论伪随机数的偏差

（提要）

李依新
（中国科学院 沈阳计算技术研究所

卢庆堂
沈阳鼓风机厂

1982．6．

摘 要

本文从计算机所能表示的最大素数为模，该模之原根为乘子，从乘同余法所生成的伪随机数分布函数发，给出计算偏度全新的方法，得到较已知结果简单而精确的偏度表达式。从而，就整体而言为最佳伪随机数的确定提供了根据。

一. 引 言

对广泛采用的乘同余法而言，生成伪随机数的递推公式为：

$$x_{i+1} \equiv a x_i \pmod{M} \qquad (1.1)$$

$$\xi_i = x_i / M \quad , \quad i = 1, 2, \cdots, M$$

由数论可知，如果M选为素数，a为其任一原根，则生成伪随机数的循环长度为$M-1$，且达到最大。这比以前用2^s为模(1)的循环长度要长个倍。以下约定，本文所言伪随机数序列$\xi_1, \xi_2, \cdots, \xi_n$均以素数为模，$a$为模的原根，用(1.1)式生成的。而且把伪随机数的整体作为子样去讨论偏度。

二. 伪随机数的均匀性和独立性

1. 一维均匀偏度

定义2.1[2] 对任意$x \in [0,1]$，号$N_n(x)$表示伪随机数序列$\xi_1, \xi_2, \cdots, \xi_n$中满足$\xi_i < x$，$i = 1, 2, \cdots, n$的个数，则

$$\delta(n) = \sup_{0 \le x \le 1} \left| \frac{N_n(x)}{n} - x \right| \qquad (2.1)$$

称作伪随机数序列的偏度。

定理2.1，伪随机数$\xi_1, \xi_2, \cdots, \xi_n$的均匀偏度为$1/M$，与$a$是M的哪一个原根无关。即

$$\delta(n)=\sup_{0\le x\le 1}\left|\frac{N_n(x)}{n}-x\right|=\frac{1}{M} \qquad (2.2)$$

2．二维偏度

定义2.2[2] 对任意 x、$y\in(0,1)$，令 $N_n(x,y)$ 表示 $(\xi_1,\xi_2),(\xi_2,\xi_3),\cdots,(\xi_n,\xi_{n+1})$，$\xi_{n+1}=\xi_1$ 中满足 $\xi_i<x$，$\xi_{i+1}<y$ 的个数。则

$$\Delta(n)=\sup_{0\le x,y\le 1}\left|\frac{N_n(x,y)}{n}-xy\right| \qquad (2.3)$$

$$\varepsilon(n)=\sup_{0\le x,y\le 1}\left|\frac{N_n(x,y)}{n}-\frac{N_n(x)}{n}\frac{N_n(y)}{n}\right| \qquad (2.4)$$

分别称作伪随机数 $\xi_1,\xi_2,\cdots,\xi_n$ 的二维均匀偏度和独立偏度。

且 $\Delta(n)$ 和 $\varepsilon(n)$ 有如下关系[2]

$$\begin{aligned}\Delta(n)&\le\varepsilon(n)+2\delta(n)\\ \varepsilon(n)&\le\Delta(n)+2\delta(n)\end{aligned} \qquad (2.5)$$

3．伪随机数对的分布定理

由(1.1)式可知，x_{i+1} 可表为

$$x_{i+1}=ax_i-qM,\quad i=1,2,\cdots,n \qquad (2.6)$$

其中 q 为正整数。上式两端用 M 除，则得

$$\xi_{i+1}=a\xi_i-q \qquad (2.7)$$

可见点 (ξ_i,ξ_{i+1}) 在直线

$$y=ax-q \qquad (2.8)$$

上，又因为 a 是 M 的原根，故 x_i 要取遍 $1,2,\cdots,M-1$ 这 M 个值；同时(2.6)式中的 q 也随之从0依次递增到某个值 λ。所以点 $(\xi_1,\xi_2),\cdots,(\xi_n,\xi_{n-1})$ 必分布在

$$\begin{aligned}y&=ax\\ y&=ax-1\\ &\cdots\cdots\\ y&=ax-\lambda\end{aligned}$$

通入+1条直线上。可以证明 $\lambda = a - 1$ 。于是推得：

定理2.2，由伪随机数 ξ_1，ξ_2，…，ξ_n 所组成的点序列 (ξ_1, ξ_2)，(ξ_2, ξ_3)，…，(ξ_n, ξ_{n+1})，在二维平面上必分布在由(2.9)式所表示的斜率为 a 的 a 条直线上。

