ASYMPTOTIC ANALYSIS OF PERES' ALGORITHM FOR RANDOM NUMBER GENERATION

ZHAO GING LIM[1], CHEN-TUO LIAO[2], AND YI-CHING YAO[3]

Abstract. von Neumann (1951) introduced a simple algorithm for generating independent unbiased random bits by tossing a (possibly) biased coin with unknown bias. While his algorithm fails to attain the entropy bound, Peres (1992) showed that the entropy bound can be attained asymptotically by iterating von Neumann's algorithm. Let $b(n,p)$ denote the expected number of unbiased bits generated when Peres' algorithm is applied to an input sequence consisting of the outcomes of n tosses of the coin with bias p. With $p=1/2$, the coin is unbiased and the input sequence consists of n unbiased bits, so that $n-b(n,1/2)$ may be referred to as the cost incurred by Peres' algorithm when not knowing $p=1/2$. We show that $\lim_{n\to\infty}\log[n-b(n,1/2)]/\log n=\theta=\log[(1+\sqrt{5})/2]$ (where log is the logarithm to base 2), which together with limited numerical results suggests that $n-b(n,1/2)$ may be a regularly varying sequence of index θ. Some open problems on the asymptotic behavior of $nh(p)-b(n,p)$ are briefly discussed where $h(p)=-p\log p-(1-p)\log(1-p)$ denotes the Shannon entropy of a random binary bit with bias p.

1. Introduction and Results

In his seminal work [11], von Neumann introduced a simple algorithm $\mathscr{A}_{\mathsf{vN}}$ (also known as an extractor) for generating independent unbiased random bits by tossing a (possibly) biased coin with unknown bias. Specifically, for $i=1,2,\ldots$, let $X_i\in\{\mathtt{0},\mathtt{1}\}$ denote the outcome of the ith toss of the coin, where $\mathtt{1}$ and $\mathtt{0}$ stand for heads and tails, respectively.

Date: June 1, 2020.

Key words and phrases. Entropy, Elias' extractor, Peres' extractor, von Neumann's extractor, analysis of algorithms, superadditivity, regularly varying sequence.

[1] Department of Agronomy, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan, ROC; Email address: zhaoging@gmail.com.

[2] Department of Agronomy, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan, ROC; Email address: ctliao@ntu.edu.tw.

[3] Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, ROC; Email address: yao@stat.sinica.edu.tw.

Assume that the input sequence $(X_1, X_2, \dots)$ is independent and identically distributed (iid) with $\mathbb{P}(X_i = \mathtt{1}) = p = 1 - \mathbb{P}(X_i = \mathtt{0})$ where the bias $p \in (0,1)$ is unknown. The algorithm $\mathscr{A}_{\mathsf{vN}}$ divides the X_i's into pairs $(X_1, X_2), (X_3, X_4), \dots$, and discards those pairs of equal values. For each pair of unequal values, $\mathscr{A}_{\mathsf{vN}}$ generates a bit equal to the first value of the pair, which is unbiased in the sense that its value is $\mathtt{1}$ or $\mathtt{0}$ with equal probability.

Let $\mathscr{A}$ denote a generic algorithm that generates independent unbiased bits from the sequence $(X_1, X_2, \dots)$. Let $\mathscr{A}(n)$ denote the set of unbiased bits generated by $\mathscr{A}$ applied to $(X_1, \dots, X_n)$, the outcomes of the first n tosses. Denote by $|\mathscr{A}(n)|$ the cardinality of $\mathscr{A}(n)$, which is an integer-valued random variable whose distribution depends on n and p. We say that $\mathscr{A}$ is *nested* if $\mathscr{A}(n_1) \subset \mathscr{A}(n_2)$ whenever $n_1 < n_2$, i.e. the set of unbiased bits generated from $(X_1, \dots, X_{n_1})$ is contained in the set of unbiased bits generated from $(X_1, \dots, X_{n_1}, X_{n_1+1}, \dots, X_{n_2})$. We write $X \sim \text{binomial}(n, p)$ if a random variable X has the binomial distribution with parameters n and p. Then given $|\mathscr{A}_{\mathsf{vN}}(n)| = k$, the k bits generated by $\mathscr{A}_{\mathsf{vN}}$ applied to $(X_1, \dots, X_n)$ are (conditionally) independent unbiased. Moreover, $|\mathscr{A}_{\mathsf{vN}(n)}| \sim \text{binomial}(\lfloor \frac{n}{2} \rfloor, 2pq)$, where $q = 1 - p$ and $\lfloor x \rfloor$ denotes the largest integer not exceeding x. When $\mathscr{A}_{\mathsf{vN}}$ is applied to $(X_1, \dots, X_n)$, the expected number of unbiased bits generated per toss equals $\mathbb{E}_p\, |\mathscr{A}_{\mathsf{vN}}(n)|/n = 2pq\lfloor \frac{n}{2} \rfloor / n$, which converges to pq as $n \to \infty$, where the subscript p in $\mathbb{E}_p$ refers to the bias of each X_i. Note that pq is less than the entropy bound $h(p) \coloneqq -p \log p - q \log q$ (the Shannon entropy of each X_i), where $\log = \log_2$ (the logarithm to base 2). This indicates that $\mathscr{A}_{\mathsf{vN}}$ does not make efficient use of information contained in $X_1, X_2, \dots$. It is also worth noting that $\mathscr{A}_{\mathsf{vN}}$ is nested.

To improve the efficiency, Elias [3] presented a more sophisticated algorithm $\mathscr{A}_{\mathsf{E}}$ which generates unbiased bits from $(X_1, \dots, X_n)$ by partitioning $S_n = \{(x_1, \dots, x_n) : x_i \in \{\mathtt{0}, \mathtt{1}\}\}$ (the set of all possible realizations of $(X_1, \dots, X_n)$) into disjoint subsets $S_{n,k} = \{(x_1, \dots, x_n) \in S_n : \sum_{i=1}^n x_i = k\}$, $k = 0, 1, \dots, n$. Write $|S_{n,k}| = \binom{n}{k} = \sum_{\ell=0}^{\lfloor \log \binom{n}{k} \rfloor} c_\ell 2^\ell$ with $c_\ell \in \{0, 1\}$ (binary representation of $\binom{n}{k}$). Then each $S_{n,k}$ is further partitioned as $S_{n,k} = \bigcup_{\{\ell : c_\ell = 1\}} S_{n,k,\ell}$, where $|S_{n,k,\ell}| = 2^\ell$ for each ℓ with $c_\ell = 1$. Specify an assignment of 2^ℓ distinct (output) sequences of $\{\mathtt{0}, \mathtt{1}\}^\ell$ to the 2^ℓ distinct sequences of $S_{n,k,\ell}$, so that if $(X_1, \dots, X_n) \in S_{n,k,\ell}$, then an output sequence of ℓ bits is generated according to the assignment. While a naive implementation of Elias' algorithm requires an exponential memory size to make a table of assignment of output sequences, Ryabko and Matchikina [10] made use of the enumerative encoding technique (*cf.* Cover [2]) to construct an assignment with

much reduced memory size and running time. Note that $\mathscr{A}_{\mathsf{E}}$ is not nested. In fact, when $\mathscr{A}_{\mathsf{E}}$ is applied to $(X_1, \dots, X_n)$, all of $X_1, \dots, X_n$ need to be observed before unbiased bits are generated. To show that $\mathscr{A}_{\mathsf{E}}$ attains the entropy bound asymptotically, Elias [3, equation (15)] proved that

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 3 \le \mathbb{E}_p\, |\mathscr{A}_{\mathsf{E}}(n)| \le \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k}. \tag{1.1}$$

Letting $H(Z)$ denote the Shannon entropy of a random variable Z and noting that $\sum_{i=1}^{n} X_i \sim$ binomial(n, p), we have

$$\begin{aligned} nh(p) - H\biggl(\sum_{i=1}^{n} X_i\biggr) &= -np \log p - nq \log q + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \biggl[\binom{n}{k} p^k q^{n-k}\biggr] \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k}, \end{aligned}$$

from which it follows that (1.1) is equivalent to

$$H\biggl(\sum_{i=1}^{n} X_i\biggr) \le nh(p) - \mathbb{E}_p\, |\mathscr{A}_{\mathsf{E}}(n)| \le H\biggl(\sum_{i=1}^{n} X_i\biggr) + 3. \tag{1.2}$$

