Computer Physics Communications 184 (2013) 2794–2802

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Monte Carlo integration with subtraction

Rudy Arthur [b,*], A.D. Kennedy [a]

[a] School of Physics and Astronomy, The University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland, United Kingdom

[b] CP3Origins & The Danish Institute for Advanced Study DIAS, University of Southern Denmark, Campusvej 55, DK5230 Odense M, Denmark

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form
12 July 2013
Accepted 7 August 2013
Available online 16 August 2013

Keywords:
Numerical integration
Monte Carlo
PANIC

ABSTRACT

This paper investigates a class of algorithms for numerical integration of a function in d dimensions over a compact domain by Monte Carlo methods. We construct a histogram approximation to the function using a partition of the integration domain into a set of bins specified by some parameters. We then consider two adaptations: the first is to subtract the histogram approximation, whose integral we may easily evaluate explicitly, from the function and integrate the difference using Monte Carlo; the second is to modify the bin parameters in order to make the variance of the Monte Carlo estimate of the integral the same for all bins. This allows us to use Student's t-test as a trigger for rebinning, which we claim is more stable than the χ^2 test that is commonly used for this purpose. We provide a program that we have used to study the algorithm for the case where the histogram is represented as a product of one-dimensional histograms. We discuss the assumptions and approximations made, as well as giving a pedagogical discussion of the myriad ways in which the results of any such Monte Carlo integration program can be misleading.

1. Introduction

We are interested in evaluating the integral of a function $f : \mathbb{R}^d \to \mathbb{R}$ over a compact domain. It is a simple matter to map any compact domain into the unit hypercube, so we need to evaluate

$$I = \int_{[0,1]^d} dx\, f(x) = \int_0^1 dx_1 \cdots \int_0^1 dx_d\, f(x_1, \ldots, x_d).$$

I is just the average value of f over a uniform probability distribution which vanishes outside the unit hypercube; we denote this average value by $\langle f \rangle$.

Defining the sample average $\bar{f}$ over a set of N uniformly distributed random points $\{x^{(i)} \in [0,1]^d, i = 1, \ldots, d\}$ to be

$$\bar{f} \equiv \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) \tag{1}$$

the weak law of large numbers [1] allows the identification $\langle f \rangle = \lim_{N\to\infty} \bar{f}$ assuming only that the integral exists. Strictly speaking the weak law of large numbers states that, with probability arbitrarily close to one, $\bar{f}$ will become arbitrarily close to $\langle f \rangle$ for sufficiently large N.

The central limit theorem makes the stronger statement that the probability distribution of $\bar{f}$ tends to a Gaussian with mean $\langle f \rangle$ and variance V/N,

$$\langle f \rangle = \bar{f} + \mathcal{O}\left(\sqrt{\frac{V}{N}}\right) \tag{2}$$

where the variance of the distribution of f values is

$$V \equiv \left\langle \left(f - \langle f \rangle\right)^2 \right\rangle = \int_{[0,1]^d} dx \left(f(x) - \langle f \rangle\right)^2 = \langle f^2 \rangle - \langle f \rangle^2,$$

but it requires stronger assumptions that we shall discuss shortly. An unbiased estimate of the variance is given by

$$\hat{V} \equiv \frac{\overline{f^2} - \bar{f}^2}{N-1}, \qquad \langle \hat{V} \rangle = V,$$

where of course

$$\overline{f^2} = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)})^2.$$

The estimate of the integral $\bar{f}$ is within one standard deviation $\sigma \equiv \sqrt{V/N}$ of the true value $I = \langle f \rangle$ about 68% of the time.

Note that the error is proportional to $1/\sqrt{N}$ independent of the dimension d of the integral. From this fact stems the great utility of Monte Carlo for integration in many dimensions compared to numerical quadrature. Generally for numerical quadrature (trapezoid rule, Simpson's rule, etc.) the error is $\mathcal{O}(\Delta^k)$ where Δ is the grid spacing and k^{-1} is the degree of the polynomial interpolation

* Corresponding author. Tel.: +45 28422030.
E-mail addresses: rudy.d.arthur@gmail.com (R. Arthur), adk@ph.ed.ac.uk (A.D. Kennedy).

0010-4655/$ – see front matter
http://dx.doi.org/10.1016/j.cpc.2013.08.003

between the grid points. With a fixed budget of function evaluations N on a regular grid each axis must be divided into $\sqrt[d]{N}$ segments. So $\Delta \propto N^{-1/d}$ and thus the error is $\mathcal{O}(N^{-k/d})$; therefore in dimension $d > 2k$ the Monte Carlo error scales better. An intuitive explanation for this is that a random sample is more homogeneous than a regular grid [2].

