Ασκήσεις Προσομοίωσης - Σειρά 2

Α. Μπουρνέτας, ΠΜΣ 'Βιοστατιστική'

1 Δειγματοληψία Σπουδαιότητας

ΠΡΟΒΛΗΜΑ 1. Έστω ότι μια βιολογική μέτρηση X ενός ατόμου ακολουθεί εκθετική κατανομή με παράμετρο 1. Η τιμή θεωρείται φυσιολογική αν $X \leq 5$. Ο βαθμός σοβαρότητας της τιμής ορίζεται ως $\max(X-5, 0)$, δηλαδή αν $X \leq 5$ δεν υπάρχει πρόβλημα ($Y = 0$), ενώ για μετρήσεις εκτός της φυσιολογικής περιοχής ο βαθμός σοβαρότητας είναι η υπέρβαση της X από το όριο του 5. Έστω $\theta = E(\max(X-5, 0))$, όπου $X \sim Exp(1)$, ο μέσος βαθμός σοβαρότητας.
(α) Να βρεθεί μια εκτίμηση του θ βασισμένη σε 1000 σενάρια προσομοίωσης, και να υπολογιστεί το αντίστοιχο διάστημα εμπιστοσύνης 95%.
(β) Να βρεθεί μια εκτίμηση του θ βασισμένη σε 1000 σενάρια προσομοίωσης, και να υπολογιστεί το αντίστοιχο διάστημα εμπιστοσύνης 95%, χρησιμοποιώντας δειγματοληψία σπουδαιότητας με κατανομή σπουδαιότητας την εκθετική $Exp(1/5)$.

Για το (α) μπορούμε να επαναλάβουμε τη διαδικασία του Προβλήματος 8 της Σειράς Ασκήσεων 1, με ρυθμό της εκθετικής κατανομής ίσο με 1, και τύπο $Y = \max(X-5, 0)$:

```
> N=1000;
> x=rexp(N,1);
> y=pmax(x-5,0);
> my=mean(y);
> my;
[1] 0.005759803
> sy=sd(y);
> sy;
[1] 0.1171935
> halflength=qt(0.975, N-1)*sy/sqrt(N);
> halflength;
[1] 0.007272406
> confint=c(my-halflength, my+halflength);
> confint;
[1] -0.001512604  0.013032209
> halflength/my
[1] 1.262614
```

Η εκτίμηση του θ από τα 1000 σενάρια είναι 0.0058, ενώ το μισό μήκος του ΔΕ είναι 0.0073, δηλαδή περίπου κατά 26% μεγαλύτερο του δειγματικού μέσου, επομένως η ακρίβεια δεν είναι καθόλου καλή με 1000 σενάρια προσομοίωσης.
(β) Πριν εφαρμόσουμε την δειγματοληψία σπουδαιότητας, παρατηρούμε ότι στην προηγούμενη προσομοίωση η μεγαλύτερη πιθανότητα της X συγκεντρώνεται στην περιοχή τιμών $X \leq 5$, για τις οποίες $Y = 0$. Επομένως στα περισσότερα σενάρια προσομοίωσης προκύπτει $Y = 0$ και για να συγκεντρωθούν αρκετές τιμές έτσι ώστε το θ να εκτιμηθεί με ακρίβεια απαιτείται πολύ μεγάλος

αριθμός σεναρίων. Πραγματικά για την εκθετική κατανομή με παράμετρο 1 ισχύει $P(X > 5) = e^{-5} = 0.0067$, δηλαδή περίπου 7 σενάρια στα 1000 θα δίνουν θετική τιμή του Y.
Για την εφαρμογή της μεθόδου δειγματοληψίας σπουδαιότητας παρατηρούμε ότι η θεωρητική τιμή του θ μπορεί να γραφεί ως ολοκλήρωμα με περισσότερους από ένα τρόπους. Συγκεκριμένα ισχύει

$$\theta = \int_0^\infty \max(x-5,0) f(x) dx,$$

όπου $f(x) = e^{-x}$ είναι η συνάρτηση πυκνότητας πιθανότητας της $Exp(1)$. Έστω μια οποιαδήποτε άλλη κατανομή με συνάρτηση πυκνότητας πιθανότητας $g(x)$ με τιμές επίσης στο διάστημα $(0, \infty)$. Τότε η θ μπορεί επίσης να εκφραστεί ως

$$\begin{aligned} \theta &= \int_0^\infty \max(x-5,0) \frac{f(x)}{g(x)} g(x) dx \\ &= E_g\left(\max(X-5,0) \frac{f(X)}{g(X)} \right) \end{aligned}$$

