THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 21, Number 1/2020, pp. 37–44

COSTS AND BENEFITS OF ADAPTIVE TESTING IN ASSESSMENT OF RANDOM NUMBER SEQUENCES

Kinga MÁRTON, Laurențiu LAZAR, Adrian COLEȘA, Alin SUCIU

Technical University of Cluj-Napoca, Computer Science Department, Cluj-Napoca, Romania
Corresponding author: Kinga MÁRTON, E-mail: kinga.marton@cs.utcluj.ro

Abstract. In the process of randomness assessment, a large number of statistical tests are applied on sequences produced by random number generators resulting in a vast number of p-values, which need to be carefully interpreted in order to accurately evaluate the quality of randomness. The interpretation of statistical results is a rather complex task, especially because we cannot draw a rigorous line and classify random number sequences only based on a p-value being inside or outside the region defined by a significance level. P-values falling outside the rejection region but close to the significance level are called *suspicious*, because it is not clear whether such a sequence appears by chance or it is an indication of a deviation from randomness in the generator. Adaptive testing strives to classify suspicious p-values by analyzing the sequence failing close to the significance level in a larger context, by iteratively increasing the sample size and reapplying the test. By conducting a thorough analysis of suspicious p-values, in this work we aim at highlighting the benefits together with the costs and the inherent limitations of adaptive testing. Through empirical experiments, we attempt to answer some of the difficult questions regarding adaptive testing, such as the required number of iterations for a relevant classification of suspicious p-values, the choice of the initial sample size, and the dependency between the chosen significance level, sample size and the costs in execution time.

***Key words**:* adaptive testing, random number sequences, statistical assessment, suspicious p-values, interpretation of results.

1. INTRODUCTION

The quality assessment of random number generators heavily relies on statistical hypothesis testing, where a set of selected statistical tests are applied on a large number of sequences produced by the tested generator. Each test when applied to a sample sequence of random numbers, summarizes the result in the form of a p-value, a probability value expressing the likelihood of finding identical or more extreme results if the null hypothesis holds, or in other words, the probability of a high quality random number generator to produce sequences with similar or weaker randomness properties. Then, based on the comparison with a chosen significance level α (generally selected as 0.001 or 0.01) the test concludes whether the p-value provides evidence against the null hypothesis, H_0.

In case of one tailed statistical tests, such as those included in the NIST statistical test suite [1], H_0 is rejected if the p-value is less than the significance level α, and in case of the two-tailed tests, such as the tests included in the TestU01 testing library [2], H_0 is rejected if the p-value is outside the interval $[\alpha, 1-\alpha]$, showing inclination towards extreme probabilities. The interpretation of p-values falling farther away from the border lines is rather obvious in either direction, however, p-values falling in the rejection region but close to the significance level are called *suspicious*, because it is not clear whether such a sequence appears by chance or it is an indication of the lack of randomness in the generator. Adaptive testing focuses on analyzing these suspicious p-values in order to highlight whether these values are statistically significant or should be considered a normal effect of chance.

Adaptive testing is not a new concept. We can find references to resolving the troublesome nature of suspicious p-values in the work of L'Ecuyer [2] or Accardi [3] but mainly as a theoretical recommendation, and in the work of Haramoto [4], where the effectiveness of adaptive testing is empirically illustrated. Our

present research work provides a more thorough analysis of suspicious p-values, highlighting the benefits together with the costs and the inherent limitations of adaptive testing. We aim to answer some of the difficult questions regarding adaptive testing, such as the required number of iterations for a relevant classification of suspicious p-values, the choice of the initial sample size, or the extent of the dependency between the chosen significance level, sample size and the costs in execution time. Empirical results are obtained by integrating adaptive testing into the Rabbit battery from the TestU01 testing library [2].

The rest of the paper is structured as follows. Section 2 provides a short introduction to the process of adaptive testing. Section 3 presents the proposed strategy together with rich experimental results showing both the effectiveness and costs of adaptive testing and is followed in Section 4 by the conclusions.

