INTRODUCTION TO MONTE CARLO METHODS

D.J.C. MACKAY
Department of Physics, Cambridge University.
Cavendish Laboratory, Madingley Road,
Cambridge, CB3 0HE. United Kingdom.

ABSTRACT

This chapter describes a sequence of Monte Carlo methods: importance sampling, rejection sampling, the Metropolis method, and Gibbs sampling. For each method, we discuss whether the method is expected to be useful for high-dimensional problems such as arise in inference with graphical models. After the methods have been described, the terminology of Markov chain Monte Carlo methods is presented. The chapter concludes with a discussion of advanced methods, including methods for reducing random walk behaviour.

For details of Monte Carlo methods, theorems and proofs and a full list of references, the reader is directed to Neal (1993), Gilks, Richardson and Spiegelhalter (1996), and Tanner (1996).

1. The problems to be solved

The aims of Monte Carlo methods are to solve one or both of the following problems.

Problem 1: to generate samples \( \{x^{(r)}\}_{r=1}^R \) from a given probability distribution \( P(x) \).

Problem 2: to estimate expectations of functions under this distribution, for example

\[
\Phi = \langle \phi(x) \rangle = \int d^N x \, P(x) \phi(x). \tag{1}
\]

1Please note that I will use the word "sample" in the following sense: a sample from a distribution \( P(x) \) is a single realization \( x \) whose probability distribution is \( P(x) \). This contrasts with the alternative usage in statistics, where "sample" refers to a collection of realizations \( \{x\} \).
The probability distribution $P(x)$, which we will call the **target density**, might be a distribution from statistical physics or a conditional distribution arising in data modelling — for example, the posterior probability of a model's parameters given some observed data. We will generally assume that $x$ is an $N$-dimensional vector with real components $x_n$, but we will sometimes consider discrete spaces also.

We will concentrate on the first problem (sampling), because if we have solved it, then we can solve the second problem by using the random samples $\{x^{(r)}\}_{r=1}^R$ to give the estimator

$$\hat{\Phi} \equiv \frac{1}{R} \sum_r \phi(x^{(r)}).$$  \hspace{1cm} (2)

Clearly if the vectors $\{x^{(r)}\}_{r=1}^R$ are generated from $P(x)$ then the expectation of $\hat{\Phi}$ is $\Phi$. Also, as the number of samples $R$ increases, the variance of $\hat{\Phi}$ will decrease as $\frac{\sigma^2}{R}$, where $\sigma^2$ is the variance of $\phi$,

$$\sigma^2 = \int d^N x \ P(x) (\phi(x) - \Phi)^2.$$  \hspace{1cm} (3)

This is one of the important properties of Monte Carlo methods.

**The accuracy of the Monte Carlo estimate (equation (2)) is independent of the dimensionality of the space sampled.** To be precise, the variance of $\hat{\Phi}$ goes as $\frac{\sigma^2}{R}$. So regardless of the dimensionality of $x$, it may be that as few as a dozen independent samples $\{x^{(r)}\}$ suffice to estimate $\Phi$ satisfactorily.

We will find later, however, that high dimensionality can cause other difficulties for Monte Carlo methods. Obtaining independent samples from a given distribution $P(x)$ is often not easy.

1.1. WHY IS SAMPLING FROM $P(x)$ HARD?

We will assume that the density from which we wish to draw samples, $P(x)$, can be evaluated, at least to within a multiplicative constant; that is, we can evaluate a function $P^*(x)$ such that

$$P(x) = P^*(x)/Z.$$  \hspace{1cm} (4)

If we can evaluate $P^*(x)$, why can we not easily solve problem 1? Why is it in general difficult to obtain samples from $P(x)$? There are two difficulties. The first is that we typically do not know the normalizing constant

$$Z = \int d^N x \ P^*(x).$$  \hspace{1cm} (5)