όπου E_g σημαίνει ότι η X τώρα ακολουθεί την κατανομή με σ.π.π. $g(x)$, και κάτω από αυτή την κατανομή υπολογίζουμε τη μέση τιμή μιας διαφορετικής συνάρτησης του X, δηλαδή της $\max(X-5,0)\frac{f(X)}{g(X)}$.
Ο εναλλακτικός τύπος για το θ σημαίνει ότι με οποιαδήποτε εναλλακτική κατανομή γίνει η προσομοίωση η μέση τιμή που θα εκτιμηθεί θα είναι η σωστή. Επομένως μπορούμε να επιλέξουμε την κατανομή $g(x)$ έτσι ώστε η εκτίμηση να γίνει με όσο το δυνατό μικρότερη διασπορά από την αρχική. Στο συγκεκριμένο παράδειγμα, αν επιλέξουμε μια κατανομή g τέτοια ώστε η περιοχή $X > 5$ να έχει μεγάλη πιθανότητα, τότε στα περισσότερα σενάρια η $\max(X-5,0)$ θα παίρνει θετικές τιμές και για το λόγο αυτό η εκτίμηση της μέσης τιμής μπορεί να γίνει με μεγαλύτερη ακρίβεια. Για παράδειγμα αν επιλέξουμε ως g την εκθετική κατανομή με μέση τιμή 10, δηλαδή παράμετρο $1/5$, $g(x) = \frac{1}{10}e^{-x/10}$. τότε η πιθανότητα του ενδεχομένου $X > 5$ είναι $P_g(X > 5) = e^{-2} = 0.61$, σημαντικά μεγαλύτερη από αυτή του (α). Αν υιοθετήσουμε την κατανομή αυτή ως κατανομή σπουδαιότητας, τότε ο λόγος

$$\frac{f(x)}{g(x)} = \frac{e^{-x}}{\frac{1}{10}e^{-x/10}} = 10e^{-9x/10},$$

επομένως

$$\theta = E_g(\max(X-5,0)10e^{-9x/10}),$$

όπου η X ακολουθεί εκθετική κατανομή με παράμετρο $1/10$. Η προσομοίωση της νέας συνάρτησης υπό τη νέα κατανομή μπορεί να γίνει με εντελώς αντίστοιχο κώδικα όπως στο (α):

```
> N=1000;
> x=rexp(N,1/10);
> y=pmax(x-5,0)*10*exp(-9*x/10);
> my=mean(y);
> my;
[1] 0.00712249
> sy=sd(y);
> sy;
```

```
[1] 0.01345978
> halflength=qt(0.975, N-1)*sy/sqrt(N);
> halflength;
[1] 0.0008352427
> confint=c(my-halflength, my+halflength);
> confint;
[1] 0.006287247 0.007957732
> halflength/my
[1] 0.1172684
```

Η εκτίμηση του θ από τα 1000 νέα σενάρια είναι 0.0071, ενώ το μισό μήκος του ΔΕ είναι 0.0008, δηλαδή μόνο το 11% του δειγματικού μέσου, σε σχέση με το 126% πριν. Δηλαδή με τη δειγματοληψία σπουδαιότητας η ακρίβεια της εκτίμησης από τον ίδιο αριθμό σεναρίων αυξήθηκε θεαματικά.

ΠΡΟΒΛΗΜΑ 2. Θεωρούμε το προηγούμενο πρόβλημα με την εξής παραλλαγή: Έστω ότι μετρώνται δύο μεγέθη και οι μετρήσεις είναι ανεξάρτητες μεταξύ τους, η μέτρηση X_1 ακολουθεί εκθετική κατανομή με παράμετρο 1 και η μέτρηση X_2 εκθετική κατανομή με παράμετρο 2. Το αποτέλεσμα είναι φυσιολογικό αν το άθροισμα των δύο μεγεθών $X_1 + X_2$ είναι μικρότερο του 7, ενώ ο βαθμός σοβαρότητας ορίζεται ως $\max(X_1 + X_2 - 7, 0)$. Έστω $\theta = E(\max(X_1 + Q_2 - 7, 0))$ ο μέσος βαθμός σοβαρότητας.
(α) Να βρεθεί μια εκτίμηση του θ βασισμένη σε 10000 σενάρια προσομοίωσης, και να υπολογιστεί το αντίστοιχο διάστημα εμπιστοσύνης 95%.
(β) Να βρεθεί μια εκτίμηση του θ βασισμένη σε 1000 σενάρια προσομοίωσης, και να υπολογιστεί το αντίστοιχο διάστημα εμπιστοσύνης 95%, χρησιμοποιώντας δειγματοληψία σπουδαιότητας με κατάλληλα επιλεγμένη κατανομή σπουδαιότητας.

