Bayesian Inference and Data Assimilation

Prof. Dr.-Ing. Sebastian Reich

Universität Potsdam

12 April 2021

3 Computational Statistics

We introduced the concept of a random variable $X : \Omega \to \mathcal{X}$, whose outcome is characterized by a probability measure μ_X on $\mathcal{X}$. For real applications, the dimension of $\mathcal{X}$ is often very large, which means that X itself is often very unwieldy. It is then convenient to communicate the uncertainty characterised by X in terms of summary statistics, many of which take the form of *approximations* (also called estimates if such approximations involve realisations of random variables) to expectation values of scalar functions $f : \mathcal{X} \mapsto \mathbb{R}$,

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\mu_X(\mathrm{d}x).$$

In most cases discussed throughout this chapter $\mathcal{X} = \mathbb{R}^{N_x}$ and μ_X is absolute continuous with respect to the Lebesgue measure on $\mathbb{R}^{N_x}$.

Then we recall that there exists a PDF π_X such that

$$\mu_X(\mathrm{d}x) = \pi_X(x)\mathrm{d}x,$$

i.e.,

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\pi_X(x)\mathrm{d}x,$$

provided the integral exists for the f under consideration.

In other words, all of statistics boils down computing/approximating integrals. There are the classical quadrature rules from numerical analysis and the Monte Carlo techniques from computational statistics.

3.1 Deterministic quadrature

We start with the simpler case of univariate random variables, i.e., $N_x = 1$.

We will often write $\bar{f}$ instead of $\mathbb{E}[f(X)]$ for simplicity.

Under appropriate assumptions on f, expectation integrals can be approximated by numerical quadrature rules.

Definition (Numerical quadrature rules)

For a particular PDF π_X, a numerical quadrature rule for an integral

$$\bar{f} = \int_{\mathbb{R}} f(x)\pi_X(x)\mathrm{d}x,$$

with any choice of function f, is given by

$$\bar{f}_M := \sum_{i=1}^{M} b_i f(c_i).$$

Here $c_i \in \mathbb{R}$, $i = 1, \ldots, M$, denote the quadrature points and $b_i > 0$ their weights.

Let $\Pi_k(\mathbb{R})$ denote the $k+1$ dimensional linear space of all polynomials of order k or less, *i.e.* of the form

$$f(x) = a_0 + a_1 x + \cdots + a_k x^k.$$

A quadrature rule is of *order* p if $\bar{f} = \bar{f}_M$ for all integrands $f(x) \in \Pi_{p-1}(\mathbb{R})$.

Example (Uniform distribution)

Consider the case of a uniform distribution on the unit interval $[0, 1]$ and recall the notation $X \sim \mathrm{U}[0, 1]$ for a random variable X with PDF

$$\pi_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Expectation values are then simply given by

$$\bar{f} = \int_0^1 f(x)\mathrm{d}x.$$

The midpoint rule

$$\bar{f} \approx \bar{f}_1 = f(1/2)$$

is the lowest-order Gauss-Legendre quadrature formula with $M = 1$, $b_1 = 1$, $c_1 = 1/2$. Its order is $p = 2$. With $M = 2$ Gauss-Legendre quadrature achieves fourth-order accuracy and the quadrature points and weights are given by

$$c_1 = \frac{1 - 1/\sqrt{3}}{2}, \quad c_2 = \frac{1 + 1/\sqrt{3}}{2}, \quad b_1 = b_2 = \frac{1}{2}.$$

Example (Continued)

In addition to (or as an alternative to) increasing the number of quadrature points M, the integral may be split into a sum of integrals over finitely many non-overlapping subintervals of $[0,1]$.

Consider, for example, the formal decomposition

$$\bar{f} = \sum_{i=1}^{N_I} \int_{(i-1)\Delta x}^{i\Delta x} f(x)\mathrm{d}x, \tag{1}$$

where each subinterval is of length $\Delta x = 1/N_I$ with $N_I > 1$. A Gauss-Legendre quadrature rule of order p can now be applied to each of the integrals in (1). Let us denote the numerical result by $\bar{f}_{M,N_I}$. If the function f is q times continuously differentiable, then

$$|\bar{f}_{M,N_I} - \bar{f}| = \mathcal{O}(\Delta x^{\min(p,q)})\,.$$

Hence high-order quadrature rules are only useful if f is sufficiently smooth.

Example (Non-smooth integrands)

Let $\phi : [0,1] \to [0,1]$ denote the standard tent map. We compute expectation values with respect to $X \sim \mathrm{U}[0,1/2]$ with PDF

$$\pi_X(x) = \begin{cases} 2 & \text{if } x \in [0,1/2], \\ 0 & \text{otherwise,} \end{cases}$$

and f given by application of ϕ n times, *i.e.*,

$$f(x) = \underbrace{\phi(\phi(\cdots(\phi(x))\cdots))}_{n \text{ times}} = \phi^n(x), \qquad n \geq 1. \tag{2}$$

Since

$$\mathbb{E}[\phi^n(X)] = \mathbb{E}[\phi^{n-1}(X)], \quad n > 1,$$

and $\mathbb{E}[\phi(X)] = 1/2$, the analytic value for $\bar{f} = \mathbb{E}[\phi^n(X)]$ is equal to $1/2$ for all $n \geq 1$.

Example (Continued)

The midpoint rule for the interval $[0, 1/2]$ can be obtained by transforming the quadrature point $c_1 = 1/2$ to $\hat{c}_1 = 1/4$ while keeping the weight $b_1 = 1$. The midpoint rule yields the exact result for $n = 1$. However, since

$$\phi^n(\hat{c}_1) = \begin{cases} 1/2 & \text{for } n = 1, \\ 1 & \text{for } n = 2, \\ 0 & \text{otherwise,} \end{cases}$$

the approximation errors for the implicit midpoint method increase to 0.5 in absolute value for $n > 1$.

Example (Continued)

Even more revealing is the behaviour of the implicit midpoint rule applied over subintervals of length $1/N$ with $f = \phi^n$. We now fix $n = 50$ and increase the number of subintervals $N_I \gg 1$.

