Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca
1.1– Outline

• Motivation.

• Introduction to Monte Carlo.
\[ \pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\Theta} \pi(\theta)f(x|\theta) \, d\theta} \]

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

- 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

\[
E[\theta|x] = \int \theta \pi(\theta|x) d\theta
\]

\[
Var[\theta|x] = \int \theta^2 \pi(\theta|x) d\theta - E^2[\theta|x].
\]

and/or marginal distributions; e.g. if \( \theta = (\theta_1, \theta_2) \) and \( \theta_2 \) 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

\[
\theta_1^{\text{MMAP}} = \arg \max \pi(\theta_1|x)
\]
2.2– Implementations problems for Bayesian inference

• If you want to predict \( Y \sim g(y|\theta) \) given \( x \) then

\[
g(y|x) = \int g(y|\theta) \pi(\theta|x) d\theta
\]

and

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

• For model selection with a infinitely countable number of models

\[
\pi(k, \theta_k|x) = \frac{\pi(k) \pi_k(\theta_k) f(x|k, \theta_k)}{\sum_{k=1}^{\infty} \pi(k) \int \pi_k(\theta_k) f(x|k, \theta_k) d\theta_k}
\]
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 $2 \times 2$ square, say $S \subset \mathbb{R}^2$, 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 $S$, i.e. the probability for a drop to fall in a region $A$ is proportional to the area of $A$.

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

$$P(D \in A) = \frac{\int_A dx \, dy}{\int_S dx \, dy}.$$ 

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

A diagram showing the distribution of points within a circle and a square.
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 $\mathbb{P}(D \in A)$ of falling in a given region $A \subset S$ (and think for example of $A = D$) would consist of using

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

- 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

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

since

\[
\int_{S = \mathcal{A} \cup S\setminus\mathcal{A}} \mathbb{I}_\mathcal{A}(x, y) \, dx \, dy = \int_{\mathcal{A}} \mathbb{I}_\mathcal{A}(x, y) \, dx \, dy + \int_{S\setminus\mathcal{A}} \mathbb{I}_\mathcal{A}(x, y) \, dx \, dy
\]

\[
= \int_{\mathcal{A}} 1 \, dx \, dy + \int_{S\setminus\mathcal{A}} 0 \, dx \, dy.
\]
3.2– Probability of this event as an expectation

• 1/4 is the probability density associated to \( P \), i.e. the density of the uniform distribution on \( S \) denoted \( \mathcal{U}_S \).

• Let us define the r.v. \( V(D) := \mathbb{I}_A(D) := \mathbb{I}_A(X,Y) \), where \( X, Y \) are the rvs representing the Cartesian coordinates of a uniformly distributed point on \( S \), denoted \( \mathcal{U}_S (D \sim \mathcal{U}_S) \), where a drop falls.

With this notation, we understand that

\[
P(d \in A) = \int_S \mathbb{I}_A(x, y) \frac{1}{4} \, dx \, dy = \mathbb{E}_{\mathcal{U}_S}(V).
\]
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

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

• 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 \( \mathbb{E}_\mathcal{U}_S(|V|) < +\infty \) here) yields

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

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

• When \( N \) is sufficiently large, this mathematically justifies our intuitive method.
3.4– Approximating pi

- As we have
  \[ \mathbb{P}(d \in D) = \int_D \frac{1}{4} dx dy = \frac{\pi}{4} \]
  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
  \[
  \text{var}(E_N) = \text{var}(S_N) = \frac{1}{N^2} \sum_{i=1}^{N} \text{var}(V_i) = \frac{1}{N} \text{var}(V_1)
  \]
as the \( \{V_i, i = 1, \ldots, N\} \) are independent.

- This means that
  \[
  \sqrt{\text{var}(S_N)} = \sqrt{\mathbb{E}[(S_N - \mathbb{E}(S_N))^2]} = \sqrt{\mathbb{E}[(S_N - \mathbb{P}(D \in D))^2]},
  \]
  which implies that the mean square error between \( S_N \) and \( \mathbb{P}(d \in D) \) 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 \( \text{var}(V) < +\infty \)). As \( N \to +\infty \),

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

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

\[
\mathbb{P}\left(|S_N - \pi/4| > 2\sqrt{\text{var}(V)/N}\right) \approx 0.05.
\]

- 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 \( \varepsilon > 0 \),

\[
\mathbb{P}\left(|S_N - \pi/4| > \varepsilon\right) \leq 2 \exp\left(-2N\varepsilon^2\right)
\]
3.5– Properties of the estimator

- For any $\alpha \in (0, 1]$, $\mathbb{P}(\lvert S_N - \pi/4 \rvert > \varepsilon) < \alpha$ is guaranteed for

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

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

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

- 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.
3.6– Simulations

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

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

• Consider the case where $\Theta = \mathbb{R}^{n_\theta}$ for any $n_\theta$, 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 $\Theta$.