Πριν κάνουμε τα πειράματα προσομοίωσης ας δούμε κάποια θεωρητικά στοιχεία του προβλήματος. Κατάρχήν επειδή $E(Q_1 + Q_2) = 3/2$, περιμένουμε η πιθανότητα $P(Q_1 + Q_2 > 7)$ να είναι πολύ μικρή και επομένως και $\theta \approx 0$. Περιμένουμε επομένως το ίδιο πρόβλημα με την απευθείας εκτίμηση όπως και στο προηγούμενο πρόβλημα. Η δυσκολία εδώ είναι επιπλέον ότι έχουμε το άθροισμα δύο ανεξάρτητων αλλά όχι ισόνομων εκθετικών τυχαίων μεταβλητών. Η κατανομή του αθροίσματος είναι γνωστή αλλά έχει όχι απλό τύπο, και ο θεωρητικός υπολογισμός του θ δεν είναι εύκολος.
(α) Για την απευθείας προσομοίωση ο αλγόριθμος δε διαφέρει πολύ από αυτόν του προηγούμενου προβλήματος. Εδώ σε κάθε σενάριο δημιουργούμε μια παρατήρηση από κάθε μια από τις δύο εκθετικές κατανομές και παίρνουμε το άθροισμά τους.

```
> N=10000; m=c(1,2); c=7;
> r=rep(0,N);
> for (i in 1:N)
+ {
+ x=-1/m*log(runif(2));
+ r[i]=max(sum(x)-c,0);
+ }
> m1=mean(r); m1;
[1] 0.001690872
> s1=sd(r); s1;
```

```
[1] 0.06462405
> h1=qt(0.975,N-1)*s1/sqrt(N); h1;
[1] 0.001266761
> ci1=c(m1-h1, m1, m1+h1);ci1;
[1] 0.0004241107 0.0016908722 0.0029576336
> h1/m1;
[1] 0.7491763
```

Παρατηρήστε ότι μέσα στο for loop η εντολή x=-1/m*log(runif(2)) δημιουργεί απευθείας ένα διάνυσμα X_1, X_2 από τις αντίστοιχες εκθετικές κατανομές, μέσω του αντίστροφου μετασχηματισμού $X = -\log(U)/\mu$. Το διάνυσμα m περιέχει τις παραμέτρους των δύο εκθετικών κατανομών. Υπενθυμίζουμε ότι στο R ο πολλαπλασιασμός/διαίρεση/δύναμη κλπ διανυσμάτων γίνεται στοιχείο προς στοιχείο, δηλαδή αν $m = (m_1, m_2), u = (u_1, u_2)$, τότε η πράξη $m * u)$ στο R έχει ως αποτέλεσμα το διάνυσμα (m_1u_1, m_2u_2). Εναλλακτικά θα μπορούσαμε να είχαμε χρησιμοποιήσει το παρακάτω ισοδύναμο for loop:

```
for (i in 1:N)
{
  x=c(0,0);
        x[1]=-1/m[1]*log(runif(1));
        x[2]=-1/m[2]*log(runif(1));
  r[i]=max(sum(x)-c,0);
}
```

Όσον αφορά τα αποτελέσματα, παίρνουμε μια εκτίμηση $\hat{\theta} = 0.00169$, και ένα διάνυσμα εμπιστοσύνης του οποίου το μισό μήκος είναι ίσο με 0.00127, δηλαδή περίπου 75% της εκτίμσης της μέσης τιμής. Επίσης με την εντολή

```
> sum(r>0)
[1] 11
```

βλέπουμε ότι από τα 10000 σενάρια μόνο στα 11 προέκυψε θετική τιμή της εκτίμησης, δηλαδή κάτω από τις συγκεκριμένες κατανομές το ενδεχόμενο $X_1 + X_2 > 7$ είναι πολύ σπάνιο. Επομένως η μέθοδος δειγματοληψίας σπουδαιότητας θα μπορούσε ενδεχομένως να βελτιώσει την ακρίβεια της εκτίμησης, αν χρησιμοποιηθούν κατάλληλα επιλεγμένες κατανομές σπουδαιότητας.

