Contents

7

Other integration methods

Most of our Monte Carlo problems involve estimating expectations and these can often be written as integrals. Sometimes it pays to treat the problems as integrals and apply non-random numerical integration methods. In other settings we may be able to combine Monte Carlo and other methods into hybrid estimators. For instance, a nearly exact numerical integral of a problem related to our own, may be used as a control variate (described in §8.9). This chapter is specialized and may be skipped on first reading.

There is a large literature on numerical integration, also called quadrature. Here, we look at a few of the main ideas. Of special importance are the midpoint rule and Simpson's rule. They are simple to use and bring enormous improvements for smooth functions in low dimensions. We will also see how the advantage of classical quadrature methods decays rapidly with increasing dimension. This phenomenon is a manifestation of Bellman's 'curse of dimensionality', with Monte Carlo versions in two classic theorems of Bakhvalov.

The connection to integration problems is as follows. Suppose that our goal is to estimate $\mathbb{E}(g(\boldsymbol{Y}))$ where $\boldsymbol{Y} \in \mathbb{R}^s$ has probability density function p. We find a transformation function $\psi(\cdot)$, using methods like those in Chapter 5, such that $\boldsymbol{Y} = \psi(\boldsymbol{X}) \sim p$ when $\boldsymbol{X} \sim \mathbf{U}(0,1)^d$. Then

$$\mathbb{E}(g(\boldsymbol{Y})) = \int_{\mathbb{R}^s} g(\boldsymbol{y}) p(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} = \int_{(0,1)^d} g(\psi(\boldsymbol{x})) \,\mathrm{d}\boldsymbol{x} = \int_{(0,1)^d} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x},$$

where $f(\cdot) = g(\psi(\cdot))$. As a result our Monte Carlo problem can be transformed into a d-dimensional quadrature. We don't always have $d = s$. This method does not work when acceptance-rejection sampling is included in the way we generate $\boldsymbol{Y}$, because there is no a priori bound on the number of uniform random variables

that we would need. Since we're doing integration, we frame the problem via

$$I = \int_{(0,1)^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

When f has a simple closed form, there is always the possibility that I can found symbolically. Tools such as Mathematica™, Maple™, and sage can solve many integration problems. When they cannot, then we might turn to quadrature.

7.1 The midpoint rule

We start with a one-dimensional problem. Suppose that we want to estimate the integral

$$I = \int_a^b f(x)\,\mathrm{d}x$$

for $-\infty < a < b < \infty$.

The value of I is the area under the curve f over the interval $[a, b]$. It is easy to compute the area under a piecewise constant curve, and so it is natural to approximate f by a piecewise constant function $\hat{f}$ and then estimate I by $\hat{I} = \int_a^b \hat{f}(x)\,\mathrm{d}x$. We let $a = x_0 < x_1 < \cdots < x_n = b$ and then take t_i with $x_{i-1} \leqslant t_i \leqslant x_i$ for $i = 1, \ldots, n$, and put $\hat{f}(x) = f(t_i)$ whenever $x_{i-1} \leqslant x < x_i$. To complete the definition, take $\hat{f}(b) = f(b)$. Then

$$\hat{I} = \int_a^b \hat{f}(x)\,\mathrm{d}x = \sum_{i=1}^n (x_i - x_{i-1}) f(t_i).$$

If f is Riemann integrable on $[a, b]$ then $\hat{I} - I \to 0$ as $n \to \infty$ as long as $\max_{1\leqslant i\leqslant n}(x_i - x_{i-1}) \to 0$.

There is a lot of flexibility in choosing $\hat{f}$ but unless we have special knowledge about f we might as well use n equal intervals of length $(b-a)/n$ and take t_i in the middle of the i'th interval. This choice yields the ***midpoint rule***

$$\hat{I} = \frac{b-a}{n} \sum_{i=1}^n f\Big(a + (b-a)\frac{i-1/2}{n}\Big). \tag{7.1}$$

If we have constructed f so that $a = 0$ and $b = 1$ then the midpoint rule simplifies to

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n f\Big(\frac{i-1/2}{n}\Big).$$

For example, if $Z \sim \mathcal{N}(0,1)$ and we want to estimate $\mathbb{E}(g(Z)) = \int_{-\infty}^{\infty} g(z)\varphi(z)\,\mathrm{d}z$ then we can use

$$\frac{1}{n} \sum_{i=1}^n g\Big(\Phi^{-1}\Big(\frac{i-1/2}{n}\Big)\Big).$$

In so doing, we are using the midpoint rule to estimate $\int_0^1 f(x)\,\mathrm{d}x$ where $f(x) = g(\Phi^{-1}(x))$. For now we will suppose that g is a bounded function so that f is also bounded. We revisit the problem of unbounded integrands on page 7.

For a smooth function and large n, the midpoint rule attains a much better rate than Monte Carlo sampling.

Theorem 7.1. *Let $f(x)$ be a real valued function on $[a,b]$ for $-\infty < a < b < \infty$. Assume that the second derivative $f''(x)$ is continuous on $[a,b]$. Let $t_i = a + (b-a)(i-1/2)/n$ for $i = 1,\dots,n$. Then*

$$\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{n}\sum_{i=1}^n f(t_i)\right| \leqslant \frac{(b-a)^3}{24n^2}\max_{a\leqslant z\leqslant b}|f''(z)|.$$

Proof. For any x between $x_{i-1} \equiv t_i - (b-a)/(2n)$ and $x_i \equiv t_i + (b-a)/(2n)$, we can write $f(x) = f(t_i) + f'(t_i)(x-t_i) + (1/2)f''(z(x))(x-t_i)^2$ where $z(x)$ is a point between x and t_i. Let $\hat{I} = ((b-a)/n)\sum_{i=1}^n f(t_i)$. Then

$$\begin{aligned}
|I-\hat{I}| &= \left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{n}\sum_{i=1}^n f(t_i)\right| \\
&= \left|\sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) - f(t_i)\,dx\right| \\
&= \left|\sum_{i=1}^n \int_{x_{i-1}}^{x_i} f'(t_i)(x-t_i) + \frac{1}{2}f''(z(x))(x-t_i)^2\,dx\right| \\
&= \frac{1}{2}\left|\sum_{i=1}^n \int_{x_{i-1}}^{x_i} f''(z(x))(x-t_i)^2\,dx\right|.
\end{aligned}$$

