计算物理——针对 LCG 的一点思考

白博臣

(四川大学　物理学拔尖计划)

摘　要

本文探讨了线性同余发生器（LCG）的特性及其在随机数生成中的应用。LCG 是一种利用求余运算的简单随机数生成器，其递推公式为 $x_n = (ax_{n-1} + c) \pmod{M}$，其中 M 是模数，a 是乘数，c 是增量，$x_0$ 是初始种子。LCG 产生的序列 $\{x_n\}$ 为非负整数，通过将 $\{x_n\}$ 除以 M 得到的 $\{R_n\}$ 可作为均匀随机数序列。

在对 LCG 进行分析的过程中，作者发现了两个问题：首先，对于参数为 $m = 2^{31} - 1, a = 4, c = 1$ 的 LCG，初始种子值对最终收敛值有显著影响；其次，对于参数为 $m = 27, a = 26, c = 5$ 的 LCG，不同的初始种子值会导致收敛值稳定在两个不同的值。这一发现挑战了随机数生成器应具有的期望值与种子值无关的特性。

为了解释这些现象，作者从数学角度对 LCG 的特性进行了深入分析，并提出了通项公式 $x_{n+1} = a^n(x_0 + \frac{c}{a-1}) - \frac{c}{a-1}$，以期找到问题的解答。

关键词：线性同余发生器，随机数生成，收敛性，数学分析

1 LCG 简介

LCG(Linear congruential generator) 即线性同余发生器，是利用求余运算的随机数发生器。其递推公式为：

$$
\begin{aligned}
x_n &= (ax_{n-1} + c) \pmod{M}, \quad n = 1, 2, \ldots \\
&m\text{为模数；} 0 < m \\
&a\text{为乘数；} 0 \le a < m \\
&c\text{为增量；} 0 \le c < m \\
x_0&\text{为初始种子；} 0 \le x_0 < m
\end{aligned}
\tag{1}
$$

得到的序列 x_n 为非负整数，$0 \le x_n \le M$。最后令 $R_n = x_n/M$，则 $R_n \in [0, 1)$，把 R_n 作为均匀随机数序列。该算法的基本思想是因为很大的整数前面的位数是重要的有效位数而后面若干位有一定随机性。因为线性同余法的递推算法仅依赖于前一项，序列元素取值只有 M 个可能取值，所以产生的序列 $x_0, x_1, x_2, \ldots$ 一定会重复。若存在正整数 n 和 m 使得 $x_n = x_m (m < n)$，则必有 $x_{n+k} = x_{m+k}, k = 0, 1, 2, \ldots$ 即 $x_n, x_{n+1}, x_{n+2}, \ldots$ 重复了 $x_m, x_{m+1}, x_{m+2}, \ldots$，称这样的 $n - m$ 的最小值 T 为此随机数发生器在初值 x_0 下的周期，易得，$T \le M$.

2 问题发现

本课程 EX12 要求复现 PPT 中的某个图像 (Figure 1):

在复现过程中，发现两个奇怪的现象：

1. 针对图像中参数为 $m = 2^{31} - 1, a = 4, c = 1$ 的折线，我们发现初始值 (种子值) 对折线最后的收敛值有影响。

2. 针对图像中参数为 $m = 27, a = 26, c = 5$ 的折线，我们发现其收敛值根据种子值的不同最后稳定在两个值。

现对两个问题进行进一步阐述。

2.1 问题一阐述

当选取种子值 $x_0 = 1$ 时，我们可以得到 Figure 2，而当我们更改种子值为 $x_0 = 12345678$ 时，得到的图像为 Figure 3.

