# Fibonacci congruences and applications

René Blacher

# Fibonacci congruences and applications

René BLACHER

Laboratory LJK
Université Joseph Fourier
Grenoble
France

**Summary :** When we consider a congruence $T(x) \equiv ax$ modulo m as a pseudo random number generator, there are several means of ensuring the independence of two successive numbers. In this report, we show that this dependence depends on the continued fraction expansion of m/a. We deduce that the congruences such that m and a are two successive elements of Fibonacci sequences are those having the weakest dependence. We will use this result to obtain truly random number sequences $x_n$. For that purpose, we will use non-deterministic sequences $y_n$ such that the conditional probabilities have Lipschitz coefficients not too large. They are transformed using Fibonacci congruences and we will get by this way sequences $x_n$. These sequences $x_n$ admit the IID model for correct model.

**Key Words :** Fibonacci sequence, Random numbers, Congruence, Dependence, Correct models.

**Résumé :** Quand on considère une congruence $T(x) \equiv ax$ modulo m comme générateur de nombres pseudo-aléatoires, il y a plusieurs moyens de s'assurer de l'indépendance de deux nombres successifs. Dans ce rapport, nous montrons que cette dépendance dépend du développement en fraction continue de m/a. On en déduit que les congruences telles que m et a sont deux éléments successifs de la suite de Fibonacci sont celles ayant la dépendance la plus faible. Nous utiliserons ce résultat pour obtenir des suites de nombres réellement aléatoires $x_n$. Pour cela, nous utiliserons des suites non-déterministes $y_n$ telles que leurs probabilités conditionnelles aient un coefficient de Lipschitz pas trop grand. On les transformera en utilisant les congruences de Fibonacci et on obtiendra ainsi les suites $x_n$. Ces suites $x_n$ admettent le modèle IID pour modèle correct.

**Mots-clefs :** Suite de Fibonacci, Nombres aléatoires, Modelès corrects, Congruence

# 1 Introduction

## 1.1 Fibonacci Congruence

One needs a transformation which transforms the elements of $\{0, 1, ...., m-1\}$ in independent numbers. Simplest is to use a congruence $T(x) \equiv ax$ mod(m). Indeed, the dependence induced by $\left(T^n(x_0), ....., T^{n+p}(x_0)\right)$ is easy to know. In particular, one can use the spectral test or the results of Dieter which allow to choose the best "a" and "m" (cf [16], [1]) .

Then, in this report, we study the set $E_2 = \left\{\ell, \overline{T(\ell)} \mid \ell \in \{0, 1, ....., m-1\}\right\}$ when $\overline{z} \equiv z$ modulo m and $0 \leq \overline{z} < m$ if $z \in \mathbb{Z}$. We will understand that this dependence depends on the continued fraction $\frac{m}{a}$, i.e. it depends on sequences $r_n$ and $h_n$ defined in the following way.

**Notations 1.1** *Let $r_0 = m$, $r_1 = a$. One denotes by $r_n$ the sequence defined by $r_n = h_{n+1}r_{n+1} + r_{n+2}$ the Euclidean division of $r_n$ by $r_{n+1}$ when $r_{n+1} \neq 0$. Moreover, one denotes by d the smallest integer such as $r_{d+1} = 0$. One sets $r_{d+2} = 0$.*

*One sets $k_0 = 0$, $k_1 = 1$ and $k_{n+2} = h_{n+1}k_{n+1} + k_n$ if $n + 1 \leq d$.*

Then, dependence depends on the $h_i$'s : more they are small, more the dependence is weak.

**Theorem 1** *Let $(x_0, y_0) \in E_2$. Let $R^0 = \left\{[x_0, x_0 + k_n] \otimes [y_0, y_0 + r_{n-2}[\right\}$ and let $R_0 = \overline{R^0}$, be the rectangle $R^0$ modulo m. Then*

*If n is even, $E_2 \cap R_0 = \left\{(\overline{x_0 + k_{n-1}\ell}\ ,\ \overline{y_0 + r_{n-1}\ell}) \middle| \ell = 0, 1, 2, ...., h_{n-1}\right\}$. Moreover the points $(\overline{x_0 + k_{n-1}\ell}\ ,\ \overline{y_0 + r_{n-1}\ell})$ are lined up modulo m .*

*If n is odd,*
*$E_2 \cap R_0 = \left\{(\overline{x_0 + k_{n-2} + k_{n-1}\ell}\ ,\ \overline{y_0 + r_{n-2} - r_{n-1}\ell}) \middle| \ell = 0, 1, 2, ...., h_{n-1}\right\}$. Moreover, the points $(\overline{x_0 + k_{n-2} + k_{n-1}\ell}\ ,\ \overline{y_0 + r_{n-2} - r_{n-1}\ell})$ are lined up modulo m.*

Of course, in general, it is only on the border that $R_0$, the rectangle modulo m, satisfies $R_0 \neq R^0$. If not, $R_0$ is a normal rectangle.

For example if $x_0 = y_0 = 0$, this theorem means that the rectangle $[0, k_n/2] \otimes [r_{n-2}/2, r_{n-2}[$ does not contain points of $E_2$ if n is even : $E_2 \cap \left\{[0, k_n/2] \otimes [r_{n-2}/2, r_{n-2}[\right\} = \emptyset$ . If $h_{n-1}$ is large, that will mean that an important rectangle of $\mathbb{R}^2$ is empty of points of $E_2$: that will mark a breakdown of independence.

As $h_i \geq 1$, the congruence which defines the best independence of $E_2$ will satisfy $h_i = 1$ and $h_d = 2$. In this case **we call it congruence of Fibonacci**. Indeed, there exists $n_0$ such that $a = fi_{n_0}$ and $m = fi_{n_0+1}$ where $fi_n$ is the sequence of Fibonnacci : $fi_1 = fi_2 = 1$, $fi_{n+2} = fi_{n+1} + fi_n$. Moreover sequences $h_n$ and $k_n$ are the sequence of Fibonacci except for the last terms.

## 1.2 Application : building of random sequence

To have IID random number [1] two methods exists : use of pseudo-random generators (for example the linear congruence) and use of random noise (for example Rap music). But, up to now *no completely reliable solution had been proposed* (cf [4]). To set straight this situation, Marsaglia has created a Cd-Rom of random numbers by using sequences of numbers provided by Rap music (cf [5], page 3 of [1]). But, it does not have proved that the sequence obtained is really random.

Congruences of Fibonacci cannot be used in order to directly generate good pseudo random sequences because $T^2 = \pm Id$ where $Id$ is the identity (cf page 141 of [9]). However, by using Fibonacci congruences, there exists simple means of obtaining random sequences whose the quality is sure : one uses the same method as Marsaglia, but one transforms the obtained sequence by using functions $T_q$ defined by the following way.

**Definition 1.2** *Let $q \in \mathbb{N}^*$. Let $T$ be the congruence of Fibonacci modulo m. We define the function of Fibonacci $T_q$ by $T_q = Pr_q \circ \widehat{T}$ where*
*1) $\widehat{T}(x) = \overline{T}(mx)/m$,*
*2) $Pr_q(z) = \overline{\overline{0, b_1 b_2 .... b_q}}$ when $z = \overline{\overline{0, b_1 b_2 ...}}$ is the binary writing of z.*

In order to build IID sequences, we will need to have a truly random sequence $\tilde{y}_n = my_n \in \{0, 1, ...., m-1\}$, n=1,2,....,N, admitting for model a sequence of random variables $\tilde{Y}_n$ defined on a probability space $(\Omega, \mathcal{A}, P)$. We will need that $Y_n = \tilde{Y}_n/m$ satisfies the following condition : the conditional probabilities of $Y_n$ admit densities with Lipschitz coefficient bounded by $K_0$ not too large.

In fact, since $Y_n$ is with discrete value, we can always assume that $Y_n$ has a continuous density with respect to $\mu_m$, where $\mu_m$ is the uniform measure : $\mu_m(k/m) = 1/m$ for all k such that

[1] By abuse of language, we will call "IID sequence" (Independent Identically Distributed) the sequences of random numbers.

$k \in \{0, 1, ...., m-1\}$. One can always admit that this density has a Lipschitz coefficient bounded by $K_0 > 0$. The assumption which we make about $Y_n$ is that $K_0$ is not too large.

We will see in Section 8 that there is such sequences $y_n$.

So, in Proposition 6.1, we shall prove easily that the conditional probabilities of $T_q(Y_n)$ check

$$P\big\{T_q(Y_n) = x_0 \,\big|\, Y_{n-1} = y'_1, Y_{n-2} = y'_2, ...., Y_{n+1} = y"_1, Y_{n+2} = y"_2, .....\big\} = (1/m)[1+O(1)K_0 2^q/m] \,.$$

By setting $X_n = T_q(Y_n)$ and $x_n = T_q(y_n)$, we shall deduce that

$$P\big\{X_n = x_0 \,\big|\, X_{n-1} = x'_1, X_{n-2} = x'_2, ...., X_{n+1} = x"_1, X_{n+2} = x"_2, .....\big\} = (1/m)[1+O(1)K_0 2^q/m] \,,$$

and that, for all Borel set $Bo \subset \{0/2^q, 1/2^q, ...., (2^q-1)/2^q\}^N$,

$$P\big\{(X_1, ...., X_N) \in Bo\big\} = L(Bo)[1 + O(1)\epsilon] \,,$$

where $|\epsilon| \leq K_0 N 2^q/m$ and where L is the measure corresponding to the Borel measure in the case of discrete space. We prove this result in Section 6.

We shall choose m and q such that $\epsilon$ is small enough. Indeed, if $\epsilon$ is small enough with respect to N, the size of sample, then $x_n = T_q(y_n)$ cannot be differentiated from an IID sequence.

Indeed, it is wellknown that, for a sample $x_n$, there is many correct models : in particular, if $x_n$ is a sample of an IID sequence of random variables, models such that $P\{(X_1, ...., X_N) \in Bo\} = L(Bo)[1 + \epsilon_{Bo}]$, $|\epsilon_{Bo}| \leq \epsilon$, are correct if $\epsilon$ is small enough. Reciprocally, if the sequence of random variables $X_n$ checks $P\{(X_1, ...., X_N) \in Bo\} = L(Bo)[1 + \epsilon_{Bo}]$, the model IID is also a correct model for the sequence $x_n$.

Thus one will be able to admit that IID model is a correct model for the sequences $x_n$. As a matter of fact, one will be even able to admit that **there exists another correct model $Y_n^{\theta_0}$ of $y_n$ such that $T_q(Y_n^{\theta_0})$ is exactly the IID sequence**. We shall prove this result in proposition 5.1.

In fact, we must know what is called a correct model. We will discuss this problem in Section 4.

So finally we can build sequences $x_n$ admitting for correct model the IID model. This means that, a priori, these sequences $x_n$ behave as random sequences. It is always possible that they do not satisfy certain tests. But it will be a very weak probability as we know that it is the case for samples of sequences of IID random variables.

By using this technique, we have created such real sequences $x_n$ by using various texts. We have tested the sequence $x_n$ with classical Diehard tests, and higher order correlation coefficients. Results are in accordance with what we waited : the hypothesis "randomness" is accepted by all these tests (cf section 9.2) . One can obtain such real random sequences in [13].

By this method, we therefore have a means to develop the technique used by Marsaglia to create a CD-ROM. **We can indeed prove mathematically that the sequence obtained can be regarded a priori as random**, what Marsaglia did not.

# 2 Dependence induced by linear congruences

In this section, we study the set $E_2 = \big\{\ell, \overline{T(\ell)} \mid \ell \in \{0, 1, ....., m-1\}\big\}$.

## 2.1 Notations

We recall that we define sequences $r_n$ and $h_n$ by the following way : we set $r_0 = m$, $r_1 = a$ and we define $r_n$ by $r_n = h_{n+1} r_{n+1} + r_{n+2}$, the Euclidean division of $r_n$ by $r_{n+1}$ when $r_{n+1} \neq 0$. One denotes by d the smallest integer such as $r_{d+1} = 0$. One sets $r_{d+2} = 0$.

One sets $k_0 = 0$, $k_1 = 1$ and $k_{n+2} = h_{n+1}k_{n+1} + k_n$ if $n + 1 \leq d$.

Then we have the writing of m/a in continued fraction :

$$\frac{m}{a} = h_1 + \frac{1}{h_2 + \frac{1}{h_3 + \frac{1}{h_4 + \ldots}}} .$$

Now, $h_n \geq 1$ for all n=1,2,...,d and $r_{d-1} = h_d r_d + r_{d+1} = h_d r_d + 0 = h_d r_d$. The full sequence $r_n$ is thus the sequence $r_0 = m$, $r_1 = a$, ....., $r_{d+1} = 0$, $r_{d+2} = 0$. Then, if T is a Fibonacci conguence, $r_n$ is the Fibonacci sequence $fi_n$, except for the last terms.

Remark that if $h_n = 1$ for n=1,2,...,d-1, $k_n$ is also the Fibonacci sequence for n=1,2,...,d. Indeed by definition, $k_0 = 0$, $k_1 = 1$, $k_2 = 1$ and $k_{n+2} = k_{n+1} + k_n$ if $n + 1 \leq d$.

## 2.2 Theorems

Now, in order to prove the theorem 1, it is enough to prove the following theorem.

**Theorem 2** *Let $n \in \{2, 3...., d\}$. Then*
*If n is even , $E_2 \cap \big\{[0, k_n[ \otimes [0, r_{n-2}[\big\} = \big\{(k_{n-1}\ell \ , \ r_{n-1}\ell) \big| \ell = 0, 1, 2, ...., h_{n-1}\big\}$. Moreover the points $(k_{n-1}\ell \ , \ r_{n-1}\ell)$ are lined up.*
*If n is odd,*
*$E_2 \cap \big\{]0, k_n] \otimes ]0, r_{n-2}]\big\} = \big\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell) \big| \ell = 0, 1, 2, ...., h_{n-1}\big\}$. Moreover, the points $(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)$ are lined up.*

Then, if there exists $h_i$ large, there is a breakdown of independence. For example suppose n=2. Then, one has a wellknown result. Indeed, $m = r_0$, $r_1 = a$, $k_1 = 1$ and $k_2 = h_1 = \lfloor m/a \rfloor$ where $\lfloor x \rfloor$ means the integer part of x. Thus, the rectangle $Rect_2 = [0, m/(2a)] \otimes [m/2, m[$ will not contain any point of $E_2$. However, this rectangle has its surface equal to $m^2/(4a)$. Thus if "a" is not sufficiently large, i.e if $h_1$ is too large, there is breakdown of independence.

### 2.2.1 Numerical examples

We confirm by graphs the previous conclusion. We suppose m=21. If a = 13, we have a Fibonacci congruence : cf figure 1. If one chooses a=10, $sup(h_i) = 20$ : cf figure 2 . If one chooses a=5, $sup(h_i) = 5$ : cf figure 3.

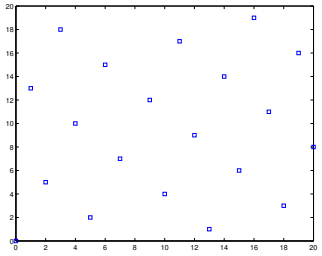

Figure 1: Fibonacci congruence

### 2.2.2 Conclusion

To avoid any dependence, it is necessary that $sup(h_i)$ is small. In the case of the Fibonacci congruence, independence is checked on all rectangles $R_{ect}$.

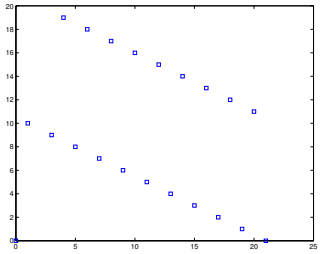

Figure 2: $sup(h_i) = 20$

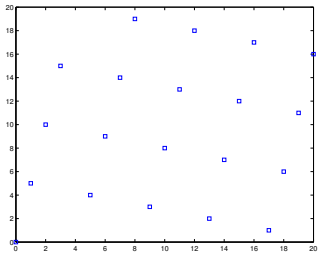

Figure 3: $sup(h_i) = 5$ Fig

### 2.3 Distribution of T([c,c'[) when T is a Fibonacci congruence

We assume that T is a Fibonacci congruence. Let $I = [c, c'[\cap\{0, 1, ...., m-1\}$ where $c, c' \in \{0, 1, ...., m-1\}$. We are interested by $\overline{T}^{-1}(I)$ or $\overline{T}(I)$ because $T^2 = \pm Id$. Since $\overline{T}(I)$ behaves as independent of I, normally, we should find that $\overline{T}(I)$ and, therefore $\overline{T}^{-1}(I)$, is well distributed in $\{0, 1, ...., m-1\}$. As a matter of fact it is indeed the case.

Indeed, let $k^n$, n=1,2,..,c'-c, be a permutation of $\{c, c+1, ..., c'-1\}$ such that $\overline{T}^{-1}(k^1) < \overline{T}^{-1}(k^2) < \overline{T}^{-1}(k^3) < ..... < \overline{T}^{-1}(k^{c'-c})$. Then, for all numerical simulations which we executed, one has always obtained, for r=0,1,2,....,c'-c-1,

$$|\overline{T}^{-1}(k^r)/m - r/N(I)| \leq \varphi(m)/N(I)$$

where $\varphi(m) << Log(m)$. In fact, it seems $\varphi(m)$ is the order of Log(Log(m)). Moreover, $Max_{r=0,1,....,N(I)-1}\left(\left|N(I)T^{-1}(k^r)/m - r\right|\right)$ seems maximum when I is large enough : $c' - c > m/2$.

For example, in figures 4, 5 and 6, we have the graphs $N(I)T^{-1}(k^r)/m - r$, r=0,1,....,N(I)-1 for various Fibonacci congruences when c'-c=100.

## 3 Proof of theorem 2

In this section, the congruences are conguences modulo m. Now the first lemma is obvious.

**Lemma 3.1** *For n=3,4,...,d+1, $k_{n+1} > k_n > k_{n-1}$ . Moreover $k_{n+2} = h_{n+1}k_{n+1} + k_n$ is the Euclidean division of $k_{n+2}$ by $k_{n+1}$.*

Now, we prove the following results.