Because f'' is continuous on $[a,b]$ and that interval is compact, f'' is absolutely continuous there and hence $M = \max_{a\leqslant x\leqslant b}|f''(x)| < \infty$. To complete the proof we write

$$|I-\hat{I}| \leqslant \frac{M}{2}\sum_{i=1}^n \int_{x_{i-1}}^{x_i}(x-t_i)^2\,dx = \frac{Mn}{2}\int_0^{x_1}(x-x_1/2)^2\,\mathrm{d}x \tag{7.2}$$

by symmetry. Then with $x_1 = (b-a)/n$,

$$\int_0^{x_1}(x-x_1/2)^2\,\mathrm{d}x = 2\int_0^{x_1/2} x^2\,\mathrm{d}x = \frac{2}{3}\left(\frac{x_1}{2}\right)^3 = \frac{(b-a)^3}{12n^3}. \tag{7.3}$$

The result follows by substituting (7.3) into (7.2). □

The midpoint rule is very simple to use and it works well on one-dimensional smooth functions. The rate $O(n^{-2})$ is much better than the $O(n^{-1/2})$ root mean square error (RMSE) from Monte Carlo. The proof in Theorem 7.1 is

fairly simple. A sharper analysis, in Davis and Rabinowitz (1984a, Chapter 4.3) shows that

$$\hat{I} - I = \frac{(b-a)^3}{24n^2} f''(\hat{z})$$

holds for some $\hat{z} \in (a, b)$, under the conditions of Theorem 7.1.

Error estimation is awkward for classical numerical integration rules. When $f''(x)$ is continuous on $[a, b]$ then the midpoint rule guarantees that $|\hat{I} - I| \leqslant (b-a)^3 M/(24n^2)$, where $M = \max_{a \leqslant z \leqslant b} |f''(z)|$. This looks like a 100% confidence interval. It would be, if we knew M, but unfortunately, we usually don't know M.

The midpoint rule is the integral of a very simple piecewise constant approximation to f. We could instead approximate f by a piecewise linear function over each interval $[x_{i-1}, x_i]$. If once again, we take equispaced values $x_i = a + i(b-a)/n$ we get the approximate function $\widetilde{f}$ that on the interval $[x_{i-1}, x_i]$ satisfies

$$\widetilde{f}(x) = f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}}\big(f(x_i) - f(x_{i-1})\big).$$

The integral of $\widetilde{f}(x)$ over $[a, b]$ yields the ***trapezoid rule***

$$\widetilde{I} = \frac{b-a}{n}\Big[\frac{1}{2}f(a) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}f(b)\Big].$$

The trapezoid rule is based on a piecewise linear approximation $\widetilde{f}$ to f instead of a piecewise constant one $\hat{f}$. For large n, the function $\widetilde{f}$ is usually much closer to f than $\hat{f}$ is, and so we might expect the trapezoid rule to attain a higher rate of convergence than the midpoint rule. But we would be wrong. Under the conditions of Theorem 7.1, $\widetilde{I} - I = -(b-a)^3 f''(\widetilde{z})/(12n^2)$ for some $\widetilde{z} \in (a, b)$ and so

$$|\widetilde{I} - I| \leqslant \frac{(b-a)^3}{12n^2} \max_{a \leqslant z \leqslant b} |f''(z)|.$$

The trapezoid rule integrates correctly any function f that is piecewise linear on each segment $[x_{i-1}, x_i]$, by using two evaluation points at the ends of the segments. The midpoint rule also integrates such a function correctly using just one point in the middle of each segment. The midpoint rule has benefitted from an error cancellation. This kind of cancellation plays a big role in the development of classical quadrature methods.

Another fact about these two methods is worth mentioning: the ***bracketing inequality***. Suppose that we know $f''(x) \geqslant 0$ for all $x \in [a, b]$. Then $\hat{I} - I = f''(\hat{z})(b-a)^3/(24n^2) \geqslant 0$ and $\widetilde{I} - I = -f''(\widetilde{z})(b-a)^3/(14n^2) \leqslant 0$ and therefore

$$\hat{I} \leqslant I \leqslant \widetilde{I}. \tag{7.4}$$

Equation (7.4) supplies us with a computable interval certain to contain the answer (when $f'' \geqslant 0$ is continuous on $[a, b]$), that is, a 100% confidence interval.

It is unusual to get such a good error estimate in a deterministic problem. Of course it requires knowledge that $f'' \geqslant 0$ everywhere. Naturally, if $f''(x) \leqslant 0$ is continuous, then the inequalities in (7.4) are reversed and we still get a bracketing.

Singularities at the endpoints

The midpoint rule has a big practical advantage over the trapezoid rule. It does not evaluate f at either endpoint a or b. Many of the integrals that we apply Monte Carlo methods to diverge to infinity at one or both endpoints. In such cases, the midpoint rule ***avoids the singularity***.

There are numerous mathematical techniques for removing singularities. These include change of variable transformations as well as methods that write $\int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int f_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int f_1(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where f_0 has a singularity but $\int f_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ is known, and f_1 has no singularity.

When we have no such analysis of our integrand, perhaps because it has a complicated problem-dependent formulation, or because we have hundreds of integrands to consider simultaneously, then avoiding the singularity is attractive. By contrast, the trapezoid rule does not avoid the endpoints $x = a$ and $x = b$. For such methods a second, less attractive principle is to ***ignore the singularity***, perhaps by using $f(x_i) = 0$ at any sample point x_i where f is singular. Davis and Rabinowitz (1984b, p. 180) remark that ignoring the singularity is a "tricky business and should be avoided where possible".

If $|f(x)| \to \infty$ as $x \to a$ or b then of course we won't have f'' bounded on $[a, b]$, and so we can no longer expect an error of $O(n^{-2})$. But the midpoint rule handles this singularity problem much more gracefully than the trapezoid rule does. See Lubinsky and Rabinowitz (1984) and references therein for information on convergence rates that are attained on integration problems containing singularities.

