STATS 230: Computational Statistics
Numerical Integration

Babak Shahbaba

Department of Statistics, UCI

Overview

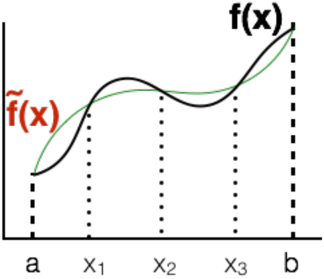
- Statistical inference quite often depends on intractable integrals

$$I = \int_a^b f(x)dx$$

- This is especially true in Bayesian statistics, where integrating functions with respect to the posterior distribution is usually not trivial
- In these cases, we typically use numerical methods to approximate integrals
- In some situations, the likelihood itself may depend on intractable integrals so frequentist methods would also require numerical integration
- In this lecture, we start by discussing some simple numerical methods that can be easily used in low dimensional problems
- Next, we will discuss several Monte Carlo strategies that could implemented even when the dimension is high

Newton-Côtes quadrature

- A common strategy for approximating integrals (especially in univariate problems) is to approximate the integrand $f(x)$ with a tractable function $\tilde{f}(x)$, which can be integrated easily
- We typically constrain the approximating function to agree with $f(x)$ on a grid of points: $x_1, \ldots, x_n$



Newton-Côtes quadrature

- Newton-Côtes methods use equally-spaced grids
- The approximating function is a polynomial
- We then approximate the integral with a weighted sum as follows

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

- In its simplest case, we can use the Riemann rule by partitioning the interval $[a, b]$ into n subintervals of length $h = \frac{b-a}{n}$; then

$$\hat{I}_\ell = h \sum_{i=0}^{n-1} f(a + ih)$$

where $\tilde{f}(x)$ is a piecewise constant function approximating $f(x)$ over each subinterval by a constant function with the same value as $f(x)$ at the left point of the interval.

Newton-Côtes quadrature

- Alternatively, the approximating function could agree with the integrand at right point of each subinterval

$$\hat{I}_r = h \sum_{i=1}^{n} f(a + ih)$$

- In either case, the approximating function is a zero-order polynomial.

- To improve the approximation, we can use the trapezoidal rule by using a piecewise linear function that agrees with $f(x)$ at both ends of subintervals

$$\hat{I} = \frac{h}{2} f(a) + h \sum_{i=1}^{n-1} f(x_i) + \frac{h}{2} f(b)$$

Newton-Côtes quadrature

- We could further improve the approximation by using higher order polynomials
- Simpson's rule uses a quadratic approximation over each subinterval

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{x_{i+1} - x_i}{6}[f(x_i) + 4f(\frac{x_i + x_{i+1}}{2}) + f(x_{i+1}))]$$

- In general, we can use any polynomial of degree k

Gaussian quadrature

- Newton-Côtes rules require equally spaced grids
- We can relax this assumption by using more points where the magnitude of the integrand is large
- This is called Gaussian quadrature, which is especially useful for the following type of integrals (e.g., integrating a function with respect to a distribution with density $g(x)$)

  $$\int_a^b f(x)g(x)dx$$

  where $g(x)$ is a nonnegative function and $\int_a^b x^k g(x)dx < \infty$

- Note that in general we can write

  $$\int_a^b f(x)dx = \int_a^b \frac{f(x)}{g(x)}g(x)dx = \int_a^b f^*(x)g(x)dx$$

Orthogonal functions

- The grids are typically specified based on the roots of a set of orthogonal polynomials

- In general, for squared integrable functions,

$$\int_a^b f(x)^2 g(x) dx < \infty$$

denoted as $f \in \mathcal{L}^2_{g,[a,b]}$, we define the inner product as

$$< f, h >_{g,[a,b]} = \int_a^b f(x)g(x)h(x)dx$$

where $f, h \in \mathcal{L}^2_{g,[a,b]}$

Orthogonal functions

- Two functions are orthogonal when $< f, h >= 0$
- For example, $\sin mx$ and $\sin nx$ are orthogonal over $[0, \pi]$ if $m \neq n$

$$\begin{aligned}
\int_0^\pi \sin mx \sin nx \; dx &= \\
\int_0^\pi 0.5 \cos(m-n)x - 0.5\cos(m+n)x \; dx &= \\
&= 0 \qquad \text{if } m \neq n
\end{aligned}$$

- See Strang (2012) for more details

Orthogonal polynomials

- For a given $g(x)$ and interval $[a, b]$, we are interested in a set of orthogonal polynomials $\{p_j(x)\}_{j=0}^{\infty}$
- Note that in general, these are not uniques since $< f, h >= 0$ implies $< af, bh >= 0$
- To make the orthogonal polynomials unique, we usually use of these standardizations:
  - make the polynomials orthonormal: $< p_j, p_j >= 1$
  - set the leading coefficient of $p_j(x)$ to 1
  - specify the values at some points: $p_j(0)$, $p_j(a)$, $p_j(b)$

