

Transform of various random noises in sequences of proved random numbers

René Blacher

Laboratory LJK,

Université de Grenoble, France

Type of Article

Received: XX December 20XX

Accepted: XX December 20XX

Online Ready: XX December 20XX

Abstract

In a previous paper, we have shown how to obtain sequences of numbers proved random : these sequences can be regarded as a sample of IID sequences of random variables. By using Fibonacci congruences, we transformed sequences of noises y_n such that the conditional probabilities have Lipschitz coefficients not too large. Then, we obtained sequences x_n which admitted the IID model for correct model, i.e. Fibonacci congruences behave as extractors. This method allowed to value the CD-ROM of Marsaglia. But we did not use Rap Music (as Marsaglia), but texts files. In this paper, we show that this technique can be applied for a vast majority of possible noises. In order to prove this, we shall provide all finite sequences of random variables with a well chosen measure. Then, with a probability very close to 1, the functions of Fibonacci are very good extractors. It is therefore a very efficient method to obtain sequences proved IID from almost any sequence of noises.

Keywords: Fibonacci sequence, Random numbers; Dependence; Correct models; Higher order correlation coefficients; Extractors

2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

1 Introduction

1.1 Random generators

In order to have IID random or pseudo-random numbers ¹ one can use pseudo-random generators. But it is admitted that they will never provide really random sequence : "Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin" (John Von Neumann, 1951). Then, in order to eliminate this defect one has created machines and chips that produce

**Corresponding author: E-mail: rene.blacher@imag.fr*

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

random noises. These Hardware-based random bit generators exploit the randomness which occurs in some physical phenomena. But they have various problems.

Thus, a true random bit generator requires a physical sources of randomness. But, in Menezes, Van Oorschoot, Vanstone (1996), it is written "Designing a hardware device or software program to exploit this randomness and produce a bit sequence that is free of biases and correlation is a difficult task. Moreover, random bit generators based on natural sources of randomness are subject to influence by external factors, and also to malfunctions. It is imperative that such devices be tested periodically".

Indeed, the major defect of all these systems is that there can be correlations and bias in the generated sequences. The underlying physical process can be random. But there are many measuring devices between the digital part of the computer and the physical device. These instruments can thus introduce bias and correlations (cf Schneier (1996) ch 17.14 Bias and Correlation).

One removes these bias and these linear correlations by various mathematical transformations using theory of extractors (Von Neumann (1951), Santha, Vazirani (1986), Blum (1984), Chor, Goldreich (1985), Shaltiel (2011), Barak, Impagliazzo, Wigderson (2006), Bougain (2007), Chor, Goldreich (1988), Gabizon, Raz (2008), Gabizon, Raz, Shaltiel (2006), Tauman, Kalai, Li, Rao (2009), Kamp, Rao, Vadhan, Zuckerman (2011), Kamp, Zuckerman (2007), Li (2011), Nisan Tashma (1999), Peres (1992), Ras (2005), Ras, Reingold, Vadhan (2002)).

However, the linear correlation is not the only one of the possible correlation. There exists correlations of higher order (quadratic cubic, etc : cf Blacher (1993)). If there are bits, the correlations of higher order are the multilinear correlations between 3,4,5,... bits. If one does not remove the correlations of higher order, one will not have independence.

All these reasons explain that it is difficult to do IID the sequences provided by the machines.

In order to remedy these defects Marsaglia has created a CD Rom by mixing modulo 1, Rap Music, a sequence furnished by a machine and a good pseudo random generator. This method gives good results because of the properties of this type of mixture (cf Deng, George (1992)) and perhaps even the sequence of this CD-ROM is really random. But if so, it is important to be sure of it. However, it seems difficult to know it (cf Blacher (2009)).

However, in Blacher (2011) and Blacher (2011), we found a way to bring out the method of Marsaglia : to transform the sequence of Marsaglia by using a congruence of Fibonacci. Indeed, we have proved that the sequence so obtained admits the IID model as correct model. In order to build this sequence, we did not use Rap Music as Marsaglia, but texts because their properties are easier to study logically. Furthermore we have not chosen the definitions of *real* random numbers which are still subject to many questions including philosophical questions (cf Knuth (1998)), but the existence of a correct model which is the basic postulate of many scientific studies.

This method meets the technical of extractors. Indeed, the extractors can solve the following problem posed by Von Neumann: being given a physical source producing random bits whose distribution is unknown, how, by applying a deterministic algorithm, can we produce bits equally distributed, with a prescribed accuracy? The deterministic algorithm is a function $E: \{0,1\}^M \rightarrow \{0,1\}^{n'}$. The source is a random variable X with values in $\{0,1\}^M$ for which only the law plays a role. The generator of random bits obtained is the random variable $E \circ X$. What interests us is its distribution. Traditionally, to define the precision we use the distance to the uniform distribution U : $dist(\mu, \nu) = \sum_x |\mu(x) - \nu(x)|$ where μ and ν are two probability measures on a finite set. In other words, the problem is to find E such that $dist(E\mu, U)$ is small for all measures μ of a family as large as possible. We then see easily that Fibonacci functions are extractors applied to some distributions which have good properties.

Now, one would like to apply this technique using Fibonacci functions to other noises, for example numbers provided by machines and chips. Unfortunately, our method requires that the conditional probabilities have Lipschitz coefficients not too large and it is not certain that this is checked by the numbers provided by the machines or chips. Then, one can think to use the Central

Limit Theorem (CLT) in order that this condition is checked. We treat this problem in Blacher (2013).

Now we can also show that the technique used is valid in many other cases than distributions with Lipschitz coefficients not too large: it is the aim of this paper

1.2 Fibonacci Congruence and Random Numbers

Linear congruences $T'(x) \equiv a'x \pmod{m'}$ are often used as pseudo-random generators. In this case, we choose a and m so that successive pseudorandom numbers, $T^n(x)$ and $T^{n+1}(x)$, behave as independent (cf Knuth (1998) Dieter (1972)). Now, in Blacher (2011) we understood that the Fibonacci congruences ensure the best independence of two successive numbers $T^n(x)$ and $T^{n+1}(x)$ (cf also Zaremba (1966), Blacher (2013)).

Definition 1.1. The congruence $T(x) \equiv ax \pmod{m}$ is a Fibonacci congruence if there exists $n_1 \in \mathbb{N}$, $n_1 \geq 3$, such that $m = fi_{n_1}$ and $a = fi_{n_1-1}$ when fi_n is the sequence of Fibonacci : $fi_1 = fi_2 = 1$ and $fi_{n+1} = fi_n + fi_{n-1}$. In this paper, we denote by T the Fibonacci congruences.

Then, by using them one can build IID sequences by transforming some random noise ¹ obtained from texts $y_n \in \{0/m, 1/m, \dots, (m-1)/m\}$, $n=1,2,\dots,N$.

Definition 1.2. Let $q \in \mathbb{N}^*$. Let T be a congruence of Fibonacci. We define the congruential function of Fibonacci T_q by $T_q = Pr_q \circ \hat{T} \in [0, 1]$ where

- 1) $\hat{T}(x) = \overline{T(xm)}/m$, when $\bar{z} \equiv z \pmod{m}$ and $0 \leq \bar{z} < m$ if $z \in \mathbb{Z}$,
- 2) $Pr_q(z) = \overline{0, b_1 b_2 \dots b_q}$ when $z = \overline{0, b_1 b_2 \dots}$ is the binary writing of z .

1.3 Application to random variables which have continuous densities

Let Y_n , $n=1,2,\dots,N$, be a sequence of random variables taking values in $\{0/m, 1/m, \dots, (m-1)/m\}$ and defined on a probability space $(\Omega, \mathcal{A}, P)$. Let $y_n = Y_n(\omega)$, $n=1,2,\dots,N$, be a realization of Y_n . Because Y_n has discrete values, we can always assume that Y_n has a continuous density.

Notations 1.3. Let μ_m be the measure defined by $\mu_m(k/m) = 1/m$ for all $k \in \{0, 1, \dots, m-1\}$. For all permutation ϕ of $\{1, 2, \dots, N\}$, for all $n \in \{1, 2, \dots, N\}$, for all $y'_1, y'_2, \dots, y'_{N-1}$, we denote by $f_{n,\phi}(\cdot | y'_1, y'_2, \dots, y'_{N-1})$ the conditional density with respect to μ_m of $Y_{\phi(n)}$ given $Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, \dots, Y_{\phi(n-N+1)} = y'_{N-1}$.

Because Y_n is discrete, we can also assume that $f_{n,\phi}(\cdot | y'_1, y'_2, \dots)$ has a Lipschitz coefficient smaller than m^2 because $|y_1 - y_2| \geq 1/m$ for all $y_1, y_2 \in \{0/m, 1/m, \dots, (m-1)/m\}$ if $y_1 \neq y_2$ and $0 \leq f_{n,\phi}(\cdot | y'_1, y'_2, \dots) \leq m$.

Notations 1.4. We denote by $K_0 < m^2$ a constant such that, for all permutation ϕ of $\{1, 2, \dots, N\}$, for all $n \in \{1, 2, \dots, N\}$, for all $y'_1, y'_2, \dots, y'_{N-1}$, $|f_{n,\phi}(y | y'_1, y'_2, \dots, y'_{N-1}) - f_{n,\phi}(y' | y'_1, y'_2, \dots, y'_{N-1})| \leq K_0 |y - y'|$ for all $y, y' \in \{0/m, 1/m, \dots, (m-1)/m\}$.

Then, our hypothesis is that K_0 is not too large. Indeed, we shall choose random noises y_n such that the following assumption holds.

Hypothesis 1. We suppose that $K_0 \geq 1$ and that $\gamma(m)NK_02^q/m$ is small ² where $\gamma(m)$ is of the order of $\log(\log(m))$.

¹ There are several terminology to mean the properties which must have the sequences y_n . In Shaltiel (2011), Shaltiel speaks about "sources of randomness". Santha and Vazirani speaks about semi random sources or slightly random source : Santha, Vazirani (1986). We can also say that y_n is not completely deterministic. Lastly, we say that y_n is the realization of a random phenomenon generally non IID

² For example $\gamma(m)NK_02^q/m = 1/1000$

Then, conditional probabilities of Y_n check $P\{T_q(Y_{\phi(n)}) = x_0 \mid Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, \dots\} = (1/2^q)[1 + O(1)\epsilon]$ where $|\epsilon| \leq \gamma(m)NK_02^q/m$. Indeed we proved in Blacher (2011) the following theorem.