2. ADAPTIVE TESTING

Considering that a high-quality random number generator produces all possible sequences of a given length with the same probability, it is natural to find some sequences that do not seem random and for which statistical tests produce p-values in the rejection region. In fact, by choosing a certain significance level α we expect two sequences (or one for one tailed tests) from $1/\alpha$ sequences to fail the test by chance, in other words, we expect sequences to fall inside the rejection region with a probability of 2α (α for a one-tailed test) and still accept the null hypothesis.

Considering a two-tailed test, if the p-value is very close to 0 or 1, such as values less than 10^{-8} or larger than $1-10^{-8}$, then the result is regarded as statistically significant and the sequence fails the test. On the other hand, if the p-value is within the interval defined by $[\alpha, 1-\alpha]$ the sequence passes the test and H_0 is accepted. *Suspicious p-values* are those falling very close to the threshold values α and $1-\alpha$. If we consider the significance level as $\alpha = 0.001$ and a limit for definitive rejection $\rho = 10^{-8}$, then a p-value of 10^{-4} is regarded as being suspicious, as it falls in the rejection region but does not provide powerful evidence against H_0 which would indicate clear rejection. The limit of definitive rejections can vary; values used in the literature are 10^{-10} in [2], 10^{-8} in [4]. In this paper, we analyzed the comparative behavior of adaptive testing using two different limits for definitive rejection, namely 10^{-6} and 10^{-8}.

In adaptive testing, when a test returns a suspicious p-value for a sample sequence then the sample is extended (usually doubled) and the test reapplied iteratively on the samples of increasing lengths, resulting in a new p-value at each iteration. Based on this sequence of p-values the nature of the original sequence can be analyzed and classified accordingly as being significant, appearing by chance or remaining suspicious.

Therefore, if the tested file of length M is divided into sub-sequences of N bytes denoted by $\{S_1, S_2, \ldots, S_i, \ldots S_n\}$ and a test T is applied on all the sub-sequences, the result is a set of $n = M/N$ p-values, namely $\{ PV_1, PV_2, \ldots, PV_i, \ldots PV_n\}$. If any of the p-values is classified as being suspicious, then adaptive testing can be configured to analyze the specific sub-sequence and the nearby region iteratively. Considering PV_i as a suspicious p-value, an increment of N bytes and maximum number of 3 iterations, the first iteration expands the subsequence length to $2N$ bytes (S_i and S_{i+1} now forms the expanded sequence) and T is reapplied on the expanded sequence resulting in $PV_i(1)$. The second iteration expands the sequence with the following N bytes and T returns $PV_i(2)$, then the third iteration expands the sequence with the following N bytes (now S_i, S_{i+1}, S_{i+2} and S_{i+3} all form together the expanded sequence) and applying T results in $PV_i(3)$. The expansion of a subsequence, as previously described, is done towards the end of the sequence, but other expansion strategies could also be considered, such as towards the beginning of the sequence, or in both directions symmetrical/asymmetrical with respect to the subsequence.

Considering the level of significance α, and the limit of definitive rejection ρ (e.g. 10^{-6} or 10^{-8}), the p-values obtained at each iteration are analyzed and classified accordingly:

- p-values below ρ and above $1-\rho$ are considered significant,
- p-values within the thresholds $[\alpha, 1-\alpha]$ are considered positive/accepted, and
- p-values between $[\rho, \alpha)$ and $(1-\alpha, 1-\rho]$ are regarded as suspicious.

By analyzing and classifying the set of p-values obtained at each iteration we get a clearer view on the nature of the initial sequence S_i and can decide on whether the suspicion on the resulted p-value can be lifted or it persists after several iterations. Thus, the testing process can be iterated depending on the user's preferences until either of the following conditions is satisfied:

- the first clear evidence is found for rejecting the sequence: the p-value obtained at the last iteration is statistically significant as it falls in the clear rejection region;
- a powerful evidence is reached for rejecting the sequence: the p-values of the last couple of iterations are significant;
- the first sign of melioration is recorded: the p-value obtained in the last iteration is acceptable as it falls within the threshold values determined by the significance level;
- the evidence of sustained amelioration is reached: the p-values of the last couple of iterations are all within the acceptable region;
- the specified number of iterations is reached: and a clear conclusion cannot be reached because the p-values are still suspicious, or their situations vary from iteration to iteration.