7.2 Simpson's rule

The midpoint and trapezoid rules are based on correctly integrating piecewise constant and linear approximations to the integrand. That idea extends naturally to methods that locally integrate higher order polynomials. The result is much more accurate integration, at least when the integrand is smooth.

As a next step, we find a three-point rule that correctly integrates any quadratic polynomial over $[0, 1]$. It is enough to correctly integrate 1, x and x^2. If we evaluate the function at points 0, 1/2 and 1 and use a rule of the form $w_1 f(0) + w_2 f(1/2) + w_3 f(1)$, the correct weights w_j can be found by solving

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 1/4 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \end{pmatrix}.$$

That is, we take

$$\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 1/4 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 1/6 \\ 2/3 \\ 1/6 \end{pmatrix}.$$

By a change of variable, this three-point rule can be adapted to the interval $[0, 2h]$ by taking

$$\int_0^{2h} f(x)\,\mathrm{d}x \doteq 2h\Big(\frac{1}{6}f(0) + \frac{2}{3}f(h) + \frac{1}{6}f(2h)\Big), \tag{7.5}$$

which is exact for quadratic functions f. Now we split the interval $[a, b]$ into $n/2$ panels of width $2h$ (so n has to be even) and apply equation (7.5) in each of them. Letting f_i be a shorthand notation for $f(a + ih)$ where $h = (b - a)/n$, the result is

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x &\doteq \frac{h}{3}\Big(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n\Big) \\ &= \frac{h}{3}\Big(f_0 + 4\sum_{j=1}^{n/2-1} f_{2j-1} + 2\sum_{j=1}^{n/2} f_{2j} + f_n\Big). \end{aligned} \tag{7.6}$$

Equation (7.5) is known as Simpson's rule. Equation (7.6) is a compound Simpson rule that is also called Simpson's rule. Equation (7.5) is exact for cubic functions, not just quadratics. As a result (7.6) is exact for a piecewise cubic approximation to f. If f has a continuous fourth derivative $f^{(4)}$ on $[a, b]$ then the error in Simpson's rule is

$$-\frac{(b-a)^4}{180\,n^4} f^{(4)}(z), \quad \text{for some } z \in (a, b).$$

7.3 Higher order rules

The idea behind Simpson's rule generalizes easily to higher orders. We split the interval $[a, b]$ into panels, find a rule that integrates a polynomial correctly within a panel, and then apply it within every panel to get a compound rule.

There are two main varieties of compound quadrature rule. For ***open rules*** we do not evaluate f at the end-points of the panel. The midpoint rule is open. For ***closed rules*** we do evaluate f at the end-points of the panel. The trapezoid rule and Simpson's rule are both closed. Closed rules have the advantage that some function evaluations get reused when we increase n. Open rules have a perhaps greater advantage that they avoid the ends of the interval where singularities often appear.

Newton-Cotes

The trapezoid rule and Simpson's rule use $m = 2$ and $m = 3$ points respectively within each panel. In general, one can use m points to integrate polynomials of degree $m-1$, to yield the Newton-Cotes formulas, of which the trapezoid rule and Simpson's rule are special cases. The Newton-Cotes rule for $m = 4$ is another of Simpson's rules, called Simpson's 3/8 rule. Newton-Cotes rules of odd order have the advantage that, by symmetry, they also correctly integrate polynomials of degree m, as we saw already in the case of Simpson's rule.

The next two odd order rules are

$$\int_0^{4h} f(x)\,\mathrm{d}x \doteq \frac{h}{45}\Big(14f_0 + 64f_1 + 24f_2 + 64f_3 + 14f_4\Big), \quad \text{and} \tag{7.7}$$
$$\int_0^{6h} f(x)\,\mathrm{d}x \doteq \frac{h}{140}\Big(41f_0 + 216f_1 + 27f_2 + 272f_3 + 27f_4 + 216f_5 + 41f_6\Big).$$

These are the 5 and 7 point closed Newton-Cotes formulas. Equation (7.7) is known as Boole's rule.

High order rules should be used with caution. They exploit high order smoothness in the integrand, but can give poor outcomes when the integrand is not as smooth as they require. In particular if a genuinely smooth quantity has some mild nonsmoothness in its numerical implementation f, then high order integration rules can behave very badly, magnifying this numerical noise.

As a further caution, Davis and Rabinowitz (1984a) note that taking f fixed and letting the order m in a Newton-Cotes formula increase does not always converge to the right answer even for f with infinitely many derivatives. Lower order rules applied in panels are more robust.

The Newton-Cotes rules can be made into compound rules similarly to the way Simpson's rule was compounded. When the basic method integrates polynomials of degree r exactly within panels, then the compound method has error $O(n^{-r})$, assuming that $f^{(r)}$ is continuous on $[a, b]$.

As noted above, open rules are valuable because they avoid the endpoints where the function may be singular. Here are a few open rules:

$$\int_0^{2h} f(x)\,\mathrm{d}x = 2hf(h) + \frac{h^3}{3}f''(z),$$
$$\int_0^{4h} f(x)\,\mathrm{d}x = \frac{4h}{3}\Big(2f(h) - f(2h) + 2f(3h)\Big) + \frac{14h^5}{45}f^{(4)}(z), \quad \text{and}$$
$$\int_0^{5h} f(x)\,\mathrm{d}x = \frac{5h}{24}\Big(11f(h) + f(2h) + f(3h) + 11f(4h)\Big) + \frac{95h^5}{144}f^{(4)}(z).$$

In each case the point z is inside the interval of integration, and the error term assumes that the indicated derivative is continuous. The first one is simply the midpoint rule after a change of variable to integrate over $[0, 2h]$. The next two are from Davis and Rabinowitz (1984a, Chapter 2.6). They both have the same order. The last one avoids negative weights but requires an extra point.