Theorem 1. *We keep the previous notations and assumptions. We assume $m/K_0 \gg 1$. Let $X_n = T_q(Y_n)$ and $x_n = T_q(y_n)$. Then, for all subsets $Bo \subset \{0/2^q, 1/2^q, \dots, (2^q - 1)/2^q\}^N$,*

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

where L is the measure corresponding to the Lebesgue measure in the case of discrete space and where " $Ob(\cdot)$ " and " $Ob'(\cdot)$ " means the classical " $O(\cdot)$ " with the additional condition : $|Ob(1)| \leq 1$ and $|Ob'(1)| \leq 1 + |d|$ when $d \approx 0$.

If we choose m and q such that $|\epsilon|$ is small, then T_q is an ϵ -extractor in the class of functions which have Lipchitz coefficients not too large. We recall that a ϵ -extractor E for a distribution B over $\{0, 1\}^n$ is a function $\{0, 1\}^n \rightarrow \{0, 1\}^M$, $n \geq M$, such that $|P\{E(B) \in Bo\} - P\{U \in Bo\}| \leq \epsilon$ for all Borel set Bo when U has the uniform distribution on $\{0, 1\}^M$. Clearly this definition can be extended by a natural way to the case $\{0, 1, \dots, m-1\}^n \rightarrow \{0, 1, \dots, 2^q-1\}^N$. Then, functions T_q can be regarded as extractors.

Now, if ϵ is small enough, then x_n cannot be differentiated with an IID sequence. Indeed, it is wellknown that, for a sample x_n , $n=1,2,\dots,N$, there is an infinity of correct models : in particular, if x_n is a sample of an IID sequence, models such that $P\{(X_1, \dots, X_N) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$ are correct also. Reciprocally, if x_n is a realization of X_n when X_n checks $P\{(X_1, \dots, X_N) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$, the model IID is also a correct model for the sequence x_n .

Thus, in Blacher (2011), we proved that the IID model is a correct model for the sequences x_n . As a matter of fact, there exists even another correct model $Y_n^{\theta_0}$ of y_n such that $T_q(Y_n^{\theta_0})$ is exactly the IID sequence (cf also Section 3.3).

By this way, we obtained a sequence x_n which we have tested and such that the model IID is a correct model. This means that, **a priori**, this sequence x_n behave as an IID sequence. It is always possible that it does not satisfy certain tests. But that occurs only with a very weak probability as we know that it is the case for samples of sequences of IID random variables.

1.4 General case

We shall now see that for almost all distributions over $\{0/m, 1/m, \dots, (m-1)/m\}^N$, functions T_q transforms associated random variables in approximately IID sequences. Then, we shall now see that functions T_q transforms the vast majority of noises in IID sequences. For this purpose, we need a measure over all distributions.

1.4.1 Definition of a measure on the set of distributions

Notations 1.5. Let $m \in \mathbb{N}^*$. Let $\Omega_N^I = \{0/m, 1/m, \dots, (m-1)/m\}^N$. One considers the probabilities spaces $(\Omega_N^I, \mathcal{A}, P_\theta)$ where $\theta \in \Theta$ and where Θ is a set which allows to define all possible probabilities over Ω_N^I .

Moreover, we set $p_{y_1, \dots, y_N}^\theta = P_\theta\{(y_1, \dots, y_N)\}$.

For example, if $N=1$, one can choose Θ by the following way : $\Theta = \{\theta_{y_1} \in [0, 1] \mid my_1 \in \{0, 1, \dots, m-1\}\}$. Then, for all $my_s \in \{0, 1, \dots, m-1\}$ for $s=1, \dots, N$, we define $P_\theta\{y_s\}$ by

$$p_{y_s}^\theta = \frac{\theta_{y_s}}{\sum_{my_1 \in \{0, \dots, m-1\}} \theta_{y_1}} \quad (1)$$

Now, one can also use random variables.

Notations 1.6. We denote Y_n^θ the random variables defined on $(\Omega_N^I, \mathcal{A}, P_\theta)$, $\theta \in \Theta : (Y_1^\theta, \dots, Y_N^\theta) = (Y_1, \dots, Y_N)$. In this case, we set $p_{y_1, \dots, y_N}^\theta = P_\theta\{(Y_1, \dots, Y_N) = (y_1, \dots, y_N)\}$.

As a matter of fact, we use the notation Y_n^θ instead of Y_n in order to know which probability P_θ is used.

It is clear that to use Y_n^θ or P_θ is equivalent in terms of definition. For example, in the case $m = 2$, this is the easiest way to define the game of heads or tails : $\{0, 1\}^N \mapsto \{0, 1\}^N$.

More generally, we prefer to use the random variables because they have more simple notations in order to translate the transformations. Thus one can easily note operations on Y_n^θ , what we do in this article with functions $T_q(Y_n^\theta)$.

Remember also that you can take other definitions of Y_n^θ and P_θ by taking space Ω more complicated.

In order to show that for almost all sequences Y_n^θ , X_n^θ is an IID sequence with a margin of ϵ , we need a measure on the space of all possible sequences Y_n^θ , $n=1, \dots, N$. We will first define it when $N = 1$.

Notations 1.7. Suppose $N=1$. Suppose that, for all $y \in \{0/m, 1/m, \dots, (m-1)/m\}$, $p_y^\theta = \frac{pu_y^\theta}{\sum_{i=0}^{m-1} pu_{i/m}^\theta}$, where pu_y^θ is a realization of an IID sequence of random variables $\mathcal{P}_y$ defined on a probability space $(\Omega m_1, \mathcal{Alge}_1, Proba_1)$: for all $y \in \{0/m, 1/m, \dots, (m-1)/m\}$, $pu_y^{\theta(\omega_1)} = \mathcal{P}_y(\omega_1)$, $\omega_1 \in \Omega m_1$.

This definition is equivalent to that of a measure about the Y_n^θ 's : for all $y \in \{0/m, 1/m, \dots, (m-1)/m\}$, $Proba_1\{\omega_1 | pu_y^{\theta(\omega_1)} \in [a, b]\} = b - a$. So, with this measure, we consider that the pu_y^θ 's are chosen randomly.

By using notations of equation 1 $pu_{k/m}^{\theta(\omega_1)} = \theta_{k+1}(\omega_1)$. Therefore, the probability $Proba_1$ defines a measure mes_1 over the set of $\theta \in \Theta : mes_1(\theta \in \Delta_1) = Proba_1(\omega_1 \in \mathcal{E}_1)$ when $\theta(\mathcal{E}_1) = \Delta_1$. Thus,

$$\begin{aligned} & mes_1\{\theta \in \Theta | (\theta_{1/m} \in [a_1, b_1], \dots, \theta_{(m-1)/m} \in [a_m, b_m])\} \\ &= Proba_1\{\omega_1 \in \Omega_1 | (pu_{1/m}^{\theta(\omega_1)} \in [a_1, b_1], \dots, pu_{(m-1)/m}^{\theta(\omega_1)} \in [a_m, b_m])\} = \prod_{s=1}^m (b_s - a_s) \end{aligned}$$

One can easily extend these notations to the general case where $N > 1$ (cf section 4.4) by taking into account the marginal probabilities. Then, one has a measure $Proba_N$ over the set of probabilities $p_{y_1, \dots, y_N}^\theta$, $\theta \in \Theta$, defined on a probability space $(\Omega m_N, \mathcal{Alge}_N, Proba_N)$ ¹.

1.4.2 Justification of the choice of the measure

By its definition, $Proba_N$ is a measure expressing the fact that probabilities can be chosen at random. This measure adapts well to the study of $X_n^\theta = T_q(Y_n^\theta)$. For example, let us choose simple texts y_n ². We will see in section 4.1 that there is no relationship neither logical nor empirical between sequences y_n and sets $\{\gamma_s\} = T_q^{-1}(k/2^q)$.

That means that, for example, $P\{T_q^{-1}(k/2^q)\} = \sum_s P\{\gamma_s\}$ where the $P\{\gamma_t\}$'s behave like probabilities randomly selected in the sets $\{P\{Y_n = y'_s\} | s = 1, 2, \dots\}$ where $y'_s \in \{0, 1, \dots, (m-1)/m\}$. Now, this is exactly like that that measure $Proba_N$ is defined. Then, the choice of this measure is relevant for the functions T_q . More generally, we shall see that this reasoning is also valid for all extractors.

¹ Let us notice well that Ωm_N is completely different of Ω : Ω is the probability space over which the sequences Y_n^θ are defined while Ωm_N is useful in order to estimate the probability that a sequence X_n^θ is close to an IID sequence.

² And not modified as we did in Blacher (2011).

1.4.3 Main theorem

Now, with the measure $Proba_N$ for the vast majority of distributions Y_n^θ , $T_q(Y_n^\theta)$ is close to the IID distribution : we shall prove in section 4 the following theorem

Theorem 2. We suppose $m \gg 2^q \gg 1$. We set $P_{X_n^\theta}(Bo) = P\{(X_1^\theta, \dots, X_N^\theta) \in Bo\}$. Then,

$$Proba_N \left\{ \bigcap_{Bo} \left\{ |P_{X_n^\theta}(Bo) - L(Bo)| \leq \frac{2\sqrt{[Log(N)+q]N^{3/2}L(Bo)}}{\sqrt{m}} \right\} \right\} \geq 1 - \frac{2\sqrt{2}.Ne^{-Log(N)N/2}e^{-Nq/2}}{\sqrt{3\pi N[Log(N)+q]}}.$$

Remark 1.1. We set $f = \frac{2\sqrt{[Log(N)+q]N^{3/2}}}{\sqrt{m}}$. Then, of course,

$$\begin{aligned} & \bigcap_{Bo} \left\{ |P_{X_n^{\theta(\omega_a)}}(Bo) - L(Bo)| \leq f.L(Bo) \right\} \\ &= \left\{ \omega_a \in \Omega m_N \mid \left| P\{(X_1^{\theta(\omega_a)}, \dots, X_N^{\theta(\omega_a)}) \in Bo\} - L(Bo) \right| \leq f.L(Bo) \quad \forall Bo \right\}. \end{aligned}$$

For example $|P_{X_n}(\{k\}) - L(\{k\})| \leq 0.000008418/2^{33N}$ with a probability larger than $1 - e^{-1012610}$ in Ωm_N when $m = O(4 * 10^{36})$, $q = 33$, $N = 10^5$, $k = (k_1/2^{33}, \dots, k_N/2^{33})$.

We deduce from this theorem that for the vast majority of models Y_n^θ of y_n , there exists a model $Y_n^{\theta_0}$ of y_n close to Y_n^θ such that $T_q(Y_n^{\theta_0})$ has the IID distribution.

Theorem 2 means that there is no need that Y_n^θ has a Lipchitz coefficient not too large in order that $T_q(y_n)$ can be regarded as a sample of an IID sequence. This result can be confirmed by tests. Thus, for simple texts, tests clearly confirm randomness.

Then, one understands easily that Fibonacci functions are extractors applied to some distributions which have good properties.

As a matter of fact, we understand in this paper that functions T_q transforms noises which are not too deterministic in IID sequences.

