Bayesian statistics, integration (MATH10093)

- Basic Bayesian statistics
  - Prior, Observations, Posterior, and Bayes' formula
  - Computational approaches
  - Priors
  - Example: Beta&Binomial
  - Credible intervals vs Confidence intervals
- Integration methods
  - Deterministic integration
  - Monte Carlo integration

Bayesian statistics

Prior, Observations, Posterior; concepts

In a basic Bayesian model, the main difference to frequentistic modelling is that the unknown parameters are now treated as *unobserved random variables* instead of unknown constants.

- The distribution for the parameters is called the *prior distribution*.
- The observation distribution conditionally on the parameters is often referred to as the observation *likelihood*, although this terminology is not universally accepted as proper.
- The *conditional distribution* for the parameters, given the observations, is called the *posterior distribution*

Prior, Observations, Posterior; mathematical probability theory

In mathematical terms, with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ and observations $\boldsymbol{y} = (y_1, \ldots, y_n)$:

- Prior density: $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, often abbreviated to just $p(\boldsymbol{\theta})$. If the prior distributions have any (fixed and known) parameters, those are often referred to as *hyper*-parameters.
- Conditional observation likelihood/density/probability mass function: $(\boldsymbol{y}|\boldsymbol{\theta}) \sim p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y})$, usually written $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Posterior density: $(\boldsymbol{\theta}|\boldsymbol{y}) \sim p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})}$ Bayes' formula
- *Marginal* observation denstity/probability mass function: $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$

For discrete variables or parameters, the integrals are replaced by sums.

Computational approaches

The Bayesian statistics framework is appealing in that (almost) everything flows from general probability theory, once a probabilistic model has been formulated. The difficuly lies in how to actually calculate the involved densities, expectations, variances, and probabilities.

- Analytical derivation; some results are easy to derive by hand (see example on later slide)
- Approximate deterministic numerical integration; deterministic integration schemes, Laplace approximation
- Basic simulation and integration; Monte Carlo integration, importance sampling
- Advanced simulation and integration; Markov chain Monte Carlo (MCMC) simulation (beyond the scope of this course)

Priors

One of the challenges in Bayesian modelling is how to specify realistic priors.

- One approach is to elicit information from application area experts and encode their *subjective* information as probability distributions.
- Other approaches include so called *objective* priors, *vague* priors, and *flat* priors (often incorrectly called "uninformative" priors) with the aim of regularising otherwise unstable results but without imposing too strict limitations.
- Before the proliferation of computational techniques in the 1990's, and indeed still today, most priors were chosen to be *conjugate* with respect to the observation distribution (the prior and posterior belong to the same distribution family), to allow analytical solutions. Modern computational methods has removed much of the motivation for such priors; they can simplify calculations, but need not be used if more realistic priors are available.

Example: Binomial observations with Beta prior

Consider the following Bayesian model (version with a single observation seen in Statistical Methodology lectures) for Binomial observations $\boldsymbol{y} = (y_1, \ldots, y_K)$ with common probability parameter ϕ:

- Parameter prior distribution: $\phi \sim \text{Beta}(a, b)$ for some *hyper*-parameters $a, b > 0$.
- Observations: $(y_k|\phi) \sim \text{Bin}(N_k, \phi)$, independent over $k = 1, \ldots, K$, known N_k.
- Posterior distribution for ϕ: $\text{Beta}\left[a + \sum_{k=1}^K y_k, b + \sum_{k=1}^K (N_k - y_k)\right]$

Proof: Moving all factors that do not depend on ϕ into a normalisation constant, we get

$$p(\phi|\boldsymbol{y}) = \frac{p(\phi)p(\boldsymbol{y}|\phi)}{p(\boldsymbol{y})} \propto p(\phi) \prod_{k=1}^K p(y_k|\phi) = \frac{\phi^{a-1}(1-\phi)^{b-1}}{B(a,b)} \prod_{k=1}^K \binom{N_k}{y_k} \phi^{y_k}(1-\phi)^{N_k - y_k}$$
$$\propto \phi^{a-1+\sum_k y_k}(1-\phi)^{b-1+\sum_k (N_k - y_k)}.$$

This takes the same form as a Beta distribution density (this shows that Beta is a conjugate prior for the probability parameter in a Binomial distribution), so the result follows by identifying the parameters.

Bayesian credible intervals

A Bayesian $(1-\alpha)\cdot 100$ *percent credible interval* $\mathsf{CI}_\theta = (L, U)$ for a parameter θ is an interval computed from the posterior distribution, such that $\mathsf{P}(L < \theta < U \mid \boldsymbol{y}) \geq 1-\alpha$.

- Often, L and U are chosen so that the tail probabilities outside the interval are both $= \alpha/2$.
- Another option is to choose the smallest set or interval with the required probability. This is achieved by finding a value c such that the set $\mathsf{CI}_\theta = \{\theta; p(\theta|\boldsymbol{y}) > c\}$ fulfils $\mathsf{P}(\theta \in \mathsf{CI}_\theta|\boldsymbol{y}) \geq 1-\alpha$. The resulting CI_θ is called a *highest posterior density* (HPD) region, and is well defined also for multi-dimensional parameter vectors.
- When c is chosen as large as possible, the HPD is the smallest credible region possible for the given model parameterisation. However, due to how probability densities are changed under transformation of the random variable, HPD region constructions are not invariant under reparameterisation.

Credible region vs. Confidence region

Posterior credible region (Bayesian concept)

A *level* $1-\alpha$ *posterior credible region* for θ, $C_\theta(\boldsymbol{y})$, is a set (usually an interval) such that

$$\mathrm{P}_{\theta\sim p(\theta|\boldsymbol{y})}\left(\theta \in C_\theta(\boldsymbol{y})\right) \geq 1-\alpha$$

for the fixed set of observations $\boldsymbol{y}$. The probability is a direct statement about our posterior beliefs about the random variable θ.

