Statistical Testing of Randomness

Boris Ryabko
Institute of Computational Technologies of SB RAS
Novosibirsk state university
Novosibirsk, Russian Federation
Email: boris@ryabko.net

***Abstract*—The problem of constructing effective statistical tests for random sequences of binary digits is considered. The effectiveness of such statistical tests is mainly estimated on the basis of experiments with various random number generators. We consider this problem in the framework of mathematical statistics and find an asymptotic estimate for the p-value of the optimal test in the case when the alternative hypothesis is an unknown stationary ergodic source.**

I. Introduction

Random numbers find many applications in various fields of information technology including information protection systems, numerical methods, computer games and many others. In practice, random numbers are generated by so-called random number generators (RNGs) and pseudo-random number generators. The quality of the generators is checked using methods of statistical hypothesis testing. For example, NIST USA recommends 15 statistical tests for use in cryptographic applications [1].

Here we consider the problem of finding optimal tests in the case when the RNG is modeled by stationary ergodic sources. We found the following asymptotic solution to this problem: we first described the asymptotic behaviour of the p-value of the optimal test for the case where the probability distribution of the RNG is a priori known, and then described a family of statistical tests that have the same asymptotic estimates of the p-value for any distribution (which is not known in advance). More precisely, we showed that in both cases, with probability 1, $\lim_{n\to\infty} -\frac{1}{n}\log \pi_\tau(x_1x_2...x_n) = 1 - h(\nu)$, where $x_1x_2...x_n$ is the sample, τ is the test, $\pi_\tau(x_1x_2...x_n)$ is the p-value, and $h(\nu)$ is the Shannon entropy of the (unknown) RNG distribution ν. It turns out that asymptotically optimal tests with the required properties are known [2], [3], and are deeply connected with so-called universal codes. Note that nowadays there are many universal codes which are based on different ideas and approaches, among which we note the PPM universal code [4], the arithmetic code [5], the Lempel-Ziv (LZ) codes [6], the Burrows-Wheeler transform [7] which is used along with the book-stack (or MTF) code [8]–[10], the class of grammar-based codes [11], [12] and some others [13], [14]. All these codes are universal. This means that, asymptotically, the ratio of compressed file length to source file length goes to the smallest possible value, i.e. the Shannon entropy ($h(\nu)$) per letter.

The main idea of randomness tests based on universal codes is rather natural: try to "compress" a test sequence by a universal code: if the sequence is significantly compressed, then it is not random, see [2], [3] and a brief description in the next part.

The rest of the paper is organised as follows. The next section contains the necessary definitions and some basic facts. Sections III, IV are devoted to investigation of the Neyman-Pearson test and tests based on universal codes, correspondingly. The proofs are given in the appendix.

II. Definitions

1) The main notations: We consider RNG which generates a sequence of letters $x = x_1x_2...x_n, n \geq 1$, from the alphabet $\{0,1\}^n$. There are two following statistical hypothesis: the null hypothesis $H_0 = \{x \quad$ *obeys uniform distribution* (μ_U) *on* $\{0,1\}^n\}$ and the alternative hypothesis $H_1 = \bar{H}_0$, that is, H_1 is negation of H_0. Let T be a test. Then, by definition, a significance level α equals probability of the Type I error. (Recall, that Type I error occurs if H_0 is true and H_0 is rejected. Type II error occurs if H_1 is true, but H_0 is accepted.) Denote a critical region of the test T for the significance level α by $C_T(\alpha)$ and let $\bar{C}_T(\alpha) = \{0,1\}^n \setminus C_T(\alpha)$. (Recall, that for a certain $x = x_1x_2...x_n$ the hypothesis H_0 is rejected if and only if $x \in C_T(\alpha)$.)

Let us assume that H_1 is true and the investigated sequence $x = x_1x_2...x_n$ is generated by (unknown) source ν. By definition, the test T is consistent (for ν), if for any significance level $\alpha \in (0,1)$ the probability of Type II error goes to 0, that is

$$\lim_{n\to\infty} \nu(\bar{C}_T(\alpha)) = 0\,. \tag{1}$$

Let us give a definition of a so-called p-value, which plays an important role in the randomness testing. Let there be a statistic τ (that is, a function on $\{0,1\}^n$) and x be a word from $\{0,1\}^n$. A p-value ($\pi_\tau(x)$) of τ and x is defined by the equation

$$\pi_\tau(x) = \mu_U\{y : \tau(y) \geq \tau(x)\} = |\{y : \tau(y) \geq \tau(x)\}|/2^n\,. \tag{2}$$

(Here and below $|X|$ is a number of elements X, if X is a set, and the length of X, if X is a word.)