$\mathbf{w}(x)$	**Rule**
$\mathbb{1}_{\lvert x\rvert\leqslant 1}$	Gauss-Legendre
$\exp(-x^2)$	Gauss-Hermite
$\mathbb{1}_{0\leqslant x<\infty}\exp(-x)$	Gauss-Laguerre
$\mathbb{1}_{\lvert x\rvert<1}(1-x^2)^{-1/2}$	Gauss-Chebyshev (1st kind)
$\mathbb{1}_{\lvert x\rvert\leqslant 1}(1-x^2)^{1/2}$	Gauss-Chebyshev (2nd kind)

Table 7.1: This table lists some weight functions and the corresponding families of Gauss quadrature rules.

Gauss rules

The rules considered above evaluate f at equispaced points. A Gauss rule takes the more general form

$$\hat{I}_G = \sum_{i=1}^{m} w_i f(x_i)$$

where both x_i and w_i can be chosen to attain high accuracy.

The basic panel for a Gauss rule is conventionally $[-1,1]$ or sometimes $\mathbb{R}$, and not $[0,h]$ as we used for Newton-Cotes rules. Also the target integration problem is generally weighted. That is we seek to approximate

$$\int_{-\infty}^{\infty} f(x)w(x)\,\mathrm{d}x$$

for a weight function $w(x)\geqslant 0$. The widely used weight functions are multiples of standard probability density functions, such as the uniform, gamma, Gaussian and beta distributions; see Table 7.1. The idea is that having f be nearly a polynomial can be much more appropriate than requiring the whole integrand $f(x)w(x)$ to be nearly a polynomial.

Choosing w_i and x_i together yields $2m$ parameters and it is then possible to integrate polynomials of degree up to $2m-1$ without error. The error in a Gauss rule is

$$\frac{(b-a)^{2m+1}(m!)^4}{(2m+1)[(2m)!]^3} f^{(2m)}(z)$$

where $a<z<b$, provided that f has $2m$ continuous derivatives. Unlike Newton-Cotes rules, Gauss rules of high order have non-negative weights. We could in principle use a very large m. For the uniform weighting $w(x)=1$ though, we could also break the region into panels. Then for n function evaluations the error will be $O(n^{-2m})$ assuming as usual that $f^{(2m)}$ is continuous on $[a,b]$. Gauss rules for uniform weights on $[-1,1]$ have the advantage that they can be used within panels. Several are listed in Table 7.2.

m	$\mathbf{x}_i$	$\mathbf{w}_i$
2	$\pm\frac{1}{\sqrt{3}}$	1
3	0	$\frac{8}{9}$
	$\pm\frac{1}{5}\sqrt{15}$	$\frac{5}{9}$
4	$\pm\frac{1}{35}\sqrt{525-70\sqrt{30}}$	$\frac{1}{36}\left(18+\sqrt{30}\right)$
	$\pm\frac{1}{35}\sqrt{525+70\sqrt{30}}$	$\frac{1}{36}\left(18-\sqrt{30}\right)$
5	0	$\frac{128}{225}$
	$\pm\frac{1}{21}\sqrt{245-14\sqrt{70}}$	$\frac{1}{900}\left(322+13\sqrt{70}\right)$
	$\pm\frac{1}{21}\sqrt{245+14\sqrt{70}}$	$\frac{1}{900}\left(322-13\sqrt{70}\right)$

Table 7.2: Gauss quadrature rules $\sum_{i=1}^m w_i f(x_i)$ to approximate $\int_{-1}^1 f(x)\,\mathrm{d}x$, for $m = 1,\dots,5$. From Weisstein, Eric W, "Legendre-Gauss Quadrature." MathWorld web site.—A Wolfram Web Resource. http://mathworld.wolfram.com/Legendre-GaussQuadrature.html

7.4 Fubini and the curse of dimensionality

Classical quadrature methods are very well tuned to one-dimensional problems with smooth integrands. A natural way to extend them to multi-dimensional problems is to write them as iterated one-dimensional integrals, via Fubini's theorem. When we estimate each of those one-dimensional integrals by a quadrature rule, we end up with a set of sample points on a multi-dimensional grid. Unfortunately, there is a curse of dimensionality that severely limits the accuracy of this approach.

To see informally what goes wrong, let $f(x,y)$ be a function on $[0,1]^2$. Now write

$$I = \int_0^1\int_0^1 f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 I(y)\,\mathrm{d}y$$

where $I(y) = \int_0^1 f(x,y)\,\mathrm{d}x$. Suppose that we use the same m point integration rule with weight u_i on point x_i, for any value of y, getting

$$\hat{I}(y) = \sum_{i=1}^m u_i f(x_i,y) = I(y) + E_1(y),$$

where $|E_1(y)| \leqslant Am^{-r}$ holds for some $A < \infty$ and all $0 \leqslant y \leqslant 1$. The value of r depends on the method we use and how smooth f is. Next we use an n point rule to average over y getting

$$\hat{I} = \sum_{j=1}^{n} v_j \hat{I}(y_j) = \sum_{j=1}^{n} v_j \big(I(y_j) + E_1(y_j)\big) = \int_0^1 I(y)\,\mathrm{d}y + E_2 + \sum_{j=1}^{n} v_j E_1(y_j)$$

where $|E_2| \leqslant Bn^{-s}$ for some $B < \infty$ and s depending on the outer integration rule and on how smooth $I(y)$ is.

The total error is a weighted sum of errors E_1 at different points y_j plus the error E_2. We suppose that the weighted sum of $E_1(y_j)$ is $O(m^{-r})$. This happens if $\sum_{j=1}^{n} |v_j| < C$ holds for all n because we assumed that $|E_1(y)| \leqslant Am^{-r}$ holds with the same $A < \infty$ for all $y \in [0,1]$.

The result is that $|\hat{I} - I| = O(m^{-r} + n^{-s})$. In other words, the convergence rate that we get is like the worst of the two one-dimensional rules that we combine.

More generally, if we are using a d-dimensional grid of points and a product rule

$$\hat{I} = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \Big(\prod_{j=1}^{d} w_{ji_j} \Big) f(x_{i_1}, x_{i_2}, \ldots, x_{i_d})$$

we cannot expect the result to be better than what we would get from the worst of the rules we have used. Suppose that rule j is the least accurate. Then $\hat{I}$ could hardly be better than

$$\sum_{i=1}^{n_j} w_{ji} I_j(x_{ji})$$

where $I_j(x_j)$ is the (exact) integral of f over all variables other than x_j.