• If we could observe a hyperrain, the same estimator could be built; the only thing we need to calculate is $I_{\mathcal{D}^{n_\theta}}(D)$ 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

\[ E_\pi(f(\theta)) \triangleq \int_{\Theta} f(\theta) \pi(\theta) d\theta, \]

where \( f : \Theta \rightarrow \mathbb{R}^{n_f} \) and \( \pi \) is a probability distribution on \( \Theta \subset \mathbb{R}^{n_x} \).

• We will assume that \( E_\pi(|f(\theta)|) < +\infty \) but that it is difficult to obtain an analytical expression for \( E_\pi(f(\theta)) \).

• Here \( \pi \) is any probability distribution and not necessary the prior.
3.7– Generalization

• Assume $N \gg 1$ 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 $\pi$, we will address the problem of sample generation later).

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

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

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

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

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

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

$$\sqrt{N} (S_N(f) - \mathbb{E}_\pi(f(\theta))) \xrightarrow{N \to +\infty} d \mathcal{N}(0, var_\pi[f(\theta)]),$$
3.7– Generalization

The conclusions drawn in the rain example are still valid here

- The rate of convergence is immune to the dimension of $\Theta$.
- 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 $\pi$. 
3.8– From the algebraic to the sample representation

• Let us introduce the delta-Dirac function $\delta_{\theta_0}$ 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 $A \subset \Theta$,
\[
\int_{\Theta} \mathbb{1}_A(\theta) \delta_{\theta_0}(\theta) d\theta = \int_A \delta_{\theta_0}(\theta) d\theta = \mathbb{1}_A(\theta_0).
\]

• Now, for $\theta^{(i)} \sim \pi$ for $i = 1, \ldots, N$, we can introduce the following mixture of delta-Dirac functions
\[
\hat{\pi}_N(\theta) := \frac{1}{N} \sum_{i=1}^N \delta_{\theta^{(i)}}(\theta),
\]
which is the empirical measure, and consider for any $A \subset \Theta$
\[
\hat{\pi}_N(A) \triangleq \int_A \hat{\pi}_N(\theta) d\theta = \sum_{i=1}^N \int_A \frac{1}{N} \delta_{\theta^{(i)}}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_A(\theta) = S_N(A)
\]
3.8– From the algebraic to the sample representation

- The concentration of points in a given region of the space represents $\pi$.

- 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 $\pi$. 
3.8 – From the algebraic to the sample representation

Sample representation of a Gaussian distribution
3.8– From the algebraic to the sample representation

Deterministic Integration

Monte Carlo Integration
3.8– From the algebraic to the sample representation

- Now consider the problem of estimating $E_\pi(f)$. We simply replace $\pi$ with its sample representation $\hat{\pi}_N$ and obtain

\[
E_\pi(f) \simeq \int_\Theta f(\theta) \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta(i)}(\theta) \, d\theta = \sum_{i=1}^{N} \frac{1}{N} \int_\Theta f(\theta) \delta_{\theta(i)}(\theta) \, d\theta = \frac{1}{N} \sum_{i=1}^{N} f(\theta(i)),
\]

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

- Clearly based on $\hat{\pi}_N$, we can easily estimate $E_\pi(f)$ for any $f$.

- For example

\[
var_\pi(f) = E_\pi(f^2) - E^2_\pi(f) \simeq \frac{1}{N} \sum_{i=1}^{N} f^2(\theta(i)) - \left( \frac{1}{N} \sum_{i=1}^{N} f(\theta(i)) \right)^2.
\]
3.8– From the algebraic to the sample representation

- Similarly, if we have

\[ \hat{\pi}_N (\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_1^{(i)}, \theta_2^{(i)}} (\theta_1, \theta_2) \]

so the marginal distribution is simply given by

\[ \hat{\pi}_N (\theta_1) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_1^{(i)}} (\theta_1) \]

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

\[ \arg \max \pi \left( \theta^{(i)} \right) \]

is a reasonable estimate.
3.9– Summary

• 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???