Robust self-unbiased, quantum random number generator based on avalanche photodiodes

Fang-Xiang Wang,¹ Chao Wang,¹ Wei Chen,¹† Shuang Wang,¹†‡ De-Yong He,¹
Fu-Sheng Lv,² Zhen-Qiang Yin,¹ Hong-Wei Li,¹ Guang-Can Guo,¹† and Zheng-Fu Han,¹
¹Key Laboratory of Quantum Information, University of Science and Technology of China, Hefei 230026, China
²Department of Mathematics and LPMC, Nankai University, Tianjin 300071, China

We propose and demonstrate a scheme to realize a high-efficiency truly quantum random number generator (RNG) at room temperature. With the effective time bin encoding method, the avalanche pulses of APD is converted into intrinsically self-unbiased random number bits that is robust to slow varying noise. Light source is not necessary in this scheme. Furthermore, the experiment result indicates that a high speed RNG chip based on the scheme is potentially available with integratable APD array.

PACS numbers:

I. INTRODUCTION

Random numbers are important in many fields of scientific research and real-life applications, such as fundamental physical research, computer science and the lottery industry. Although pseudo random numbers can be generated by computer software and hardware with very high speed, high quality truly random number generations (TRNGs) must be adopted in some important applications. For example, TRNGs play important roles in information security, in which quantum cryptography is an emerging technology with potential applications to the next generation information security infrastructure. As an assumption of protocol security, truly random number generations are necessary in many parts of a quantum cryptography system, such as quantum state preparation and measurement and raw key sampling.

The unpredictability of a physical procedure is the resource for generating truly random bits, and two steps are typically necessary to generate random bits with these procedures. First, signals related to a random physical procedure must be effectively generated and gathered. There have been many TRNGs based on physics, for example, circuit noise [1–4] and radioactive decay [5]. In all available elements, quantum mechanics is good for generating a nondeterministic signal. Some RNGs are designed with a quantum procedure in nature, such as wave function collapse of single photon due to measurement [6–10], entangled state measurement [11], effects of vacuum fluctuation [12, 13] and quantum phase fluctuation [14]. In many of these schemes, an almost single photon light source is necessary for generating quantum signals [7, 9, 10, 15]. A quantum random number generator (QRNG), in which the light source is not necessary but compatible may have advantages in integration and usage.

The second step of a physical RNG is to implement an effective encoding method to transform these physical signals into random bits. The encoding method will greatly affect the performance of the RNG. The efficiency of encoding methods is a key limitation for the output rate of RNGs. As devices and environments may vary in real time, self-unbiasedness is not a natural result of real-life RNGs, and post-processing will be necessary to generate high-quality random bits. Even commercial products, like IdQuantique Quantis random number generator, for which Photons - light particles - are sent one by one onto a semi-transparent mirror and detected, can not avoid post-processing. Some situations require that post-processing not be used, as algorithms applied to raw bits may reduce the efficiency of the final bits and increase the complexity and cost of the system. Thus, an effective, intrinsically bias-free encoding method that is immune to slowly varying noise interference is a kernel of a TRNG. What should be noted here is the boundary between the encoding method and the post-processing method. Although the boundary is not clearly defined, we adopt the principle that post-processing takes a large amount of resources [16]. There have been many bias-free RNGs, efficiency is a dominant limitation for most ones [10, 15, 17], for example, the efficiency in reference [10] is 40%. Some other one achieved very high rates by encoding the amplitude of the probe current of the detector into multi-bit RNs [18]. However, the amplitude of the probe current is sensitive to devices and environ-
ments, so stable devices with high resolution are required when implementing these RNGs, which indicates higher cost and greater complexity.