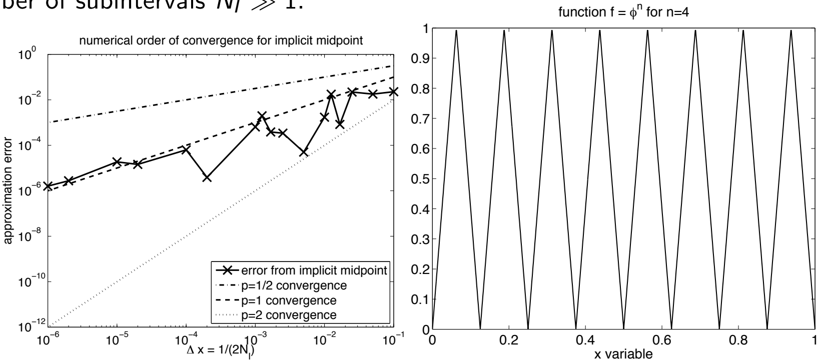


Figure: Left: Approximation errors from the composite midpoint rule with $f = \phi^n$, $n = 50$. The approximation error is displayed as a function of $\Delta x = 1/(2N_I)$. The standard order of the implicit midpoint is $p = 2$; but $p = 1$ is observed in this example, where f is a highly irregular function. Right: Plot of the function $\phi^n(x)$ for $n = 4$.

Example (Gaussian distribution)

If X' is a univariate Gaussian random variable with mean zero and variance one, then expectation values take the form

$$\mathbb{E}[f(X')] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x') e^{-(x')^2/2} \mathrm{d}x'.$$

With $M = 2$, the quadrature points and weights for the Gauss-Hermite quadrature rule of order $p = 4$ are given by

$$c_1 = -1, \quad c_2 = 1, \quad b_1 = b_2 = 1/2.$$

Assume that we want to use the same quadrature rule for approximating expectation values of another Gaussian random variable X with mean $\bar{x} \neq 0$ and variance $\sigma^2 \neq 1$.

How to achieve this? There are two ways to go about this; which we illustrate next.

Example (Continued)

We may rewrite $\mathbb{E}[f(X)]$ as an expectation over X' as follows,

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)\pi_X(x)\mathrm{d}x, = \int_{\mathbb{R}} f(x)\frac{\pi_X(x)}{\pi_{X'}(x)}\pi_{X'}(x)\mathrm{d}x, = \mathbb{E}\left[f(X')\frac{\pi_X(X')}{\pi_{X'}(X')}\right].$$

This suggests the following quadrature formula for $\bar{f} = \mathbb{E}[f(X)]$ given by

$$\bar{f}_M = \sum_{i=1}^{M} b_i f(c_i)\frac{\pi_X(c_i)}{\pi_{X'}(c_i)} = \sum_{i=1}^{M} \hat{b}_i f(c_i),$$

with new weights

$$\hat{b}_i = b_i \frac{\pi_X(c_i)}{\pi_{X'}(c_i)}.$$

For constant $f = f_0$, we expect to obtain $\bar{f}_M = f_0$, but in general $\sum_i \hat{b}_i \neq 1$: the order of the quadrature rule has dropped from $p = 2M$ to $p = 0$! In order to recover $p = 1$, we could use normalised weights.

Example (Continued)

A much better remedy is to apply a change-of-variables formula to adjust the quadrature points c_i instead of the weights b_i. Utilising the linear transformation $x' = (x - \bar{x})/\sigma$ and $\mathrm{d}x = \sigma \mathrm{d}x'$, we get

$$\begin{aligned}\mathbb{E}[f(X)] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} f(x) e^{-(x-\bar{x})^2/2\sigma^2}\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\sigma x' + \bar{x}) e^{-(x')^2/2}\mathrm{d}x' = \mathbb{E}[f'(X')] \\ &\approx \sum_{i=1}^{M} b_i f(\sigma c_i + \bar{x})\,,\end{aligned}$$

where $f'(x) = f(\sigma x + \bar{x})$. We conclude that Gauss-Hermite quadrature points for a Gaussian random variable with mean $\bar{x}$ and variance σ^2 are given by

$$\hat{c}_i = \sigma c_i + \bar{x}.$$

Note that a linear transformation does not change the order of a polynomial $f' \in \Pi_k(\mathbb{R})$ and, hence, the order of a quadrature rule is preserved.

Remark

We observed that when transforming between two Gaussian random variables with different means and variances, changing the quadrature points preserves the order of the quadrature rule, whereas changing the quadrature weights does not.

The choice between either changing the weights of a quadrature rule, or the location of its quadrature points (or both) will be a recurring theme of this course.

Changing the weights in the quadrature method is related to importance sampling while changing the quadrature points is ultimately connected to the concept coupling of measures.

Remark (Empirical measure)

A quadrature rule with quadrature points c_i and weights b_i yields exact expectation value with respect to the empirical measure of the form

$$\mu_M(\mathrm{d}x) := \sum_{i=1}^{M} w_i \mu_{x_i}(\mathrm{d}x) = \sum_{i=1}^{M} w_i\, \delta(x - x_i)\mathrm{d}x.$$

The empirical measure μ_M is a probability measure if $\sum_{i=1}^{M} w_i = 1$.

Proceeding, insertion of the empirical measure into the expectation formula gives,

$$\bar{f} = \int_{\mathcal{X}} f(x)\mu_M(\mathrm{d}x) = \sum_{i=1}^{M} w_i f(x_i) = \sum_{i=1}^{M} b_i f(c_i),$$

as required.

Hence, an interpretation of quadrature rules for expectation integrals is that we first approximate the measure μ_X by the empirical measure, and then evaluate all expectations exactly with respect to that measure.

Remark (Curse of dimensionality)

Gauss-type quadrature rules are extremely powerful whenever they are applicable. However, the data assimilation problems lead to random variables in very high-dimensional spaces. Under those circumstances, standard quadrature rules, such as Gauss-Hermite or Gauss-Legendre, are no longer useful.

For example, a straightforward approximation of an integral

$$\bar{f} = \int_0^1 \cdots \int_0^1 f(x_1, \ldots, x_{N_x}) \mathrm{d}x_1 \cdots \mathrm{d}x_{N_x} = \int_{[0,1]^{N_x}} f(x) \mathrm{d}x$$

in $\mathbb{R}^{N_x}$ by a product of N_x one-dimensional quadrature rules, each with M quadrature points, leads to a regular array of M^{N_x} quadrature points in $\mathbb{R}^{N_x}$.

This number quickly becomes very large as the dimension of the problem increases and the associated computational effort cannot be justified, especially in combination with a possible reduction of order as encountered in the previous (non-smooth) tent map example. This is often referred to as the "curse of dimensionality".

Remark (ANOVA)

A popular method for reducing the computational effort of high-dimensional integrals is the analysis of variance (ANOVA) representation of an integrand f.

In order to demonstrate the basic idea, we set $N_x = 3$ and $x = (x_1, x_2, x_3)^T$ as on the previous slide. We seek a decomposition of f of the form

$$
\begin{aligned}
f(x_1, x_2, x_3) = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2) +& \\
f_{13}(x_1, x_3) + f_{23}(x_2, x_3) + f_{123}(x_1, x_2, x_3).&
\end{aligned}
$$

Once such a decomposition is available, the first three non-trivial terms can be integrated by standard univariate quadrature rules. The next three terms require more work but still reduce to two-dimensional quadrature rules, while only the last term requires a full three-dimensional quadrature approach.