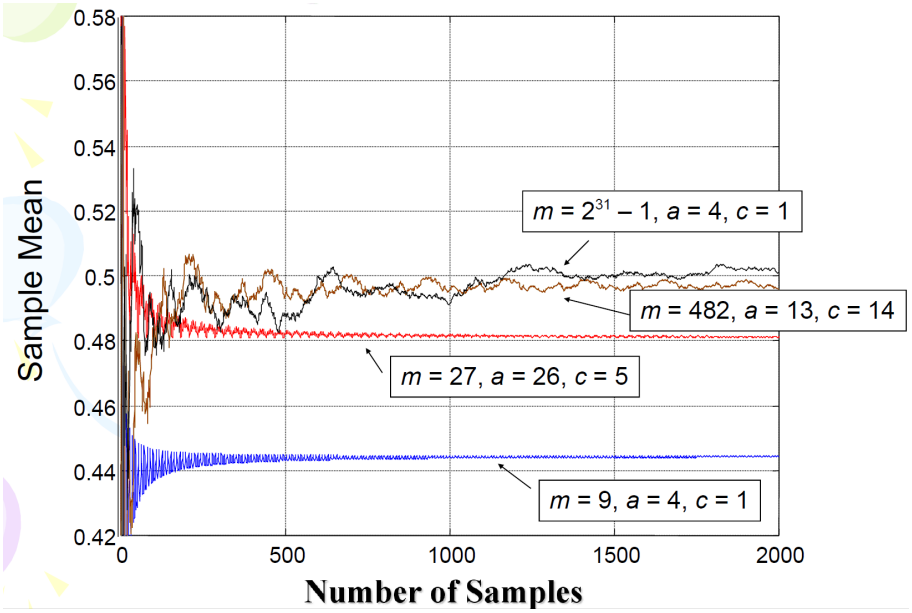

Figure 1: EX12 要求复现四组参数值下 LCG 的期望收敛性

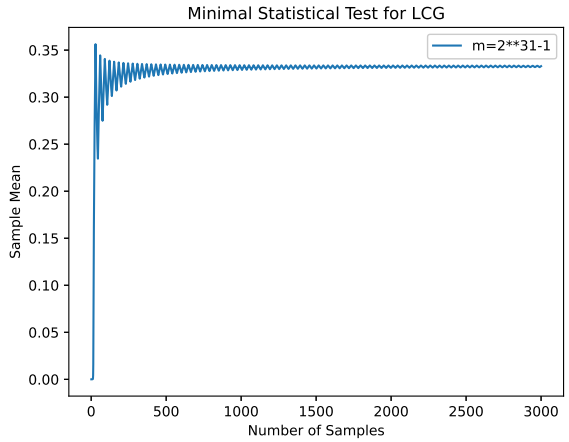

Figure 2: 种子值取 1 时，最后并没有收敛到 0.5，而是在 0.3∼ 0.35 之间

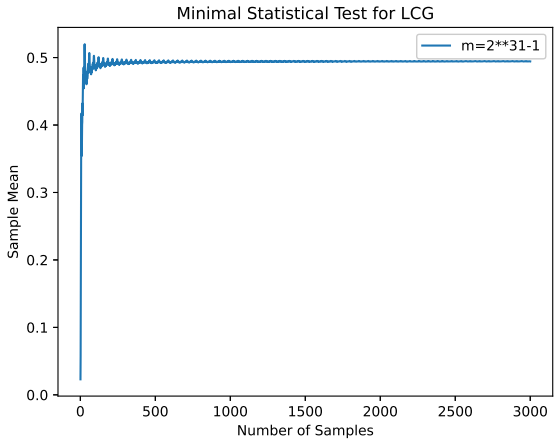

Figure 3: 种子值取 12345678 时，最后收敛到 0.5 左右，“成功”复现图像

通过调整种子值的大小我们可以控制最终随机数的期望，但是按照随机数的要求，我们不应该令随机数的期望与种子值有关 (至少应维持在 0.5 左右)。所以我对该组参数的选取持质疑态度，在后面问题解答中我将提出我的看法。

2.2 问题二阐述

当参数为 $m = 27, a = 26, c = 5$ 时，我们先选取种子值为 $x_0 = 4$ 得到 Figure 4，再选取种子值为 $x_0 = 6$ 得到 Figure 5.

