Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

1.1– Outline

- Motivation.
- Introduction to Monte Carlo.

2.1– Summary of Previous Lectures

• Bayesian model: Prior $\pi(\theta)$ and likelihood $f(x|\theta)$

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_\Theta \pi(\theta) f(x|\theta) d\theta}$$

• Except for simple cases -conjugate priors-, there is no closed form expression for the posterior.

• Bayes rule requires being able to compute the potentially high dimensional integral

$$\int_\Theta \pi(\theta) f(x|\theta) d\theta.$$

2.2– Implementations problems for Bayesian inference

- In practice, point estimates are computed

$$
\begin{aligned}
E[\theta|\,x] &= \int \theta\pi(\theta|\,x)\,d\theta \\
Var[\theta|\,x] &= \int \theta^2\pi(\theta|\,x)\,d\theta - E^2[\theta|\,x].
\end{aligned}
$$

and/or marginal distributions; e.g. if $\theta = (\theta_1, \theta_2)$ and θ_2 are so-called nuisance parameters then

$$\pi(\theta_1|\,x) = \int \pi(\theta_1, \theta_2|\,x)\,d\theta_2.$$

- We might also be interested in

$$\theta_1^{\mathrm{MMAP}} = \arg\max\ \pi(\theta_1|\,x)$$

2.2– Implementations problems for Bayesian inference

- If you want to predict $Y \sim g\left(y \middle| \theta \right)$ given x then

$$g\left(y \middle| x \right) = \int g\left(y \middle| \theta \right) \pi\left(\theta \middle| x \right) d\theta$$

and

$$E\left[Y \middle| x \right] = \int \int y g\left(y \middle| \theta \right) \pi\left(\theta \middle| x \right) d\theta.$$

- For model selection with a infinitely countable number of models

$$\pi\left(k, \theta_k \middle| x \right) = \frac{\pi\left(k \right) \pi_k\left(\theta_k \right) f\left(x \middle| k, \theta_k \right)}{\sum_{k=1}^{\infty} \pi\left(k \right) \int \pi_k\left(\theta_k \right) f\left(x \middle| k, \theta_k \right) d\theta_k}$$

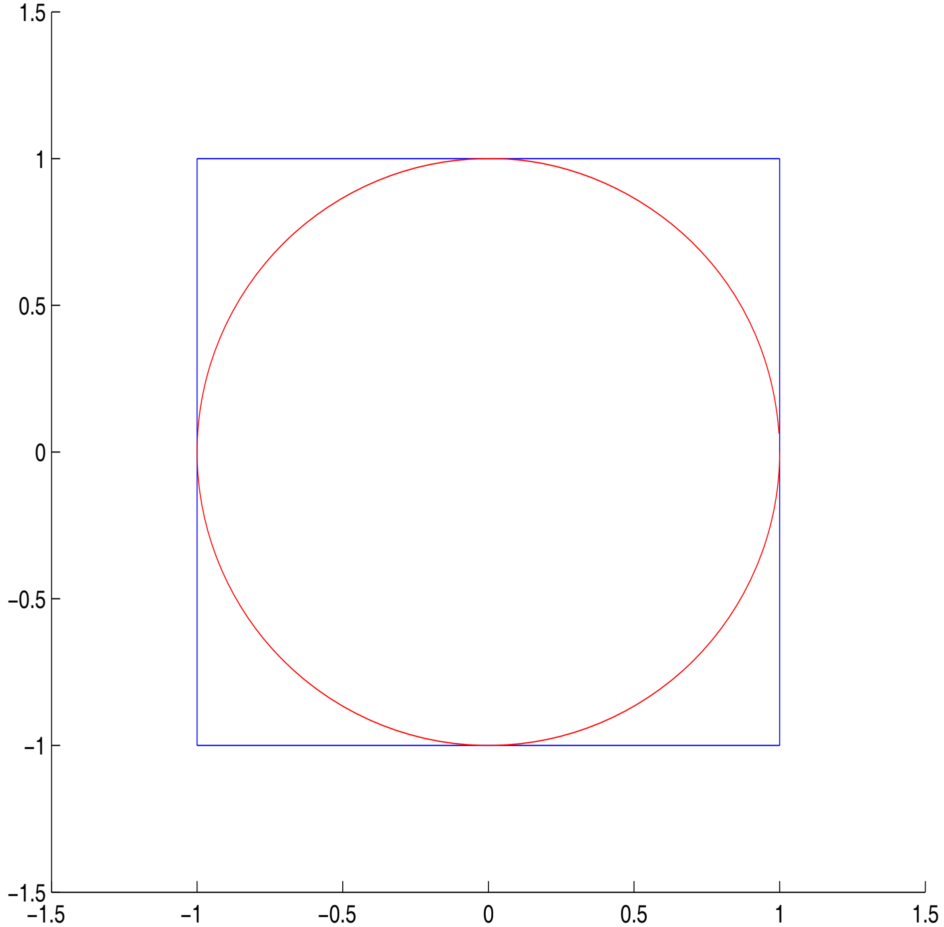
2.2– Implementations problems for Bayesian inference

• Bayesian inference is conceptually simple (once the model is set) but how do you perform Bayesian inference for complex models??? It requires computing high dimensional integrals.

• In practice, Bayesian inference is not only used to determine whether coins are biased and for Gaussian models.