There are several aspects that need to be addressed regarding the adaptive testing procedure in order to capture its true merits and limitations, and the next section combines the discussion of these aspects with experimental results illustrating the practical behavior of different approaches in adaptive testing.

3. EXPERIMENTAL STRATEGY AND RESULTS

In order to have a thorough understanding of the impact adaptive testing can have on randomness assessment we have conducted an extensive statistical assessment using the Rabbit battery of statistical tests on a series of input data comprised of 23 000 MB of sequences generated using a mixed set of 14 pseudo and true random number generators.

Rabbit is a battery of statistical tests which can be applied on binary files and is included in the TestU01 statistical testing library [2]. The battery contains 26 statistical tests which apply a number of 40 test statistics to the input bit sequences. Most of the test statistics require at least 500 bits of data and the parameters of each test are chosen automatically as a function of the number of bits, in order to make the test reasonably sensitive. This characteristic makes Rabbit a perfect candidate for an adaptive testing strategy.

The generators considered in the experiments are from all the three major categories of random number generators: true-, pseudo- and unpredictable-RNGs, and are mentioned in the following:

- 6 PRNGs are provided by the TestU01 library [2], namely AES (uses the AES cipher as a source of random numbers), ISAAC (the version of ISAAC recommended for cryptography, with RANDSIZL = 8), CombWH3 (combines three LCGs using the method of L'Ecuyer), Mersenne Twister (the 2002 version of the Mersenne Twister generator designed by Matsumoto and Nishimura, which has a better initialization procedure than the original 1998 version), SHA1 (uses SHA-1 as a source of random numbers), XorShift13 (a full-period xorshift generator of order 8 with 13 xorshifts).
- 3 TRNGs provided by the following TPMs (Trusted Platform Modules): Dell E6420, Gigabyte GA-EP45-DS3R and HP DC7800.
- 2 PRNGs provided by programming languages: the *rand* function provided by the C standard library and the *SecureRandom* class provided by Java 7.
- 2 URNGs (unpredictable RNG): one provided by an operating system: /dev/urandom and HAVEGE, a user-level software URNG for general-purpose computers [5].
- A quantum TRNG: Qauntis designed by idQuantique [6].

From each of these generators we have obtained at least 1 000 MB of random bit sequences.

All results were obtained by averaging the run times of 1 000 adaptive tests with 25 iterations on a machine with Fedora 27, 64 bits version, having an Intel i5 CPU running at 2.60 GHz with 2 cores and 4 hardware threads, 2×32 KB L1 cache for data and 2×32 KB L1 cache for instructions, 2×256 KB of L2 cache, 3 MB of L3 cache, and 8 GB of RAM.

3.1. Adaptive testing strategy

The 23 input files of 1 000 MB each, obtained from the above-described generators, were analyzed considering two sequence sizes of 128 KB and 512 KB, therefore the original bit-sequences from the input files are split into 2 000 or 8 000 samples. Each one of the obtained sequences was then tested using the Rabbit battery, obtaining 38 p-values per sequence.

After all sequences were tested with Rabbit, the resulting p-values were classified using the approach described in Section 2 into *significant*, *positive/accepted*, or *suspicious*.

We have considered three values for the significance level α: 0.005, 0.0025 and 0.001, and ρ has 2 possible values: 10^{-6} or 10^{-8}.

For each suspicious p-value, the initial sequence is retested using the following adaptive strategy:

- only the test which leads to the suspicious p-value is reapplied,
- the sequence size is increased by its initial size,
- the resulting p-value is classified again,
- if the p-value is still suspicious, we continue retesting with increased sequence size up to 25 times.