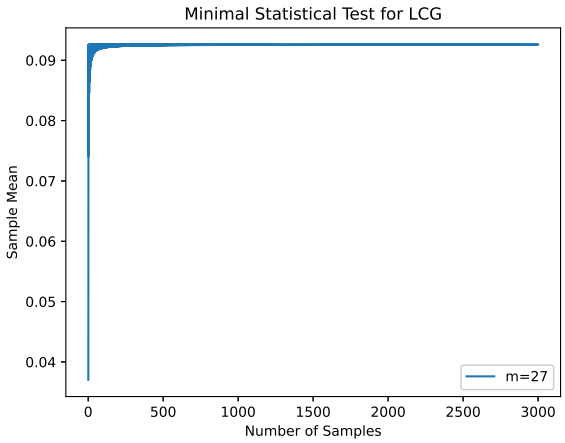

Figure 4: 种子值取 4 时，期望收敛到 0.1 左右

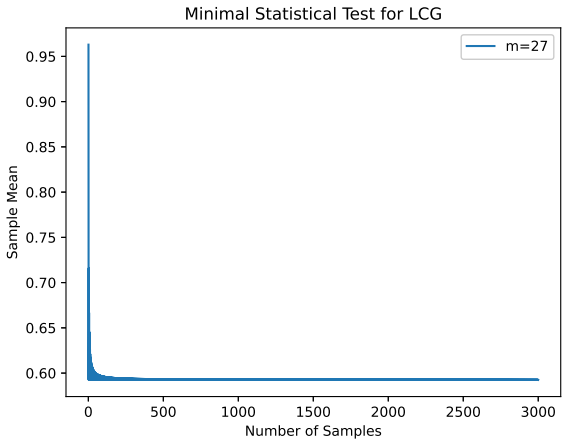

Figure 5: 种子值取 6 时，期望收敛到 0.6 左右

后续我们选取不同种子值，发现在 $x_0 \leq 5$ 时，最终期望均收敛到 0.1 左右，而对于 $x_0 > 5$ 时，最终期望值均收敛到 0.6 左右，不管如何改变种子值的选取，我们都发现该组参数下的期望值始终不会趋于 0.5 左右，无法复现题目中要求的图像。

3 问题解答

针对上述两个问题，我不得不从数学角度思考线性同余本身的一些特性，希望能够从中得出问题的解答。

3.1 数学求解

首先，我们可以先不考虑递推关系后的取余运算，而是直接先根据递推关系求得通项公式再取余运算，证明如下：

设 $x_n = x' + d * m$，其中 x' 是 x_n 模 m 后的余数，d 是整数，根据递推关系，

$$x_n = (ax_{n-1} + c)(modM)$$

代入后得：

$$x_n = (a(x'_{n-1} + d * m))(modM)$$

在取余运算中 $a * d * m$ 项被约去，所以先代入通项公式再取模与先取模再代入递推关系得到的结果是一样的。

所以线性同余递推关系可改写为：

$$x_{n+1} = a^n(x_0 + \frac{c}{a-1}) - \frac{c}{a-1} \tag{2}$$

不难判断：

1) 如果某 LCG 产生的随机序列的周期 T 小于 m，则选取不同的初始值 x_0 产生的 LCS(Linear Congruential Scquence; 线性同余序列) 可能有不同的周期。

2) 如果其周期 T=m，则即使选取不同的 x_0，产生的这些 LCS 具有相同的周期且必定为 T。

下列定理给出了混合同余发生器达到满周期的一个充分条件：

定理 3.1 当下列三个条件都满足时，混合同余发生器可以达到满周期：

(1)c 与 M 互素；

(2) 对 M 的任一个素因子 P，a-1 被 P 整除

(3) 如果 4 是 M 的因子，则 a-1 被 4 整除。

我们常取 $M = 2^L$，L 是计算机中整数的尾数字长。按照上述定理的建议可取 $a = 4\alpha + 1, c = 2\beta + 1$，$\alpha, \beta$ 为任意正整数。x_0 为任意非负整数，这样的 LCG 是满周期的，周期为 2^L.

