Στατιστική Απόσταση

Για τους παρακάτω αλγορίθμους, να υπολογιστεί η στατιστική απόσταση της εξόδου τους από την ομοιόμορφη κατανομή U στο $S = \{0, 1, 2, \ldots, A-1\}$.

Sampler 1. $n := \lceil \log_2 A \rceil$
$x_0, x_1, \ldots, x_{n-1} \leftarrow \{0, 1\}$
$y := \sum_{i=0}^{n-1} 2^i \cdot x_i$
return y

Sampler 2. $x_0, x_1, \ldots, x_{A-1} \leftarrow \{0, 1\}$
$y := \sum_{i=0}^{A-1} x_i$
return y

Sampler 3. $n := \lceil \log_2 A \rceil$
repeat:
  $x_0, x_1, \ldots, x_{n-1} \leftarrow \{0, 1\}$
  $y := \sum_{i=0}^{n-1} 2^i \cdot x_i$
  if $y < A$: return y

Sampler 1. Ορίζουμε $B = 2^n$, και ελέγχουμε ότι η έξοδος του sampler είναι μια τυχαία μεταβλητή Y που ακολουθεί την ομοιόμορφη κατανομή στο $\{0, 1, 2, \ldots, B-1\}$. Παρατηρούμε επίσης ότι $B >= A$, άρα το σύνολο τιμών της Y υπερκαλύπτει αυτό της U. Από τον ορισμό της στατιστικης απόστασης έχουμε:

$$
\begin{aligned}
\Delta(U, Y) =& \frac{1}{2} \sum_{i=0}^{B} |\Pr(U = 1) - \Pr(Y = i)| \\
&= \frac{1}{2} \sum_{i=0}^{A} |\Pr(U = 1) - \Pr(Y = i)| + \frac{1}{2} \sum_{i=A}^{B} |\Pr(U = 1) - \Pr(Y = i)| \\
&= \frac{1}{2} \sum_{i=0}^{A} \left| \frac{1}{A} - \frac{1}{B} \right| + \frac{1}{2} \sum_{i=A}^{B} \left| 0 - \frac{1}{B} \right| \\
&= \frac{1}{2} A \left| \frac{1}{A} - \frac{1}{B} \right| + \frac{1}{2} (B - A) \left| 0 - \frac{1}{B} \right| \\
&= \frac{1}{2} A \left(\frac{1}{A} - \frac{1}{B} \right) + \frac{1}{2} (B - A) \frac{1}{B} \\
&= \frac{1}{2} \left(\frac{A}{A} - \frac{A}{B} \right) + \frac{1}{2} \frac{B - A}{B} \\
&= \frac{1}{2} \left(1 - \frac{A}{B} + 1 - \frac{A}{B} \right) \\
&= 1 - \frac{A}{B}
\end{aligned}
$$

Οπως αναμένουμε, όταν $A = B$ η στατιστική απόσταση είναι μηδενική, ενώ όταν $A = 2^k - 1$ η διαφορά είναι σχεδόν $\frac{1}{2}$. Όταν το A είναι κοντά στο B, παρατηρούμε επίσης ότι η απόσταση είναι αμελητέα ως προς το n.

Sampler 2. Αναγνωρίζουμε ότι η κατανομή Y του sampler είναι η διωνυμική με παραμέτρους $p = q = \frac{1}{2}$ και $n = A$ το πλήθος δοκιμές. Ξέρουμε ότι η κατανομή αυτή έχει σχήμα καμπύλης, άρα πιστεύουμε ότι μακριά από το μέσο όρο, θα εμφανίζει χαμηλή πιθανότητα. Αφού ο μέσος όρος είναι $p \cdot A = \frac{A}{2}$, επιλέγουμε να εξετάσουμε τις τιμές από το $\frac{3A}{4}$ ως το A. Προφανώς για τη στατιστική απόσταση θα ισχύει ότι:

$$\begin{aligned}
\Delta(U,Y) =& \frac{1}{2}\sum_{i=0}^{A} |\Pr(U=1) - \Pr(Y=i)| \\
\leq& \frac{1}{2}\sum_{i=3A/4}^{A} |\Pr(U=1) - \Pr(Y=i)| \quad \text{(Παραλείπουμε θετικούς όρους, το άθροισμα δεν αυξάνεται)} \\
\leq& \frac{1}{2}\sum_{i=3A/4}^{A} (\Pr(U=1) - \Pr(Y=i)) \quad \text{(Αφαιρούμε απόλυτα, άρα το άθροισμα δεν αυξάνεται)} \\
=& \frac{1}{2}\sum_{i=3A/4}^{A} \Pr(U=1) - \frac{1}{2}\sum_{i=3A/4}^{A} \Pr(Y=i) \\
=& \frac{1}{2}\cdot\frac{1}{4} - \frac{1}{2}\sum_{i=3A/4}^{A} \Pr(Y=i) \\
=& \frac{1}{8} - \frac{1}{2}\sum_{i=3A/4}^{A} \Pr(Y=i)
\end{aligned}$$