Since $H(\sum_{i=1}^{n} X_i) = \frac{1}{2} \log n + \frac{1}{2} \log e + \log \sqrt{2\pi pq} + O(\frac{1}{n})$ (*cf.* [4]), we have

$$nh(p) - \mathbb{E}_p\, |\mathscr{A}_{\mathsf{E}}(n)| = \frac{1}{2} \log n + O(1). \tag{1.3}$$

Consequently, $\lim_{n\to\infty} \mathbb{E}_p\, |\mathscr{A}_{\mathsf{E}}(n)|/n = h(p)$. Later Pae and Loui [7] established the exact optimality of $\mathscr{A}_{\mathsf{E}}$ that for any algorithm $\mathscr{A}$, $\mathbb{E}_p\, |\mathscr{A}_{\mathsf{E}}(n)| \ge \mathbb{E}_p\, |\mathscr{A}(n)|$ for all $p \in (0, 1)$ and $n \ge 1$.

While $\mathscr{A}_{\mathsf{vN}}$ fails to attain the entropy bound asymptotically, Peres [8] pointed out that the entropy bound can be attained asymptotically by iterating $\mathscr{A}_{\mathsf{vN}}$. To describe Peres' ingenious idea, we consider an infinite iid Bernoulli sequence $\boldsymbol{X} = (X_1, X_2, \dots)$ with bias p, which is decomposed into three infinite Bernoulli sequences $\psi_i(\boldsymbol{X})$, $i = 1, 2, 3$, as follows. First divide $X_1, X_2, \dots$ into pairs $(X_1, X_2), (X_3, X_4), \dots$. The ith bit of $\psi_1(\boldsymbol{X})$ is $\mathtt{1}$ or $\mathtt{0}$ according as the ith pair (X_{2i-1}, X_{2i}) is of equal values or of unequal values. Then separate those pairs of equal values from the other pairs of unequal values. The ith bit of $\psi_2(\boldsymbol{X})$ is the common value of the ith pair of equal values. The ith bit of $\psi_3(\boldsymbol{X})$ is the first value of the ith pair of unequal values. As an example, let $\boldsymbol{X} = (\mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{0}, \dots)$. Then $\psi_1(\boldsymbol{X}) = (\mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \dots)$, $\psi_2(\boldsymbol{X}) = (\mathtt{1}, \mathtt{0}, \mathtt{0}, \dots)$ and $\psi_3(\boldsymbol{X}) = (\mathtt{0}, \mathtt{1}, \mathtt{1}, \dots)$. It

is readily seen that (i) $\psi_1(\boldsymbol{X})$, $\psi_2(\boldsymbol{X})$ and $\psi_3(\boldsymbol{X})$ are mutually independent, (ii) $\psi_1(\boldsymbol{X})$, $\psi_2(\boldsymbol{X})$ and $\psi_3(\boldsymbol{X})$ are each an iid Bernoulli sequence with respective biases $f_1(p) \coloneqq p^2+q^2$, $f_2(p) \coloneqq p^2/(p^2+q^2)$ and $1/2$, (iii) $\boldsymbol{X}$ can be recovered from $\psi_1(\boldsymbol{X})$, $\psi_2(\boldsymbol{X})$ and $\psi_3(\boldsymbol{X})$, implying that they together contain all information in $\boldsymbol{X}$. The first iteration of Peres' algorithm $\mathscr{A}_{\mathsf{P}}$ yields $\psi_i(\boldsymbol{X})$, $i = 1, 2, 3$, where $\psi_3(\boldsymbol{X})$ is precisely the output sequence generated by $\mathscr{A}_{\mathsf{vN}}$ applied to $\boldsymbol{X}$. On the second iteration of $\mathscr{A}_{\mathsf{P}}$, $\psi_i(\boldsymbol{X})$, $i = 1, 2$ are each decomposed into three iid Bernoulli sequences $\psi_1(\psi_i(\boldsymbol{X}))$, $\psi_2(\psi_i(\boldsymbol{X}))$ and $\psi_3(\psi_i(\boldsymbol{X}))$ with respective biases $f_1(f_i(p))$, $f_2(f_i(p))$ and $1/2$. Thus, after 2 iterations, there are $7(= 2^3-1)$ Bernoulli sequences, $3(= 2^2-1)$ of which have bias $1/2$. More generally, after ν iterations ($\nu = 1, 2, \dots$), there are $2^{\nu+1}-1$ Bernoulli sequences, $2^\nu - 1$ of which have bias $1/2$. We refer to the $2^\nu - 1$ Bernoulli sequences having bias $1/2$ as *unbiased* Bernoulli sequences, and refer to the other 2^ν Bernoulli sequences as *biased* Bernoulli sequences. Note that the $2^{\nu+1}-1$ Bernoulli sequences are all mutually independent, from which $\boldsymbol{X}$ can be recovered.

We now consider the finite setting where only the first n terms of the infinite input sequence $\boldsymbol{X}$ are available. Let $(\boldsymbol{X})_n = (X_1, \dots, X_n)$, the subsequence of $\boldsymbol{X}$ consisting of the first n terms. Then $(\boldsymbol{X})_n$ induces the first n_i terms of $\psi_i(\boldsymbol{X})$, $i = 1, 2, 3$, where $n_1 = n_2 + n_3 = \lfloor \frac{n}{2} \rfloor$, $n_2 \sim \text{binomial}(\lfloor \frac{n}{2} \rfloor, p^2+q^2)$ and $n_3 \sim \text{binomial}(\lfloor \frac{n}{2} \rfloor, 2pq)$. In shorthand notation, $(\boldsymbol{X})_n$ induces $(\psi_i(\boldsymbol{X}))_{n_i}$, $i = 1, 2, 3$. While the infinite sequences $\psi_i(\boldsymbol{X})$, $i = 1, 2, 3$ are mutually independent, the subsequences $(\psi_i(\boldsymbol{X}))_{n_i}$, $i = 1, 2, 3$ are no longer independent. Indeed, the numbers of `1`'s and `0`'s in $(\psi_1(\boldsymbol{X}))_{n_1}$ are equal to n_2 and n_3, respectively. It is readily seen that, given the value of n_3, the bits in $(\psi_3(\boldsymbol{X}))_{n_3}$ are (conditionally) independent unbiased. In fact, given the values of the bits in $(\psi_1(\boldsymbol{X}))_{n_1}$ and $(\psi_2(\boldsymbol{X}))_{n_2}$, the bits in $(\psi_3(\boldsymbol{X}))_{n_3}$ remain (conditionally) independent unbiased. Furthermore, for even n, $(\boldsymbol{X})_n$ can be recovered from $(\psi_i(\boldsymbol{X}))_{n_i}$, $i = 1, 2, 3$, but for odd n, the last term of $(\boldsymbol{X})_n$ cannot be recovered, resulting in a loss of information. After ν iterations ($\nu = 1, 2, \dots$), $(\boldsymbol{X})_n$ induces a (possibly empty) subsequence of each of the $2^{\nu+1}-1$ infinite Bernoulli sequences as decomposed from $\boldsymbol{X}$. Let $\mathscr{A}_{\mathsf{P},\nu}(n)$ denote the set of all unbiased bits contained in the subsequences of those $2^\nu - 1$ infinite unbiased Bernoulli sequences. Since after $\lfloor \log n \rfloor$ iterations, the longest biased subsequence has length 1, no more unbiased bits can be produced by further iteration. We have $\mathscr{A}_{\mathsf{P},\nu}(n) = \mathscr{A}_{\mathsf{P},\lfloor \log n \rfloor}(n)$ for $\nu \geq \lfloor \log n \rfloor$. Let $\mathscr{A}_{\mathsf{P}}(n) = \mathscr{A}_{\mathsf{P},\lfloor \log n \rfloor}(n)$, the set of all unbiased bits generated by $\mathscr{A}_{\mathsf{P}}$ applied to $(\boldsymbol{X})_n$. Consider again the example where $\boldsymbol{X} = (\mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{0}, \dots)$. For $n = 12$, we have