Therefore, in this adaptive strategy the expansion is done by considering a larger sequence that includes the original sequence by adding increments of the same size as the original suspicious sequence. Instead, in the strategy proposed by Haramoto [4], where only PRNGs are considered, the suspicious sequence is resolved by testing newly generated sequences with doubled sample sizes.

3.2. Choosing the sequence sizes

Aiming to highlight the comparative behavior of adaptive testing on different input sizes, the experiments were conducted using two sequence sizes: 128 KB and 512 KB. The reason behind choosing these sample sizes has to do with the minimum amount of sample bits needed to apply all Rabbit test statistics and the amount of time needed to run adaptive tests on a large number of samples.

In order to cover 38 out of the 40 test statistics applied by Rabbit, the starting sequence size must be at least 36.68 KB. To cover all 40 test statistics, the starting sequence size would have to be at least 6.25 MB (the Matrix Rank tests require at least 625 KB and 6.25 MB sequences respectively).

Yet, the cost in execution time for applying 25 iterations of adaptive testing on only the first test of Rabbit (Multinomial Bits Over) using a starting sequence of 512 KB and 6000 adaptive targets would require up to almost 42 days of processing time. Hence, we decided to give up on the two tests which require sequences larger than 625 KB, and stick to the 38 test statistics that require less.

3.3. Adaptive testing in action

The first simple experiment demonstrates how adaptive testing works. We used a PRNG based on AES to generate a 10 MB file of random data. The file is split in sequences of 128 KB and each sequence is tested with Rabbit using an α value of 0.001 and an ρ value of 10^{-8}. The p-values obtained by applying the first test from Rabbit and three iterations of adaptive testing on two of these sequences can be seen in Table 1.

Table 1

Adaptive testing results for two suspicious 128 KB sequences

Test	Rabbit	Adaptive #1	Adaptive #2	Adaptive #3	Result
Sequence size	128 KB	256 KB	384 KB	512 KB	
1st	4.06×10^{-6}	5.46×10^{-4}	1.42×10^{-4}	2.50×10^{-1}	Accepted
2nd	$1-5\times10^{-6}$	$1-8.3\times10^{-8}$	$1-9.6\times10^{-12}$		Rejected

We observe that both resulting p-values are in the suspicion area, the first one is smaller than 0.001, but not in the clear rejection area below 10^{-8}, and the second one is greater than 0.999, but not as close to 1 as $1-10^{-8}$. So, both sequences / p-values are perfect candidates for adaptive testing.

We start the adaptive testing of the first sequence by increasing its size by another 128 KB, and obtain a sequence of 256 KB which includes the initial 128 KB sequence and the next 128 KB from the file. This larger sequence is tested again resulting in a p-value of 5.45×10^{-4} which is closer to our acceptance area than the initial p-value, but still suspicious. The procedure is repeated two more times. In the third iteration of our adaptive testing process, the sequence has 512 KB (the initial 128 KB and the next 384 KB from the file), and by applying the test again, we obtain a p-value of 2.50×10^{-1}. We accept the result as being positive, as it is inside the acceptance region, and stop the testing by succesfully classifying the sequence.

The second suspicious sequence goes through the same adaptive testing process. First, its size is increased by 128 KB, and is retested using only the first Rabbit test. This time, the resulting p-value gets closer to the clear rejection area, but is still suspicious. In the second iteration, the test result for the now 384 KB sequence is $1-9.6\times10^{-12}$, hence it is now considered significant, therefore the sequence is rejected, and the testing process is stopped.

3.4. Determining the required number of iterations

The second experiment is aimed at identifying the number of iterations needed by adaptive testing in order to classify the majority of suspicious p-values that result from the application of the Rabbit test battery. The input files consisting of 23000 MB of random data from a mixed set of 14 RNGs (described above) were tested in subsequences of 128 KB and 512 KB. After identifying all suspicious p-values the adaptive testing procedure was applied six times to each sequence - for every combination of ρ and α.