The key idea of ANOVA is now to choose the decomposition such that the significant contributions to the integral come from the lower order terms.

Remark (Continued)

We define

$$f_0 = \mathbb{E}[f(X)] = \int_{[0,1]^3} f(x_1, x_2, x_3)\, \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3,$$

and set

$$f_l(x_l) = \int_{[0,1]^2} f(x) \prod_{j \neq l} \mathrm{d}x_j - f_0, \quad l = 1, 2, 3.$$

We then continue at the next level with

$$f_{lk}(x_l, x_k) = \int_{[0,1]} f(x) \mathrm{d}x_j - f_l(x_l) - f_k(x_k) - f_0, \quad j \neq k,\ j \neq l,\ k < l \leq 3.$$

We finally obtain

$$f_{123}(x) = f(x) - f_{12}(x_1, x_2) - f_{13}(x_1, x_3) - f_{23}(x_2, x_3) - \sum_{i=1}^{3} f_i(x_i) - f_0.$$

Remark (Continued)

The ANOVA decomposition has the desirable property that all terms in the decomposition are mutually orthogonal under the inner product

$$\langle g, f\rangle = \int_{[0,1]^3} g(x) f(x) \mathrm{d}x,$$

for two functions $f, g : [0,1]^3 \to \mathbb{R}$. *Hence the variance of the integrand, i.e.,*

$$\sigma^2 = \int_{[0,1]^3} \left(f(x) - \mathbb{E}[f(X)]\right)^2 \mathrm{d}x,$$

is equivalent to

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_{12}^2 + \sigma_{23}^2 + \sigma_{13}^2 + \sigma_{123}^2,$$

with, for example,

$$\sigma_1^2 = \int_0^1 \left(f_1(x_1) - \mathbb{E}[f_1(X_1)]\right)^2 \mathrm{d}x_1,$$

and the other variances defined accordingly.

Remark (Continued)

Let us now assume that

$$\sigma_i^2 \gg \sigma_{ij} \gg \sigma_{123} \, ,$$

Then a highly accurate quadrature rule is applied to $f_i(x_i)$, a less accurate to $f_{ij}(x_i, x_j)$, and an even less accurate to $f_{123}(x_1, x_2, x_3)$.

This line of thinking leads to sparse quadrature rules. When applicable, these methods can be highly efficient; but we will turn our attention to more easily applicable approximations.

Remark (Laplace's method)

There are two key assumptions behind Laplace's method. First, there is a function g such that

$$f(x)\pi_X(x) = e^{-g(x)},$$

and second, g has a unique global minimum denoted by x_0. In that case we may expand g about x_0 to obtain

$$\begin{aligned} g(x) &= g(x_0) + \underbrace{g'(x_0)}_{=0}(x - x_0) + \frac{1}{2}g''(x_0)(x-x_0)^2 + \cdots \\ &= g(x_0) + \frac{1}{2}g''(x_0)(x-x_0)^2 + \cdots \,, \end{aligned}$$

where we have assumed $x \in \mathbb{R}$ for notational simplicity, and where primes denote differentiation with respect to x. Since x_0 is the unique minimum, we have $g''(x_0) > 0$.

Remark (Continued)

Substituting this expansion into $\bar f$ gives

$$\bar f \approx e^{-g(x_0)} \int_{\mathbb{R}} e^{-g''(x_0)(x-x_0)^2/2}\mathrm{d}x.$$

Then, making use of the fact that

$$\int_{\mathbb{R}} e^{-g''(x_0)(x-x_0)^2/2}\mathrm{d}x = \sqrt{\frac{2\pi}{g''(x_0)}},$$

we finally obtain the approximation

$$\bar f \approx e^{-g(x_0)}\sqrt{\frac{2\pi}{g''(x_0)}}.$$

This approximation becomes very useful in higher dimensions, since it always leads to a global integral of a multivariate Gaussian function which can be evaluated analytically.

Lemma (Laplace's method)

Assume that $\tilde{g} : [a,b] \to \mathbb{R}$ is twice differentiable and has a unique global minimum at $x_0 \in [a,b]$. Then

$$\lim_{\varepsilon \to 0} \frac{\int_a^b e^{-\tilde{g}(x)/\varepsilon} \mathrm{d}x}{e^{-\tilde{g}(x_0)/\varepsilon} \sqrt{\frac{2\pi\varepsilon}{\tilde{g}''(x_0)}}} = 1.$$

Remark

We formally have $g(x) = \tilde{g}(x)/\varepsilon$ in our description of Laplace's method.

In other words, Laplace's method become increasingly accurate as $\varepsilon \to 0$. One can think of this phenomena as follows: Most contributions to the integral arise from where the integrant looks like a quadratic function. In high dimensions, this is often referred to a concentration of measure.

Example (Laplace's method)

We consider the approximation of

$$\mathbb{E}[\cos(X)] = \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} \cos(x) e^{-x^2/(2\varepsilon)} \mathrm{d}x$$

by Laplace's method, where X is a Gaussian with mean zero and variance $\varepsilon > 0$. We first rewrite the integral as

$$\begin{aligned}\mathbb{E}[\cos(X)] &= \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} (\cos(x)+1) e^{-x^2/(2\varepsilon)} \mathrm{d}x - \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} e^{-x^2/(2\varepsilon)} \mathrm{d}x, \\ &= \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} (\cos(x)+1) e^{-x^2/(2\varepsilon)} \mathrm{d}x - 1,\end{aligned}$$

and apply Laplace's method to the remaining integral,

$$\begin{aligned}I &= \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} (\cos(x)+1) e^{-x^2/(2\varepsilon)} \mathrm{d}x, \\ &= \frac{1}{\sqrt{2\pi\varepsilon}} \int_{\mathbb{R}} e^{-x^2/(2\varepsilon)+\log(\cos(x)+1)} \mathrm{d}x.\end{aligned}$$

Example (Continued)

Hence, using the notation of previous lemma,

$$\tilde{g}(x) = \frac{x^2}{2} - \varepsilon \log(\cos(x) + 1),$$

with a unique global minimum $x_0 = 0$ for all $\varepsilon > 0$ sufficiently small, since

$$\tilde{g}'(x) = x + \varepsilon \frac{\sin(x)}{1 + \cos(x)}.$$

Hence $\tilde{g}(x_0) = -\varepsilon \log(2)$, and the second derivative satisfies

$$\tilde{g}''(x_0) = 1 + \frac{\varepsilon}{2}.$$

Finally we obtain the Laplace approximation

$$\mathbb{E}[\cos(X)] \approx \frac{e^{\log(2)}}{\sqrt{1 + \varepsilon/2}} - 1 = \frac{2}{\sqrt{1 + \varepsilon/2}} - 1,$$

for ε sufficiently small.