**Lemma 3.2** *Let n=0,1,2,...,d. If n is even, $\overline{k_n a} = m - r_n$. If n is odd, $\overline{k_n a} = r_n$.*

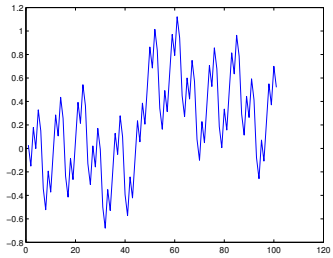


Figure 4: a= 1346269, m=2178309

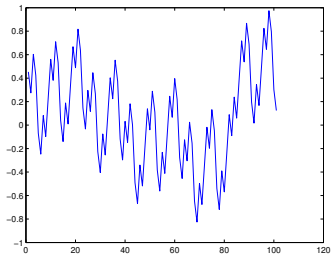


Figure 5: a= 121393, m=196418

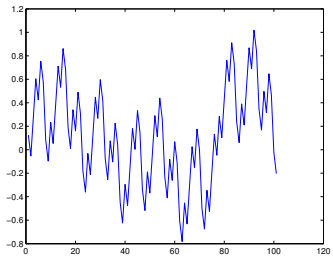


Figure 6: a= 10946, m=17711

**Proof** : We prove this lemma by recurrence.
For n=0, $\overline{k_n a} = \overline{0} = 0 = m - m = m - r_0$. For n=1, $\overline{k_n a} = \overline{a} = a = r_1$.

We suppose that it is true for n.
One supposes n even. Then, $k_{n+1}a \equiv ah_n k_n + ak_{n-1} \equiv -h_n r_n + r_{n-1} = r_{n+1}$.
One supposes n odd. Then, $k_{n+1}a \equiv ah_n k_n + ak_{n-1} \equiv h_n r_n - r_{n-1} = -r_{n+1} \equiv m - r_{n+1}$.
Therefore, $\overline{k_{n+1}a} = m - r_{n+1}$. ■

**Lemma 3.3** *Let n=2,3,...,d+1. Let $t \in \{1, 2, ..., k_n - 1\}$. If $n \geq 2$ is even, $r_{n-1} \leq \overline{at} < m - r_n$ . If $n \geq 3$ is odd, $m - r_{n-1} \geq \overline{at} > r_n$.*
*Moreover, if $n \geq 2$ is even, $\overline{k_n a} = m - r_n$. If $n \geq 3$ is odd, $\overline{k_n a} = r_n$.*

**Proof** : The second assertion is lemma 3.2. Now, we prove the first assertion by recurrence.
*One supposes n=2.* Then, $m = r_0 = h_1 r_1 + r_2 = h_1 a + r_2$. Moreover, $k_2 = h_1$. If $1 \leq t < h_1 = k_2$, $r_1 = a \leq at < h_1 a = m - r_2$ .

**One supposes that the first assertion is true for n** where $2 \leq n \leq d$.

Let $0 < t' < k_{n+1}$. Let $t' = fk_n + e$ be the Euclidean division of t' by $k_n$ : $e < k_n$.
Then, $f \leq h_n$ . If not, $t' \geq (h_n + 1)k_n + e \geq h_n k_n + k_{n-1} = k_{n+1}$.

**One supposes n even.**
In this case, $r_{n-1} \leq \overline{at} < m - r_n$ for $t \in \{1, 2, ..., k_n - 1\}$.
Moreover, $at' \equiv fak_n + ae \equiv f(m - r_n) + ae \equiv -fr_n + ae$.

First, one supposes $e = 0$. Then, $f \geq 1$.
Moreover, because $n \geq 2$, $m - r_n \geq m - fr_n \geq m - h_n r_n = m - (r_{n-1} - r_{n+1}) = r_0 - r_{n-1} + r_{n+1} \geq r_0 - r_1 + r_{n+1} > r_{n+1}$ .
Therefore, because $at' \equiv -fr_n$, $\overline{at'} = m - fr_n$ .
Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$ .

Now, one supposes $f < h_n$ and $e > 0$ .
By recurrence, $m - r_n \geq \overline{ae} \geq \overline{ae} - \underline{fr_n} \geq r_{n-1} - fr_n \geq r_{n-1} - (h_n - 1)r_n = r_n + r_{n+1} > r_{n+1}$.
Therefore, because $at' \equiv -fr_n + ae$, $\overline{at'} = \overline{ae} - fr_n$ .
Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$ .

One supposes $f = h_n$, $e \neq k_{n-1}$ and $e > 0$.
If $e \neq k_{n-1}$, $\overline{ae} \neq \overline{k_{n-1}a}$. Indeed, if not, $\overline{a(e - k_{n-1})} = 0$. For example, if $e - k_{n-1} > 0$, $k_n > e - k_{n-1} > 0$. Then, because our recurence, $\overline{a(e - k_{n-1})} > r_{n-1} > 0$ : it is impossible.
Now, if $n = 2$, $\overline{k_{n-1}a} = \overline{k_1 a} = \overline{a} = r_1 = r_{n-1}$.
Moreover, if $n > 2$, $n \geq 4$. Then, by recurence $\overline{k_{n-1}a} = r_{n-1}$.
Then, if $e \neq k_{n-1}$, $\overline{ae} \neq \overline{k_{n-1}a} = r_{n-1}$. Then, $\overline{ae} > r_{n-1}$.
Moreover, $m - r_n \geq \overline{ae} \geq \overline{ae} - fr_n > r_{n-1} - fr_n \geq r_{n-1} - h_n r_n = r_{n+1}$.
Therefore, because $at' \equiv -fr_n + ae$, $\overline{at'} = \overline{ae} - fr_n$ .
Therefore, $m - r_n \geq \overline{at'} > r_{n+1}$ .

One supposes $f = h_n$ and $e = k_{n-1}$. Then, $t' = h_n k_n + k_{n-1} = k_{n+1}$. It is oppositite to the assumption.

Then, in all the cases, for $t' \in \{1, 2, ...., k_{n+1} - 1\}$, $m - r_n \geq \overline{at'} > r_{n+1}$. Therefore, the lemma is true for n+1 if n is even. Then, it is also true for n+1=3.

**One supposes n odd** with $n \geq 3$. In this case, $r_n < \overline{at} \leq m - r_{n-1}$ for $t \in \{1, 2, ..., k_n - 1\}$. Moreover, $\overline{ak_n} = r_n$. Therefore, $at' \equiv fak_n + ae \equiv fr_n + ae$.

Assume $e = 0$. Then, $f \geq 1$.
Then, $r_n \leq fr_n \leq h_n r_n = r_{n-1} - r_{n+1} < m - r_{n+1}$.
Then, because $at' \equiv fr_n$, $r_n \leq \overline{at'} = fr_n < m - r_{n+1}$.

Assume $e > 0$ and $f \leq h_n - 1$.
By recurrence, $r_n < \overline{ae} + fr_n \leq m - r_{n-1} + fr_n \leq m - r_{n-1} + (h_n - 1)r_n = m - (r_{n-1} - h_n r_n) - r_n = m - r_{n+1} - r_n < m - r_{n+1}$.
Then, because $at' \equiv fr_n + ae$, $r_n < \overline{at'} = \overline{ae} + fr_n < m - r_{n+1}$

Assume $e > 0$, $e \neq k_{n-1}$ and $f = h_n$.
Because, $e \neq k_{n-1}$, $\overline{ae} \neq m - r_{n-1}$. If not, $\overline{ae} = \overline{ak_{n-1}} = m - r_{n-1}$. For example, if $e > k_{n-1}$, $a(e - k_{n-1}) = 0$ where $0 < e - k_{n-1} < k_n$. Then, by the assumption of recurrence, $\overline{a(e - k_{n-1})} > 0$. It is impossible.
Then, $\overline{ae} < m - r_{n-1}$.
Then, by recurrence, $r_n \leq \overline{ae} + h_n r_n \leq m - r_{n-1} + h_n r_n = m - r_{n+1}$.
Then, because $at' \equiv h_n r_n + ae$, $r_n \leq \overline{at'} = \overline{ae} + h_n r_n < m - r_{n+1}$

One supposes $f = h_n$ and $e = k_{n-1}$. Then, $t' = h_n k_n + k_{n-1} = k_{n+1}$. It is oppositite to the assumption.

Then the lemma is true for n+1. ■

**Lemma 3.4** *The following inequalities holds :* $k_{d+1} \leq m$.

**Proof** If $t \in \{1, 2, ..., k_{d+1} - 1\}$, by lemma 3.3, $r_d \leq \overline{at} < m - r_{d+1}$ or $m - r_d \geq \overline{at} > r_{d+1}$, i.e. $r_d \leq \overline{at} < m$ or $m - r_d \geq \overline{at} > 0$ where $r_d > 0$. Then, $0 < \overline{at} < m$ or $m > \overline{at} > 0$.
Then, if $k_{d+1} > m$, there exists $t_0 \in \{1, 2, ..., k_{d+1} - 1\}$ such that $t_0 = m$, i.e. $\overline{at_0} = \overline{am} = 0$. It is impossible. ■

**Lemma 3.5** *Let* $t,\ t' \in \{1, 2, ..., k_{d+1} - 1\}$ *such that* $\overline{at} = \overline{at'}$. *Then, t=t'.*

**Proof** Suppose $t > t'$. Then, $a(t - t') \equiv 0$ and $\overline{a(t - t')} = 0$. Then, by lemma 3.3, $r_d \leq \overline{a(t - t')} < m - r_{d+1}$ or $m - r_d \geq \overline{a(t - t')} > r_{d+1} = 0$ where $r_d > 0$. Then, $0 < \overline{a(t - t')}$. It is a contradiction. ■

**Lemma 3.6** *Let n=1,2,...,d. Let* $H_n = h_1k_1 + h_2k_2 + h_3k_3 + ..... + h_nk_n$. *Then,* $H_n = k_{n+1} + k_n - 1$.

**Proof** We have $H_n = h_1k_1 + h_2k_2 + h_3k_3 + ..... + h_{n-1}k_{n-1} + h_nk_n$

$= k_2 - k_0 + k_3 - k_1 + k_4 - k_2 + k_5 - k_3 + k_6 - k_4 + k_7 - k_5 + ..... + k_n - k_{n-2} + k_{n+1} - k_{n-1}$.

Therefore, if n=2m,

$$H_n =$$

$$k_2 - k_0 + k_3 - k_1 + k_4 - k_2 + k_5 - k_3 + k_6 - k_4 + ..... + k_{2m} - k_{2m-2} + k_{2m+1} - k_{2m-1}$$
$$= k_2 - k_0 + k_4 - k_2 + k_6 - k_4 + ..... + k_{2m} - k_{2m-2}$$
$$+k_3 - k_1 + k_5 - k_3 + k_7 - k_5 + ..... + k_{2m+1} - k_{2m-1}$$
$$= k_{2m} - k_0 + k_{2m+1} - k_1 = k_{n+1} + k_n - k_1 - k_0 = k_{n+1} + k_n - 1.$$

If n=2m+1

$$H_n =$$
$$k_2 - k_0 + k_3 - k_1 + k_4 - k_2 + k_5 - k_3 + k_6 - k_4 + ..... + k_{2m+1} - k_{2m-1} + k_{2m+2} - k_{2m}$$
$$= k_2 - k_0 + k_4 - k_2 + k_6 - k_4 + ..... + k_{2m+2} - k_{2m}$$
$$+k_3 - k_1 + k_5 - k_3 + k_7 - k_5 + ..... + k_{2m+1} - k_{2m-1}$$
$$= k_{2m+2} - k_0 + k_{2m+1} - k_1 = k_{n+1} + k_n - 1 \ . \ \blacksquare$$

**Lemma 3.7** *Let n=1,2,3,...,d-1 . Let* $L_n = \left\{t \middle| t = 0, 1, 2, ...., H_n\right\}$ . *Then, for all* $n \geq 1$, $L_{n+1} = \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.

**Proof** Let $l \in L_n$ , $l \leq H_n$. Let $g \leq h_{n+1}$.
Therefore, if $t = l + gk_{n+1}$, $t \leq H_n + h_{n+1}k_{n+1} = H_{n+1}$.
Therefore, $\left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\} \subset L_{n+1}$ .

Reciprocally, let $t \in L_{n+1}$ and let $t = fk_{n+1} + e$ , $e < k_{n+1}$ be the Euclidean division of t by $k_{n+1}$.

We know that $H_n = k_{n+1} + k_n - 1 \geq k_{n+1}$. Therefore, $e \leq H_n$. Therefore, $e \in L_n$.

Therefore, if $f \leq h_{n+1}$ , $t = fk_{n+1} + e \in \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$ .

Moreover, if $f > h_{n+1} + 1$ , $t = fk_{n+1} + e \geq (h_{n+1} + 2)k_{n+1} + e \geq h_{n+1}k_{n+1} + 2k_{n+1} = H_{n+1} - H_n + 2k_{n+1} = H_{n+1} - k_{n+1} - k_n + 1 + 2k_{n+1} = H_{n+1} + k_{n+1} - k_n + 1 \geq H_{n+1} + 1$ .
Therefore, $t \notin L_{n+1}$.

Then, suppose $f = h_{n+1}+1$. Then, $t = fk_{n+1}+e = (h_{n+1}+1)k_{n+1}+e = h_{n+1}k_{n+1}+k_{n+1}+e = H_{n+1} - H_n + k_{n+1} + e = H_{n+1} - k_{n+1} - k_n + 1 + k_{n+1} + e = H_{n+1} - k_n + 1 + e$.
Because $t \in L_{n+1}$ and $t = H_{n+1} - k_n + 1 + e$, $e + 1 - k_n \leq 0$. Therefore, $e \leq k_n - 1$.
Therefore, $t = fk_{n+1} + e = h_{n+1}k_{n+1} + k_{n+1} + e$,
where $k_{n+1} + e \leq k_{n+1} + k_n - 1 = H_n$
Therefore, $t = h_{n+1}k_{n+1} + e'$ where $e' \leq H_n$.
Therefore, $t \in \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$ .
Therefore, $L_{n+1} \subset \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$.

Therefore, $L_{n+1} = \left\{t = l + gk_{n+1} \middle| l \in L_n, g \leq h_{n+1}\right\}$ . $\blacksquare$.

**Lemma 3.8** *Let* $F_n = \left\{\overline{at} \middle| t = 0, 1, 2, ...., H_n\right\}$ .
*Let* $E_n = \left\{\overline{at} + km \middle| t = 0, 1, 2, ...., H_n, k \in \mathbb{Z}\right\}$ . *We set* $E_n = \{o_s^n | s \in \mathbb{Z}\}$ *where* $o_0^n = 0$ *et* $o_{s+1}^n > o_s^n$ *for all* $s \in \mathbb{Z}$.
*Then, for all* $s \in \mathbb{Z}$, $o_{s+1}^n - o_s^n = r_n$ *or* $o_{s+1}^n - o_s^n = r_{n+1}$.

**Proof** We prove this lemma by recurrence.

*Suppose n=1.* Then, $r_1 = a$, $H_1 = h_1k_1 = k_2 = h_1$. Therefore,
$F_1 = \{\overline{at}|t = 0, 1, 2, ..., h_1\} = \{0, a, 2a, ..., h_1a\} = \{0, r_1, 2r_1, ..., h_1r_1 = m - r_2\}$. Therefore, the lemma is true for n=1.

*Suppose that the lemma is true for n.*

Then, $E_{n+1} = \{\overline{at} + km|t = 0, 1, 2, ...., H_{n+1}, k \in \mathbb{Z}\}$,
where $H_{n+1} = h_1k_1 + h_2k_2 + h_3k_3 + ..... + h_{n+1}k_{n+1} = H_n + h_{n+1}k_{n+1}$.

Because $t \in \{0, 1, 2, ...., H_{n+1}\}$, $t \in L_{n+1}$. By lemma 3.7, si $t \in L_{n+1}$, $t = l + gk_{n+1}$ where $g \leq h_{n+1}$. By lemma 3.2, $\overline{at} \equiv \overline{a(l + gk_{n+1})} \equiv \overline{al} + (-1)^{n+2}gr_{n+1} \equiv \overline{al} + (-1)^n gr_{n+1}$ .

Therefore,

$$\begin{aligned} E_{n+1} &= \{\overline{at} + km|t \in L_{n+1}, k \in \mathbb{Z}\} \\ &= \{\overline{at} + km|t = l + gk_{n+1}, l \in L_n, g \leq h_{n+1}, k \in \mathbb{Z}\} \\ &= \{\overline{al} + (-1)^n gr_{n+1} + km|l \in L_n, g \leq h_{n+1}, k \in \mathbb{Z}\} \\ &= \{f + (-1)^n gr_{n+1} + km|f \in F_n, g \leq h_{n+1}, k \in \mathbb{Z}\} \\ &= \{o_s^n + (-1)^n gr_{n+1} + km|s \in Z, g \leq h_{n+1}, k \in \mathbb{Z}\}. \end{aligned}$$

*Suppose that n is even.*

Then, $o_s^n + (-1)^n gr_{n+1} = o_s^n + gr_{n+1} \leq o_s^n + r_n - r_{n+2}$ because $gr_{n+1} \leq h_{n+1}r_{n+1} = r_n - r_{n+2}$ .

Use the recurrence. Suppose $o_{s+1}^n - o_s^n = r_n$ . Then, $o_s^n + (-1)^n gr_{n+1} \leq o_s^n + r_n - r_{n+2} = o_{s+1}^n - r_{n+2}$.
Therefore,
$\{o_t^{n+1} \mid o_s^n \leq o_t^{n+1} < o_{s+1}^n\} = \{o_s^n < o_s^n + r_{n+1} < .... < o_s^n + h_{n+1}r_{n+1} < o_{s+1}^n\}$ .

Therefore, $o_{t+1}^{n+1} - o_t^{n+1} = r_{n+1}$ or $r_{n+2}$ if $o_s^n \leq o_t^{n+1} < o_{t+1}^{n+1} \leq o_{s+1}^n$.