(β) Πριν προχωρήσουμε στην εφαρμογή της μεθόδου, πρέπει να δούμε πώς εφαρμόζεται όταν στην εκτίμηση εμπλέκονται περισσότερες από μια κατανομές. Στην περίπτωσή μας η άγνωστη μέση τιμή θ είναι ίση με

$$\theta = E(\max(X_1 + X_2 - c, 0)) = \int_{x_1=0}^{\infty} \int_{x_2=0}^{\infty} \max(x_1 + x_2 - c, 0) f_1(x_1) f_2(x_2) dx_2 dx_1,$$

όπου $f_1(x_1) = e^{-x_1}, f_2(x_2) = 2e^{-2x_2}$ οι πυκνότητες πιθανότητας των δύο κατανομών. Για να εφαρμόσουμε τη μέθοδο δειγματοληψίας σπουδαιότητας, πρέπει να επιλέξουμε επίσης δύο νέες

κατανομές σπουδαιότητας $g_1(x_1), g_2(x_2)$, για τις X_1, X_2, αντίστοιχα. Τότε η θ μπορεί να εκφραστεί ως εξής:

$$\begin{aligned}\theta &= \int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty}\max(x_1+x_2-c,0)f_1(x_1)f_2(x_2)dx_2dx_1 \\ &= \int_{x_1=0}^{\infty}\int_{x_2=0}^{\infty}\max(x_1+x_2-c,0)\frac{f_1(x_1)}{g_1(x_1)}\frac{f_2(x_2)}{g_2(x_2)}g_1(x_1)g_2(x_2)dx_2dx_1 \\ &= E_{g_1,g_2}\left(\max(X_1+X_2-c,0)\frac{f_1(X_1)}{g_1(X_1))}\frac{f_2(X_2)}{g_2(X_2)}\right)\end{aligned}$$

Δηλαδή σε κάθε σενάριο δημιουργούμε παρατηρήσεις από τις νέες κατανομές g_1, g_2, και εκτιμούμε τη μέση τιμή της συνάρτησης που μας ενδιαφέρει $\max(X_1+X_2-c,0)$ πολλαπλασιασμένης με το λόγο πιθανοφάνειας και των δύο παρατηρήσεων.
Το επόμενο ερώτημα είναι πώς πρέπει να επιλεγούν οι κατανομές g_1, g_2. Όπως και στο προηγούμενο παράδειγμα, η ιδέα είναι να επιλεγούν έτσι ώστε κάτω από αυτές το ενδεχόμενο $X_1+X_2>c$ να μην είναι σπάνιο. Για παράδειγμα αν επιλέξουμε $X_1, X_2 \sim Exp(1/4)$ ανεξάρτητες, τότε $E(X_1+X_2)=4+4=8$, και περιμένουμε κάτωαπό αυτές το άθροισμα να είναι συχνά μεγαλύτερο του 7. Μπορούμε επομένως να χρησιμοποιήσουμε ως κατανομή σπουδαιότητας την εκθετική με παράμετρο 4 και για τις δύο μεταβλητές.
Τέλος, αντί να υπολογίσουμε αναλυτικά το λόγο πιθανοφάνειας μέσω των τύπων των κατανομών, όπως κάναμε στο προηγούμενο πρόβλημα, μπορούμε να χρησιμοποιήσουμε απευθείας την πιθανότητας της εκθετικής κατανομής που υλοποιείται στο R μέσω της συνάρτησης δεξπ(ξ,λ).

```
> N=10000; m=c(1,2); v=c(1/4,1/4); c=7;
> r=rep(0,N);
> for (i in 1:N)
+ {
+ x=-1/v*log(runif(2));
+          r[i]=max(sum(x)-c,0)*dexp(x[1],m[1])*dexp(x[2],m[2])/(dexp(x
   [1],v[1])*dexp(x[2],v[2]));
+ }
> m2=mean(r);m2;
[1] 0.001885922
> s2=sd(r);s2;
[1] 0.007554311
> h2=qt(0.975,N-1)*s2/sqrt(N);
> ci2=c(m2-h2, m2, m2+h2);
> ci2;
[1] 0.001737843 0.001885922 0.002034002
> h2/m2;
[1] 0.07851845
```

Βλέπουμε ότι τώρα το μισό μήκος του διαστήματος εμπιστοσύνης είναι μόνο 7.9% της εκτίμησης της μέσης τιμής, δηλαδή και εδώ πετύχαμε μεγάλη βελτίωση της ακρίβειας μέσω της δειγματοληψίας σπουδαιότητας.