• Monte Carlo methods have appeared in the 90's in statistics and have truly revolutionized the whole field.

Consider the 2×2 square, say $\mathcal{S} \subset \mathbb{R}^2$, with inscribed disc $\mathcal{D}$ of radius 1.



3.1– Introduction to Monte Carlo: A simple example

• An "idealised" rain falls uniformly on the square $\mathcal{S}$, *i.e.* the probability for a drop to fall in a region $\mathcal{A}$ is proportional to the area of $\mathcal{A}$.

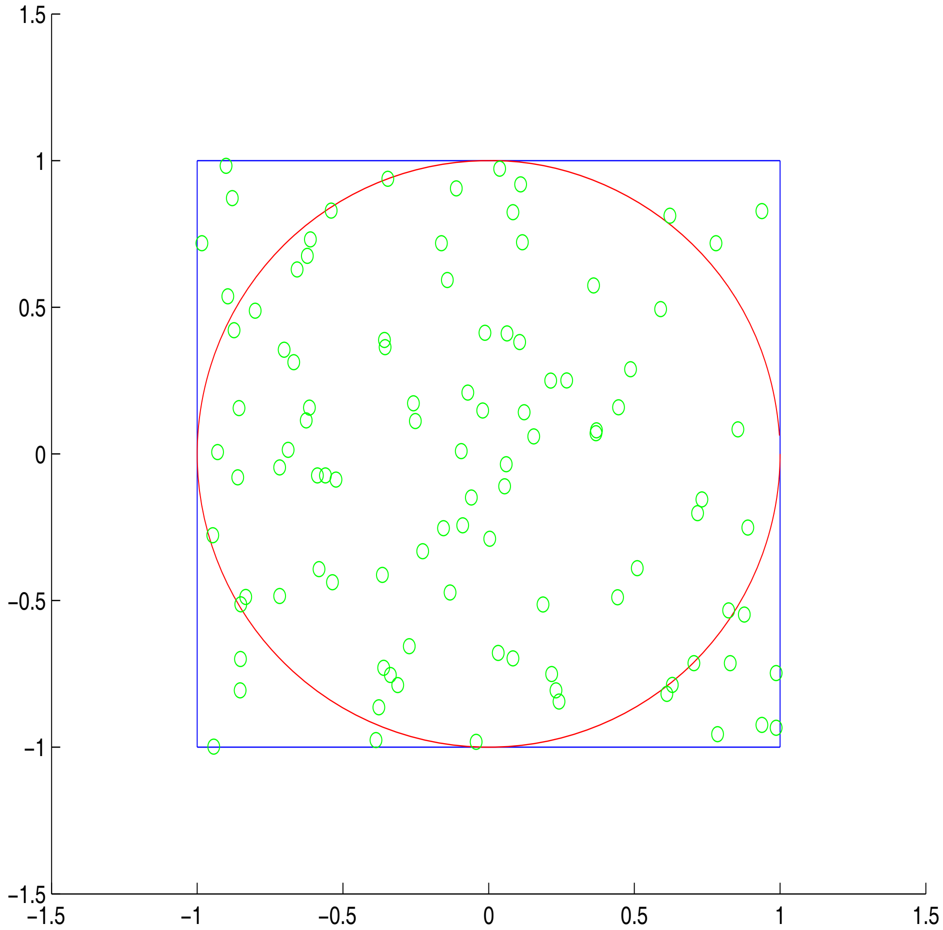
• Let D be the random variable defined on $\Theta = \mathcal{S}$ representing the location of a drop and $\mathcal{A}$ a region of the square, then

$$\mathbb{P}(D \in \mathcal{A}) = \frac{\int_{\mathcal{A}} dxdy}{\int_{\mathcal{S}} dxdy}.$$

where x and y are the Cartesian coordinates.

• Assume we observe N such *independent* drops, say $\{D_i, i = 1, \ldots, N\}$.

3.1– Introduction to Monte Carlo: A simple example



3.1– Introduction to Monte Carlo: A simple example

- Intuitively, imagining that you have never followed any statistics course, a sensible technique to estimate the probability $\mathbb{P}(D \in \mathcal{A})$ of falling in a given region $\mathcal{A} \subset \mathcal{S}$ (and think for example of $\mathcal{A} = \mathcal{D}$) would consist of using

$$\mathbb{P}(d \in \mathcal{A}) \simeq \frac{\text{number of drops that fell in } \mathcal{A}}{N}.$$

- We want a statistical justification to it.

3.2– Probability of this event as an expectation

- Let us denote the indicator function of a set $\mathcal{A}$ as follows,

$$\mathbb{I}_{\mathcal{A}}(x,y) = \begin{cases} 1 & \text{if point } d = (x,y) \in \mathcal{A}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

- We have

$$\mathbb{P}(D \in \mathcal{A}) = \frac{\int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y)dxdy}{\int_{\mathcal{S}} dxdy} = \frac{\int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y)dxdy}{4} = \int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y)\frac{1}{4}dxdy.$$

since

$$\int_{\mathcal{S}=\mathcal{A}\cup\mathcal{S}\setminus\mathcal{A}} \mathbb{I}_{\mathcal{A}}(x,y)dxdy = \int_{\mathcal{A}} \mathbb{I}_{\mathcal{A}}(x,y)dxdy + \int_{\mathcal{S}\setminus\mathcal{A}} \mathbb{I}_{\mathcal{A}}(x,y)dxdy$$

$$= \int_{\mathcal{A}} 1dxdy + \int_{\mathcal{S}\setminus\mathcal{A}} 0dxdy.$$

3.2– Probability of this event as an expectation

• $1/4$ is the probability density associated to $\mathbb{P}$, *i.e.* the density of the uniform distribution on $\mathcal{S}$ denoted $\mathcal{U}_{\mathcal{S}}$.