Confidence region (Frequentist concept)

A *level* $1-\alpha$ *confidence region* procedure for θ, $C_\theta(\boldsymbol{y})$, generates sets (usually intervals) in such a way that

$$\mathrm{P}_{\boldsymbol{y}\sim p(\boldsymbol{y};\theta)}\left(\theta \in C_\theta(\boldsymbol{y})\right) \geq 1-\alpha$$

for every possibe θ (or at least for the true value). The probability statement concerns the confidence region construction *procedure* under repeated experiments; the observations $\boldsymbol{y}$ are random.

Numerical integration

Numerical integration

Non-statistical and statistical problems involving integrals:

- Integrate $f(\cdot)$ over some interval/set/domain Ω:

$$I = \int_{\Omega} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

- Compute a *marginal density* $p_Y(y)$ from a *joint density* $p_{X,Y}(x, y)$, $x \in \Omega$:

$$p_Y(y) = \int_{\Omega} p_{X,Y}(x, y) \, \mathrm{d}x$$

- For some function $\phi(\boldsymbol{x})$ and density $p_X(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, compute the expectation

$$\mathsf{E}_X[\phi(\boldsymbol{x})] = \int_{\Omega} \phi(\boldsymbol{x}) p_X(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

- Challenges: High-dimensional Ω and computationally expensive integrands

Deterministic integration methods

- General idea: Find points and weights $(\boldsymbol{x}_k, w_k)$ such that

$$I = \int_\Omega f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \sum_k f(\boldsymbol{x}_k) w_k = \widehat{I}$$

- Midpoint rule for $\Omega = (a, b)$: $x_k = a + (k - \frac{1}{2})\frac{b-a}{N}$, $w_k = \frac{b-a}{N}$, for $k = 1, \ldots, N$.
- Trapezoidal rule
- Simpson's rule
- Gaussian quadrature: For some special function $\psi(x)$, there is a specific set $\{(x_k, w_k)\}_N$ such that $\int g(x)\psi(x)\,\mathrm{d}x = \sum_{k=1}^N g(x_k) w_k$ for *any polynomial* of degree $\leq 2N - 1$. Use

$$\widehat{I} = \sum_{k=1}^N \frac{f(x_k)}{\psi(x_k)} w_k; \quad \text{good approximation of } I \text{ if } f(x)/\psi(x) \text{ is close to a polynomial.}$$

- Laplace approximation for $f(\boldsymbol{x}) \geq 0$ on $\Omega = \mathbb{R}^d$: Scale a Gaussian density to match the amplitude and shape at the mode of the integrand:

$$\widehat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}), \quad \widehat{\boldsymbol{H}} = -\nabla^2 \log f(\boldsymbol{x})\big|_{\boldsymbol{x}=\widehat{\boldsymbol{x}}}, \quad \widehat{w} = \frac{(2\pi)^{d/2}}{(\det \widehat{\boldsymbol{H}})^{1/2}}, \quad \widehat{I} = f(\widehat{\boldsymbol{x}})\widehat{w}$$

Monte Carlo integration

- We know that if we have a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from a distribution with density $p_X(\boldsymbol{x})$ we can estimate the expectation $\mathsf{E}_{\boldsymbol{x} \sim p_X}[\phi(\boldsymbol{x})] = \int \phi(\boldsymbol{x}) p_X(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ with the average of $\phi(\boldsymbol{x}_k)$:

$$\widehat{\mathsf{E}}_{\boldsymbol{x} \sim p_X}[\phi(\boldsymbol{x})] = \frac{1}{N} \sum_{k=1}^{N} \phi(\boldsymbol{x}_k)$$

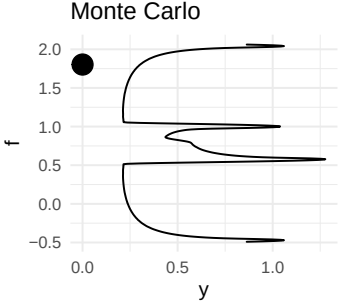
- For bounded Ω we can write a Monte Carlo integration scheme for

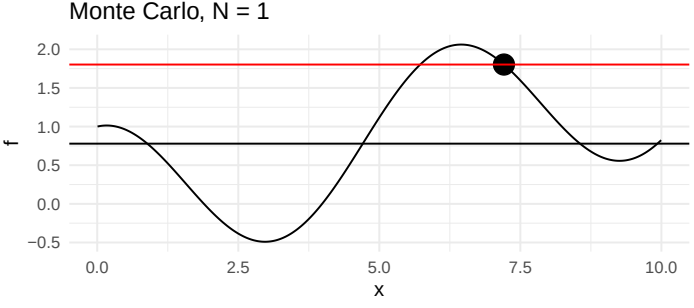
$$I = \int_\Omega f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_\Omega |\Omega| f(\boldsymbol{x}) \frac{1}{|\Omega|} \, \mathrm{d}\boldsymbol{x} \approx \frac{|\Omega|}{N} \sum_{k=1}^{N} f(\boldsymbol{x}_k), \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \mathsf{Unif}(\Omega)$$

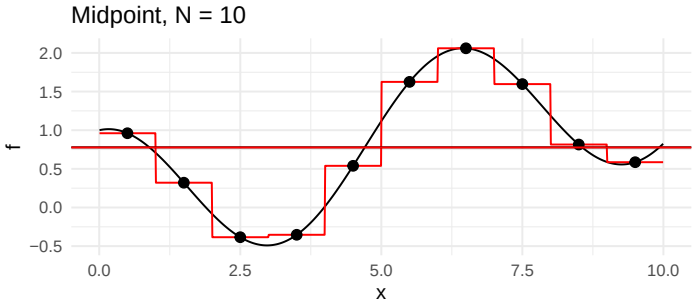
- Plain MC estimates are unbiased, with variance

