

Monte Carlo and MC-estimates

Geoff Nicholls

11/11/10

1. Motivation, Bayes Factors
2. Rejection
 - (a) Algorithm
 - (b) Stopping
 - (c) BF's and rejection.
3. Importance sampling
 - (a) Algorithm (known normalising constants)
 - (b) Variance of weights
 - (c) Algorithm (unknown normalising constants)
4. MCMC
 - (a) Algorithm
 - (b) Convergence
5. Bridge Estimation
 - (a) Meng-Wong identity, estimator, MSE
 - (b) Examples and extensions

1. Motivation for Monte Carlo

A special case of generic interest: $\theta \rightarrow x \rightarrow y$
Observe $Y \sim f_Y(y|x)$. Know something $p(x|\theta)$
about x given θ and something $p(\theta)$ about θ .

Need to estimate $q_S = \Pr(X \in S|y, \theta)$ say.

$$p(x|y, \theta) = \frac{f_Y(y|x)p(x|\theta)}{m(y, \theta)}$$
$$m(y, \theta) = \sum_{x \in \Omega} f_Y(y|x)p(x|\theta)$$

For $q_S = E_{X|Y, \theta}(\mathbb{I}_{X \in S})$, simulate $\{X_t = x_t\}_{t=1}^n$,
 $X_t \sim X|y, \theta$ iid and estimate

$$\hat{q}_{S,n} = \frac{1}{n} \sum_{t=1}^n \mathbb{I}_{x_t \in S}$$

with

$$\hat{q}_{S,n} \sim N(q_S, \text{var}(\mathbb{I}_{X \in S})/n)$$

good at large n .

Expectations and inference

Model 1: $\theta = \theta_1$. Model 2: $\theta = \theta_2$.

Bayes factor is fundamental. Let $p(\theta_1) = p(\theta_2)$.

Then

$$\begin{aligned} B(1v.2) &= \frac{p(\theta_1|y)}{p(\theta_2|y)} \\ &= \frac{m(y, \theta_1)}{m(y, \theta_2)} \end{aligned}$$

Hard to provide reliable confidence intervals.

Take n samples $X_t \sim X|\theta$ from the prior

$$\begin{aligned} m(y, \theta) &= E_{X|\theta}(f_Y(y|X)) \\ \widehat{m}(y, \theta) &= \frac{1}{n} \sum_{t=1}^n f_Y(y|x'_t) \end{aligned}$$

and generate garbage.

We need efficient Monte Carlo algorithms to simulate given distributions, and estimators which make efficient use of simulated values. The two (algorithms and estimators) are linked.

2/2a Rejection

Want to simulate $X \sim p(x)$ with $p(x) = \tilde{p}(x)/C_p$

Can simulate $Z \sim q(z)$ for some $q(z) = \tilde{q}(z)/C_q$.

Suppose we can find a constant M satisfying $M \geq \tilde{p}(x)/\tilde{q}(x)$ for all $x \in \Omega$. The following 'Rejection algorithm' returns $X \sim p$.

Algorithm 1 *Let $Z \sim q$ and $U \sim U(0, 1)$.*

1. *Simulate $Z = z$ and $U = u$.*
2. *If*

$$u \leq \frac{\tilde{p}(z)}{M\tilde{q}(z)}$$

then stop and return $X = z$, and otherwise, start again at 1.

2b) we can 'stop' the RA at the end of a trial, instead of stopping after a success. The stopped algorithm doesn't 'while'.

Suppose we run the RA n times to realise X_1, X_2, \dots, X_n , iid realisations of $X \sim p$.

T_i = number of trials Step 1 \rightarrow Step 2 taken to realise X_i .

T_i and X_i are independent.

Algorithm 2

Simulate J pairs $Z = (Z_1 = z_1, \dots, Z_J = z_J)$, $U = (U_1 = u_1, \dots, U_J = u_J)$, collect

$$S = \left\{ z \in Z; u_i \leq \frac{p(z)}{Mq(z)} \right\},$$

and identify $(X_1, X_2, \dots, X_N) = S$ (N random).

Cant we stop the RA on the clock?

Algorithm 3 *Let a be the start time (of the day), let t be the current time and let b be tea-time.*

Repeat until tea time

- 1. call the RA, accumulating $S = (X_1, X_2, \dots, X_N)$.*
- 2. If $t = b$ press Ctrl-C and stop the process.*

This algorithm may return biased realisations. The time $\tau(z)$ to compute $p(z)/Mq(z)$ may depend on z . Might be relevant for massively parallel rejection.

Take an example with two states $X = 0$ and $X = 1$. Suppose $\tau(0) < (b - a) < \tau(1)$. Then $X_1 = X_2 = \dots = X_N = 0$ (no 1's).