Informally, $\pi_\tau(x)$ is the probability to meet a random point y which is "worse" than the observed when considering the null hypothesis.

2) The consistent tests for stationary ergodic sources and universal codes: First let us give a short informal description of the universal codes. For any integer m a lossless code ϕ is defined as such a map from the set of m-letter words to the set of all binary words that for any sequence of encoded m-letter words $\phi(v_1)\phi(v_2)...$ the initial sequence $v_1v_2...$ can be found without mistakes; the formall definition can be found, for example, in [15].

We will consider universal codes which have the following property: for any stationary ergodic ν defined on the set of all infinite binary words $x = x_1x_2...$, with probability one

$$\lim_{n\to\infty} \frac{1}{n}|\phi(x_1x_2...x_n)| = h(\nu)\,, \tag{3}$$

where $h(\nu)$ is the Shannon entropy of ν (see for definition [15]). Such codes exist, see, for example, [2], [3]. Note, that a goal of codes is to " compress " sequences, i.e. make a length of the codeword $\phi(x_1x_2...x_n)$ as small as possible. The property (3) shows that the universal codes are asymptotically optimal, because the Shannon entropy is a lower bound of the length of the compressed sequence per letter, see [15].

Let us back to considered problem of hypothesis testing. Suppose, it is known that a sample sequence $x = x_1x_2...$ was generated by stationary ergodic source and, as before, we consider the same H_0 against the same H_1. Let ϕ be a universal code. The following test is a particular case of a goodnes-of-fit test suggested in [2], [3]:

If $n - |\phi(x_1...x_n)| \geq -\log_2\alpha$ then H_0 is rejected, otherwise accepted. Here, as before, α is the significance level, $|\phi(x_1...x_n)|$ is the length of encoded ("compressed") sequence.

We denote this test by T_ϕ and its statistic by τ_ϕ, i.e.

$$\tau_\phi(x_1...x_n) = n - |\phi(x_1...x_n)|\,. \tag{4}$$

It turns out that this test is consistent for any stationary ergodic source. More precisely, the following theorem is proven in [2], [3]:

For each stationary ergodic ν, $\alpha \in (0,1)$ and a universal code ϕ, the Type I error of the described test is not larger than α and the Type II error goes to 0, when $n \to \infty$.

III. Asymptotic behaviour of a p-value of the Neyman-Pearson test.

Suppose, that H_1 is true and sequences $x \in \{0,1\}^n$ are obey a certain distribution ν. It is well-known in mathematical statistics that the optimal test (NP-test or likelihood-ratio test) is described by Neyman-Pearson lemma and the critical region of this test is defined as follows:

$$C_{NP}(\alpha) = \{x:\ \mu_U(x)/\nu(x) \leq \lambda_\alpha\}\,,$$

where $\alpha \in (0,1)$ is the significance level and the constant λ_α is chosen in such a way that $\mu_U(C_{NP}(\alpha)) = \alpha$, see [16]. (We did not take into account that the set $\{0,1\}^n$ is finite. Strictly speaking, in such a case a randomized test should be used, but in what follows we will consider asymptotic behaviour of the tests for large n and this effect will be negligible). The p-value for the NP-test can be derived from the definition (2), if we put $\tau(x) = \nu(x)$ and take into account that by definition, $\mu_U(x) = 2^{-n}$ for any $x \in \{0,1\}^n$. So,

$$\pi_{NP}(x) = \mu_U\{y : \nu(y) \geq \nu(x)\} = |\{y : \nu(y) \geq \nu(x)\}|/2^n\,. \tag{5}$$

The following theorem describes an asymptotic behaviour of p-values for stationary ergodic sources for NP test.

Theorem 1: If ν is a stationary ergodic measure, then, with probability 1,

$$\lim_{n\to\infty} -\frac{1}{n}\log\pi_{NP}(x) = 1 - h(\nu)\,, \tag{6}$$

where $h(\nu)$ is the Shannon entropy of ν, see for definition [15].

The NP-test is optimal in the sense that its probability of a Type II error is minimal, but when testing RNG the alternative distribution is unknown, and, hence, some other tests should be used. It turns out that the above described test T_ϕ has the same asymptotic behaviour as NP-test.