1.5 Detailed results.

We have carried out many numerical and mathematical studies in order to obtain the results of this paper. We cannot detail them here. Most these results are in Blacher (2009), Blacher (2010), Blacher (2010) and Blacher (2011).

1.6 Symbols

CLT : CLT for Central Limit Theorem

μ_m : μ_m is the uniform measure over $\{0, 1, \dots, m-1\}/m$.

L : L is the measure corresponding to the Lebesgue measure in the discrete case.

$Ob'(1)$: Ob' is the classical "O" with the condition $|Ob'(1)| \leq 1 + |d|$ when $d \approx 0$.

$Ob(1)$: Ob is the classical "O" with the condition $|Ob(1)| \leq 1$.

Bo : Bo means a Borel set of $\mathbb{R}^p$.

$I_t = [k_t/2^q, k'_t/2^q[$, $k_t, k'_t \in \{0, 1, \dots, 2^q - 1\}$, $k_t < k'_t$.

$N(Bo)$: $N(Bo)$ is the number of points of a sample ζ_n which belongs to Bo .

ϵ : $\epsilon > 0$ is a real very small.

fi_n : fi_n is the sequence of Fibonacci : $fi_{n+2} = fi_{n+1} + fi_n$, $fi_1 = fi_2 = 1$.

T : T means a Fibonacci congruence $T(x) \equiv ax \text{ modulo } m$.
 T_q : T_q is the congruential function of Fibonacci
 $\bar{h}$: $\bar{h} \equiv h \text{ modulo } m$ and $0 \leq \bar{h} < m$.
 $N(0, \sigma^2)$: $N(0, \sigma^2)$ means the normal distribution with mean 0 and variance σ^2 .
 $Y \sim N(0, 1)$: $Y \sim N(0, 1)$ if Y is a random variable which has the distribution $N(0, 1)$.
 $\Gamma(b)$: $\Gamma(b) = P\{|X_G| \geq b\}$ where $X_G \sim N(0, 1)$.
 $\overline{0, d_1 d_2 \dots}$: $\overline{0, d_1 d_2 \dots}$ means the writing base 2 of a number $z = \overline{0, d_1 d_2 \dots}$.
 $E\{X\}$: $E\{X\}$ means the expectation of the random variables X
 σ_s^2 : σ_s^2 means always variances.
 $\text{Card}(E)$: $\text{Card}(E)$ is the cardinality of the set E .

2 Fibonacci congruences

We are interested in the dependence of two successive random numbers of a linear congruence $T'(x) \equiv ax \text{ mod } (m)$. In Blacher (2011), we understood that this dependence depends on sequences r_n defined as follows.

Notations 2.1. Let $r_0 = m$, $r_1 = a$. If $n \in \mathbb{N}^*$, one denotes by $r_n \in \mathbb{N}$ and $h_n \in \mathbb{N}^*$ the sequences defined by recurrence by the Euclidean division of r_n by r_{n+1} when $r_{n+1} \neq 0$: $r_n = h_{n+1}r_{n+1} + r_{n+2}$, $r_{n+2} < r_{n+1}$. Moreover, one denotes by d the smallest integer such as $r_{d+1} = 0$.

Remark 2.1. By induction we know r_n and r_{n+1} . Then, h_{n+1} and r_{n+2} are fully defined by the Euclidean division of r_n by r_{n+1} .

Then, dependence is weaker if the h_i 's are small (cf Blacher (2011)). So the congruence which provides the best independence is such that $h_i = 1$ (except for the two last terms). It is the Fibonacci congruence : r_n is the Fibonacci sequence f_i : $m = r_0 = f_{i_{n_0+1}}$, $a = r_1 = f_{i_{n_0}}$, $r_2 = f_{i_{n_0-1}}, \dots$ (cf Blacher (2011) or Blacher (2011)).

Therefore it's with Fibonacci congruences that dependence is the weakest. In particular, let $I = [c, c'[\cap\{0, 1, \dots, m-1\}$ where $c, c' \in \{0, 1, \dots, m-1\}$. Let k^n , $n=1, 2, \dots, c'-c$, be the permutation of $\{c, c+1, \dots, c'-1\}$ such that $\bar{T}^{-1}(k^1) < \bar{T}^{-1}(k^2) < \bar{T}^{-1}(k^3) < \dots < \bar{T}^{-1}(k^{c'-c})$. Then, $|\bar{T}^{-1}(k^r)/m - r/N(I)| \leq \varphi(m)/N(I)$, where $\varphi(m)$ is the order of $\text{Log}(\text{Log}(m))$ and where $N(I) = c' - c$ (cf Blacher (2011)). Then, we set $\gamma(m) = [2 + \varphi(m)]$: cf Hypothesis 1.

3 Models ϵ -equivalent

3.1 Correct models

One can always suppose that y_n is the realization of a sequence of random variables Y_n defined on a probability space (Ω, A, P) : $y_n = Y_n(\omega)$, $\omega \in \Omega$, where Y_n is a correct model of y_n . As a matter of fact, there exist an infinity of correct models Y_n^θ of y_n ¹.

It thus raises the question to define what is a correct model. Indeed, if a model Y_n^θ is not correct, it is however possible that $y_n = Y_n^\theta(\omega)$. Now, in the case where the model Y_n^θ is IID, to say that Y_n^θ is correct amounts to say that the sequence of *real* numbers y_n is IID. But this problem is a very complex problem which needs mathematical logic and philosophical concepts and which has never been completely resolved: various definitions were given by different authors and none is truly satisfactory (cf Knuth (1998)). It is therefore clear that we cannot in a few lines give a satisfactory solution for everybody.

¹ There are m^N sequences y_n : $(y_1 = k_1/m, \dots, y_N = k_N/m)$ while there is a not countable infinity of models $P\{(Y_1 = k_1/m, \dots, Y_N = k_N/m)\} = p_{k_1, \dots, k_N}$.

However, it is admitted generally that correct models exist. In fact, it is a traditional assumption in science. Then, the most simple way is to admit it like a postulate without defining exactly what is a correct model (as it is admitted that IID numbers exist) : cf Blacher (2010). However, in this article we will require that the proposition 3.2 is valid. Then, the simplest way is to admit the following conjecture.

Conjecture 3.1. *There are models called correct models such that the proposition 3.2 is checked*

It is certain that correct models defined by this way exist ². But, as a matter of fact, there is no indication that a such correct model has the sense which we usually think to give to it. Let us suppose therefore that it is the case (cf Blacher (2010)).

Therefore, we do not detail more here what we can define as correct model. Note however that the IID correct models differs from the usual definitions of random sequences of numbers : the latest one concern sequences of *real* numbers while the correct models use random variables. However as regards the randomness of $T_q(y_n)$, in Blacher (2009), we demonstrated in 2010 that these sequences checked the definitions of random numbers give by Knuth in Knuth (1998), but translated in terms of models.

Now we can give empirical definitions. They have the default to represent not exactly what is meant by correct models. We can realize it in Blacher (2010) where we made a first study of this concept. For example one can use functions of chi square.

Example 3.1. *Let T_1 be a test in order to test the hypothesis $\ll$ the sequence $y_n^1, n = 1, 2, \dots, n_I$, is a realization of the sequence of random variables $Y_n^1, n = 1, 2, \dots, n_I$, $\gg$. One could say that Y_n^1 is a correct model for the sequence $y_n, n=1,2,\dots,N$, if the test is accepted for this sequence y_n .*

Note that the test can be more complicated than an usual chi-square test based on a statistic of type $\sum_s (\mathbb{1}_{E_s}(y_n) - P\{E\})^2$. One can for example also tests at the same time distribution of Y_n and dependence of order 2 with $\sum_s (\mathbb{1}_{E_s}(y_n) - P\{E_s\})^2 + \sum_{s,t} (\mathbb{1}_{E_s}(y_n)\mathbb{1}_{E_t}(y_{n+1}) - P\{E_s \otimes E_t\})^2$. etc. We can therefore get extremely sophisticated tests, but theoretical, because the distributions of statistics can be very difficult to obtain. But, thus, we can test almost all theoretically possible dependences.

Of course, to check all the required conditions is extremely difficult in order to define only the IID model (cf Knuth (1998), Blacher (2009)). One can easily imagine the complications which will occur for some models which in addition, have their distribution unknown.

This is why it is easier to say that there are correct models without defining more them. Now, it is necessary that proposition 3.2 is checked. But we can assume that it is the case because it is very natural. We will see why now.

Note that a more detailed study is in Blacher (2010).

3.2 Equivalent Models

3.2.1 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 ,

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

One supposes that $Y_n^{\theta_1}$ is a correct model of the sequence $y_n, n=1,2,\dots,N$. One wants to show that $Y_n^{\theta_2}$ is also a correct model of y_n if ϵ is small enough.

² It suffices to take as correct models, all models Y_n^θ

3.2.2 The problem of the definition

Of course in order to prove that $Y_n^{\theta_2}$ is a correct model, it would require a definition of a correct model. Now we said that the easiest way is to admit the existence of correct model without giving a definition. Yet, even without precise definition, it is also an extremely plausible conjecture: thus, the theory of extractors uses models ϵ -close : $|P\{(Y_1^{\theta_2}, \dots, Y_N^{\theta_2}) \in Bo\} - P\{(Y_1^{\theta_1}, \dots, Y_N^{\theta_1}) \in Bo\}| \leq \epsilon$ (but with less emphasis about the statistical aspect, although this one is implicit in this definition). In fact, it suffices to study examples to be convinced of the validity of this conjecture (cf also Blacher (2010)).

3.2.3 Example

Let us suppose that we have a really IID sequence of random variables X_n^ϵ 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 absolutely not understand that X_n^ϵ has not the uniform distribution on $[0, 1]$. It is wellknown that one needs samples with size larger than $N=1000$ minimum in order to test this difference.

For example, one cannot test significantly H_0 : " X_n^θ has the uniform distribution" against $H_1(\epsilon)$: " $P\{X_n^\theta \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$ " if $\sqrt{N} \epsilon \leq 1/10$. Indeed, 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^θ is really IID (cf also section 4.3 of Blacher (2010)).

Then it is no possible to differentiate the IID model and models such that $P\{(X_1^\theta, \dots, X_N^\theta) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$.

3.2.4 Border of correct models

Now there is a problem : for example, let use a realization x_n of the IID model, and let $X_n^{\theta_1}$ be a model checking $P\{(X_1^{\theta_1}, \dots, X_N^{\theta_1}) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon^1]$ where ϵ^1 is small enough but not very small. Let $X_n^{\theta_2}$ be a model such that $P\{(X_1^{\theta_2}, \dots, X_N^{\theta_2}) \in Bo\} = P\{(X_1^{\theta_1}, \dots, X_N^{\theta_1}) \in Bo\}[1 + Ob(1)\epsilon]$. Then, $X_n^{\theta_2}$ can be not a correct model of x_n : it is enough that ϵ^1 is in extreme case of the possible values of the ϵ 's such that $P\{(X_1^\theta, \dots, X_N^\theta) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$, $\sup_{Bo}(Ob(1)) = 1$, imply that X_n^θ is a correct model.