定理2.3 由伪随机数 ξ_1，ξ_2，…，ξ_n 所组成的点序列 (ξ_1, ξ_2)，…，(ξ_n, ξ_{n+1})，也在

$$\begin{aligned} y &= -(M-a)x + 1 \\ y &= -(M-a)x + 2 \\ &\cdots\cdots\cdots \\ y &= -(M-a)x + (M-a) \end{aligned} \qquad (2.10)$$

所表示的斜率为 $-(M-a)$ 的 $M-a$ 条斜线上。

必须注意，上述二定理是从不同角度揭示同一点序列的分布性质。前者称为正斜率分布，后者称为负斜率分布。显然，直线族(2.9)、(2.10)在区域 $(0,1)\times(0,1)$ 内的交点，即为全部点序列。

定理2.4，对正斜率分布，任二相邻直线间的横向距离均为 $1/a$ 。且在同一条直线上分布有 $n_0 = M/a + \theta_0$ 个点，$|\theta_0| < 1$ 。

三 二维均匀偏度的计算

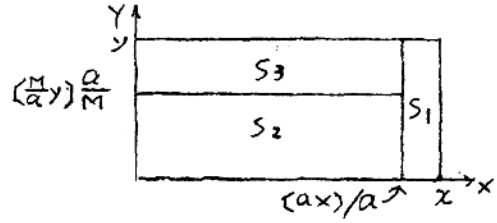
为计算方便，把点 $(0,0)$ 也计入 n 个点序列中，相应的均匀偏度可用下式近似代替

$$\Delta(M) = \sup_{0 \le x, y \le 1} \left| \frac{N(x,y)}{M} - xy \right| \qquad (3.1)$$

其中 $N(x,y)$ 表示在 $(0,x)\times(0,y)$ 区域内的已包括 $(0,0)$ 点的点数。

可以证明 $\Delta(M)$ 和(2.3)式所定义的 $\Delta(n)$ 之差小于等于 $1/M$，M 是相当大的数，故可用 $\Delta(M)$ 近似代替 $\Delta(n)$。

下面计算$\Delta(M)$，先把矩形区域：



$S:(0,x)\times(0,y)$分为三个区域：

$S_1:(0,(ax)/a)\times(0,[\frac{M}{a}y]\frac{a}{M})$

$S_2:((ax)/a,x)\times(0,y)$

$S_3:(0,(ax)/a)\times([\frac{M}{a}y]\frac{a}{M},y)$

图 3.1

其中$0\le x\le 1$，$0\le y\le 1$，如图3.1所示。显然，

$$S=S_1+S_2+S_3$$

在S区域内的点数

$$N(S)=N(x,y)=N(S_1)+N(S_2)+N(S_3)$$

再令 $\Delta_i=\frac{N(S_i)}{M}-|S_i|$，$i=1,2,3$.

其中$N(S_i)$表示在区域S_i中的点数，$|S_i|$表示区域S_i的测度。下面分别计算Δ_i。

1. Δ_1的计算

由定理2.4推知，在S_1内一条直线上分布有$[y/\frac{a}{M}]=[\frac{M}{a}y]$个点。因此，

$$N(S_1)=[ax]\,[\frac{M}{a}y]$$

又 $$|S_1|=\frac{(ax)}{a}[\frac{M}{a}y]\frac{a}{M}$$

所以 $$\Delta_1=\frac{N(S_1)}{M}-|S_1|=0 \qquad (3.3)$$

2. Δ_2的计算

由图3.1看到$|S_2|=\frac{\{ax\}}{a}y$，其中$\{x\}$表示取x的小数部分。再由点分布图可知

$$N(S_2)=\begin{cases}\frac{M}{a}\{ax\}+\theta_0' & \text{当 } y\ge\{ax\}\\ \frac{M}{a}y+\theta_0' & \text{当 } y<\{ax\}\end{cases}$$