$\mathscr{A}_{\mathsf{P},1}(12) = \{0,1,1\}$, $\mathscr{A}_{\mathsf{P},2}(12) = \{0,1,1,0,1\}$, and $\mathscr{A}_{\mathsf{P}}(12) = \mathscr{A}_{\mathsf{P},3}(12) = \{0,1,1,0,1,0\}$. It is shown in Peres [8] that (i) for each ν, given $|\mathscr{A}_{\mathsf{P},\nu}(n)| = k$, the k bits in $\mathscr{A}_{\mathsf{P},\nu}(n)$ are independent unbiased, (ii) the rates $r_\nu(p) \coloneqq \lim_{n\to\infty} \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)|/n$ satisfy $r_1(p) = pq$ and the recursion

$$r_\nu(p) = pq + \frac{1}{2} r_{\nu-1}(p^2+q^2) + \frac{1}{2}(p^2+q^2) r_{\nu-1}\Big(\frac{p^2}{p^2+q^2}\Big) \quad \text{for } \nu \ge 2, \tag{1.4}$$

and (iii) $r_\nu(p)$ increases as $\nu \to \infty$ to $h(p)$ uniformly in $p \in (0,1)$. As a consequence, $\mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|/n \to h(p)$ as $n \to \infty$, showing that $\mathscr{A}_{\mathsf{P}}(n)$ attains the entropy bound asymptotically. Moreover, $\mathscr{A}_{\mathsf{P}}(n)$ is nested.

While $\mathscr{A}_{\mathsf{E}}$ and $\mathscr{A}_{\mathsf{P}}$ both attain the entropy bound asymptotically, (1.2) and (1.3) provide a precise (second-order) behavior of $nh(p) - \mathbb{E}_p\,|\mathscr{A}_{\mathsf{E}}(n)|$. In contrast, there is not much known about the behavior of $nh(p) - \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|$ for large n. In this regard, Pae [6] gave a formula to compute $\mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|$, which is not convenient for deriving the asymptotic behavior of $nh(p) - \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|$. Recently, Prasitsupparote *et al.* [9] showed, based on some heuristics, that for $p = 1/2$,

$$nh(p) - \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)| = n - \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)| \ge n^{\log 3 - 1}. \tag{1.5}$$

To derive (1.5), they assumed, without rigorous justification, that

$$\frac{1}{n}\mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)| \le r_\nu(p) \quad \text{for } p \in (0,1),\ n \ge 1,\ \nu \ge 1. \tag{1.6}$$

In this paper, we establish the following results.

Proposition 1. *Let $a(n,p,\nu) = \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)|$. Then for $p \in (0,1)$ and $\nu = 1,2,\dots$, the sequence $(a(1,p,\nu), a(2,p,\nu), \dots)$ is superadditive, i.e. $a(n,p,\nu)+a(m,p,\nu) \le a(n+m,p,\nu)$ for $n,m \ge 1$. Consequently, $\lim_{n\to\infty} a(n,p,\nu)/n$ exists and is equal to $\sup_{n\ge1} a(n,p,\nu)/n$. That is,*

$$r_\nu(p) \coloneqq \lim_{n\to\infty} \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)|/n = \sup_{n\ge1} \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)|/n,$$

which implies (1.6).

Proposition 2. *For $p = 1/2$, let $b(n) = \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|$. Then*

(i) *the $b(n)$ satisfy $b(0) = b(1) = 0$ and the recursion*

$$b(n) = \Big\lfloor \frac{n}{2} \Big\rfloor /2 + b\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, b(B_{\lfloor n/2 \rfloor, 1/2}) \quad \textit{for } n = 2,3,\dots, \tag{1.7}$$

where $B_{n,p}$ denotes a binomial(n,p) *random variable;*

(ii)

$$\lim_{n\to\infty} \frac{\log(n - b(n))}{\log n} = \log\Big(\frac{1+\sqrt{5}}{2}\Big).$$

The next section contains the proofs of Propositions 1 and 2. In addition, for completeness, a rigorous proof of (1.5) is also given, which is needed for the proof of Proposition 2(ii). Section 3 presents numerical results, and Section 4 concludes the paper with some open problems.

We close this section by remarking that while $\mathscr{A}_{\mathsf{E}}$ generates more unbiased bits than $\mathscr{A}_{\mathsf{P}}$, $\mathscr{A}_{\mathsf{P}}$ is much simpler to implement. Prasitsupparote *et al.* [9] made an extensive numerical study of $\mathscr{A}_{\mathsf{E}}$ and $\mathscr{A}_{\mathsf{P}}$ regarding their memory and running time requirements and concluded that $\mathscr{A}_{\mathsf{P}}$ is superior to $\mathscr{A}_{\mathsf{E}}$ in practical applications.

2. Proofs of Propositions 1 and 2 and (1.5)

Proof of Proposition 1. Recall that when $\mathscr{A}_{\mathsf{P}}$ is applied to $(X_1, \ldots, X_n)$, 3 subsequences $(\psi_i(\boldsymbol{X}))_{n_i}$, $i = 1, 2, 3$, are induced where $n_1 = n_2 + n_3 = \lfloor n/2 \rfloor$ and $n_3 \sim \text{binomial}(\lfloor n/2 \rfloor, 2pq)$. It follows that $\mathscr{A}_{\mathsf{P},\nu}(n)$, the set of unbiased bits generated after ν iterations, is the union of 3 disjoint subsets $\mathcal{S}_i$, $i = 1, 2, 3$, where $\mathcal{S}_3 = (\psi_3(\boldsymbol{X}))_{n_3}$, and $\mathcal{S}_i$, $i = 1, 2$, is the set of unbiased bits generated when $\mathscr{A}_{\mathsf{P}}$ is applied to $(\psi_i(\boldsymbol{X}))_{n_i}$ with $\nu - 1$ iterations. We have

$$a(n, p, \nu) = \mathbb{E}_p\, |\mathscr{A}_{\mathsf{P},\nu}(n)| = \mathbb{E}_p\, |\mathcal{S}_1| + \mathbb{E}_p\, |\mathcal{S}_2| + 2pq\lfloor n/2 \rfloor. \tag{2.1}$$

Noting that $(\psi_1(\boldsymbol{X}))_{n_1}$ is a sequence of $n_1 = \lfloor n/2 \rfloor$ iid Bernoulli random variables with bias $f_1(p)$, we have

$$\mathbb{E}_p\, |\mathcal{S}_1| = \mathbb{E}_{f_1(p)}\, |\mathscr{A}_{\mathsf{P},\nu-1}(\lfloor n/2 \rfloor)| = a(\lfloor n/2 \rfloor, f_1(p), \nu - 1). \tag{2.2}$$

Similarly, conditioning on n_2, $(\psi_2(\boldsymbol{X}))_{n_2}$ is a sequence of n_2 iid Bernoulli random variables with bias $f_2(p)$, so that the conditional expectation of $|\mathcal{S}_2|$ given n_2 equals $\mathbb{E}_{f_2(p)}\, |\mathscr{A}_{\mathsf{P},\nu-1}(n_2)| = a(n_2, f_2(p), \nu - 1)$. Since $n_2 \sim \text{binomial}(\lfloor n/2 \rfloor, 1 - 2pq)$, we have

$$\mathbb{E}_p\, |\mathcal{S}_2| = \mathbb{E}\, a(B_{\lfloor n/2 \rfloor, 1-2pq}, f_2(p), \nu - 1), \tag{2.3}$$

where the expectation operator $\mathbb{E}$ on the right-hand side is on $B_{\lfloor n/2 \rfloor, 1-2pq}$ (a binomial$(\lfloor n/2 \rfloor, 1 - 2pq)$ random variable). By (2.1), (2.2) and (2.3),

$$a(n, p, \nu) = a(\lfloor n/2 \rfloor, f_1(p), \nu - 1) + \mathbb{E}\, a(B_{\lfloor n/2 \rfloor, 1-2pq}, f_2(p), \nu - 1) + 2pq\lfloor n/2 \rfloor. \tag{2.4}$$