So far, the output rate limit of a bias-free RNG is a research focus. Bias-free physical processes are perfect for RNG so that the encoding method will be mostly simple and will use minimal resources. However, processes used in RNG are always biased, so that the encoding method plays a key role in the RNG to obtain unbiased random numbers. John von Neumann first proposed an unbiased encoding method for biased Bernoulli trials [19]. It has been used in QRNG to obtain bias-free random number sequences [17]. However, the efficiency limitation of the von Neumann method is 0.25. The method was subsequently developed in order to obtain a higher efficiency [20–22], among which Elias produced a very high efficiency for infinite situations [20]. Ren et al. proposed another encoding method based on the precise discrimination of photon numbers of two consecutive pulses [10]. This scheme needs high precision devices to discriminate photon numbers, and the efficiency limitation of the method is 0.5.

In this study, we propose a TRNG scheme based on the discrimination avalanche pulses of APD. These pulses can be generated by the dark current of APD or incident photons, so that a light source is compatible but not necessary in the scheme. An intrinsic bias-free encoding method with higher efficiency than previous works is proposed. With the scheme, a post-processing free QRNG is implemented and tested. Several raw sequences of more than $10^9$ bits generated by the QRNG under different conditions with noise interference were then tested using the NIST statistical test suite. The testing results demonstrate that the QRNG can generate a self-unbiased random bit sequence. Furthermore, we tested the system with multi-APDs, and the results demonstrate that APDs can be combined in parallel to increase the random number generation rate, while preserving the post-processing free characteristic of the scheme. The experiment results also indicate the feasibility of implementation of high-quality, robust QRNG chips using an integrated APD array.

II. RNG SCHEME

According to the quantum theory of lasers, the photon statistics of a laser pulse operating above threshold follows the Poisson distribution [23], which can be preserved after drastic attenuation [24]. The Poisson distribution is

$$P_\lambda(n) = \frac{\lambda^n}{n!} e^{-\lambda},$$

where $\lambda$ is the mean photon number of a laser pulse and $P_\lambda(n)$ is the probability that the pulse contains $n$ photons. Thus, the coherent state produces an unpredictable photon number for every detection, and this quantum property can be used to implement QRNG.

If the detection efficiency of the APD is not taken into account, the probability of an avalanche pulse caused by a laser pulse produces a photon number $n > 0$ in the pulse, which can be described as

$$\sum_{n>0} P_\lambda(n) = 1 - P_\lambda(n = 0) = 1 - e^{-\lambda}.$$  \hspace{1cm} (2)

The avalanche pulses of an APD can be generated by dark currents. Because of thermal fluctuation, electrons of the APD may transit from the top of the valence band to the conduction band. Electrons in the conduction band are sped up by the high reverse-bias electric field and lead to avalanche pulses. Because the thermal fluctuation at room temperature is much smaller than the energy gap between the valence band and the conduction band, the transiting probability is very small. We use the tight-binding approximation here. Thus, the transiting events of different atoms are independent identically distributed (IID), and the statistics of total events follow the Bernoulli distribution

$$P(n_1 = k) = \binom{N_1}{k} p_1^k (1 - p_1)^{N_1-k}$$

where $n_1$ is the total electron number transiting to the conduction band during a certain period $\tau$, $N_1$ is the total electron number at the top of valence band, $p_1$ is the transiting probability of a single electron, and $P(n_1 = k)$ is the transiting probability of $k$ electrons. As $p_1$ is much smaller than 1 and $N_1$ is large in the material, the transiting process follows the Poisson limit theorem

$$\lim_{N_1 \to \infty, p_1 \to 0} \binom{N_1}{k} p_1^k (1 - p_1)^{N_1-k} = \frac{\lambda_k}{k!} e^{-\lambda_1} = P_\lambda(k),$$

where $\lambda_1$ is the mean electron number transiting to the conduction band during $\tau$. Thus, the transiting process follows the Poisson distribution, and so does the dark count. The dark count is then as usable as a laser pulse. Dark counts of APD were first used by Tawfeeq to propose a RNG scheme [25]. The scheme was easier and provided a new idea regarding RNG based on APD. However, that scheme did not show an effective encoding method and could not output bias-free random number sequence. And RNG with high randomness based on dark counts of APD has not been implemented before.