Getting the worst one-dimensional rate leads to a curse of dimensionality. Suppose that we use the same n point one-dimensional quadrature rule on each of d dimensions. As a result we use $N = n^d$ function evaluations. If the one-dimensional rule has error $O(n^{-r})$, then the combined rule has error

$$|\hat{I} - I| = O(n^{-r}) = O(N^{-r/d}).$$

Even modestly large d will give a bad result. The value r is the smaller of the number of continuous derivatives f has and the number that the quadrature rule is able to exploit. Taking $r \gg d$ won't help in practice, because high order rules get very cumbersome and many of them are prone to roundoff errors.

This curse of dimensionality is not confined to sampling on grids formed as products of one-dimensional rules. Any quadrature rule in high dimensions will suffer from the same problem. Two important theorems of Bakhvalov, below, make the point.

Theorem 7.2 (Bakhvalov I)**.** *For $0 < M < \infty$ and integer $r \geqslant 1$, let*

$$C_M^r = \left\{ f : [0,1]^d \to \mathbb{R} \mid \left| \frac{\partial f(\boldsymbol{x})}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right| \leqslant M, \quad \forall\, \alpha_j \geqslant 0 \text{ with } \sum_{j=1}^{d} \alpha_j = r \right\}.$$

Then there exists $k > 0$ such that for any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in [0,1]^d$ and any $w_1, \dots, w_n \in \mathbb{R}$, there is an $f \in C_M^r$ with

$$\left| \sum_{i=1}^n w_i f(\boldsymbol{x}_i) - \int_{[0,1]^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| \geqslant k n^{-r/d}$$

Proof. This is given as Theorem 3.1 in Dimov (2008). □

Bakhvalov's theorem makes high-dimensional quadrature seem intractable. There is no way to beat the rate $O(n^{-r/d})$, no matter where you put your sampling points $\boldsymbol{x}_i$ or how cleverly you weight them. At first, this result looks surprising, because we have been using Monte Carlo methods which get an RMSE $O(n^{-1/2})$ in any dimension. The explanation is that in Monte Carlo sampling we pick one single function $f(\cdot)$ with finite variance σ^2 and then in sampling n uniform random points, get an RMSE of $\sigma n^{-1/2}$ for the estimate of that function's integral. Bahkvalov's theorem works in the opposite order. We pick our points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, and their weights w_i. Then Bakhvalov finds a function f with r derivatives on which our rule makes a large error.

When we take a Monte Carlo sample, there is always some smooth function for which we would have got a very bad answer. Such worst case analysis is very pessimistic because the worst case functions could behave very oddly right near our sampled $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, and the worst case functions might look nothing like the ones we are trying to integrate.

Bakhvalov has a counterpart to Theorem 7.2 which describes random sampling. Whatever way we choose to sample our input points, there exists a smooth function with a large RMSE:

Theorem 7.3 (Bakhvalov II)**.** *For $0 < M < \infty$ and integer $r \geqslant 1$, let C_M^r be as given in Theorem 7.2. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be random elements of $[0,1]^d$ and let $w_1, \dots, w_n \in \mathbb{R}$. Then there exists $k > 0$ such that some function $f \in C_M^r$ satisfies*

$$\mathbb{E}\left(\left(\sum_{i=1}^n w_i f(\boldsymbol{X}_i) - \int_{[0,1]^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right)^2 \right)^{1/2} \geqslant k n^{-1/2-r/d}.$$

Proof. This is given as Theorem 3.2 of Dimov (2008). □

In Theorem 7.3, the worst case function f is chosen knowing how we will sample $\boldsymbol{X}_i$, but not knowing the resulting values $\boldsymbol{x}_i$ that we will actually use. Here we see an RMSE of at least $O(n^{-1/2-r/d})$ which does not contradict the MC rate.

7.5 Hybrids with Monte Carlo

We can hybridize quadrature and Monte Carlo methods by using each of them on some of the variables. For example, to approximate

$$I = \int_{[0,1]^d} \int_{[0,1]^s} f(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}$$

we might take $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in [0,1]^s$, $w_1, \dots, w_m \in \mathbb{R}$ and draw $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n \sim \mathbf{U}(0,1)^d$ independently. Then the hybrid estimate is

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_j f(\boldsymbol{x}_j, \boldsymbol{Y}_i). \tag{7.8}$$

Our hybrid has a curse of dimension driven by the size of s and it has variance σ^2/n where

$$\sigma^2 = \mathrm{Var}\biggl(\sum_{j=1}^{m} w_j f(\boldsymbol{x}_j, \boldsymbol{Y})\biggr) = \sum_{j=1}^{m}\sum_{j'=1}^{m} w_j w_{j'} \mathrm{Cov}(f(\boldsymbol{x}_j, \boldsymbol{Y}), f(\boldsymbol{x}_{j'}, \boldsymbol{Y})).$$

We might expect this hybrid to be useful when s is not too large and $f(\boldsymbol{x}, \boldsymbol{y})$ is very smooth in $\boldsymbol{x}$. Then the inner sum in (7.8) is well handled by quadrature. If additionally, the f has large variance in $\boldsymbol{X}$ for fixed $\boldsymbol{Y} = \boldsymbol{y}$ then our quadrature may be much better than using Monte Carlo on both $\boldsymbol{x}$ and $\boldsymbol{y}$.

If we could integrate out $\boldsymbol{x}$ in closed form then we could use the estimate $\hat{I} = (1/n)\sum_{i=1}^{n} h(\boldsymbol{Y}_i)$ where $h(\boldsymbol{y}) = \mathbb{E}(f(\boldsymbol{X}, \boldsymbol{Y}) \,|\, \boldsymbol{Y} = \boldsymbol{y})$ for $\boldsymbol{X} \sim \mathbf{U}(0,1)^d$. This is the method called conditioning in §8.7. The hybrid (7.8) is conditioning with a numerical approximation to h.