IV. Asymptotically optimal tests for randomness.

The following theorem describes an asymptotic behaviour of p-values for stationary ergodic sources for tests which are based on universal codes.

Theorem 2: Let ϕ be a universal code and the test T_ϕ with statistic τ_ϕ (4) is applied. Then for any stationary ergodic measure ν, with probability 1,

$$\lim_{n\to\infty} -\frac{1}{n}\log\pi_{\tau_\phi}(x) = 1 - h(\nu)\,, \tag{7}$$

where π_{τ_ϕ} is the p-value.

Note that this theorem gives some idea of the relation between the Shannon entropy of the (unknown) process ν and the required sample size. Indeed, suppose that a NP test is used and the desired significance level is α. Then, we can see that (asymptotically) α should be less than $\pi_{NP}(x)$ and from (6) we obtain $n > -\log\alpha/(1 - h(\nu))$ (for the most powerful test). It is known that the Shannon entropy is 1 if and only if ν is the uniform measure μ_u. Therefore, in a certain sense, the difference $1 - h(\nu)$ estimates the distance between the distributions, and the last inequality shows that the required sample size goes to infinity if ν approaches the uniform distribution.

The next simple example illustrates the theorems. Let there be a test κ and a generator (a measure ν) that generates sequences of independent binary digits with, say, $\nu(0) = 0.501, \nu(1) = 0.499$. Suppose that $\lim_{n\to\infty} -\frac{1}{n}\log\pi_\kappa(x) = c$, where c is a positive constant. Let us consider the following "decimation test" $\kappa^{1/2}$: an input sequence $x_1x_2....x_n$ is transformed into $x_1x_3x_5...x_{2\lfloor n/2\rfloor-1}$ and then κ is applied to this transformed sequence. Obviously, for this test $\lim_{n\to\infty} -\frac{1}{n/2}\log\pi_{\kappa^{1/2}}(x) = c$, and, hence, $\lim_{n\to\infty} -\frac{1}{n}\log\pi_{\kappa^{1/2}}(x) = c/2$. Thus, the value $-\frac{1}{n}\log\pi_\kappa(x_1...x_n)$ seems to be a reasonable estimate of the power of the test for large n.

V. Appendix

Proof of Theorem 1. The well-known Shannon-McMillan-Breiman (SMB) theorem states that for the stationary ergodic source ν and any $\epsilon > 0, \delta > 0$ there exists such $n'(\epsilon, \delta)$ that

$$\nu\{x \in \{0,1\}^n : \; h(\nu) - \epsilon < -\frac{1}{n}\log \nu(x) < h(\nu) + \epsilon \; \} > 1 - \delta \quad for \;\; n > n'(\epsilon, \delta)\,, \tag{8}$$

see [15]. From this we obtain

$$\nu\{x \in \{0,1\}^n : \; 2^{-n(h(\nu)-\epsilon)} > \nu(x) > 2^{-n(h(\nu)+\epsilon)}\} > 1 - \delta \tag{9}$$

for $n > n'(\epsilon, \delta)$. It will be convenient to define

$$\Phi_{\epsilon,n} = \{x \in \{0,1\}^n : \; h(\nu) - \epsilon < -\frac{1}{n}\log \nu(x) < h(\nu) + \epsilon \; \} \tag{10}$$

From this definition and (9) we obtain

$$(1-\delta)\, 2^{n(h(\nu)-\epsilon)} \le |\Phi_{\epsilon,n}| \le 2^{n(h(\nu)+\epsilon)}\,. \tag{11}$$

For any $x \in \Phi_{\epsilon,n}$ define

$$\Lambda_x = \{y : \nu(y) \ge \nu(x) \; \} \bigcap \Phi_{\epsilon,n}\,. \tag{12}$$

Note that, by definition, $|\Lambda_x| \le |\Phi_{\epsilon,n}|$ and from (11) we obtain

$$|\Lambda_x| \le 2^{n(h(\nu)+\epsilon)}\,. \tag{13}$$

For any $\rho \in (0,1)$ we define $\Psi_\rho \subset \Phi_{\epsilon,n}$ such that

$$\nu(\Psi_\rho) = \rho \;\&\; \forall u \in \Psi_\rho, \forall v \in (\Phi_{\epsilon,n} \setminus \Psi_\rho) \; : \nu(u) \ge \nu(v)\,. \tag{14}$$