Because the sum of Poisson-distributed random variables follows the Poisson distribution, the sum probability of an avalanche during $\tau'$ follows

$$p = P_\lambda(n > 0) = 1 - e^{-\lambda'},$$

where $\tau'$ is the detection window, $\lambda' = \eta(\lambda+\lambda_1)$, and $\eta$ is the detection efficiency of APD. Clearly, the probability of no avalanche pulse is $q = 1 - p = e^{-\lambda'}$. The detection process is then a Bernoulli trial. A bias-free encoding method for biased Bernoulli trials is then proposed.
here. It is an extension of the von Neumann method but with much higher efficiency. The encoding method is constructed as follows.

We consider the physical system of an avalanche photodiode (APD) working on the Geiger mode and treat a detection window of APD as a time bin and sequence these time bins with time. According to the discussion above, avalanches caused by laser and thermal fluctuation in different time bins are IID if experimental parameters are constant, namely, $p = p_0$. In fact, experimental parameters are hardly constant. We claim that our encoding method constructed here also applies to the condition that experimental parameters vary slowly so that the encoding method is effective and robust. We mark a "1" in a time bin if an avalanche happened in the corresponding detection window; otherwise we mark a "0" in it. Considering $N$ time bins happened successively as a time-bin block. There are totally $(N \choose k)$ possible combinations when $k$ "1" are marked in the block if we do not get additional information about the block, namely, the uncertainty of these $N$ time bins are $(N \choose k)$. These equiprobable $(N \choose k)$ possible combinations are then encoded into uniform random numbers from 0 to $(N \choose k) - 1$. The encoding processes are one-to-one mapping and the mapping function is

$$f(k_1, k_2, \cdots, k_k) = \sum_{j=1}^{k} \left( N - k_j \right) \left( k - j + 1 \right), \quad (6)$$

where $k_j$ means that the $j$-th "1" happened in the $k_j$-th time bin.

Then, we go to the interpretation of the mapping function. As discussed, if we only know that there are $k$ "1" in the time-bin block, the uncertainty is $(N \choose k)$. If we get the temporal information in the time bin sequence of the first "1", namely, $k_1$ is known, the uncertainty reduces to $(N-k_1)\choose{k_1-1}$, in other words, the information content we get is $(N \choose k) - (N-k_1 \choose k_1-1)$. As similar, when $k_2$ is also known, the information content we get increases by $(N-k_1)\choose{k_1-1} - (N-k_2)\choose{k_2-1}$. The uncertainty will reduce further if $k_3, k_4, \cdots$ are also known. In the extreme case, if $k_1, k_2, \cdots, k_k$ are all known, the uncertainty remaining becomes zero, and we get all information content about these $(N \choose k)$ combinations. We sum all information content got with $k_1$ to $k_k$ and the summation is the random number we want. The mapping function is $f(k_1, k_2, \cdots, k_k) = \sum_{j=1}^{k} \left( N - k_j \right) \left( k - j + 1 \right) - (N-k_j)\choose{k-j-1}$. The mapping function is monotonically increasing.

The maximum possible number got from mapping function is $f(k_1 = 1, k_2 = 2, \cdots, k_k = k) = \sum_{j=1}^{k} (k-j+1) = (N \choose k) - 1$. And the minimum possible number is $f(k_1 = N-k+1, k_2 = N-k+2, \cdots, k_k = N) = \sum_{j=1}^{k} (N-(N-k+j)) = 0$. Thus, the encoding process is one-to-one mapping and the random number is in $(N \choose k)$ representation.