• Let us define the r.v. $V(D) := \mathbb{I}_{\mathcal{A}}(D) := \mathbb{I}_{\mathcal{A}}(X,Y)$, where X, Y are the rvs representing the Cartesian coordinates of a uniformly distributed point on $\mathcal{S}$, denoted $\mathcal{U}_{\mathcal{S}}$ $(D \sim \mathcal{U}_{\mathcal{S}})$, where a drop falls.
With this notation, we understand that

$$\mathbb{P}(d \in \mathcal{A}) = \int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y)\frac{1}{4}dxdy = \mathbb{E}_{\mathcal{U}_{\mathcal{S}}}(V).$$

3.3– Law of large numbers

• Introduce $\{V_i := V(D_i), i = 1, \ldots, N\}$ the r.v.s associated to the drops $\{D_i, i = 1, \ldots, N\}$ and consider the sum

$$S_N = \frac{\sum_{i=1}^N V_i}{N} = \frac{\text{number of drops that fell in } \mathcal{A}}{N}$$

• This expression shows that our suggested approximation of $\mathbb{P}(D \in \mathcal{A})$ is the empirical average of i.i.d. r.v.s $\{V_i, i = 1, \ldots, N\}$.

• Assuming that the rain lasts forever (i.e. $N \to +\infty$) then the *law of large numbers* (since $\mathbb{E}_{\mathcal{U}_\mathcal{S}}(|V|) < +\infty$ here) yields

$$\lim_{N \to +\infty} S_N = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V), \text{ (almost surely)},$$

where we have already proved that $\mathbb{P}(D \in \mathcal{A}) = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V)$.

• When N is sufficiently large, this mathematically justifies our intuitive method.

- As we have
$$\mathbb{P}(d \in \mathcal{D}) = \int_{\mathcal{D}} \frac{1}{4} dx dy = \frac{\pi}{4}$$
then S_N is an (unbiased) estimator of $\pi/4$.

- It is a r.v., *i.e.* $S_N = \pi/4 + E_N$ where E_N is an error term.

- To characterise the precision of our estimator, we can use
$$var(E_N) = var(S_N) = \frac{1}{N^2} \sum_{i=1}^{N} var(V_i) = \frac{1}{N} var(V_1)$$
as the $\{V_i, i = 1, \dots, N\}$ are independent.
- This means that
$$\sqrt{var(S_N)} = \sqrt{\mathbb{E}\left[(S_N - \mathbb{E}(S_N))^2\right]} = \sqrt{\mathbb{E}\left[(S_N - \mathbb{P}(D \in \mathcal{D}))^2\right]},$$
which implies that the *mean square error* between S_N and $\mathbb{P}(d \in \mathcal{D})$ decreases as $1/\sqrt{N}$.

3.5– Properties of the estimator

• One can invoke an asymptotic result, the *central limit theorem* (which can be applied here as $var(V) < +\infty$). As $N \to +\infty$,

$$\sqrt{N} S_N \to_d \mathcal{N}(\pi/4, var(V))$$

which implies that for N large enough the probability of the error being larger than $2\sqrt{var(V)/N}$ (here $2\sqrt{var(V)} = 0.8211$) is

$$\mathbb{P}\left(|S_N - \pi/4| > 2\sqrt{var(V)/N}\right) \simeq 0.05.$$

• We are sampling here from a Bernoulli distribution so we can establish a non-asymptotic result. Using a Bernstein type inequality, one can prove that for any integer $N \geq 1$ and $\varepsilon > 0$,

$$\mathbb{P}\left(|S_N - \pi/4| > \varepsilon\right) \leq 2\exp\left(-2N\varepsilon^2\right)$$

• For any $\alpha \in (0,1]$, $\mathbb{P}\left(|S_N - \pi/4| > \varepsilon\right) < \alpha$ is guarenteed for