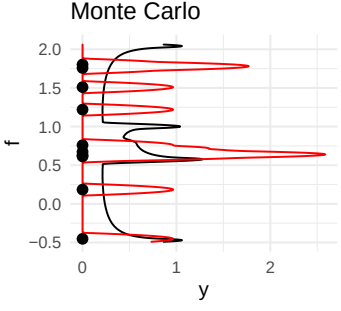
$$\mathsf{Var}_{\{\boldsymbol{x}_k \sim \mathsf{Unif}(\Omega)\}} \left[\frac{|\Omega|}{N} \sum_{k=1}^{N} f(\boldsymbol{x}_k) \right] = \frac{|\Omega|^2}{N^2} N \, \mathsf{Var}_{\boldsymbol{x} \sim \mathsf{Unif}(\Omega)} [f(\boldsymbol{x})] \propto N^{-1}$$

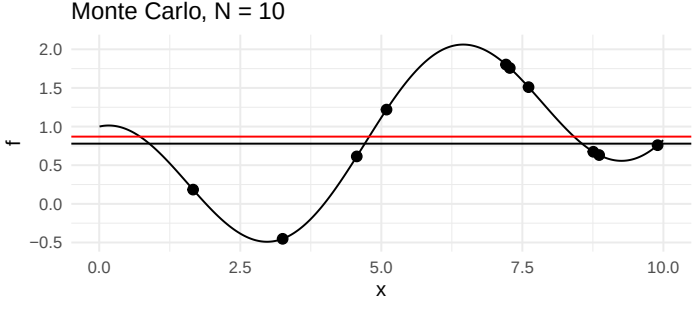

Midpoint, N = 1
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

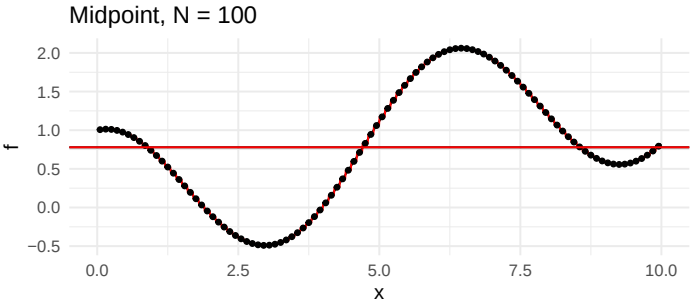

Monte Carlo
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
0.5
1.0
y

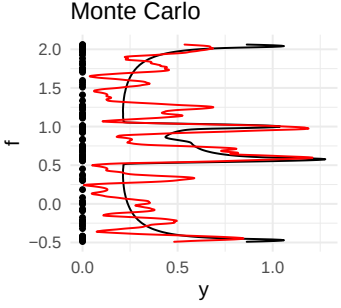

Monte Carlo, N = 1
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

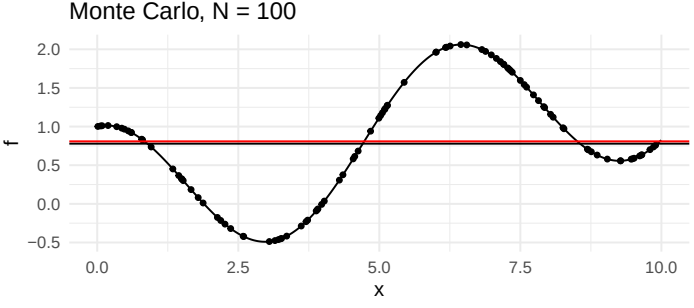

Midpoint, N = 10
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

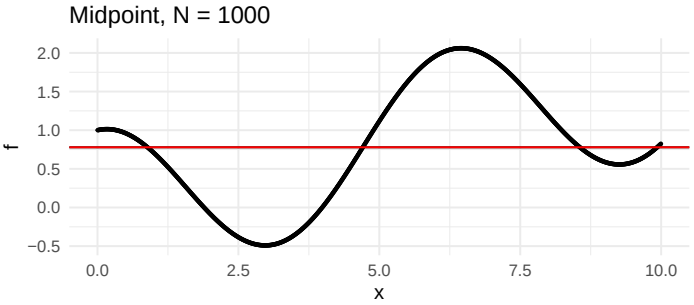

Monte Carlo
2.0
1.5
1.0
0.5
0.0
−0.5
f
0
1
2
y

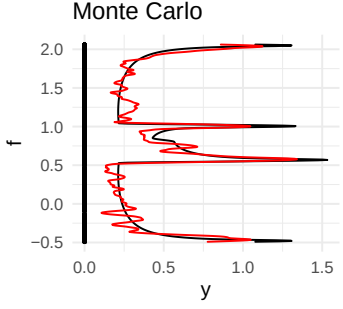

Monte Carlo, N = 10
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

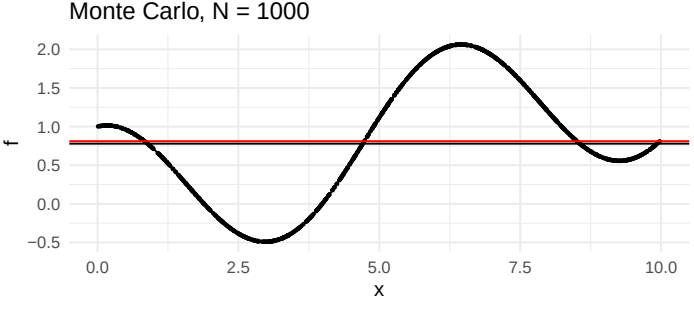

Midpoint, N = 100
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

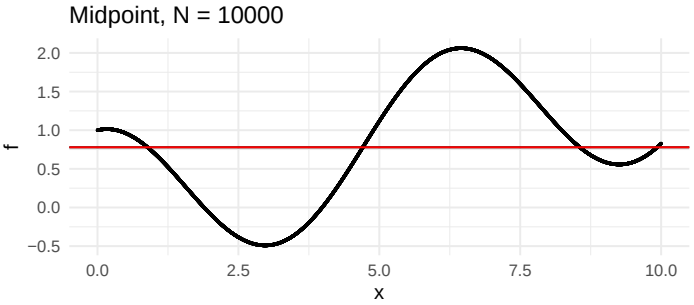

Monte Carlo
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
0.5
1.0
y

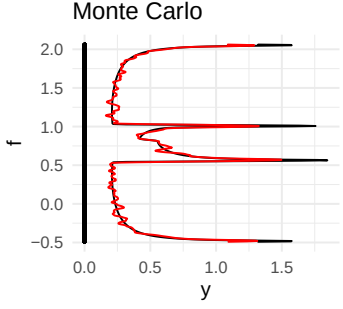

Monte Carlo, N = 100
