Προσομοίωση

Στο προηγούμενο παράδειγμα καταφέραμε να παράγουμε τυχαίες τιμές με την βοήθεια της συνάρτησης rgamma στην R.

Υπάρχουν γενικοί αλγόριθμοι προσομοίωσης από οποιαδήποτε κατανομή; Εδώ θα αναφερθούμε σε έναν εξ αυτών την **Μέθοδο Απόρριψης (Rejection Sampling – von Neumann, 1951)**

Παράδειγμα

- Την άνοιξη του 1993 μια έρευνα έγινε στο “University of Berkeley”, για την χρήση ποδηλάτων στο χώρο του πανεπιστημίου. Σε δείγμα n οχημάτων τα s ήταν ποδήλατα. Για να βγάλουμε συμπεράσματα για το πραγματικό ποσοστό θ των ποδηλάτων στον χώρο του πανεπιστημίου χρησιμοποιούμε το ακόλουθο μοντέλο:

Παράδειγμα

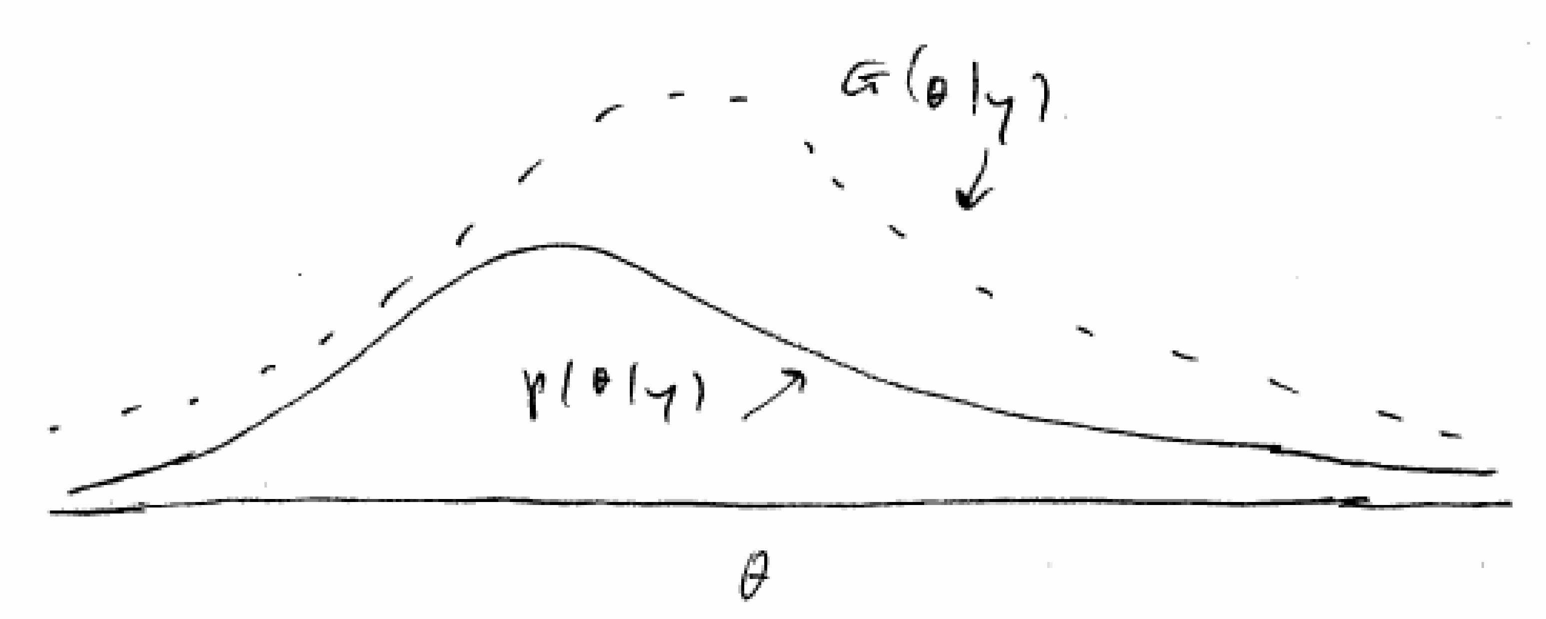
$$\begin{Bmatrix} \theta \sim \mathrm{Beta}(\alpha_0, \beta_0) \\ (S \mid \theta) \sim \mathrm{Binomial}(n, \theta) \end{Bmatrix} \to (\theta \mid s) \sim \mathrm{Beta}(\alpha_0 + s, \beta_0 + n - s)$$

