $x$  given  $y$* ,  $p_{x|y}$ , is defined as

$$p_{x|y}(x | y) = \frac{p_{x,y}(x, y)}{p_y(y)} \quad \forall x \in X, y \in Y : p_y(y) > 0 \quad (1.36)$$

which can be seen to be analogous to the definition of conditional probability in (1.4). Note the definition of conditional expectation in (1.27) was not dependent on a joint density  $p_{x,y}$  being defined and so is more general than (1.35).

## 1.2 GRAPHICAL MODELS

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

### 1.2.1 Directed and undirected graphical models

### 1.2.2 Factor graphs

### 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