Then, there are models of y_n more correct than others. These are models $Y_n^{\theta_0}$ of y_n such that, if $P\{(Y_1^\theta, \dots, Y_N^\theta) \in Bo\} = P\{(Y_1^{\theta_0}, \dots, Y_N^{\theta_0}) \in Bo\}[1 + Ob(1)\epsilon]$, Y_n^θ is also a correct model of y_n when ϵ is small enough, but not too small. For example, $\epsilon = 1/10$, $1/100$ or at worst $\epsilon = 1/N$ if need be (cf section 5-7 of Blacher (2011)).

It seems clear that such models exist. For example it is assumed that this is the case when x_n is a sample of the independent sequence X_n^ϵ which is a good realization of X_n^ϵ . On the other hand, we know that there 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 : close models will also correct models.

All these points are detailed in Blacher (2010).

3.3 Exact IID model

Then, generally, if Y_n^θ is a correct model such that $T_q(Y_n^\theta)$ cannot be differentiated with the IID model, one will be able to choose another correct model $Y_n^{\theta_0}$ close to Y_n^θ and such that $T_q(Y_n^{\theta_0})$ is exactly the IID model. Indeed, in Blacher (2010), one proved easily the following proposition.

Proposition 3.2. *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^θ is a model satisfying, for all Borel set Bo , $P\{(Y_1^\theta, \dots, Y_N^\theta) \in Bo\} = P\{(Y_1^{\theta_c}, \dots, Y_N^{\theta_c}) \in Bo\}[1 + Ob(1)\epsilon_Y]$, then Y_n^θ is a correct model of y_n .*

One assumes also that, for all $(k_1, \dots, k_N)$, $P\{\{T_q(Y_1^{\theta_c}) = k_1/2^q\} \cap \dots \cap \{T_q(Y_N^{\theta_c}) = k_N/2^q\}\} = \frac{1}{2^{qN}}[1 + \epsilon_{k_1, \dots, k_N}(q)]$, where $\sup_{k_1, \dots, k_N} |\epsilon_{k_1, \dots, k_N}(q)| = \epsilon_X(q)$. 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 $\{y_n\}_{n=1, \dots, N}$ such that, for all $(k_1, \dots, k_N)$,

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

3.4 Discussion about the definitions of random sequence

The results and the above considerations are sufficient to ensure that the sequence of IID random variables is a correct model of x_n .

This avoids any complications due to the extremely sophisticated definitions of a random sequence of real numbers (cf Knuth (1998)). It's better this way. Indeed, these definitions are sometimes contradictory, and some philosophical definitions are incompatible with the statistical reality (a detailed study is in Blacher (2009)). Thus one of these definitions implies that there is no finite random sequence (cf Knuth (1998)). We avoid discussions related to this problem by saying that the IID model $U_n \in \{0, 1, \dots, 2^q - 1\}/2^q$, $n=1, 2, \dots, N$, is a correct model of the sequence x_n ¹. However, one can also note that the results obtained are compatible with other definitions of random sequences of real number quoted in Knuth (1998) (cf also Blacher (2009)).

In conclusion the technique used here avoids these problems of definition because it admits as a conjecture the existence of correct models and ϵ -equivalent models, assumptions which are extremely likely. A more detailed study of all these problems is in Blacher (2009) and Blacher (2010). But of course it can not fully answer such complex problems which are discussed for many years.

4 Introduction of a measure over the set of probabilities

4.1 Sums of probabilities randomly chosen

In Blacher (2011), we saw that if we assume that K_0 is not too large $P\{T_q(Y_n^\theta) = k/2^q | Y_{n-s}^\theta = y'_s, s = 1, \dots, p\} = \sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s | Y_{n-s}^\theta = y'_s, s = 1, \dots, p\} \approx 1/2^q$ where $\mathcal{G}_k = \{\gamma_1, \gamma_2, \dots\} = \bar{T}^{-1}(mI_k \cap \{0, 1, \dots, m-1\})/m$. We proved this result by an approximation of the Riemann integration. But it is enough to read the proof in order to see that this sum behaves like a sum randomly chosen and that there is no need of the hypothesis " K_0 not too large" (cf section 5).

This method can be applied as soon as $\sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s | Y_{n-s}^\theta = y'_s, s = 1, \dots, p\}$ behaves as sums randomly chosen (cf section 5). It will be shown below that it is

¹ This may indeed at first sight contradictory to the definition quoted just before: some logicians have worried about this contradiction. But the answer is simple: our result is valid for some definition (statistical) and not others (philosophical). Note also that to use sequences of random variables is more reasonable than using the philosophical definition about real numbers : one can not deny the existence of a sequence of random variables U_n (because of *philosophical definition*) without declaring that many of Statistics would be unfounded because based precisely on the study of sequences U_n : the theory of Statistics began with the study of sequences of numbers such as sequences of dice. Another remark: the philosophical definition concerned implies that an infinite sequence might be random, whereas none of its finite subsequences would be it, what is a shocking concept of randomness.

generally the case for the vast majority of models. To prove this, we need a measure on all models Y_n^θ which takes account of this fact.

4.2 One dimensional case

Therefore, we will define a measure on all models Y_n^θ , which is equivalent to defining a measure on the set of associated probabilities.

Notations 4.1. For all $\theta \in \Theta$, for all $(y_1, y_2, y_3, \dots, y_N) \in \{0/m, 1/m, \dots, (m-1)/m\}^N$, we set $p_{y_1, \dots, y_N}^\theta = P\{\{Y_1^\theta = y_1\} \cap \{Y_2^\theta = y_2\} \cap \dots \cap \{Y_N^\theta = y_N\}\}$.

Then, we shall use a measure of the set of the $p_{y_1, \dots, y_N}^\theta$'s where the probabilities are chosen randomly.

For example, if $N=1$, we have supposed that, for all $y \in \{0/m, 1/m, \dots, (m-1)/m\}$, $p_y^\theta = \frac{pu_y^\theta}{\sum_{i=0}^{m-1} pu_{i/m}^\theta}$, where pu_y^θ is a realization of an IID sequence of random variables $\mathcal{P}_{y_s}$ defined on a probability space $(\Omega m_1, \mathcal{Alge}_1, \text{Proba}_1)$.

By this way, one provides the set of the p_y^θ with a measure : i.e. the set of possible probabilities p_y^θ is itself the realization of a sequence of random variables defined on a probability space. Therefore, in this paper, one thus uses two type of probability spaces : Ω and Ωm_N . On Ω , one defines the random variables associated with the sample $y_n : y_n = Y_n^\theta(\omega)$, $\omega \in \Omega$. On Ωm_1 , one defines a measure intended for knowing the quantity of probability p_y^θ such that $T_q(Y_n^\theta)$ will be close to the IID model : $p_y^{\theta(\omega_1)} = \frac{\mathcal{P}_y(\omega_1)}{\sum_{i=1}^m \mathcal{P}_{i/m}(\omega_1)}$, $\omega_1 \in \Omega m_1$.

4.3 Two dimensional case

If $N \geq 2$, it is necessary to take in account the marginal distributions. For example, if $N=2$ we have to consider that $\sum_{y_2} p_{y_1, y_2}^\theta = p_{y_1}^\theta$ where the $p_{y_1}^\theta$'s mean the probabilities of the first marginal distribution. With this aim, we have introduced following notations.

Notations 4.2. For all $(y_1, y_2) \in \{0/m, 1/m, \dots, (m-1)/m\}^2$, we set

$$P\{(Y_1^\theta, Y_2^\theta) = (y_1, y_2)\} = p_{y_1, y_2}^\theta = \frac{pu_{y_1, y_2}^\theta}{\sum_{i_2=0}^{m-1} pu_{y_1, i_2/m}^\theta}.$$

where $\{pu_y^\theta\}_{m y \in \{0, 1, \dots, m-1\}}$ is a sample of an IID sequence of random variables $\mathcal{P}_y$, defined on a probability space $(\Omega m_2, \mathcal{Alge}_2, \text{Proba}_2) : pu_y^{\theta(\omega_2)} = \mathcal{P}_y(\omega_2)$ for all $y \in \{0/m, 1/m, \dots, (m-1)/m\}$. We assume also that, for each y , $\{pu_{y, y'}^\theta\}_{m y' \in \{0, 1, \dots, m-1\}}$ is a sample of an IID sequence of random variables $\mathcal{P}_{y, y'}$, $y' \in \{0/m, 1/m, \dots, (m-1)/m\}$, defined on $\Omega m_2 : pu_{y, y'}^{\theta(\omega_2)} = \mathcal{P}_{y, y'}(\omega_2)$.

Therefore $\mathcal{P}_{y_1, y_2}$ and $\mathcal{P}_{y_1}$ have the uniform distribution U . Let E_U and σ_U^2 be the associated expectation and variance : $E_U = 1/2$ and $\sigma_U^2 = 1/12$. Then, we can study intervals I_t defined by the following way.

Notations 4.3. Let $I_t = [k_t/2^q, k'_t/2^q[$, $k_t, k'_t \in \{0, 1, \dots, 2^q - 1\}$, $k_t < k'_t$. Let $N(I_t)$ be the number of points $h/m \in [k_t/2^q, k'_t/2^q[$, $h \in \mathbb{N}$.

We set $\Gamma_1(b) = \text{Max}_{n \geq h_0} (\text{Proba}\{|S_n| \geq b\})$ where $h_0 = \lfloor m/2^q \rfloor$, the integer part of $m/2^q$, and where $S_n = \frac{\sum_{i=1}^{n-1} \mathcal{P}_{i_1/m}}{\sigma_U \sqrt{n}}$ where $b \in \mathbb{R}_+$ is not too large

Proposition 4.1. We suppose $m \gg 2^q$ and $2^q \gg 1$. Let $\epsilon'_{(2)} \approx \frac{2b\sigma_U}{E_U\sqrt{h_0}}$. Then, the subspace $\mathcal{F} \subset \Omega m_2$ such that

$$P\left\{\{X_1^{\theta(\omega_2)} \in I_1\} \cap \{X_2^{\theta(\omega_2)} \in I_2\}\right\} = \frac{N(I_1)N(I_2)}{m^2}[1 + Ob(1).\epsilon'_{(2)}],$$

for all $\omega_2 \in \mathcal{F}$ has a probability larger than $1 - 4\Gamma_1(b)$ where $\Gamma_1(b) \approx \Gamma(b) = P\{|X_G| \geq b\}$ when $X_G \sim N(0, 1)$.