Θα ένιωθα μεγάλη έκπληξη αν το ποσοστό των ποδηλάτων ήταν μικρότερο του 5% και μεγαλύτερο του 50%, οπότε μια αρκετά καλή επιλογή θα ήταν να χρησιμοποιήσω ως εκ των προτέρων κατανομή την Beta(2.0, 6.4) η οποία έχει μέση τιμή 0.24 και τυπική απόκλιση 0.14. Ας υποθέσουμε ότι το δείγμα μας ήταν μεγέθους n=74 και ο αριθμός των ποδηλάτων ήταν s=16. Άρα η εκ των υστέρων κατανομή είναι η Beta(18.0, 64.4). Πως μπορούμε να προσομοιώσουμε ένα IID Monte Carlo δείγμα από αυτή την κατανομή;

Μέθοδος Απόρριψης

- Ας υποθέσουμε ότι θέλουμε να παράγουμε τυχαίους αριθμούς από την $p(\theta \mid \mathbf{y})$ αλλά είναι δύσκολο. Ταυτόχρονα, είναι εύκολο να παράγουμε τυχαίους αριθμούς από μία άλλη συνάρτηση $G(\theta \mid \mathbf{y})$ (φάκελος) τέτοια ώστε:
  1. $G(\theta \mid \mathbf{y}) \geq p(\theta \mid \mathbf{y}) \geq 0$
  2. Η $g(\theta \mid \mathbf{y})$ είναι η σ.π.π. που προκύπτει αφού κανονικοποιήσουμε την G και καλείται κατανομή εισήγησης (proposal).

Μέθοδος Απόρριψης



Τότε για να παράγουμε τυχαίους αριθμούς από την p, παράγουμε τυχαίους αριθμούς θ^* από την g (**proposal distribution**) και τους δεχόμαστε ή τους απορρίπτουμε (όταν απορρίψεις το θ^*, συνεχίζεις μέχρι να αποδεχτείς κάποιο άλλο) με πιθανότητα αποδοχής

$$\alpha_R(\theta^* \mid \mathbf{y}) = \frac{p(\theta^* \mid \mathbf{y})}{G(\theta^* \mid \mathbf{y})}$$

Μέθοδος Απόρριψης

Ο von Neumann έδειξε ότι η παραπάνω επιλογή πιθανότητας αποδοχής, πράγματι παράγει IID τυχαίους αριθμούς από την p.

Γιατί;

Για να προσομοιώσεις τυχαίες τιμές από την p πρέπει να επιλέξεις τυχαία σημεία του επιπέδου κάτω από την καμπύλη p(θ|**y**) με τέτοιο τρόπο ώστε όλα τα πιθανά σημεία να έχουν την ίδια πιθανότητα και εν συνεχεία να αναφέρεις την τιμή θ. Αν αντί αυτού παράγουμε τυχαίες τιμές από την G, για να είναι αυτά σημεία και της p θα πρέπει να εξαιρέσουμε όσα είναι μεταξύ της G(θ|**y**) και της p(θ|**y**) και αυτό επιτυγχάνεται με την παραπάνω πιθανότητα επιλογής.