The aggregated results for sequences of size 128 KB is shown in Table 2. A number of 38627 adaptive tests were performed for each of the first three combinations having $\rho=10^{-6}$, and 45477 for each of the last three combinations having $\rho=10^{-8}$. Regardless of the values of ρ and α, the results show that a rather small number of iterations, between 2 and 3 iterations, are needed in order to classify up to 95% of all suspicious p-values. The first iteration of adaptive testing succeds in classifying between 73% and 87% of all suspicious p-values. Going above 95% starts to get a little harder. The number of required iterations to reach up to at least a 99.5% classification success rate for an α value of 0.001 is situated in the range of 9 to 12. Yet, 25 adaptive testing iterations were not enough to reach a 99.5% classification rate for an α value of 0.005. As anticipated, results indicate that a smaller value of α (a bigger acceptance area) and a larger value of ρ (a smaller suspicion area) requires fewer adaptive iterations to achieve a higher classification rate.

Table 2

The evolution of the percentage of classified suspicious p-values for 128KB sequences

		Iteration							
ρ	**α**	**1**	**2**	**3**	**9**	**12**	**18**	**23**	**25**
10^{-6}	**0.005**	76.7	88.4	95.4	98.3	98.9	99.2	99.3	99.4
	0.0025	80.4	94.4	96.7	99.1	99.3	99.5	99.6	99.6
	0.001	87.8	96.4	98.0	99.5	99.7	99.8	99.8	99.9
10^{-8}	**0.005**	73.5	88.2	95.3	98.0	98.6	99.0	99.2	99.2
	0.0025	76.7	93.4	96.5	98.8	99.1	99.4	99.5	99.5
	0.001	83.1	95.2	97.8	99.3	99.5	99.7	99.8	99.8

One notable example of the effectiveness of the adaptive testing procedure in bringing clarifications to the interpretation of statistical results is highlighted by the results for the AES generator. After testing each one of the 8000 sequences of size 128 KB with Rabbit using $\rho=10^{-6}$ and $\alpha=0.001$, a number of 304000 p-values are obtained. The majority of the resulted p-values, 99.6164%, are in the acceptance area, another 0.0155% are in the clear rejection area, and 1119 p-values representing 0.3681% are in the suspicion area. The rejected and suspicious p-values together represent 0.3796% of all p-values which is almost double than the allowed 0.2% which could fall due to natural chance. After performing one adaptive testing iteration on each suspicious p-value, that allowed us to classify 87.8% of all suspicious p-values, the percentage of accepted sequences raises to 99.93%. Now the number of rejected and still suspicious p-values together represent under 0.07% which is acceptable considering our α value.

Looking at the results from the perspective of the applied Rabbit tests, 60% of the 38 analyzed test statistics required 10 iterations or less to classify 100% of all tested sequences for a ρ value of 10^{-6}, regardless of the selected α. The percentage is even higher for an α value of 0.001 for which all suspicious p-values from 29 out of 38 (76%) test statistics can be successfully classified in 10 or less iterations. The percentage for $\rho=10^{-8}$ is lower; regardless of the chosen α value, 55% of test statistics require 10 iterations or less to classify 100% of p-values. The same tendency can also be seen here, a lower number of iterations is required when the acceptance interval is larger, and the suspicion area is smaller.

Great improvements can be achieved for the first test statistic of Rabbit, Multinomial Bits Over. 30% of all suspicious p-values for $\rho = 10^{-6}$ and $\alpha = 0.001$ are obtained from the first test statistics, and 100% of them are classified in 10 adaptive iterations. The procedure managed to classify all suspicious p-values from test statistic 6 (Lempel Ziv), representing approximately 9.5% of all suspicious values, in 3 to 7 iterations, regardless of the chosen ρ and α. Similarly, for test statistics 33 and 34 (RandomWalk1 $L = 1024$ R & C) in 4 iterations, cumulating 22.5% of all suspicious p-values.