Proof At first, the following lemma is obvious : it is enough to use the CLT.

Lemma 4.1. We have $\Gamma_1(b) = \text{Max}_{n \geq h_0} \left(\text{Proba}\{|S_n^\mathcal{E}| \geq b\} \right)$ where $S_n^\mathcal{E} = \frac{\sum_{i_1 \in \mathcal{E}} \mathcal{P}_{i_1/m}}{\sigma_U \sqrt{n}}$ for all $\mathcal{E} \subset \{0, 1, \dots, m-1\}$ such that $\text{card}(\mathcal{E}) = n$.

Then, for all y_1 , $\Gamma_1(b) = \text{Max}_{n \geq h_0} \left(\text{Proba}\{|S_n^{\mathcal{E}, y_1}| \geq b\} \right)$ where $S_n^{\mathcal{E}, y_1} = \frac{\sum_{i_2 \in \mathcal{E}} \mathcal{P}_{y_1, i_2/m}}{\sigma_U \sqrt{n}}$ for all $\mathcal{E} \subset \{0, 1, \dots, m-1\}$ such that $\text{card}(\mathcal{E}) = n$.

Then, we use the CLT. Then, with a probability larger than $1 - 2\Gamma_1(b)$ in Ωm_2 , the two following relations hold

$$\begin{aligned} \frac{1}{N(I_2)} \sum_{y_2 \in I_2} pu_{y_1, y_2}^{\theta(\omega_2)} &= E_U + \frac{Ob(1).b\sigma_U}{\sqrt{N(I_2)}}, \\ \frac{1}{m} \sum_{i_2=0}^{m-1} pu_{y_1, i_2/m}^{\theta(\omega_2)} &= E_U + \frac{Ob(1).b\sigma_U}{\sqrt{m}}. \end{aligned}$$

Then, with a probability larger than $1 - 2\Gamma_1(b)$ in Ωm_2 ,

$$pu_{y_1}^{\theta(\omega_2)} \frac{\sum_{y_2 \in I_2} pu_{y_1, y_2}^{\theta(\omega_2)}}{\sum_{i_2=0}^{m-1} pu_{y_1, i_2/m}^{\theta(\omega_2)}} = \frac{N(I_2)pu_{y_1}^{\theta(\omega_2)}}{m} \frac{1 + \frac{Ob(1).b\sigma_U}{E_U\sqrt{N(I_2)}}}{1 + \frac{Ob(1).b\sigma_U}{E_U\sqrt{m}}} = \frac{N(I_2)pu_{y_1}^{\theta(\omega_2)}}{m} [1 + Ob(1).\epsilon_2],$$

where $\epsilon_2 \approx \frac{b\sigma_U}{E_U\sqrt{N(I_2)}} + \frac{b\sigma_U}{E_U\sqrt{m}}$.

By the same way

$$\frac{\sum_{y_1 \in I_1} pu_{y_1}^{\theta(\omega_2)}}{\sum_{i_1=0}^{m-1} pu_{i_1/m}^{\theta(\omega_2)}} = \frac{N(I_1)}{m} \frac{E_U + \frac{Ob(1).b\sigma_U}{\sqrt{N(I_1)}}}{E_U + \frac{Ob(1).b\sigma_U}{\sqrt{m}}} = \frac{N(I_1)}{m} [1 + Ob(1).\epsilon_1],$$

where $\epsilon_1 \approx \frac{b\sigma_U}{E_U\sqrt{N(I_1)}} + \frac{b\sigma_U}{E_U\sqrt{m}}$.

Then, with a probability larger than $1 - 4\Gamma_1(b)$ in Ωm_2 ,

$$\begin{aligned} \sum_{y_1 \in I_1} \sum_{y_2 \in I_2} p_{y_1, y_2}^{\theta(\omega_2)} &= \sum_{y_1 \in I_1} \frac{pu_{y_1}^{\theta(\omega_2)} \frac{\sum_{y_2 \in I_2} pu_{y_1, y_2}^{\theta(\omega_2)}}{\sum_{i_2=0}^{m-1} pu_{y_1, i_2/m}^{\theta(\omega_2)}}}{\sum_{i_1=0}^{m-1} pu_{i_1/m}^{\theta(\omega_2)}} = \sum_{y_1 \in I_1} \frac{\frac{N(I_2)pu_{y_1}^{\theta(\omega_2)}}{m} [1 + Ob(1).\epsilon_2]}{\sum_{i_1=0}^{m-1} pu_{i_1/m}^{\theta(\omega_2)}} \\ &= \frac{N(I_1)N(I_2)}{m^2} [1 + Ob(1).\epsilon_1][1 + Ob(1).\epsilon_2] = \frac{N(I_1)N(I_2)}{m^2} [1 + Ob(1).\epsilon''_{(2)}] \end{aligned}$$

where $\epsilon''_{(2)} \approx \frac{b\sigma_U}{E_U\sqrt{N(I_1)}} + \frac{b\sigma_U}{E_U\sqrt{m}} + \frac{b\sigma_U}{E_U\sqrt{N(I_2)}} + \frac{b\sigma_U}{E_U\sqrt{m}} \approx Ob(1) \frac{2b\sigma_U}{E_U\sqrt{h_0}}$. ■

We wonder if it is better to choose the sums $\sum_{y_2 \in A_2} p_{y_1, y_2}^\theta$ or $\sum_{y_1 \in A_1} p_{y_1, y_2}^\theta$ in order to define the measure. But we understand that, by proposition 4.1, it does not have importance : whatever the chosen sum, the result remains the same one : $\epsilon'_{(2)} \approx \frac{2b\sigma_U}{E_U \sqrt{\hbar_0}}$. However, one can choose a representation not giving priority to a sum rather than the other.

Notations 4.4. One generalizes the notations 4.2 by changing the roles of y_1 and y_2 : in particular, we set

$$p_{y_1, y_2}^{\prime\theta} = \frac{p_{y_1, y_2}^{\prime\theta} \sum_{i_1=0}^{m-1} p_{i_1/m, y_2}^{\prime\theta}}{\sum_{i_2=0}^{m-1} p_{i_2/m}^{\prime\theta}}.$$

Then, we set

$$\pi_{y_1, y_2}^\theta = (1/2)[p_{y_1, y_2}^\theta + p_{y_1, y_2}^{\prime\theta}].$$

4.4 General case

It is easy to generalize the previous results.

Notations 4.5. We generalize by natural way notations 4.4. to probabilities $p_{y_1, y_2, \dots, y_N}^\theta$. Then we generalize the $p_{y_1}^\theta$ and p_{y_1, y_2}^θ in various $p_{y_1, y_2, \dots, y_p}^\theta$, $p=1, 2, 3, \dots, N$, which are samples of sequences of IID random variables defined on the probability space $(\Omega_m, \text{Alge}_p, \text{Proba}_p)$.

Proposition 4.2. Let $\epsilon'_{(N)} \approx \frac{Nb\sigma_U}{E_U \sqrt{\hbar_0}}$. Then, with a probability larger than $1 - 2N\Gamma_1(b)$ in the space Ω_{m_N} ,

$$P\left\{\{X_1^{\theta(\omega_N)} \in I_1\} \cap \{X_2^{\theta(\omega_N)} \in I_2\} \cap \dots \cap \{X_N^{\theta(\omega_N)} \in I_N\}\right\} = \frac{\prod_{r=1}^N N(I_r)}{m^N} [1 + Ob(1) \cdot \epsilon'_{(N)}].$$

We deduce the following proposition.

Proposition 4.3. There exists $\epsilon_{(N)} \approx \epsilon'_{(N)}$ such that, with a probability larger than $1 - 2N\Gamma_1(b)$,

$$P\left\{X_1^{\theta(\omega_N)} = \frac{k_1}{2^q}, \dots, X_N^{\theta(\omega_N)} = \frac{k_N}{2^q}\right\} = \frac{1}{2^{Nq}} [1 + Ob(1) \cdot \epsilon_{(N)}].$$

Now, we study the Borel sets of $\{0, 1, \dots, 2^q - 1\}^N$: $Bo = \cup_{(k_1, \dots, k_N) \in \Theta} \{\{k_1\} \otimes \dots \otimes \{k_N\}\} = \cup_{(k_1, \dots, k_N) \in \Theta} \{I_{k_1} \otimes \dots \otimes I_{k_N}\}$. Then, $L(Bo) = \frac{\text{Card}(\Theta)}{2^{Nq}}$. Now we need the following lemma.

Lemma 4.2. . Let $\epsilon > 0$. Assume that, for all $(k_1, \dots, k_N)$, $P\{X_1 = k_1/2^q, \dots, X_N = k_N/2^q\} = \frac{1}{2^{Nq}} [1 + Ob(1)\epsilon]$. Then, $P\{(X_1, \dots, X_N) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$, for all Borel set $Bo \subset \{0/m, 1/m, \dots, (m-1)/m\}^N$.

Then, we can prove the following proposition.

Proposition 4.4. The probability in Ω_{m_N} that, for all Borel set $Bo \subset \{0, 1, \dots, m-1\}^N$,

$$|P\{(X_1^{\theta(\omega_N)}, \dots, X_N^{\theta(\omega_N)}) \in Bo\} - L(Bo)| \leq L(Bo)\epsilon_{(N)}$$

is larger than $1 - 2N2^{qN}\Gamma_1(b)$.

Proof For each $\{k_1\} \otimes \dots \otimes \{k_N\}$, the probability in Ωm_N that

$$\left| P\{X_1^{\theta(\omega_N)} = k_1/2^q, \dots, X_N^{\theta(\omega_N)} = k_N/2^q\} - \frac{1}{2^{Nq}} \right| > \frac{\epsilon_{(N)}}{2^{Nq}}$$

is smaller than $2N\Gamma_1(b)$.

There is $(2^q)^N$ possible $\{k_1\} \otimes \dots \otimes \{k_N\}$. Therefore, the probability in Ωm_N that, for all $\{k_1\} \otimes \dots \otimes \{k_N\}$,

$$\left| P\{X_1^{\theta(\omega_N)} = k_1/2^q, \dots, X_N^{\theta(\omega_N)} = k_N/2^q\} - \frac{1}{2^{Nq}} \right| > \frac{\epsilon_{(N)}}{2^{Nq}}$$

is smaller than $2N2^{qN}\Gamma_1(b)$.

