Numerical Calculus II: Integration

Anthony Lee

December 2024

Outline

Outline

Introduction

- Integration is involved in, for example:
  - integrating random effects out of a joint distribution to get a likelihood,
  - evaluating expectations, including posterior expectations in Bayesian inference.
- Unfortunately, integration is typically intractable and accurate approximations are often computationally expensive.
- There is no simple rule for obtaining the integral of a composition of functions, cf. differentiation and the chain rule.
- We will look at two approaches:
  - one more classical, for integrating 1D functions and,
  - one based on viewing integrals as expectations of random variables.

Outline

Integrals

- The definite integral of a function f over the interval (a, b) is

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

- Our goal in quadrature is to approximate this integral with a sum

$$\sum_{i=1}^{N} w_i f(x_i),$$

  for some choice of $\{(x_i, w_i) : i \in \{1, \ldots, N\}\}$.
- How can we come up with some suitable points and weights?

Polynomial approximations

- Key high-level idea is that if we approximate f with a polynomial p then we can compute

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

  as an approximation of $I(f)$, since polynomials can be integrated exactly.
- One particularly simple choice is, with k points $x_1, \ldots, x_k$, to use an interpolating polynomial of degree at most $k-1$.
- The interpolating polynomial is unique and it is convenient to express it as a Lagrange polynomial:

$$p_{k-1}(x) := \sum_{i=1}^{k} \ell_i(x) f(x_i),$$

  where the Lagrange basis polynomials are

$$\ell_i(x) = \prod_{j=1, j\neq i}^{k} \frac{x - x_j}{x_i - x_j} \qquad i \in \{1, \ldots, k\}.$$

Example interpolating polynomials

- Course website

Piecewise interpolating polynomials

- For finite k, the error $\|f - p_{k-1}\|$ may be large.
- Options: increase k or split the domain into subintervals.
  - Piecewise polynomial approximation.
- Simple version: split into m subintervals and use k equally spaced points in each subinterval.
  - Closed if we put points at the interval boundaries. Open if we don't.
- Course website again.

Integrating the polynomials

- Consider integrating an interpolating polynomial p_{k-1}:

$$\begin{aligned} I(p_{k-1}) &= \int_a^b p_{k-1}(x)\mathrm{d}x \\ &= \int_a^b \sum_{i=1}^{k} \ell_i(x) f(x_i)\mathrm{d}x \\ &= \sum_{i=1}^{k} f(x_i) \int_a^b \ell_i(x)\mathrm{d}x \\ &= \sum_{i=1}^{k} w_i f(x_i), \end{aligned}$$

  where $w_i := \int_a^b \ell_i(x)\mathrm{d}x$ and we recall $\ell_i(x) = \prod_{j=1, j\neq i}^{k} \frac{x-x_j}{x_i-x_j}$.

- The ℓ_i can be integrated analytically by hand / ahead of time.

A note on the domain

- For constants $a < b$ and $c < d$, we can accommodate a change of finite interval via

$$\int_a^b f(x)\mathrm{d}x = \int_c^d g(y)\mathrm{d}y,$$

  by defining

$$g(y) := \frac{b-a}{d-c} f\left(a + \frac{b-a}{d-c}(y-c)\right).$$

- One can also accommodate integrating over $(0, \infty)$ or $(-\infty, \infty)$ by a similar change of variables.
- Idea: just sort out how to integrate over $(-1, 1)$ or $(0, 1)$.
- Map the problem to this domain if you have some other domain.

Some examples

- $k = 1$, closed: $I(p_0) = (b - a)f(a)$.
- $k = 1$, open: $I(p_0) = (b - a)f(\frac{a+b}{2})$.
- $k = 2$, closed: $I(p_1) = \frac{b-a}{2}\{f(a) + f(b)\}$.
- $k = 3$, closed: $I(p_2) = \frac{b-a}{6}\left\{f(a) + 4f(\frac{a+b}{2}) + f(b)\right\}$.
- Integration error bounds depend on the derivative $f^{(k+1)}$ on (a, b).
- The bounds get really very good for large k, for sufficiently smooth functions.

Composite rules

- As with interpolation, we often split into subintervals.
- If $A_1, \ldots, A_m$ partition (a, b)

$$I(f) = \int_a^b f(x)\mathrm{d}x = \sum_{i=1}^{m} \int f(x) \cdot 1_{A_i}(x)\mathrm{d}x = \sum_{i=1}^{m} I(f_i),$$

  so we can compute approximations of each integral separately.
- Course website for some plots.