Για να έχουμε λοιπόν μια εκτίμηση για την απόσταση, μένει να υπολογίσουμε (προσεγγιστικά) το άθροισμα πιθανοτήτων της Y για τιμές από $3A/4$ ως A, ή ισοδύναμα την πιθανότητα $P[Y >= \frac{3A}{4}]$. Από τις σημειώσεις γνωρίζουμε την ανισότητα του Chebychev:

$$\Pr[|Y - E(Y)| \geq t] \leq \frac{Var[Y]}{t^2}$$

Γνωρίζουμε ότι για τη διωνυμική ο μέσος όρος $E(Y)$ είναι $n \cdot p = \frac{A}{2}$, η διακύμανση $Var[Y]$ είναι $n \cdot p \cdot q = \frac{A}{4}$. Αντικαθιστούμε για $t = \frac{A}{4}$ και έχουμε:

$$\Pr[Y \geq \frac{3A}{4}] = \Pr[Y - \frac{A}{2} \geq \frac{A}{4}] = \Pr[Y - E(Y) \geq t] \leq \Pr[|Y - E(Y)| \geq t] \leq \frac{\frac{A}{8}}{\frac{A}{4}\cdot\frac{A}{4}} = \frac{\frac{1}{2}}{\frac{A}{4}} = \frac{1 \cdot 4}{2 \cdot A} = \frac{2}{A}$$

Επιστρέφοντας στο φράγμα για τη στασιστική απόσταση έχουμε ότι:

$$\begin{aligned}
\Delta(U,Y) =& \frac{1}{2}\sum_{i=0}^{A} |\Pr(U=1) - \Pr(Y=i)| \\
\leq& \frac{1}{8} - \frac{1}{2}\sum_{i=3A/4}^{A} \Pr(Y=i) \\
\leq& \frac{1}{8} - \frac{1}{2}\frac{2}{A} \\
=& \frac{1}{8} - \frac{1}{A}
\end{aligned}$$

Άρα για μεγάλες τιμές του A, η στατιστική απόσταση είναι σημαντική, και αυξάνεται όσο μεγαλώνει το A.

Παρατήρηση: Σε τρία σημεία της λύσης μπορούμε επικαλούμενοι τη συμμετρία να πάρουμε σημαντικά καλύτερες τιμές για τα φράγματα (περίπου 3 διπλασιασμούς). Πρώτα: μπορούμε να εξετάσουμε και το διάστημα από 0 έως $\frac{A}{4}$. Μετά: μπορούμε να ισχυριστούμε ότι αφού οι πιθανότητες και των δύο κατανομών αθροίζονται στο 1, η διαφορά που έχουν στο διάστημα $[0, A]$ θα είναι συνολικά 0, άρα η διαφορά στο κεντρικο διάστημα (χωρίς απόλυτη τιμή) θα είναι η αντίθετη της διαφοράς στα άκρα (οπότε εφαρμόζοντας την απόλυτη τιμή διπλασιάζουμε τη διαφορά που είχαμε). Τέλος, στην ανισότητα Chebychev φράξαμε την πιθανότητα να ξεπεράσουμε τα 3/4 με την πιθανότητα να απομακρυνθούμε από το 1/2 κατά 1/4 ή παραπάνω, η οποία λόγω συμμετρίας θα είναι περίπου διπλάσια.

Sampler 3. Θα κάνουμε χρήση της δεσμευμένης πιθανότητας. Οι τιμές του y επιλέγονται ομοιόμορφα από το $\{0, 1, 2, \ldots, B-1\}$ αλλά επιστρέφονται μόνο εάν ισχύει $y < A$. Ονομάζουμε Z την τυχαία μεταβλητή της εξόδου του sampler, και θεωρούμε την Y όπως στον πρώτο sampler . Από την κατασκευή του προγράμματος έχουμε ότι: $\Pr(Z = x) = \Pr(Y = x|Y < A)$. Από τον ορισμό της δεσμευμένης πιθανότητας:

$\Pr(Y = x|Y < A) = \Pr(Y = x|x < A) = \frac{\Pr((Y=x)\cap(x<A))}{\Pr(Y<A)}$

Όταν $x >= A$ η παραπάνω πιθανότητα είναι μηδενική. Στη μη τετριμμένη περίπτωση όπου $x < A$, έχουμε $\frac{\Pr(Y=x)}{\Pr(Y<A)} = \frac{\frac{1}{B}}{\frac{A}{B}} = \frac{1}{B} \cdot \frac{B}{A} = \frac{1}{A}$.

Άρα, για $x >= A$, $\Pr(Z = x) = 0$ και για $x < A$, $\Pr(Z = x) = \frac{1}{A}$, και η στατιστική απόσταση είναι προφανώς 0. Παρατηρούμε όμως ότι ο χρόνος εκτέλεσης δεν είναι σταθερός.