Taking into account the wide applications of binary random numbers, the $(N \choose k)$-ary encoding method can go further and be modified by the binary method proposed by Elias [20]. Elias expand $(N \choose k)$ into subblocks as follows:

$$\left( N \choose k \right) = \alpha_m 2^m + \alpha_{m-1} 2^{m-1} + \cdots + \alpha_0 2^0. \quad (7)$$

So that $\alpha_m, \alpha_{m-1}, \cdots, \alpha_0$ is the binary expansion of integer $(N \choose k)$, where $\alpha_m = 1, \alpha_i = 0$ or 1 for $0 \leq i < m$. The subblock related to the $m$ term should be abandoned, as it contains either one or no members and could not be encoded into unbiased bits. Surmise the non-zero binary expansion coefficients are $\alpha_m, \alpha_{m-1}, \cdots, \alpha_i$. If $f(k_1, k_2, \cdots, k_k) < 2^m$, convert $f(k_1, k_2, \cdots, k_k)$ into a m-bit binary number directly. If $2^m + \sum_{s=1}^{r} 2^s \leq f(k_1, k_2, \cdots, k_k) < 2^m + \sum_{s=1}^{r+1} 2^s$, then convert $f(k_1, k_2, \cdots, k_k)$ into a $(r+1)$-bit number directly (the schematic graph of encoding process is shown in Figure 1). $f(k_1, k_2, \cdots, k_k)$ is abandoned if $i_{r+1} = 0$.

The encoding method constructed requires $p$ to be constant among these $N$ time bins (detection windows) of the same block. But the $p$ values in different blocks are not necessarily identical, so that the method is robust to environment noise. As detection interval of APD can be as less as $\sim n_s$, it is only $\sim \mu s$ when $N \sim 100$. It is reasonable to consider that parameters, depending on environments, which are slowly varying, are invariable in such a short time. These parameters can be laser intensity, temperature, etc. In addition, the raw random number sequence output remains uniform even the slowly varying interferences are periodic (see sections III and IV). Thus, the encoding method is effective, efficiency and robust in practice.

The encoding method is effective for any $k \neq 0, N$. Taking into account all possible value of $k$, the average
encoding efficiency per time bin before binary expansion is

\[ H(N, p) = -\frac{1}{N} \sum_{k=1}^{N-1} \binom{N}{k} p^k (1 - p)^{N-k} (\log_2 \frac{1}{\binom{N}{k}}). \] (8)

A higher \( H(N, p) \) indicates a higher efficiency. For any integer \( N \geq 2 \), the optimal \( p \) for the average encoding efficiency \( H(N, p) \) is \( \frac{1}{2} \), and \( H(N, p) \rightarrow S(p) \) as \( N \rightarrow \infty \) (see Figure 2(a)), where \( S(P) \) is the Shannon entropy of a single Bernoulli trial. In addition, \( H(N, \frac{1}{2}) \) increases with \( N \) and converges to 1, as shown in Figure 2(a), 2(b). For instance, \( H(5, \frac{1}{2}) = 0.5604 \), while \( H(10, \frac{1}{2}) = 0.7294 \), the efficiency is much higher than perivous ones based on single photon discrimination [10, 15, 17]. The theorems are proven in the appendix.

The more subblocks \( \binom{N}{k} \) divides into, the fewer possible combinations and thus the less uncertainty of the subblock there will be. In addition, the uncertainties among different subblocks (blocks) are not utilized in both encoding methods above. Thus, more subblocks indicate more uncertainty among subblocks and hence less extracted entropy and encoding efficiency, as the total entropy is conserved. The unbiased output sequences in binary representation are therefore obtained at the cost of entropy or efficiency, and the efficiency becomes \( H(N, p) = \frac{1}{N} \sum_{k=1}^{N-1} \binom{N}{k} p^k (1 - p)^{N-k} (\alpha_m 2^m k m + \alpha_{m-1} 2^{m-1} k (m - 1) + \cdots + \alpha_0 2^k \cdot 1) \), where the subscript \( k \) means there are \( k \) “1” in the block. The efficiency after expansion is shown in Figure 2(b). Moreover, more blocks mean more resources to be required.

It should be noted that a similar spatial encoding method has been proposed for a different physical system by Marangon et al. [26].

