

# Large Deviations Performance of Interval Algorithm for Random Number Generation

Akisato KIMURA\*

Tomohiko UYEMATSU\*

**Abstract**— We investigate large deviations performance of the interval algorithm for random number generation, especially for intrinsic randomness. First, we show that the length of output fair random bits per the length of input sequence approaches to the entropy of the source almost surely. Next, we consider to obtain the fixed number of fair random bits from the input sequence with fixed length. We show that the approximation error measured by the variational distance and divergence vanishes exponentially as the length of input sequence tends to infinity, if the number of fair bits per input sample is below the entropy of the source. Contrarily, the approximation error measured by the variational distance approaches to two exponentially, if the number of fair bits per input sample is above the entropy.

**Keywords**—random number generation, interval algorithm, intrinsic randomness, variational distance, error exponent, divergence.

## I. Introduction

Random number generation is a problem of simulating some prescribed target distribution by using a given source. This problem has been investigated in computer science, and has a close relation to information theory [1, 2, 3]. Some practical algorithms for random number generation have been proposed so far, i.e. [1, 3, 4, 5]. In this paper, we consider the interval algorithm proposed by Han and Hoshi [3].

Performance of the interval algorithm has already been investigated in [3, 6, 7]. Han and Hoshi [3] have showed that the expected length of input sequence per the length of output sequence can be characterized by the ratio of entropy of the input and output distributions. Uyematsu and Kanaya [6] have investigated large deviations performance of the interval algorithm where the distribution of input source is uniform. Further, Uchida and Han [7] have extended the result of Uyematsu and Kanaya to stationary ergodic Markov process. We investigate large deviations performance, where the distribution of target random number is uniform.

We first show that the length of output sequence per input sample approaches to the entropy of the source almost surely. Next, we consider to obtain the fixed number of fair random bits from the input sequence with fixed length. We show that the approximation

error measured by the variational distance and divergence vanishes exponentially as the length of input sequence tends to infinity, if the number of fair bits per input sample is below the entropy of the source. Contrarily, the approximation error measured by the variational distance approaches to two exponentially, if the number of fair bits per input sample is above the entropy of the source.

## II. Basic Definitions

### (a) Discrete Memoryless Source

Let  $\mathcal{X}$  be a finite set. We denote by  $\mathcal{M}(\mathcal{X})$  the set of all probability distributions on  $\mathcal{X}$ . Throughout this paper, by a source  $X$  with alphabet  $\mathcal{X}$ , we mean a discrete memoryless source (DMS) of distribution  $P_X \in \mathcal{M}(\mathcal{X})$ . To denote a source we will use both notations  $X$  and  $P_X$  interchangeably.

For random variable  $X$  which has a distribution  $P_X$ , we shall denote this entropy as  $H(P_X)$  and  $H(X)$ , interchangeably. Further, for arbitrary distributions  $P, Q \in \mathcal{M}(\mathcal{X})$ , we denote by  $D(P \parallel Q)$  the information divergence

$$D(P \parallel Q) \triangleq \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Lastly, we denote by  $d(P, Q)$  the variational distance or  $l_1$  distance between two distributions  $P$  and  $Q$  on  $A$

$$d(P, Q) \triangleq \sum_{a \in A} |P(a) - Q(a)|.$$

From now on, all logarithms and exponentials are to the base two.

### (b) Intrinsic Randomness

In this paper, we especially investigate the problem to generate a uniform random number with as large size as possible from a source  $X$ . This problem is called *intrinsic randomness problem* [8]. Here, we shall introduce basic definitions and a result for intrinsic randomness problem.

*Definition 1:* For arbitrary source  $X$ ,  $R$  is *achievable Intrinsic Randomness (IR) rate* if and only if there exists a map  $\varphi_n : \mathcal{X}^n \rightarrow \mathcal{U}_{M_n}$  such that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log M_n \geq R$$
$$\lim_{n \rightarrow \infty} d(\mathcal{U}_{M_n}, \varphi_n(\mathcal{X}^n)) = 0,$$

\* Dept. of Electrical and Electronic Eng., Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan. E-mail: akisato@ss.titech.ac.jp, uyematsu@ss.titech.ac.jp

where  $\mathcal{U}_{M_n} \triangleq \{1, 2, \dots, M_n\}$  and  $U_{M_n}$  is a uniform distribution on  $\mathcal{U}_{M_n}$ .

