08 - Approximate Bayesian inference and Monte-Carlo - 02

Master in Foundations of Data Science
Bayesian Statistics and Probabilistic Programming
Fall 2018-2019

Josep Fortiana

Wednesday, November 14, 2018

Universitat de Barcelona

08 - Monte-Carlo - 02

Monte-Carlo computation of expectations

Rejection sampling

Importance sampling

SIR algorithm (Sampling Importance Resampling)

08 - Monte-Carlo - 02

One-dimensional function

For a function $g(X)$ of a r.v. X, with pdf $f(x)$,

$$\gamma \equiv \mathrm{E}[g(X)] = \int g(x) \cdot f(x)\, dx.$$

Generate $x_1, \ldots, x_n \sim f$. The average:

$$\overline{g}_n = \frac{1}{n} \sum_{j=1}^{n} g(x_i),$$

approximates γ.

One-dimensional function

Under usual conditions the sequence $\{\overline{g}_n\}$ is convergent to γ.

The simulation standard error of the estimate is estimated by:

$$\mathsf{se}_{\overline{g}_n} = \sqrt{\frac{\sum_{i=1}^{n}(g(x_i) - \overline{g}_n)}{(n-1)\,n}}.$$

For a d-dimensional function

For a function $g(\boldsymbol{X})$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$, where:

$$g : \mathbb{R}^d \longrightarrow \mathbb{R}, \quad \text{and } \operatorname{E}[|g(\boldsymbol{X})|] < +\infty,$$

generate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \boldsymbol{X}$, evaluate $g_i = g\left(\boldsymbol{X}_i\right)$, $1 \leq i \leq n$, and set:

$$\overline{g}_n = \frac{g_1 + \cdots + g_n}{n}.$$

Properties

$\overline{g}_n$ is a good estimator of $\gamma \equiv \mathrm{E}[g(\boldsymbol{X})]$, as it is:

- Unbiased:

$$\mathrm{E}[\overline{g}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[g\left(\boldsymbol{X}_i\right)] = \frac{n\gamma}{n} = \gamma.$$

- Consistent: SLLN $\implies$ $\overline{g}_n \xrightarrow[n\to\infty]{a.s.} \gamma$.

Example: The quadrature of a function

The Monte Carlo integral (quadrature) of a function $g : \mathbb{R}^2 \to \mathbb{R}$:

$$\theta := \int_0^1 \int_0^1 g(x_1, x_2)\, dx_1\, dx_2.$$

on a rectangle $[0, 1] \times [0, 1]$ is a particular case of this procedure.

Indeed, $\theta = \mathrm{E}[g(\boldsymbol{X})]$, where $\boldsymbol{X} := (U_1, U_2)$, U_1 and U_2 are i.i.d. r.v. $\sim \mathrm{Unif}(0, 1)$.

Pseudocode for the quadrature of a function

From $i = 1$ to n,

1. Generate independent $U_1 \sim \mathsf{Unif}(0,1)$ and $U_2 \sim \mathsf{Unif}(0,1)$.
2. Set $g_i = g(U_1, U_2)$.

Finally:

$$\widehat{\theta}_n = \frac{g_1 + \cdots + g_n}{n}.$$

Error estimation

Monte Carlo integration of functions has an error of order $n^{-1/2}$, where n is the number of samples, independently from dimension.

Usually determinist numerical integration methods have error rates of order $n^{-2/d}$, where d is the integral dimension and n is the number of points in the partition.

Thus, Monte Carlo is a competitive method for large-dimensional problems.

Curse of dimensionality

A sample of uniform points on $[0,1]^d$, with a large d, does not fill efficiently the whole d-cube.

08 - Monte-Carlo - 02

A candidate function, given the target pdf g

We seek (or design) another pdf p such that:

- g, p have equal support ($g(x) > 0 \iff p(x) > 0$).
- There is a constant M such that $g(x) \leq M \cdot p(x)$ for all x.
- We know how to generate RN distributed as p

Accept-reject procedure

Generate a pair of random numbers (u, v),

u from the distribution with pdf p,

v from a $\mathsf{Unif}(0, 1)$,

If $v \leq \dfrac{g(u)}{M \cdot p(u)}$, then keep u in the list,

Otherwise throw it away.

Continue until the list is sufficiently long.

Rejection sampling for the stomach cancer mortality dataset

Simulating the posterior pdf $g(\theta_1, \theta_2|\boldsymbol{y})$ from a Beta-Binomial likelihood and a non-informative prior as in last session.

Choice of candidate function $p(\,\cdot\,)$:

- The bivariate normal resulting from the Laplace approximation (fails: does not majorize g)
- A bivariate Student's t with the same location and scale (succeeds: it has heavier tails)

Student's t pdf

The univariate Student's t pdf with ν *degrees of freedom* is:

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\,\nu}\cdot\Gamma(\nu/2)} \times \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

for $t \in \mathbb{R}$.