2 Δειγματολήπτης Gibbs

ΠΡΟΒΛΗΜΑ 3. Έστω το ζεύγος τυχαίων μεταβλητών (X, Y) με από κοινού συνάρτηση πυκνότητας πιθανότητας

$$f(x,y) = \begin{cases} c & \text{αν } 0 \le y \le x \le 1 \\ 0 & \text{διαφορετικά} \end{cases}$$

Να δημιουργηθεί ένα δείγμα 1000 ζευγών από την παραπάνω κατανομή χρησιμοποιώντας τον αλγόριθμο Gibbs Sampler και να εκτιμηθεί ο συντελεστής συσχέτισης των X, Y.

Πρώτα πρέπει να υπολογίσουμε τις δεσμευμένες κατανομές $f_{Y|X}(y|x), f_{X|Y}(x|y)$. Για τον υπολογισμό χρειάζονται οι περιθώριες κατανομές. Προσοχή στις αντίστοιχες περιοχές ολοκλήρωσης. Και οι δύο μεταβλητές παίρνουν τιμές μεταξύ 0 και 1, αλλά ΔΕΝ είναι ανεξάρτητες: Δεδομένης της τιμής της $X = x$, η περιοχή τιμών της Y είναι $0 \le y \le x$, και αντίστοιχα δεδομένου $Y = y$, η περιοχή τιμών της X είναι $y \le x \le 1$.
Με βάση τα παραπάνω, η περιθώρια κατανομή της X είναι

$$f_X(x) = \int_{y=0}^{1} f(x,y)dy = \int_{y=0}^{x} cdy = cx, \quad 0 \le x \le 1,$$

και η περιθώρια κατανομή της Y:

$$f_Y(y) = \int_{x=0}^{1} f(x,y)dx = \int_{x=y}^{1} cdx = c(1-y), \quad 0 \le y \le 1.$$

Επομένως οι δεσμευμένες πυκνότητες πιθανότητας είναι:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{c}{c(1-y)} = \frac{1}{1-y}, \quad y \le x \le 1.$$

Βλέπουμε επομένως ότι για κάθε y η δεσμευμένη πυκνότητα της X είναι **σταθερή** ως προς x στο διάστημα $y \le x \le 1$, δηλαδή δεδομένου $Y = y$, η X ακολουθεί ομοιόμορφη κατανομή $U(y, 1)$.
Αντίστοιχα υπολογίζουμε και τη δεσμευμένη πυκνότητα

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{c}{cx} = \frac{1}{x}, \quad 0 \le y \le x,$$

επομένως δεδομένου $X = x$, η Y ακολουθεί ομοιόμορφη κατανομή $U(0, x)$.
Οι δεσμευμένες κατανομές είναι εύκολα προσομοιώσιμες, επομένως η μέθοδος Gibbs Sampler μπορεί να εφαρμοστεί για την δημιουργία παρατηρήσεων. Ο αλγόριθμος αυτός περιγράφεται περιληπτικά ως εξής:

1. Ξεκινάμε με αυθαίρετα (X_0, Y_0), από την επιτρεπτή περιοχή τιμών $0 \le Y_0, \le X_0 \le 1$.
2. Στο γενικό βήμα t δημιουργούμε το ζεύγος (X_t, Y_t) ως εξής: $X_t \sim f_{X|Y}(x|Y_{t-1})$, και $Y_t \sim f_{Y|X}(y|X_t)$.
3. Επαναλαμβάνουμε τη διαδικασία μέχρι $t = T$ (για μεγάλο T).