Suppose $o_{s+1}^n - o_s^n = r_{n+1}$. Then, s is fixed .
Let $T = min\{t = 0, 1, ..., \ |o_{s+t+1}^n - o_{s+t}^n = r_n\}$. Therefore, $o_{s+T+1}^n - o_{s+T}^n = r_n$.
Let $O = \cup_{t=0}^{T}\{o_{s+t}^n + gr_{n+1} \mid 0 \leq g \leq h_{n+1}\}$.
Then, $O = \{o_s^n, o_{s+1}^n, ....., o_{s+T-1}^n\} \cup \{o_{s+T}^n + gr_{n+1}| \ 0 \leq g \leq h_{n+1}\}$.
Therefore, $O = \{o'_s, o'_{s+1}, ....., o'_{s+K}\}$ where $o'_{s'+1} - o'_{s'} = r_{n+1}$. Moreover, $o_{s+T+1}^n - o'_{s+K} = r_n - h_{n+1}r_{n+1} = r_{n+2}$.
Therefore, if $o_{t'}^{n+1}$ and $o_{t'+1}^{n+1} \in \{o_t^{n+1} \mid o_s^n \leq o_t^{n+1} \leq o_{s+T+1}^n\}$, $o_{t'+1}^{n+1} - o_{t'}^{n+1} = r_{n+1}$ or $r_{n+2}$.

*Suppose that n is odd.*

Then, $o_s^n + (-1)^n gr_{n+1} = o_s^n - gr_{n+1} \geq o_s^n - r_n + r_{n+2}$ because $gr_{n+1} \leq h_{n+1}r_{n+1} = r_n - r_{n+2}$ .

Suppose $o_s^n - o_{s-1}^n = r_n$ . Then, $o_s^n + (-1)^n gr_{n+1} \geq o_s^n - r_n + r_{n+2} = o_{s-1}^n - r_{n+2}$.
Therefore,
$\{o_t^{n+1} \mid o_s^n \geq o_t^{n+1} > o_{s-1}^n\} = \{o_s^n > o_s^n - r_{n+1} > ..... > o_s^n - h_{n+1}r_{n+1} > o_{s-1}^n\}$.
Therefore, $o_t^{n+1} - o_{t-1}^{n+1} = r_{n+1}$ or $r_{n+2}$ if $o_s^n \geq o_t^{n+1} > o_{t-1}^{n+1} \geq o_{s-1}^n$.

Suppose $o_s^n - o_{s-1}^n = r_{n+1}$. Let $T = min\{t = 0, 1, ..., \ |o_{s-t}^n - o_{s-t-1}^n = r_n\}$. Therefore, $o_{s-T}^n - o_{s-T-1}^n = r_n$
Let $O = \cup_{t=0}^{T}\{o_{s-t}^n - gr_{n+1} \mid 0 \leq g \leq h_{n+1}\}$.
Then, $O = \{o_s^n, o_{s-1}^n, ....., o_{s-T+1}^n\} \cup \{o_{s-T}^n - gr_{n+1}| \ 0 \leq g \leq h_{n+1}\}$.
Therefore, $O = \{o'_s, o'_{s-1}, ....., o'_{s-K}\}$ where $o'_{s'} - o'_{s'-1} = r_{n+1}$. Moreover, $o'_{s-K} - o_{s-T-1}^n = r_n - h_{n+1}r_{n+1} = r_{n+2}$.
Therefore, if $o_{t'}^{n+1}$ and $o_{t'-1}^{n+1} \in \{o_t^{n+1} \mid o_s^n \geq o_t^{n+1} \geq o_{s-T-1}^n\}$, $o_{t'}^{n+1} - o_{t'-1}^{n+1} = r_{n+1}$ or $r_{n+2}$. ■

**Proof 3.9** *Now one proves theorem 2.*

*Suppose that n is even.*
Then, $\overline{k_{n-1}a} = r_{n-1}$, $2\overline{k_{n-1}a} = 2r_{n-1}$, .....$h_{n-1}\overline{k_{n-1}a} = h_{n-1}r_{n-1} = r_n - r_{n-2}$.

Now, $\overline{ak_{n-1}\ell} = \overline{\ell r_{n-1}} = \ell r_{n-1}$ for $\ell = 0, 1, 2, ...., h_{n-1}$.

Therefore,

$\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\} = \big\{(k_{n-1}\ell, \overline{ak_{n-1}\ell})\big|\ell = 0, 1, 2, ...., h_{n-1}\big\} \subset E_2$ .

Moreover, $r_{n-2} = h_{n-1}r_{n-1} + r_n$. On the other hand, by lemma 3.8 , all the points of $E_2 = (t, \overline{at})$, $t \leq H_{n-1}$, have ordinates distant of $r_n$ or $r_{n-1}$.

Therefore, if there is other points of $E_2 \cap \big\{[0, H_{n-1}] \otimes [0, r_{n-2}[\big\}$ that the points $\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}$, there exists $\ell_0 \in \{1, 2, ...., h_{n-1}\}$ and $(x_1, y_1) \in E_2 \cap \big\{[0, H_{n-1}] \otimes [0, r_{n-2}]\big\}$ such that $r_{n-1}\ell_0 - y_1 = r_n$.

Because $H_{n-1} = k_n + k_{n-1} - 1 < k_{n+1} \leq k_{d+1}$, by lemma 3.5, there exists an only $t \in \{1, ...., H_{n-1}\}$, such that $\overline{at} = y_1 : t = x_1$. Because $y_1 \neq 0$, there exists an only $t \in \{0, 1, ...., H_{n-1}\}$, such that $\overline{at} = y_1$.

Now, $r_{n-1}\ell_0 - y_1 = \overline{a\ell_0 k_{n-1}} - \overline{at} = r_n = \overline{-ak_n}$. Then, $\overline{a\ell_0 k_{n-1}} - \overline{-ak_n} = \overline{at}$. Then, $\overline{a(\ell_0 k_{n-1} + k_n)} = \overline{at}$.

Because $r_{d-1} = h_d r_d$ with $r_{d-1} > r_d$, $h_d \geq 2$. Moreover, $d \geq n \geq 2$. Then, $d - 1 > 0$. Then, $k_{d-1} > 0$.

Then, by lemma 3.4, $0 < k_{n-1} + k_n \leq \ell_0 k_{n-1} + k_n \leq h_{n-1}k_{n-1} + k_n \leq k_n - k_{n-2} + k_n = 2k_n - k_{n-2} \leq 2k_d < 2k_d + k_{d-1} \leq h_d k_d + k_{d-1} = k_{d+1} \leq m$. Then, $0 < \ell_0 k_{n-1} + k_n < k_{d+1}$.

Now $0 < t \leq H_{n-1} = k_n + k_{n-1} - 1 < k_d + k_{d-1} \leq k_{d+1}$. Moreover, $0 < \ell_0 k_{n-1} + k_n < k_{d+1}$.

Then, because $\overline{a(\ell_0 k_{n-1} + k_n)} = \overline{at}$, by lemma 3.5, $t = \ell_0 k_{n-1} + k_n$.

Then, $t = \ell_0 k_{n-1} + k_n \geq k_{n-1} + k_n > H_{n-1}$. It is a contradiction.

Therefore, there is not other points of $E_2 \cap \big\{[0, H_{n-1}] \otimes [0, r_{n-2}[\big\}$ that $\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}$.

Therefore, there is not other points of $E_2 \cap \big\{[0, k_n[ \otimes [0, r_{n-2}[\big\}$ that the points $\big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}$.

Therefore,

$$E_2 \cap \big\{[0, k_n[ \otimes [0, r_{n-2}[\big\} = \big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\} .$$

According to what precedes,

$$\big\{(k_{n-1}\ell, \overline{ak_{n-1}\ell})\big|\ell = 0, 1, 2, ...., h_{n-1}\big\} = \big\{(k_{n-1}\ell, r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}$$

is located on the straight line $y = (r_{n-1}/k_{n-1})x$ if n is even.

*Suppose that n is odd.* Then, $\overline{k_{n-2}a} = r_{n-2}$, $\overline{k_{n-2}a + k_{n-1}a} = r_{n-2} - r_{n-1}$, $\overline{k_{n-2}a + 2k_{n-1}a} = r_{n-2} - 2r_{n-1}$, ....., $\overline{k_{n-2}a + h_{n-1}k_{n-1}a} = r_{n-2} - h_{n-1}r_{n-1}$.

Therefore,

$$\big\{(k_{n-2} + k_{n-1}\ell\ ,\ r_{n-2} - r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}$$
$$= \big\{(k_{n-2} + k_{n-1}\ell\ , \overline{k_{n-2}a + \ell k_{n-1}a}\big|\ell = 0, 1, 2, ...., h_{n-1}\big\}\ \subset E_2.$$

For $\ell = 0, 1, 2, ...., h_{n-1}$, $k_{n-2} + k_{n-1}\ell \leq k_{n-2} + h_{n-1}k_{n-1} = k_n$. Therefore,

$$\big\{(k_{n-2} + k_{n-1}\ell\ ,\ r_{n-2} - r_{n-1}\ell)\big|\ell = 0, 1, 2, ...., h_{n-1}\big\} \subset E_2 \cap \big\{]0, k_n] \otimes ]0, r_{n-2}]\big\} .$$

Moreover, $r_{n-2} - h_{n-1}r_{n-1} = r_n$. On the other hand, by lemma 3.8 , all the points of $E_2 = (t, \overline{at})$, $t \leq H_{n-1}$, have ordinates distant of $r_n$ or $r_{n-1}$.

Therefore, if there is other points of $E_2 \cap \left\{[0, H_{n-1}] \otimes ]0, r_{n-2}]\right\}$ that the points $\left\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\}$, there exists $\ell_0 \in \{1, 2, ...., h_{n-1}\}$ and $(x_1, y_1) \in E_2 \cap \left\{[0, H_{n-1}] \otimes [0, r_{n-2}]\right\}$ such that $y_1 - (r_{n-2} - r_{n-1}\ell_0) = r_n$.

Because $H_{n-1} = k_n + k_{n-1} - 1 < k_{n+1} \leq k_{d+1}$, by lemma 3.5, there exists an only $t \in \{1, ...., H_{n-1}\}$, such that $\overline{at} = y_1$. Because $y_1 \neq 0$, there exists an only $t \in \{0, 1, ...., H_{n-1}\}$, such that $\overline{at} = y_1$.

Then, $y_1 - (r_{n-2} - r_{n-1}\ell_0) = \overline{at} - \overline{k_{n-2}a + \ell_0 k_{n-1}a} = r_n = \overline{ak_n}$. Then, $\overline{at} = \overline{k_{n-2}a + \ell_0 k_{n-1}a} + \overline{ak_n}$. Then, $\overline{at} = \overline{a(k_{n-2} + \ell_0 k_{n-1} + k_n)}$.

Now, because $r_{d-1} = h_d r_d$ with $r_{d-1} > r_d$, $h_d \geq 2$. Now, $n \geq 3$. Then, $d - 1 \geq n - 1 > 1$. Then, $k_{d-1} > 0$.

Then $0 < k_{n-2} + \ell_0 k_{n-1} + k_n \leq k_{n-2} + h_{n-1}k_{n-1} + k_n \leq 2k_n \leq 2k_d < 2k_d + k_{d-1} \leq h_d k_d + k_{d-1} = k_{d+1} \leq m$.

Now $0 < t \leq H_{n-1} = k_n + k_{n-1} - 1 < k_d + k_{d-1} \leq k_{d+1}$.

Then, because $\overline{a(k_{n-2} + \ell_0 k_{n-1} + k_n)} = \overline{at}$, by lemma 3.5, $t = k_{n-2} + \ell_0 k_{n-1} + k_n$.

Then, $t = k_{n-2} + \ell_0 k_{n-1} + k_n \geq k_{n-1} + k_n > H_{n-1}$. It is a contradiction.

Therefore, there is not other points of $E_2 \cap \left\{[0, H_{n-1}] \otimes ]0, r_{n-2}]\right\}$ that the points $\left\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\}$.

Therefore, there is not other points of $E_2 \cap \left\{]0, k_n] \otimes ]0, r_{n-2}]\right\}$ that the points $\left\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\}$ : i.e.

$$E_2 \cap \left\{]0, k_n] \otimes [0, r_{n-2}[\right\} = \left\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\} .$$

According to what precedes,

$$\left\{(k_{n-2} + k_{n-1}\ell \ , \ r_{n-2} - r_{n-1}\ell)\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\}$$

$$= \left\{(k_{n-2} + k_{n-1}\ell \ , \overline{ak_{n-2} + ak_{n-1}\ell})\middle|\ell = 0, 1, 2, ...., h_{n-1}\right\}$$

is located on a straight line. ■

# 4 Correct models

## 4.1 General study

One can always suppose that $y_n$, n=1,2,...,N, is the realization of a sequence of random variables $Y_n$ defined on a probability space $(\Omega, A, P)$ : $y_n = Y_n(\omega)$ where $\omega \in \Omega$ and where $Y_n$ is a correct model of $y_n$.

As a matter of fact, there exist an infinity of correct models of $y_n$. It is thus necessary to be placed in the set of all the possible random variables.

**Notations 4.1** *Let $m \in \mathbb{N}^*$. One considers the sequences of random variables $Y_n^\theta$, n=1,.....,N, defined on the probabilities spaces $(\Omega, \mathcal{A}, P_\theta)$, $\theta \in \Theta$ : $(Y_1^\theta, \ Y_2^\theta, \ Y_3^\theta, \ Y_4^\theta, ....., Y_N^\theta) : \Omega \to \{0/m, 1/m, ...., (m-1)/m\}^N$. One assumes that $Y_n^\theta = Y_n$ for all $\theta \in \Theta$.*

For example, one can assume that $\Omega = \{0, 1, ...., m-1\}^N$ and $(Y_1, ...., Y_N) = (Id, ...., Id)$.

## 4.2 Definitions

It thus raises the question to define correctly what is a correct model. Indeed, if a model $Y_n^\theta$ is not correct, it is however possible that $y_n = Y_n^\theta(\omega)$ , where $Y_n^\theta$ is a sequence of random variable defined on a probability space $(\Omega, \mathcal{A}, P)$.

In the case where the model $Y_n^\theta$ is IID, to define a correct model is a generalization of the already very complex problem of the definition of an IID sequence (cf [1]). However one can have a solution because one wants only to prove that the correct models $T_q(Y_n^\theta)$ will be close to the IID model.

### 4.2.1 A scientific assumption

Generally, one feels well that correct models exist. In fact, it is a traditional assumption in science. In weather for example, the researchers seek a correct model, which implies its existence (if not, why to try to make forecasts?). One could thus admit that like a conjecture or a postulate without defining exactly what is a correct model.

### 4.2.2 Definition using tests

Now, in order to know if a sequence $y_n$ is a realization of a model $Y_n$, tests are generally used. So we could say that the model $Y_n$ is a correct model of a sequence $y_n$ if the sequence $y_n$ satisfies all the tests that one could make about hypothesis "$y_n$ is a realization of $Y_n$" with a frequency equivalent [2] to that real realizations of model $Y_n$.

### 4.2.3 Definitions using estimate

In fact, this is a case a little similar to the definition using the tests. We will estimate the parameters of the distribution of $Y_n$. In order to do this the easiest way is to estimate the marginal distributions and higher order correlation coefficients : cf [6]. This will define the dependence of the model $Y_n$.

In some cases, it is indeed possible. So for some texts, one can admit the Q dependence when Q is not too large (cf [9] section 10) and perhaps the stationarity. Indeed, in this case, there will need to define only some dependences between $Y_n$, $Y_{n+1}$ and $Y_{n+2}$, for example.

However, in more general cases, this may pose difficulties. Estimators may then be less clear. It is the relation between the coefficients of p variables p and p' variables, $p' > p$, which is a problem. For example if we have a sample of size N = 4, and points on each interval [0,1/2[ and [1/2,1[ of $X_1$ and $X_2$, it is possible that in the associated squares, there is one with an empirical probability equal to 0. If p is large, it is possible that there are many such hypercubes. This raises problems.

But in concrete cases, it is always possible to get estimates for models $Y_n$, n=,2,...,N. It can be assumed that such estimates exist.

### 4.2.4 Other definition

One can also define a correct model by the following way : if $Y_n$ is a correct model for the sequence $y_n$, n=1,2,...,N, that means that the event "the sequence $y_n$ is the result of a choice at random of $\omega \in \Omega$ where $y_n = Y_n(\omega)$" is an event which has reasonable probability to be carried out.

[2] It agree to know what we call "equivalent frequency". Then, in this definition, we assume that there is a mathematical answer, which is reasonable, but not completely sure. Note that there may exist infinitely from them and that we may be in the same problem (there are an infinite number of possible tests). In this case, the easiest will be to choose one from them. It is therefore assumed here that there are mathematical definitions of this fact without giving more details. For example the test of uniformity by the chi-square could be not verified : the probability of result which vould be found would be only one percent that uniform assumption is verified. But if we do 100 tests, it is possible that such an event happens. It would be necessary still that these tests are independent and it would then be necessary to know what "independent tests" means. Such a study might be long.

We thus find the same problem as the definition using tests : what is the probability that it is reasonable [3]? But we feel that a such definition means and it is reasonable. The simplest is to assume that it exists.

### 4.2.5 To predict the future

In fact, a correct model depends on its usefulness. For example, in meteorology, its usefulness is to predict weather.

One can transpose that to unspecified sequences of real numbers $y_n$, n=1,2,....,N. The usefulness of a model will be in general to predict the future. That applies perfectly to the research which we carry out in order to obtain IID sequences : if a sequence is IID random, one will not be able to predict the future knowing the past.

One could thus admit like definition of a correct model this one : a correct model is a model such as, knowing the past $Y^{\theta}_{n-s} = y'_{n-s}$, s=1,2,..., this one makes possible to predict the best possible the future. To be more complete, it is necessary to extend this definition to the sequences $y_{\phi(n)}$ where $\phi$ is a permutation of $\{1, 2, ...., N\}$.

It is necessary thus that the forecast is good : it has to be the most precise possible, but, if knowing the past, one predicts the future in a too precise way and that it is not real, the model will be bad.

Let us notice, that, under this condition, we suppose that one does not know the future $y_{\phi(n+s)}$, s=1,2,... : if not, the empirical probability would be a correct model.