### III. EXPERIMENTAL SETUP

Three scenarios were designed in order to evaluate the feasibility and the robustness of this scheme. (a) Avalanche singles from a single APD was acquired and data were encoded according to the method of part II, which can verify if the method can generate bias-free random number. A laser diode (LD) was added in this setup as an optional light source to increase the random number generation rate. By modulating the power of LD, we can simulate an additional noise and the variation of the avalanche efficiency, so that the robustness of scheme can be evaluated. (b) We added an additional APD to the system of setup (a). Two APDs were grouped, and the avalanche pulses were gathered and processed parallelly to generate random numbers. This setup is to evaluate the possibility to increase the random number generation rate with APD arrays while keeping the bias-free feature of random numbers. (c) In order to demonstrate the system can work properly without light source, the LD of setup (b) was removed and the avalanche pulses were generated only by dark counts of APDs.

The system diagram is shown in Figure 3. A pulse LD with the wavelength of 1550 nm is used as an optional light source and triggered by 1 MHz electronic pulses. An intensity modulator (IM) following the LD was used to modulate the power of light pulses from LD. The output light pulses from IM were divided into two parts by a beam splitter (BS) and were attenuated to the single photon level by two electronically variable optical attenuators (EVOA), then coupled to two APDs (PGA-300, Princeton Lightwave), individually.

APDs worked in Geiger mode and were triggered by gating pulses with width of 2.5 ns and the frequency of 1 MHz, individually. APDs used to detect single photons are commonly cooled from -30 °C to -50 °C in order to reduce the dark count rate, such as in quantum key distribution applications. In our experiments, dark counts of APDs can be used as a resource to generate random bits as well as external photons. Thus, the cooling processes for APDs are not necessary, and APDs can work at room temperature, which makes the system more practical and less expensive. The working temperature of APDs in our experiments was approximately 23 °C.

The avalanche pulses of the two APDs were discriminated and amplified, then were sent into a time-to-digital converter (Agilent U1051A Acqiris TC890) to be processed. The TDC has one input channel of start signal and 6 input channels of stop signals, and can convert the time intervals between the stop and start signals into 32-bit numbers. The discrimination results from TDC were sequentially numbered with the time bin of 1s, and consequently encoded into random numbers according to the method of part II in real time. As a proof-of-principle experiment, the encoding process was executed every four successive detection windows. All trigger signals in the system were synchronized by a home-made circuit and the delay among them can be adjusted in the step of 10 ps, respectively.

The system can be divided into three parts, as shown in Figure 3. In scenario (a), part II was removed, thus single APD was used to generate random numbers consequently. In scenario (c), part I was removed, and the random numbers were exclusively generated by dark counts of APDs. We added a sinusoidal driving signal with the frequency of 0.05 Hz and amplitude of 3 V to the IM in scenario (a) and (b) to simulate an external noise. The average counting probabilities of the two APDs were initialized to 0.5 by adjusting the EVOAs ahead of them individually. According to the sinusoidal modulation of IM, the probabilities of the APDs varied from 0.3 to 0.7, which can be regarded as an external noise to APDs.

### IV. RESULTS AND DISCUSSION

In our experiments, we set \( N \) as 4 as discussed. The total 16 types of detection results can be classified as 5 subsets according to the \( k \) value. Two of these sixteen types of detection results, subsets with \( k = 0 \) and \( k = 4 \),
were abandoned, while the other fourteen were set to output random numbers.

First, we analyzed the uniformity of the random bit output from these experiments. Only the uniformity of detection results of the scenario (a) are demonstrated, as shown in Figure 4, because the other two experiments contain very similar properties. Let \( p(x, y) \) represent the population of elements of a 16 \( \times \) 16 matrix, which indicates that neighboring detection results \( x \) and \( y \) happen successively, and integers \( x, y \in \{0, 1, 2, \ldots, 15\} \) represent the 16 possible detection results for every four consecutive detections.