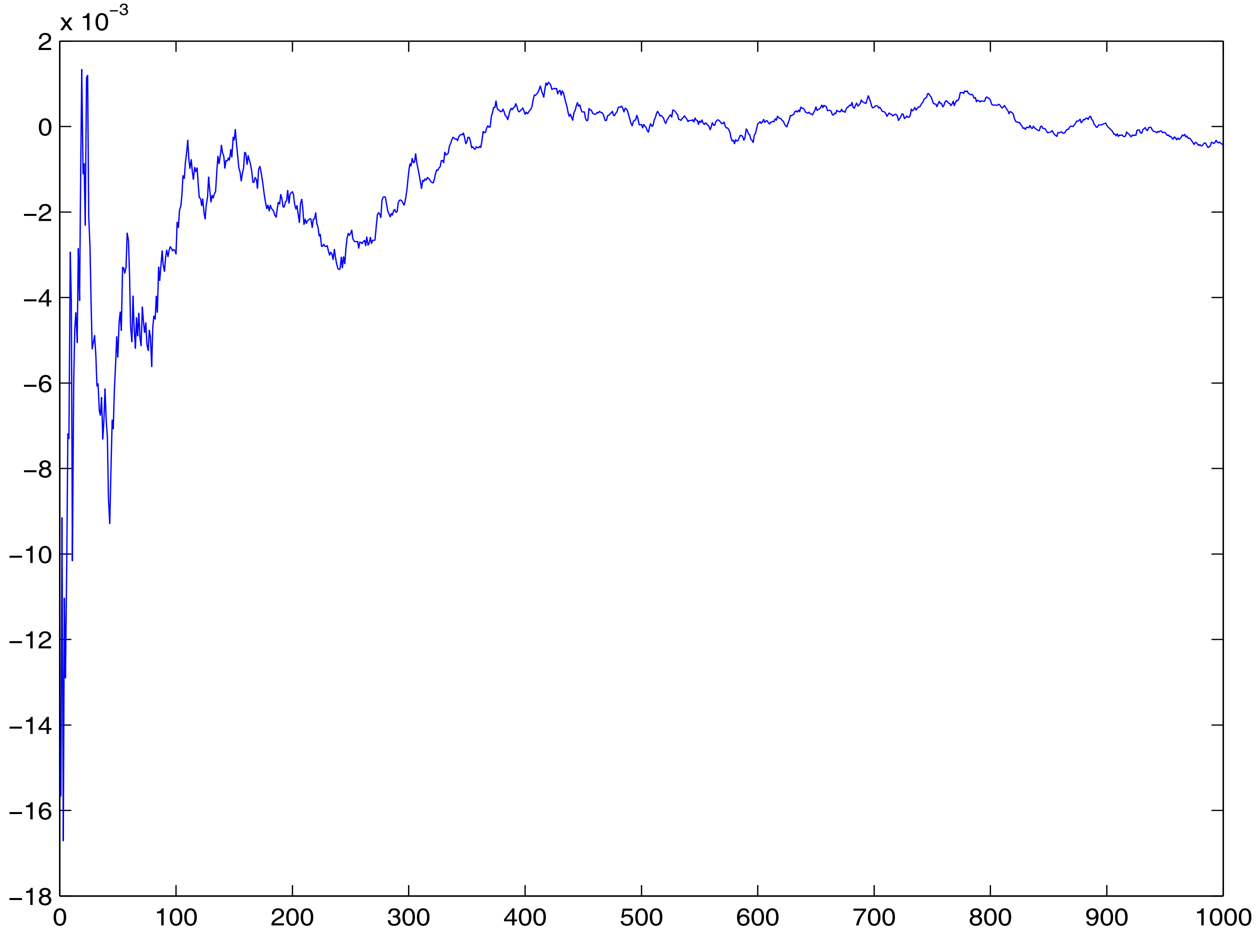
$$N \geq \left\lceil \frac{\log\left(2/\alpha\right)}{2\varepsilon^2} \right\rceil,$$

Alternatively, it tells us that for any $N \geq 1$,

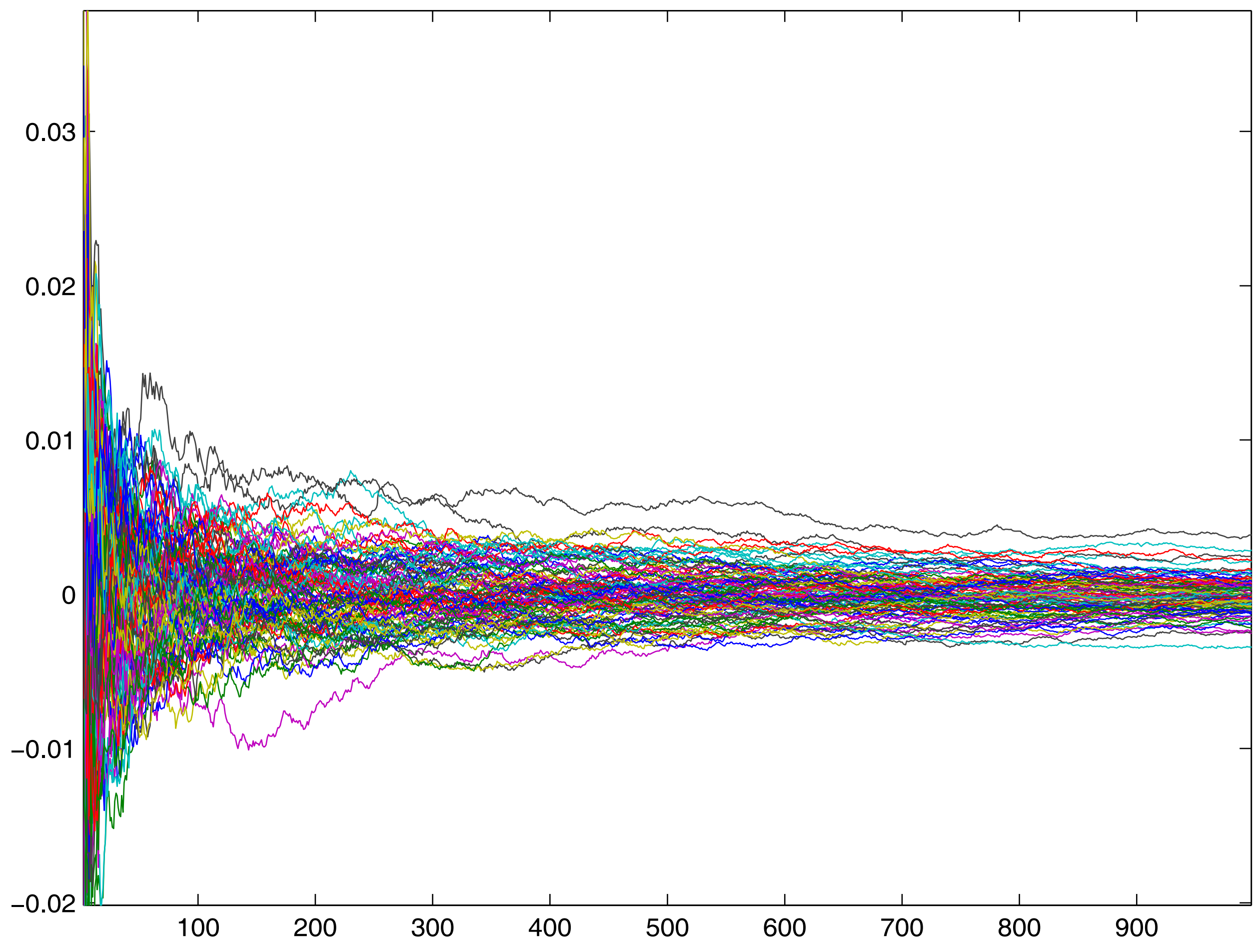
$$\mathbb{P}\left(|S_N - \pi/4| > \sqrt{\frac{\log\left(40\right)}{2N}}\right) \leq 0.05$$