Then, the probability in Ωm_N that, for all Borel set $Bo \subset \{0, 1, \dots, m-1\}^N$,

$$|P\{(X_1^{\theta(\omega_N)}, \dots, X_N^{\theta(\omega_N)}) \in Bo\} - L(Bo)| \leq L(Bo)\epsilon_{(N)}$$

is larger than $1 - 2N2^{qN}\Gamma_1(b)$. ■

Then, we can prove theorem 2 : the previous results hold for various b. Then, we can choose $b = \sqrt{3N[Log(N)N + q]}$. Now, $\Gamma_1(b) \approx \Gamma(b) \approx \frac{\sqrt{2}}{\sqrt{\pi b}} e^{-b^2/2}$ when b is big (cf (28)' page 56 Johnson, Kotz (1970)). We deduce the result. ■

Remark 4.3. Remark that if we impose $\frac{N^3 q 2^q}{m} \ll 1$ one can choose $b = \sqrt{2Nq}$: in this case, with a probability in Ωm_N larger than $1 - 2\sqrt{N/[\pi q]}e^{-[1-Log(2)]qN}$,

$$|P\{(X_1^{\theta(\omega_N)}, \dots, X_N^{\theta(\omega_N)}) \in Bo\} - L(Bo)| \leq L(Bo)\sqrt{\frac{N^3 q 2^q}{m}}. \quad \blacksquare$$

Then, theorem 2 show that, for the vast majority of models Y_n^θ of a sequence y_n , $n=1, \dots, N$, $T_q(Y_n^\theta)$ is ϵ -equivalent to the IID model, i.e. $T_q(y_n)$ admits the IID models for correct model if m and q are well chosen. As a matter of fact we shall see below that it is the case for the models Y_n^θ which are not too deterministic.

5 Case where K_0 is not too large

Now we shall need use the proof of theorem 1. Then, we summarize at first this proof.

In theorem 1, it is written that $P\{(X_1, \dots, X_N) \in Bo\} = L(Bo)\left[1 + \frac{\gamma(m)Ob'(1)K_0N2^q}{m}\right]$ for all Borel set Bo. Then, if K_0 is not too large, there is no difficulty in order to choose m and q in such a way that $\epsilon \leq \gamma(m)2^qNK_0/m$ is small enough. Therefore, $P\{(X_1, \dots, X_N) \in Bo\} = L(Bo)[1 + Ob(1)\epsilon]$.

Let h_N be a density with a Lipschitz coefficient not too large. Because, the points of $\bar{T}^{-1}(mI)$ are well distributed in $\{0, 1, \dots, m-1\}$, the sum of points of $h_N(\bar{T}^{-1}(mI))$ will be close to $card(mI \cap \{0, 1, \dots, m-1\})/m$ (e.g. cf figure 1). Then, $P\{X_{\phi(n)} \in I_k \mid Y_{\phi(n-1)} = y'_1, Y_{\phi(n-2)} = y'_2, \dots\} = L(I_k)\left[1 + \frac{\gamma(m)Ob'(1)K_0}{m/2^q}\right]$ (cf Blacher (2011)).

Now, one proves easily that, for all $I_{k_1} \otimes \dots \otimes I_{k_N}$, $P\{(X_1, \dots, X_N) \in I_{k_1} \otimes \dots \otimes I_{k_N}\} = \prod_{s=1}^N (L(I_{k_s})[1 + Ob(1)\epsilon'])$, where $|\epsilon'| \leq \frac{\gamma(m)|Ob'(1)K_0|}{m/2^q}$. We deduce $P\{(X_1, X_2, \dots, X_N) \in$

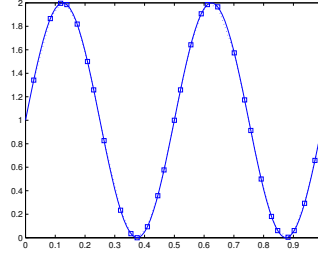


Figure 1: Points of $h'_N[\bar{T}^{-1}(mI_k)]$ when $h'_N(t) = \sin(4\pi t) + 1$

$I_{k_1} \otimes I_{k_2} \otimes \dots \otimes I_{k_N} \} = \frac{1}{2^{Nq}} \left[1 + \frac{\gamma(m)Ob'(1)NK_0}{m/2^q} \right]$, and therefore, theorem 1.

Then, by theorem 1, si K_0 is not too large, $T_q(Y_n)$ is *always* approximately IID. Yet even with a such model, the sample $T_q(y_n)$ may behave not as an IID sample. For example if $y_n \in T^{-1}(k_0/2^q)$ for all n : x_n is constant. This contradiction is due to what such a model is good *a priori* in the same way that a sample extracted from a phenomenon really IID may also be constant. As a matter of fact, it is exactly that could be desired: the use of functions T_q amounts to extract a sample of a phenomenon really IID.

Now, we are interested in the behavior of the sets $T^{-1}(k/2^q)$ with respect to densities of Y_n . This means to be interested in points of densities of type $h'_N[\bar{T}^{-1}(mI_k)]$: cf figure 1. Now $P\{T^{-1}(k/2^q)\} = \sum_{s: y_s \in T^{-1}(k/2^q)} P\{y_s\}$ and we have the same type of equalities for conditional probabilities. Now, these $P\{y_s\}$ are well distributed in the set $\mathcal{T} = \{P\{t/m\} | t \in \{0, 1, \dots, m-1\}\}$. In fact we can consider that the $\{P\{y_s\} | y_s \in T^{-1}(k/2^q)\}$'s are a set randomly selected $\mathcal{T}$, especially if we are interested in sums $\sum_{s: y_s \in T^{-1}(k/2^q)} P\{y_s\}$. Indeed, tests show that one can admit that we use sums randomly chosen (cf also section 6.3). This shows that we can make assumptions weaker than K_0 not too large.

6 Argument for the choice of this measure

We have seen that if K_0 is not too large $P\{T_q(Y_n^\theta) = k/2^q | Y_{n-t}^\theta = y_t'\} = \sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s | Y_{n-t}^\theta = y_t'\} \approx 1/2^q$ and that, in order to prove this result, it suffices that this sum behaves like a random randomly chosen. This justifies the use of the measure $Proba_N$. We shall show why now.

6.1 Sets or Noises randomly chosen

That the sums behave as random is normal since it is what happens if the noise or the set E of size $|E| = d$ is chosen randomly : $E = \cup_i \{h_i/m\}$, $h_i \in \{0, 1, \dots, m-1\}$, $E \subset \{0, 1, \dots, m-1\}/m$. To say that E is chosen at random is translated in saying that the h_i 's are randomly chosen : $Prob\{h_i = k\} = 1/m$.

Suppose then we have a sample of y_n all distinct and consider the associated empirical probabilities $P_e\{y_n \in E\}$.

Let $\mathcal{Y} = \{y_1, \dots, y_N\}$. Then, with a high probability, $P_e\{y_n \in E\} = (1/N) \sum_{n=1}^N 1_E(y_n) = (1/N) \sum_{n=1}^N \sum_{i=1}^d 1_{h_i}(y_n) = (1/N) \sum_{n=1}^N \sum_{i=1}^d 1_{y_n}(h_i) = (1/N) \sum_{i=1}^d 1_{\mathcal{Y}}(h_i) \approx (d/N) \frac{|\mathcal{Y}|}{m} = \frac{|E|}{m}$. Therefore we have indeed $P_e\{y_n \in E\} \approx \frac{|E|}{m}$. We can show the same result for the empirical

conditional probabilities : $P_e(y_n \in E | y_{n-1} \in A_1, \dots, y_{n-p} \in A_p) \approx \frac{|E|}{m}$ with a high probability.

This will be the same thing if for a given set E , it is the noise y_n which is taken at random among the set of all noises. Indeed, to take a sequence y_n at random means that $\text{Prob}\{(y_1 = k_1/m, \dots, y_N = k_N/m)\} = 1/m^N$. Then, with a high probability, $P_e(y_n \in E | y_{n-1} \in A_1, \dots, y_{n-p} \in A_p) \approx \frac{|E|}{m}$.

This is especially true for texts: a text is a noise randomly chosen¹. Then, with a high probability, $P_e(y_n \in E | y_n \in A_1, \dots, y_n \in A_p) \approx \frac{|E|}{m}$.

6.2 Empirical Probability

Now, in order to study the behavior of empirical probabilities one can also use the following property (cf section 6.1.2 of Blacher (2010)).

Property 6.1. Let $z_n \in \{0, 1, \dots, m-1\}$, $n = 1, 2, \dots, n_0$, be a sequence of integers such that all the z_n 's are different. For all the sets not too small $F \subset \{0, 1, \dots, m-1\}$ (except a tiny minority), $\frac{1}{\text{card}(F)} \sum_{n \in F} z_n \approx L = \frac{1}{n_0} \sum_{n=1}^{n_0} z_n$.

Moreover, for almost all sets $F_s \subset \{0, 1, \dots, m-1\}$, $s=1, 2, \dots, P$, not too small such that $F_s \cap F_t = \emptyset$ if $s \neq t$, $\left\{ \frac{1}{\sqrt{\text{card}(F_s)}} \sum_{n \in F_s} (z_n - L) \right\}_{s=1, 2, \dots, P}$ behaves approximately as a Gaussian sample.

Then $\sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s\}$ converges to the same limit L for all $k \in \{0, 1, \dots, 2^q - 1\}$. At last, $\sum_k \sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s\} = 1$. Therefore, $\sum_{\gamma_s \in \mathcal{G}_k} P\{Y_n^\theta = \gamma_s\} \approx 1/2^q$.

We have seen that when K_0 is not too large, the sums of probabilities behave as randomly selected, and in this case, $T_q(Y_n)$ is always approximately IID.

Now the same thing happens to models not checking K_0 not too large. Indeed if we try to imagine other models, we obtain the same type of results except if conditional probabilities are concentrated near some points only. In fact, in order that conditional probabilities do not behave as randomly selected, it is necessary that they are specially built for this : for example that the conditional probabilities are concentrated near n_1 points which are precisely the points of $T_q^{-1}(k/2^q)$.

Like most models not too deterministic seem to verify that the sums of conditional probabilities behave like sums randomly selected, it is clear that the only thing which can distinguish that empirical probabilities $P\{T_q(Y_n^\theta) = k/2^q | Y_{n-t}^\theta = y_t'\}$ does not behave as randomly chosen is the behavior of the samples y_n . The only way to know if the y_n behave as predicted by the model is to do tests.

6.3 Tests

Therefore we tested the empirical probabilities $P_e(y_n \in E | y_{n-s} \in A_s, s = 1, \dots, p)$, $p=0, 1, 2$, for different sets $E = T_q^{-1}(k/2^q)$ when the noises are texts.

In this regard, we must first make a preliminary observation: take the empirical probability implies that there have some convergence properties. This eliminates a number of models. But if we take the models without these properties, it is much more difficult to obtain conclusions.