In Figure 4, different colors indicate different population \( p(x, y) \). Clearly, the matrix is symmetric and is partitioned into subblocks. The symmetric matrix shows that \( p(x, y) = p(y, x) \) for any \( x, y \). It means that there is no time correlation between successive \( x \) and \( y \). \( p(x, y) \) of different elements in the same subblock are identical. It means the population is uniform in subblocks. Thus \( p(x, y) = p(y, x) = p(x)p(y) \). This represents that each detection event is independent, and the imperfections of the beam splitter and detection efficiencies make no difference to the uniformity of the randomness extraction system. The experimental results are consistent with the theory we proposed in Section II.

The raw binary random number samples output from the RNG scheme. For every four detection windows, \( x \) or \( y = 0 \) indicates that no avalanche pulse is detected, \( x \) or \( y \in \{1, 4\} \) indicates that one avalanche pulse is detected, \( x \) or \( y \in \{5, 10\} \) indicates that two avalanche pulses are detected, \( x \) or \( y \in \{11, 14\} \) indicates that three avalanche pulses are detected and \( x \) or \( y = 15 \) means four avalanche pulses are detected, where \( x, y \in integer \). The altitude represents the population of detection results in which \( x \) and \( y \) happen successively.

FIG. 2: (Color online) The average encoding efficiency per time bin (in red) increases with \( N \). (a) \( H(N, p) \) (in blue) converges to \( S(p) \) (in red) with infinite \( N \). (b) The encoding efficiency will converge to 1 with infinite \( N \) when \( p = \frac{1}{2} \). The corresponding efficiency after the binary expansion (in blue) is lower (\( N > 2 \)) but will converge to the previous one at large \( N \).

FIG. 3: (Color online) Schematic setup of the experiment. LD: Laser diode; IM: optical intensity modulator; BS: beam splitter; Att: electronically variable optical attenuators (EVOA); OF: optical fiber; APD: avalanche photodiode; TDC: time-to-digital converter; PC: personal computer.

FIG. 4: (Color online) The uniformities of random numbers output from the RNG scheme. For every four detection windows, \( x \) or \( y = 0 \) indicates that no avalanche pulse is detected, \( x \) or \( y \in \{1, 4\} \) indicates that one avalanche pulse is detected, \( x \) or \( y \in \{5, 10\} \) indicates that two avalanche pulses are detected, \( x \) or \( y \in \{11, 14\} \) indicates that three avalanche pulses are detected and \( x \) or \( y = 15 \) means four avalanche pulses are detected, where \( x, y \in integer \). The altitude represents the population of detection results in which \( x \) and \( y \) happen successively.

FIG. 4: (Color online) The uniformities of random numbers output from the RNG scheme. For every four detection windows, \( x \) or \( y = 0 \) indicates that no avalanche pulse is detected, \( x \) or \( y \in \{1, 4\} \) indicates that one avalanche pulse is detected, \( x \) or \( y \in \{5, 10\} \) indicates that two avalanche pulses are detected, \( x \) or \( y \in \{11, 14\} \) indicates that three avalanche pulses are detected and \( x \) or \( y = 15 \) means four avalanche pulses are detected, where \( x, y \in integer \). The altitude represents the population of detection results in which \( x \) and \( y \) happen successively.
FIG. 5: (Color online) (a) (b) (c) Min-entropy of samples (data point) output from scenarios (a), (b) and (c), respectively. The linear fitting function (fitting line) is shown. (d) The relative deviations between min-entropy and Shannon entropy of uniform distribution of all three scenarios.

parameters in the NonOverlappingTemplate item, each of them indicates a corresponding occurrence of pre-specified 9-bit target strings. Too many or too few occurrences of a certain string would lead to the failure of the corresponding test with this parameter. It is worth mentioning that only one of the 148 items was responsible for the failures of the NonOverlappingTemplate tests, and we did not find obvious commonality in these failure items.