其中$|\theta_0|<1$。于是

$$\Delta_2=\frac{N(S_2)}{M}-|S_2|$$

$$=\begin{cases}\frac{\{ax\}}{a}(1-y)+\frac{\theta_0'}{M} & \text{当 } y\geqslant\{ax\}\\ \frac{y}{a}(1-\{ax\})+\frac{\theta_0'}{M} & \text{当 } y<\{ax\}\end{cases}\qquad(3.4)$$

因为$y=\{ax\}=\frac{1}{2}$时Δ_2达到最大，所以

$$\Delta_2\leqslant\frac{1}{4a}+\frac{1}{M}\qquad(3.5)$$

Δ_3的计算比较复杂，下面将予专门讨论。

四 带和带内的点分布

1，纵带和横带

定义4.1 称区域$\left(\frac{j-1}{a_i},\frac{j}{a_i}\right)\times(0,1)$为第$i$级第$j$个纵带。其中

$$a_{i+1}=a_{i-1}-a_i\,(\mathrm{mod}\,a_i),\quad i=1,2,\cdots\qquad(4.1)$$

而$a_0=M$，$a_1=a_0$。

定义4.2，称区域$(0,1)\times\left(\sum_{l=0}^{i}\frac{k_{l-1}a_{l-1}}{M}+\frac{j-1}{M}a_i,\ \sum_{l=0}^{i}\frac{k_{l-1}a_{l-1}}{M}+\frac{j}{M}a_i\right)$为第$i$级第$j$个横带。这里规定$a_{-1}=k_{-1}=k_0=0$，而且$k_l$和$j$要满足不等式$k_la_{l-1}\leqslant a_{l-2}$，$ja_i\leqslant a_{i-1}$。

此外，当$i=0$时，j只取1，所以区域$(0,1)\times(0,1)$本身就是零级横带。还称$x=j/a_i$，$j=0,1,\cdots,a_i$和$y=\sum_{l=0}^{i}\frac{k_{l-1}a_{l-1}}{M}+\frac{j}{M}a_i$为带边界线。

类似于S_1，S_2和S_3的规定，来定义区域$S_j^{(i)}$：

$$S_j^{(i)}:\left(0,\frac{j_{i+1}}{a_{i+1}}\right)\times\left(\sum_{l=0}^{i+1}\frac{k_{l-1}a_{l-1}}{M},\sum_{l=0}^{i+2}\frac{k_{l-1}a_{l-1}}{M}\right)$$

$$S_2^{(i)}:\left[\frac{j_{i+1}}{a_{i+1}},\frac{j_i}{a_i}\right)\times\left(\sum_{l=0}^{i+1}\frac{K_{l-1}a_{l-1}}{M},\sum_{l=0}^{i+1}\frac{K_{l-1}a_{l-1}}{M}+\frac{a_i}{M}\theta_i\right)$$

$$S_3^{(i)}:\left[0,\frac{j_{i+1}}{a_{i+1}}\right)\times\left(\sum_{l=0}^{i+2}\frac{K_{l-1}a_{l-1}}{M},\sum_{l=0}^{i+2}\frac{K_{l-1}a_{l-1}}{M}+\frac{a_{i+1}}{M}\theta_{i+1}\right)$$

其中 $0<\theta_i,\theta_{i+1}<1$，$i=1,2,\cdots,K_{l-1},a_{-1},j_{i+1}$ 等均满足定义4·2的规定。

由上述定义和点的分布图看到：

（1）同级横带中所包括的点数均相等；

（2）任二同级横带中的 $S_3^{(i)}$ 区，只要该二 $S_3^{(i)}$ 区全等，则它们所包括的点数也相等。

定理4.1. 任何 i 级横带中 $S_3^{(i)}$ 区对偏度的贡献 $\Delta_3^{(i)}$，与它是 i 级中的第几个横带无关，只决定于 $S_3^{(i)}$ 的形状和大小，即都等于

$$\Delta_3^{(i)}=\frac{N(S_3^{(i)})}{M}-|S_3^{(i)}| \qquad (4.3)$$

由该定理可知，只要对比包括原点的区域 $S_3^{(i)}:\left[0,\frac{j_{i-1}}{a_{i-1}}\right)\times\left[0,\frac{a_{i+1}}{M}\theta_{i+1}\right)$ 也就够了。这将使计算简化。