Gaussian quadrature

- In practice, one can do even better.
- The main issue with what we've seen is the selection of points within each subinterval.
- One can use some nice mathematics involving orthogonal polynomials to show that
  - choosing a special set of points will improve the approximation accuracy for the polynomial integral approximations
  - even when the interpolating polynomial is not that close to the true function!
- In practice you can get the nodes and weights using software packages.

Multiple integrals

- Consider an integral over $D = [a_1, b_1] \times \cdots \times [a_d, b_d]$

$$I(f) = \int_D f(x_1, \ldots, x_d) \mathrm{d}(x_1, \ldots, x_d).$$

- Letting $D' = [a_2, b_2] \times \cdots \times [a_d, b_d]$, we rewrite $I(f)$ as an iterated integral

$$I(f) = \int_{a_1}^{b_1} \int_{D'} f(x_1, \ldots, x_d) \mathrm{d}(x_2, \ldots, x_d) \mathrm{d}x_1 = \int_{a_1}^{b_1} g(x_1) \mathrm{d}x_1,$$

  where taking $h_{x_1}(x_2, \ldots, x_d) = f(x_1, \ldots, x_d)$ we have

$$g(x_1) = I(h_{x_1}) = \int_{D'} h_{x_1}(x_2, \ldots, x_d) \mathrm{d}(x_2, \ldots, x_d).$$

- Suggests a recursive algorithm, which calls a quadrature method...curse of dimensionality!

Outline

Laplace approximation

- Imagine you have a latent variable model with joint density $f(y, b)$, $b \in \mathbb{R}^d$.
- We want to evaluate $f(y) = \int f(y, b) \mathrm{d}b$.
- For a fixed y, use a Taylor expansion

$$\log f(y, b) = \log f(y, \hat{b}_y) - \frac{1}{2}(b - \hat{b}_y)^T H (b - \hat{b}_y) + \cdots,$$

  where $\hat{b}_y$ maximizes $f(y, \cdot)$.
- Then, if the expansion about $\hat{b}_y$ is accurate for all b,

$$f(y, b) \approx f(y, \hat{b}_y) \exp\left\{ -\frac{1}{2}(b - \hat{b}_y)^T H (b - \hat{b}_y) \right\}.$$

- We find

$$f(y) \approx f(y, \hat{b}_y) \int e^{-\frac{1}{2}(b - \hat{b}_y)^T H (b - \hat{b}_y)} \mathrm{d}b = f(y, \hat{b}_y) \frac{(2\pi)^{d/2}}{\det(H)^{1/2}}.$$

Laplace approximation: alternative version

- We have
$$f(y, b) = f(y)f(b \mid y).$$
- Now assume $b \mid y$ is $N(\hat{b}_y, H^{-1})$, and compute at $b = \hat{b}_y$,
$$f(y) = \frac{f(y, \hat{b}_y)}{f(\hat{b}_y \mid y)} = f(y, \hat{b}_y)\frac{(2\pi)^{d/2}}{\det(H)^{1/2}},$$
since
$$f(\hat{b}_y \mid y) = \frac{1}{(2\pi)^{d/2}\det(H)^{-1/2}} = \frac{\det(H)^{1/2}}{(2\pi)^{d/2}}.$$
- So really this amounts to approximating the conditional with a Gaussian.

Outline

Something different...

- Quadrature is not always appropriate. E.g.,
  - high-dimensional integrals
  - non-smooth functions.
- In quadrature we compute a weighted sum

$$\sum_{i=1}^{N} w_i f(x_i),$$

  where x_i, w_i are fixed values.
- A strange idea: what if we randomize the X_i and just use $w_i = \frac{1}{N}$?
  - Surprisingly, this can be a very good idea.

Fundamental idea

- The fundamental idea behind Monte Carlo integration is to view the integral of g over some set X

$$I = \int_{\mathrm{X}} g(x)\mathrm{d}x,$$

  as the expectation of a random variable.
- Find a PDF π such that $\pi(x) > 0$ whenever $g(x) \neq 0$, then we can write

$$I = \int_{\mathrm{X}} g(x)\mathrm{d}x = \int_{\mathrm{X}} f(x)\pi(x)\mathrm{d}x = \mathbb{E}_{\pi}[f(X)],$$

  where $f(x) = g(x)/\pi(x)$.
- Idea: simulate independent $X_i \sim \pi$, compute $N^{-1}\sum_{i=1}^{N} f(X_i)$.