Furthermore, min-entropy evaluation was employed for all these experiments. Min-entropy, defined as

\[ H_\infty = -\log_2 \left( \max_i p(x_i) \right), \]  

(9)

is the evaluation of the worst situation, where \( p(x_i) \) is the probability of possible output \( x_i \), and \( \max p(x_i) \) is the maximal value of all \( p(x_i) \). It is a strong way to measure the information content, while Shannon entropy is a weighted average evaluation. Both min-entropy and Shannon entropy are special cases of Renyi entropy \([28, 29]\). Shannon entropy is the upper bound of min-entropy, and they coincide if and only if the distribution of the variable is uniform \([30, 31]\). Min-entropy evaluation is therefore a good way to evaluate the quality of randomness of random numbers.

The min-entropy of raw data from all scenarios were evaluated, where \( i = 0, 1, \ldots, d - 1 \) and \( p(x_0), p(x_1), \ldots, p(x_{d-1}) \) represent the probabilities of "0", "1", \ldots, "d-1" in binary representation respectively. As shown in Figure 5, the results show that deviations between min-entropy and Shannon entropy of uniform distribution are all of the order of 0.001. This is within the statistical error (~3·10^{-3}, the red dot line in Figure 5(d)), as the statistical amount is 10^5·2^d bits, which indicates a good quality of randomness of the random numbers.

It is worth noting that samples to be tested are extracted continuously by days. Thus, interferences that may affect the experiments are more complex. Despite a few failed tests, the results of the tests and analysis above indicate good quality of randomness of data from all of these three scenarios and no need for any type of complex post-processing. The results of the scenario (a) were consistent with the theory analysis and show the robust of our RNG from slowly varying interferences. Scenarios (b) and (c), with double APD, suggest an APD array scheme, which is promising to break through the output rate limitation. The scenario (c) also gave a relatively strict proof that the dark count of APD is usable for RNG. In all scenarios, no special measures were adopted to reduce the interference of background light noise and temperature fluctuation. The light source for this scheme, to be or not to be, is not a question any more.

V. CONCLUSION

In conclusion, we have proposed and realized a bias-free TRNG scheme, which is robust and high-efficiency. Dark counts of APD were used as a resource in this scheme, so that a deep cooling process is not necessary and the system can work at room temperature. The photons arriving at the APD and slowly varying interferences affected only the efficiency of the RNG rather than the randomness of the bit series, so the scheme is compatible with light source and background photons. The experimental results also indicate the feasibility of integrating an APD array into a RNG chip, which can effectively increase the output rate of random bits and make the scheme have more practical value.

Acknowledgements

This work has been supported by the National Basic Research Program of China (Grants No. 2011CBA00200 and No. 2011CB921200), the National Natural Science Foundation of China (Grant Nos. 61101137, 61201239, 61205118, 11304397). Fang-Xiang Wang and Chao Wang contributed equally to this work.

[1] W. T. Holman, J. a. Connelly, and a. B. Dowlabadi, IEEE Trans. Circuits Syst. I Fundam. Theory Appl. 44, 521 (1997).
[2] C. S. Petrie and J. a. Connelly, IEEE Trans. Circuits Syst. I Fundam. Theory Appl. 47, 615 (2000).
[3] B. Jun and P. Kocher, Cryptogr. Res. Inc. White Pap. 27, 1 (1999).
[4] M. Bucci, L. Germani, R. Luzzi, A. Trifiletti, and M.
TABLE I: The standard statistical test results of NIST. Twenty samples of 1 Gbit are tested for the first two experiments and 7 samples for the third experiment, as the dark count rate is lower. For the tests outputting multiple p-values and proportions, the worst case is adopted.