2 包括原点的横带中点的分布

（1）一级横带的点分布

定理4.2，在任何一个一级横带中，均有 $a_1=a$ 个点分布在 α_2 条斜率为 K_1 的斜线上，而且一条斜线上等距离地分布有 $a_1/a_2+\theta_1$ 个点。

这里 $K_1=\frac{a_2}{M}\left(\frac{1}{a_1}+\eta_1\right)$。

（2）求 i 级横带 $[0,1)\times[0,a_i/M)$ 的点分布。

定理4.3 在任何一个 i 级横带中都有 a_i 个点，分布在 a_{i+1} 条斜率为 K_i 的斜线上，而且一条斜线上等距离分布有 $a_i/a_{i+1}+\theta_i$ 个点。

推论：在 i 级横带中的 a_i 个点，分布在 $a'_{i-1}=a_i-a_{i+1}$ 条斜率为 $-K'_i$ 的斜线上，且一条斜线上等距分布有 $a_i/a_{i+1}+\theta_i$

个点。

(三) 均匀偏度的计林

把区域$(0,1)\times(0,1)$分成子区，直到$a_{i+1}=1$为止。则所求偏度应等于各级子区域$S_1^{(i)}$的贡献之和。

(1). $S_1^{(i)}$对偏度的贡献。

计林可得

$$\Delta_1^{(i)}=\frac{1}{M}N(S_1^{(i)})-|S_1^{(i)}|=0,\quad i=1,2\cdots \tag{4.13}$$

可见任何级的$S_1^{(i)}$区域对偏度都没贡献。

(11). $S_2^{(i)}$对偏度的贡献

应用分布图可以林得

$$\Delta_2^{(i)}=\frac{a_i}{Ma_{i+1}}\theta_i(1-\theta_i)-\frac{1}{M}(e_i-a_i\theta_i t_i)+\frac{\theta_i'}{M} \tag{4.13}$$

其中$0<\theta_i<1$，$t_i\leqslant\frac{1}{a_i}$，或$t_i=\frac{1}{a_i}\theta_{ti}$，$0\leqslant\theta_{ti}\leqslant 1$。

证明可知$|e_i-a_i\theta_i t|\leqslant 1$，于是

$$\Delta_2^{(i)}=\frac{a_i}{Ma_{i+1}}\theta_i(1-\theta_i)+O(\frac{1}{M}) \tag{4.14}$$

由此得到S_3区对偏度的贡献为

$$\Delta_3=\sum_{i=1}^{S}\Delta_2^{(i)}=\sum_{i=1}^{S}\frac{a_i}{Ma_{i+1}}\theta_i(1-\theta_i)+O(\frac{S}{M}) \tag{4.15}$$

其中S满足$a_{S+1}=1$。考虑(3.4)式，则得

$$\frac{1}{M}N(S)-|S|=\frac{1}{M}N(x,y)-xy=\Delta_2+\Delta_3 \tag{4.16}$$

为求该式上确界，可知当

$$x=\{ax\}=\theta_1=\theta_2=\cdots=\theta_S=\frac{1}{2} \tag{4.17}$$

$|\Delta_2+\Delta_3|$可达最大。总可以找到一个η_0，$|\eta_0|<1$，在$x_0=\frac{1}{2}+\frac{a_1}{M}\eta_0=\frac{1}{2}+\frac{a_1}{a_0}\eta_0$时使得$\theta_1=\frac{1}{2}$。依此使各级都有

$$\theta_i = \frac{1}{2} + \frac{a_{i+1}}{a_i}\eta_i, \quad i = 0, 1, 2, \cdots, S-1$$

$$\theta_S = \frac{1}{2} \tag{4.18}$$

代入(4.16)式，立刻得到

$$\tilde{\Delta} = \Delta_2 + \Delta_3 = \sum_{i=0}^{S} \frac{a_i}{4Ma_{i+1}} - \sum_{i=0}^{S} \frac{a_{i+1}}{Ma_i}\eta_i^2 + O\left(\frac{S+1}{M}\right)$$