• Both results tell us that in some sense the approximation error is inversely proportional to $\sqrt{N}$.

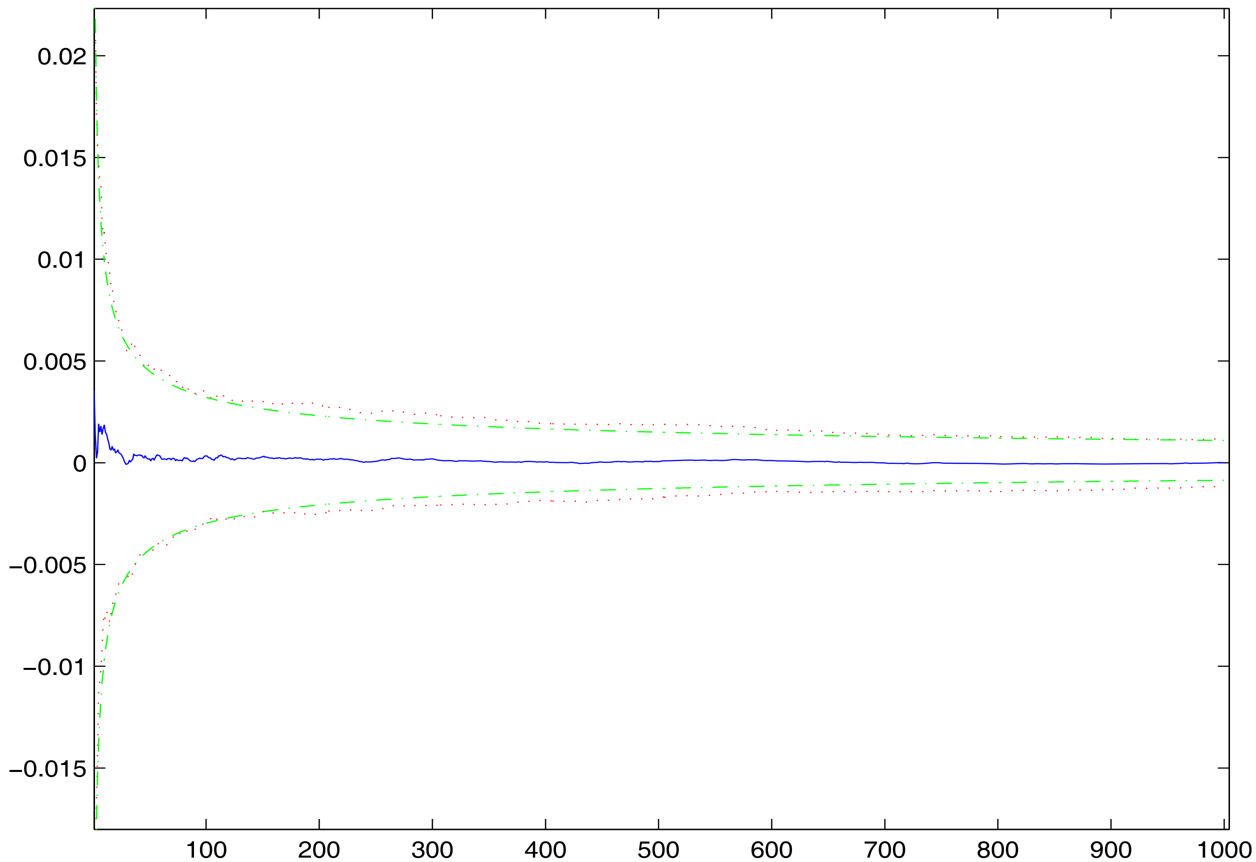
Convergence of $S_N - \frac{\pi}{4}$ as a function of N, 1 realisation.