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

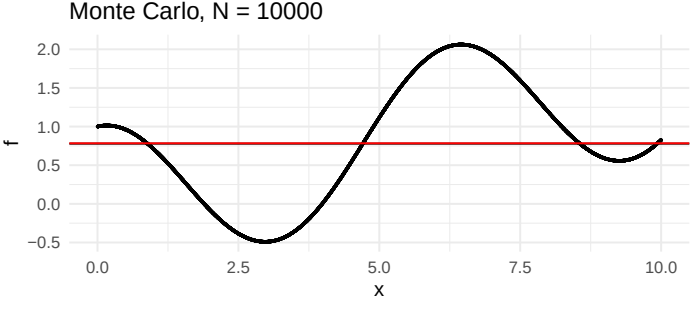

Midpoint, N = 1000
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x


Monte Carlo
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
0.5
1.0
1.5
y


Monte Carlo, N = 1000
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x


Midpoint, N = 10000
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x


Monte Carlo
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
0.5
1.0
1.5
y


Monte Carlo, N = 10000
2.0
1.5
1.0
0.5
0.0
−0.5
f
0.0
2.5
5.0
7.5
10.0
x

Importance sampling

- We can often reduce the MC variance by sampling with a density $p_Z(\cdot)$ that is similar to $\phi(\boldsymbol{x})p_X(\boldsymbol{x})$ or $f(\boldsymbol{x})$, a technique called *Importance sampling*:

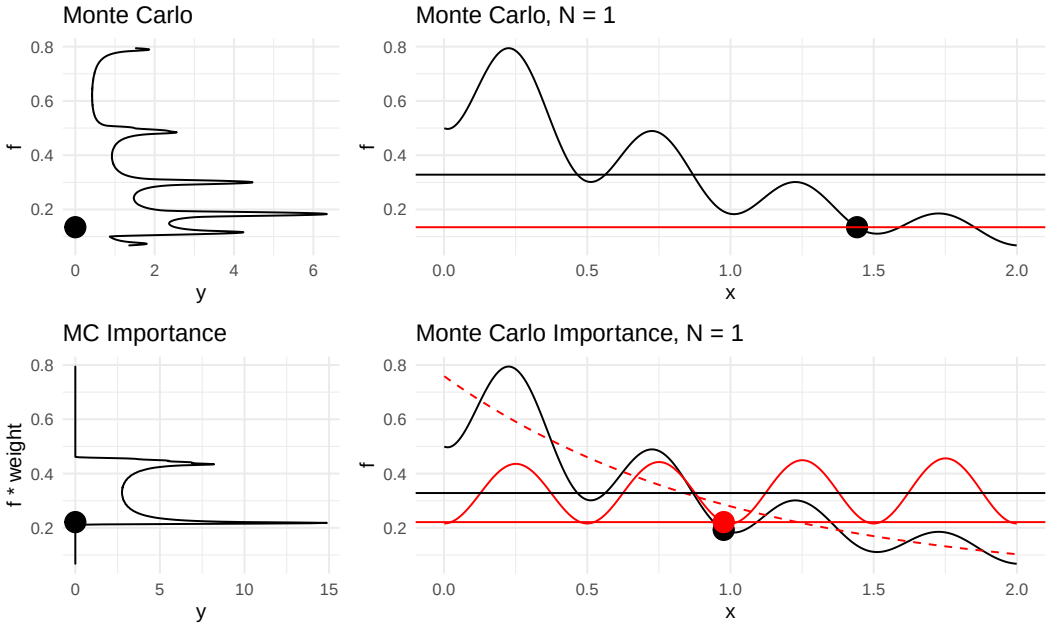
$$I = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{f(\boldsymbol{x})}{p_Z(\boldsymbol{x})} p_Z(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathsf{E}_{\boldsymbol{x}\sim p_Z}\left[\frac{f(\boldsymbol{x})}{p_Z(\boldsymbol{x})}\right] \approx \frac{1}{N}\sum_{k=1}^{N}\frac{f(\boldsymbol{x}_k)}{p_Z(\boldsymbol{x}_k)} = \widehat{I}$$

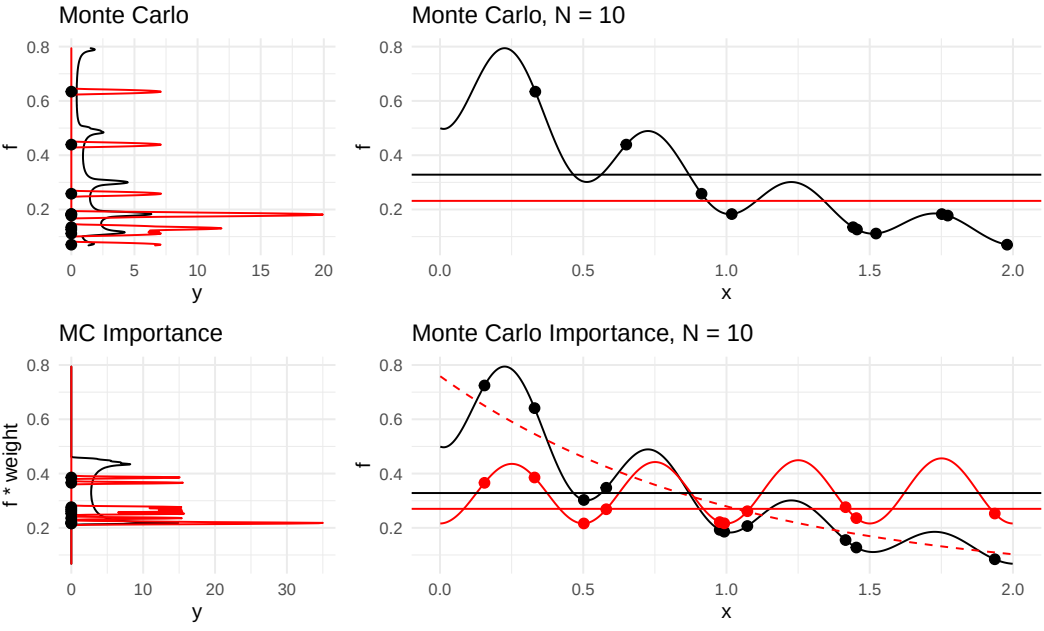
  where $\boldsymbol{x}_k$, $k = 1, \ldots, N$ are sampled from the $p_Z(\cdot)$ density.