For example, let $c_j \in \{0, 1, \dots, 31\}$, $j = 1, 2, \dots, N_3$, be a sequence of letters². One gathers the

¹ We mean that text can be regarded as a realization of some sequences of random variables. It is not surprising : there is no logical connection between the text and the distribution of the points of $\mathcal{G}_k$: the sequence $\mathcal{G}_k$ is built by a specific and relatively simple mathematical application whereas the data y_n are the realization of a specific sequence of random variables $Y_n^{\theta_0}$ unpredictable in an exact way.

² 32 corresponds to 26 letters, punctuations, signs and space : cf Blacher (2009).

letters by group of $r_0 : d_n = \sum_{r=1}^{r_0} c_{r_0(n-1)+r} 32^{r-1}$. Let m such that $m = f_{i_{n_0}} \geq 32^{r_0} > f_{i_{n_0-1}}$. We set $y_n = \lfloor md_n/32^{r_0} \rfloor$, the integer part of $md_n/32^{r_0}$. We made tests about sequences y_n , $n=1, \dots, N$, when $N = 10^7$.

For example, for $m \approx 10^{20}$ for different $E_k = T^{-1}(k/2^q)$ and different A_s , we are interested by $P_e\{y_n \in E | y_{n-1} \in A_1 \cap \dots \cap y_{n+p} \in A_p\}$.

When $p=0$ (i.e. for $P_e\{y_n \in E\}$), we made a chi-squared test of normality with 20 degrees of freedom for the sample provided by $P_e\{y_n \in E_k\}$, $k=1, \dots, 500$ with $q=18 : 2^q = 262144$ and $10^7/2^{18} = 38.146972656250000$. We have obtained 17.321.

When $p=1$, we made a chi-squared test of normality with 5 degrees of freedom about the $P_e(Y_n \in E | Y_{n-1} \in A_1)$'s with $q=15 : 2^q = 32768$ for the sample provided by 100 different k for different $A_1 = [h/10, (h+1)/10[$ for different h (one test for each h). We have obtained 4.111, 7.211, 3.1181, 5.24, 5.78,

When $p=2$, we made a chi-squared test of normality with 5 degrees of freedom about the $P_e(Y_n \in E | Y_{n-1} \in A_1, Y_{n-2} \in A_2)$'s for the sample provided by 100 different k for different A_1, A_2 with $q=12 : 2^q = 32768$ and $A_s = [h/10, (h+1)/10[$ for different h . We have obtained 3.789, 4.548, 5.468, 7.01, 6001.

We also calculated the number of points belonging to $T_q^{-1}(k/2^q)$ when $2^q \gg N$. So we find N sets $T_q^{-1}(k/2^q)$ containing one point and the rest $2^q - N$ containing 0 points.

As we can see the test of normality is accepted, which allows to admit that the probability that $y_n \in E$ behaves as being independent of E when $|E|$ is constant.

Then, the measure on the Y_n^θ 's corresponds indeed to the properties which must be checked by the models.

6.4 Study of the measure $Proba_N$

Indeed, the measure which we use is a well chosen measure. For example, under notations 4.2,

$$P\{T_q(Y_2^\theta) = k/2^q | Y_1^\theta = y_1\} = \frac{\sum_{y_2 \in T_q^{-1}(k/2^q)} p_{y_1, y_2}^\theta}{\sum_{i_2=1}^m p_{y_1, i_2/m}^\theta}.$$

Of course, with this measure, $\sum_{y_2 \in T_q^{-1}(k/2^q)} p_{y_1, y_2}^\theta$ is a sum randomly chosen.

Now, we have seen that for most models, the sums of $P\{T_q(Y_n^\theta) = \gamma_s | Y_{n-1}^\theta = y_1'\}$ behave as being randomly selected. The measure chosen behaves exactly as one would wish. We can conclude that this is a good measure in order to quantify the proportion of model Y_n^θ such that $T_q(Y_n^\theta)$ behaves approximately as IID.

Note also that, because of these properties one can admit $P\{T_q(Y_{\phi(n)}^\theta) = k/2^q | Y_{\phi(n-1)}^\theta = y_1', Y_{\phi(n-2)}^\theta = y_2', \dots\} = (1/2^q)[1 + Ob(1)\epsilon]$ where $|\epsilon| \leq \epsilon'_{(1)} \approx \frac{b\sigma_U}{E_U \sqrt{h_0}}$. We deduce by the same way as in Blacher (2011) (cf also section 5), that, for all $I_{k_1} \otimes \dots \otimes I_{k_N}$, $P\{(X_1^\theta, \dots, X_N^\theta) \in I_{k_1} \otimes \dots \otimes I_{k_N}\} = \prod_{s=1}^N (L(I_{k_s})[1 + \epsilon_s])$, where $|\epsilon_s| \leq \epsilon'_{(1)}$. Then, $P\{(X_1^\theta, \dots, X_N^\theta) \in Bo\} = L(Bo)[1 + Ob(1)N\epsilon_{(1)}]$, with $\epsilon_{(N)} = N\epsilon_{(1)}$. This also is in conformity with our assumptions.

Then, to use measure $Proba_N$ means each probability $p_{y_1, \dots, y_n}^\theta$ is randomly chosen. That means that the noises which are too deterministic are a negligible set with this measure. They are these noises for which $T_q(Y_n^\theta)$ can be not IID.

Indeed, a way in order to understand what we want to mean when we say "noises not too deterministic" is to compare with min-entropy : $H_\infty(X) = \min_x \log\left(\frac{1}{P\{X=x\}}\right)$. Indeed, by Shaltiel

(2011), if $H_\infty(X) \leq qN$, T_q cannot be an 0-extractor : i.e. a necessary condition in order that $T_q(Y_n^\theta)$ is IID is that $P\{(Y_1^\theta, \dots, Y_N^\theta) = (\eta_1, \dots, \eta_N)\} \leq \frac{1}{2^{qN}}$ for all $m(\eta_1, \dots, \eta_N) \in \{0, 1, \dots, m-1\}^N$. If it is not case, we think that "the noise y_n is too deterministic".

For example, assume $m = O(4 \cdot 10^{36})$, $q = 33$ and $N_0 = 10^5$. Then, if y_n is not too deterministic, $P\{(Y_1^\theta, \dots, Y_N^\theta) = (\eta_1, \dots, \eta_N)\} = O(\beta)$ where $\beta = 1/m^N$. Then, if $H_\infty(X) \leq qN$, $\frac{m^N}{2^{qN}} = \left(\frac{m}{2^q}\right)^N \geq 5^{33N}$. Then, there exists $(\eta_1^0, \dots, \eta_N^0)$ such that $P\{(Y_1^\theta, \dots, Y_N^\theta) = (\eta_1^0, \dots, \eta_N^0)\} = O(5^{33N}\beta)$. Of course, in this case, there is a very big breakdown of randomness and one can say that Y_N^θ is very deterministic.

In conclusion, the measure defined in notations 4.5 adapts well to the results obtained and it is quite selected.

Let us notice that these theorems remain the same one for other functions than the functions T_q . For example, let T' be a congruence and $T'_q(x) = Pr_q(T'(xm)/m)$. Now assume that y_n is a text and that T' is the identity. Then the measure defined by the notation 4.5 is not well chosen. In this case, there is connection between the text and the sets $T_q'^{-1}(I)$ where I is an interval: $T_q'(I)$ represents generally the first letters of the text. The reasoning thus does not apply and the measure defined on Ωm_N is thus not interesting : cf also pages 200-202 of Blacher (2009).

7 Fibonacci functions and extractors

7.1 Functions T_q and extractors

We recall that a ϵ -extractor E for a distribution B over $\{0, 1\}^n$ is an application $E : \{0, 1\}^n \rightarrow \{0, 1\}^M$, $n \geq M$, such that $E(B)$ is ϵ -close to U when U has the uniform distribution on $\{0, 1\}^M$, i.e. $|P\{E(B) \in Bo\} - P\{U \in Bo\}| \leq \epsilon$ for all Borel set Bo . Clearly this definition can be extended by a natural way to the case $\{0, 1, \dots, m-1\}^n \rightarrow \{0, 1, \dots, 2^q-1\}^N$. Of course, functions T_q can be considered as extractors.

Note first that the condition " K_0 not too large" has a form similar to the condition about sources of Santha and Varizani : $p_0 \leq P\{B_i = 1 | B_1 = b_1, \dots, B_{i-1} = b_{i-1}\} \leq 1 - p_0$ (cf Santha, Vazirani (1986)). But it is clearly different: in this case, K_0 can be large.

At last, in theory of extractors, ϵ -close distributions are used and in our method, we use distributions ϵ -equivalent what is a stronger condition.

Now we are interested in the behavior of the measure $Proba_N$ relative to extractors. Remember at first that it is commonly accepted that it is an application random chosen among all possible applications $\{0, 1\}^{\ell m'} \rightarrow \{0, 1\}^n$ which gives the best result as multisource extractor : Pansu (2008).

Now let $E^{-1}(k/2^q) = \{b_s \in \{0, 1\}^{m'}\}$. So if we choose randomly extractors $\{0, 1\}^{m'} \rightarrow \{0, 1\}^n$, $m' > n$, for a distribution B , the b_t 's can therefore be regarded as random. So the $P\{b_t\}$'s can therefore be regarded as randomly chosen in the complete set $\{P\{b_s\} | s \in \{0, 1\}^{m'}\}$.

Now, $P\{E^{-1}(k/2^q)\} = \sum_s P\{b_s\}$. We can apply the same reasonings as those we have used above and get the same type of results. They show indeed that extractors randomly chosen are the best extractors possible.

Now we have seen that $T_q^{-1}(k/2^q) = \cup_s \{\gamma_s\}$ and $P\{T_q^{-1}(k/2^q)\} = \sum_s P\{\gamma_s\}$ where the γ_t 's can therefore be regarded as randomly chosen. The functions T_q behave therefore in the same way as extractors randomly chosen, what explains their performances. We have already studied in a way more detailed, but still incomplete, the connection between the sets $T_q^{-1}(k/2^q)$ and distributions

in Blacher (2009). We will complete this question in a more detailed way later.

So here we see some connections between Fibonacci functions and basic properties of extractors. In fact, the extractors are rather used in cryptography. So far we have mainly studied the statistical aspect of the distributions ϵ -equivalent proving that, practically the $T_q(Y_n)$'s can not be differentiated from IID sequences. But the question arises whether the functions T_q can be used in cryptography. It is currently being studied.

8 Building of IID sequences

8.1 Choice of parameters

We are interested in the choice of parameters in the building of IID sequences. Then, we consider the case of sequences of noises $y'_n \in \{0, 1, \dots, m-1\}$, $n = 1, 2, \dots, N_0$. We set $y_n = \lfloor y'_n + \text{rand}_0(n) \rfloor / m$ where $\text{rand}_0(n)$ is pseudo random generator¹. Then, by applying Remark 4.3, we shall impose that

$$\sqrt{\frac{N_0^3 q 2^q}{m}} \leq 1/N_0.$$

Remark that this condition is less good than the condition used when one impose that the Lipschitz coefficients K_0 are small : $\frac{\gamma(m)K_0N_02^q}{m} \leq 10^{-3}$ (cf Blacher (2011)).