(That is, Ψ_ρ contains the most probable words whose total probability equals ρ. If there are several such sets we can take any of them.) Let us consider any $x \in (\Phi_{\epsilon,n} \setminus \Psi_\rho)$. Taking into account the definition (14) and (11) we can see that for this x

$$|\Lambda_x| \ge \rho |\Phi_{\epsilon,n}| \ge \rho(1-\delta) 2^{n(h(\nu)-\epsilon)}\,. \tag{15}$$

So, from this inequality and (13) we obtain

$$\rho(1-\delta) 2^{n(h(\nu)-\epsilon)} \le |\Lambda_x| \le \; 2^{n(h(\nu)+\epsilon)}\,. \tag{16}$$

From equation (9), (10) and (14) we can see that $\nu(\Phi_{\epsilon,n} \setminus \Psi_\rho) \ge (1-\delta)(1-\rho)$. Taking into account (16) and this inequality, we can see that

$$\nu\{x : \; h(\nu) - \epsilon + \log(\rho(1-\delta))/n \le \log|\Lambda_x|/n \le h(\nu) + \epsilon\} \ge (1-\delta)(1-\rho). \tag{17}$$

From the definition (5) of $\pi_{NP}(x)$ and the definition (12) of Λ_x, we can see that $\pi_{NP}(x) = |\Lambda_x|/2^n$. Taking into account this equation and (17) we obtain the following:

$$\nu\{x : \; 1 - (h(\nu) - \epsilon + \log(\rho(1-\delta))/n) \le -\log \pi_{NP}(x)/n \ge 1 - (h(\nu) + \epsilon)\} \ge (1-\delta)(1-\rho). \tag{18}$$

Clearly, there exists such $n^*(\rho)$ that for $n > n^*(\rho)$ $-\log(\rho(1-\delta))/n < \epsilon$. Taking into account (8) we can see that

$$\nu\{x : \; 1 - (h(\nu) - 2\epsilon) \le -\log \pi_{NP}(x)/n \ge 1 - (h(\nu) + \epsilon)\} \ge (1-\delta)(1-\rho) \tag{19}$$

for $n > \max(n'(\epsilon, \delta), n^*(\rho))$. This inequality is valid for any $\rho \in (0,1)$ and, in particular, for $\rho = \delta$. So, from (19) we obtain

$$\nu\{x : \; 1 - (h(\nu) - 2\epsilon) \le -\log \pi_{NP}(x)/n \ge 1 - (h(\nu) + \epsilon)\} \ge (1-2\delta).$$

for $n > \max(n'(\epsilon, \delta), n^*(\delta))$.

Having taken into account that this inequality is valid for all positive ϵ and δ, we obtain the statement of the theorem.

Proof of Theorem 2 is similar to the previous one. First, for any $\epsilon > 0, \delta > 0$ we define

$$\hat{\Phi}_{\epsilon,n} = \{x : h(\nu) - \epsilon < |\phi(x_1 ... x_n)|/n < h(\nu) + \epsilon \; \}\,. \tag{20}$$

Note that from (3) we can see that there exists such $n''(\epsilon, \delta)$ that, for $n > n''(\epsilon, \delta)$,

$$\nu(\hat{\Phi}_{\epsilon,n}) > 1 - \delta\,. \tag{21}$$

We will use the set $\Phi_{\epsilon,n}$ (see (10)). Having taken into account the SMB theorem (8) and (21), we can see that

$$\nu(\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n}) > 1 - 2\delta\,, \tag{22}$$

if $n > \max(n'(\epsilon, \delta), n''(\epsilon, \delta))$.

From this moment, the proof begins to repeat the proof of the first theorem, if we use the set $(\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n})$ instead of $\Phi_{\epsilon,n}$. Namely, define

$$\hat{\Lambda}_x = \{y : |\phi(y)| \le |\phi(x)| \; \} \cap (\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n}) \tag{23}$$

and $\hat{\Psi}_\rho$ is such a subset of $(\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n})$ that

$$\nu(\hat{\Psi}_\rho) = \rho \;\&\; \forall u \in \Psi_\rho\,, \forall v \in ((\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n}) \setminus \Psi_\rho) : \; |\phi(u)| \le |\phi(v)|\,. \tag{24}$$

Let us consider any $x \in ((\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n}) \setminus \hat{\Psi}_\rho)$. Taking into account the definition (23) and (22), we obtain

$$\rho(1-2\delta) 2^{n(h(\nu)-\epsilon)} \le |\hat{\Lambda}_x| \le \; 2^{n(h(\nu)+\epsilon)}\,. \tag{25}$$