The aggregated results for sequences of 512 KB can be seen in Table 3. A number of 4 756 adaptive tests were performed for each one of the first three combinations having $\rho = 10^{-6}$, and 4 860 for each one of the last three combinations having $\rho = 10^{-8}$. The results show a significant increase in the number of adaptive iterations required to obtain a classification rate of at least 95% of all tested p-values than in the case of 128 KB sequences. For 128 KB sequences, 3 adaptive iterations were enough to classify 95% of all p-values, regardless of ρ and α, but the 512 KB sequences require up to 12 iterations, ranging from 3 to 12 depending more tightly on the chosen test parameters. Nonetheless, reaching a 99.5% classification rate for a bigger sequence size requires a smaller multiplication factor compared to the number of iterations needed for obtaining a 95% rate. The number of iterations needed to classify 99.5% of all suspicious p-values when starting with sequences of 128 KB have to be increased at least 4.5 times (going up to 8 times for $\alpha = 0.005$), whereas the 512 KB sequences require at most 3 times more iterations to reach a 99.5% classification rate.

The results for sequences of 512 KB show the same tendency as the ones for 128 KB sequences. A bigger acceptance interval and a smaller suspicion area require fewer adaptive iterations to obtain a higher p-value classification rate.

Table 3

The evolution of the percentage of classified suspicious p-values for 512KB sequences

		Iteration												
ρ	**α**	1	2	3	4	5	6	7	9	10	12	14	15	25
10^{-6}	**0.005**	64.4	76.6	88.9	91.0	93.1	94.6	95.4	96.8	97.5	98.5	99.1	99.2	99.5
	0.0025	69.5	88.0	91.6	93.7	95.5	96.4	96.9	98.4	98.9	99.1	99.5	99.5	99.7
	0.001	82.9	92.0	95.0	96.3	98.0	98.6	98.9	99.4	99.5	99.6	99.7	99.7	99.8
10^{-8}	**0.005**	62.5	72.9	84.9	87.1	89.0	90.7	91.7	93.2	94.1	96.3	99.1	99.1	99.5
	0.0025	67.5	84.3	88.0	90.0	91.7	929	93.7	95.3	96.2	97.4	99.4	99.5	99.7
	0.001	80.7	88.4	91.6	93.1	95.1	96.1	96.7	97.5	97.9	98.7	99.7	99.7	99.8

The same improvements regarding the effectiveness of adaptive testing sequences can also be seen in the case of larger starting sequences, of 512 KB. After we apply the Rabbit tests battery to a 1 000 MB file generated using the TPM from a Gigabyte GAEP45-DS3R motherboard, split in 2 000 sequences, using a ρ value of 10^{-8} and an α value of 0.005, we obtain 72 000 p-values. The suspicious and rejected p-values amount to 1.3013%, a value which is greater than the allowed percentage of 1%. After 2 adaptive iterations on all suspicious p-values, the accepted p-values proportion raises to 99.03% which is acceptable under the test conditions.

Considering the same test on a 1 000 MB generated using Linux's /dev/urandom, but with an α value of 0.0025, we apply Rabbit on 2 000 512 KB sequences and obtain an acceptance percentage of 99.3211% from 72 000 p-values, and a 0.6789% of suspicious p-values, which is greater than the accepted 0.5%. 6 adaptive iterations raise the acceptance percentage to 99.5025%, and the rest of the suspicious p-values can be left unclassified.

Looking at the results from the perspective of the applied Rabbit tests, up to 71% of the 38 test statistics analyzed required 10 iterations or less to classify 100% of all adaptively tested sequences for a ρ value of 10^{-6} and regardless of the selected α. The percentage is even higher for an α value of 0.001 for which all suspicious p-values from 33 out of 38 (86%) test statistics can be successfully classified in 10 or less iterations. The percentage for $\rho = 10^{-8}$ is lower; regardless of the chosen α value, 68% of test statistics require 10 iterations or less to classify 100% of p-values. An increase of about 5 to 15% can be seen compared to the results obtained for 128 KB sequences.