8.2 Example

By using this technique, we have created real sequences $xe_n = x_n$, $n=1,2,\dots,100000$: cf also page 93 of Blacher (2010). One can obtain it in a Matlab file in Blacher (2009). In this case sequences y_n , $n=1,2,\dots,100000$, is obtained from texts : dictionary, encyclopedia, and Bible.

Let fi_n be the sequence of Fibonacci. We choose $m = fi_{n_2}$ when $fi_{n_2-1} < 4 * 10^{36} \leq fi_{n_2}$, $q = 33$ and $N_0 = 10^5$. For all the models $Y_n^{\theta_{\omega N_0}}$ with a probability larger than, $1 - e^{-10^{12610}}$ in Ωm_{N_0} , $|P\{(X_1^{\theta_{\omega N_0}}, \dots, X_{N_0}^{\theta_{\omega N_0}}) = (k_1/2^{33}, \dots, k_{N_0}/2^{33})\} - 1/2^{33N_0}| \leq 0.000008418/2^{33N_0}$.

8.3 Tests

We have tested the sequence $xe(n)$. We have used the classical Diehard tests (cf Knuth (1998), Gentle (1984)), and the higher order correlation coefficients (cf Blacher (1993)). Results are in accordance with what we waited: the hypothesis "randomness" is accepted by all these tests.

For example, we have used the gap test. With the notations of Knuth (1998) page 62, we choose $[\alpha, \beta[= [0, 1/2[$. Then, one uses the chi squared statistics with t degrees of freedom (with the notations de Knuth (1998)) : we denote them by $\chi_{N_1}^2$. One chooses $t = 5, 6, 7, 8, 9, 10$. We test 100 samples with various sizes N_0 : $N_0 = 10^5, 10^4, 10^3$. We are interested in the maximum of these 100 $\chi_{N_1}^2$.

The following table is that of the maxima of $\chi_{N_1}^2$ obtained for each t . Moreover α_0 is the selected percentage points with probability 0,95 of chi squared statistics.

t	5	6	7	8	9	10
$Max(\chi_{N_1}^2)$	11.21	13.1	13.22	17.08	16.87	17.22
α_0	11.07	12.59	14.07	15.51	16.92	18.31

¹ We use $\text{rand}_0(n)$ because this method decreases also K_0 (cf Blacher (2011) or Blacher (2011)) in the case where we can use theorem 1. By the study made here, there would be no need to add $\text{rand}_0(n)$. But it can improve the approximation in many cases. There is no reason to go without since our goal is to get sequences proved IID with the greatest possible safety.

Because, we used the maximum of the statistics associated in these 100 tests, it is not thus surprising that they are sometimes larger than α .

Remark that we have carried out tests with samples built by adding $rand_0(n)$, but also by removing it : $y_n = (y'(n) + rand_0(n))/m$ or $y_n = y'(n)/m$. We have carried out 100 tests in each category with significance level 5/100. We denote by n_{tp}^r and n_{tf}^r , n_{tp} and n_{tf} , the number of tests passed or failed, with or without the pseudo random generator $rand_0(j)$. Then, we have the following results.

	n_{tp}	n_{tf}	n_{tp}^r	n_{tf}^r
Equidistribution Test	97	3	95	5
Serial Test	95	5	93	7
Poker Test	96	4	96	4
Coupon collectors Test	94	6	95	5
Run Test	91	9	93	7
Maximum-of-t Test	97	3	92	8
Collision Test	94	6	97	3
Birthday spacings Test	93	7	95	5
Serial Correlation Test	98	2	92	8
Higher order correlation coefficient Test	94	6	97	3

As a matter of fact, we have carried out many of other tests during this study. We have given some results in page 345-349 of Blacher (2009) ¹.

9 Conclusion

Theorem 2 means that there is no need that Y_n^θ has a Lipchitz coefficient not too large in order that $T_q(y_n)$ can be regarded as a sample of an IID sequence. Therefore, most of the sequences of noises y_n can be used to create IID sequences $x_n = T_q(y_n)$.

As a matter of fact it is enough that y_n is a noise not too deterministic. For example, one can choose the following noises

- 1) Texts or transformed texts (cf Blacher (2011))
- 2) Files systems of the computers
- 3) Music (e.g. Rap music as Marsaglia (1995))
- 4) Elapsed time between emission of particles during radioactive decay
- 5) Thermal noise from a semiconductor diode or resistor
- 6) The frequency instability of a free running oscillator
- 7) Quantum phenomena

References

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

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

¹ In Blacher (2009), the numbers are obtained with a method slightly different. In Blacher (2009), we do not have unfortunately noted all the results and all the parameters (number and size samples, etc).

- MENEZES A., VAN OORSCHOT P. , VANSTONE S. (1996) Handbook of Applied Cryptography, CRC Press, 1996.
- SCHNEIER B (1996) Applied Cryptography 2nd Edition, John Wiley and sons, Inc
- BLACHER R. (1993) Higher Order Correlation Coefficients. Statistics 25, 1-15.
- BLACHER R. (2004) Solution complete au probleme des nombres aleatoires. Journes statistiques de Montpellier. <http://www.agro-montpellier.fr/sfds/CD/textes/blacher1.pdf>
- BLACHER R. (2009) A Perfect Random Number Generator. Rapport de Recherche LJK Universite de Grenoble. <http://hal.archives-ouvertes.fr/hal-00426555/fr/>
- BLACHER R. (2010) A Perfect Random Number Generator II. Rapport de Recherche LJK Universite de Grenoble. <http://hal.archives-ouvertes.fr/hal-00443576/fr/>.
- BLACHER R. (2009) File of random Number.
<http://www-ljk.imag.fr/membres/Rene.Blacher/GEAL/node3.html>.
- BLACHER R. (2010) Correct models. Rapport de Recherche LJK Universite de Grenoble.
<http://hal.archives-ouvertes.fr/hal-00521529/fr/>
- BLACHER R. (2011) Fibonacci congruences and applications. Rapport de Recherche LJK Universite de Grenoble. <http://hal.archives-ouvertes.fr/hal-00587108/fr/>.
- BLACHER R. (2011) Fibonacci congruences and applications. Open Journal of Statistics, Vol 1 n2, <http://www.scirp.org/journal/PaperInformation.aspx?paperID=6551>
- BLACHER R. (2011) A method for Building true random sequences. Far east Journal of Theoretical Statistics. Volume 36 n2, 75-104.
- BLACHER R. (2013) Dependence of two simulated generated by congruences. Research Report <http://hal.archives-ouvertes.fr/docs/00/82/13/91/PDF/GealRapportFibo8313.pdf>
- BLACHER R. (2013) Proved random numbers obtained from hardware devices. Communications in Statistics - Simulation and Computation. <http://www.tandfonline.com/doi/abs/10.1080/03610918.2012.725143>
- JOHNSON N.L. KOTZ S. (1970) Continuous univariate distributions. Wiley, New York
- 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
- 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.
- MARSAGLIA G (1995) CD ROM. Florida State University, <http://stat.fsu.edu/pub/diehard/>
- VON NEUMANN J . (1951) Various techniques used in connection with random digits. Monte Carlo method, Applied Mathematics series N12, US National Bureau of Standards, Washington DC, 36-38.
- SANTHA M. , VAZIRANI U. V. (1986). Generating quasi-random sequences from semi-random sources, Journal of Computer and System Sciences, v.33 n.1, p.75-87.

- ELIAS P. (1972) The efficient construction of an unbiased random sequence. *Annals Math Stat*, vol 43, n3, 865-870.
- BLUM M (1984) Independent unbiased coin clips form a correlated biased source : A finite state Markov chain. *Proc. 25 th FOCS Florida* 425-433.
- CHOR B. GOLDREICH O. (1985) Unbiased bits from weak sources of randomness *Proc 26th FOCS Portland Oregon* 429-442.
- SHALTIEL R. (2011) An introduction to randomness extractors, invited paper for ICALP 2011. : [http : //cs.haifa.ac.il/ ronen/online_papers/ICALPinvited.pdf](http://cs.haifa.ac.il/~ronen/online_papers/ICALPinvited.pdf)
- BARAK.B IMPAGLIAZZO R. WIGDERSON A. (2006) Extracting randomness using few independent sources. *SIAM J. Comput.*, 36(4):10951118, .
- BOUGAIN J. (2007) On the construction of affine extractors. *Geometric And Functional Analysis*, 17(1):3357.
- CHOR B. GOLDREICH O. (1988) Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230261, April 1988. Special issue on cryptography.
- GABIZON A. RAZ R.(2008) Deterministic extractors for affine sources over large fields. *Combinatorica*, 28(4):415440.
- (2006) GABIZON A. RAZ R. SHALTIEL R. Deterministic extractors for bit-xing sources by obtaining an independent seed. *SIAM J. Comput.*, 36(4):10721094.
- TAUMAN KALAI Y. LI X. RAO R. (2009) 2-source extractors under computational assumptions and cryptography with defective randomness. In *FOCS*, pages 617626.
- KAMP J. RAO A. VADHAN S. P. ZUCKERMAN D. (2011) Deterministic extractors for small-space sources. *J. Comput. Syst. Sci.*, 77(1):191220.
- KAMP J. ZUCKERMAN D. (2007) Jesse Kamp and David Zuckerman. Deterministic extractors for bit-xing sources and exposure-resilient cryptography. *SIAM J. Comput.*, 36(5):1231 1247.
- LI X.(2011) Improved constructions of three source extractors. In *IEEE Conference on Computational Complexity*.
- NISAN N. TA-SHMA A. (1999) Extracting randomness: A survey and new constructions. *J. Comput. Syst. Sci.*, 58(1):148173.
- PERES Y. (1992) Iterating von neumanns procedure for extracting random bits. *Ann. Statist.*, 20:590597.
- RAS R.(2005) Extractors with weak random seeds. In *STOC*, pages 1120.
- RAS R. REINGOLD O. VADHAN S. P. (2002) Extracting all the randomness and reducing the error in trevisans extractors. *J. Comput. Syst. Sci.*, 65(1):97128.
- PANSU P. (2008) Application des estimées sommes-produits aux extracteurs, d'après Barak, Impagliazzo et Wigderson. [http://www.math.u-psud.fr/ pansu/28jan08.pdf](http://www.math.u-psud.fr/~pansu/28jan08.pdf)
- ZAREMBA S. K. (1966) Good lattice points, discrepancy and numerical integration. *Annali di Mathematica Pura ed Applicata*. IV 73, p 293-317.

-
©2013 René Blacher; This is an Open Access article distributed under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/2.0>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.