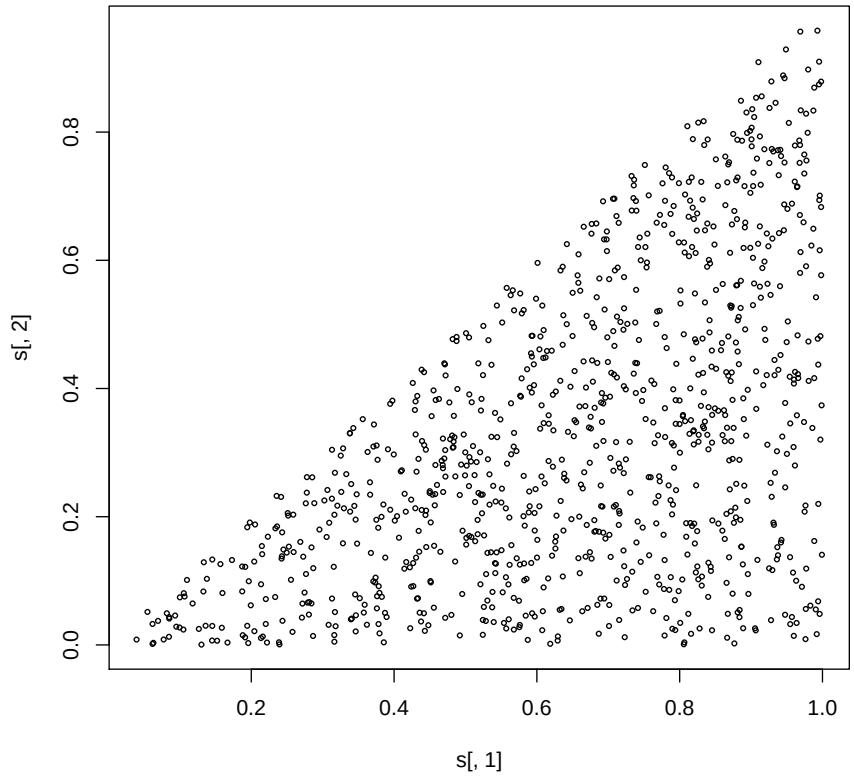
Η θεωρία στοχαστικών ανελίξεων εξασφαλίζει ότι για μεγάλη τιμή του T, η τελευταία παρατήρηση (X_T, Y_T) θα ακολουθεί με καλή προσέγγιση τη ζητούμενη κατανομή $f(x, y)$.
Για την υλοποίηση του αλγορίθμου υπολογιστικά μπορούμε να κάνουμε το εξής: Έστω ότι θέλουμε να δημιουργήσουμε N ζεύγη από την κατανομή αυτή. Χρησιμοποιούμε μια 'μεγάλη τιμή' για το T, π.χ. $T = 100$. Ξεκινώντας με αυθαίρετο (X_0, Y_0), επαναλαμβάνουμε τον αλγόριθμο για T βήματα και η πρώτη παρατήρησή μας είναι το ζεύγος (X_T, Y_T). Τώρα αντί να αρχίσουμε εκ νέου με αυθαίρετο αρχικό ζεύγος για να δημιουργήσουμε τη δεύτερη παρατήρηση, μπορούμε (αυθαίρετα) να πάρουμε ως αρχικό ζεύγος για τη δεύτερη επανάληψη την πρώτη παρατήρηση, δηλαδή το (X_T, Y_T), να επαναλάβουμε τον αλγόριθμο για T επιπλέον βήματα και ως δεύτερη παρατήρηση να πάρουμε το ζεύγος (X_{2T}, Y_{2T}). Συνεχίζουμε με το ζεύγος αυτό ως αρχική τιμή για την τρίτη παρατήρηση κ.ο.κ.
Για να δημιουργήσουμε λοιπόν N παρατηρήσεις μπορούμε να εφαρμόσουμε τον αλγόριθμο για συνολικά NT βήματα, και να πάρουμε ως δείγμα τα ζεύγη $(X_T, Y_T), (X_{2T}, Y_{2T}), \ldots, (X_{NT}, Y_{NT})$, δηλαδή τα ζεύγη που δημιούργησε ο αλγόριθμος με διαλείμματα T βημάτων μεταξύ τους.
Η θεωρία στοχαστικών ανελίξεων επίσης εξασφαλίζει ότι για μεγάλο T τα ζεύγη που θα δημιουργηθούν με τον παραπάνω τρόπο θα ακολουθούν την επιθυμητή κατανομή και επίσης ασυμπτωτικά θα είναι ανεξάρτητα μεταξύ τους, επομένως θα συνιστούν τυχαίο δείγμα από την κατανομή. Αν το T είναι μικρό και το N μεγάλο, δηλαδή το διάλειμμα μεταξύ των δειγμάτων είναι μικρό, τότε και πάλι τα ζεύγη ασυμπτωτικά θα ακολουθούν την κατανομή που θέλουμε, αλλά δε θα είναι ανεξάρτητα μεταξύ τους.
Με βάση τα παραπάνω μπορούμε να δημιουργήσουμε $N = 1000$ παρατηρήσεις από την παραπάνω κατανομή με $T = 100$, χρησιμοποιώντας τον παρακάτω κώδικα. Η μεταβλητή i δείχνει τον αριθμό παρατηρήσεων που έχουν δημιουργηθεί, ενώ η j τον αριθμό βημάτων που έχουν εκτελεστεί για τη δημιουργία της i-οστής παρατήρησης. Οι εντολές $x = y + (1 - y) * runif(1)$ και $y = x * runif(1)$ δημιουργούν τις γεννήτριες από τις αντίστοιχες ομοιόμορφες δεσμευμένες κατανομές. Ο πίνακας s διαστάσεων $N \times 2$ περιέχει το τελικό δείγμα που έχει δημιουργηθεί, στη στήλη 1 τις τιμές της X και στη στήλη 2 τις τιμές της Y.

```
> N=1000; T=100;
> s=rep(0,2*N);
> s=matrix(s,N,2);
> x=1/2; y=1/2;
> for (i in 1:N)
+ {
+     for (j in 1:T)
+     {
+         x=y+runif(1)*(1-y);
+         y=x*runif(1);;
+     }
+     s[i,]=c(x,y);
+ }
^[[A> cor(s);
           [,1]      [,2]
[1,] 1.0000000 0.4556465
[2,] 0.4556465 1.0000000
> mean(s[,1]);
[1] 0.6644191
> mean(s[,2]);
```

```
[1] 0.3264041
```