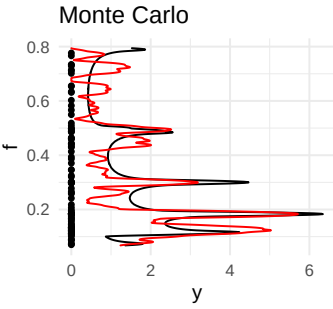
- The variance is still $\propto N^{-1}$, but with a potentially much smaller constant:

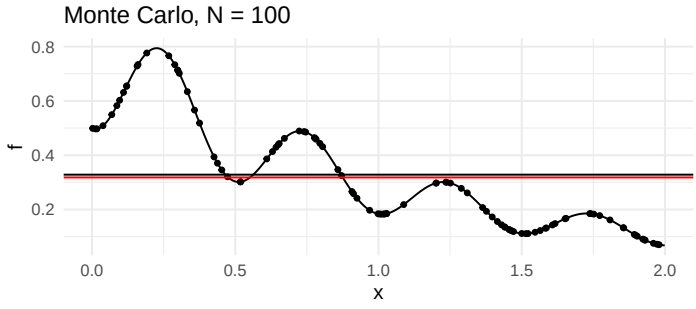
$$\mathsf{Var}_{\{\boldsymbol{x}_k\sim p_Z\}}\left[\widehat{I}\right] = \mathsf{Var}_{\{\boldsymbol{x}_k\sim p_Z\}}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{f(\boldsymbol{x}_k)}{p_Z(\boldsymbol{x}_k)}\right] = \frac{1}{N}\mathsf{Var}_{\boldsymbol{x}\sim p_Z}\left[\frac{f(\boldsymbol{x})}{p_Z(\boldsymbol{x})}\right]$$

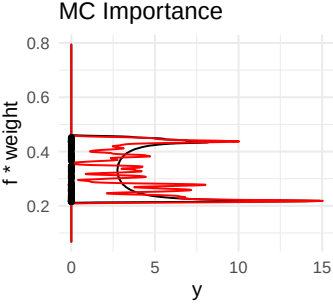
- Note: If we were able to choose $p_Z(\boldsymbol{x}) \propto f(\boldsymbol{x})$, the variance would be zero!

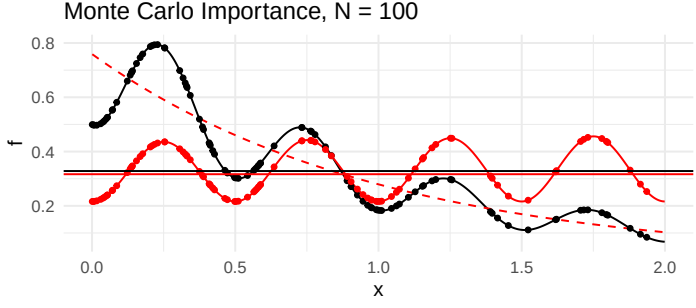

Monte Carlo
f
y
Monte Carlo, N = 1
f
x
MC Importance
f * weight
y
Monte Carlo Importance, N = 1
f
x

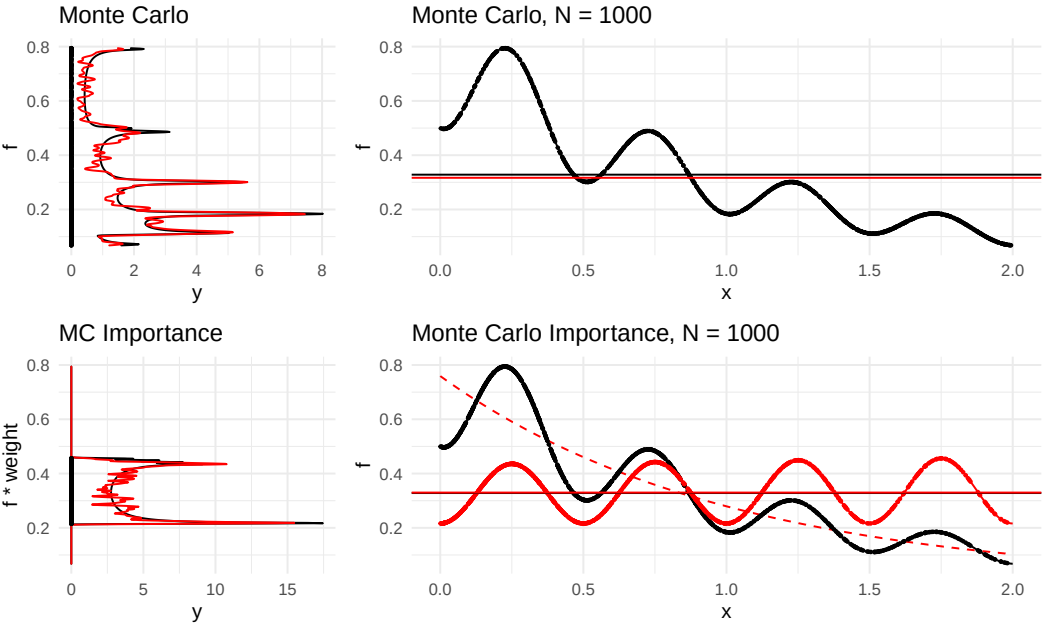

Monte Carlo
f
y
0.8
0.6
0.4
0.2
0
5
10
15
20
Monte Carlo, N = 10
f
x
0.0
0.5
1.0
1.5
2.0
MC Importance
f * weight
y
30
Monte Carlo Importance, N = 10