1.1. Singular integrands

If the integrand has a singularity within or on the boundary of the integration region extra care is required. Let us consider the proof of the central limit theorem. The probability distribution P_f for the values $F = f(x)$ when x is chosen from the distribution P is

$$P_f(F) \equiv \int dx\, P(x)\, \delta\Big(F - f(x)\Big) \tag{3}$$

for which

$$\int dF\, P_f(F) = \int dx\, P(x) = 1 \quad \text{and}$$

$$\int dF\, P_f(F)\, F = \int dx\, P(x) f(x) = \langle f\rangle.$$

We define the generating function for connected moments as the logarithm of the Fourier transform of P_f

$$W_f(ik) \equiv \ln \int dF\, P_f(F) e^{ikF} = \ln \int dx\, P(x) e^{ikf(x)} = \ln\langle e^{ikf}\rangle. \tag{4}$$

Assuming that (4) can be expanded in an asymptotic series

$$W_f(k) \sim \sum_{m=1}^{\infty} \frac{k^m C_m}{m!} \tag{5}$$

where coefficients C_m (cumulants)

$$C_0 = 1$$
$$C_1 = \langle f\rangle$$
$$C_2 = V = \left\langle \Big(f - \langle f\rangle\Big)^2 \right\rangle$$
$$C_3 = \left\langle \Big(f - \langle f\rangle\Big)^3 \right\rangle$$
$$C_4 = \left\langle \Big(f - \langle f\rangle\Big)^4 \right\rangle - 3C_2$$
$$C_5 = \left\langle \Big(f - \langle f\rangle\Big)^5 \right\rangle - 10C_3C_2$$
$$C_6 = \left\langle \Big(f - \langle f\rangle\Big)^6 \right\rangle - 15C_4C_2 - 10C_3^2 - 15C_2^2$$
$$\vdots$$

are all finite. We consider the distribution function $P_{\bar f}$ for the sample average $\bar f$ defined in Eq. (1)

$$P_{\bar f}(F) = \int dx^{(1)} \cdots dx^{(N)}\, P(x^{(1)}) \cdots P(x^{(N)}) \times \delta\left(F - \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\right)$$

and the corresponding generating function $W_{\bar f}$

$$W_{\bar f}(k) \equiv \ln \int dF\, P_{\bar f}(F)\, e^{ikF}. \tag{6}$$

Since the points $x^{(i)}$ were chosen independently, Eq. (6) factorises to give

$$W_{\bar f}(k) \sim \ln \left[\int dx\, P(x) e^{ikf(x)/N}\right]^N = NW_f\left(\frac{k}{N}\right),$$

which can be expanded as an asymptotic series in powers of $1/N$

$$W_{\bar f}(k) \sim \sum_{m=1}^{\infty} \frac{k^m C_m}{N^{m-1} m!} = kC_1 + \frac{k^2 C_2}{2N} + \frac{k^3 C_3}{6N^2} + \mathcal{O}\left(\frac{1}{N^3}\right).$$

Ignoring terms of $\mathcal{O}(N^{-2})$ and taking the inverse Fourier transform gives

$$P_{\bar f}(F) = \frac{\exp\left(\frac{-(F-\langle f\rangle)^2}{2V/N}\right)}{\sqrt{2\pi V/N}},$$

namely a Gaussian with mean $\langle f\rangle$ and variance V/N.

However, if any of the cumulants C_n are not finite then the expansion (5) is not meaningful. A simple class of integrals with divergent higher moments is $\int_0^1 dx\, x^\alpha$ with $\alpha < 0$. If $-1 < \alpha < 0$ this integral is well-defined but has an infinite number of divergent moments. Following Eq. (3)

$$P(y) = \int_\epsilon^1 dx\, \frac{\delta(y - x^\alpha)}{1-\epsilon} = \frac{y^{\frac{1}{\alpha}-1}}{\alpha(1-\epsilon)}$$

for $y \in [\epsilon^\alpha, 1]$, and zero elsewhere. The cumulants are combinations of the moments

$$\int dy\, P(y) y^n = \frac{1 - \epsilon^{1+\alpha n}}{(1-\epsilon)(1+n\alpha)}$$

which diverge as $\epsilon \to 0$ if $1 + \alpha n \le 0$, or equivalently every cumulant C_n with $n > -1/\alpha$ diverges. What this means in practice is that although estimating such integrals by Monte Carlo is allowed – the weak law of large numbers assures us that as long as N is large enough the estimate will converge but it gives no indication of how large N should be – it is misleading to estimate the error from the variance alone. This is because the distribution of the estimates of the integral is not Gaussian even if the variance exists. In practice one obtains non-Gaussian distributions with "fat tails"; for some examples see Fig. 1.

We may also observe that a singularity in the integrand does not necessarily lead to infinite cumulants: for the function $f(x) = -\ln x$ a similar analysis to that given above shows that $P_f(F) = e^{-F}$ for $F \in [0, \infty)$ and hence $\langle f^m\rangle = m!$ so all its cumulants are finite, and therefore the Monte Carlo estimates of the integral do have a Gaussian distribution as the number of samples $N \to \infty$.

For all convergent integrals Monte Carlo provides an estimate of the mean. However, if any moment diverges the distribution of the means is not Gaussian, and in general even if the variance exists it gives an underestimate of the "width" of the probability distribution. This is often the case in practice, for example when evaluating Feynman parameter integrals which are integrable but not square integrable. If the Monte Carlo integrator is treated as a black box the quoted error from the standard deviation will often be an underestimate of the true error. With these provisos on the applicability of Monte Carlo integration we now turn to our main topic, a new method of adaptive Monte Carlo integration.

2. Variance reduction

The Monte Carlo scheme of Section 1 (naïve Monte Carlo) can be improved by variance reduction schemes [3,4]: the basic idea is to use some information about the integral in order to reduce the variance sample average. We describe two methods: importance sampling and subtraction.