Hybrids of Monte Carlo and quasi-Monte Carlo methods are often used. See Chapter 17. They take the form $\hat{I} = (1/n)\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{Y}_i)$ for a quadrature rule with $\boldsymbol{x}_i \in [0,1]^s$ and Monte Carlo samples $\boldsymbol{Y}_i \sim \mathbf{U}(0,1)^d$.

7.6 Additional methods

There are many other approaches to approximating an integral.

Laplace approximations

The Laplace approximation is a classical device for approximate integration. Suppose that we seek to estimate $I = \int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $\log(f(\boldsymbol{x}))$ has Taylor expansion $b + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^{\mathsf{T}} H (\boldsymbol{x}-\boldsymbol{x}')$ around its maximizer $\boldsymbol{x}'$, where H is the Hessian matrix for $\log(f(\cdot))$ at $\boldsymbol{x} = \boldsymbol{x}'$. If H is negative definite, then $\Sigma = -H^{-1}$ is a convariance matrix and we may write

$$I \simeq e^b \int e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{x}')}\,\mathrm{d}\boldsymbol{x} = e^b (2\pi)^{d/2} |\Sigma|^{-1/2} \tag{7.9}$$

using the known normalization of the $\mathcal{N}(\boldsymbol{x}', \Sigma)$ distribution. Note that $|\Sigma| = |-H|$ which is the absolute value of $|H|$.

The Laplace approximation is very accurate when $\log(f(\boldsymbol{x}))$ is smooth and the quadratic approximation is good where f is not negligible. Such a phenomenon often happens when $\boldsymbol{x}$ is a statistical parameter subject to the central limit theorem, $f(\boldsymbol{x})$ is its posterior distribution, and the sample size is large enough for the CLT to apply.

If we want the integral of $g(\boldsymbol{x})f(\boldsymbol{x})$ for smooth g, then we may multiply (7.9) by $g(\boldsymbol{x}')$. The following theorem gives the details:

Theorem 7.4. *The asymptotic equivalence*

$$\int_A g(\boldsymbol{x})e^{-\lambda h(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \sim g(\boldsymbol{x}')(2\pi)^{d/2}|\lambda H(\boldsymbol{x}')|^{-1/2}e^{-\lambda h(\boldsymbol{x}')}$$

holds as $\lambda \to \infty$, if

i) $A \subseteq \mathbb{R}^d$ is open,

ii) $\int_A |g(\boldsymbol{x})|e^{-\lambda h(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} < \infty$ for all $\lambda \geqslant \lambda_0$ for some $\lambda_0 < \infty$,

iii) there exists $\boldsymbol{x}' \in A$ such that for all $\epsilon > 0$

$$\inf\{h(\boldsymbol{x}) - h(\boldsymbol{x}') \mid \boldsymbol{x} \in A, \|\boldsymbol{x} - \boldsymbol{x}'\| > \epsilon\} > 0,$$

iv) g is continuous in a neighborhood of $\boldsymbol{x}'$ with $g(\boldsymbol{x}') \neq 0$,

v) h has two continuous derivatives on A and $H(\boldsymbol{x}') \equiv \partial^2 h(\boldsymbol{x})/\partial\boldsymbol{x}\partial\boldsymbol{x}^{\mathsf{T}}$ is positive definite.

Proof. This is Theorem 4.14 of Evans and Swartz (2000) which is based on a result in Wong (1989). □

The parameter λ is a measure of how strongly spiked f is. In statistical applications λ is often the number n of observations in a sample. Similarly the argument $\boldsymbol{x}$ is usually a parameter vector θ in statistical applications, and f is a distribution for θ. For the rest of this discussion we switch to the notation used for Bayesian statistical applications.

Suppose that a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ has prior distribution $\pi_0(\theta)$, and the data are independent draws from the density or mass function $p(\boldsymbol{x};\theta)$. We will use $\mathcal{X}$ to represent the full set of observations $\boldsymbol{X}_i = \boldsymbol{x}_i \in \mathbb{R}^d$, for $i = 1, \ldots, n$. The log likelihood function is

$$\ell(\theta) = \ell(\theta;\mathcal{X}) = \sum_{i=1}^n \log(p(\boldsymbol{x}_i;\theta))$$

and so the posterior distribution of θ given $\mathcal{X}$ is

$$\pi(\theta) = \pi(\theta \mid \mathcal{X}) \propto \pi_0(\theta)e^{\ell(\theta)}.$$

The value $\hat\theta = \arg\max_\theta \ell(\theta)$ is the maximum likelihood estimate of θ. When $\hat\theta$ is interior to Θ and ℓ is smooth, then we have the Taylor approximation $\ell(\theta) \doteq \ell(\hat\theta) - \frac{1}{2}(\theta - \hat\theta)^{\mathsf{T}}H(\hat\theta)(\theta - \hat\theta)$ where

$$H(\theta) = -\sum_{i=1}^n \frac{\partial^2 \log(p(\boldsymbol{x}_i;\theta))}{\partial\theta\partial\theta^{\mathsf{T}}} \equiv n\hat I(\theta).$$

The quantity $\hat{I}(\hat{\theta})$ is known as the empirical Fisher information for θ. Because H is a sum of n independent and identically distributed terms, we find by the law of large numbers that $H(\theta) \approx -n\,\mathbb{E}\left(\partial^2 \log(p(\boldsymbol{X};\theta))/\partial\theta\partial\theta^{\mathsf{T}}\right) \equiv nI(\theta)$ where $I(\theta)$ is the ordinary (non-empirical) Fisher information for θ. In asymptotic expansions, it is most convenient to work with $nI(\hat{\theta})$, but to present the method we work with $H(\hat{\theta})$.

The best predictor of $g(\theta)$, with respect to squared error, is

$$\mathbb{E}(g(\theta) \mid \mathcal{X}) = \frac{\int g(\theta)\pi_0(\theta)e^{\ell(\theta)}\,\mathrm{d}\theta}{\int \pi_0(\theta)e^{\ell(\theta)}\,\mathrm{d}\theta},$$

which we may write as

$$\frac{\int g_{\mathrm{N}}(\theta)e^{-H_{\mathrm{N}}(\theta)}\,\mathrm{d}\theta}{\int g_{\mathrm{D}}(\theta)e^{-H_{\mathrm{D}}(\theta)}\,\mathrm{d}\theta} \tag{7.10}$$

where, for example, $g_{\mathrm{N}}(\theta)$ is either 1 or $g(\theta)$ or $g(\theta)\pi_0(\theta)$ with $H_{\mathrm{N}}(\theta)$ adjusted accordingly, and similar choices are available for the denominator. We investigate several specific formulations of (7.10) below.