We now prove by induction on ν that

$$a(n,p,\nu) + a(m,p,\nu) \le a(n+m,p,\nu). \tag{2.5}$$

For $\nu = 1$, $a(n,p,1) = 2pq\lfloor n/2 \rfloor$. Since $\lfloor n/2 \rfloor + \lfloor m/2 \rfloor \le \lfloor (n+m)/2 \rfloor$ for $n, m \ge 1$, we have $a(n,p,1) + a(m,p,1) \le a(n+m,p,1)$, implying that (2.5) holds for $\nu = 1$. Suppose that for an integer $k > 0$, (2.5) holds for all $n, m \ge 1$, all $p \in (0,1)$, and $\nu = k$. We need to show that (2.5) holds for $n, m \ge 1$, $p \in (0,1)$ and $\nu = k+1$. By the induction hypothesis,

$$\begin{aligned} a(\lfloor n/2 \rfloor, f_1(p), k) + a(\lfloor m/2 \rfloor, f_1(p), k) &\le a(\lfloor n/2 \rfloor + \lfloor m/2 \rfloor, f_1(p), k) \\ &\le a(\lfloor (n+m)/2 \rfloor, f_1(p), k), \end{aligned} \tag{2.6}$$

where the second inequality follows from the fact that $a(n,p,\nu)$ is non-decreasing in n. Let U and V be independent random variables with $U \sim \text{binomial}(\lfloor n/2 \rfloor, 1-2pq)$ and $V \sim \text{binomial}(\lfloor m/2 \rfloor, 1-2pq)$. Then $U+V \sim \text{binomial}(\lfloor n/2 \rfloor + \lfloor m/2 \rfloor, 1-2pq)$. If n and m are both odd, let W be independent of U and V with $W \sim \text{binomial}(1, 1-2pq)$. If at least one of n and m is even, let W be identically 0. Then $U+V+W \sim \text{binomial}(\lfloor (n+m)/2 \rfloor, 1-2pq)$. We have by the induction hypothesis that

$$\begin{aligned} &\mathbb{E}\, a(B_{\lfloor n/2 \rfloor, 1-2pq}, f_2(p), k) + \mathbb{E}\, a(B_{\lfloor m/2 \rfloor, 1-2pq}, f_2(p), k) \\ &\quad = \mathbb{E}\left\{ a(U, f_2(p), k) + a(V, f_2(p), k) \right\} \\ &\quad \le \mathbb{E}\, a(U+V, f_2(p), k) \\ &\quad \le \mathbb{E}\, a(U+V+W, f_2(p), k) \\ &\quad = \mathbb{E}\, a(B_{\lfloor (n+m)/2 \rfloor}, f_2(p), k). \end{aligned} \tag{2.7}$$

Moreover,

$$2pq\lfloor n/2 \rfloor + 2pq\lfloor m/2 \rfloor \le 2pq\lfloor (n+m)/2 \rfloor. \tag{2.8}$$

By (2.4) and (2.6)–(2.8),

$$a(n,p,k+1) + a(m,p,k+1) \le a(n+m,p,k+1),$$

showing that (2.5) holds for $n, m \ge 1$, $p \in (0,1)$ and $\nu = k+1$. The proof is complete. □

Proof of (1.5). The following argument is taken from the proof of Theorem 1 in Prasitsupparote [9]. With $p = 1/2$, we have $r_1(1/2) = pq = \frac{1}{4}$ and, by (1.4)

$$r_\nu(1/2) = \frac{1}{4} + \frac{3}{4} r_{\nu-1}(1/2) \quad \text{for } \nu \ge 2,$$

from which it follows that $r_\nu(1/2) = 1 - (\frac{3}{4})^\nu$, $\nu \geq 1$. By Proposition 1,

$$1 - \Big(\frac{3}{4}\Big)^\nu = r_\nu(1/2) \geq \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P},\nu}(n)|/n,$$

so that with $\nu = \lfloor \log n \rfloor$ and $b(n) = \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|$, we have

$$1 - \Big(\frac{3}{4}\Big)^{\lfloor \log n \rfloor} \geq \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P},\lfloor \log n \rfloor}(n)|/n = \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|/n = b(n)/n,$$

implying that

$$n - b(n) \geq n\Big(\frac{3}{4}\Big)^{\lfloor \log n \rfloor} \geq n\Big(\frac{3}{4}\Big)^{\log n} = n^{\log 3 - 1},$$

proving (1.5). □

Proof of Proposition 2(i). For $p = 1/2$, $f_1(1/2) = f_2(1/2) = 1/2$, and $B_{\lfloor n/2 \rfloor, 1-2pq} = B_{\lfloor n/2 \rfloor, 1/2}$. Letting $a(n, p, \nu) = \mathbb{E}_p\,|\mathscr{A}_{\mathsf{P},\nu}(n)|$, we have by (2.4) that

$$a(n, 1/2, \nu) = a(\lfloor n/2 \rfloor, 1/2, \nu - 1) + \mathbb{E}\,a(B_{\lfloor n/2 \rfloor, 1/2}, 1/2, \nu - 1) + \lfloor n/2 \rfloor/2. \tag{2.9}$$

Recall that $\mathscr{A}_{\mathsf{P}}(n) = \mathscr{A}_{\mathsf{P},\nu}(n)$ for $\nu \geq \lfloor \log n \rfloor$. By (2.9),

$$\begin{aligned} b(n) &= \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)| = \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P},\lfloor \log n \rfloor}(n)| \\ &= a(n, 1/2, \lfloor \log n \rfloor) \\ &= a(\lfloor n/2 \rfloor, 1/2, \lfloor \log n \rfloor - 1) + \mathbb{E}\,a(B_{\lfloor n/2 \rfloor, 1/2}, 1/2, \lfloor \log n \rfloor - 1) + \lfloor n/2 \rfloor/2 \\ &= b(\lfloor n/2 \rfloor) + \mathbb{E}\,b(B_{\lfloor n/2 \rfloor, 1/2}) + \lfloor n/2 \rfloor/2, \end{aligned}$$

proving (1.7). □

To prove Proposition 2(ii), we need the following lemmas. Proposition 2(ii) follows immediately from Lemmas 4 and 5 below. For the rest of this section, to simplify notation, we write $B_n = B_{n,1/2}$ for a binomial$(n, 1/2)$ random variable. Let $g(n) = n - b(n) \geq 0$, for $n = 0, 1, \ldots$. We have $g(0) = 0$, $g(1) = 1$, and by Proposition 2(i), for even $n \geq 0$,

$$\begin{aligned} g(n) = n - b(n) &= n - \Big[\frac{n}{4} + b\Big(\frac{n}{2}\Big) + \mathbb{E}\,b(B_{\frac{n}{2}})\Big] \\ &= \Big[\frac{n}{2} - b\Big(\frac{n}{2}\Big)\Big] + \mathbb{E}\,\big[B_{\frac{n}{2}} - b(B_{\frac{n}{2}})\big] \\ &= g\Big(\frac{n}{2}\Big) + \mathbb{E}\,g(B_{\frac{n}{2}}), \end{aligned}$$

and for odd $n \geq 1$,

$$\begin{aligned} g(n) = n - b(n) &= n - \left[\frac{n-1}{4} + b\Big(\frac{n-1}{2}\Big) + \mathbb{E}\, b(B_{(n-1)/2})\right] \\ &= 1 + \left[\frac{n-1}{2} - b\Big(\frac{n-1}{2}\Big)\right] + \mathbb{E}\left[B_{(n-1)/2} - b(B_{(n-1)/2})\right] \\ &= 1 + g\Big(\frac{n-1}{2}\Big) + \mathbb{E}\, g(B_{(n-1)/2}). \end{aligned}$$

So, for $n \geq 0$,

$$g(n) = g\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, g(B_{\lfloor \frac{n}{2} \rfloor}) + \mathbf{1}_{\{n \text{ is odd}\}}, \tag{2.10}$$

where $\mathbf{1}$ denotes the indicator function.