*Definition 2* (IR rate):

$$S(X) = \sup\{R \mid R \text{ is achievable}\}$$

As for the characterization of IR rate, Vembu and Verdú [2] proved the following fundamental theorem.

*Theorem 1:* For any stationary source  $X$ ,

$$S(X) = H(X), \quad (1)$$

where  $H(X)$  is the entropy rate of  $X$ .

### III. Interval Algorithm

In this chapter, we introduce the interval algorithm for random number generation proposed by Han and Hoshi [3]. For our purpose, we rephrase their algorithm for intrinsic randomness.

Let us consider to produce a sequence of fair bits of length  $n$  by using an i.i.d. sequence  $X_1, X_2, \dots$  with a generic distribution  $\mathbf{p} = (p_1, p_2, \dots, p_M)$ .

*Interval Algorithm for Generating Fair Bits*

- 1a) Partition an unit interval  $[0, 1)$  into disjoint subinterval  $J(0), J(1)$  such that

$$J(i) = [i/2, (i+1)/2) \quad i = 0, 1.$$

- 1b) Set

$$P_j = \sum_{k=1}^j p_k \quad j = 1, 2, \dots, M; \quad P_0 = 0.$$

- 2) Set  $s = t = \lambda$  (null string),  $\alpha_s = \gamma_t = 0$ ,  $\beta_s = \delta_t = 1$ ,  $I(s) = [\alpha_s, \beta_s)$ ,  $J(t) = [\gamma_t, \delta_t)$ ,  $l = 0$  and  $m = 1$ .

- 3) Obtain an output symbol from the source to have a value  $a \in \{1, 2, \dots, M\}$ , and generate the subinterval of  $I(s)$

$$I(sa) = [\alpha_{sa}, \beta_{sa}),$$

where

$$\begin{aligned} \alpha_{sa} &= \alpha_s + (\beta_s - \alpha_s)P_{a-1} \\ \beta_{sa} &= \alpha_s + (\beta_s - \alpha_s)P_a, \end{aligned}$$

and set  $l = l + 1$ .

- 4a) If  $I(sa)$  is entirely contained in some  $J(ti)$  ( $i = 0, 1$ ), then set  $t = ti$ . Otherwise, go to 5).
- 4b) If  $m = n$  then output  $t$  as the output sequence  $Y^n$ , and stop the algorithm. Otherwise, partition the interval  $J(t) \equiv [\gamma_t, \delta_t)$  into disjoint subinterval  $J(t0), J(t1)$  such that

$$J(tj) = [\gamma_{tj}, \delta_{tj}) \quad j = 0, 1$$

where

$$\begin{aligned} \gamma_{tj} &= \gamma_t + j(\delta_t - \gamma_t)/2 \\ \delta_{tj} &= \gamma_t + (j+1)(\delta_t - \gamma_t)/2, \end{aligned}$$

and set  $m = m + 1$  and go to 4a).

- 5) Set  $s = sa$  and go to 3).

As for the property of the above algorithm, Han and Hoshi have shown that

$$\lim_{n \rightarrow \infty} \frac{E(L)}{n} = \frac{1}{H(\mathbf{p})}, \quad (2)$$

where  $E(L)$  is the average length of input sequences to obtain fair bits of length  $n$ .

### IV. Almost Sure Convergence of Number of Fair Bits per Input Sample

We shall investigate large deviations performance of the interval algorithm for random number generation.