If we take $\hat{\theta}_{\mathrm{N}}$ and $\hat{\theta}_{\mathrm{D}}$ to be the minimizers of H_{N} and H_{D} respectively, then a straightforward use of the Laplace approximation yields

$$\widehat{\mathbb{E}}(g(\theta) \mid \mathcal{X}) = \frac{g_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})|\bar{H}_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})|^{-1/2}e^{-H_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})}}{g_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})|\bar{H}_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})|^{-1/2}e^{-H_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})}} \tag{7.11}$$

where $\bar{H}_{\mathrm{N}}$ and $\bar{H}_{\mathrm{D}}$ are the Hessian matrices of H_{N} and H_{D} respectively.

The Laplace approximation is said to be in ***standard form*** when $H_{\mathrm{N}}(\theta) = H_{\mathrm{D}}(\theta)$. In the standard form, the estimate (7.11) simplifies to

$$\widehat{\mathbb{E}}(g(\theta) \mid \mathcal{X}) = \frac{g_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})}{g_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})} = \frac{g_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})}{g_{\mathrm{D}}(\hat{\theta}_{\mathrm{N}})} \tag{7.12}$$

because $\hat{\theta}_{\mathrm{D}} = \hat{\theta}_{\mathrm{N}}$ in the standard form. If we take $H_{\mathrm{N}}(\theta) = H_{\mathrm{D}}(\theta) = \ell(\theta)$, so that $g_{\mathrm{N}}(\theta) = g(\theta)\pi_0(\theta)$ and $g_{\mathrm{D}}(\theta) = \pi_0(\theta)$, we obtain $\widehat{\mathbb{E}}(g(\theta) \mid \mathcal{X}) = g(\hat{\theta})$ where $\hat{\theta}$ is the maximum likelihood estimator of θ. If instead we take $H_{\mathrm{N}}(\theta) = \ell(\theta) - \log(\pi_0(\theta))$, with $g_{\mathrm{N}}(\theta) = g(\theta)$ and $g_{\mathrm{D}}(\theta) = 1$, then we obtain $\widehat{\mathbb{E}}(g(\theta) \mid \mathcal{X}) = g(\widetilde{\theta})$, where $\widetilde{\theta}$ is the maximum a posteriori (MAP) estimate of θ.

The Laplace approximation is in ***fully exponential form*** when $g_{\mathrm{N}}(\theta) = g_{\mathrm{D}}(\theta)$. For example, we might have $g_{\mathrm{N}}(\theta) = g_{\mathrm{D}}(\theta) = 1$ with $H_{\mathrm{N}}(\theta) = \ell(\theta) - \log(\theta) - \log(g(\theta))$ and $H_{\mathrm{D}}(\theta) = \ell(\theta) - \log(\theta)$. The fully exponential form requires $g(\theta) > 0$. In the fully exponential form, the estimate (7.11) becomes

$$\widehat{\mathbb{E}}(g(\theta) \mid \mathcal{X}) = \frac{|\bar{H}_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})|^{-1/2}e^{-H_{\mathrm{N}}(\hat{\theta}_{\mathrm{N}})}}{|\bar{H}_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})|^{-1/2}e^{-H_{\mathrm{D}}(\hat{\theta}_{\mathrm{D}})}}. \tag{7.13}$$

It now requires two separate optimizations and the determinants of two different Hessian matrices. In return for this extra work, the method attains higher

accuracy. While the standard form has errors of order n^{-1}, the exponential form has errors of order n^{-2} (Tierney and Kadane, 1986). This extra accuracy arises because the numerator and denominator estimate their corresponding quantities with very nearly the same relative error, and so much of that error cancels.

It is possible to attain an error of order n^{-2} from the standard form. To do so, we replace the numerator and denominator in the right hand side of (7.11) by a Taylor expansion taken as far as third order mixed partial derivatives of H_{N} and H_{D}. See Evans and Swartz (2000) for a formula. This process is less convenient than the fully exponential one, especially for a large number p of parameters in θ.

The positivity constraint on $g(\theta)$ from the fully exponential form is a nuisance. Tierney et al. (1989) consider several ways around it. One way is to replace $g(\theta)$ by $g(\theta) + c$ for some $c > 0$, apply the fully exponential approximation, subtract c from the resulting estimate, and use the limit of this process as $c \to \infty$. Another is to work with the moment generating function $M(t) \equiv \mathbb{E}(e^{tg(\theta)} \,|\, \mathcal{X})$. When $M(t)$ exists we can estimate it using the fully exponential form because $e^{tg(\theta)} > 0$. Then $\mathbb{E}(g(\theta) \,|\, \mathcal{X}) = M'(0)$ which can be estimated numerically.

The Laplace approximation is now overshadowed by MCMC. One reason is that the Laplace approximation is designed for unimodal functions. When $\pi(\theta \,|\, \mathcal{X})$ has two or more important modes, then the space Θ can perhaps be cut into pieces containing one mode each, and Laplace approximations applied separately and combined, but such a process can be cumbersome. MCMC by contrast is designed to find and sample from multiple modes, although on some problems it will have difficulty doing so. The Laplace approximation also requires finding the optimum of a d-dimensional function and working with the Hessian at the mode. In some settings that optimization may be difficult, and when d is extremely large, then finding the determinant of the Hessian can be a challenge. Finally, posterior distributions that are discrete or are mixtures of continuous and discrete parts can be handled by MCMC but are not suitable for the Laplace approximation.

The Laplace approximation is not completely superceded by MCMC. In particular, the fully exponential version is very accurate for problems with modest dimension d and large n. When the optimization problem is tractable then it may provide a much more automatic and fast answer than MCMC does.