Consistency, lack-of-bias

- Averages of i.i.d. random variables are simple to analyze!
- Consistency: if I is finite,

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i) \to_p \mathbb{E}_\pi[f(X)] = I.$$

- Lack-of-bias:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right] = \mathbb{E}_\pi[f(X)] = I.$$

Variance, CLT

- Variance: if $\mathrm{var}_\pi(f) = \mathbb{E}_\pi[f(X)^2] - \mathbb{E}_\pi[f(X)]^2 < \infty$,

$$\mathrm{var}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right) = \frac{1}{N}\mathrm{var}_\pi(f(X)).$$

- CLT:

$$\sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N} f(X_i) - I\right\} \to_d N\left(0, \mathrm{var}_\pi(f)\right).$$

- Quantitative control from $\int f(x)^2\pi(x)\mathrm{d}x$.
- Smoothness is irrelevant.
- $L^2(\pi)$ is the space of finite variance functions under π.

Simple example

- Let $I = \int_0^1 x^2 \mathrm{d}x$. Of course this is $1/3$.
- Let $\pi(x) = \mathbb{I}(0 < x < 1)$, i.e. π is Uniform$(0, 1)$.
- Simulate $X_1, \ldots, X_N \sim \pi$ independent and compute

$$\frac{1}{N} \sum_{i=1}^{N} X_i^2.$$

- Variance is

$$\frac{1}{N} \left[\int_0^1 x^4 \mathrm{d}x - \left\{ \int_0^1 x^2 \mathrm{d}x \right\}^2 \right] = \frac{1}{N} \left\{ \frac{1}{5} - \frac{1}{9} \right\} = \frac{1}{N} \cdot \frac{4}{45}.$$

What about a different π?

- Instead, consider $\pi(x) = 2x\mathbb{I}(0 < x < 1)$.
- Then we need $f(x) = g(x)/\pi(x) = x^2/(2x) = x/2$ on $(0, 1)$.
- Then simulate $X_1, \ldots, X_N \sim \pi$ independent and compute

$$\frac{1}{N}\sum_{i=1}^{N}\frac{X_i}{2}.$$

- Now variance is

$$\frac{1}{N}\left[\int_0^1\left(\frac{x}{2}\right)^2 2x\mathrm{d}x - \left\{\int_0^1 x^2\mathrm{d}x\right\}^2\right] = \frac{1}{N}\left\{\frac{1}{8} - \frac{1}{9}\right\} = \frac{1}{N}.\frac{1}{72}.$$

- That's better!

Importance sampling

- If we have $I = \int f(x)\pi(x)\mathrm{d}x$, then importance sampling refers to the identity

$$I = \int f(x)\pi(x)\mathrm{d}x = I = \int f(x)w(x)\mu(x)\mathrm{d}x,$$

  where $w(x) = \pi(x)/\mu(x)$, and we assume $\mu(x) > 0$ whenever $\pi(x) > 0$.
- Basically the same thing we did when coming up with

$$\int g(x)\mathrm{d}x = \int f(x)\pi(x)\mathrm{d}x.$$

- Often emphasis is on changing hard to sample π to easy to sample μ.
- Possibly consideration of several f's, hence the stronger constraint.

Optimal importance distribution

- Consider $I = \int g(x)\mathrm{d}x$.
- Importance sampling variance is

$$\int \left|\frac{g(x)}{\mu(x)}\right|^2 \mu(x)\mathrm{d}x - \left(\int g(x)\mathrm{d}x\right)^2,$$

  where the second term does not depend on μ.
- Jensen's inequality gives

$$\int \left|\frac{g(x)}{\mu(x)}\right|^2 \mu(x)\mathrm{d}x \geq \left\{\int \left|\frac{g(x)}{\mu(x)}\right| \mu(x)\mathrm{d}x\right\}^2 = \left\{\int |g(x)|\,\mathrm{d}x\right\}^2.$$

- Now observe that if $\mu(x) = |g(x)| \,/ \int |g(x)|\,\mathrm{d}x$ then

$$\int \left|\frac{g(x)}{\mu(x)}\right|^2 \mu(x)\mathrm{d}x = \left\{\int |g(x)|\,\mathrm{d}x\right\}^2,$$

  so this is an optimal choice!

Practical example

- Often we can't really go for optimality.
- Need to be able to actually sample according to μ.
- Consider π the density of some complicated distribution, e.g. a Bayesian posterior density.
- If we believe π is close to $\mathrm{Normal}(m, \Sigma)$, we could take $\mu = \mathrm{Normal}(m, \Sigma)$.
- In fact, we might take for m and Σ the maximizer of π and the inverse Hessian of $\log \pi$ at m.
- This is the Laplace approximation again!
- Try it in the lab...

