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

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.

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

Formally, a probability space is defined as a triplet $(S, \mathcal{E}, \mu)$ 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),
- μ is the *probability measure*, a finite measure satisfying $\mu(S) = 1$, 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:* $\mu(E) \geq 0 \; \forall E \in \mathcal{E}$,
2. *Normalisation:* $\mu(S) = 1$,
3. *Countable additivity: for any countable set of disjoint events* $\{E_i\}_i : E_i \in \mathcal{F} \; \forall i$, $E_i \cap E_j = \emptyset \; \forall i, j$, $\mu(\cup_i E_i) = \sum_i \mu(E_i)$.

Given this definition of a probability space, Kolmogorov's axioms [] can be used to derive a measure-theoretic formulation of probability theory. The 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

When modelling real-world processes, rather than considering abstract sample spaces, it is usually more helpful to consider *random variables* which represent the observed and unobserved variables in the model of interest. Formally a random variable $\mathsf{x} : S \to X$ is a measurable function from the sample space to a measurable space $(X, \mathcal{F})$. Often X is the reals $\mathbb{R}$ and $\mathcal{F}$ is the Borel σ-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 $\mathbf{x} : S \to X$. A final special case is when X is (a subset of) the integers $\mathbb{Z}$ and $\mathcal{F}$ is the power set $\mathcal{P}(X)$ in which case we will refer to x as a *discrete random variable*.

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

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

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

The measure μ_{x} therefore defines the probability that the random variable x takes a value in a measurable set $A \in \mathcal{F}$ as $\mu_{\mathsf{x}}(A)$.

1.1.2 Probability densities

So far we have ignored how the probability measure μ (or by consequence the pushforward measure on a random variable μ_{x}) is defined.

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

$$\nu(A) = \int_A f \, \mathrm{d}\mu \quad \forall A \in \mathcal{F}. \tag{1.2}$$

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 (which the probability measure will be absolutely continuous with respect to).

For the common case of real random variables (vectors), an appropriate choice of base measure is the *Lebesgue measure*, λ, on the reals $\mathbb{R}$ ($\mathbb{R}^D$ for random vectors). The pushforward measure μ_{x} defining the probability distribution of a real random variable x can then be defined via a *probability density function* $p_{\mathsf{x}} : X \to [0, \infty)$ by

$$\mu_{\mathsf{x}}(A) = \int_A p_{\mathsf{x}} \, \mathrm{d}\lambda = \int_A p_{\mathsf{x}}(x) \, \mathrm{d}x \qquad \forall A \subseteq \mathcal{B}(\mathbb{R}) \tag{1.3}$$

with an equivalent definition for a random vector $\mathbf{x}$ with density $p_{\mathbf{x}}$. The notation in the second equality uses 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.

For discrete random variables, an appropriate base measure is instead the *counting measure*, $\#$. The probability distribution of a discrete random variable can then be defined via a *probability mass function* $P_{\mathsf{x}} : X \to [0, 1]$ by

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

$$\mu_{\mathsf{x}}(A) = \int_A P_{\mathsf{x}} \, \mathrm{d}\# = \sum_{x \in A} P_{\mathsf{x}}(x) \qquad \forall A \subseteq \mathcal{P}(X). \tag{1.4}$$

Note that technically P_{x} is a density with respect to the counting measure, however we will follow the common convention of reserving density for real random variables. Unlike a probability density p_{x}, the codomain of a probability mass function P_{x} is restricted to $[0, 1]$ due to the normalisation requirement $\mu_{\mathsf{x}}(X) = 1$.

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

You cannot do inference without making assumptions
—David Mackay

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

4 THERMODYNAMIC METHODS