3.5. Runtimes

In order to provide a meaningful insight into the costs, measured in processing time, of the adaptive testing process, we provide the time measurements in seconds for the most expensive tests applied by Rabbit. Table 4 shows the run times for several adaptive iterations applied on sequences of 128 KB, and an increment size of 128 KB per iteration. The most relevant numbers of iterations were chosen based on the results from the previous experiments. A 0 for the number of adaptive iterations means that the test includes just the initial application of Rabbit. Table 5 shows the same results, but for sequences of 512 KB with an increment size of 512 KB per iteration.

Table 4

Processing times for adaptive testing sequences of 128 KB using a number of iterations between 0 and 25

	Iteration							
Statistic name	**0**	**2**	**3**	**9**	**12**	**18**	**23**	**25**
#00 – MultinomialBitsOver	0.389s	2.478s	4.206s	23.926s	39.470s	1m 22.879s	2m 10.616s	2m 32.856s
#01 – ClosePairsBitMatch, $t=2$	0.004s	0.029s	0.050s	0.344s	0.612s	1.414s	2.370s	2.831s
#02 – ClosePairsBitMatch, $t=4$	0.004s	0.024s	0.041s	0.267s	0.470s	1.042s	1.699s	2.013s
#05-06 – LinearComp	0.122s	0.417s	0.585s	1.827s	2.576s	4.297s	5.940s	6.648s
#07 – LempelZiv	0.015s	0.094s	0.196s	1.253s	2.000s	4.769s	7.610s	8.746s
#09 – Fourier3	0.010s	0.064s	0.107s	0.615s	1.029s	2.162s	3.417s	3.998s
#25-29 – RandomWalk1	0.009s	0.052s	0.086s	0.472s	0.779s	1.616s	2.541s	2.969s
#30-34 – RandomWalk1 ($L=1024$)	0.009s	0.051s	0.084s	0.463s	0.763s	1.583s	2.489s	2.909s
#36-39 – RandomWalk1 ($L=10016$)	0.009s	0.051s	0.085s	0.462s	0.762s	1.577s	2.478s	2.895s

Following the adaptive testing strategy described in the above sections, the time cost of the whole testing procedure of our AES generator increased by about 16% when we included the adaptive testing of all suspicious p-values resulting from the initial Rabbit battery application. The time required to apply Rabbit with $\rho = 10^{-8}$ and $\alpha = 0.001$ to a 1 000 MB file split in 8 000 sequences (128 KB sequences) is somewhere around 1 hour and 17 minutes. The tests resulted in 304 000 p-values from which 1 160 were in the suspicion area, which required a further 11 minutes and 26 seconds for at most 25 adaptive iterations, classifying 99.4% of them. The same test applied to sequences of 512 KB (the file is split in 2 000 sequences) takes around 1 hour 11 minutes and 45 seconds for the initial Rabbit tests. A number of 72 000 p-values were obtained from which 302 were suspicious and required a further 11 minutes and 8 seconds of adaptive testing, which succeeded in classifying 98.3% of them.

Table 5

Processing times for adaptive testing sequences of 512 KB

	Iteration							
Statistic name	**0**	**2**	**3**	**4**	**6**	**10**	**15**	**25**
#00 – MultinomialBitsOver	1.718s	10.348s	17.264s	25.872s	48.398s	1m 54.138s	3m 54.504s	10m 4.674s
#01 – ClosePairsBitMatch, $t=2$	0.021s	0.163s	0.292s	0.463s	0.936s	2.395s	5.283s	14.829s
#02 – ClosePairsBitMatch, $t=4$	0.017s	0.125s	0.214s	0.331s	0.652s	1.675s	3.701s	10.418s
#05-06 – LinearComp	0.168s	0.631s	0.914s	1.228s	1.943s	3.695s	6.442s	13.639s
#07 – LempelZiv	0.102s	0.590s	1.149s	1.707s	2.821s	7.770s	15.438s	42.581s
#09 – Fourier3	0.043s	0.274s	0.456s	0.684s	1.275s	2.604s	6.582s	17.126s
#25-29 – RandomWalk1	0.035s	0.206s	0.341s	0.509s	0.946s	2.217s	4.552s	11.710s
#30-34 – RandomWalk1 ($L=1024$)	0.034s	0.202s	0.334s	0.499s	0.926s	2.171s	4.457s	11.464s
#36-39 – RandomWalk1 ($L=10016$)	0.034s	0.201s	0.333s	0.497s	0.922s	2.157s	4.426s	11.377s