Example (Continued)

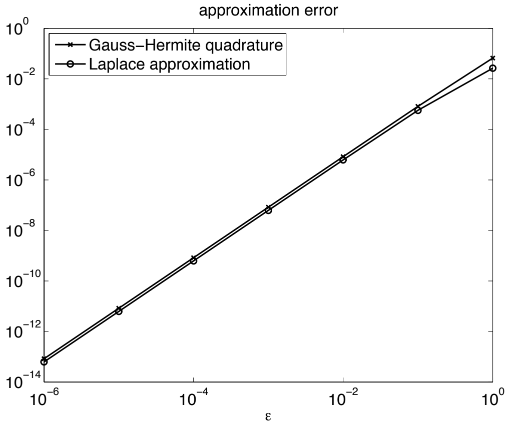


Figure: Approximation errors for Laplace's method and the 4th-order Gauss-Hermite quadrature rule. Both methods display second-order accuracy in ε. However, while Laplace's method extends easily to higher dimensions, the same does not hold true for Gauss-Hermite quadrature rules.

3.2 Monte Carlo quadrature

We now introduce an extremely flexible and powerful family of methods for approximating integrals: Monte Carlo methods. Monte Carlo methods have been developed as random and alternatives to the numerical quadrature rules considered so far.

Monte Carlo methods can be used to approximate statistics, *e.g.* expectation values $\mathbb{E}[f(X)]$, of a random variable X. We begin by discussing the special case $f(x) = x$, *i.e.* the mean.

Definition (Empirical mean)

Given a sequence X_i, $i = 1, \dots, M$, of independent random variables with identical PDF π_X (so that they have a joint measure equal to the product PDF $\prod_{i=1}^M \pi_X(x_i)$), the *empirical mean* is

$$\bar{x}_M = \frac{1}{M} \sum_{i=1}^{M} x_i,$$

for independent samples $(x_1, x_2, \dots, x_M) = (X_1(\omega), X_2(\omega), \dots, X_M(\omega))$.

Remark

The empirical mean $\bar{x}_M$ constitutes a Monte Carlo approximation to the integral $\int_{\mathbb{R}} x \mu_X(\mathrm{d}x)$.

We will generalise this approximation to general f using the concept of random probability measures.

Remark (Mean squared error)

Before we discuss the application of Monte Carlo approximations to more general integrals (in particular, expectation values), we need to understand the sense in which $\bar{x}_M$ provides an approximation to the mean $\bar{x}$ as $M \to \infty$.

We note that $\bar{x}_M$ itself is the realisation of a random variable

$$\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i .$$

We quantify the error in the empirical mean by using the mean squared error (MSE), given by

$$\begin{aligned}\mathrm{MSE}(\bar{X}_M) &= \mathbb{E}[(\bar{X}_M - \bar{x})^2] \\ &= (\mathbb{E}[\bar{X}_M] - \bar{x})^2 + \mathbb{E}\left[(\bar{X}_M - \mathbb{E}[\bar{X}_M])^2\right],\end{aligned}$$

as a measure for the approximation error.

Remark (Continued)

We have decomposed the MSE into two components: squared bias and variance. Such a decomposition is possible for any estimator and is known as the bias-variance decomposition.

The bias measures the systematic deviation of the estimator from the true expectation value, whilst the variance measures the fluctuations in the estimator due to the use of random samples.

The particular estimator $\bar{X}_M$ is called unbiased since $\mathbb{E}[\bar{X}_M] = \bar{x}$ for any $M \geq 1$. Furthermore, the variance of the estimator satisfies

$$\mathbb{E}\left[(\bar{X}_M - \mathbb{E}[\bar{X}_M])^2\right] = \frac{\sigma^2}{M}$$

provided that $\sigma^2 = \mathbb{E}[(X - \bar{x})^2] < \infty$. This result holds since the random variables X_i are independent and identically distributed.

Example (Empirical covariance matrix)

Consider the task of estimating the covariance matrix P of a random variable X with PDF π_X. We again assume that X_i, $i = 1, \ldots, M$, are independent and identically distributed random variables with PDF π_X. In analogy with the estimator X_M for the mean, we first consider the estimator

$$\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} (X_i - \bar{x}_M)(X_i - \bar{x}_M)^T.$$

However, while the estimator for the mean is unbiased, the same does not hold for $\hat{P}_M$. An unbiased estimator is given by the modification

$$\hat{P}_M = \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \bar{x}_M)(X_i - \bar{x}_M)^T,$$

Since $(M-1)/M \to 1$ as $M \to \infty$, both estimators agree in the limit $M \to \infty$. Realisations $\hat{P}_M(\omega)$ of the estimator $\hat{P}_M$, *i.e.* actual estimates based on samples x_i, will be denoted by P_M.

We now discuss the convergence of a sequence of random variables. We may first ask in what sense such a sequence of random variables may converge.

Definition (Convergence of sequences of random variables)

Let X_M, $M \geq 1$, denote a sequence of (univariate) random variables. Such a sequence converges with probability one to a random variable X if

$$\mathbb{P}(\lim_{M\to\infty} X_M = X) = 1.$$

The sequence is said to converge in probability to X if for every $\varepsilon > 0$ it holds that

$$\lim_{M\to\infty} \mathbb{P}(|X_M - X| > \varepsilon) = 0.$$

Finally, the sequence converges weakly (or in distribution) to X if

$$\lim_{M\to\infty} \mathbb{E}[g(X_M)] = \mathbb{E}[g(X)],$$

for any bounded and continuous function g.

We also recall the central limit theorem and Chebychev's inequality, which are essential tools for studying the asymptotic behavior of estimators.

Theorem (Central limit theorem)

Given a sequence X_i, $i = 1, \dots, M$, of independent univariate random variables with identical PDF π_X, mean $\bar{x}$, and finite variance σ^2, then the random variable X_M, defined as

$$X_M = \sqrt{\frac{M}{\sigma^2}} \left[\frac{1}{M} \sum_{i=1}^{M} X_i - \bar{x} \right] = \sqrt{\frac{M}{\sigma^2}} (\bar{X}_M - \bar{x}),$$

converges weakly to a Gaussian random variable with mean zero and variance one as $M \to \infty$.

Remark (Convergence of the empirical mean)

For quadrature rules, we are able to quantify convergence of expectations as the number of quadrature points goes to infinity, if the function f and PDF π_X are sufficiently smooth.