### 4.2.6 Mathematical definition

Mathematically, one can thus specify that: it will be said that $Y^{\theta}_n$ is a correct model, if, for any permutation $\phi$ of $\{1, 2, ..., N\}$, for all sequence $y'_s$, for all n, it makes possible to give the conditional probability of $Y^{\theta_c}_{\phi(n)}$ knowing the past $Y^{\theta_c}_{\phi(n-1)} = y'_1$, $Y^{\theta_c}_{\phi(n-2)} = y'_2$,....., which is the best possible one.

It will be thus true in particular when $y'_s = y_{\phi(n-s)}$ for s=1,2,3,.... It will thus be known that $P\{Y^{\theta_c}_{\phi(n)} \in Bo \mid Y^{\theta_c}_{\phi(n-1)} = y_{\phi(n-1)}, Y^{\theta_c}_{\phi(n-2)} = y_{\phi(n-2)}, .....\}$ will be the most precise possible by taking account of what one really knows, i.e the sequence $y_{\phi(n-s)}$.

Therefore, one can nothing object to this conditional probability in order to define the future when what one really knows, it is the sequence $y_n$. Of course it is in question conditional probabilities which one could really deduce from the sample $y_n$ if all the mathematical properties were known and if one had an infinite computing power.

### 4.2.7 Some difficulties

Unfortunately, in these definitions, one made only to move the problem: mathematically, what means "probabilities the most precise possible" and "the best possible"? One understands well what one wishes. But to define it mathematically seems complicated.

However, one can do our study without knowing it. Indeed, which interests us, it is that the $X^{\theta}_n = T_q(Y^{\theta}_n)$ have a law close to an IID distribution.

Now, if $Y^{\theta_c}_n$ is a correct model, $P\{Y^{\theta_c}_{\phi(n)} \in Bo \mid Y^{\theta_c}_{\phi(n-1)} = y'_1, Y^{\theta_c}_{\phi(n-2)} = y'_2, .....\}$ defines the future $Y^{\theta_c}_{\phi(n)} \in Bo$ sufficiently well for all Borel set Bo, when, which one knows, it is the sequence $y_{\phi(n)}$. It will be thus true in particular for $P\{T_q(Y^{\theta_c}_{\phi(n)}) \in Bo' \mid Y^{\theta_c}_{\phi(n-1)} = y'_1, Y^{\theta_c}_{\phi(n-2)} = y'_2, .....\}$, and, therefore, for $P\{X^{\theta_c}_{\phi(n)} \in Bo' \mid X^{\theta_c}_{\phi(n-1)} = x'_1, X^{\theta_c}_{\phi(n-2)} = x'_2, ....\}$ (cf proposition A1 [11]). Therefore, this conditional probability defines a good forecast of the future. That means that if one knows $x_{\phi(n-s)}$, s=1,2,.., a good prediction of $x_{\phi(n)}$ will be given by this conditional probability.

[3] Remark that if $y_n = Y_n(\omega)$ is an event which has reasonable probability to be carried out, the tests will have to be checked with a good frequency.

However we will prove in theorem 3 that $P\{X^{\theta_c}_{\phi(n)} \in Bo' \mid X^{\theta_c}_{\phi(n-1)} = x'_1, X^{\theta_c}_{\phi(n-2)} = x'_2, ....\} = L(Bo')[1+Ob(1)\epsilon]$ where $\epsilon$ is small enough for the models with a continuous density and a coefficient of Lipschitz $K'_0$ not too large. Moreover, one will understand in section 8 that one can admit that such models are correct if $y_n$ is obtained from texts. At last, we shall prove in section 5.9 that, in this case, there exists a correct model $Y^{\theta_c}_n$ such that $P\{X^{\theta_c}_{\phi(n)} \in Bo' \mid X^{\theta_c}_{\phi(n-1)} = x'_1, X^{\theta_c}_{\phi(n-2)} = x'_2, ....\} = L(Bo')$ if $\epsilon$ is small enough.

That means that if one knows $x_{\phi(n-s)}$, s=1,2,.., a good prediction of $x_{\phi(n)}$ will be given by uniform probability. Then, we have proved that, there exists a correct model $Y^{\theta_c}_n$ such that $T_q(Y^{\theta_c}_n)$ is exactly the IID random sequence.

#### 4.2.8 A Problem

It raises a problem : according to the definition which one chooses, it is often the empirical model which is the best model possible. It is a known problem of the definition of an IID sequence : some say there is no random sequence of finite dimension (cf [1] and section 10.2 of [9]).

In this case, the IID model is one of the worst model. For example if we use the definition of Section 4.2.4, an increasing sequence $x_n$ has as much chance to be realized than a real IID sample IID. In spite of this difficulty, we happens to prove that this model is correct for the sequences $T_q(y_n)$. It's almost a feat.

In order to adress this problem, we can try to impose conditions - known a priori or not - like the Lipschitz coefficient not too large. But in this case again, it will be the models whose marginal densities are bumps near each point of the sample $y_n$ which are the best. This is also true for a real IID sample : the model with bumps will be the best model possible.

Even if we assumed that the $Y_n$'s have the same distribution, one could assume that each $Y_n$ is concentrated only on the points of the sample. One might assume that there is in addition a function g such that $Y_{n+1} = g(Y_n)$ (in this case, $K_0$ would be great) .

#### 4.2.9 Other definition

To remedy this, we can introduce a new definition : we say that a model $Y_n$ is correct if the hypothesis "$y_n = Y_n(\omega)$" is an assumption that nothing prevents. This means that $y_n$ is perfectly plausible as realization of $Y_n$. Then, expected properties of a such sample should be checked, especially the tests and estimates.

#### 4.2.10 Connection between some definitions

Some definitions have common points : the definition which defines the best conditional probability for all sequences $y_{\phi(n)}$ and the definition saying that the hypothesis "$y_n = Y_n(\omega)$" is a plausible hypothesis (especially using the tests and estimates). In order to understand this connection, it is enough to assume that the conditional probabilities are defined by estimate (e.g. by using higher order correlation coefficients cf [6]). It can be difficult to define all the conditional probabilities depending on the choice of the permutations $\phi$ in some cases. But in others cases, it is possible as is the case for texts. In this case the two definitions are almost equivalent because how is it possible to know the conditional probabilities of the model if we are not able to estimate?

In conclusion, a correct model would make possible to obtain the best conditional probability for all sequences $y_{\phi(n)}$, In particular, by taking into account the estimate

Does there exist such a model? Presumably, because generally $y_n$ represents a physical phenomenon. It is thus normal to suppose its existence. Moreover, in certain cases, one can show such models : it is the case for texts.

### 4.3 Texts

Now, we consider the particular case where the $y_n$'s result from texts.

A priori, a correct model would be a model which makes possible to predict the following letters ($y_n$, $y_{n+1}$,.....) with a satisfactory probability if one knows the preceding letters $y_{n-1}$, $y_{n-2}$ ..... One could thus say that the model will predict all the possible texts which follows the beginning of the text.

We will choose therefore as model the model of all possible texts with the uniform probability (or another: see below). We can also choose more perfect models as all possible texts of the author.

This seems a good model because knowing any subsequence, we can predict the following letters with reasonable probability [4].

However such a model is too precise: indeed, for sequences representing a text, to suppose that one is in an English text is a priori which is wrong : cf 6) page 307 of [9]. For example, one could logically predict words invented not existing. A model in modern English language would be a correct model. But a model in a possible evolution of the English language would be it too.

These model can be refined besides: if a novel is used, it would be astonishing to find texts speaking about mathematical theorem. Therefore, there are models which make possible to better predict the continuation than others. But it is necessary that is explained by the text which precedes. If one takes only 100 words, one will not deduce from it the style of the author.

In fact in order to admit that only the English texts can represent the $y_n$, it would be necessary that sequence $y_n$ consists of a very large number of books which make possible to decode the language. In this case, it is possible that the only correct models are texts, even texts of the author.

Let us suppose that it is the case. That makes possible to define precise correct models. Indeed, in this case, one can admit that the correct model will be that representing all the possible texts written according to the style of the author. Of course, there is an almost infinite number of possible texts as soon as N, the sample size of $y_n$ is large.

Concerning the associated probabilities, one can suppose that all the texts are equiprobable. That seems a correct model.

But it is not the alone one. One can choose other probabilities than the equiprobable probability, for example a close probability, even another. Indeed, it seems that certain text are likely more to exist than the different ones. The equiprobable model is thus not the best inevitably. In order to find the best models it would be necessary to find those whose probabilities correspond the best to all which one knows about texts of the author. That seems impossible to realize. But theoretically, it could exist. In fact, there are several suitable models.

It thus seems difficult to find exactly all the possible correct models and especially to find a better model. However, it is felt well that these models including all the texts which the author can write seems rather correct and that there are from them which are better than others.

Therefore, for the texts, one can show correct models. All the possible texts of the author with an about uniform probability seems be a good model. Then this model defines conditional probabilities $P\{Y^{\theta_t}_{\phi(n)} \in Bo | Y^{\theta_t}_{\phi(n-1)} = y'_1, Y^{\theta_t}_{\phi(n-2)} = y'_2, ....\}$ for all n, for all $y'_s$, s=1,2,..., and for all permutation $\phi$.

Now, if we use the definition by estimates, we obtain similar results if we assume that we have a very large number of texts at our disposal.

Of course, this is not the case. But it is not serious : what matters to us is being able to increase the Lipschitz coefficients of conditional probabilities. We reach this result by adding a pseudo random and a text written backward (cf section 8).

[4] It is not embarrassing to be limited to the following letters : it is enough to take a subsequence containing the letters preceding and following a portion of text to get a correct estimate.

### 4.4 Conclusion

Thus in certain cases, there exist correct models which enable us to predict the future correctly. One can suppose that the method described for the texts is good and can be generalized.

If this assumption is refused, it may be easier to admit that there exists such correct models defining correctly the conditional probabilities without more precise details as one does it in weather and elsewhere. It was understood that it is enough in order to prove that the IID model is a correct model of $x_n = T_q(y_n)$.

## 5 Models equivalent with a margin of $\epsilon$

### 5.1 The problem

Let $Y_n^{\theta_2}$ and $Y_n^{\theta_1}$ be two sequences of random variables such that, for all Borel set Bo,

$$P\big\{(Y_1^{\theta_2}, ....., Y_N^{\theta_2}) \in Bo\big\} = P\big\{(Y_1^{\theta_1}, ....., Y_N^{\theta_1}) \in Bo\big\}\big[1 + Ob(1)\epsilon\big] \ ,$$

where $\epsilon$ is small enough and where Ob(.) means the classical O(.) with the additional condition $|Ob(1)| \leq 1$. One supposes that $Y_n^{\theta_1}$ is a correct model of the sequence $y_n$, n=1,2,....,N. One wants to prove that $Y_n^{\theta_2}$ is also a correct model of $y_n$ if $\epsilon$ is small enough.

### 5.2 Example

Let us suppose that we have a really IID sequence of random variables $X_n^\epsilon$ with uniform distribution on [0,1/2] and [1/2,1] and with a probability such as $P\{X_n^\epsilon \in [1/2, 1]\} = 0,500[1 + \epsilon]$ where $\epsilon = 0,001$. Then, this sequence has not the uniform distribution on [0,1]. However, if we have a sample with size 10, we will absoluetely not understand that $X_n^\epsilon$ has not the uniform distribution on [0,1]. It is wellknown that one need samples with size larger than N=1000 minimum in order to test this difference.

More precisely, by the CLT (Central Limit Theorem), $P\Big\{\frac{|\sum_{n=1}^N (\mathbb{1}_{[1/2,1]}(X_n^\epsilon) - 1/2 - \epsilon/2)|}{\sqrt{N(1-\epsilon^2)/4}} \geq b\Big\} \approx \Gamma(b)$ where $\Gamma(b) = P\{|X_G| \geq b\}$ when $X_G \sim N(0,1)$. Then, $P\Big\{\frac{|\sum_{n=1}^N (\mathbb{1}_{[1/2,1]}(X_n^\epsilon) - 1/2)|}{\sqrt{N/4}} \geq b\Big\} \approx \Gamma\big(b[1 - \eta(\epsilon)]\big)$ where $\eta$ is continuous with $\eta(0) = 0$.

More generally, one cannot test significantly $H_0$ : "$X_n^\theta$ has the uniform distribution" against $H_1(\epsilon)$ : "$P\{X_n^\theta \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$ " if $\sqrt{N}\ \epsilon \leq 1/10$.

For example, if $\sqrt{N}\ \epsilon = 1/10$ and b=2, the probability of obtaining $\frac{\sum_{n=1}^N [1_{[1/2,1]}(X_n^\theta) - 1/2]}{\sqrt{N/4}} \geq 2$ is about 0.0466 under $H_1(\epsilon)$ and about 0.0455 under $H_0$ : i.e. the probability of rejecting the assumption IID, $H_0$, under $H_1(\epsilon)$ is not much bigger than that of rejecting $H_0$ if $X_n^\theta$ is really IID (cf also section 4.3 of [11]).

### 5.3 IID models with a margin of $\epsilon$

These results hold in dimension p , i.e. for $\frac{1}{N-p}\sum_n \mathbb{1}_{Bo_1}(Y_{n+j_1}^{\theta_1}).....\mathbb{1}_{Bo_p}(Y_{n+j_p}^{\theta_1})$. One deduces from what precedes that, if $x_n$ is the realization of a sequence of random variables $X_n^\theta$ such that $P\{(X_1^\theta, ...., X_N^\theta) \in Bo\} = L(Bo)[1+Ob(1)\epsilon]$ for all Borel set Bo, one will not be able to differentiate this model from an IID model if $\epsilon$ is rather small with respect to N.

Reciprocally, if $x_n$, n=1,2,....,N, is really an IID sample, a model such that $P\{(X_1^\theta, ...., X_N^\theta) \in Bo\} = L(Bo)[1 + \epsilon]$ is also a correct model of the sequence $x_n$.

Because we shall obtain $P\{(X_1^\theta, ...., X_N^\theta) \in Bo\} = L(Bo)[1 + \epsilon]$ in theorem 3 if m and q are well chosen, one will be able to admit that the IID model is a correct model of the sequences $x_n$ which we built in this report.

## 5.4 Case where the CLT holds

One can adopt assumptions more general than those of the IID case by only supposing that the CLT is checked. For example, assume that the CLT holds and that the $Y_n^{\theta_1}$'s have the same distribution for n=1,2,....,N. Let $P_{Y_1}(I) = P\{Y_n^{\theta_1} \in I\}$ where I is an interval. Let $P_e^1 = (1/N)\sum_n \mathbb{1}_I(Y_n^{\theta_1})$ and $P_e^2 = (1/N)\sum_n \mathbb{1}_I(Y_n^{\theta_2})$. Let $\sigma_B^2$ the variance of $P_e^1$. Then, if N is big enough, by the CLT,

$$P\{|P_e^1 - P_{Y_1}(I)| > \sigma_B b\} \approx \Gamma(b) \ ,$$

where $\Gamma(b) = P\{|X_G| \geq b\}$ when $X_G \sim N(0,1)$. Then, it is easy to prove (cf page 8 of [11])

$$P\{|P_e^2 - P_{Y_1}(I)| > \sigma_B b\} \approx \Gamma(b)[1 + Ob(1)\epsilon] \ .$$

Then, there will not be possible to conclude that $y_n$ is a realization of $Y_n^{\theta_1}$ rather than of $Y_n^{\theta_2}$ by testing $P_{Y_1}(I)$. For example, let us suppose $N = 10^4$, $\epsilon = 0.00001$. In this case, for b=2,

$$P\{|P_e^1 - P_{Y_1}(I)| > 2\sigma_B\} \approx 0.0455,$$

$$P\{|P_e^2 - P_{Y_1}(I)| > 2\sigma_B\} \leq c_2, \ \textit{where } c_2 \approx 0.0500 \ .$$

Now, if $y_n$ is a realization of $Y_n^{\theta_1}$, it is known that $(1/N)\sum_n \mathbb{1}_I(y_n)$ is close to $P_{Y_1}(I)$ with a certain probability : it is completely possible that $(1/N)\sum_n \mathbb{1}_I(y_n)$ is enough different from $P_{Y_1}(I)$, but the probability that occurs is weak.

Moreover, if $y_n$ is a realization of $Y_n^{\theta_2}$, it is also possible that $(1/N)\sum_n \mathbb{1}_I(y_n)$ is enough different from $P_{Y_1}(I)$, but that is not likely much more to occur than if $y_n$ is a realization of $Y_n^{\theta_1}$.

Then, for the test associated to $P_{Y_1}(I)$, it will be thus impossible to differentiate the model $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$ as good model for the sequence $y_n$.

These results are not only true for the estimate of only one $P_{Y_1}(I)$, but of several (cf page 9 of [11] with $p_s = P\{Y_n^{\theta_1} \in I_s\}$) :

$$P\Big\{N\sum_s \Big[\frac{1}{N}\sum_n \mathbb{1}_{I_s}(Y_n^{\theta_2}) - p_s\Big]^2 \geq a\Big\} = P\Big\{N\sum_s \Big[\frac{1}{N}\sum_n \mathbb{1}_{I_s}(Y_n^{\theta_1}) - p_s\Big]^2 \geq a\Big\}[1 + Ob(1)\epsilon] \ .$$

Then, if $\epsilon$ is small enough, one cannot differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$ by this chi squared test.

One can generalize these results in dimension p : one uses $\sum_n \mathbb{1}_{Bo_1}(Y_{n+j_1}^{\theta_1})$.....$\mathbb{1}_{Bo_p}(Y_{n+j_p}^{\theta_1})$. Of course, one can also generalize to other functions, i.e. to about the totality of the known tests. Because of it, it seems impossible to differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$ as models of $y_n$ .

Then, we have just studied the tests associated to these models. In order to be able to apply them it is useful to be able to use the CLT. Now, in general, the sequences $y_n$ which we use are asymptotically independent (for example texts or numbers provided by machines). The models where the CLT is checked are thus correct. The conclusions that we deduce of it are thus correct too : it is impossible to differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$ as models of $y_n$ .

## 5.5 Another case

As a matter of fact, the relation $P\{(Y_1^{\theta_2}, ....., Y_N^{\theta_2}) \in Bo\} = P\{(Y_1^{\theta_1}, ....., Y_N^{\theta_1}) \in Bo\}$ $[1 + Ob(1)\epsilon]$ for all Borel set $Bo \subset \{0/m, 1/m, ...., (m-1)/m\}^N$ is a very strong relation. Because of it, it seems impossible to differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$ as models of $y_n$ in other cases than the case where the CLT holds.