Μέθοδος Απόρριψης

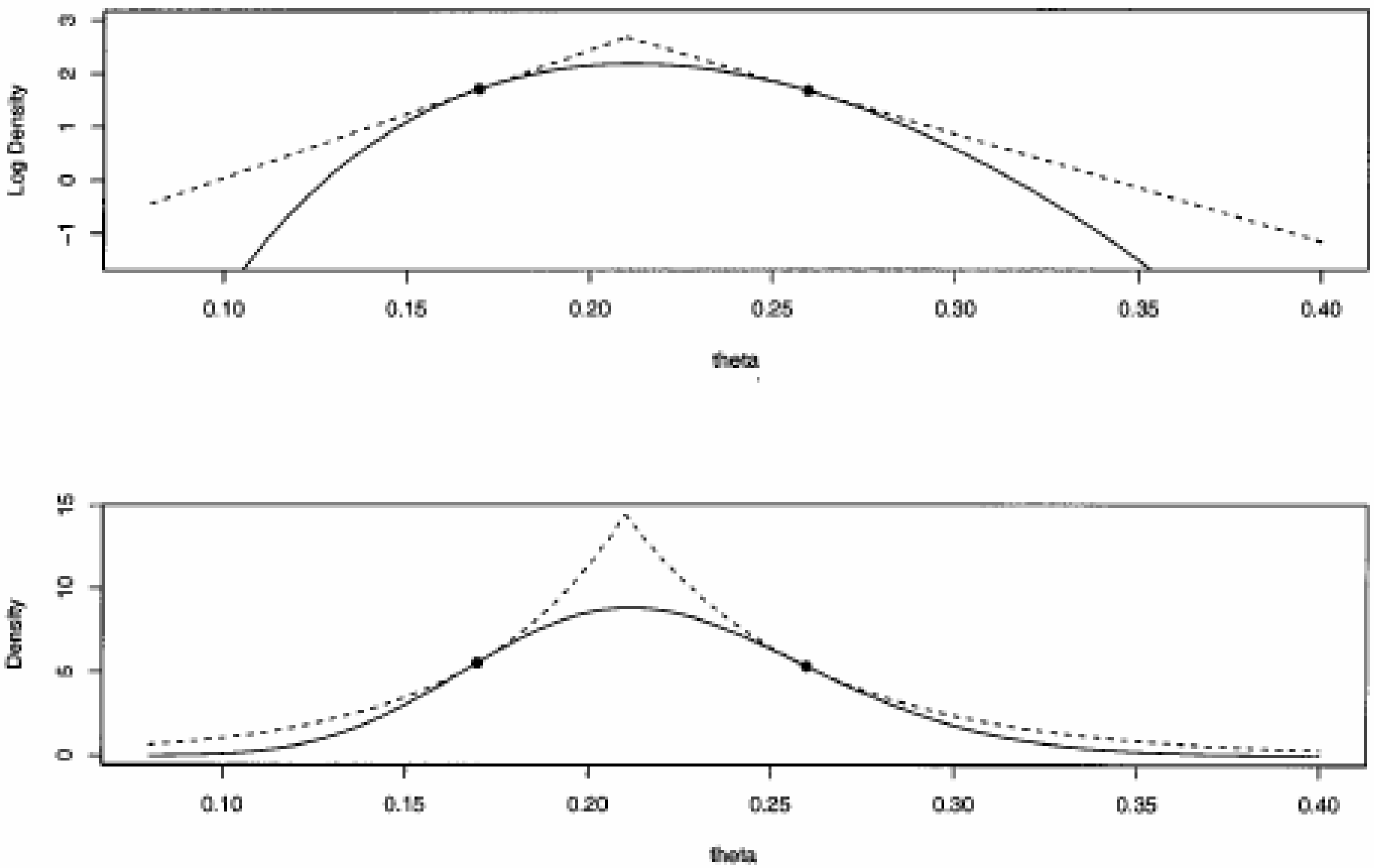
Algorithm (rejection sampling). To make m IID draws at random from the density $p(\theta|y)$ for real-valued θ, select an integrable **envelope function** G—which when normalized to integrate to 1 is the **proposal distribution** g—such that $G(\theta|y) \geq p(\theta|y) \geq 0$ for all θ; define the acceptance probability $\alpha_R(\theta^*|y) = \frac{p(\theta^*|y)}{G(\theta^*|y)}$; and

```
Initialize t ← 0
Repeat {
  Sample θ* ~ g(θ|y)
  Sample u ~ Uniform(0,1)
  If u ≤ α_R(θ*|y) then
    { θ_{t+1} ← θ*;  t ← (t + 1) }
}
until t = m.
```