For the empirical mean $\bar{X}_M$, this is replaced by the concept of a confidence interval and its scaling as $M \to \infty$. Under repeated sampling of the empirical mean $\bar{x}_M$, a confidence interval is constructed such that the true mean value $\bar{x}$ is contained within the confidence interval with some chosen probability (typically 95% is used).

Example (Confidence intervals)

Suppose X_i, $i = 1, \ldots, M$, are independent and identically distributed random variables with mean $\bar{x}$ and variance σ^2. Consider the empirical mean $\bar{X}_M$ with $M \gg 1$ such that the distribution of $\bar{X}_M$ can be well approximated by a Gaussian with mean $\bar{x}$ and variance

$$\sigma_M^2 = \frac{\sigma^2}{M},$$

according to the central limit theorem. The constant c in the 95% *confidence interval* $I = [\bar{x}_M - c, \bar{x}_M + c]$ for a given estimate $\bar{x}_M = \bar{X}_M(\omega)$ is defined by the condition

$$\mathbb{P}\left(-c \le \bar{X}_M - \bar{x} \le c\right) \approx \int_{\bar{x}-c}^{\bar{x}+c} \mathrm{n}(x; \bar{x}, \sigma_M^2)\mathrm{d}x = 0.95.$$

It follows that

$$c \approx 1.96\sigma_M = 1.96\frac{\sigma}{M^{1/2}},$$

for a Gaussian distribution $\mathrm{N}(\bar{x}, \sigma_M^2)$. Hence, averaged over many estimates $\bar{x}_M$, the true mean $\bar{x}$ will fall within the confidence interval $I \approx [\bar{x}_M - 1.96\sigma_M, \bar{x}_M + 1.96\sigma_M]$ in 95% of the cases.

Theorem (Chebychev's inequality)

Let the random variable $\bar{X}_M$ denote the empirical mean. Then Chebychev's inequality states that

$$\mathbb{P}\left(|\bar{X}_M - \bar{x}| \geq k\sigma_M\right) \leq \frac{1}{k^2}, \quad k > 0$$

with $\sigma_M^2 = \sigma^2/M$.

Remark

For example, set $k = 1/\sqrt{0.05} \approx 4.47$. Then the 95% confidence interval I of X_M satisfies $I \subset [\bar{x}_M - k\sigma_M, \bar{x}_M + k\sigma_M]$ independently of whether X_M is Gaussian or not, and goes to zero as $M \to \infty$ with rate $p = 1/2$.

Proof.

See Theorem 3.12 on page 78 of the textbook. □

Remark (Strong and weak law of large numbers)

We also mention that the strong law of large numbers states that

$$\mathbb{P}(\lim_{M\to\infty} \bar{X}_M = \bar{x}) = 1,$$

provided $\mathbb{E}[|X|] < \infty$*. In other words, the sequence* $\bar{X}_M$ *converges with probability one to the mean* $\bar{x}$ *of the underlying (univariate) random variables* $X_i \sim \pi_X$*.*

The weak law of large numbers is a statement of convergence in probability, i.e.,

$$\lim_{M\to\infty} \mathbb{P}(|\bar{X}_M - \bar{x}| > \varepsilon) = 0$$

for all $\varepsilon > 0$*. Both laws imply that Monte Carlo approximations* $\bar{x}_M$ *converge to the mean* $\bar{x}$ *as* $M \to \infty$*.*

However, in contrast to the central limit theorem and Chebychev's inequality, they do not provide a rate of convergence.

Remark

Of course, Monte Carlo methods can be used to approximate the expectation value of functions more general than $f(x) = x$. In fact, in classical quadrature methods, for which certain smoothness assumptions have to be made on f in order to achieve higher-order convergence, Monte Carlo methods can be applied to any f as along as, for example, the second moment $\mathbb{E}[Y^2]$ of $Y = f(X)$ is bounded.

Definition (Monte Carlo approximation)

For a given PDF π_X and a measurable function f, let X_i, $i = 1, \ldots, M$, be independent random variables with identical PDF π_X. Then the Monte Carlo approximation to $\mathbb{E}[f(X)]$ is given by

$$\bar{f}_M = \frac{1}{M} \sum_{i=1}^{M} f(x_i),$$

where $(x_1, x_2, \ldots, x_M) = (X_1(\omega), X_2(\omega), \ldots, X_M(\omega))$ are the i.i.d. samples.

Example (Monte Carlo approximation)

We return to the integral

$$\bar{f} = \int_0^{1/2} 2f(x)\mathrm{d}x,$$

with f defined by $f = \phi^n$, and $n = 50$. In order to implement a Monte Carlo approximation, we need to simulate samples x_i from $\mathrm{U}[0, 1/2]$. Most scientific computing software packages contain a pseudo random number generator which will simulate samples $\hat{x}_i$ from the $\mathrm{U}[0, 1]$ distribution. A simple change of variables can be used to define the Monte Carlo approximation

$$\bar{f}_M = \frac{1}{M}\sum_{i=1}^{M} f(\hat{x}_i/2).$$

Example (Continued)

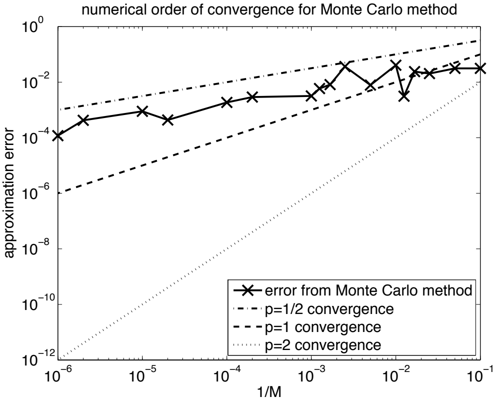


Figure: Expectation value for f defined by $f = \phi^n$ with $n = 50$ resulting from a Monte Carlo approximation using M independent samples from $X \sim \mathrm{U}[0, 1/2]$. We observe an order $1/\sqrt{M}$ convergence on average. Note that any specific error depends on the drawn samples and is therefore the outcome of a random experiment.

Definition (Generalised Monte Carlo approximation)

A generalised Monte Carlo approximation to $\mathbb{E}[f(X)]$ uses

$$\bar{f}_M = \sum_{i=1}^{M} w_i f(x_i)$$

with a different (and possibly non-independent) joint distribution with PDF $\tilde{\pi}_M(x_1, \ldots, x_M)$ for the sequence of random variables $\{X_i\}$, and non-uniform weights $\{w_i\}$ subject to the conditions $w_i > 0$, $i = 1, \ldots, M$, and

$$\sum_{i=1}^{M} w_i = 1.$$

Definition (random probability measures)

Monte Carlo methods lead to empirical measures of the form

$$\hat{\pi}_X(x) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i) \qquad x_i \sim \pi_X$$

Since the x_i' are realisations of i.i.d. random variables X_i with PDF π_X, the associated measure $\hat{\pi}_X$ is called a random probability measure.