Let us consider to produce a sequence of fair bits by using a sequence  $X^n = (X_1, X_2, \dots, X_n)$  of length  $n$ . Each random variable  $X_i$  ( $i = 1, 2, \dots, n$ ) is subject to a generic distribution  $P_X$  on  $\mathcal{X} = \{1, 2, \dots, M\}$ . We denote by  $L_n(\mathbf{x})$  the number of generated fair bits from the input sequence  $\mathbf{x} \in \mathcal{X}^n$ . Here, we define the following functions:

$$E_r(R, P_X) = \min_{\substack{Q \in \mathcal{M}(\mathcal{X}): \\ H(Q) \leq R}} D(Q \parallel P_X), \quad (3)$$

$$E_{sp}(R, P_X) = \min_{\substack{Q \in \mathcal{M}(\mathcal{X}): \\ D(Q \parallel P_X) + H(Q) \leq R}} D(Q \parallel P_X), \quad (4)$$

$$F(R, P_X) = \min_{\substack{Q \in \mathcal{M}(\mathcal{X}): \\ D(Q \parallel P_X) + H(Q) \geq R}} D(Q \parallel P_X), \quad (5)$$

$$G(R, P_X) = \min_{\substack{Q \in \mathcal{M}(\mathcal{X}): \\ H(Q) \geq R}} D(Q \parallel P_X). \quad (6)$$

Then, we obtain the following large deviations performances of the interval algorithm:

*Theorem 2:* For  $R > 0$ ,

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} L_n(X) \leq R \right\} \right] \geq E_r(R, P_X). \quad (7)$$

For  $R > R_{min} = \min_{x \in \mathcal{X}} \log \frac{1}{P_X(x)}$ ,

$$\limsup_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} L_n(X) \leq R \right\} \right] \leq E_{sp}(R, P_X). \quad (8)$$

Further,  $E_r(R, P_X) > 0$  if and only if  $R < H(X)$ ,  $E_{sp}(R, P_X) > 0$  if and only if  $R_{min} < R < H(X)$ , and  $E_r(R, P_X) < E_{sp}(R, P_X)$  for  $R < H(X)$ .

*Theorem 3:* For  $R < R_{max} = \max_{x \in \mathcal{X}} \log \frac{1}{P_X(x)}$ ,

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} L_n(X) \geq R \right\} \right] \geq F(R, P_X). \quad (9)$$

For  $R < \log |\mathcal{X}|$ ,

$$\limsup_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} L_n(X) \geq R \right\} \right] \leq G(R, P_X). \quad (10)$$

Further,  $F(R, P_X) > 0$  if and only if  $H(X) < R < R_{max}$ ,  $G(R, P_X) > 0$  if and only if  $H(X) < R < \log |\mathcal{X}|$ , and  $F(R, P_X) < G(R, P_X)$  for  $R > H(X)$ .

Combining these theorems and Borel-Cantelli's lemma (e.g. [9]) we immediately obtain the following corollary.

*Corollary 1:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} L_n(X) = H(X) \quad \text{a.s.} \quad (11)$$

*Remark 1:* Let us consider to produce a specified number of fair bits by using a sequence from the source  $X$ . We denote by  $T_n(X)$  the length of sequences to obtain fair bits of length  $n$ . Then, we can obtain similar relations as (7)-(10). For example, corresponds to (7), we have

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \Pr \left\{ \frac{1}{n} T_n(X) \geq R \right\} \right] \geq \widetilde{E}_r(R, P_X), \quad (12)$$

where

$$\widetilde{E}_r(R, P_X) = \min_{\substack{Q \in \mathcal{M}(\mathcal{X}): \\ H(Q) \leq 1/R}} R D(Q \| P_X). \quad (13)$$

Similarly, corresponding to Corollary 1, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} T_n(X) = \frac{1}{H(X)} \quad \text{a.s.} \quad (14)$$

## V. Error Exponent for random number generation

In this chapter, let us consider to produce fixed number of random bits with an input sequence of length  $n$ . In this case, we cannot generate fair bits exactly but approximately.