4. CONCLUSIONS AND FUTURE WORK

In this research work, we provided a thorough analysis of adaptive testing, highlighting its benefits together with costs and the inherent limitations. We aimed to answer some of the difficult questions regarding adaptive testing, such as the required number of iterations for a relevant classification of

suspicious p-values, the choice of the initial sample size, or the extent of the dependency between the chosen significance level, sample size and the costs in execution time.

Empirical results are obtained by integrating adaptive testing into the Rabbit battery of statistical tests from the TestU01 testing library. Experiments were conducted using two sequence sizes: 128 KB and 512 KB, considering three values for the significance level α: 0.005, 0.0025 and 0.001, and 2 possible values for ρ, the limit for definitive rejection: 10^{-6} or 10^{-8}.

Results show that for 128 KB sequences, 3 adaptive iterations were enough to classify 95% of all p-values, regardless of ρ and α, but the 512 KB sequences require up to 12 iterations, ranging from 3 to 12 depending more tightly on the chosen test parameters.

Following the proposed adaptive testing strategy, the costs in processing time for the whole testing procedure increased in average by approximately 13.27% to 22.14% when testing 128 KB sequences, and by 15.58% to 36.22% when 512 KB sequences are used.

The effectiveness of adaptive testing in allowing to reach a clearer conclusion in randomness assessment is highlighted in many experiments. The suspicious p-values once classified enabled the tester to arrive at a more relevant conclusion regarding the quality of randomness in the tested sequences.

In the future we intend to analyze other expansion strategies, using different increment sizes and expansion directions (such as towards the beginning of the sequence, or in both directions symmetrical/asymmetrical with respect to the subsequence). Another interesting future development is the study of the evolution of suspicious p-values considering a larger number of iterations and performing the final classification only after the expanded sequence maintains its affiliation to a category in several successive iterations.

ACKNOWLEDGMENTS

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI-UEFISCDI, project number PN-III-P2-2.1-PED-2016-2073, within PNCDI III.

REFERENCES

1. A. RUKHIN, J. SOTO, J. NECHVATAL, M. SMID, E. BARKER, *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, Technical report – NIST Special Publication 800-22, National Institute of Standards and Technology, US, 2010.
2. P. LECUYER, R. SIMARD, *TestU01: A C library for empirical testing of random number generators*, ACM Transactional on Mathematical Software (TOMS), **33**, *4*, p. 22, 2007.
3. L. ACCARDI, M. GÄBLER, *Statistical analysis of random number generators*, Quantum Bio-Informatics IV – From Quantum Information to Bio-Informatics, **28**, pp. 117-128, 2011, World Scientific Publishing.
4. H. HARAMOTO, *Automation of statistical tests on randomness to obtain clearer conclusion*, Monte Carlo and Quasi-Monte Carlo Methods 2008, Springer, pp. 411-421, 2009.
5. A. SEZNEC, N. SENDRIER, *HArdware Volatile Entropy Gathering Expansion: generating unpredictable random number at user level*, INRIA Research Report, 2002.
6. idQuantique White Paper, *QUANTIS – When random numbers cannot be left to chance*, True Random Number Generator, Available online: https://marketing.idquantique.com/acton/attachment/11868/f-021f/1/-/-/-/-/Quantis%20QRNG_Brochure.pdf, December 2019.

Receiving August 18 2019