Orthogonal polynomials

- Orthogonal polynomials form a basis for $\mathcal{L}^2_{g,[a,b]}$ so any function in this space can be written as a

$$f(x) = \sum_{j=1}^{\infty} a_j p_j(x)$$

  where $a_j = \frac{<f,p_j>}{<p_j,p_j>}$

- Orthogonal have many attractive properties including:
  - We can write $p_j = (A_j + xB_j)p_{j-1} - C_j p_{j-2}$ for some easy to find constants A_j, B_j, and C_j
  - If the coefficients are real, the zeros are real and are located in the interior of $[a, b]$

Gaussian quadrature

- Two commonly used orthogonal polynomials are Legendre and Hermite

	Interval	$g(x)$	Standardization
Legendre	$[-1, 1]$ $[0, 1]$	1	$p_j(1) = 1$
Hermite	$(-\infty, \infty)$	$\exp(-x^2/2)$	leading coefficient $= 1$

- Note that, when the interval is $[a, b]$, we can still use the Legendre method after change of variable: $x = \alpha y + \beta$

$$\int_a^b f(x)dx = \int_{-1}^{1} f(\frac{b-a}{2}y + \frac{b+a}{2})\frac{b-a}{2}dy$$

Gaussian quadrature

- To use Gaussian quadrature, we use a set of orthogonal polynomials $\{p_j(x)\}$ with the corresponding $g(x)$ and $[a, b]$

- Denote the n zeros of $p_n(x)$ by $a < x_1 < \ldots, < x_n < b$, and the corresponding weights by $w_1, \ldots, w_n$

- Then,

$$\int_a^b f(x)g(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

- The approximation is exact if f is a polynomial of degree at most $2n - 1$

- We can obtain the zeros and their corresponding weights using R (e.g., gauss.quad function in the statmod package) or MATLAB

Monte Carlo method

- We now discuss the Monte Carlo method mainly in the context of statistical inference
- From now on, we use $f(x)$ to denote density functions
- As before, we are interested in finding integrals of the form $I = \int_a^b h(x)dx$
- If we can draw iid samples, $x^{(1)}, x^{(2)}, ..., x^{(m)}$ uniformly from (a, b), we can approximate this integral as

$$\hat{I}_m = (b-a)\frac{1}{m}[h(x^{(1)}) + h(x^{(2)}) + \ldots + h(x^{(m)})]$$

- Note that we can think about the integral as

$$(b-a)\int_a^b h(x)f(x)dx$$

, where $f(x) = 1/(b-a)$ is the density of Uniform(a, b)

Monte Carlo method

- In general, we are interested in integrals of the form $\int_{\mathcal{X}} h(x)f(x)dx$, where $f(x)$ is a probability density function, i.e., the integral is $\mu = E_f(h(x))$

- Analogous to the above argument, we can approximate this integral (or expectation) by drawing iid samples $x^{(1)}, x^{(2)}, ..., x^{(m)}$ from the density $f(x)$ and then

$$\hat{I} = \frac{1}{m}[h(x^{(1)}) + h(x^{(2)}) + ... + h(x^{(m)})]$$

- Based on the law of large numbers, we know that

$$\lim_{m\to\infty} \hat{I}_m = I, \qquad \text{with probability 1}$$

- And based on the central limit theorem

$$\sqrt{m}(\hat{I}_m - I) \to N(0, \sigma^2), \qquad \sigma^2 = \mathrm{Var}(h(x))$$

Monte Carlo method

- For sampling x from f, we can sample $u \sim \mathrm{Uniform}(0,1)$, and set $x = F^{-1}(u)$, where F^{-1} is the inverse CDF of f
- This would of course work if the CDF has a closed form and we can find its inverse.
- For example, assume we want to find the expectation of the function $h(x) = \sqrt{x}$ with respect to the exponential distribution Exp(3) where $f(x) = \theta \exp(-\theta x)$ is the density and $F(x) = 1 - \exp(-\theta x)$ is the CDF
- We can sample $u^{(i)} \sim \mathrm{Uniform}(0,1)$ for $i = 1, ..., m$, and set

$$x^{(i)} = -\frac{log(1 - u^{(i)})}{\theta}$$

- We can then estimate $E_f(\sqrt{x})$ as

$$\hat{\mu} = \frac{1}{m}[\sqrt{x^{(1)}} + \sqrt{x^{(2)}} + ... + \sqrt{x^{(m)}}]$$

Rejection sampling

- If it is difficult or computationally intensive to sample directly from $f(x)$ (as described above), we need to use other strategies.