Convergence of $S_N - \frac{\pi}{4}$ for 100 realisations.



Square root empirical mean square error $S_N - \frac{\pi}{4}$ accross 100 realisations as a function of N (dashed) and $\pm\sqrt{var\,(V)\,/N}$ (dotted).



3.7– Generalization

• Consider the case where $\Theta = \mathbb{R}^{n_\theta}$ for any n_θ, and in particular $n_\theta >> 1$. Replace the $\mathcal{S}$ and $\mathcal{D}$ above with a hypercube $\mathcal{S}^{n_x}$ and an inscribed hyperball $\mathcal{D}^{n_\theta}$ in Θ.

• If we could observe a hyperrain, the same estimator could be built; the onyly thing we need to calculate $\mathbb{I}_{\mathcal{D}^{n_\theta}}(D)$ pointwise.
Arguments that lead earlier to the formal validation of the Monte Carlo approach remain identical here.
• In particular the rate of convergence of the estimator in the mean square sense is again in $1/\sqrt{N}$ and *independent of the dimension* n_x.

• This would not be the case using a deterministic method on a grid of regularly spaced points where the CV rate is typically of the form $1/N^{r/n_\theta}$ where r is related to the smoothness of the contours of $\mathcal{A}$.

$\Rightarrow$ Monte Carlo methods are thus extremely attractive when n_x is large.

3.7– Generalization

- Now we generalise this idea to tackle the generic problem of estimating

$$\mathbb{E}_\pi(f(\theta)) \triangleq \int_\Theta f(\theta)\pi(\theta)d\theta,$$

where $f : \Theta \to \mathbb{R}^{n_f}$ and π is a probability distribution on $\Theta \subset \mathbb{R}^{n_x}$.

- We will assume that $\mathbb{E}_\pi(|f(\theta)|) < +\infty$ but that it is difficult to obtain an analytical expression for $\mathbb{E}_\pi(f(\theta))$.

- Here π is any probability distribution and not necessary the prior.

3.7– Generalization

• Assume $N >> 1$ *i.i.d.* samples $\theta^{(i)} \sim \pi$ $(i = 1, \ldots, N)$ are available to us (since it is unlikely that rain can generate samples from any distribution π, we will address the problem of sample generation later).

• Now consider any set $\mathcal{A} \subset \Theta$ and assume that we are interested in $\pi(\mathcal{A}) = \mathbb{P}(\theta \in \mathcal{A})$ for $\theta \sim \pi$. We naturally choose the following estimator

$$\pi(\mathcal{A}) \simeq \frac{\text{number of samples in } \mathcal{A}}{\text{total number of samples}},$$

which by the law of large numbers is a consistent estimator of $\pi(\mathcal{A})$ since

$$\lim_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\mathcal{A}}(\theta^{(i)}) = \mathbb{E}_{\pi}(\mathbb{I}_{\mathcal{A}}(\theta)) = \pi(\mathcal{A}).$$

3.7– Generalization

- A way of generalising this in order to evaluate $\mathbb{E}_\pi(f(\theta))$ consists of considering the unbiased estimator

$$S_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)}),$$

- From the law of large numbers $S_N(f)$ will converge and

$$\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)}) = \mathbb{E}_\pi(f(\theta)) \; a.s.$$

- A good measure of the approximation is the variance of $S_N(f)$,

$$var_\pi\left[S_N(f)\right] = var_\pi\left[\frac{1}{N}\sum_{i=1}^{N} f(\theta^{(i)})\right] = \frac{var_\pi\left[f(\theta)\right]}{N}.$$

Now the central limit theorem applies if $var_\pi\left[f(\theta)\right] < \infty$ and tells us that

$$\sqrt{N}\left(S_N(f) - \mathbb{E}_\pi(f(\theta))\right) \overset{N\to+\infty}{\longrightarrow}_d \mathcal{N}\left(0, var_\pi\left[f(\theta)\right]\right),$$

The conclusions drawn in the rain example are still valid here

- The rate of convergence is immune to the dimension of Θ.

- It is easy to take complex integration domains into account.

- It is easily implementable and general. The requirements are

  to be able to evaluate $f(\theta)$ for any $\theta \in \Theta$,

  to be able to produce samples distributed according to π.

3.8– From the algebraic to the sample representation

- Let us introduce the delta-Dirac function δ_{θ_0} for $\theta_0 \in \Theta$ defined for any $f : \Theta \to \mathbb{R}^{n_f}$.as follows

$$\int_\Theta f(\theta)\delta_{\theta_0}(\theta)d\theta = f(\theta_0).$$

- Note that this implies in particular that for $\mathcal{A} \subset \Theta$,

$$\int_\Theta \mathbb{I}_\mathcal{A}(\theta)\delta_{\theta_0}(\theta)d\theta = \int_\mathcal{A} \delta_{\theta_0}(\theta)d\theta = \mathbb{I}_\mathcal{A}(\theta_0).$$