We introduce the operator $\mathcal{S}^M$ to denote the transition from π_X to $\hat{\pi}_X$, i.e.

$$\hat{\pi}_X = \mathcal{S}^M \pi_X.$$

We like to characterise the error between π_X and its Monte Carlo approximation $\hat{\pi}_X$ in terms of a distance between random probability measures.

Expectations over the randomness in an empirical measure will be denoted by $\mathbb{E}_\Omega$.

We note, for example, that

$$\mathbb{E}_\Omega \mathbb{E}_{\hat{X}}[f] = \mathbb{E}_\Omega \left[\frac{1}{M} \sum_{i=1}^{M} f(X_i) \right] = \mathbb{E}_X[f].$$

We also introduce the shorthand

$$\mathbb{E}_X[f] = \pi_X[f], \qquad \mathbb{E}_{\hat{X}}[f] = \hat{\pi}_X[f],$$

etc. to make the dependence of the expectation value on the measure more explicit.

Definition (root mean square distance between random probability measures)

Given two random probability measures π_X and $\hat{\pi}_X$, we define their root mean square distance by

$$d(\pi_X, \hat{\pi}_X) := \sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}_\Omega |\pi_X[f] - \hat{\pi}_X[f]|^2}.$$

Lemma

Let $\mathcal{S}^M$ denote the MC resampling operator that draws realisations from M i.i.d. random variables $X_i \sim \pi_X$, then

$$d(\pi_X, \mathcal{S}^M \pi_X) \leq \frac{1}{\sqrt{M}}$$

for the associated empirical measure $\hat{\pi}_X = \mathcal{S}^M \pi_X$.

Proof.

See Explain Video and Section 3.5 from the textbook. □

3.3 Sampling Algorithms

We now discuss some practical aspects of how to implement Monte Carlo methods. Monte Carlo methods are easy to implement provided that it is possible to simulate the random variables $\{X_i\}$.

Most scientific computing software packages provide (pseudo) random number generators for the uniform and the standard Gaussian distribution. The following techniques will allow us to move beyond those distributions.

Lemma (Transform method for sampling)

Let X and Y be random variables, and let T be a transport map that defines a coupling between them. Let $\{x_i\}$ be a sequence of independent and identically distributed samples from X. Then $\{T(x_i)\}$ is a sequence of independent and identically distributed samples from Y.

Proof.

First, it is clear that $\{T(X_i)\}$ are independent, identically distributed, since $\{X_i\}$ are. Further, Y and $T(X)$ have the same law (weakly) if $\mathbb{E}[f(Y)] = \mathbb{E}[f(T(X))]$ for all suitable test functions f (*i.e.*, functions for which the expectations are finite). This is guaranteed by the definition of transport maps. □

Example (Transform method for uniform random numbers)

Given a univariate random variable X with PDF π_X and cumulative distribution function

$$F_X(x) = \int_{-\infty}^{x} \pi_X(x')\mathrm{d}x',$$

we first draw M samples u_i from independent random variables U_i with uniform distribution $\mathrm{U}[0,1]$ and then solve the implicit equation

$$F_X(x_i) = u_i,$$

for $x_i \in \mathbb{R}$, $i = 1, \ldots, M$. The samples x_i provide realisations from M independent and identically distributed random variables with PDF π_X.

Example (Gaussian random samples)

We define a transport map between a pair of independent Gaussian random variables $X_1,\ X_2 \sim \mathrm{N}(0,1)$ and a pair of independent uniform random variables $U_1,\ U_2 \sim \mathrm{U}[0,1]$, as follows,

$$X_1 = \sqrt{-2\ln U_1}\sin(2\pi U_2), \quad X_2 = \sqrt{-2\ln U_1}\cos(2\pi U_2).$$

This is a transport map since

$$\begin{aligned}\mathrm{d}x_1\mathrm{d}x_2 &= \det\left(\begin{array}{cc} -\frac{\sin(2\pi u_2)}{u_1\sqrt{-2\ln u_1}} & 2\pi\sqrt{-2\ln u_1}\cos(2\pi u_2), \\ -\frac{\cos(2\pi u_2)}{u_1\sqrt{-2\ln u_1}} & -2\pi\sqrt{-2\ln u_1}\sin(2\pi u_2) \end{array}\right)\mathrm{d}u_1\mathrm{d}u_2, \\ &= \left(2\pi\frac{\cos(2\pi u_2)^2}{u_1} + 2\pi\frac{\sin(2\pi u_2)^2}{u_1}\right)\mathrm{d}u_1\mathrm{d}u_2 = \frac{2\pi}{u_1}\mathrm{d}u_1\mathrm{d}u_2.\end{aligned}$$

We also find that $u_1 = e^{-(x_1^2+x_2^2)/2}$, and therefore

$$\mathrm{d}u_1\mathrm{d}u_2 = \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\mathrm{d}x_1\,\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}\mathrm{d}x_2.$$

The resulting sampling method is called the Box-Muller algorithm.

Example (Multivariate Gaussian random samples)

Mathematical software packages typically provide pseudo random number generators for multivariate Gaussian random variables with mean zero and covariance matrix $P = I$.

As an example, consider two multivariate Gaussian distributions $\mathrm{N}(\bar{x}_1, P_1)$ and $\mathrm{N}(\bar{x}_2, P_2)$ in $\mathbb{R}^{N_x}$ with means $\bar{x}_1 = 0$ and $\bar{x}_2 \neq 0$, and covariance matrices $P_1 = I$ and $P_2 \neq I$, respectively. The coupling

$$X_2 = \bar{x}_2 + P_2^{1/2} P_1^{-1/2}(X_1 - \bar{x}_1)$$

leads then to the well-known transformation

$$X_2 = \bar{x}_2 + P_2^{1/2} X_1.$$

This transformation can be used to generate samples from a Gaussian distribution $\mathrm{N}(\bar{x}, P)$ based on available random number generators.

Remark (Transform method and optimal couplings)

As we have already discussed in Chapter 2 in the context of optimal transportation, there exist mappings T which transform a random variable X with PDF π_X into another random variable Y with PDF π_Y under rather general assumptions.

Furthermore, those maps can be generated by convex potentials ψ, i.e., $Y = T(X) = \nabla_x \psi(X)$. This result allows us to extend the transform method to more general multivariate random variables, at least theoretically.

Finding approximations to $T(X)$ is an active area of research, which is beyond the scope of the course. But see Section 5.3 of the book. We will also return to this topic when introducing the ensemble Kalman filter in Chapter 7.