2c) better models give more efficient rejection

$T_i \sim \text{Geometric}(r)$, with $r = C_p/(MC_q)$, so

$$\begin{aligned} r &= \Pr\left(U \leq \frac{\tilde{p}(Z)}{M\tilde{q}(Z)}\right) \\ &= \mathbb{E}\left[\Pr\left(U \leq \frac{\tilde{p}(Z)}{M\tilde{q}(Z)} \mid Z\right)\right] \\ &= \mathbb{E}\left[\frac{\tilde{p}(Z)}{M\tilde{q}(Z)}\right] \\ &= C_p/(MC_q) \end{aligned}$$

and mean $\mu = \mathbb{E}(T_i)$ with $\mu = MC_q/C_p$. Then

$$\hat{\mu} = n^{-1} \sum_{i=1}^n T_i$$

estimates MC_q/C_p .

In particular if the target dbn is a posterior dbn

$$p(x|y, \theta) = \frac{f_Y(y|x)p(x|\theta)}{m(y, \theta)}$$

and the proposal $Z \sim p(z|\theta)$ is its prior, and $M \geq f_Y(y|x)$ all x , then

Algorithm 4 *Let $Z \sim p(z|\theta)$ and $U \sim U(0, 1)$.*

1. *Simulate $Z = z$ and $U = u$.*
2. *If*

$$u \leq \frac{f_Y(y|z)}{M}$$

then stop and return $X = z$, and otherwise, start again at 1.

Now $C_q/C_p = 1/m(y, \theta)$, $E(T) = (Mm)^{-1}$, and $M\hat{\mu}$ estimates (the inverse of) the marginal likelihood.

Better model, larger $m(y, \theta)$, smaller $\mu = E(T_i)$, longer tea breaks. But, prior and posterior are typically 'far apart' (in units of rejection trials).

Related estimator for $m(y, \theta_1)/m(y, \theta_2)$.

$p(x|y, \theta_2)$ and $p(x|y, \theta_1)$ may be closer, so if

$$p(x|y, \theta_2) = \frac{f_Y(y|x)p(x|\theta_2)}{m(y, \theta_2)}$$

and $Z \sim p(z|y, \theta_1)$ with $M \geq p(x|\theta_2)/p(x|\theta_1)$ all x , then

Algorithm 5 Let $Z \sim p(z|y, \theta_1)$, $U \sim U(0, 1)$.

1. Simulate $Z = z$ and $U = u$.
2. If

$$u \leq \frac{p(z|\theta_2)}{Mp(z|\theta_1)}$$

then stop and return $X = z$, and otherwise, start again at 1.

simulates $X \sim p(x|y, \theta_2)$ with

$$ME(T) = \frac{m(y, \theta_1)}{m(y, \theta_2)}.$$

Week 8, see that this estimator (and some other published estimators for BF's) belongs to a larger family of estimators. That family includes estimators (functions of *the same* output) that beat $\hat{\mu}/M$ in MSE (by miles).

Importance sampling

Want $E(f(X))$ for $X \sim p(x)$.

Can simulate $Z \sim q(z)$, $p(x) > 0 \Rightarrow q(x) > 0$.

Let $Z_i \sim q$, $i = 1, 2, \dots, n$ iid and set

$$W_i = p(Z_i)/q(Z_i).$$

Then

$$\bar{f} = \frac{1}{n} \sum_{i=1}^n W_i f(Z_i)$$

is an unbiased for $E(f(X))$.

Example 1 (Recycling) Suppose $Y_i \sim \text{Gamma}(a, b)$ and need $E_p(f(X))$ for $X \sim \text{Gamma}(\alpha, \beta)$.

$$p(x) = x^{\alpha-1} \exp(-\beta x) \beta^\alpha / \Gamma(\alpha).$$

$$W_i = Z_i^{\alpha-a} e^{-(\beta-b)Z_i} \times \frac{\Gamma(a)\beta^a}{\Gamma(\alpha)b^a}.$$

If $0 < \text{var}(\bar{f}) < \infty$ then have CLT for \bar{f} .

$$\begin{aligned}
 \text{var}(\bar{f}) &= \frac{1}{n} \text{var} \left(\frac{p(Z_1)}{q(Z_1)} f(Z_1) \right) \\
 &= \frac{1}{n} \left(E_q \left(\frac{p^2}{q^2} f^2 \right) - E_q \left(\frac{p}{q} f \right)^2 \right) \\
 &= \frac{1}{n} \left(E_p \left(\frac{p}{q} f^2 \right) - E_p(f)^2 \right). \quad (1)
 \end{aligned}$$

Check variance is finite! (check $E(p f^2 / q)$).

Example 2 (*Recycling Gammas... see R notes*)

$$E_p(p f^2 / q) \propto \int_0^\infty f(x)^2 x^{\alpha-1} e^{-\beta x} x^{\alpha-a} e^{-(\beta-b)x} dx.$$

$$E_p(f^2) \propto \int_0^\infty f(x)^2 x^{\alpha-1} \exp(-\beta x) dx$$

$E_p(p f^2 / q) < \infty$ if $a < \alpha$ and $b < \beta$ (so the extra factor under the integrand is bounded). This is sufficient for p/q bounded so could do rejection.