Η εκτίμηση του συντελεστή συσχέτισης $\rho(X,Y)$ από το δείγμα που δημιουργήθηκε είναι 0.456. Το scatterplot των X, Y φαίνεται στο Σχήμα 1.

```
> plot(s[,1],s[,2], cex=0.5)
```



Σχήμα 1: Δείγμα μεγέθους $N = 1000$ από την από κοινού συνάρτηση πυκνότητας πιθανότητας $f(x,y) = c, 0 \leq y \leq x \leq 1$ με μέθοδο Gibbs Sampler

ΠΡΟΒΛΗΜΑ 4. Θα μπορούσε να υλοποιηθεί το προηγούμενο πρόβλημα χωρίς τη μέθοδο Gibbs;

Στο συγκεκριμένο πρόβλημα έχουν υπολογιστεί εκτός από τις δεσμευμένες κατανομές και οι περιθώριες των X, Y, συγκεκριμένα βρήκαμε ότι

$$f_X(x) = \int_{y=0}^{1} f(x,y)dy = \int_{y=0}^{x} cdy = cx, \quad 0 \leq x \leq 1,$$

και

$$f_Y(y) = \int_{x=0}^{1} f(x,y)dx = \int_{x=y}^{1} cdx = c(1-y), \quad 0 \leq y \leq 1.$$

Και από τις δύο αυτές κατανομές μπορούμε εύκολα να δημιουργήσουμε γεννήτριες τυχαίων αριθμών, αρκεί να υπολογιστεί η σταθερά c, η οποία εύκολα εδώ προκύπτει από τη συνθήκη

$$\int_{x=0}^{1} \int y = 0^1 f(x,y)dydx = 1$$

ότι $c = 2$. Έτσι για παράδειγμα, για να δημιουργήσουμε ένα ζεύγος (X, Y) από τη δοσμένη κατανομή, μπορούμε να δημιουργήσουμε μια παρατήρηση από την περιθώρια κατανομή της X, έστω $X = x$ και με βάση αυτή να δημιουργήσουμε μια παρατήρηση από τη δεσμευμένη κατανομή της Y δοθέντος $X = x$, δηλαδή από την $f_{Y|X}(y|x$. Για την περιθώρια της X βρίσκουμε την αθροιστική συνάρτηση κατανομής