Παράδειγμα

- Πίσω στο παράδειγμα μας όπου θέλουμε να να προσομοιώσουμε ένα IID Monte Carlo δείγμα από την Beta(18.0, 64.4). Η συγκεκριμένη Beta σ.π.π. είναι log-concave (ο λογάριθμος της είναι κοίλη συνάρτηση, δηλαδή $\frac{\partial^2 \log f(x)}{\partial x^2} \leq 0 \quad \forall x$), οπότε είναι εύκολο να φτιάξουμε ένα τμηματικά γραμμικό φάκελο, απλά φέρνοντας εφαπτόμενες σε κάθε πλευρά του μεγίστου.

Παράδειγμα



Παράδειγμα

- Η βέλτιστη επιλογή των εφαπτόμενων σημείων θα πρέπει να γίνεται με τρόπο τέτοιο ώστε να μεγιστοποιείται η περιθώρια πιθανότητα επιλογής:

$$P(\text{δέχομαστε το } \theta) = \int_{-\infty}^{+\infty} \alpha_R(\theta \mid y) g(\theta \mid y) d\theta =$$

$$\int_{-\infty}^{+\infty} \frac{p(\theta \mid y)}{Mg(\theta \mid y)} g(\theta \mid y) d\theta = 1/M = \left[\int_{-\infty}^{+\infty} G(\theta \mid y) d\theta \right]^{-1}$$

Άρα θα πρέπει να ελαχιστοποιήσουμε την περιοχή κάτω από τον φάκελο G, με την προϋπόθεση βέβαια πως υπερκαλύπτει την p. Στο παράδειγμα μας τα εφαπτόμενα σημεία έχουν x-συνιστώσες 0.17 και 0.26 αντίστοιχα και η περιθώρια πιθανότητα επιλογής είναι 0.75. Βεβαίως μπορούμε να αυξήσουμε αυτή την πιθανότητα, απλά προσθέτοντας και άλλες εφαπτόμενες. Οι τμηματικοί γραμμικοί φάκελοι σε log κλίμακα είναι μία καλή επιλογή διότι τότε η g αποτελείται από δύο εκθετικές κατανομές από τις οποίες μπορούμε εύκολα να προσομοιώσουμε τιμές (με την μέθοδο αντιστροφής).

Η προϋπόθεση για IID

- Στην πράξη η δημιουργία IID δείγματος από μία κατανομή δεν είναι και τόσο εύκολη ειδικά όταν το **θ** είναι πολυδιάστατο (σκεφθείτε να πρέπει να βρείτε ένα φάκελο για την p(**θ**|**y**) όταν π.χ. το **θ** έχει διάσταση 10). Το ευτύχημα είναι ότι στην αρχική τους ιδέα για προσομοίωση οι Metropolis και Ulam δεν είχαν αναφερθεί σε IID δείγμα. Αντί λοιπόν $\theta_i^* \overset{\text{IID}}{\sim} p(\theta \mid \mathbf{y})$ μπορούμε να θεωρήσουμε δείγμα $\theta_1^*, \ldots, \theta_m^*$ που διαμορφώνει μια χρονοσειρά (δηλαδή το θ_j^* μπορεί να εξαρτάται από το $\theta_{j'}^*$ για j'<j).
- Οι Metropolis et al. (1953) συνδύασαν την ιδέα του von Neumann με τις Μαρκοβιανές Αλυσίδες. Ο συνδυασμός του Monte Carlo δείγματος με τις αλυσίδες Markov (Markov chains) γεννά την τεχνική για την οποία θα αναφερθούμε στην διάρκεια του μαθήματος **Markov chain Monte Carlo** (**MCMC**).