Lemma 1. *For $\delta \in (0,1)$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(B_n > \frac{n}{2}(1+\delta)\Big) = -\frac{1}{2}\Big[(1-\delta)\log(1-\delta) + (1+\delta)\log(1+\delta)\Big] < 0.$$

Lemma 2. *If $f(0) \leq f(1) \leq \cdots \leq f(n+1)$, then $\mathbb{E}\, f(B_{n+1}) \geq \mathbb{E}\, f(B_n)$.*

Lemma 1 is a standard result in large deviation theory; see e.g. [5, pages 539–540]. Lemma 2 follows from the fact that B_n is stochastically smaller than B_{n+1}. By (2.10),

$$g(0) = 0, \qquad g(1) = 1, \qquad g(2) = \frac{3}{2}, \qquad g(3) = \frac{5}{2}, \qquad g(4) = \frac{19}{8} < g(3), \tag{2.11}$$

so that $g(n)$ is not non-decreasing. Lemma 3 below constructs two non-decreasing sequences G and H that are closely related to g and satisfy $0 \leq H(n) \leq g(n) \leq G(n) \leq n$.

Lemma 3. *Let $G(n)$ and $H(n)$, $n = 0, 1, \ldots$ be defined by*

$$\begin{aligned} G(n) &= g(n) \quad \text{for } n = 0, 1, 2, 3, \\ H(n) &= g(n) \quad \text{for } n = 0, 1, \end{aligned}$$

and recursively

$$G(n) = G\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, G(B_{\lfloor \frac{n}{2} \rfloor}) + 1 \quad \text{for } n \geq 4, \tag{2.12}$$

$$H(n) = H\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, H(B_{\lfloor \frac{n}{2} \rfloor}) \qquad \text{for } n \geq 2. \tag{2.13}$$

Then (i) *G is non-decreasing and $g(n) \leq G(n) \leq n$ for all n, and* (ii) *H is non-decreasing and $g(n) \geq H(n) \geq 0$ for all n.*

Proof. In view of (2.10) and (2.12), it is easily shown by induction that $g(n) \le G(n)$ for all n. By (2.11), $G(n) = g(n) \le n$ and $G(n)$ is non-decreasing for $n \le 3$. For $n \ge 4$, if $G(\ell) \le \ell$ for all $\ell < n$, then

$$\begin{aligned} G(n) &= G\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, G(B_{\lfloor \frac{n}{2} \rfloor}) + 1 \\ &\le \Big\lfloor \frac{n}{2} \Big\rfloor + \mathbb{E}\, B_{\lfloor \frac{n}{2} \rfloor} + 1 \\ &\le \frac{3}{4} n + 1 \le n. \end{aligned}$$

It follows by induction that $G(n) \le n$ for all n. That $G(n)$ is non-decreasing in n also follows by induction and Lemma 2. This proves part (i). In view of (2.10) and (2.13), part (ii) can be proved similarly. □

Lemma 4. *For each $\delta \in (0,1)$, there exists an $N \ge 4$ and a non-decreasing sequence $(G'(0), G'(1), \dots)$ such that $G'(n) \ge g(n)$ for all n and*

$$G'(n) = G'\Big(\Big\lfloor \frac{n}{c} \Big\rfloor\Big) + \frac{4}{c^2} G'\Big(\Big\lfloor \frac{n}{c^2} \Big\rfloor\Big), \quad \textit{for all } n \ge N,$$

where $c = c(\delta) = 2/\sqrt{1+\delta}$. Moreover,

$$\limsup_{n\to\infty} \frac{\log g(n)}{\log n} \le \limsup_{n\to\infty} \frac{\log G'(n)}{\log n} \le \frac{1}{\log c} \log\left(\frac{1}{2} + \sqrt{\frac{4}{c^2} + \frac{1}{4}} \right).$$

Consequently, letting $\delta \to 0$ so that $c = c(\delta) \to 2$, we have

$$\limsup_{n\to\infty} \frac{\log g(n)}{\log n} \le \log\left(\frac{1+\sqrt{5}}{2} \right).$$

Proof. Let $\delta \in (0,1)$ be fixed. Let G be defined as in Lemma 3, so that G is non-decreasing and $0 \le g(n) \le G(n) \le n$ for all n. We have

$$\begin{aligned} \mathbb{E}\, G(B_{\lfloor \frac{n}{2} \rfloor}) + 1 &\le 1 + G\Big(\Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} \le \Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) \\ &\quad + G\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} > \Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) \\ &\le 1 + G\Big(\Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) + \Big\lfloor \frac{n}{2} \Big\rfloor \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} > \Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big). \end{aligned} \tag{2.14}$$

By (1.5), $G(\lfloor \frac{\lfloor n/2 \rfloor}{2}(1+\delta) \rfloor) \ge g(\lfloor \frac{\lfloor n/2 \rfloor}{2}(1+\delta) \rfloor) \to \infty$ as $n \to \infty$, and by Lemma 1,

$$\lim_{n\to\infty} \Big\lfloor \frac{n}{2} \Big\rfloor \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} > \Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) = 0,$$

so that by (2.14), there is a (large) $N \geq 4$ such that

$$\mathbb{E}\, G(B_{\lfloor \frac{n}{2} \rfloor}) + 1 \leq (1+\delta)\, G\Big(\Big\lfloor \Big\lfloor \frac{n}{2} \Big\rfloor \Big(\frac{1+\delta}{2}\Big) \Big\rfloor\Big) \quad \text{for all } n \geq N. \tag{2.15}$$

Letting $c = 2/\sqrt{1+\delta}$, we have by (2.12) and (2.15) that for all $n \geq N (\geq 4)$,

$$\begin{aligned} G(n) &= G\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, G(B_{\lfloor \frac{n}{2} \rfloor}) + 1 \\ &\leq G\Big(\Big\lfloor \frac{n}{c} \Big\rfloor\Big) + \frac{4}{c^2} G\Big(\Big\lfloor \frac{n}{c^2} \Big\rfloor\Big). \end{aligned} \tag{2.16}$$

Define $G'(n)$, $n = 0, 1, \ldots$ by $G'(n) = G(n)$ for $n < N$ and recursively

$$G'(n) = G'\Big(\Big\lfloor \frac{n}{c} \Big\rfloor\Big) + \frac{4}{c^2} G'\Big(\Big\lfloor \frac{n}{c^2} \Big\rfloor\Big) \quad \text{for } n \geq N. \tag{2.17}$$

(Note that for $c > \sqrt{2}$ and $n \geq N \geq 4$, $\lfloor n/c^2 \rfloor \leq \lfloor n/c \rfloor \leq n-1$, so G' is well defined.) Since $G(n)$ is non-decreasing and $G(n) = G'(n)$ for all $n < N$, we have by (2.16), (2.17) and induction that $G'(n) \geq G(n) (\geq g(n))$ for all n. To show that $G'(n)$ is non-decreasing, note that $G'(N) \geq G(N) \geq G(N-1) = G'(N-1)$. Since $G'(0) \leq G'(1) \leq \cdots \leq G'(N)$, it follows by (2.17) and induction that $G'(n) \leq G'(n+1)$ for all $n \geq N$.