好的均匀分布随机数发生器应该周期足够长，统计性质符合均匀分布，序列之间独立性好。把同余法生成的数列看成随机变量序列 $\{x_n\}$，在满周期时，可认为 x_n 是从 $0 \sim M-1$ 中随机等可能选取的，即

$$P\{x_n = i\} = 1/M, i = 0, 1, \ldots, M-1$$

此时

$$E(x_n) = \sum_{i=0}^{M-1} i \cdot \frac{1}{M} = (M-1)/2$$

$$S^2(x_n) = E\left(x_n^2\right) - (E(x_n))^2 = \sum_{i=0}^{M-1} i^2 \frac{1}{M} - \frac{(M-1)^2}{4} = \frac{1}{12}\left(M^2 - 1\right)$$

于是当 M 很大时,

$$E(x_n) = \frac{1}{2} - \frac{1}{2M} \approx \frac{1}{2}$$

$$S^2(x_n) = \frac{1}{12} - \frac{1}{12M^2} \approx \frac{1}{12}$$

可见 M 充分大时从一、二阶矩看生成的数列很接近均匀分布。

随机数序列还需要有很好的随机性。数列的排列不应该有规律，序列中的两项不应该有相关性。

因为序列由确定性公式生成，所以不可能真正独立。至少我们要求是序列自相关性弱。对于满周期的混合同余发生器可以得序列中前后两项自相关系数的近似公式

$$\rho(1) \approx \frac{1}{a} - \frac{6c}{aM}\left(1 - \frac{c}{M}\right)$$

所以应该选取较大 a 值。

3.2 问题求解

回到问题中来，问题一、二之所以与种子值相关，均是因为其并没有达到满周期，由于糟糕的 a 的取值，导致 LCS 的周期较小 (比方说问题二的 T 只有 2)，最后的序列期望与均匀分布的期望相差甚远。

我们可以更改问题一中的取值，使用 C++11 中的随机数参数 $a = 16807$，种子数随机选取，得到的期望图像见 Figure 6.

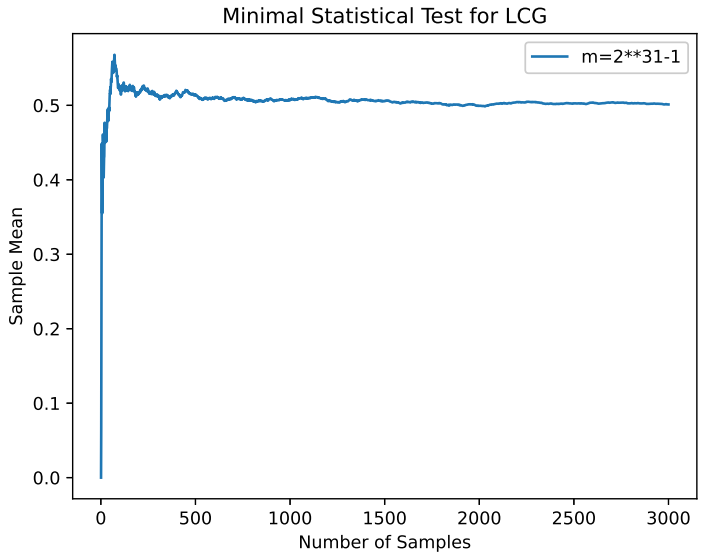

Figure 6: 参数 $a = 16087, x_0 = 1$

改变参数 a 后其期望趋近于 0.5 左右，与均匀分布的期望接近。

4 总结

简单总结几点即是：

1) 模数 m 应该尽可能大，通常至少大于 2^{30}，为了计算效率，通常会结合计算机的字长考虑选取 m 的值。

2) 如果 m 选取为 2 的幂，也即 $m = 2^l$，则选取的 a 通常应该满足 a 模 8 等于 5。

3) 当参数 m 和 a 的选定比较合理时，对于 c 的选择约束性不是很强烈，但要保证 c 与 m 互质。例如 c 可以选择 1 或者 11。

4) 种子 seed 应该是随机选取的，可以将时间戳作为种子。