From equations (22) and (24) we can see that $\nu((\hat{\Phi}_{\epsilon,n} \cap \Phi_{\epsilon,n}) \setminus \hat{\Psi}_\rho) \ge (1-2\delta)(1-\rho)$. Taking into account (25) and this inequality, we can see that

$$\nu\{x : \; h(\nu) - \epsilon + \log(\rho(1-2\delta))/n \le \log|\hat{\Lambda}_x|/n \le h(\nu) + \epsilon\} \ge (1-2\delta)(1-\rho). \tag{26}$$

From the definition of p-value (2) and the definition (23), we can see that $\pi_{\tau_\phi}(x) = |\hat{\Lambda}_x|/2^n$. Taking into account this equation and (26) we obtain the following:

$$\nu\{x : \; 1 - (h(\nu) - \epsilon + \log(\rho(1-\delta))/n) \le -\log \pi_{\tau_\phi}(x)/n \ge 1 - (h(\nu) + \epsilon)\} \ge (1-2\delta)(1-\rho). \tag{27}$$

Clearly, there exists such $n^{**}(\rho)$ that for $n > n^{**}(\rho)$ $-\log(\rho(1-2\delta))/n < \epsilon$. Taking it account we can see from (27) that

$$\nu\{x:\ 1-(h(\nu)-2\epsilon) \geq$$

$$-\log \pi_{\tau_\phi}(x)/n \geq 1-(h(\nu)+\epsilon)\} \geq (1-2\delta)(1-\delta) \quad (28)$$

for $n > \max(n'(\epsilon,\delta), n''(\epsilon,\delta), n^{**}(\rho))$. So, from (28) we obtain

$$\nu\{x:\ 1-(h(\nu)-2\epsilon) \geq$$

$$-\log \pi_{\tau_\phi}(x)/n \ \geq\ 1-(h(\nu)+\epsilon)\} \geq (1-3\delta).$$

for $n > \max(n'(\epsilon,\delta), n''(\epsilon,\delta), n^{**}(\delta))$.

Having taken into account that this inequality is valid for all positive ϵ and δ, we obtain the statement of the theorem.

ACKNOWLEDGMENT

Research was supported by Russian Foundation for Basic Research (grant no. 18-29-03005).

REFERENCES

[1] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, and S. Vo, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. National Institute of Standards and Technology, 2010.

[2] B. Ryabko and J. Astola, " Universal Codes as a Basis for Time Series Testing," *Statistical Methodology*, vol.3, pp. 375-397, 2006.

[3] B. Ryabko, J. Astola, M. Malyutov, *Compression-Based Methods of Statistical Analysis and Prediction of Time Series,* Springer International Publishing Switzerland, 2016.

[4] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, no. 4, pp. 396–402, 1984.

[5] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM Journal of research and development*, vol. 23, no. 2, pp. 149–162, 1979.

[6] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on information theory*, vol. 23, no. 3, pp. 337–343, 1977.

[7] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," 1994.

[8] B. Y. Ryabko, "Data compression by means of a book stack," *Problems of Information Transmission*, vol. 16, no. 4, pp. 265–269, 1980.

[9] J. Bentley, D. Sleator, R. Tarjan, and V. Wei, " A locally adaptive data compression scheme," *Communications of the ACM*, vol. 29, no. 4, pp. 320–330, 1986.

[10] B. Ryabko, N. R. Horspool, G. V. Cormack, S. Sekar, and S. B. Ahuja, "Technical correspondence," *Communications of the ACM*, vol. 30, no. 9, pp. 792–797, 1987.

[11] J. C. Kieffer and E.-H. Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 737–754, 2000.

[12] E.-H. Yang and J. C. Kieffer, "Efficient universal lossless data compression algorithms based on a greedy sequential grammar transform. I. without context models," *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 755–777, 2000.

[13] M. Drmota, Yu. Reznik, and W. Szpankowski, " Tunstall code, Khodak variations, and random walks," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2928–2937, 2010.

[14] B. Ryabko, "Twice-universal coding," *Problems of Information Transmission*, vol. 3, pp. 173–177, 1984.

[15] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 2006.

[16] M. Kendall, A. Stuart, *The advanced theory of statistics; Vol.2: Inference and Relationship*; Hafner Publishing Company: New York, NY, USA, 1961.