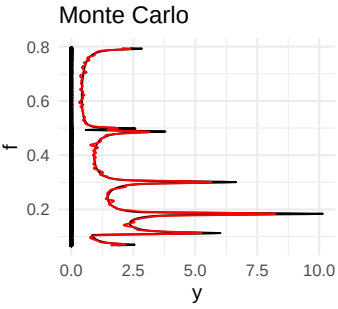

Monte Carlo
f
y
0.8
0.6
0.4
0.2
0
2
4
6

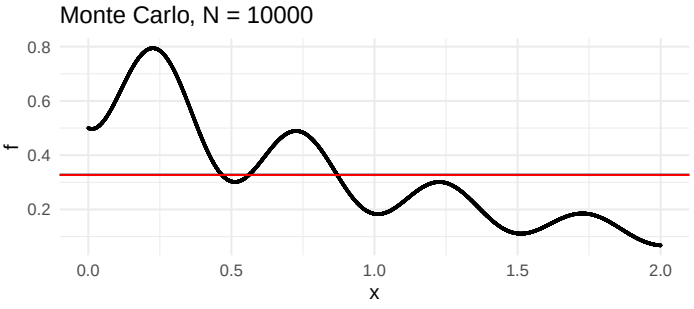

Monte Carlo, N = 100
f
x
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0

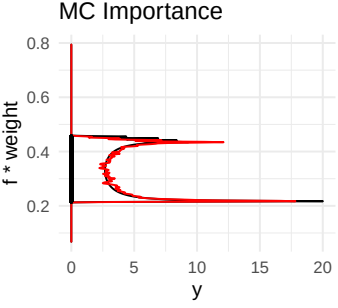

MC Importance
f * weight
y
0.8
0.6
0.4
0.2
0
5
10
15

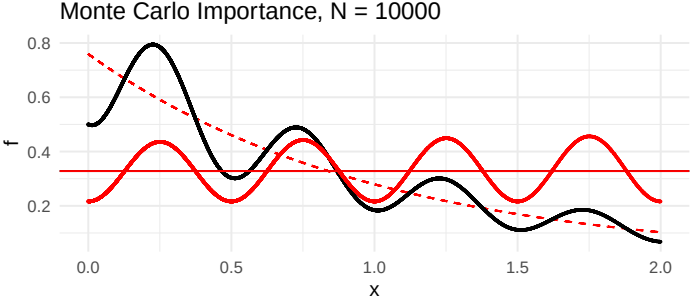

Monte Carlo Importance, N = 100
f
x
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0

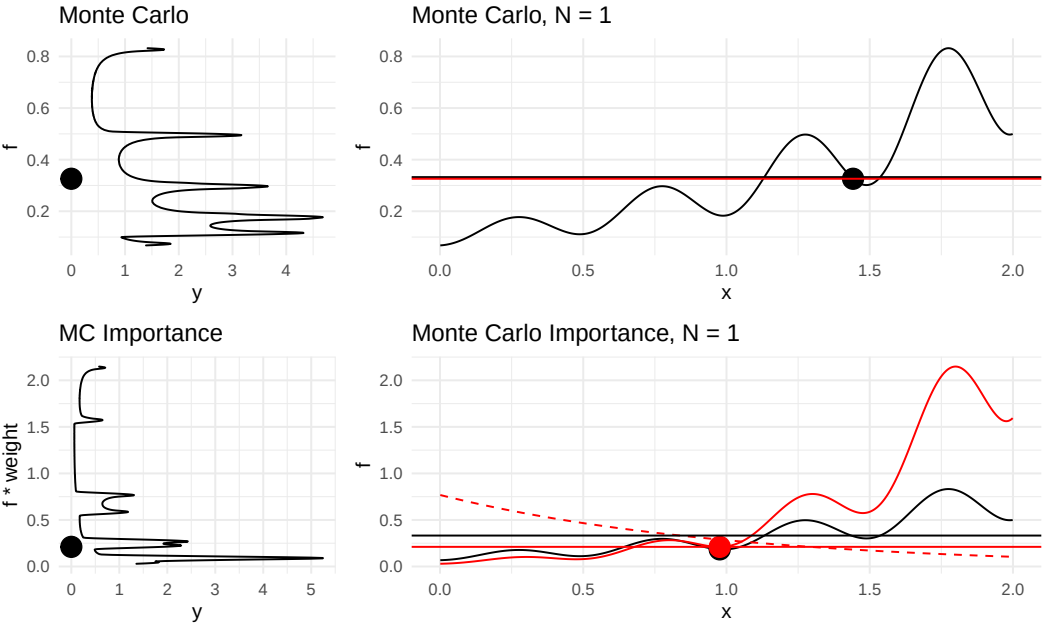

Monte Carlo
f
y
0.8
0.6
0.4
0.2
0
2
4
6
8
Monte Carlo, N = 1000
f
x
0.0
0.5
1.0
1.5
2.0
MC Importance
f * weight
y
0
5
10
15
Monte Carlo Importance, N = 1000