It remains to prove that

$$\limsup_{n \to \infty} \frac{\log G'(n)}{\log n} \leq \frac{1}{\log c} \log \left(\frac{1}{2} + \sqrt{\frac{4}{c^2} + \frac{1}{4}} \right). \tag{2.18}$$

Let $\ell_k = \lfloor c^k N \rfloor$, $k = 0, 1, \ldots$. Let $x_0 = G'(\ell_0)$, $x_1 = G'(\ell_1)$, and

$$x_k = x_{k-1} + \frac{4}{c^2} x_{k-2}, \quad k = 2, 3, \ldots. \tag{2.19}$$

By (2.17) and monotonicity of G', we have for $k \geq 2$

$$\begin{aligned} G'(\ell_k) = G'(\lfloor c^k N \rfloor) &= G'\left(\Big\lfloor \frac{\lfloor c^k N \rfloor}{c} \Big\rfloor \right) + \frac{4}{c^2} G'\left(\Big\lfloor \frac{\lfloor c^k N \rfloor}{c^2} \Big\rfloor \right) \\ &\leq G'\big(\lfloor c^{k-1} N \rfloor\big) + \frac{4}{c^2} G'\big(\lfloor c^{k-2} N \rfloor\big) \\ &= G'(\ell_{k-1}) + \frac{4}{c^2} G'(\ell_{k-2}). \end{aligned} \tag{2.20}$$

Since $x_k = G'(\ell_k)$ for $k = 0, 1$, it follows by (2.19), (2.20) and induction that

$$G'(\ell_k) \leq x_k \quad \text{for all } k \geq 0. \tag{2.21}$$

Since x_k satisfies the difference equation (2.19), we have

$$x_k = \alpha_1 \lambda_1^k + \alpha_2 \lambda_2^k, \quad k = 0, 1, \ldots$$

where

$$\lambda_1 = \frac{1}{2}(1+\gamma), \qquad \lambda_2 = \frac{1}{2}(1-\gamma)$$
$$\alpha_1 = \frac{1}{\gamma}\Big(\frac{1}{2}(\gamma-1)x_0 + x_1\Big), \qquad \alpha_2 = \frac{1}{\gamma}\Big(\frac{1}{2}(\gamma+1)x_0 - x_1\Big)$$

and $\gamma = \sqrt{1+\frac{16}{c^2}}$. Noting that $-1 < \lambda_2 < 0 < 1 < \lambda_1$ (since $\sqrt{2} < c < 2$) and $\alpha_1 > 0$, it follows that

$$\lim_{k\to\infty} \frac{\log x_k}{k} = \log \lambda_1 = \log\left(\frac{1}{2} + \sqrt{\frac{4}{c^2}+\frac{1}{4}}\right). \tag{2.22}$$

By (2.21) and (2.22),

$$\limsup_{k\to\infty} \frac{\log G'(\ell_k)}{\log \ell_k} \le \limsup_{k\to\infty} \frac{\log x_k}{\log\lfloor c^k N\rfloor} = \frac{\log \lambda_1}{\log c}.$$

Since G' is non-decreasing, for each $n \ge 1$, let $k = k(n)$ be such that $\ell_k \le n < \ell_{k+1}$, so that

$$\frac{\log G'(n)}{\log n} \le \frac{\log G'(\ell_{k+1})}{\log \ell_k},$$

implying that

$$\begin{aligned}\limsup_{n\to\infty} \frac{\log G'(n)}{\log n} &\le \limsup_{k\to\infty} \frac{\log G'(\ell_{k+1})}{\log \ell_k} \\ &= \limsup_{k\to\infty} \frac{\log G'(\ell_{k+1})}{\log \ell_{k+1}} \frac{\log \ell_{k+1}}{\log \ell_k} \\ &\le \frac{\log \lambda_1}{\log c} = \frac{1}{\log c}\log\left(\frac{1}{2} + \sqrt{\frac{4}{c^2}+\frac{1}{4}}\right),\end{aligned}$$

proving (2.18). The proof is complete. □

Lemma 5. *For each $\delta \in (0,1)$, there exists an $N \ge 4$ and a non-decreasing sequence $(H'(0), H'(1), \dots)$ such that $0 \le H'(n) \le g(n)$ for all n and*

$$H'(n) = H'\Big(\Big\lceil \frac{n}{d}\Big\rceil\Big) + \frac{4}{d^2} H'\Big(\Big\lceil \frac{n}{d^2}\Big\rceil\Big), \quad \text{for all } n \ge N,$$

where $d = d(\delta) = 2+\delta$ and $\lceil x\rceil$ denotes the smallest integer not less than x. Moreover,

$$\liminf_{n\to\infty} \frac{\log g(n)}{\log n} \ge \liminf_{n\to\infty} \frac{\log H'(n)}{\log n} \ge \frac{1}{\log d}\log\left(\frac{1}{2} + \sqrt{\frac{4}{d^2}+\frac{1}{4}}\right).$$

Consequently, letting $\delta \to 0$ so that $d = d(\delta) \to 2$, we have

$$\liminf_{n\to\infty} \frac{\log g(n)}{\log n} \ge \log\left(\frac{1+\sqrt{5}}{2}\right).$$

Proof. The following proof is similar to that of Lemma 4. Let $\delta \in (0,1)$ be fixed. Let H be defined as in Lemma 3, so that H is non-decreasing and $0 \le H(n) \le g(n)$ for all n. Also $H(0) = g(0) = 0$, $H(1) = g(1) = 1$.

For $d = 2+\delta > 2$, by the law of large numbers, $\mathbb{P}(B_{\lfloor \frac{n}{2} \rfloor} < \lceil \frac{n}{d^2} \rceil) \to 0$ as $n \to \infty$. So there exists an $N \ge 4$ such that for all $n \ge N$,

$$H\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) \ge H\Big(\Big\lceil \frac{n}{d} \Big\rceil\Big) \qquad \text{and} \qquad \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} \ge \Big\lceil \frac{n}{d^2} \Big\rceil\Big) \ge \frac{4}{d^2}.$$

By (2.13), for $n \ge N \ge 4$,

$$\begin{aligned} H(n) &= H\Big(\Big\lfloor \frac{n}{2} \Big\rfloor\Big) + \mathbb{E}\, H(B_{\lfloor \frac{n}{2} \rfloor}) \\ &\ge H\Big(\Big\lceil \frac{n}{d} \Big\rceil\Big) + \mathbb{P}\Big(B_{\lfloor \frac{n}{2} \rfloor} \ge \Big\lceil \frac{n}{d^2} \Big\rceil\Big) H\Big(\Big\lceil \frac{n}{d^2} \Big\rceil\Big) \\ &\ge H\Big(\Big\lceil \frac{n}{d} \Big\rceil\Big) + \frac{4}{d^2} H\Big(\Big\lceil \frac{n}{d^2} \Big\rceil\Big). \end{aligned} \tag{2.23}$$

Define $H'(n)$, $n = 0, 1, \dots$ by $H'(0) = 0$, $H'(1) = \cdots = H'(N-1) = 1$ and recursively

$$H'(n) = H'\Big(\Big\lceil \frac{n}{d} \Big\rceil\Big) + \frac{4}{d^2} H'\Big(\Big\lceil \frac{n}{d^2} \Big\rceil\Big) \quad \text{for } n \ge N. \tag{2.24}$$

(Note that $\lceil \frac{n}{d^2} \rceil \le \lceil \frac{n}{d} \rceil \le n-1$ for $n \ge N \ge 4$, so that the recursion is well defined.) Since by (2.24), $H'(N) = H'(\lceil \frac{N}{d} \rceil) + \frac{4}{d^2} H'(\lceil \frac{N}{d^2} \rceil) = 1 + \frac{4}{d^2} > 1$, we have $H'(0) < H'(1) = \cdots = H'(N-1) < H'(N)$. It follows by (2.24) and induction that H' is a non-decreasing sequence. Since $H(n) \ge H'(n)$ for all $n < N$, we have by (2.23), (2.24) and induction that $H'(n) \le H(n)$ for all n.