For example, this results holds also if only the Weak Law of Large Number holds. Indeed one does not know the exact law of $P_e - P_{Y_1}(I)$. But it exists theoretically. However, to know this law

is not important : it is enough that one has the relation $P\{|P_e^2 - P_{Y_1}(I)| > b\} = P\{|P_e^1 - P_{Y_1}(I)| > b\}[1 + Ob(1)\epsilon]$ for all $b > 0$ in order to be able to conclude from it that one will cannot differentiate the models $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$.

Moreover, the inequality of Bienaymé-Tschebischeff shows that the sums divided by the variance are normalized. One deduced from it that one cannot differentiate the effects of these models.

## 5.6 General Case

One now asks if to prove this result in the general case is possible, i.e. if, whatever the model $Y_n^{\theta_1}$ (for example without tests), the relation $P\{(Y_1^{\theta_2}, ...., Y_N^{\theta_2}) \in Bo\} = P\{(Y_1^{\theta_1}, ...., Y_N^{\theta_1}) \in Bo\}[1 + Ob(1)\epsilon]$ implies always that one cannot differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$. It is maybe the case. But, in order to prove it, there is likely philosophical or other problems of the type of the definition of the randomness of Franklin (cf [1], [17]). That is thus likely a complicated study.

But one can say still a certain number of thing in the general case.

### 5.6.1 Empirical probability

It is observed now that, if a model $Y_n^{\theta_1}$ is correct and a model $Y_n^{\theta_2}$ is not correct, it would be necessary that a variation of the probability which would be smaller than $P\big\{(Y_1^{\theta_1}, ....., Y_N^{\theta_1}) \in Bo\big\}\epsilon$ exchange something sufficiently important so that one understands a difference of the models with respect to the sample. Therefore, the probability in question will be close to the empirical probability. Thus the model would be very close to the empirical model.

However, the empirical model is in general a bad theoretical model. Thus, in the case of texts, it is known a priori that the empirical probability is not the good model because it will fail as soon as one increases N. One thus arrives at a contradiction.

Then, even if the empirical probability can be selected like correct model, a probability of a model $Y_n^{\theta_2}$ where one changes only a little this probability is also correct.

It would be thus astonishing that a model as special as the empirical model $Y_n^{\theta_1}$ satisfies effectively that, if $Y_n^{\theta_1}$ is correct, an approximate model $Y_n^{\theta_2}$ will be it also and that an unspecified model does not check this implication. In particular, it would be astonishing for models with continuous density and coefficient of Lipschitz not too large. It would be even astonishing for models with unspecified coefficient of lipschitz, i.e. in the general case. Of course astonishing means that this is intuitive.

### 5.6.2 Presentation of the intuition

In fact, this intuition is based on the following reasoning: if $Y_n^{\theta_1}$ is a correct model for the sequence $y_n$, that means that the event "the sequence $y_n$ is the result of a choice at random of $\omega$ where $y_n = Y_n^{\theta_1}(\omega)$" is an event which has reasonable probability to be carried out. Then, it is not understood what can prevent that $y_n = Y_n^{\theta_2}(\omega)$ is a realization equally probable if one changes only a little the probabilities (except in the case studied in section 5.8).

The only cases where they could have problem seem those of the probability concentrated close to some points like the empirical probability. But one has just understood that even in this case, it is still true.

One thus understands well what leads to think that, in all the cases, one will not be able to differentiate $Y_n^{\theta_1}$ and $Y_n^{\theta_2}$.

## 5.7 Value of $\epsilon$

To get an idea of the value of $\epsilon$, the best is to return to the definition using tests.
Because $P\{(Y_{_1}^{\theta},....,Y_N^{\theta}) \in Bo\} = P\{(Y_1^{\theta_0},....,Y_N^{\theta_0}) \in Bo\}[1+Ob(1)\epsilon]$ implies that $P\{g(Y_{_1}^{\theta},....,Y_N^{\theta}) \in Bo'\} = P\{g(Y_1^{\theta_0},....,Y_N^{\theta_0}) \in Bo'\}[1+Ob(1)\epsilon]$, we can consider that all tests defined by a function of type g, will produce results not very different if $\epsilon$ is small enough. Then, one can choose $\epsilon = 1/10$, $1/100$ ou $1/1000$, ..... Now, if we wanted to avoid any doubt, intuition would dictate that we choose $\epsilon$ as a function of N. Intuitively, one might therefore wish to impose $\epsilon = 1/N$. But this is probably exaggerated and there is nothing which justifies this intuition. Moreover, in theorem 3, it already imposed $\epsilon = K_0 N 2^q/m$. The idea of choosing $\epsilon$ as a function of N is already realized

## 5.8 A problem

### 5.8.1 The problem

But it is necessary to add something to these assertions. If the model $Y_n^{\theta_1}$ is correct and that the model $Y_n^{\theta_2}$ is also correct, a model $Y_n^{\theta_3}$ equivalent with a margin of $\epsilon$ to $Y_n^{\theta_2}$ would be it also correct with the relation $P\{(Y_{_1}^{\theta_3},....,Y_N^{\theta_3}) \in Bo\} = P\{(Y_1^{\theta_1},....,Y_N^{\theta_1}) \in Bo\}[1+\mathbf{2}\mathrm{Ob}(1)\epsilon]$? A priori not inevitably!

If it is admitted, one would manage to find that the models $Y_n^{\theta_p}$ checking $P\{(Y_{_1}^{\theta_p},....,Y_N^{\theta_p}) \in Bo\} = P\{(Y_1^{\theta_1},....,Y_N^{\theta_1}) \in Bo\}[1+\mathbf{p}\mathrm{Ob}(1)\epsilon]$ would be also correct. One would end up finding models which would not be correct.

Therefore, there is no reason that $Y_n^{\theta_3}$ is also correct. It cannot be differentiated of $Y_n^{\theta_2}$, but not of $Y_n^{\theta_1}$. In other words, this relation is not transitive.

### 5.8.2 IID Case

That thus poses a problem because if one uses for example a realization $y_n$ of the IID model, and that if one takes for sequence $Y_n^{\theta_1}$ a model checking $P\{(Y_{_1}^{\theta_1},....,Y_N^{\theta_1}) \in Bo\} = L(Bo)[1+Ob(1)\epsilon^1]$ where $\epsilon^1$ is small enough but not very small, there are no reasons a priori that $Y_n^{\theta_2}$ is a correct model. Indeed, in order that $Y_n^{\theta_2}$ is not correct, it is enough that $Y_n^{\theta_1}$ is in extreme cases of the correct models, i.e. it is enough that $\epsilon^1$ is in extreme cases of the possible values of the $\epsilon$'s such that $P\{(Y_{_1}^{\theta},....,Y_N^{\theta}) \in Bo\} = L(Bo)[1+Ob(1)\epsilon]$, $sup_{Bo}(Ob(1)) = 1$, imply that $Y_n^{\theta}$ is a correct model.

### 5.8.3 What we want

But what interests us is that there exists correct models $Y_n^{\theta_1}$ such that all models close $Y_n^{\theta}$ , i.e. checking $P\{(Y_{_1}^{\theta},....,Y_N^{\theta}) \in Bo\} = P\{(Y_1^{\theta_1},....,Y_N^{\theta_1}) \in Bo\}[1+Ob(1)\epsilon]$ would be also correct. But we need that $\epsilon$ is small but not too, i.e. of the order of what we saw : $\epsilon = 1/10$, $1/100$ or at worst $\epsilon = 1/N$ if need be (cf section 5.7)

### 5.8.4 Case of a known model

To understand that this is the case, suppose first that we have an sequence $x_n$, sample of an IID sequence $X_n$ and that it is a good realization of $X_n$. So we know that models checking $P\{(X_1^{\theta},....,X_N^{\theta}) \in Bo\} = L(Bo)[1+Ob(1)\epsilon]$ will also correct models. In this case, we can admit that there is a model (the IID model) such as all close models are correct models.

If we are not in the IID case but in any case and if one knows the model : it is the same matter.

However we can always accept that a sequence $y_n$ is the realization of a given model $Y_n^1$ : this is indeed the usual hypothesis in Statistics. The model $Y_n^1$ will be thus "at the center of some models close" : $P\{(Y_1^{\theta},....,Y_N^{\theta}) \in Bo\} = P\{(Y_1^1,....,Y_N^1) \in Bo\}[1+Ob(1)\epsilon]$ implies that $Y_n^{\theta}$ will be also a correct model.

Intuitively, one feels that it is general : there are many correct models, $Y_n^\theta$ of a sequence $y_n$, n=1,2,...,N, such that models close with a margin of $\epsilon$ are also correct.

### 5.8.5 Use of estimates

Suppose we take the IID example of section 5.2 : $P\{X_n^\epsilon \in [1/2,1]\} = 0,5[1+\epsilon]$ when $\epsilon = 0,001$. Suppose that the sample $x_n$ is a good sample. It is then clear that there are many correct models close to the model $X_n^\epsilon$ if $\epsilon$ is small enough.

But it is possible that there are other such models. Thus, we can choose for correct model, the IID model $X_n^2$ such that $P\{X_n^2 \in [0,1/2[\} = p_e$, the empirical probablity of [0,1/2[.

In this case, it is clear that there are many close models which are correct. For example, consider as a model $P\{X_n^3 \in [0,1/2[\} = p_e(1+0.0003)$. It is also a correct model and the close models $P\{(X_1^\theta,....,X_N^\theta) \in Bo\} = P\{(X_1^3,....,X_N^3) \in Bo\}[1+Ob(1)\epsilon]$ will be also correct models if $\epsilon$ is small.

More generally, we know that it should exist estimates of models (these estimates are easier to calculate in some cases as texts). Then, we can choose as model $Y_n^{\theta_1}$, the model provided by these estimates. If these estimates are correct, then it is clear that all close models checking $P\{(Y_1^{\theta_2},....,Y_N^{\theta_2}) \in Bo\} = P\{(Y_1^{\theta_1},....,Y_N^{\theta_1}) \in Bo\}[1+Ob(1)\epsilon]$ will be also correct models.

## 5.9 Exact IID model

Then if $Y_n^\theta$ is a correct model such as $T_q(Y_n^\theta)$ cannot be differentiated with the IID model, one will be able to choose another correct model $Y_n^{\theta_0}$ close to $Y_n^\theta$ and such that $T_q(Y_n^{\theta_0})$ is exactly the IID model.

**Proposition 5.1** *One assumes that m is large enough. Let $Y_n^{\theta_c}$ be a correct model of the sequence $y_n$. One assumes that there exists $\epsilon_Y > 0$ such that if $Y_n^\theta$ is a model satisfying, for all Borel set Bo, $P\big\{(Y_1^\theta,.....,Y_N^\theta) \in Bo\big\} = P\big\{(Y_1^{\theta_c},.....,Y_N^{\theta_c}) \in Bo\big\}[1+Ob(1)\epsilon_Y]$, then $Y_n^\theta$ is a correct model of $y_n$.*

*One assumes also that, for all $(k_1,....,k_N)$,*

$$P\big\{\{T_q(Y_1^{\theta_c}) = k_1/2^q\} \cap ..... \cap \{T_q(Y_N^{\theta_c}) = k_N/2^q\}\big\} = \frac{1}{2^{qN}}[1+\epsilon_{k_1,....,k_N}(q)]$$

*where $sup_{k_1,....,k_N}|\epsilon_{k_1,....,k_N}(q)| = \epsilon_X(q)$. One assumes that $\epsilon_X(q)$ is increasing, that $\epsilon_X(1) << \epsilon_Y$ and that there exists $q_1 \in \mathbb{N}^*$ such that $\epsilon_X(q_1)$ is small enough.*

*Then, there exists $q_0 \in \mathbb{N}^*$ and a correct model $Y_n^{\theta_0}$ of the sequence $\{y_n\}_{n=1,...,N}$ such that, for all $(k_1,....,k_N)$,*

$$P\big\{\{T_{q_0}(Y_1^{\theta_0}) = k_1/2^q\} \cap ..... \cap \{T_{q_0}(Y_N^{\theta_0}) = k_N/2^q\}\big\} = \frac{1}{2^{q_0N}} \ .$$

**Proof** There exists $q_0 \leq q_1$ such that $\epsilon_X(q_0) \leq (1/2)\epsilon_Y$. Then, one uses the model $Y_n^{\theta_0}$ such that, for all $(k_1,....,k_N)$,

$$P\big\{(Y_1^{\theta_0},.....,Y_N^{\theta_0}) = (y_1',.....,y_N')\big\} = \frac{P\big\{(Y_1^{\theta_c},.....,Y_N^{\theta_c}) = (y_1',.....,y_N')\big\}}{1+\epsilon_{k_1,....,k_N}(q_0)}$$

for all $y_1' \in T_{q_0}^{-1}(k_1/2^{q_0}),.....,y_N' \in T_{q_0}^{-1}(k_N/2^{q_0})$. It ckecks

$$P\big\{\{T_{q_0}(Y_1^{\theta_0}) = k_1/2^{q_0}\} \cap ..... \cap \{T_{q_0}(Y_N^{\theta_0}) = k_N/2^{q_0}\}\big\} = \frac{1}{2^{q_0N}} \ .$$

It ckecks also : for all $(y_1',.....,y_N')$,

$$P\big\{(Y_1^{\theta_0},.....,Y_N^{\theta_0}) = (y_1',.....,y_N')\big\} = P\big\{(Y_1^{\theta_c},.....,Y_N^{\theta_c}) = (y_1',.....,y_N')\big\}[1+Ob(1)\epsilon_Y']$$

where $|\epsilon'_Y| \leq C_0 \approx \epsilon_X(q_0)$. Then, $|\epsilon'_Y| < \epsilon_Y$. Then, for all Borel sets Bo,

$$P\big\{(Y_1^{\theta_0}, ....., Y_N^{\theta_0}) \in Bo\big\} = P\big\{(Y_1^{\theta_c}, ....., Y_N^{\theta_c}) \in Bo\big\}[1 + Ob(1)\epsilon'_Y] \ .$$

Then, $Y_n^{\theta_0}$ is a correct model of $y_n$. Moreover $T_{q_0}(Y_n^{\theta_0})$ is the IID model. ■

Now, generally, by theorem 3 one can find models correct $Y_n^{\theta_c}$ such that $P\{(X_1^{\theta_c}, ..., X_N^{\theta_c}) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon']$ where $\epsilon'$ is increasingly small if q decreases when $K_0$ is not too large.

Then, we shall understand in section 10 that it is possible to build in practical terms a such sequence $y_n$, i.e. a sequence $y_n$ such that the model IID is a correct model of $x_n$.

# 6 Approximation theorem

## 6.1 Theorem

In this section, we assume that T is a Fibonacci congruence and we use Fibonacci function $T_q$ in order to build IID sequences.

**Notations 6.1** *Let $Y_n \in \{0/m, 1/m, ...., (m-1)/m\}$, n=1,2,...,N, be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$. We define the sequence $X_n$, n=1,2,...,N, by $X_n = T_q(Y_n)$*

Now we define a measure equivalent to the Borel measure in the discrete case.

**Notations 6.2** *For all $p \in \mathbb{N}^*$, let L be the measure defined on $\{0/2^q, 1/2^q, ...., (2^q-1)/2^q\}^p$ by $L(Bo) = \frac{Card(\Theta)}{2^{pq}}$ when $Bo = \cup_{(k_1,...,k_p)\in\Theta}\{(k_1/2^q, ..., k_p/2^q)\}$.*

For example, if p=1, and if $I = \{k/2^q, (k+1)/2^q, ....., (k'-1)/2^q\}$, $L(I) = (k'-k)/2^q$ the length of interval $[k/2^q, k'/2^q[$.

Because $Y_n$ is a sequence with values in a discrete space, it always admits a density with respect to the discrete uniform measure.

**Notations 6.3** *We denote by $\mu_m$ the uniform measure defined on $\{0/m, 1/m, ...., (m-1)/m\}$ by $\mu_m(k/m) = 1/m$ for all $k \in \{0, 1, ..., m-1\}$.*

*For all permutation $\phi$ of $\{1, 2, ..., N\}$, for all $n \in \{1, 2, ..., N\}$, we denote by $f_{n,\phi}(.|y'_1, y'_2, ....)$ the conditional density with respect to $\mu_m$ of $Y_{\phi(n)}$ given $Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, .....$*

Since $Y_n$ is discrete, we can also assume that $f_{n,\phi}(.|y'_1, y'_2, ....)$ has a finite Lipschitz coefficient.

**Notations 6.4** *We denote by $K_0$ a constant such that, for all permutation $\phi$ of $\{1, 2, ..., N\}$, for all $n \in \{1, 2, ..., N\}$, $|f_{n,\phi}(y|y'_1, y'_2, ...) - f_{n,\phi}(y'|y'_1, y'_2, ...)| \leq K_0|y - y'|$. In order to simplify the proofs we suppose $K_0 > 1$.*

Under these conditions, if m and q are well chosen, $X_n$ is approximately the IID sequence : for all Borel set $Bo \subset \{0/2^q, 1/2^q, ...., (2^q-1)/2^q\}^N$, $P\big\{(X_1, ...., X_N) \in Bo\big\} = L(Bo)[1 + Ob(1)\epsilon]$, where $\epsilon \approx 0$.

**Theorem 3** *Let $\gamma(m) = [2 + \varphi(m)]$. We assume $\gamma(m)NK_0 2^q/m \approx 0$ and $m/K_0 >> 1$. Then, for all Borel set Bo,*

$$P\big\{(X_1, ...., X_N) \in Bo\big\} = L(Bo)\Big[1 + \frac{\gamma(m)Ob'(1)NK_0}{m/2^q}\Big] \ .$$

*where $|Ob'(1)|$ is increased by a number close to 1.*

If $K_0$ is not too large, there is no difficulty to choose m and q in such a way that $\epsilon \leq \gamma(m)2^q NK_0/m$ is small enough. Therefore, $P\big\{(X_1, ...., X_N) \in Bo\big\} = L(Bo)\big[1 + Ob'(1)\epsilon\big]$.

## 6.2 First proposition

In order to prove theorem 3, we shall use the following proposition.