Remark (Importance sampling)

Whilst discussing Gauss-Hermite quadrature, we explored reformulating integrals with respect to a PDF π_X as

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^{N_x}} f(x) \frac{\pi_X(x)}{\pi_{X'}(x)} \pi_{X'}(x) \mathrm{d}x.$$

This formula is also useful for generalised Monte Carlo approximation, in the case where $\pi_{X'}(x)$ denotes the PDF of a random variable X' which we can easily draw samples from. We then use weights

$$w_i = \frac{\pi_X(x_i)/\pi_{X'}(x_i)}{\sum_{j=1}^{M} \pi_X(x_j)/\pi_{X'}(x_j)},$$

where $\{x_i\}$ are samples from M independent and identically distributed random variables X_i' with PDF $\pi_{X'}$. This is referred to as importance sampling, our second type of Monte Carlo method.

Definition (Importance sampling)

Let X be a random variable with PDF π_X and X' be a second random variable with PDF $\pi_{X'}$ such that

$$\pi_X(x) = 0 \quad \text{if} \quad \pi_{X'}(x) = 0,$$

i.e., the measure $\mu_X(\mathrm{d}x) = \pi_X(x)\mathrm{d}x$ is *absolutely continuous* with respect to $\mu_{X'}(\mathrm{d}x) = \pi_{X'}(x)\mathrm{d}x$. Assume that we wish to obtain expectation formulas for X, but that it is much easier or efficient to sample from X'. Then, the importance sampling estimate of $\mathbb{E}[f(X)]$ is given by

$$\mathbb{E}[f(X)] \approx \sum_{i=1}^{M} w_i f(x_i'),$$

where $\{x_i'\}$ are M independent samples from the random variable X'. The weights $\{w_i\}$ are given by

$$w_i \propto \frac{\pi_X(x_i')}{\pi_{X'}(x_i')},$$

where the constant of proportionality is chosen such that $\sum_i w_i = 1$.

Example (Variance reduction)

We consider the approximation of the expectation value

$$\mathbb{E}[e^{-10X}\cos(X)] = \int_0^1 e^{-10x}\cos(x)\mathrm{d}x, \tag{3}$$

with respect to $X \sim \mathrm{U}[0,1]$. Its analytic value is given by

$$\mathbb{E}[e^{-10X}\cos(X)] = \frac{10}{101} - \frac{10\cos(1)-\sin(1)}{101e^{10}}.$$

We can approximate this value by drawing M independent samples u_i from the uniform distribution $\mathrm{U}[0,1]$, *i.e.*,

$$\bar{x}_M = \frac{1}{M}\sum_{i=1}^{M}\cos(u_i)e^{-10u_i}.$$

Example (Continued)

Since e^{-10x} decays very rapidly away from zero, it seems reasonable to replace the uniform samples u_i by samples $x_i \in [0,1]$, which are concentrated around zero. For example, we may take X' to have PDF

$$\pi_{X'}(x) = \begin{cases} \frac{10e^{-10x}}{1-e^{-10}} & \text{if } x \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Samples x_i from this distribution can be generated using the transform method with cumulative distribution function

$$F_{X'}(x) = \begin{cases} 1 & \text{if } x \geq 1, \\ \frac{1-e^{-10x}}{1-e^{-10}} & \text{if } x \in [0,1), \\ 0 & \text{otherwise.} \end{cases}$$

We obtain the explicit formula

$$x_i = -0.1 \log(1 - u_i + u_i e^{-10})$$

for independent samples u_i from the uniform distribution $\mathrm{U}[0,1]$. The expectation value (3) is now approximated by

Example (Continued)

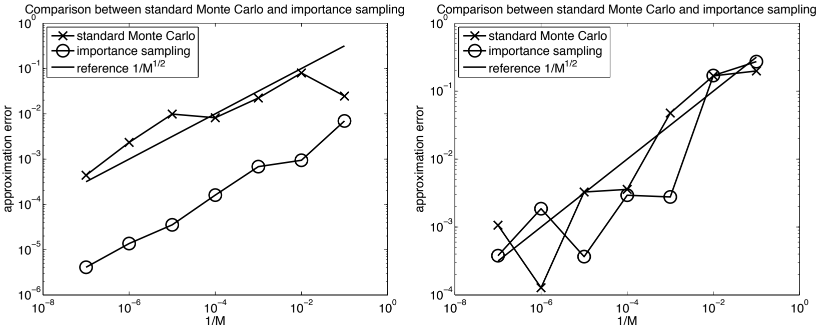


Figure: Comparison between a standard Monte Carlo approximation to the desired expectation value and an importance sampling Monte Carlo approximation. The plot on the left shows that an error reduction by a factor of about one hundred is achieved in this case. However, the same importance sampling approach applied to the integrand $f(x) = e^{-10x}\sin(x)$ does not lead to an improvement. See figure on the right. This result indicates that importance sampling for reducing the variance of estimators needs to be handled with great care.

We wish to produce i.i.d. samples X_i from a PDF π_X while only be able to sample from a PDF π_P. We assume we can find a constant $m > 0$ such that

$$\frac{\pi_X(x)}{m\,\pi_P(x)} \leq 1 \qquad \text{for all} \quad x\,.$$

Definition (Rejection Sampling)

Rejection sampling consists of the following algorithm:

For $i = 1, \ldots, M$ do

(i) Use a random number generator to generate a sample $x \in \mathbb{R}^{N_x}$ from a random variable X' with PDF π_P and draw a u from the uniform distribution $\mathrm{U}[0,1]$.

(ii) If

$$u < \frac{\pi_X(x)}{m\,\pi_P(x)},$$

then set $x_i = x$, increase i by one, and go back to (i). Otherwise reject the proposal, x, and return to (i).

Lemma (Consistency of rejection sampling)

Rejection sampling generates samples x_i, $i = 1, \ldots, M$, from M independent and identically distributed random variables X_i with PDF π_X.

Proof.

See Lemma 3.23 from the textbook. □

Remark (Rejection sampling - geometric interpretation)

Step (i) of the algorithm generates pairs (x, u). We introduce the variable $y = u\, m\, \pi_P(x)$ and plot the associated (x, y) pairs in the plane. These pairs fill the area underneath $f(x) := m\, \pi_P(x)$ uniformly.

In step (ii), we only retain those samples x for which the associated y is also underneath the graph of $g(x) := \pi_X(x)$. Hence the ratio of generated to accepted samples converges to the ratio of the two definite integrals associated with f and g.

Hence, we may apply rejection sampling to any pair of functions $f(x) > g(x) \geq 0$. If the area enclosed by f is known, then rejection sampling can be used to approximate the area defined by g.