Self-normalized importance sampling I

- Try it in the lab...
- Except that we often don't know π precisely, but only up to a normalizing constant.
- Posterior:

$$\pi(\theta) = \frac{1}{Z}\pi_0(\theta)L(\theta; y),$$

  where $Z = \int \pi_0(\theta)L(\theta; y)\mathrm{d}\theta$.
- We can compute unnormalized importance weights

$$\tilde{w}(\theta) = L(\theta; y) \propto \frac{\pi(\theta)}{\pi_0(\theta)},$$

  but not $w(\theta)$ as we can't compute Z.

Self-normalized importance sampling II

- Consider general setting, $\int f(x)\pi(x)\mathrm{d}x$ the objective.
- Imagine we can simulate from μ with $\tilde{w}(x) \propto \pi(x)/\mu(x)$.
- Then consider the identity

$$\frac{\int f(x)\tilde{w}(x)\mu(x)\mathrm{d}x}{\int \tilde{w}(x)\mu(x)\mathrm{d}x} = \frac{\int f(x)w(x)\mu(x)\mathrm{d}x}{\int w(x)\mu(x)\mathrm{d}x} = \int f(x)\pi(x)\mathrm{d}x,$$

  so we can approximate numerator and denominator on LHS!
- I.e., with samples $X_i \sim \mu$ independent, compute

$$\frac{\frac{1}{N}\sum_{i=1}^N \tilde{w}(X_i)f(X_i)}{\frac{1}{N}\sum_{i=1}^N \tilde{w}(X_i)} = \frac{\sum_{i=1}^N \tilde{w}(X_i)f(X_i)}{\sum_{i=1}^N \tilde{w}(X_i)} = \sum_{i=1}^N \bar{W}_i f(X_i),$$

  where the self-normalized weights are:

$$\bar{W}_i = \frac{\tilde{w}(X_i)}{\sum_{i=1}^N \tilde{w}(X_i)}.$$

- Cf. quadrature rules.

Self-normalized importance sampling III

- SNIS is feasible in many scenarios.
- Consistency follows from law of large numbers and continuous mapping.
- Not unbiased in general.
- Asymptotic normality also holds:

$$\sqrt{N}\left\{\sum_{i=1}^{N} \bar{W}_i f(X_i) - I\right\} \to N(0, \sigma^2(f)),$$

where $I = \int f(x)\pi(x)\mathrm{d}x$ and

$$\sigma^2(f) = \int (f(x) - I)^2 w(x)^2 \mu(x)\mathrm{d}x.$$

A measure of sample quality I

- One simple measure of sample quality is the so-called "effective sample size".
- It was inspired originally by quantifying the ratio of asymptotic variances with μ and with π.
  - But this is not what it actually approximates consistently...
  - It is (intentionally) function independent.
- The effective sample size is the random variable

$$\mathcal{E}_N = N \cdot \frac{\left\{ \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(X_i) \right\}^2}{\left\{ \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(X_i)^2 \right\}},$$

  where the fraction on the right tends to $R(\pi, \mu) = \left\{ \int w(x)^2 \mu(x) \mathrm{d}x \right\}^{-1} \in (0, 1]$.
- The effective sample size takes values in $[1, N]$.

A measure of sample quality II

- One interpretation:

$$\begin{aligned}\sigma^2(f) &= \int (f(x) - I)^2 w(x)^2 \mu(x) \mathrm{d}x \\ &\leq \|f\|_{\mathrm{osc}}^2 \int w(x)^2 \mu(x) \mathrm{d}x \\ &= \frac{\|f\|_{\mathrm{osc}}^2}{R(\pi, \mu)},\end{aligned}$$

  and the effective sample size is an approximation of the denominator.

- We also have a relationship to the χ^2-divergence between π and μ:

$$R(\pi, \mu) = \frac{1}{1 + d_{\chi^2}(\pi, \mu)}.$$

Outline

Wrapping up

- Low-dimensional, smooth integrands: give quadrature a go!
- High-dimensional, $L^2(\pi)$ functions: give Monte Carlo a go!
- More seriously, these methods work for suitably simple problems, which can arise in practice.
- When suitable, certainly easier to explain their use.
  - Some statistical models have a lot more regularity than arbitrary functions.
- For more challenging integrals, there are more advanced Monte Carlo methods.
  - Markov chains, interacting particles, diffusions, auxiliary distributions, etc.