Student's t with location and scale

A translation and scale transformation gives the univariate Student's t with ν *degrees of freedom*, *location* μ and *scale* σ:

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\,\nu}\cdot\Gamma(\nu/2)\cdot\sigma} \times \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2},$$

for $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$.

Multivariate Student's t pdf

Given $\boldsymbol{\mu} \in \mathbb{R}^p$ *(location)*, $\boldsymbol{\Sigma}$ a $p \times p$ positive definite (nonsingular) matrix *(scale)*, the p-dimensional Student's t pdf:

$$f(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)\cdot(\pi\,\nu)^{p/2}\cdot(\det|\boldsymbol{\Sigma}|)^{1/2}}$$

$$\times \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})' \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})\right)^{-(\nu+p)/2},$$

$\boldsymbol{x} \in \mathbb{R}^p$.

Computations

See the notebook.

08 - Monte-Carlo - 02

Computing posterior expectations

Expectation $\mathrm{E}(h(\theta)|y)$ of a function $h(\theta)$ of the parameters θ from the posterior pdf of a model with prior $g(\theta)$ and likelihood $f(y|\theta)$, given the data y.

Often the posterior pdf will be non-normalized:

$$g(\theta|y) \propto g(\theta) \cdot f(y|\theta)$$

We want to estimate this expectation by the method described in the previous section.

An auxiliary pdf

We substitute a pdf $p(\theta)$ for the posterior:

$$\begin{aligned}
\mathrm{E}(h(\theta)|y) &= \frac{\int h(\theta) \cdot g(\theta) \cdot f(y|\theta)\, d\theta}{\int g(\theta) \cdot f(y|\theta)\, d\theta} \\
&= \frac{\int h(\theta) \cdot \left(\frac{g(\theta) \cdot f(y|\theta)}{p(\theta)}\right) \cdot p(\theta)\, d\theta}{\int \left(\frac{g(\theta) \cdot f(y|\theta)}{p(\theta)}\right) \cdot p(\theta)\, d\theta}.
\end{aligned}$$

Thus we avoid sampling from the presumably difficult posterior, doing it instead from $p(\theta)$.

Weights

Define the *weight function:*

$$w(\theta) \equiv \frac{g(\theta) \cdot f(y|\theta)}{p(\theta)}.$$

For an n-sample $(\theta_1, \ldots, \theta_n)$ drawn from $p(\theta)$, the *importance sampling estimate* of the posterior expectation is the average:

$$\overline{h}_{IS} = \frac{\sum_{i=1}^{n} h(\theta_i) \cdot w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)}$$

Simulation standard error

Estimated by:

$$\mathrm{se}_{\overline{h}_{IS}} = \frac{\sqrt{\sum_{i=1}^{n}\left(\left(h(\theta_i) - \overline{h}_{IS}\right) \cdot w(\theta_i)\right)^2}}{\sum_{i=1}^{n} w(\theta_i)}.$$

Choice of a suitable candidate $p(\theta)$

Hopefully it should be:

- Easy to sample from
- As close as possible to the target pdf
- Heavier tails than the target pdf (otherwise weights become very large)

Computations

See the notebook.

08 - Monte-Carlo - 02

The SIR procedure

- Start as in importance sampling: choose a proposal pdf, generate an n-sample $(\theta_1, \ldots, \theta_n)$ from it, and the corresponding weights: $w_i = w(\theta_i)$, $1 \leq i \leq n$.

- Convert weights to probabilities:

$$p_i = \frac{w_i}{\sum_{j=1}^{n} w_j}, \qquad 1 \leq i \leq n.$$

- Generate a new n-sample *(resample)* from the n values $(\theta_1, \ldots, \theta_n)$ with probabilities $(p_1, \ldots, p_n)$.

Comparison (mostly intuitive) to rejection sampling

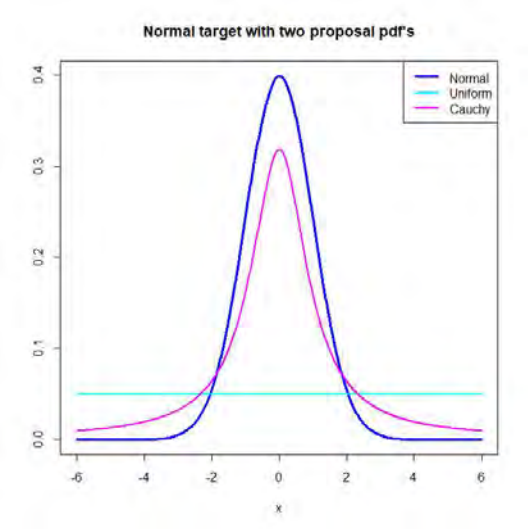
In rejection sampling, a given sample θ_i drawn from the proposal pdf $p(\theta)$ is accepted with probability equal to:

$$g(\theta_i)\,/\,c\,p(\theta_i)\,.$$

In SIR θ_i appears in the resample with probability proportional to the weight:

$$w_i = g(\theta_i)\,/\,p(\theta_i)\,.$$

Example: SIR vs. Rejection sampling



Example: SIR vs. Rejection sampling

See notebook.