First, we modify the interval algorithm for generating fair bits so that the algorithm outputs a specified sequence  $11 \cdots 1 \in \{0, 1\}^{nR}$  whenever the algorithm does not stop with a input sequence of length  $n$ . The modified algorithm can be described below. Since steps 1), 2), 4) and 5) are the same as the original algorithm, we omit them in the modified algorithm.

*Modified Interval Algorithm for Generating Fair Bits*

- 3) If  $l = n$  then output  $11 \cdots 1$  as the output sequence  $Y^{nR}$ , and stop the algorithm. Otherwise obtain an output symbol from the source  $X$  to have a value  $a \in \mathcal{X} = \{1, 2, \dots, M\}$ , and generate the subinterval of  $I(s)$

$$I(sa) = [\alpha_{sa}, \beta_{sa})$$

where

$$\begin{aligned} \alpha_{sa} &= \alpha_s + (\beta_s - \alpha_s) P_{a-1} \\ \beta_{sa} &= \alpha_s + (\beta_s - \alpha_s) P_a, \end{aligned}$$

and set  $l = l + 1$ .

We first measure the approximation error by the variational distance between the desired and approximated output distributions. Then, we obtain the following theorems.

*Theorem 4:* If the modified interval algorithm is used for random number generation, then we have

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log d(U_{\exp(nR)}, \widetilde{P}_Y^{nR}) \right] \geq E_r(R, P_X), \quad (15)$$

where  $\widetilde{P}_Y^{nR}$  denotes the output distribution of the modified interval algorithm, and  $E_r(R, P_X)$  is given by (3). Further, for  $R > R_{min}$

$$\limsup_{n \rightarrow \infty} \left[ -\frac{1}{n} \log d(U_{\exp(nR)}, \widetilde{P}_Y^{nR}) \right] \leq E_{sp}(R, P_X), \quad (16)$$

where  $E_{sp}(R, P_X)$  is given by (4).

This theorem implies that if the length of output sequence per input sample is below the entropy of the source, the approximation error measured by the variational distance vanishes exponentially as the length of input sequence tends to infinity.

Next theorem shows the upper bound of the error exponent.

*Theorem 5:* Let  $\widetilde{P}_Y^{nR}$  denote a distribution over  $U_{\exp(nR)}$  using any algorithm for random number generation with fixed input length  $n$ . Then, for  $R > R_{min}$

$$\limsup_{n \rightarrow \infty} \left[ -\frac{1}{n} \log d(U_{\exp(nR)}, \widetilde{P}_Y^{nR}) \right] \leq E_{sp}(R, P_X), \quad (17)$$

where  $E_{sp}(R, P_X)$  is given by (4).

Note that  $E_r(R, P_X) < E_{sp}(R, P_X)$ . Hence, it is still an open problem to obtain the exact error exponent of the proposed algorithm.

Next theorem shows the converse result.

*Theorem 6:* If the modified interval algorithm is used for random number generation, then for  $R < R_{max}$ ,

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \{2 - d(U_{\exp(nR)}, \widetilde{P}_Y^{nR})\} \right] \geq F(R, P_X), \quad (18)$$

where  $F(R, P_X)$  is given by (5). Further, for  $R < \log |\mathcal{X}|$

$$\limsup_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \{2 - d(U_{\exp(nR)}, \tilde{P}_Y^{nR})\} \right] \leq G(R, P_X), \quad (19)$$

where  $G(R, P_X)$  is given by (6).

This theorem implies that if the length of output sequence per input sample is above the entropy of the source, the approximation error measured by the variational distance approaches to two exponentially as the length of input sequence tends to infinity.

Next theorem was due to Ohama [5].