**Proposition 6.1** *Let $h_N$ be the probability density function of $Y \in \{0/m, 1/m, ....., (m-1)/m\}$, with respect to $\mu_m$ : $\int_0^1 h_N(u)\mu_m(du) = 1$. Let $h'_N = (1/c_0)h_N$ such that $\int_0^1 h'_N(u)du = 1$.*

*Let $K_0 \in \mathbb{R}_+$ such that $|h_N(r) - h_N(r')| \leq K_0|r' - r|$ and $|h'_N(r) - h'_N(r')| \leq K_0|r' - r|$ when $r, r' \in [0, 1]$. One supposes $2^q/m \approx 0$, and $m/K_0 >> 1$ .*

*Then, the following equality holds :*

$$P\{\overline{T}(mY)/m \in I_k\} = L(I_k)\Big[1 + \frac{\gamma(m)Ob'(1)K_0}{m/2^q}\Big] \ ,$$

*where $I_k = [k/2^q, (k+1)/2^q[$, $L(I_k) = 1/2^q$.*

**Proof** The proof of this proposition is simple : the points of $\overline{T}^{-1}(mI_k)$ are well distributed in $\{0, 1, ...., m-1\}$. Thus in figure 7, it is easy to understand that the sum of points of $h'_N(\overline{T}^{-1}(mI_k))$ will be close $card\big(mI_k \cap \{0, 1, ...., m-1\}\big)/m$ because $\int_0^1 h'_N(u)du = 1$.

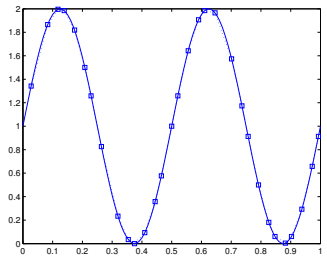

Figure 7: Points of $h'_N[\overline{T}^{-1}(mI_k)]$ when $h'_N(t) = sin(4\pi t) + 1$

Now in order to prove mathematicaly this proposition, we need the following lemmas.

**Lemma 6.1** *Let $N(I_k)$ be the number of $t/m \in \{0/m, 1/m, ...., (m-1)/m\}$ such that $k/2^q \leq t/m < (k+1)/d^q$. Then, $N(I_k) = m/2^q + Ob(1)$.*

**Proof** We define the interval $[c_k/m, c'_k/m[$ with $c_k, c'_k \in \{0, 1, ...., m-1\}$ by

$$[c_k/m, c'_k/m[\cap\{0/m, 1/m, ...., (m-1)/m\} = [k/2^q, (k+1)/2^q[\cap\{0/m, 1/m, ...., (m-1)/m\}.$$

Then, $(c_k - 1)/m < k/2^q \leq c_k/m$ and $(c'_k - 1)/m < (k+1)/2^q \leq c'_k/m$.

Let $m/2^q = h_0 + e$ where $0 \leq e < 1$ and $h_0 \in \mathbb{N}$. Then, $N(I_k) = c'_k - c_k = h_0$ or $N(I_k) = h_0 + 1$.

By our definition $h_0 \leq m/2^q \leq (h_0 + 1)$. Then, $N(I_k) = m/2^q + Ob(1)$. ■

**Lemma 6.2** *The following equality holds :*

$$c_0 = 1 + \frac{Ob'(1)K_0}{m} \ .$$

**Proof** The following equalities hold :

$$1 = \sum_t \int_{t/m}^{(t+1)/m} h'_N(u)du = \sum_t \int_{t/m}^{(t+1)/m} \big[h'_N(t/m) + Ob(1)K_0/m\big]du$$

$$= \frac{1}{m}\sum_t h'_N(t/m) + \frac{Ob(1)K_0}{m} = \int_0^1 h'_N(u)\mu_m(du) + \frac{Ob(1)K_0}{m} \ .$$

Then, $\int_0^1 h'_N(u)\mu_m(du) = 1 + \frac{Ob(1)K_0}{m}$. Therefore,

$$1 = \int_0^1 h_N(u)\mu_m(du) = c_0 \int_0^1 h'_N(u)\mu_m(du) = c_0\Big[1 + \frac{Ob(1)K_0}{m}\Big] \ . \ \blacksquare$$

**Lemma 6.3** *The following equality holds :* $\frac{1}{N(I_k)}\sum_r h_N(r/N(I_k)) = 1 + \frac{Ob'(1)K_0}{N(I_k)}$ .

**Proof** The following equalities hold :

$$1 = \sum_r \int_{r/N(I_k)}^{(r+1)/N(I_k)} h'_N(u)du = \sum_r \int_{r/N(I_k)}^{(r+1)/N(I_k)} \big[h'_N(r/N(I_k)) + Ob(1)K_0/N(I_k)\big]du$$

$$= \frac{1}{N(I_k)}\sum_r h'_N(r/N(I_k)) + \frac{Ob(1)K_0}{N(I_k)} \ .$$

Therefore $c_0 = \frac{1}{N(I_k)}\sum_r h_N(r/N(I_k)) + \frac{Ob(1)c_0K_0}{N(I_k)}$.

Therefore, by lemma 6.2,

$$c_0 = 1 + \frac{Ob'(1)K_0}{m} = \frac{1}{N(I_k)}\sum_r h_N(r/N(I_k)) + \frac{Ob(1)[1 + \frac{Ob'(1)K_0}{m}]K_0}{N(I_k)} \ .$$

Because $K_0/m \approx 0$ and $N(I_k)/m \approx 0$, we deduce the lemma. $\blacksquare$

**Lemma 6.4** *Let* $g_N(k) = h_N\big(\overline{T}^{-1}(k)/m\big)$. *The following approximation holds*

$$\frac{1}{N(I_k)}\sum_{k=c_k}^{c'_k-1} g_N(k) = 1 + \frac{[1+\varphi(m)]Ob(1)K_0}{N(I_k)} \ .$$

**Proof** Let $k^n$, $n = 1, 2, .., c'_k - c_k$, be a permutation of $\{c_k/m, (c+1)/m, ..., (c'_k-1)/m\}$ such that $\overline{T}^{-1}(k^1) < \overline{T}^{-1}(k^2) < \overline{T}^{-1}(k^3) < ..... < \overline{T}^{-1}(k^{c'_k-c_k})$. Then, by definition of section 2.3,

$$|\overline{T}^{-1}(k^r)/m - r/N(I_k)| \leq \varphi(m)/N(I_k) \ .$$

We deduce that $|g_N(k^r) - h_N(r/N(I_k))| \leq K_0\varphi(m)/N(I_k)$.

Therefore, by lemma 6.3,

$$\frac{1}{N(I_k)}\sum_{k=c_k}^{c'_k-1} g_N(k) = \frac{1}{N(I_k)}\sum_r g_N(k^r)$$

$$= \frac{1}{N(I_k)}\sum_r h_N(r/N(I_k)) + \frac{1}{N(I_k)}\sum_r \big[g_N(k^r) - h_N(r/N(I_k))\big]$$

$$= \frac{1}{N(I_k)} \sum_r h_N(r/N(I_k)) + \frac{Ob(1)\varphi(m)K_0}{N(I_k)} = 1 + \frac{Ob'(1)K_0}{N(I_k)} + \frac{Ob(1)\varphi(m)K_0}{N(I_k)} \ . \blacksquare$$

**Proof of proposition 6.1** By the previous equalities,

$$P\{\overline{T}(Y)/m \in I_k\} = \frac{1}{m} \sum_k g_N(k) = \frac{N(I_k)}{m} \Big[1 + \frac{[1+\varphi(m)]Ob'(1)K_0}{N(I_k)}\Big]$$

$$= \Big[L(I_k) + \frac{Ob(1)}{m}\Big]\Big[1 + \frac{[1+\varphi(m)]Ob'(1)K_0}{N(I_k)}\Big] = L(I_k)\Big[1 + \frac{Ob(1)}{mL(I_k)}\Big]\Big[1 + \frac{[1+\varphi(m)]Ob'(1)K_0}{N(I_k)}\Big]$$

$$= L(I_k)\Big[1 + \frac{2^q Ob(1)}{m}\Big]\Big[1 + \frac{[1+\varphi(m)]Ob'(1)K_0}{m/2^q + Ob(1)}\Big]$$

$$= L(I_k)\Big[1 + \frac{2^q Ob(1)}{m}\Big]\Big[1 + \frac{2^q[1+\varphi(m)]Ob'(1)K_0}{m[1 + Ob(1)2^q/m]}\Big]$$

$$= L(I_k)\Big[1 + \frac{2^q[2+\varphi(m)]Ob'(1)K_0}{m}\Big] \ . \blacksquare$$

## 6.3 Other propositions

**Proposition 6.2** *Let $X_n$, n=1,2,...,N, be a sequence of random variables. Assume that, for all $p \in N^*$, for all sequence $x_s$, s=1,...,p, for all $n \in N^*$, for all sequence of intervals, $J_s$, s=1,2,..,p, for all injective sequence $j_s$, s=1,2,...,p, such that $j_1 = 0$ and $j_s + n \in \{1, 2, ..., N\}$,*

$$P\big\{X_{n+j_1} \in J_s | X_{n+j_2} = x_2, ...., X_{n+j_p} = x_p\big\} = L(J_1) + Ob(1)\epsilon \ .$$

*Then, for all injective sequence $j_s \in \mathbb{Z}$ such that $j_1 = 0$ ,*

$$P\Big\{\{X_{n+j_1} \in J_1\} \cap ..... \cap \{X_{n+j_p} \in J_p\}\Big\} = \big[L(J_1) + Ob(1)\epsilon\big].....\big[L(J_p) + Ob(1)\epsilon\big] \ .$$

**Proof** Let Q be the distribution of $(X_{n+j_1}, X_{n+j_2}, ...., X_{n+j_p})$ and let $Q^-$ be the distribution of $(X_{n+j_2}, ...., X_{n+j_p})$. Let $Q(.|x_2, ...., x_p)$ be the distribution of $X_{n+j_1}$ given $X_{n+j_s} = x_s$, for s=1,2,...,p.

Then,

$$P\Big\{\{X_{n+j_1} \in J_1\} \cap ..... \cap \{X_{n+j_p} \in J_p\}\Big\} = \int \mathbb{1}_{J_1}(x_1).....\mathbb{1}_{J_p}(x_p) Q(dx_1, ..., dx_p)$$

$$= \int \Big( \int \mathbb{1}_{J_1}(x_1) Q(dx_1 | x_2, ...., x_p)\Big) \mathbb{1}_{J_2}(x_2).....\mathbb{1}_{J_p}(x_p) Q^-(dx_2, ..., dx_p)$$

$$= \int P\big\{X_{n+j_1} \in J_1 | X_{n+j_2} = x_2, ...., X_{n+j_p} = x_p\big\} \mathbb{1}_{J_2}(x_2).....\mathbb{1}_{J_p}(x_p) Q^-(dx_2, ..., dx_p)$$

$$= L(J_1) \int \mathbb{1}_{J_2}(x_2).....\mathbb{1}_{J_p}(x_p) Q^-(dx_2, ..., dx_p) + \int Ob(1)\epsilon \mathbb{1}_{J_2}(x_2).....\mathbb{1}_{J_p}(x_p) Q^-(dx_2, ..., dx_p)$$

$$= \big(L(J_1) + Ob(1)\epsilon\big) P\Big\{\{X_{n+j_2} \in J_2\} \cap ..... \cap \{X_{n+j_p} \in J_p\}\Big\} \ .$$

Then, we prove the proposition by recurence. $\blacksquare$

**Proposition 6.3** *Let $Y_n \in \{0/m, 1/m, ...., (m-1)/m\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ and let $X_n = T_q(Y_n)$. Then, for all Borel set Bo,*

$$P\{X_n \in Bo \mid X_{n-s} = x_s, s = 1, 2, ..., p\}$$

$$= \sum_{y_{s_1} \in T_q^{-1}(x_1)} .... \sum_{y_{s_p} \in T_q^{-1}(x_p)} \eta_{y_{s_1},...,y_{s_p}} P\{X_n \in Bo | Y_{n-j} = y_{s_j}, j = 1, 2, ..., p\}$$

*where*

$$\sum_{y_{s_1} \in T_q^{-1}(x_1)} .... \sum_{y_{s_p} \in T_q^{-1}(x_p)} \eta_{y_{s_1},...,y_{s_p}} = 1 \ .$$

**Proof** We have :

$$P\{X_n \in Bo | X_{n-s} = x_s, s = 1, 2, ..., p\}$$

$$= \frac{P\big\{\{X_n \in Bo\} \cap \{X_{n-1} = x_1\} \cap .... \cap \{X_{n-p} = x_p\}\big\}}{P\big\{\{X_{n-1} = x_1\} \cap .... \cap \{X_{n-p} = x_p\}\big\}}$$

$$= \frac{P\Big\{\{X_n \in Bo\} \cap \big\{ \cup_{y_{s_1}} \{Y_{n-1} = y_{s_1}\}\big\} \cap .... \cap \big\{ \cup_{y_{s_p}} \{Y_{n-p} = y_{s_p}\}\big\}\Big\}}{P\Big\{\big\{ \cup_{y_{s_1}} \{Y_{n-1} = y_{s_1}\}\big\} \cap .... \cap \big\{ \cup_{y_{s_p}} \{Y_{n-p} = y_{s_p}\}\big\}\Big\}}$$

where $\cup_{y_{s_t}} \{Y_{n-t} = y_{s_t}\} = \cup_{y_{s_t} \in T_q^{-1}(x_t)} \{Y_{n-t} = y_{s_t}\}$.

Then,

$$P\{X_n \in Bo | X_{n-s} = x_s, s = 1, 2, ..., p\}$$

$$= \frac{P\Big\{ \cup_{y_{s_1}} .... \cup_{y_{s_p}} \{X_n \in Bo\} \cap \{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}{P\Big\{ \cup_{y_{s_1}} .... \cup_{y_{s_p}} \{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}$$

$$= \frac{\sum_{y_{s_1}} .... \sum_{y_{s_p}} P\Big\{\{X_n \in Bo\} \cap \{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}{\sum_{y_{s_1}} .... \sum_{y_{s_p}} P\Big\{\{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}$$

$$= \sum_{y_{s_1}} .... \sum_{y_{s_p}} \eta_{y_{s_1},...,y_{s_p}} \frac{P\Big\{\{X_n \in Bo\} \cap \{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}{P\Big\{\{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}$$

where

$$\eta_{y_{s_1},...,y_{s_p}} = \frac{P\Big\{\{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}}{\sum_{y_{s_1}} .... \sum_{y_{s_p}} P\Big\{\{Y_{n-1} = y_{s_1}\} \cap .... \cap \{Y_{n-p} = y_{s_p}\}\Big\}} \ .$$

Of course,

$$\sum_{y_{s_1}} .... \sum_{y_{s_p}} \eta_{y_{s_1},...,y_{s_p}} = 1 \ . \ \blacksquare$$

## 6.4 Proof of theorem 3

We apply proposition 6.1 when $Y$ has for distribution the distribution of the conditional probability of $Y_{\phi(n)}$ given $Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, .....$ Let $K'_0$ which satisfies the condition of Lipschitz coefficient of proposition 6.1. By lemma 6.2, $K'_0 = Ob'(1)K_0$. Then, we have

$$P\big\{\overline{T}(mY)/m \in I_k\big\} = P\big\{T_q(Y) = k/2^q\big\}$$

$$= P\big\{X_{\phi(n)} = k/2^q \;\big|\; Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, ....\big\}$$

$$= P\big\{X_{\phi(n)} \in I_k \;\big|\; Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, ....\big\}$$

$$= L(I_k)\Big[1 + \frac{\gamma(m)Ob'(1)K'_0}{m/2^q}\Big]$$

$$= L(I_k)\Big[1 + \frac{\gamma(m)Ob'(1)K_0}{m/2^q}\Big] ,$$

where $L(I_k) = 1/2^q$.

By applying proposition 6.3 to the sequence $Y_{\phi(n-s)}$,

$$P\big\{X_{\phi(n)} \in I_k \;\big|\; X_{\phi(n-1)} = x'_1, X_{\phi(n-2)} = x'_2, ....\big\}$$

$$= \sum_{y'_{s_1} \in T_q^{-1}(x'_1)} ..... \sum_{y'_{s_{N-1}} \in T_q^{-1}(x'_{N-1})} \eta_{y'_{s_1},...,y'_{s_{N-1}}} P\big\{X_{\phi(n)} \in I_k \;\big|\; Y_{\phi(n-1)} = y'_{s_1}, Y_{\phi(n-2)} = y'_{s_2}, ....\big\}$$

$$= \sum_{y'_{s_1} \in T_q^{-1}(x'_1)} ..... \sum_{y'_{s_{N-1}} \in T_q^{-1}(x'_{N-1})} \eta_{y'_{s_1},...,y'_{s_{N-1}}} L(I_k)\Big[1 + \frac{\gamma(m)Ob'(1)K_0}{m/2^q}\Big]$$

$$= L(I_k)\Big[1 + \frac{\gamma(m)Ob'(1)K_0}{m/2^q}\Big] .$$

Then, by proposition 6.2 used with $\epsilon = L(I_k)\epsilon$, for all $I_{k_1} \otimes ..... \otimes I_{k_N}$,

$$P\big\{(X_1, ...., X_N) \in I_{k_1} \otimes ..... \otimes I_{k_N}\big\} = \prod_{s=1}^{N} \Big(L(I_{k_s})[1 + Ob(1)\epsilon]\Big) ,$$

where $|\epsilon| \leq \frac{\gamma(m)|Ob'(1)|K_0}{m/2^q}$. Because $\gamma(m)NK_0 2^q/m \approx 0$, we deduce that

$$P\big\{(X_1^\theta, ...., X_N^\theta) \in I_{k_1} \otimes ..... \otimes I_{k_N}\big\} = \frac{1}{2^{Nq}}\Big[1 + \frac{\gamma(m)Ob'(1)NK_0}{m/2^q}\Big] .$$

Now, we study the Borel sets of $\{0/2^q, , ...., (2^q-1)/2^q\}^N$ : $Bo = \cup_{(k_1,...,k_N)\in\Theta}\{(k_1/2^q, ..., k_N/2^q)\}$. Then, $L(Bo) = Card(\Theta)/2^{Nq}$. We deduce, that, for all Borel set Bo

$$P\big\{(X_1, ...., X_N) \in Bo\big\} = L(Bo)\Big[1 + \frac{\gamma(m)Ob'(1)NK_0}{m/2^q}\Big] . \blacksquare$$

Then, by using results of section 5, because $\gamma(m)NK_0 2^q/m \approx 0$, $X_n$ cannot be differentiated with the IID model.