即 $$\Delta(M) = \frac{1}{4M}\sum_{i=0}^{S}\frac{a_i}{a_{i+1}} + O\left(\frac{S+1}{M}\right) \tag{4.20}$$

五 $\Delta(M)$收敛性的改进

(4.20)式当$a_i/a_{i+1} < 2$时，求和项就会很多，特别是当此值接近一时，误差会大到难以接受的程度。为克服这困难引入如下概念。

定义5.1. 对于包括原点的i级横带a_i/M，$i=0,1,2,\cdots,S$，连同定上面的带边界线在内，以直线$y = a_i/2M$为轴，旋转180°(对零级带原点上的序列点不动)，则新分布称为原分布的准对称分布。

定理5.1. 包括原点的i级横带内，任何半宽区域$S^{(i)}: [0,x)\times(0, a_i/2M)$，其成分布和准对称分布对偏度的贡献，在误差为$O(\frac{1}{M})$的范围内，大小相等，符号相反。

2. 对(4.20)式的改进

注意到(4.20)式相当于在各级横带内仍取半宽区域得到的。故根据定理5.1，可用准对称分布代替原来的计林。只要把(4.20)式中a_{i+1}换为$a'_{i+1} = a_i - a_{i+1}$，然后再加一符号就可用准对称分布代替成分布。当然a_{i+2}的计林也要变成$a_{i+2} = a'_{i+1} - a_i \pmod{a'_{i+1}}$再加一负号，直到某一整数i使得$a'_{i+i_1} = 1$或$a'_{i+i_1-1}/a'_{i+i_1} < 2$为止。对后一情况再用准对称分布代替，重复前面计林。于是(4.20)式可写为

$$\Delta' = \frac{1}{4M}\sum_{\sigma=0}^{V}(-1)^{\sigma}\sum_{i=i_\sigma+1}^{i_{\sigma+1}}\frac{a_i}{a_{i+1}} \tag{5.3}$$

其中 $i_0=-1$，$i_{v+1}=s$，σ 表示经过准对称分布代替[illegible]而且凡是 $a_i/a_{i+1}<2$ 的 a_{i+1} 均换为 a_i-a_{i+1}，[illegible]

$\frac{a_i}{a_i-a_{i+1}}=\frac{1}{1-a_{i+1}/a_i}<2$，从而解决了收敛问题。

再令 $$A_\sigma=(-1)^\sigma\sum_{i=i_\sigma+1}^{i_{\sigma+1}}\frac{a_i}{a_{i+1}} \qquad (5.4)$$

则有 $$\Delta'=\frac{1}{4M}\sum_{\sigma=0}^{v}A_\sigma \qquad (5.5)$$

为使此时 Δ' 仍有最大可能值，可令

$$A^{(+)}=\frac{1}{4M}\sum_{\sigma=0}^{\sigma_1}A_\sigma,\ \sigma_1=v \qquad (5.6)$$

表示 σ_1 从零到 v 变化时，其和为正值当中最大者。

$$A^{(-)}=\frac{1}{4M}\sum_{\sigma=0}^{\sigma_2}A_\sigma-\frac{1}{4a_1},\ \sigma_2\leq v \qquad (5.7)$$

表示其和为负值当中，绝对值最大者。显然

$$\Delta'\leq\max\ (A^{(+)},A^{(-)}) \qquad (5.8)$$

此即改进后的均匀偏度一般表达式。

计算表明其误差项满足

$$O\left(\frac{S+1}{M}\right)\leq O\left(\frac{1+\log_2 a_1}{M}\right) \qquad (5.9)$$

以上结果比U.Dieter 的结果要简单得多，精度也相当高。依此就整体而言为判别伪随机数的优劣提供了根据，也必将改进众多与伪随机数相关的结果。

参 考 文 献

(1) D.E.KNUTH, The Art of Computer programming vol2: seminumerical Algorithms. Addison—wesley—Reading. mass, 1969

(2) U.Dieter, J.Ahrens, Numer. Math, 16.101 (1970)

(3) U.Dieter, Math, comp, 25 835 (1971)