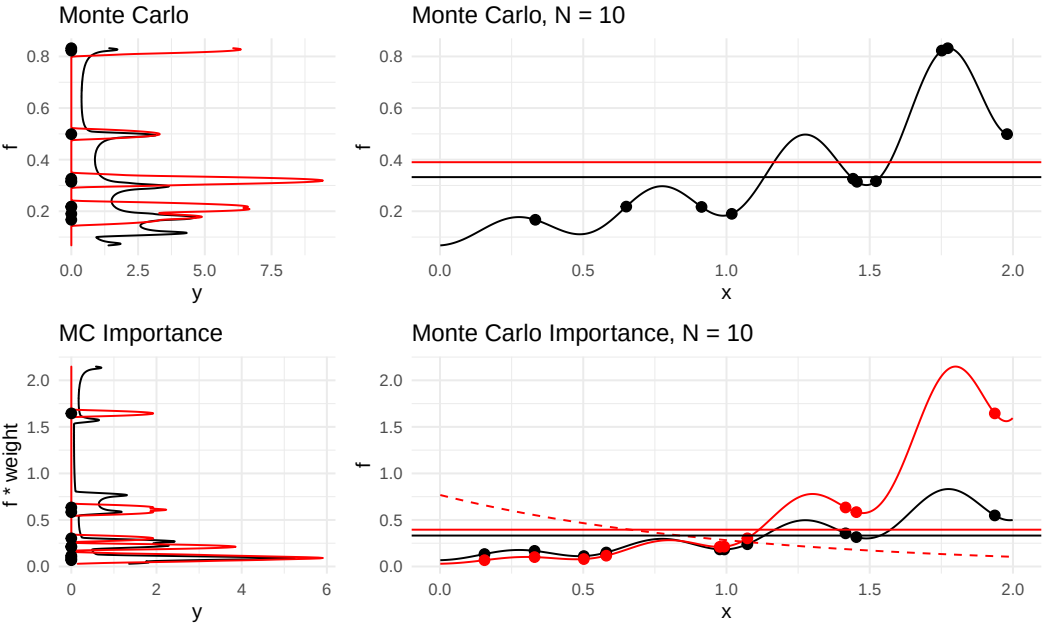

Monte Carlo
0.8
0.6
0.4
0.2
f
0.0
2.5
5.0
7.5
10.0
y

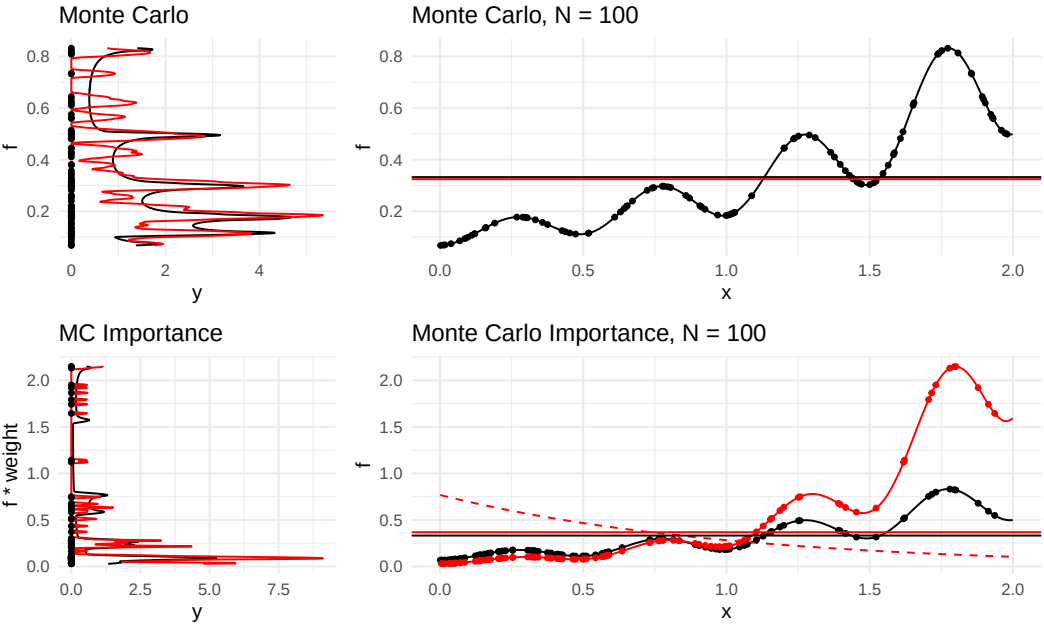

Monte Carlo, N = 10000
0.8
0.6
0.4
0.2
f
0.0
0.5
1.0
1.5
2.0
x

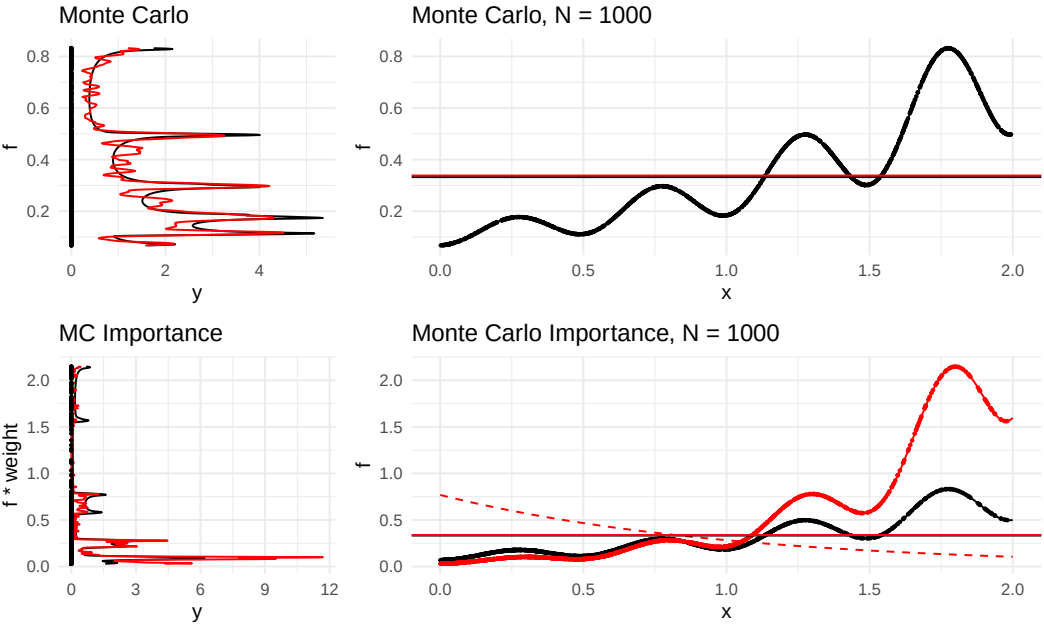

MC Importance
0.8
0.6
0.4
0.2
f * weight
0
5
10
15
20
y

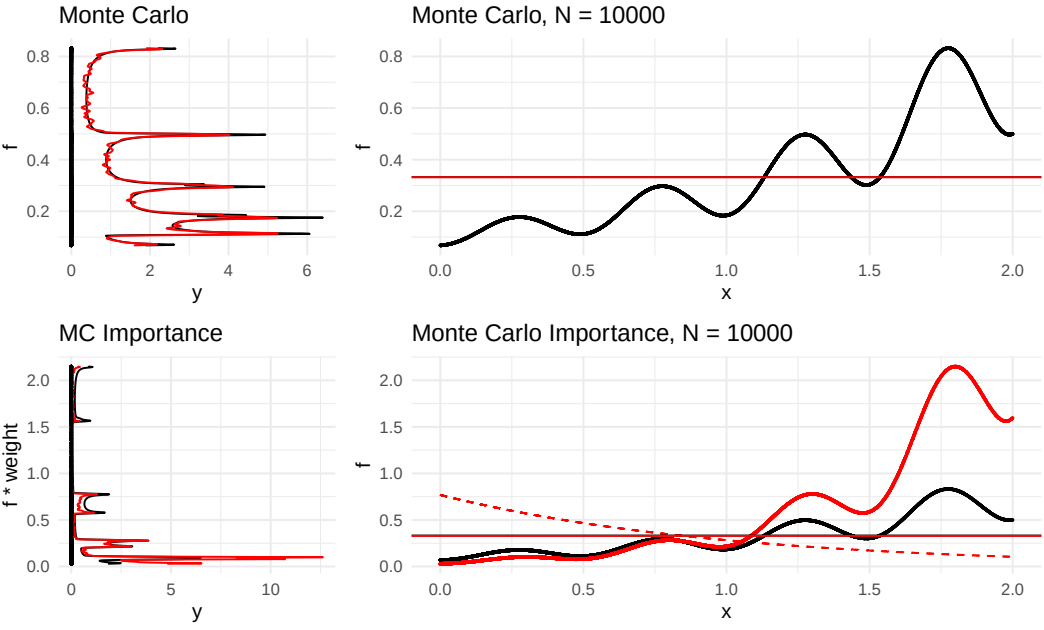

Monte Carlo Importance, N = 10000
0.8
0.6
0.4
0.2
f
0.0
0.5
1.0
1.5
2.0
x


Monte Carlo
f
y
Monte Carlo, N = 1
f
x
MC Importance
f * weight
y
Monte Carlo Importance, N = 1
f
x


Monte Carlo
f
y
Monte Carlo, N = 10
f
x
MC Importance
f * weight
y
Monte Carlo Importance, N = 10
f
x


Monte Carlo
Monte Carlo, N = 100
MC Importance
Monte Carlo Importance, N = 100
f
f * weight
y
x
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
4.0
5.0
7.5


Monte Carlo
f
y
Monte Carlo, N = 1000
f
x
MC Importance
f * weight
y
Monte Carlo Importance, N = 1000
f
x


Monte Carlo
f
y
Monte Carlo, N = 10000
f
x
MC Importance
f * weight
y
Monte Carlo Importance, N = 10000
f
x

Bias/Variance tradeoff

- The bias for the deterministic Midpoint rule is $\mathcal{O}(h^2)$, where h is the distance between each integration point.
- In d dimensions, the spacing in regular lattice for a cube $[0,1]^d$ is $h = N^{-1/d}$, which means that the bias is $\mathcal{O}(N^{-2/d})$.
- The standard deviation for plain Monte Carlo integration is $\mathcal{O}(N^{-1/2})$.