- Now, for $\theta^{(i)} \sim \pi$ for $i = 1, \ldots, N$, we can introduce the following mixture of delta-Dirac functions

$$\widehat{\pi}_N(\theta) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^{(i)}}(\theta),$$

which is the *empirical measure*, and consider for any $\mathcal{A} \subset \Theta$

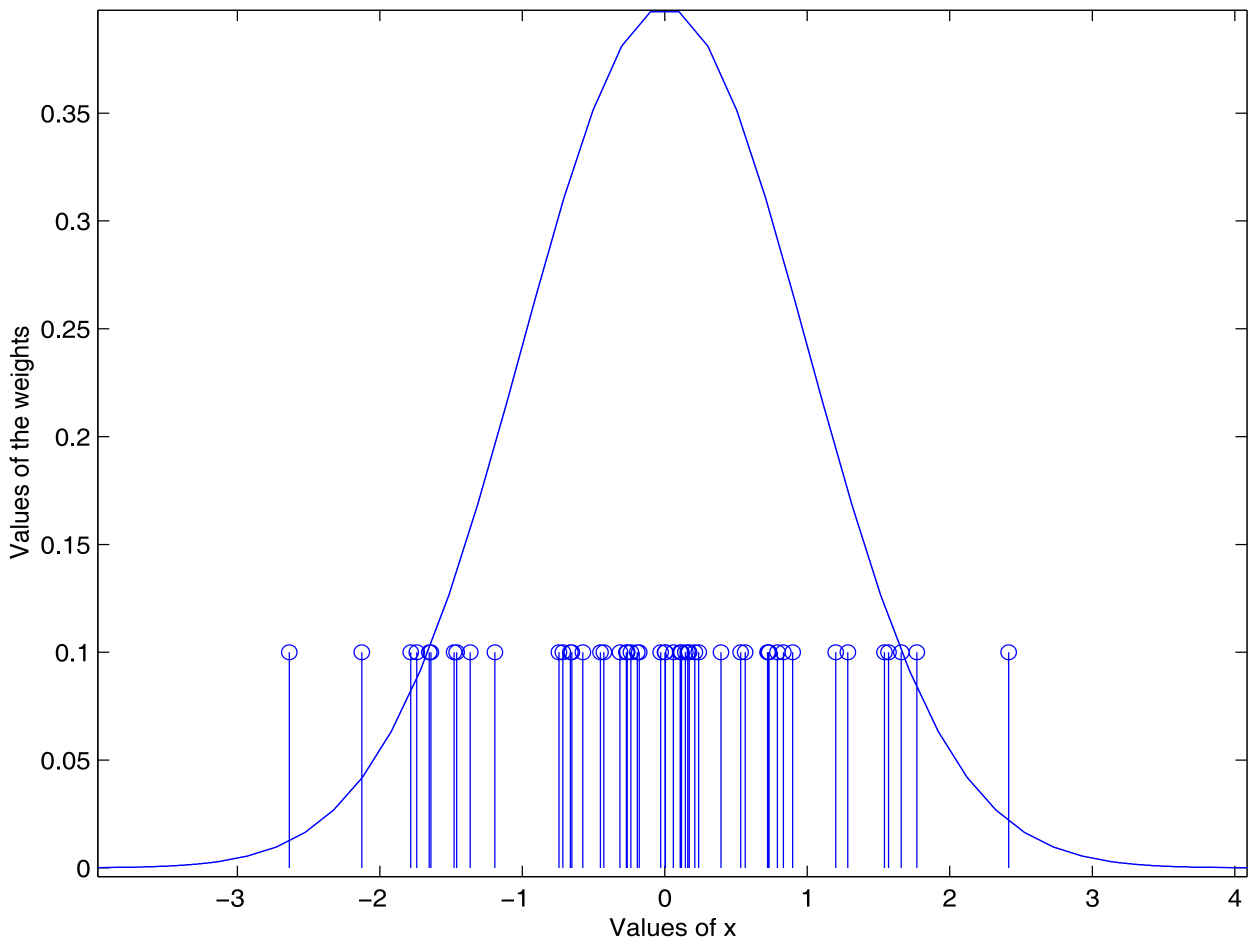
$$\widehat{\pi}_N(\mathcal{A}) \triangleq \int_\mathcal{A} \widehat{\pi}_N(\theta)\, d\theta = \sum_{i=1}^{N} \int_\mathcal{A} \frac{1}{N}\delta_{\theta^{(i)}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_\mathcal{A}(\theta) = S_N(\mathcal{A})$$