Importance sampling v. Rejection (I. v. R.)

It is nice to see iid samples. However, I. beats R. for statistical efficiency. Suppose we use Algorithm 2 (fix $(Z_i, U_i)_{i=1}^J$).

$$\begin{aligned}\bar{f}_R &= \frac{\sum_{j=1}^J \mathbb{I}(U_i \leq p(Z_i)/Mq(Z_i))f(Z_i)}{\sum_{j=1}^J \mathbb{I}(p(Z_i)/Mq(Z_i))} \\ &= A/B\end{aligned}$$

$$E(B) = J/M$$

$$\bar{f}'_R = \frac{M}{J} \sum_{j=1}^J \mathbb{I}(U_i \leq p(Z_i)/Mq(Z_i))f(Z_i)$$

$$\text{var}(\bar{f}'_R) = \frac{1}{J} \left[E(Mf(X)^2) - E(f(X))^2 \right]$$

$$\text{var}(\bar{f}_I) = \frac{1}{J} \left[E(W(X)f(X)^2) - E(f(x))^2 \right]$$

with $W(X) = p(X)/q(X)$. Now $M > p/q$ so I beats R.

Variance reduction:

Same output, different estimating function, lower variance.

For example, $X \sim N(0, \sigma^2)$ with $\sigma \ll 1$ and $f(X) = \mathbb{I}_{X>1}$. Suppose $Z_i \sim N(0, 1)$.

$$\bar{f}' = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\sigma Z_i > 1}$$

and

$$\bar{f} = \frac{1}{n\sigma} \sum_{i=1}^n e^{-Z_i^2(1/2\sigma^2-1/2)} \mathbb{I}_{\sigma Z_i > 1}.$$

IS estimate \bar{f} has lower variance if $E_p(pf^2/q) < E_p(f^2)$

$$E_p(f^2) = \int_1^\infty p(x) dx$$
$$E_p(pf^2/q) = \int_1^\infty p(x) \sigma^{-1} e^{-x^2(1/2\sigma^2-1/2)} dx$$

and $\sigma^{-1} e^{-x^2(1/2\sigma^2-1/2)} < 1$ when $x > 1$ all sufficiently small σ .

Choice of q .

$$\text{var}(\bar{f}) = \frac{1}{n} \left(E_p \left(\frac{p}{q} f^2 \right) - E_p(f)^2 \right)$$

$q(x) = p(x)f(x)/E_p(f)$: zero variance!

Weights $W_i = p(Z_i)/q(Z_i)$ so we need

$$W_i = E_p(f)/f(Y_i)$$

for \bar{f} , but $E_p(f)$ unknown. Guides choice of q .

Importance Sampling II

Dont know NC's C_p and C_q

$$\begin{aligned} E(f(X)) &= E_q(pf/q) \\ &= \frac{E_q(\tilde{p}f/\tilde{q})}{C_p/C_q} \\ &= \frac{E_q(\tilde{p}f/\tilde{q})}{E_q(\tilde{p}/\tilde{q})} \end{aligned}$$

Algorithm 6 *IS.II*

1 Simulate $Z_i \sim q$ iid for $i = 1, 2, \dots, n$.

2 Set $\tilde{W}_i = \tilde{p}(Z_i)/\tilde{q}(Z_i)$ and

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n \tilde{W}_i f(Z_i)$$

and

$$\bar{b} = \frac{1}{n} \sum_{i=1}^n \tilde{W}_i.$$

3 The IS-estimator for $E_p(f)$ is $\bar{f} = \bar{a}/\bar{b}$.

\bar{f} biased but consistent: for each $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr(|\bar{f} - E_p(f)| > \epsilon) = 0.$$

The asymptotic distribution can be calculated using the δ -method.

$$\begin{aligned} \bar{f} &= \frac{E_q(\tilde{w}f)}{E_q(\tilde{w})} + (\hat{a}_N - E_q(\tilde{w}f)) \frac{1}{E_q(\tilde{w})} \\ &\quad - (\hat{b}_N - E_q(\tilde{w})) \frac{E_q(\tilde{w}f)}{E_q(\tilde{w})^2} + \text{remainder} \\ &\rightarrow E_p(f) + 0 + 0 + 0 \quad \text{in probability as } n \rightarrow \infty \end{aligned}$$

Example 3 (*Recycling Gammas... see R notes*)

Surprisingly, simulation studies show that importance sampling using method II often gives more stable estimates of $E_p(f)$ than does method I - some authors recommend using II even when you know and can easily calculate all the normalizing constants. In my experience it is easy to make programming errors, even for simple normalizing functions, when using method I, so I tend to use method II.