- Although it is difficult to sample from $f(x)$, suppose that we can evaluate the density at any given point up to a constant $f(x) = f^*(x)/Z$, where Z could be unknown (remember that this makes Bayesian inference convenient since we usually know the posterior distribution only up to a constant).

- Furthermore, assume that we can easily sample from another distribution with the density $g(x) = g^*(x)/Q$, where Q is also a constant.

- Now we choose the constants c such that $cg^*(x)$ becomes the envelope (blanket) function for $f^*(x)$:

$$cg^*(x) \geq f^*(x), \qquad \forall x$$

- Then, we can use a strategy known as *rejection sampling* in order to sample from $f(x)$ indirectly.

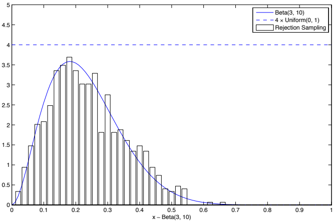
Rejection sampling

- The rejection sampling method works as follows:
  1. draw a sample x from g(x)
  2. generate $u \sim \mathrm{Uniform}(0, cg^*(x))$
  3. if $u \leq f^*(x)$ we accept x as the new sample, otherwise, reject x (discard it) and start with a new sample from $g(x)$.

An illustrative example

- Assume that it is difficult to sample from the Beta(3, 10) distribution (this is not the case of course!).

- We use the Uniform(0, 1) distribution with $g(x) = 1, \forall x \in [0, 1]$, which has the envelop property: $4g(x) > f(x), \forall x \in [0, 1]$. The following graph shows the result after 3000 iterations.



- Finding an appropriate distribution $g(x)$ becomes very difficult (and sometimes impossible) as the dimensionally of x increases, and it might not be efficient in general if there is a high rejection rate.

Importance sampling

- Importance sampling is used to find the expectation of a function $h(x)$ with respect to a distribution, with the density $f(x)$, from which we cannot directly sample.
- Assume again that we can sample from another distribution with the density $g(x)$ that is close to $f(x)$.
- Note that unlike the rejection sampling, we do not need the envelop property.
- The only requirement is that $g(x)$ must not be zero anywhere that $f(x)$ is not zero.
- As before, we only need to know $f(x)$ and $g(x)$ up to a constant.

Importance sampling

- Now we can write $E_f(h(x))$ as follows:

$$\begin{aligned}
\mu = E_f(h(x)) &= \int_{\mathcal{X}} h(x) f(x) dx \\
&= \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx \\
&= \int_{\mathcal{X}} [h(x) w(x)] g(x) dx \\
&= E_g(h(x) w(x))
\end{aligned}$$

Importance sampling

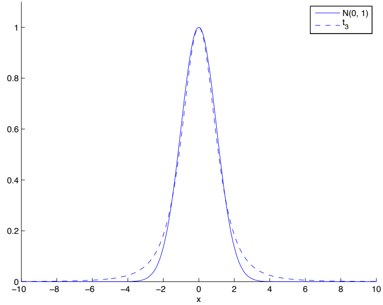
- We can then approximate the original expectation as follows:
  1. draw samples $x^{(1)}, ..., x^{(m)}$ from $g(x)$
  2. Find the *importance weight* $w^{(j)} = \frac{f(x^{(j)})}{g(x^{(j)})}$, where $j = 1, ..., m$
  3. Approximate the original expectation, $\mu = E_f(h(x))$, as follows

$$\hat{\mu} = \frac{[w^{(1)}h(x^{(1)}) + ... + w^{(m)}h(x^{(m)})]}{w^{(1)} + ... + w^{(m)}}$$

- In general, $f(x)$ and $g(x)$ do not need to be normalized. We only need to know them up to a constant. Whatever those constants are, they will be canceled out from the numerator and denominator.

An illustrative example

- We want to approximate a N(0, 1) distribution with $t(3)$ distribution:



- We use the unnormalized forms where $f(x) = \exp(-\frac{x^2}{2})$ and $g(x) = (1 + \frac{x^2}{3})^{-2}$.
- We generated 500 samples and estimated $\mu = E(x^2)$ as 0.97, which is close to the true value 1.

Potential problems

- The efficiency of this approach depends on how good $g(x)$ approximates $f(s)$.
- If the samples do not include the areas where f is large, or they include only a few samples from the high probability region, the estimation would not be accurate.
- To see this, as an exercise, repeat the above example, but this time approximate $t(3)$ with $N(0, 1)$.
- The estimate of $E(x^2)$ this time would be systematically smaller than the true value 3.