3.8– From the algebraic to the sample representation

• **The concentration of points in a given region of the space represents π.**

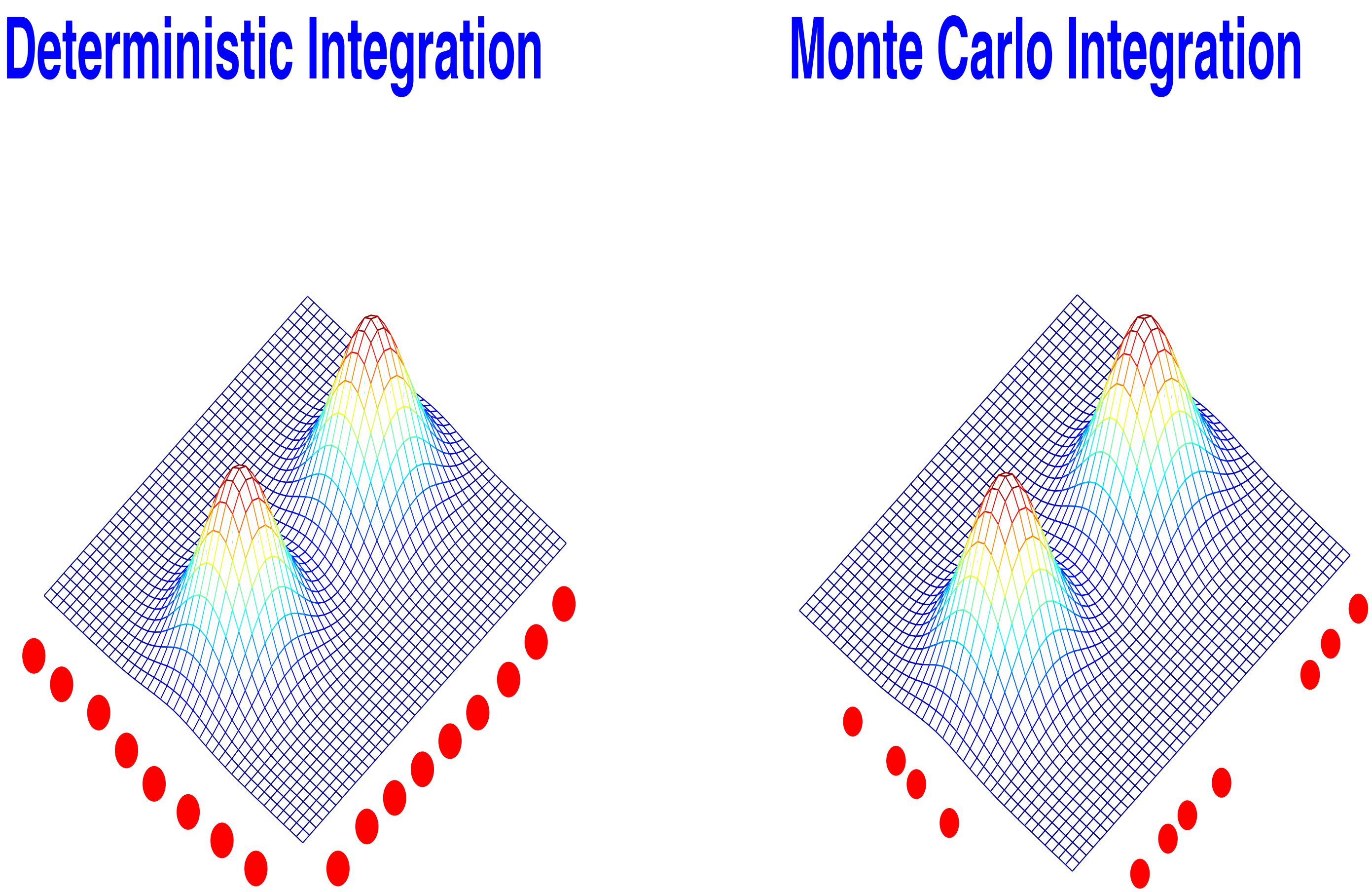
• This approach is in contrast with what is usually done in parametric statistics, *i.e.* start with samples and then introduce a distribution with an algebraic representation for the underlying population.

• Note that here each sample $\theta^{(i)}$ has a weight of $1/N$, but that it is also possible to consider weighted sample representations of π.



Sample representation of a Gaussian distribution

3.8– From the algebraic to the sample representation



3.8– From the algebraic to the sample representation

- Now consider the problem of estimating $\mathbb{E}_\pi(f)$. We simply replace π with its sample representation $\widehat{\pi}_N$ and obtain

$$\mathbb{E}_\pi(f) \simeq \int_\Theta f(\theta) \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta^{(i)}}(\theta)\, d\theta = \sum_{i=1}^{N} \frac{1}{N} \int_\Theta f(\theta)\, \delta_{\theta^{(i)}}(\theta)\, d\theta = \frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)}),$$

which is precisely $S_N(f)$, the Monte Carlo estimator suggested earlier.

- Clearly based on $\widehat{\pi}_N$, we can easily estimate $\mathbb{E}_\pi(f)$ for any f.

- For example

$$var_\pi(f) = \mathbb{E}_\pi(f^2) - \mathbb{E}_\pi^2(f) \simeq \frac{1}{N} \sum_{i=1}^{N} f^2(\theta^{(i)}) - \left(\frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)}) \right)^2 .$$

3.8– From the algebraic to the sample representation

- Similarly, if we have

$$\widehat{\pi}_N(\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_1^{(i)}, \theta_2^{(i)}}(\theta_1, \theta_2)$$

so the marginal distribution is simply given by

$$\widehat{\pi}_N(\theta_1) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_1^{(i)}}(\theta_1)$$

- If we want to estimate $\arg\max \pi(\theta)$ and $\pi(\theta)$ is known up to a normalizing constant then

$$\arg\max_{\{\theta^{(i)}\}} \pi\left(\theta^{(i)}\right)$$

is a reasonable estimate.

- If you could sample easily from an arbitrary probability distribution,

then you could easily estimate all the quantities you are interested in.

- **Problem**: How do you sample from an arbitrary probability distribution???