$$F_X(x) = \int_{s=0}^{x} 2sds = x^2, 0 \leq x \leq 1$$

και χρησιμοποιώντας τη γεννήτρια αντίστροφου μετασχηματισμού

$$F_X(X) = U_1 \Rightarrow X^2 = U_1 \Rightarrow X = \sqrt{U_1},$$

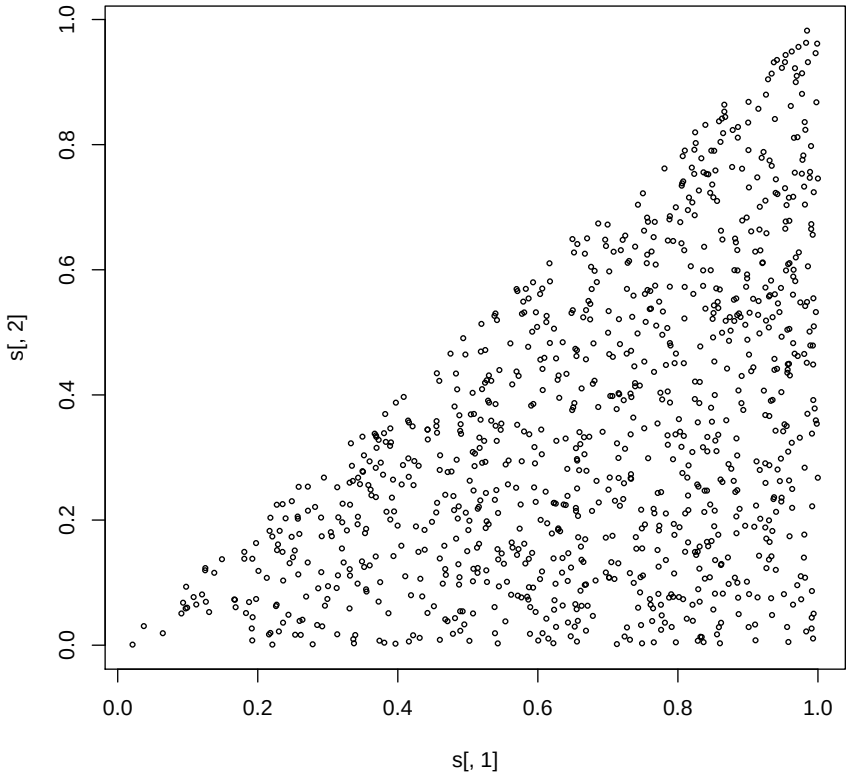
όπου $U_1 \sim U(0,1)$. Για τη δεσμευμένη κατανομή έχουμε δει στο προηγούμενο πρόβλημα ότι $Y|X = x \sim U(0,x)$, επομένως η γεννήτρια είναι $Y = xU_2$, όπου $U_2 \sim U(0,1)$.
Με τον παρακάτω κώδικα δημιουργούμε 1000 ανεξάρτητες παρατηρήσεις από την (X, Y).

```
> N=1000;
> s=rep(0,2*N);
> s=matrix(s,N,2);
> for (i in 1:N)
+ {
+      x=sqrt(runif(1));
+      y=x*runif(1);
+      s[i,]=c(x,y);
+ }
> cor(s);
          [,1]      [,2]
[1,] 1.0000000 0.4930739
[2,] 0.4930739 1.0000000
> mean(s[,1]);
[1] 0.6777101
> mean(s[,2]);
[1] 0.3390096
```

και το αντίστοιχο διάγραμμα στο Σχήμα 2.

```
> plot(s[,1],s[,2], cex=0.5)
```

Με τον τρόπο αυτό δημιουργούμε κάθε ζεύγος από την (X, Y), χρησιμοποιώντας μόνο δύο τυχαίους αριθμούς και μια επανάληψη αντί για τις T επαναλήψεις που απαιτούνται στη μέθοδο Gibbs. Επίσης εδώ οι παρατηρήσεις είναι πραγματικά ανεξάρτητες μεταξύ τους γιατί από τη μια στην άλλη



Σχήμα 2: Δείγμα μεγέθους $N = 1000$ από την από κοινού συνάρτηση πυκνότητας πιθανότητας $f(x, y) = c, 0 \leq y \leq x \leq 1$ με απευθείας προσομοίωση.

δεν διατηρείται καμιά τιμή, ενώ στη μέθοδο Gibbs ακόμα και για μεγάλες τιμές του T υπάρχει μια μικρή εξάρτηση μεταξύ των παρατηρήσεων και η ανεξαρτησία είναι προσεγγιστική. Επομένως θα προτιμούσε κανείς να χρησιμοποιεί αυτό τον απλούστερο τρόπο. Το πρόβλημα εδώ είναι ότι σε γενικότερα προβλήματα αφενός μπορεί να είναι δύσκολο έως αδύνατο να υπολογιστεί η τιμή της σταθεράς c, και αφετέρου η δημιουργία γεννήτριας από τις περιθώριες μπορεί επίσης να είναι πολύ δύσκολη. Η μέθοδος Gibbs έχει το πλεονέκτημα ότι χρησιμοποιεί μόνο τις δεσμευμένες και όχι τις περιθώριες κατανομές.

Άσκηση: Υπολογίστε τη θεωρητική τιμή του συντελεστή συσχέτισης (προκύπτει $\rho = 1/2$).