It remains to prove that

$$\liminf_{n\to\infty} \frac{\log H'(n)}{\log n} \ge \frac{1}{\log d} \log\left(\frac{1}{2} + \sqrt{\frac{4}{d^2} + \frac{1}{4}}\right) \tag{2.25}$$

Let $\ell_k = \lceil d^k N \rceil$, $k = 0, 1, \dots$. Let $x_0 = H'(\ell_0)$, $x_1 = H'(\ell_1)$, and

$$x_k = x_{k-1} + \frac{4}{d^2} x_{k-2}, \quad k = 2, 3, \dots. \tag{2.26}$$

By (2.24) and monotonicity of H', we have for $k \ge 2$

$$\begin{aligned} H'(\ell_k) = H'(\lceil d^k N \rceil) &= H'\left(\left\lceil \frac{\lceil d^k N \rceil}{d} \right\rceil\right) + \frac{4}{d^2} H'\left(\left\lceil \frac{\lceil d^k N \rceil}{d^2} \right\rceil\right) \\ &\ge H'\big(\lceil d^{k-1} N \rceil\big) + \frac{4}{d^2} H'\big(\lceil d^{k-2} N \rceil\big) \\ &= H'(\ell_{k-1}) + \frac{4}{d^2} H'(\ell_{k-2}). \end{aligned} \tag{2.27}$$

Since $x_k = H'(\ell_k)$ for $k = 0, 1$, it follows by (2.26), (2.27) and induction that

$$H'(\ell_k) \geq x_k \quad \text{for all } k \geq 0. \tag{2.28}$$

Note that the difference equation (2.26) is the same as (2.19) with c replaced by d. Solving (2.26) yields (*cf.* (2.22))

$$\lim_{k\to\infty} \frac{\log x_k}{k} = \log\left(\frac{1}{2} + \sqrt{\frac{4}{d^2} + \frac{1}{4}}\right).$$

By (2.28),

$$\liminf_{k\to\infty} \frac{\log H'(\ell_k)}{\log \ell_k} \geq \liminf_{k\to\infty} \frac{\log x_k}{\log\lceil d^k N\rceil} = \frac{1}{\log d} \log\left(\frac{1}{2} + \sqrt{\frac{4}{d^2} + \frac{1}{4}}\right).$$

Since H' is non-decreasing, for each $n \geq 1$, let $k = k(n)$ be such that $\ell_k \leq n < \ell_{k+1}$, so that

$$\frac{\log H'(n)}{\log n} \geq \frac{\log H'(\ell_k)}{\log \ell_{k+1}},$$

implying that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{\log H'(n)}{\log n} &\geq \liminf_{k\to\infty} \frac{\log H'(\ell_k)}{\log \ell_{k+1}} \\
&= \liminf_{k\to\infty} \frac{\log H'(\ell_k)}{\log \ell_k} \frac{\log \ell_k}{\log \ell_{k+1}} \\
&\geq \frac{1}{\log d} \log\left(\frac{1}{2} + \sqrt{\frac{4}{d^2} + \frac{1}{4}}\right),
\end{aligned}$$

proving (2.25). The proof is complete. □

3. Numerical results

Recall that $g(n) = n - b(n) = n - \mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|$. By (2.10), we computed $g(n)$ for all $n \leq 16384$. Figure 1 plots $\log g(n)/\log n$ versus n for $n \leq 16384$ where $\theta = \log[(1+\sqrt{5})/2] \approx 0.694$. It shows that $\log g(n)/\log n$ is slightly greater than θ and appears to converge to θ slowly. Figure 2 plots $g(n)/n^\theta$ versus n for $n \leq 16384$. By Proposition 2(ii),

$$\lim_{n\to\infty} \log\left[g(n)/n^\theta\right]/\log n = 0.$$

While it is unclear whether $g(n)/n^\theta$ converges to some constant eventually, it appears that $g(n)/n^\theta$ fluctuates less when n becomes larger. Figure 3 plots $g(2n)/g(n)$ versus n for $n \leq 8192$. It appears that $g(2n)/g(n)$ is close to 2^θ for large n. Figure 4 plots $g(3n)/g(n)$ versus n for $n \leq 5461$, where $g(3n)/g(n)$ oscillates around 3^θ. Our limited numerical results provide weak evidence that $g(3n)/g(n)$ converges to 3^θ eventually.

4. Concluding Remarks

Recall that $\mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|$ denotes the expected number of unbiased bits generated when Peres' algorithm is applied to the input sequence $(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are iid with $\mathbb{P}(X_i=\mathtt{1})=p=1-\mathbb{P}(X_i=\mathtt{0})$. When $p=1/2$, $X_1,\ldots,X_n$ are unbiased, so that $g(n)=n-\mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|$ may be referred to as the cost incurred by Peres' algorithm when not knowing $p=\frac{1}{2}$. We derived

$$\lim_{n\to\infty}\log\left[n-\mathbb{E}_{1/2}\,|\mathscr{A}_{\mathsf{P}}(n)|\right]/\log n=\theta=\log\left(\frac{1+\sqrt{5}}{2}\right) \tag{4.1}$$

by exploiting the recursion in (1.7). It is a challenging task to obtain more refined results. A positive sequence $L(n)$ is said to be regularly varying of index θ if $\lim_{n\to\infty}L(\lfloor\lambda n\rfloor)/L(n)=\lambda^\theta$ for all $\lambda>0$. (See Bojanic and Seneta [1] for a unified theory of regularly varying sequences.) Figures 3 and 4 suggest that $g(n)$ may be regularly varying of index θ. Furthermore, it is of interest to see if $g(n)/n^\theta$ converges to a constant (which would imply that $g(n)$ is regularly varying of index θ). If so, how can this constant be characterized?

For $p\neq 1/2$, no recursion like (1.7) is available. It seems difficult to obtain an asymptotic result on $nh(p)-\mathbb{E}_p\,|\mathscr{A}_{\mathsf{P}}(n)|$ similar to (4.1). Furthermore, $\mathrm{Var}_p(|\mathscr{A}_{\mathsf{P}}(n)|)$, the variance of $|\mathscr{A}_{\mathsf{P}}(n)|$, is also of interest and importance. Even for $p=1/2$, it seems challenging to derive the asymptotic behavior of $\mathrm{Var}_{1/2}(|\mathscr{A}_{\mathsf{P}}(n)|)$ as $n\to\infty$.

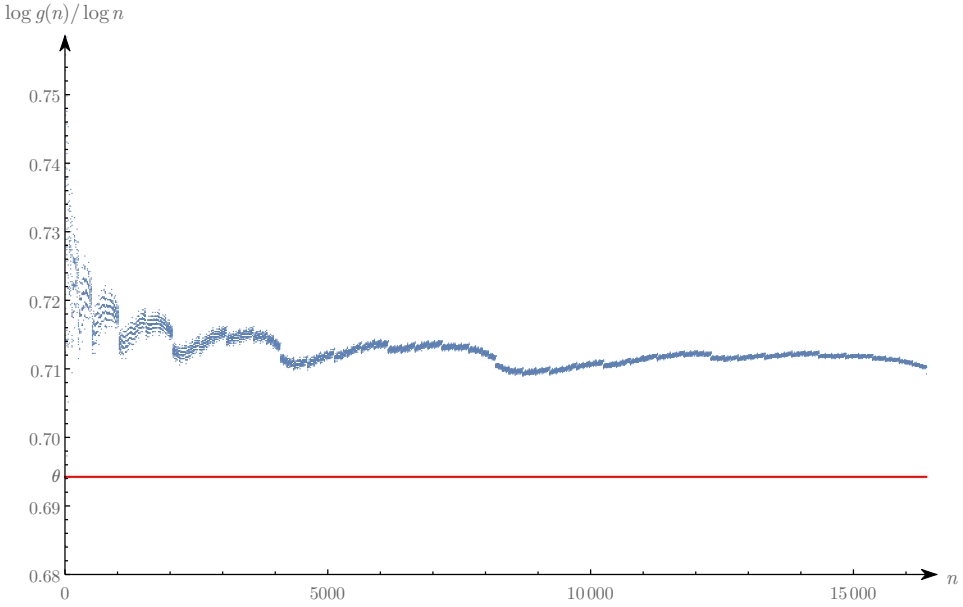

Figure 1. Plot of $\log g(n)/\log n$ versus n.