- For $1 \leq d < 4$, the bias for the Midpoint rule decreases faster in N than the MC std.dev.
- For $d > 4$, the MC std.dev. decreases faster.
- Conclusion: *Monte Carlo integration is relatively efficient for high dimensional integrals.*

Bias/Variance tradeoff

Can we get the best of both worlds? Partly, via *stratified* Monte Carlo:

- Split the integration domain into N disjoint subdomains: $\Omega = \cup_{k=1}^{N} \Omega_k$
- Sample each $\boldsymbol{x}_k \sim \mathsf{Unif}(\Omega_k)$ and let $\widehat{I} = \sum_{k=1}^{N} |\Omega_k| f(\boldsymbol{x}_k)$, which is unbiased
- Take a regular lattice with $|\Omega_k| = |\Omega|/N$, $h = \mathcal{O}(N^{-1/d})$. If f has bounded gradients, then $\mathsf{Var}_{\boldsymbol{x} \sim \mathsf{Unif}(\Omega_k)}[f(\boldsymbol{x})] = \mathcal{O}(h^2)$, and

$$\mathsf{Var}_{\{\boldsymbol{x}_k \sim \mathsf{Unif}(\Omega_k)\}}\left[\sum_{k=1}^{N} |\Omega_k| f(\boldsymbol{x}_k)\right] = \cdots = \frac{|\Omega|^2}{N^2} N \mathcal{O}(h^2) = \mathcal{O}(N^{-2/d-1})$$

  so the std.dev. $\mathcal{O}(N^{-1/d-1/2})$ decays faster in N than for plain MC, and also faster than plain Midpoint integration for $d > 2$. The gain over plain MC diminishes with increasing d.
- *Quasi Monte Carlo* integration uses a deterministic space filling sequence, converting variance to bias.