# 7 Use of text witten backward

There exists noises $y_n$ such that $K_0$ is not too large and m large enough. In order to obtain these noises, one can use texts. This choice is justified because it is easier to study their properties logically (cf section 10.3 of [9]) : For example asymptotical independence holds. Of course, one could use other random noises, for example, noises provided by machines. One could also use rap music as Marsaglia but its properties are more difficult to study logically.

## 7.1 Use of texts

Now, we suppose that we use sequences $y_n$ obtained from texts. As a matter of fact, in this section we define $y_n$ by the following way.

We choose two consecutive elements a and m of the Fibonacci sequence : m can be chosen with respect to N, the size of the sample. Then, we choose $r_1$ such that $a < 32^{r_1} \leq m$.

It is supposed that one has a sequence of data a(j) obtained from texts and translated in number: a(j), $j = 1, 2, ...., N_3$, $a(j) \in \{0, 1, ..., 255\}$. Let $N_0 = \lfloor N_3/r_1 \rfloor$, the integer part of $N_3/r_1$.
a) We set $c(j) = \overline{a(j)} \bmod \kappa = 32$ [5].
b) We set $d(n) = \sum_{r=1}^{r_1} c(r_1(n-1)+r)\kappa^{r-1}$ for $j = 1, 2, ...., N_0$.
c) We set $y'_n = \left\lfloor d(n)m/\kappa^{r_1} \right\rfloor /m$ for $j = 1, 2, ...., N_0$.

## 7.2 Use of a pseudo-random sequence

Moreover, a pseudo-random sequence $rand_0(n)$ is added to used texts : $y_n = \overline{my'(n) + rand_0(n)}/m$. That makes possible to have sequences $y_n$ which have a good randomness (cf [15], or chapter 3 of [9]).

Now, it is necessary that a priori all the possible values of $\{0/m, 1/m, ...., (m-1)/m\}$ can exist in a sample. It is reasonably the case when one adds modulo m a pseudo-random sequence $rand_0(n)$ of period m. Normally any value k/m has a chance reasonable to be realized a priori. There is no reason that can not occur. Moreover, a priori all k/m has about as much chance to be an image than any other k/m. Therefore, a priori, " $P\{Y'_n = y\}$ is not too different from 1/m " is a reasonable assumption. Now if we use simulations, they confirm this result. While, it is always possible that this is not the case. But it has a weak probability to happen.

Recall also that, for the texts, as soon as one takes as sequence $y'_n$ a sequence of group of Q=10 or 20 letters for example, one finds the Q-dependence statistically (chapter 10 of [9]).

**Now, we suppose that $(rand_0(n), rand_0(n+1))$ has a distribution close to independence.** So, normally this will be the case also for $(\overline{my'(n) + rand_0(n)}\ ,\ \overline{my'(n+1) + rand_0(n+1)})$. This can be understood by simulation. But a priori, it is always possible that this is not checked with, it seems, a very weak probability [6].

In this case, a two-dimensional model $(Y_n, Y_{n+1})$ with a continuous density and a Lipschitz coefficient not too big will be a good model. By the same way, $P\{Y_n = y | Y_{n-1} = y_1\}$ will have a continuous density with a coefficient of Lipschitz $K_{y_1}$ checking $K_{y_1} \leq K"_0$ for $y_1 = 0, 1, ...., m-1$ where $K"_0$ is small . Therefore, $P\{Y_n = y | Y_{n-1} = y_1\}$ is not too different of $P\{Y_n = y\}$ which is not too different from 1/m. Then, it is normal to accept this hypothesis for sequences $Y_n$ [7].

[5] There are only 26 letters. But it is necessary to add the capital letters, the ":" , ";" , etc. Also, we will write these numbers in base 32 so that each number has a reasonable probability to appear.

[6] If one wants to build random numbers, one can always check if this hypothesis holds. If this is not the case, we choose other generators or other texts

[7] Of course, if we want to be sure from it, we can confirm it by tests.

## 7.3 Text written backward

In an obvious way, the texts are realizations of sequences of random variables: for example, one can take as model, the set of the possible texts provided with the uniform probability. In this model, if one knows a text until the letter "n-1", there are a large number of alternatives for the r following letters as soon as r is rather large. That means indeed that the conditional probability of $Y_n$ knowing the past, is not concentrated in a too small number of points.

However there is a problem for some subsequences $y'_{\phi(n)}$ : if one knows a text until the letter "n-1" and the text after the letter "n+r", (for example r=18), there will be much less possibilities for the r letters ranging between the two parts of texts than if only the past is known. To answer this point, we will add modulo m a text and a text written backward.

But that seems exaggerated because it is not known a priori that we are in an English text if one has only a few texts [8]. Moreover, a priori all the words possible of the English language are not known : one cannot thus predict them. That does not prevent from concluding : if the conditional probabilities of the texts are not concentrated in some points in a model of English text, a fortiori, it is also the case if it is not known that one is in a English text.

Now, it is encore easier to prove that the conditional probability of $Y_n$ knowing the past, is not concentrated in a too small number of points if $y_n = [\overline{my'(n) + rand_0(n) + my"(n) + rand_1(n)}]/m$ where y"(n) represent a text written backward independent of $y'_n$ and $rand_j(n)$ pseudo-random sequences for j=0,1 (which have good empirical independence assumptions for p successive pseudo random numbers with $p \geq 3$). In this case, one can show that this condition is correct.

## 7.4 Theorem

Indeed, now we suppose that the sequences $x_n$ and $y_n$ represent two independent texts at which one adds to each one a good pseudo-random sequences. Let $Y_n$ and $X_n$ be two correct models. One is interested to the sequence $\overline{X_{n+s} + Y_{n-s}}$ , $s = 0, \pm 1, \pm 2, .....$ As matter of fact, one adds a text to a text written backward

Then, we will understand that the probability that $\overline{X_n + Y_n} = a_0$ given $\overline{X_{n+s} + Y_{n-s}} = a_s$ for s=1,-1, will be about that of $\overline{X_n + Y_n} = a_0$ given $X_{n-1} = b_1$ and $Y_{n-1} = c_1$.

At first, we have the following theorem

**Theorem 4** *Let $Y_n$ and $X_n$ be two independent sequences of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$ such that $X_n,\ Y_n \in \{0/m, 1/m, ...., (m-1)/m\}$. Then,*

$$P\{X_n + Y_n \equiv a_0 \mid X_{n-1} + Y_{n+1} \equiv a_1, X_{n+1} + Y_{n-1} \equiv a_2\}$$
$$= \sum_{x_1, y_1} \eta_{x_1,y_1} \alpha_{x_1,y_1} P\big\{X_n + Y_n \equiv a_0 \ \big|\ Y_{n+1} \equiv a_1 - x_1, X_{n+1} \equiv a_2 - y_1\big\} \ ,$$

*where*

$$\alpha_{x_1,y_1} = \frac{P\big\{X_{n-1} = x_1, Y_{n-1} = y_1 \ \big|\ X_n + Y_n \equiv a_0, Y_{n+1} \equiv a_1 - x_1, X_{n+1} \equiv a_2 - y_1\big\}}{P\big\{X_{n-1} = x_1, Y_{n-1} = y_1 \ \big|\ Y_{n+1} \equiv a_1 - x_1, X_{n+1} \equiv a_2 - y_1\big\}} \ ,$$

$$\eta_{x_1,y_1} = \frac{P\big\{\{X_{n-1} = x_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\big\}\ P\big\{\{Y_{n+1} \equiv a_1 - x_1\} \cap \{Y_{n-1} = y_1\}\big\}}{\sum_{x_1,y_1} P\big\{\{X_{n-1} = x_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\big\}\ P\big\{\{Y_{n+1} \equiv a_1 - x_1\} \cap \{Y_{n-1} = y_1\}\big\}} \ ,$$

$\sum_{x_1,y_1} = \sum_{(x_1,y_1) \in \{0/m, 1/m, ...., (m-1)/m\}^2}$ , $\sum_{x_1,y_1} \eta_{x_1,y_1} = 1$.

[8] Let us recall difficulties in order to discover the meaning of certain languages in archeology : all are not identifed. Let us recall also the hieroglyphs on the Rosetta Stone whose one had however 3 translations.

**Proof** We have

$$P\{X_n + Y_n \equiv a_0 \mid X_{n-1} + Y_{n+1} \equiv a_1, X_{n+1} + Y_{n-1} \equiv a_2\}$$

$$= \frac{P\big\{\{X_n + Y_n \equiv a_0\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\big\}}{P\big\{\{X_{n-1} + Y_{n+1} \equiv a_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\big\}}$$

$$= \frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \Big\{\cup_{x_1}\, \{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\cup_{y_1}\, \{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}{P\Big\{\Big\{\cup_{x_1}\, \{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\cup_{y_1}\, \{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}$$

$$= \sum_{x_1, y_1} \frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}{\sum_{x_1, y_1} P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}$$

$$= \sum_{x_1, y_1} \frac{P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}{\sum_{x_1, y_1} P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}$$

$$\frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}{P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{X_{n-1} + Y_{n+1} \equiv a_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} + Y_{n-1} \equiv a_2\}\Big\}\Big\}}$$

$$= \sum_{x_1, y_1} \eta_{x_1, y_1} \frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \Big\{\{X_{n-1} = x_1\} \cap \{Y_{n+1} \equiv a_1 - x_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}\Big\}}{P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{Y_{n+1} \equiv a_1 - x_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}\Big\}}$$

$$= \sum_{x_1, y_1} \eta_{x_1, y_1} P\Big\{X_n + Y_n \equiv a_0 \;\Big|\; X_{n-1} = x_1, Y_{n+1} \equiv a_1 - x_1, Y_{n-1} = y_1, X_{n+1} \equiv a_2 - y_1\Big\}\,.$$

On the other hand,

$$P\Big\{X_n + Y_n \equiv a_0 \;\Big|\; X_{n-1} = x_1, Y_{n+1} \equiv a_1 - x_1, Y_{n-1} = y_1, X_{n+1} \equiv a_2 - y_1\Big\}$$

$$= Cx_1, y_1 P\big\{X_n + Y_n \equiv a_0 \big| Y_{n+1} \equiv a_1 - x_1, X_{n+1} \equiv a_2 - y_1\big\}\,,$$

where

$$c_{x_1, y_1} = \frac{P\Big\{X_n + Y_n \equiv a_0 \;\Big|\; X_{n-1} = x_1, Y_{n+1} \equiv a_1 - x_1, Y_{n-1} = y_1, X_{n+1} \equiv a_2 - y_1\Big\}}{P\Big\{X_n + Y_n \equiv a_0 \;\Big|\; Y_{n+1} \equiv a_1 - x_1, X_{n+1} \equiv a_2 - y_1\Big\}}$$

$$= \frac{\frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \Big\{\{X_{n-1} = x_1\} \cap \{Y_{n+1} \equiv a_1 - x_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}\Big\}}{P\Big\{\Big\{\{X_{n-1} = x_1\} \cap \{Y_{n+1} \equiv a_1 - x_1\}\Big\} \cap \Big\{\{Y_{n-1} = y_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}\Big\}}}{\frac{P\Big\{\{X_n + Y_n \equiv a_0\} \cap \{Y_{n+1} \equiv a_1 - x_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}}{P\Big\{\{Y_{n+1} \equiv a_1 - x_1\} \cap \{X_{n+1} \equiv a_2 - y_1\}\Big\}}}$$

$$= \frac{\frac{P\Big\{\{X_n+Y_n\equiv a_0\}\cap\Big\{\{X_{n-1}=x_1\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}\cap\Big\{\{Y_{n-1}=y_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}\Big\}}{P\Big\{\{X_n+Y_n\equiv a_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}}{\frac{P\Big\{\Big\{\{X_{n-1}=x_1\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}\cap\Big\{\{Y_{n-1}=y_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}\Big\}}{P\Big\{\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}}$$

$$= \frac{P\Big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\Big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\Big\}}{P\Big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\Big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\Big\}} \;.\; \blacksquare$$

## 7.5 Application

Let us suppose again that the sequences $x_n$ and $y_n$ represents texts at which one adds to each one a good pseudo-random sequence. It is supposed that $Y_n$ and $X_n$ are two correct models. One is interested by $\overline{X_{n+s}+Y_{n-s}}$ , $s=0,\pm1,\pm2,....$ : one adds a text and a text written backward.

**Study of** $P\big\{X_n+Y_n\equiv a_0 \;\big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ We know that $P\big\{X_n+Y_n\equiv a_0 \;\big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ is the conditional probability that $X_n+Y_n\equiv a_0$ given the futures $Y_{n+1}$ and $X_{n+1}$.

There will be thus a probability which will not be more concentrated that of a text knowing the future. But it is an increase: the probability of the sum $\overline{X_n+Y_n}$ knowing the future $Y_{n+1}\equiv a_1-x_1$ and $X_{n+1}\equiv a_2-y_1$ is probably less concentrated than, for example, the probability of $X_n$ knowing the future $X_{n+1}\equiv a_2-y_1$.

In fact, the conditional probability will be much less concentrated than that: it is not known that one is in a text. Moreover, because a good pseudo-random generator is added, this probability will be rather close to that of independence : $P\big\{X_n+Y_n\equiv a_0 \;\big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ is not too distant from $P\big\{X_n+Y_n\equiv a_0\big\}$ which is not too distant from 1/m.

Therefore, the probability of the sum $X_n+Y_n$ knowing the future is not concentrated close to some points. That means that there will be no points where it is close to 0, and not points where it is close to 1. That means that, in the case of models with continuous density, the coefficient of Lipschitz will not be too large.

**Study of** $P\{X_{n-1}=x_1, Y_{n-1}=y_1 \mid Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\}$**.** Now considering the independence of texts $X_n$ and $Y_n$, $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\} = P\big\{X_{n-1}=x_1 \;\big|\; X_{n+1}\equiv a_2-y_1\big\}P\big\{Y_{n-1}=y_1 \;\big|\; Y_{n+1}\equiv a_1-x_1\big\}$.

Therefore, $P\Big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\Big|\; Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\Big\} \approx 1/m^2$ if m is large enough.

**Study of** $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ **.** One understands, by simulation, that $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ is not too different from $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0\big\}$.

It is not astonishing: $X_{n-1}$ is almost independent of $X_{n+1}$. Therefore, $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ depends especially on $X_n+Y_n$ [9].

One can also understand it because of following relations

$$P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$$

$$=\frac{P\Big\{\{X_{n-1}=x_1\}\cap\{Y_{n-1}=y_1\}\cap\{X_n+Y_n\equiv a_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}{P\Big\{\{X_n+Y_n\equiv a_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}$$

$$=\frac{P\Big\{\{X_{n-1}=x_1\}\cap\{Y_{n-1}=y_1\}\cap\Big\{\cup_{x_0}\{X_n=x_0\}\cap\{X_n+Y_n\equiv a_0\Big\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}{P\Big\{\Big\{\cup_{x_0}\{X_n=x_0\}\cap\{X_n+Y_n\equiv a_0\Big\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}$$

$$=\sum_{x_0}\frac{P\Big\{\{X_{n-1}=x_1\}\cap\{Y_{n-1}=y_1\}\cap\Big\{\{X_n=x_0\}\cap\{Y_n\equiv a_0-x_0\}\Big\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}{\sum_{x_0}P\Big\{\Big\{\{X_n=x_0\}\cap\{Y_n\equiv a_0-x_0\}\Big\}\cap\{Y_{n+1}\equiv a_1-x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}}$$

$$=\sum_{x_0}\frac{P\Big\{\{X_{n-1}=x_1\}\cap\{X_n=x_0\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}P\Big\{\{Y_{n-1}=y_1\}\cap\{Y_n\equiv a_0-x_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}}{\sum_{x_0}P\Big\{\{X_n=x_0\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}P\Big\{\{Y_n\equiv a_0-x_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}}$$

$$=\sum_{x_0}\beta_{x_0}\frac{P\Big\{\{X_{n-1}=x_1\}\cap\{X_n=x_0\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}P\Big\{\{Y_{n-1}=y_1\}\cap\{Y_n\equiv a_0-x_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}}{P\Big\{\{X_n=x_0\}\cap\{X_{n+1}\equiv a_2-y_1\}\Big\}P\Big\{\{Y_n\equiv a_0-x_0\}\cap\{Y_{n+1}\equiv a_1-x_1\}\Big\}}$$

$$=\sum_{x_0}\beta_{x_0}P\big\{X_{n-1}=x_1|X_n=x_0, X_{n+1}\equiv a_2-y_1\big\}P\big\{Y_{n-1}=y_1|Y_n\equiv a_0-x_0, Y_{n+1}\equiv a_1-x_1\big\}$$

where $\sum_{x_0}\beta_{x_0}=1$.

It is not too difficult to understand, that, for example, $P\{X_{n-1}=x_1|X_n=x_0, X_{n+1}\equiv a_2-y_1\big\}$ is hardly more concentrated than $P\{X_{n-1}=x_1|X_n=x_0\big\}$ if $x_n$ represents only texts. It is even truer if $x_n$ represents a text to which one adds a good pseudo random sequence, and it is even truer in the case which interests us considering than one summons on all the $x_0$.

Then, it is not astonishing that $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0, Y_{n+1}\equiv a_1-x_1, X_{n+1}\equiv a_2-y_1\big\}$ is not too different from $P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0\big\}$.

