Importance sampling: Summary

As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $g(x)$ matches the target distribution $f(x)$.

Lecture 6: Feedback so far ...

- Participation 17 out of 41 ($\approx$ 41%) registered students.
- Difficult:
  R (3), Monte Carlo integration (3), importance sampling (1), ratio-of-uniforms (1), Box-Muller(1), mixtures (1), ...
- Easy:
  Method of inversion (2), R (1), rejection sampling (1)
- General comments:
  - Sometimes too fast
  - More examples (on blackboard)
  - More interactive
  - "Almost everything easy"
  - Slides are good

Bayesian concept

... The essence of the Bayesian approach is to provide a mathematical rule explaining how you change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. ...

The Economist, September 30th 2000

Bayes Theorem I

named after the English theologian and mathematician Thomas Bayes [1701–1761]

The theorem relies on the asymmetry of the definition of conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(B)\,P(A|B) \qquad (3)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A)\,P(B|A) \qquad (4)$$

for any two events A and B under regularity conditions, i.e. $P(B) \neq 0$ in (3) and $P(A) \neq 0$ in (4).

Bayes Theorem II

Thus, from $P(A|B)P(B) = P(B|A)P(A)$ follows

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \overset{\text{Law of tot. prob.}}{=} \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

More general, let $A_1, \ldots, A_n$ be *exclusive* and *exhaustive* events, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Interpretation

$P(A_i)$ prior probabilities

$P(A_i|B)$ posterior probabilities

After observing B the prob. of A_i changes from $P(A_i)$ to $P(A_i|B)$.

Towards inference

A more general formulation of Bayes theorem is given by

$$f(X = x|Y = y) = \frac{f(Y = y|X = x)f(X = x)}{f(Y = y)}$$

where X and Y are random variables.

(Note: Switch of notation from $P(.)$ to $f(.)$ to emphasise that we do not only relate to probabilities of events but to general probability functions of the random variables X and Y.)

Even more compact version

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

Posterior distribution

The posterior distribution is the most important quantity in Bayesian inference. It contains all information about the unknown parameter θ after having observed the data $X = x$.

Let $X = x$ denote the observed realisation of a random variable or random vector X with density function $f(x|\theta)$. Specification of a prior distribution with density function $f(\theta)$ allows to compute the density function of the posterior distribution using Bayes theorem:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}.$$

For discrete parameter space the integral has to be replaced with a sum.

Posterior distribution (II)

Since the denominator in

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

does not depend on θ, the density of the posterior distribution is proportional to

$$\underbrace{f(\theta|x)}_{\text{Posterior}} \propto \underbrace{f(x|\theta)}_{\text{Likelihood}} \times \underbrace{p(\theta)}_{\text{Prior}}$$

where $1/\int f(x|\theta)f(\theta)d\theta$ is the corresponding normalising constant to ensure $\int f(\theta|x)d\theta = 1$.

Reminder:

A likelihood approach uses only the likelihood and calculated Maximum Likelihood estimate (MLE), defined as the particular value of θ that maximises the likelihood.

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathrm{E}(\theta|x)$:

$$\mathrm{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

- Posterior mode $\mathrm{Mod}(\theta|x)$:

$$\mathrm{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$$

- Posterior median $\mathrm{Med}(\theta|x)$ is defined as the value a which satisfies

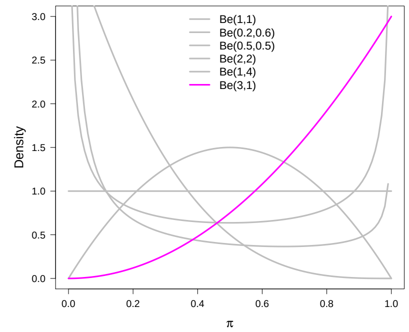
$$\int_{-\infty}^{a} f(\theta|x)d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} f(\theta|x)d\theta = 0.5$$

Binomial experiment

Let $X \sim \mathrm{Bin}(n, p)$ with n known and $p \in \Pi = (0, 1)$ unknown.

Since p is constrained to be within 0 and 1, a usual prior distribution is a beta distribution, so that

$p \sim \mathrm{Be}(\alpha, \beta)$ with $\alpha, \beta > 0$ and $\mathcal{T} = (0, 1)$.



Binomial experiment (2)

$$X \sim \mathrm{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \mathrm{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\mathrm{L}(p) = f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad f(p) = \frac{1}{\mathrm{B}(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto p^x (1-p)^{n-x} \qquad\qquad \propto p^{\alpha-1}(1-p)^{\beta-1}$$

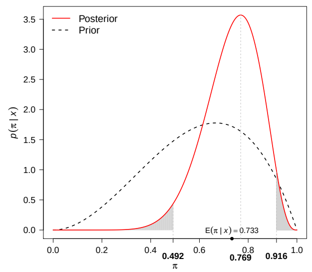
Thus, the posterior distribution results as:

$$\begin{aligned} f(p|x) &\propto f(x|p) \times f(p) \\ &= p^x (1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

This corresponds to the core of a beta distribution, so that

$$p|x \sim \mathrm{Be}(\alpha + \underbrace{x}_{\text{successes}}, \beta + \underbrace{n-x}_{\text{failures}})$$

Binomial experiment: Simple example



Posterior density of $p|x$ for a $\mathrm{Be}(3, 2)$ prior and observation $x = 8$ in a binomial experiment with $n = 10$ trials. An equi-tailed 95% credible interval is also shown.

Using a Be(1,1) the posterior mode equals the Maximum Likelihood (ML) estimate.