*Theorem 7:* Consider the optimum algorithm for random number generation with fixed input length  $n$ , let  $\tilde{P}_Y^{nR}$  denote the distribution over  $\mathbf{y} \in \{0, 1\}^{nR}$  which minimizes the variational distance. Then, we have

$$\lim_{n \rightarrow \infty} \left[ -\frac{1}{n} \log \{2 - d(P_Y^{nR}, \tilde{P}_Y^{nR})\} \right] = F'(R, P_X), \quad (20)$$

where

$$F'(R, P_X) = \min_{Q \in \mathcal{M}(\mathcal{X})} \{D(Q \| P_X) + |R - H(Q) - D(Q \| P_X)|^+\}. \quad (21)$$

Further,  $F'(R, P_X) \geq F(R, P_X)$  and equality holds for  $R \geq R_0$ , where

$$R_0 \triangleq D(Q_0 \| P_X) + \log |\mathcal{X}| \quad (22)$$

and  $Q_0$  is a uniform distribution on  $\mathcal{X}$ .

Theorem 6 and 7 imply that the interval algorithm with fixed input length is not optimum if  $R \geq R_0$ .

Next, consider to measure the approximation error by the divergence between the desired and approximated output distributions. First, we show the following lemma.

*Lemma 1:* Let  $P^n, Q^n$  be arbitrary distributions on  $\mathcal{X}^n$ . If  $d(P^n, Q^n) \leq \epsilon$ , then  $D(P^n \| Q^n) \leq -\epsilon \log P_{\min}^n Q_{\min}^n$ , where  $P_{\min}^n$  (resp.  $Q_{\min}^n$ ) is the minimum of  $P^n$  (resp.  $Q^n$ ) over  $\mathcal{X}^n$ .

From Theorem 4 and Lemma 1, we immediately obtain the following corollary.

*Corollary 2:* If the modified interval algorithm is used for random number generation, then

$$\liminf_{n \rightarrow \infty} \left[ -\frac{1}{n} \log D(U_{\exp(nR)} \| \tilde{P}_Y^{nR}) \right] \geq E_r(R, P_X), \quad (23)$$

where  $E_r(R, P_X)$  is given by (3).

This theorem implies that if the length of output sequence per input sample is below the entropy of the

source, the approximation error measured by the divergence also vanishes exponentially as the length of input sequence tends to infinity.

*Remark 2:* Han has showed that there exists an algorithm for random number generation of which normalized divergence vanishes [8]. However, as shown in Corollary, for DMS (more generally unifilar sources), even divergence can vanishes as the input length tends to infinity.

## VI. Conclusion

We have investigated large deviations performance of interval algorithm for random number generation. We have clarified some asymptotic properties, when target random number have been subject to uniform distribution. As future researches, we are going to generalize our results to more complex sources.

## References

- [1] D. Knuth and A. Yao, "The complexity of nonuniform random number generation," *Algorithm and Complexity, New Directions and Results*, pp.357-428, ed. by J. F. Traub, Academic Press, New York, 1976.
- [2] S. Vembu and S. Verdú, "Generating random bits from an arbitrary source: Fundamental limits," *IEEE Trans. on Inform. Theory*, vol.IT-41, pp.1322-1332, 1995.
- [3] T. S. Han and M. Hoshi, "Interval algorithm for random number generation," *IEEE Trans. on Inform. Theory*, vol.43, pp.599-611, 1997.
- [4] F. Kanaya, "An asymptotically optimal algorithm for generating Markov random sequences," *Proc. of SITA '97*, pp.77-80, Matsuyama, Japan, Dec., 1997 (in Japanese).
- [5] Y. Ohama, "Fixed to fixed length random number generation using one dimensional piecewise linear maps," *Proc. of SITA '98*, pp.57-60, Gifu, Japan, Dec., 1998 (in Japanese).
- [6] T. Uyematsu and F. Kanaya, "Channel simulation by interval algorithm: A performance analysis for interval algorithm," *submitted to IEEE Trans. on Inform. Theory*.
- [7] O. Uchida and T. S. Han, "Performance analysis of interval algorithm for generating Markov processes," *Proc. of SITA '98*, pp.65-68, Gifu, Japan, Dec., 1998.
- [8] T. S. Han: *Information-Spectrum Methods in Information Theory*, Baifukan, Tokyo, 1998 (in Japanese).
- [9] P. C. Shields: *The ergodic theory of discrete sample paths*, Graduate Studies in Math. vol.13, American Math. Soc. (1996).