Now, $P\{X_n+Y_n\equiv a_0\}\approx 1/m$ because one adds a pseudo random sequence to text (cf pages 199-202 of [9]). Therefore,

$$P\big\{X_{n-1}=x_1, Y_{n-1}=y_1 \;\big|\; X_n+Y_n\equiv a_0\big\}=\frac{P\{X_{n-1}=x_1, Y_{n-1}=y_1, X_n+Y_n\equiv a_0\}}{P\{X_n+Y_n\equiv a_0\}}$$

$$\approx m.P\big\{X_{n-1}=x_1, Y_{n-1}=y_1\big\}\frac{P\big\{X_{n-1}=x_1, Y_{n-1}=y_1, X_n+Y_n\equiv a_0\big\}}{P\big\{X_{n-1}=x_1, Y_{n-1}=y_1,\big\}}$$

[9] In the general case, that could be false : e.g. cf the properties of higher order correlation coefficients (cf [6])

$$= m.P\big\{X_{n-1}=x_1\big\}P\big\{Y_{n-1}=y_1\big\}P\big\{X_n+Y_n\equiv a_0\big|X_{n-1}=x_1,Y_{n-1}=y_1,\big\}$$
$$\approx (1/m)P\big\{X_n+Y_n\equiv a_0\big|X_{n-1}=x_1,Y_{n-1}=y_1,\big\}\ .$$

Of course, $P\big\{X_n+Y_n\equiv a_0\big|X_{n-1}=x_1,Y_{n-1}=y_1,\big\}$ is, this time, the conditional probability knowing the past. There are thus about the same results that above for $P\big\{X_n+Y_n\equiv a_0\ \big|\ Y_{n+1}\equiv a_1-x_1,X_{n+1}\equiv a_2-y_1\big\}$. Therefore, $P\big\{X_n+Y_n\equiv a_0\big|X_{n-1}=x_1,Y_{n-1}=y_1,\big\}$ will be not too different from 1/m.

**Conclusion** By joining together all these results, one understands that

$$\alpha_{x_1,y_1}P\big\{X_n+Y_n\equiv a_0\ \big|\ Y_{n+1}\equiv a_1-x_1,X_{n+1}\equiv a_2-y_1\big\}$$

will be not too different from 1/m.

Now,

$$\eta_{x_1,y_1}=\frac{P\big\{\{X_{n-1}=x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\big\}P\big\{\{Y_{n+1}\equiv a_1-x_1\}\cap\{Y_{n-1}=y_1\}\big\}}{\sum_{x_1,y_1}P\big\{\{X_{n-1}=x_1\}\cap\{X_{n+1}\equiv a_2-y_1\}\big\}P\big\{\{Y_{n+1}\equiv a_1-x_1\}\cap\{Y_{n-1}=y_1\}\big\}}$$
$$\approx\frac{P\{X_{n-1}=x_1\}P\{X_{n+1}\equiv a_2-y_1\}P\{Y_{n+1}\equiv a_1-x_1\}P\{Y_{n-1}=y_1\}}{\sum_{x_1,y_1}P\{X_{n-1}=x_1\}P\{X_{n+1}\equiv a_2-y_1\}P\{Y_{n+1}\equiv a_1-x_1\}P\{Y_{n-1}=y_1\}}$$
$$\approx\frac{1/m^4}{\sum_{x_1,y_1}(1/m^4)}\approx 1/m^2\ .$$

Therefore,

$$\sum_{x_1,y_1}\eta_{x_1,y_1}\alpha_{x_1,y_1}P\big\{X_n+Y_n\equiv a_0\ \big|\ Y_{n+1}\equiv a_1-x_1,X_{n+1}\equiv a_2-y_1\big\}$$

is not too different from

$$(1/m^2)\sum_{x_1,y_1}P\big\{X_n+Y_n\equiv a_0\ \big|\ Y_{n+1}\equiv a_1-x_1,X_{n+1}\equiv a_2-y_1\big\}\ .$$

However in general, to make a sum on $x_1,y_1$ standardizes the probabilities (it is true as soon as one can consider that they are randomly selected cf section 6.1.2 of of [10]). Therefore, in most case, $P\{X_n+Y_n\equiv a_0\mid X_{n-1}+Y_{n+1}\equiv a_1,X_{n+1}+Y_{n-1}\equiv a_2\}$ will be even more close to (1/m) that the previous reasonings which is carry out without the sums $\sum_{x_1,y_1}$ did not let it suppose.

Finally, all this confirms that $P\{X_n+Y_n\equiv a_0\mid X_{n-1}+Y_{n+1}\equiv a_1,X_{n+1}+Y_{n-1}\equiv a_2\}$ is not too different from 1/m. One deduces from it that the coefficient of Lipschitz will not be too large. Then, it is enough to apply $T_q$ in order to have sequences proved IID.

## 7.6 Increase of $K_0$

Similar results can be obtained for $P\{X_n+Y_n\equiv a_0\mid X_{n-1}+Y_{n+1}\equiv a_1,\ X_{n-2}+Y_{n+2}\equiv a_2,,....,X_{n+1}+Y_{n-1}\equiv b_1,X_{n+2}+Y_{n-2}\equiv b_2,.....\}$.

This is not surprising because when one tests the sequences of groups of letters $y'_n$, we understand that models $Y'_n$ behave like Q-dependent sequences and also like Markov chain as soons as $r_1\geq 20$. Of course, that is even more true for sequences $Y_n$.

This leads to the conclusion that we can increase the Lipschitz coefficient of this conditional probability by a $K_0$ which is not too large.

### 7.7 Important remark

One might wonder if the sequence built (by adding text, a text written backward and pseudo-random sequences) is not an lID sequence. It is a similar hypothesis which Marsaglia does by building its CD-Rom. It is a such tendency that matches the result of [15]. But in fact, nothing is proved.

One might then wonder if you can not apply to these sequences, the same technique as that used for sequences $T_q(Y_n)$ and to prove $P\big\{(X_1,....,X_N)\in Bo\big\}=L(Bo)\big[1+Ob(1)\epsilon\big]$. But it seems very difficult to prove in a sure way.

Finally, it is much easier to apply the functions $T_q$ : in this case, it requires only that $K_0$ the Lipschitz coefficient is not too big. It's an hypothesis much simpler to be verified and it does not require many efforts in some cases. That is why we choose to build IID sequences using this technique

## 8 Use of Central limit theorem

There exists other noises $y_n$ such that $K_0$ is not too large and m large enough. For example, one can use software programs noises provided by machines and chips, etc..

In these cases, on can use the CLT (Central Limit theorem). That is a possibility which is natural when one wants to increase reliably Lipschitz coefficient of conditional probabilities. Moreover, this technique has the advantage of giving a clear idea of $K_0$ since the conditional densities are so close to Gaussian densities and since convergence is quick (e.g. cf [9]). On the other hand, CLT holds under hypotheses rather weak if we use the following decomposition.

**Notations 8.1** *We denote by $\kappa(n)\in\mathbb{N}$, an increasing sequence such that $\kappa(1)=0$, $\kappa(n)\leq n$ and $\kappa(n)/n\to 0$ . We define the sequences u(n) and t(n) by : u(1)=1, $u(n)=max\big\{m\in\mathbb{N}^*\big|2m+\kappa(m)\leq n\big\}$, and t(1)=0, $t(n)=n$-$2u(n)$ if $n\geq 2$. Let $\sigma(u)^2=\mathbb{E}\{(X_1+X_2+.....+X_u)^2\}$. One sets $S_u=\frac{X_1+X_2+.....+X_u}{\sigma(u)}$, $\xi_u=\frac{X_{u+1}+X_{u+2}+.....+X_{u+t}}{\sigma(u)}$ and $S'_u=\frac{X_{u+t+1}+X_{u+t+2}+.....+X_{u+t+u}}{\sigma(u)}$.*

Then, one can define assumptions of asymptotic independence.

**Notations 8.2** *: Let $k\in\mathbb{N}^*$. We define conditions $H_{mS}(k)$ and $H_{mI}(k)$ by the following way :*
*$H_{mS}(k)$ : $\forall p\in\mathbb{N}$ , $p<k+1$ , $E\big\{(S_u)^p\big\}-E\big\{(S'_u)^p\big\}\to 0$ as $n\to\infty$.*
*$H_{mI}(2k)$ : $\forall(p,q)\in(\mathbb{N}^*)^2$ , $p+q<k+1$ , $E\big\{(S_u)^p(S'_u)^q\big\}-E\big\{(S_u)^p\big\}E\big\{(S'_u)^q\big\}\to 0$ as $n\to\infty$.*

**Notations 8.3** *: Let $I_{k,j}=\big[j.4^{-k},(j+1)4^{-k}\big[\big\}$. Let $A_{k,j}=\big\{S_u\in I_{k,j}\big\}$ and $B_{k,j}=\big\{S'_u\in I_{k,j}\big\}$. Then, we define condition $H_S$ and $H_I$ by the following way :*
*$H_S:\forall k\in\mathbb{N},\forall j\in\mathbb{N},P\{A_{k,j}\}-P\{B_{k,j}\}\to 0$ as $n\to\infty$ ,*
*$H_I:\forall k\in\mathbb{N},\forall(j,j')\in\mathbb{N}^2,P\{A_{k,j}\cap B_{k,j'}\}-P\{A_{k,j}\}P\{B_{k,j'}\}\to 0$ as $n\to\infty$ .*

Then, if $H_{mS}(\infty)$ and $H_{mI}(\infty)$ the CLT holds and an equivalent condition holds for $H_I$ and $H_S$ cf ([7], [8]). For example, the following theorem holds.

**Theorem 5** *We assume that $H_S$ , $H_I$ , $H_{mS}(4)$ and $H_{mI}(4)$ hold. We assume also that $\mathbb{E}\{(S_u)^2\}-\mathbb{E}\{(S'_u)^2\}\to 0$ and $\mathbb{E}\{\xi_u^2\}\to 0$ as $n\to\infty$ . Then, $S_n\xrightarrow{D}N(0,1)$ .*

We can then apply these results to random sequences obtained from machines, chips, various electronic files which have a certain asymptotic independence. More if one adds modulo m a good pseudo random, tests show that the conditions $H_{mS}(\infty)$ and $H_S(\infty)$ are checked.

We therefore use lines of noises $y_{i,n}\in\{0,1,....,m-1\}$ which we sum : $y_n=\sum_{i=1}^{S}y_{i,n}/m$ and we apply $T_q$. Indeed, in this case, the conditional densities are approximately Gaussian. We can then increase $K_0$ by estimating the coefficients of linear correlation. Thus we obtain coefficient $K_0$ whose we are sure and which are in general not too large.

Then, by applying $T_q$ and choosing good parameters depending on N, we obtain sequences $x_n$ which we can consider as a sample of an IID sequence of random variables

Now if we sum the random variables modulo m : $y_n = \overline{\sum_{i=1}^{S} y_{i,n}}/m$, we get even a better result because in this case, the $Y_n$ are asymptotically independent (cf section 5.2 of [9]). This improves again, if it was necessary, the quality of the increase of $K_0$.

# 9 Building of random sequence

## 9.1 Choice of m and q

We are interested in the choice of parameters in the building of IID sequences.

We consider the case of texts $y_n = \overline{[my'_n + rand_0(n) + my"_n + rand_1(n)]}/m$ defined in section 8. Then, we shall impose that the sample size $N_0$ satisfies $\frac{\gamma(m)N_0K_0}{m/2^q} = 1/1000$ (cf theorem 3 ). We could take $\frac{\gamma(m)N_0K_0}{m/2^q} = 1/10$ or $1/100$ without problem. We choose $1/1000$ in order to be sure that there will be absolutely no mistake.

Likewise, because the $y_n$ are obtained using the method described in section 8, we choose$K_0 = 1000$ although the calculations made in section 8 shows that we can probably choose it much smaller.

So finally, we choose m and q so that $2^q/m = \frac{1}{\gamma(m)N_0 10^6}$ .

Then, the particular choices of m and, therefore also of q, depends on questions of convenience. For example, is that the computer has a program to perform multiplication and division of numbers which have more than 30 digits? In this case, we can choose m of order of $10^{30}$ if the other conditions permit.

If we use the CLT, we proceed in exactly the same way except that we may normally choose $K_0$ smallest : $K_0 \leq 10$.

## 9.2 Example

By using this technique, we have created sequences $x_n$ which admit the IID model for correct model. We have used dictionary, encyclopedia, and Bible. As a matter of fact, we combine both methods : we are made sums of 10 lines including 5 written backwards [10]. We have estimated $K_0 = 0.01$. In order to avoid any error we have choose $K_0 = 10^4$ in the building of $x_n$.

One can download these real random sequences written in Matlab files in [13].

We have tested this sequence $x_n$. We have used the classical Diehard tests (cf [1], [2]), and the higher order correlation coefficients (cf [6]). Results are in accordance with what we waited : the hypothesis "randomness" is accepted by all these tests.

For example, we have used the Coupon collector's Test. We keep the notations of [1] page 64. We choose d=3,4,5,6,7,8 (with the notations of [1] ). Then, one uses chi squared statistics : we denote them by $\chi^2_{N_1}$. We use various t (with the notations of [1] ). We choose t as a function of d. We lump a few categories of low probability together.

We use samples with various sizes $N_1$. We are interested in the maximum of these various $\chi^2_{N_1}$.

The following table is that of the maxima of $\chi^2_{N_1}$ obtained for each d.

| d | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| $max(\chi^2_{N_1})$ | 11.02 | 13.59 | 18.08 | 16.73 | 20.08 | 22.84 |
| $\alpha_5$ | 11.07 | 12.59 | 14.07 | 15.51 | 16.92 | 18.31 |

[10] It's probably too much. But in [9], we demonstrated theorems less performing about the increase of conditional empirical probabilities. We obtained only $P\{Y_{\phi(n)} \in I_k | Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, .....\} = L(I_k) + Ob(1)\epsilon$. This means that the method employed in order to obtain these sequences is still much safer than what we assumed at this time.

For this test, we took many different samples (more than 100). It is not surprising that maximum are close to $\alpha_5$.

We have carried out 100 tests in each category of other tests with significance level 5/100. We denote by $n_{tp}^r$ and $n_{tf}^r$, the number of tests passed or failed. Then, we have the following results.

| | $n_{tp}^r$ | $n_{tf}^r$ |
|---|---|---|
| Equidistribution Test | 96 | 4 |
| Serial Test | 94 | 6 |
| Poker Test | 92 | 8 |
| Coupon collectors Test | 95 | 5 |
| Run Test | 97 | 3 |
| Maximum-of-t Test | 95 | 5 |
| Collision Test | 98 | 2 |
| Birthday spacings Test | 96 | 4 |
| Serial Correlation Test | 92 | 8 |
| Higher order correlation coefficient Test | 96 | 4 |

# 10 Conclusion

By theorem 3 one can find models correct $Y_n^\theta$ such that $P\{(X_1^\theta, ..., X_N^\theta) \in Bo\} = L(Bo)[1+Ob(1)\epsilon]$ where $\epsilon$ is small and it is possible to build such sequences concretely.

Now, $K_0$ increases very little when $r_1$ increases. Even, in some cases, it seems that it decreases. It seems to be the case as soon as there is asymptotic independence. Then, at most $2^q/m$ decreases much more quickly than $K_0$ increases.

So by taking m large enough and by choosing well q, we found $\epsilon$ small enough in a way that there exists correct models which checks the conditions of proposition 5.1. Then, there exists a correct model $Y_n^{\theta_0}$ of $\{y_n\}$ such that $T_q(Y_n^{\theta_0})$ is the IID model.

Then, this result show that **one can build sequences $x_n$ such that the model IID is a correct model of $x_n$.**

That means that $x_n$ behaves like any IID sample : a priori, $x_n$ can check not the properties which one awaits from a IID sample like certain tests, but that occurs only with a probability equal to that of any IID sample.

By this method, we therefore have a mean to value the technique used by Marsaglia to create its CD-ROM. We arrive in fact *to prove mathematically* that the sequence obtained can be regarded a priori as random, what Marsaglia did not.

# References


[1] KNUTH D.E. (1998) the Art of Computer Programming; Vol 2. Third Edition Addison-Wesley, Reading, Massachusetts.

[2] GENTLE J. (1984) Random Number Generation and Monte Carlo Method, Springer 13, 61-81.

[3] MENEZES A., VAN OORSCHOT P. , VANSTONE S. (1996) Handbook of Applied Cryptography, CRC Press, 1996.

[4] SCHNEIER B (1996) Applied Cryptography 2nd Edition, John Wiley and sons, Inc

[5] MARSAGLIA G (1995) CD ROM. Florida State University, site internet http://stat.fsu.edu/pub/diehard/

[6] BLACHER R. (1993) Higher Order Correlation Coefficients. Statistics 25, 1-15.

[7] BLACHER R. (2007) Central Limit Theorem by moments. Statistics and Probability Letters, 2007; 77 (17) 1647-1651

[8] BLACHER R. (2007) Une nouvelle condition d'independance pour le theoreme de la limite centrale http://hal.archives-ouvertes.fr/hal-00144878/en/ HAL: hal-00144878, version 1

[9] BLACHER R. (2009) A Perfect Random Number Generator. Rapport de Recherche LJK Universite de Grenoble. http://hal.archives-ouvertes.fr/hal-00426555/fr/

[10] BLACHER R. (2010) A Perfect Random Number Generator II. Rapport de Recherche LJK Universite de Grenoble. http://hal.archives-ouvertes.fr/hal-00443576/fr/.

[11] BLACHER R. (2010) Correct models. Rapport de Recherche LJK Universite de Grenoble. http://hal.archives-ouvertes.fr/hal-00521529/fr/

[12] BLACHER R. (2004) Solution complete au probleme des nombres aleatoires. Journes statistiques de Montpellier. http://www.agro-montpellier.fr/sfds/CD/textes/blacher1.pdf

[13] BLACHER R. (2009) File of random Number.

http://www-ljk.imag.fr/membres/Rene.Blacher/GEAL/node3.html.

[14] FRANKLIN J. N. (1963). Deterministic simulation of random processes. Math. Comp., 17, 28-59.

[15] DENG L. Y. and GEORGE, E. O. (1992) Some characterizations of the uniform distribution with applications to random number generation. Ann. Inst. Statist. Math. Vol. 44, No. 2, pp. 379-385

[16] DIETER U. (1972) Statistical interdependence of pseudo-random numbers generated by the linear congruential method, Applications of Number Theory to Numerical Analysis (S. K. Zaremba, ed.), Academic Press, New York, 1972, pp. 287-317.

[17] FRANKLIN J. N. (1963). Deterministic simulation of random processes. Math. Comp., 17, 28-59.