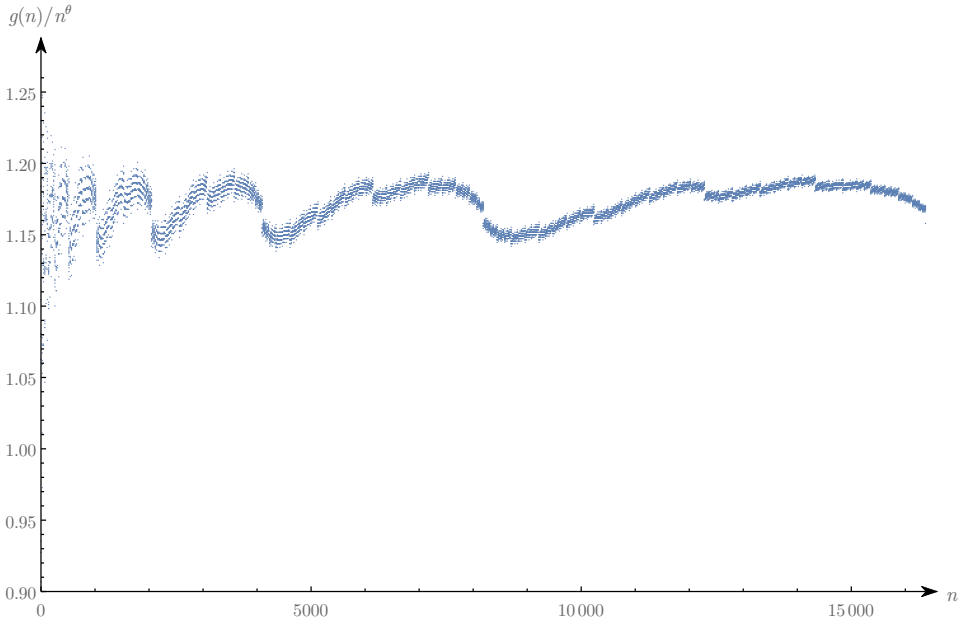

FIGURE 2. Plot of $g(n)/n^\theta$ versus n with $\theta = \log[(1+\sqrt{5})/2]$.

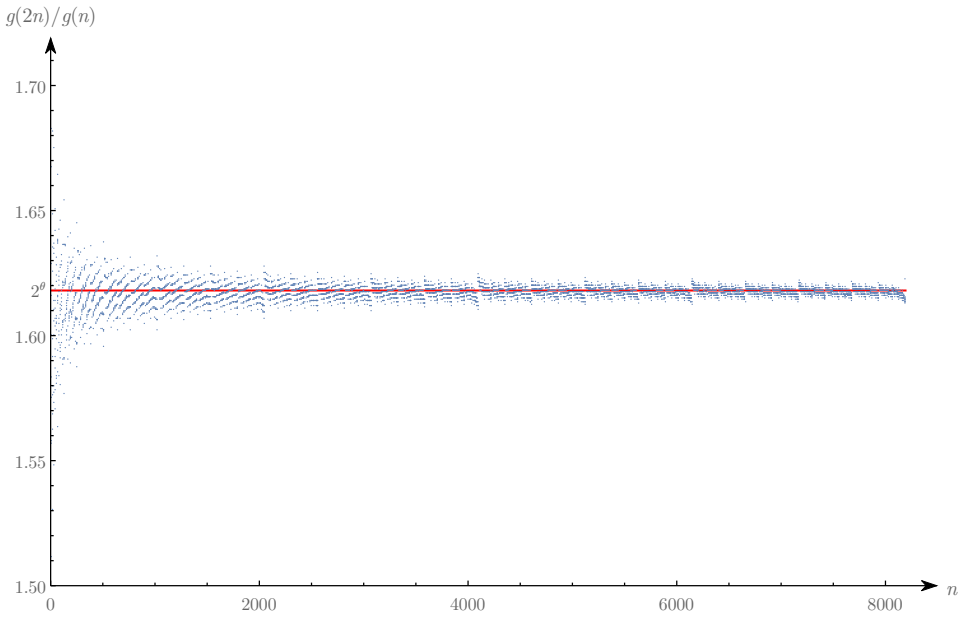

FIGURE 3. Plot of $g(2n)/g(n)$ versus n.

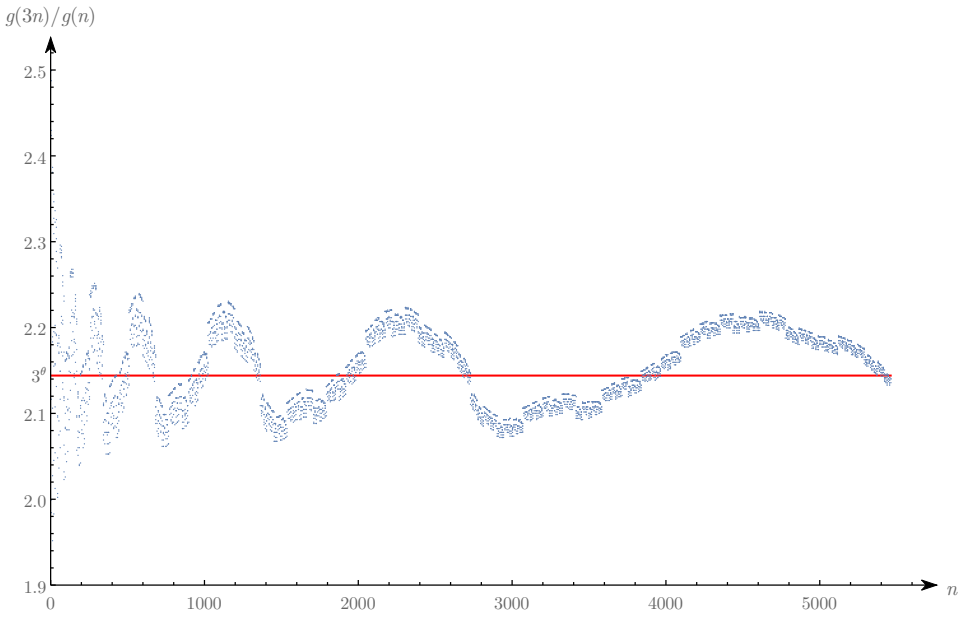

FIGURE 4. Plot of $g(3n)/g(n)$ versus n.

Acknowledgment. The authors gratefully acknowledge support from the Ministry of Science and Technology, Taiwan, ROC.

REFERENCES

1. Bojanic, R. & Seneta, E. (1973). A unified theory of regularly varying sequences. *Mathematische Zeitschrift* 134(2): 91–106.
2. Cover, T. (1973). Enumerative source encoding. *IEEE Transactions on Information Theory* 19(1): 73–77.
3. Elias, P. (1972). The efficient construction of an unbiased random sequence. *Annals of Mathematical Statistics* 43(3): 865–870.
4. Jacquet, P. & Szpankowski, W. (1999). Entropy computations via analytic depoissonization. *IEEE Transactions on Information Theory* 45(4): 1072–1081.
5. Kallenberg, O. (2002). *Foundations of Modern Probability*. New York: Springer.
6. Pae, S.-i. (2013). Exact output rate of Peres's algorithm for random number generation. *Information Processing Letters* 113(5): 160–164.

7. Pae, S.-i. & Loui, M. C. (2006). Randomizing functions: simulation of a discrete probability distribution using a source of unknown distribution. *IEEE Transactions on Information Theory* 52(11): 4965–4976.
8. Peres, Y. (1992). Iterating von Neumann's procedure for extracting random bits. *The Annals of Statistics* 20(1): 590–597.
9. Prasitsupparote, A., Konno, N., & Shikata, J. (2018). Numerical and non-asymptotic analysis of Elias's and Peres's extractors with finite input sequences. *Entropy* 20(10): article #729, 19 pages.
10. Ryabko, B. & Matchikina, E. (2000). Fast and efficient construction of an unbiased random sequence. *IEEE Transactions on Information Theory* 46(3): 1090–1093.
11. Von Neumann, J. (1951). Various techniques used in connection with random digits. *National Bureau of Standards Applied Math Series* 12: 36–38.