Chapter end notes

Davis and Rabinowitz (1984b, Chapters 2–4) have a very comprehensive discussion of one-dimensional quadrature rules. The emphasis is on problems where one can obtain highly accurate answers. They give special attention to integration over unbounded intervals and to integrands that oscillate. Their Chapter 2.12 covers unbounded integrands over bounded intervals. Evans and Swartz (2000) describe many multivariate quadrature methods.

Press et al. (2007) present some adaptive quadrature methods. The idea is to take a rule like the midpoint rule and use it with a higher density of points in some subintervals than others. For example subintervals where the function varies more (having a larger $|f''|$) get more points while other intervals get fewer points. The information on where the function varies most is gathered along with the function values. They also present some multivariate adaptive quadrature methods.

There is some mild controversy about the use of adaptive methods. There are theoretical results showing that adaptive methods cannot improve significantly over non-adaptive ones. There are also theoretical and empirical results showing that adaptive methods may do much better than non-adaptive ones. These results are not contradictory, because they make different assumptions about the problem. For a high level survey of when adaptation helps, see Novak (1996).

Exercises

7.1. A test for Churg-Strauss syndrome (Masi et al., 1990) correctly detects it in 99.7% of affected patients. The test falsely reports Churg-Strauss in 15% of patients without the disease. Suppose that we sample m people at a clinic and x of them test positive for Churg-Strauss. We are interested in the fraction $p \in (0, 1)$ of visitors to the clinic that really have Churg-Strauss. For a uniform prior distribution on p the posterior distribution of p is proportional to

$$\pi_u(p \,|\, x) = (p \times 0.997 + (1-p) \times 0.15)^x ((1-p) \times 0.85 + p \times 0.003)^{m-x}.$$

The clinic finds that $x = 10$ out of $m = 30$ patients test positive.

a) Use the midpoint rule with $n = 1000$ to estimate the posterior mean of p given the data:

$$\mathbb{E}(p \,|\, x) = \frac{\int_0^1 \pi_u(p \,|\, x) p \,\mathrm{d}p}{\int_0^1 \pi_u(p \,|\, x) \,\mathrm{d}p}.$$

b) Use the midpoint rule with $n = 1000$ to estimate the posterior variance

$$\mathrm{Var}(p \,|\, x) = \frac{\int_0^1 \pi_u(p \,|\, x)(p - \mathbb{E}(p \,|\, x))^2 \,\mathrm{d}p}{\int_0^1 \pi_u(p \,|\, x) \,\mathrm{d}p}.$$

c) One third of the patients tested positive. Use the midpoint rule with $n = 1000$ to estimate the posterior probability that $p \leqslant 1/3$,

$$\mathbb{P}(p \leqslant 1/3 \,|\, x) = \frac{\int_0^{1/3} \pi_u(p \,|\, x) \,\mathrm{d}p}{\int_0^1 \pi_u(p \,|\, x) \,\mathrm{d}p}.$$

7.2. Our theoretical understanding of the midpoint rule suggests that the error in the first two parts of Exercise 7.1 should decrease as $O(1/n^2)$. The third part should not attain this rate because $f(p) = \mathbb{1}_{p\leqslant 1/3}$ is not twice differentiable. Use the midpoint rule with $n = 10^6$ as if it were the exact answer. Then plot the absolute error versus n of the midpoint rule for $n = 10^j$ for $j = 1, 2, 3, 4, 5$, for all three of the parts of Exercise 7.1. Does the predicted n^{-2} rate hold for the first two parts? What rate appears to hold in the third part?

7.3. Solve the system of equations

$$\int_0^n x^p \,\mathrm{d}x = \sum_{i=0}^{n} w_{in} i^n, \quad p = 0, \ldots, n$$

for w_{in}, for $n = 1, 2, 3, 4, 5, 6$. Use your results to give the next symmetric rule after Bode's rule.

7.4. Mike uses the midpoint rule with n points to approximate $\int_0^1 f(x)\,\mathrm{d}x$ by $\hat{I}_n$, and Trish uses the trapezoid rule with the same intervals on the same problem to get $\widetilde{I}_n$.

a) How many function values does Trish use?

b) How many distinct function values did the two people need?

c) Describe how they could combined their data to fit a larger midpoint rule.

d) If f is very smooth, do you expect differences in accuracy among choices $(\hat{I} + \widetilde{I})/2$, $(2\hat{I} - \widetilde{I})$ and $(2\widetilde{I} - \hat{I})$?

7.5. Verify that equation (7.5) is correct for $f(x) = x^3$. Show that it fails for $f(x) = x^4$.

Bibliography

Davis, P. J. and Rabinowitz, P. (1984a). *Methods of Numerical Integration*. Academic Press, San Diego, 2nd edition.

Davis, P. J. and Rabinowitz, P. (1984b). *Methods of Numerical Integration*. Academic Press, San Diego, 2nd edition.

Dimov, I. T. (2008). *Monte Carlo methods for applied scientists*. World Scientific, Singapore.

Evans, M. J. and Swartz, T. (2000). *Approximating integrals by Monte Carlo and deterministic methods*. Oxford University Press, Oxford.

Lubinsky, D. S. and Rabinowitz, P. (1984). Rates of convergence of Gaussian quadrature for singular integrands. *Mathematics of Computation*, 43(167):219–242.

Masi, A. T., Hunder, G. G., Lie, J. T., Michel, B. A., Bloch, D. A., Arend, W. P., Calabrese, L. H., Edworthy, S. M., Fauci, A. S., Leavitt, R. Y., Lightfoot Jr., R. W., McShane, D. J., Mills, J. A., Stevens, M. B., Wallace, S. L., and Zvaifler, N. J. (1990). The American College of Rheumatology 1990 criteria for the classification of Churg-Strauss syndrome (allergic granulomatosis and angiitis). *Arthritis & Rheumatism*, 33(8):1094–1100.

Novak, E. (1996). On the power of adaption. *Journal of Complexity*, 12(3):199–238.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical recipes: the art of scientific computing*. Cambridge University Press, Cambridge, 3rd edition.

Tierney, L. and Kadane, J. B. (1986). Accurate approximateions for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

Wong, R. (1989). *Asymptotic approximation of integrals*. Academic Press, San Diego.