Example (Approximating π)

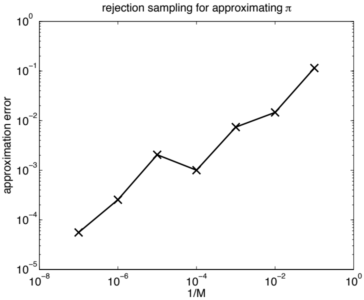
We consider the following specific example. Define f and g by

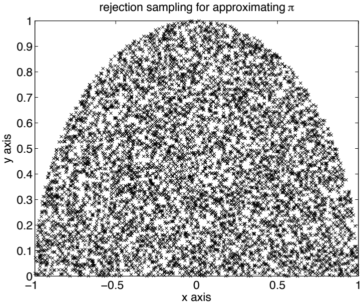
$$f(x) = \begin{cases} 1 & \text{if } x \in [-1,1], \\ 0 & \text{otherwise,} \end{cases} \qquad g(x) = \begin{cases} \sqrt{1-x^2} & \text{if } x \in [-1,1], \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\int_{-1}^{1} f(x)\mathrm{d}x = 2 \quad \text{and} \quad \int_{-1}^{1} g(x)\mathrm{d}x = \frac{\pi}{2},$$

and rejection sampling can be used to approximate π.





We finally combine importance sampling with a resampling step in order to produce equally weighted samples.

Definition (Resampling)

Let μ_M be an empirical measure of the form

$$\mu_M(\mathrm{d}x) = \sum_{i=1}^{M} w_i \delta(x - x_i)\mathrm{d}x$$

Resampling replaces μ_M by another empirical measure $\tilde{\mu}_M$ of the form

$$\tilde{\mu}_M(\mathrm{d}x) = \frac{1}{L}\sum_{i=1}^{M} \xi_i \delta(x - x_i)\mathrm{d}x,$$

where $L > 0$ is a positive integer, and the weights $\{\xi_i\}$ are realisations of univariate discrete random variables $\Xi_i : \Omega \to \mathcal{S}$ with integer-valued realisations in $\mathcal{S} = \{0, 1, \dots, L\}$ subject to

$$\sum_{i=1}^{M} \xi_i = L.$$

Lemma (Unbiased resampling)

Take the sequence $\Xi = \{\Xi_i\}$ of discrete random variables such that

$$\mathbb{E}[\Xi_i] = L w_i.$$

Then,

$$\mathbb{E}\left[\frac{1}{L}\sum_{i=1}^{M} \Xi_i f(x_i)\right] = \sum_{i=1}^{M} w_i f(x_i).$$

This means that, upon averaging over all possible realisations of $\{\xi_i\}$, the resampled empirical measure produces the same statistics as μ_M. Any errors are due to the variance in the estimator.

Proof.

See Lemma 3.26 from the textbook. □

Remark (Multinomial distribution)

The interpretation of unbiased resampling is that we have replaced an empirical measure with non-uniform weights w_i by a new empirical measure with uniform weights $\tilde{w}_j = 1/L$ and each sample x_i being replaced by ξ_i identical offspring. The total number of offspring is equal to L.

For example, the offspring $\{\xi_i\}_{i=1}^M$ may follow a multinomial distribution defined by

$$\mathbb{P}(\xi_i = n_i, i = 1, \ldots, M) = \frac{M!}{\prod_{i=1}^M n_i!} \prod_{i=1}^M (w_i)^{n_i},$$

with $n_i \geq 0$ such that $\sum_{i=1}^M n_i = L$.

We introduce the notation $\mathrm{Mult}(L; w_1, \ldots, w_M)$ to denote the multinomial distribution of L independent trials, where the outcome of each trial is distributed among M possible outcomes according to probabilities $\{w_i\}_{i=1}^M$. The following algorithm draws random samples from $\mathrm{Mult}(L; w_1, \ldots, w_M)$.

Definition (Multinomial samples)

The integer-valued variable $\bar{\xi}_i$, $i = 1, \ldots, M$, is set equal to zero initially.
For $l = 1, \ldots, L$ do:

(i) Draw a number $u \in [0, 1]$ from the uniform distribution $\mathrm{U}[0, 1]$.

(ii) Determine the integer $i^* \in \{1, \ldots, M\}$ which satisfies

$$i^* = \arg \min_{i \geq 1} \sum_{j=1}^{i} w_j \geq u.$$

(iii) Increment $\bar{\xi}_{i^*}$ by one.

The final integers $\{\bar{\xi}_i\}$ are distributed according to $\mathrm{Mult}(L; w_1, \ldots, w_M)$.

Theorem (Convergence of multinomial resampling)

We have

$$d(\mu_M, \tilde{\mu}_M) \leq \frac{1}{\sqrt{L}}$$

for $\tilde{\mu}_M$ defined by multinomial resampling.

Definition (Residual resampling)

Given a weighted empirical measure residual resampling generates a total of M offspring with

$$\xi_i = \lfloor M w_i \rfloor + \bar{\xi}_i,$$

offspring for each sample x_i, $i = 1, \ldots, M$. Here $\lfloor x \rfloor$ denotes the integer part of x and the values $\{\bar{\xi}_i\}$ follow the multinomial distribution with

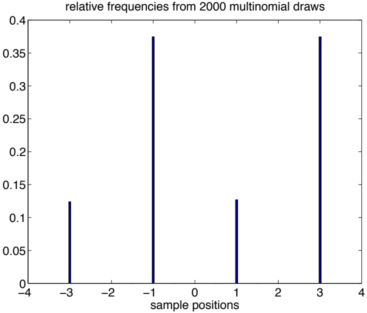
$$L = M - \sum_{i=1}^{M} \lfloor M w_i \rfloor$$

and new weights

$$\overline{w}_i = \frac{M w_i - \lfloor M w_i \rfloor}{\sum_{j=1}^{M} (M w_j - \lfloor M w_j \rfloor)}.$$

Example (Multinomial resampling)

Consider four samples $x_1 = -3$, $x_2 = -1$, $x_3 = 1$, and $x_4 = 3$ with associated weights $w_1 = 1/16$, $w_2 = 3/16$, $w_3 = 5/16$, and $w_4 = 7/16$, respectively. Residual resampling leads to $\lfloor Mw_i \rfloor$ equal to zero for $i = 1, 2$ and $\lfloor Mw_i \rfloor = 1$ for $i = 3, 4$. The new weights $\overline{w}_i$ are then given by $\overline{w}_{1,3} = 1/8$, $\overline{w}_{2,4} = 3/8$, and $L = 2$. Drawing a total of two thousand samples from the associated multinomial distribution, we find that the relative frequency of the outcomes indeed reproduces the probabilities given by the probability vector $\overline{w} = (1/8, 3/8, 1/8, 3/8)$.


relative frequencies from 2000 multinomial draws
sample positions