| Testing item | First Experiment | Second Experiment | Third Experiment |
|--------------|------------------|------------------|-----------------|
|              | Proportion       | p-value          | Proportion      | p-value          |
| Frequency    | 20/20            | 20/20            | 20/20           | 7/7              |
| BlockFrequency | 20/20          | 20/20            | 20/20           | 7/7              |
| CumulativeSums | 20/20          | 20/20            | 20/20           | 7/7              |
| Run          | 20/20            | 20/20            | 20/20           | 7/7              |
| LongestRun   | 20/20            | 20/20            | 20/20           | 7/7              |
| Rank         | 20/20            | 20/20            | 20/20           | 7/7              |
| FFT          | 20/20            | 20/20            | 20/20           | 7/7              |
| NonOverlappingTemplate | 16/20   | 20/20            | 18/20           | 6/7              |
| OverlappingTemplate | 20/20      | 20/20            | 20/20           | 7/7              |
| Universal    | 20/20            | 20/20            | 20/20           | 7/7              |
| ApproximateEntropy | 20/20    | 20/20            | 20/20           | 7/7              |
| RandomExcursions | 20/20    | 20/20            | 20/20           | 7/7              |
| RandomExcursionsVariant | 20/20   | 20/20            | 20/20           | 7/7              |
| Serial       | 20/20            | 20/20            | 20/20           | 7/7              |
| LinearComplexity | 20/20      | 20/20            | 20/20           | 7/7              |

Varanou, 52, 403 (2003).
[5] H. I. Masatugu Isida, Ann. Inst. Stat. Math. 8, 119 (1956).
[6] A. Stefanov, N. Gisin, and O. Guinnard, Optical quantum random number generator, J. Mod. Opt. 47, 595-598 (2000).
[7] T. Jennewein, U. Achleitner, G. Weihs, H. Weinfurter, and A. Zeilinger, Rev. Sci. Instrum. 71, 1675 (2000).
[8] H. Ma, Y. Xie, and L. Wu, Appl. Opt. 44, 7760-7763 (2005).
[9] M. Ren, E. Wu, Y. Liang, Y. Jian, G. Wu, and H. Zeng, Phys. Rev. A 83, 023820 (2011).
[10] S. Pironio, A. Acn, and S. Massar, Nature 464, 1021 (2010).
[11] C. Gabriel, C. Wittmann, D. Sych, R. Dong, W. Mauerer, U. L. Andersen, C. Marquardt, and G. Leuchs, Nat. Photonics 4, 711 (2010).
[12] M. Jofre, M. Curty, F. Steinlechner, G. Anzolin, J. P. Torres, M. W. Mitchell, and V. Pruneri, Opt. Express 18, 13029 (2010).
[13] F. Xu, B. Qi, X. Ma, H. Xu, H. Zheng, and H.-K. Lo, Opt. Express 20, 12366 (2012).
[14] J. F. Dynes, Z. L. Yuan, a. W. Sharpe, and a. J. Shields, Appl. Phys. Lett. 93, 031109 (2008).
[15] Stipcevic, Mario, and etin Kaya Ko, http://www.cs.ucsb.edu/~koc/cren/docs/w06/trng.pdf
[16] W. Wei and H. Guo, Opt. Lett. 34, 1876 (2009).
[17] Y. Liu, M.-Y. Zhu, B. Luo, J.-W. Zhang, and H. Guo, Laser Phys. Lett. 10, 045001 (2013).
[18] J. Von Neumann, Appl. Math Ser. (1951).
[19] P. Elias, Ann. Math. Stat. (1972).
[20] W. Hoefding and G. Simons, Ann. Math. Stat. 41, 341 (1970).
[21] P. Samuelson, J. Am. Stat. Assoc. 63, (1968).
[22] M. Scully, Quantum Optics (Cambridge university press, 1997).
[23] Y. Hu, X. Peng, T. Li, and H. Guo, Phys. Lett. A 367, 173 (2007).
[24] S. Tawfeeq, J. Light. Technol. 27, 5665 (2009).
[25] D. G. Marangon, Vallone, and P. Villaresi, Scientific reports, 4, 5490 (2014).
[26] http://csrc.nist.gov/groups/ST/toolkit/rng/
[27] A. Renyi, Fourth Berkeley Symp. Math. Stat. 63, 547 (1961).
[28] P. Bromley, Stat. Inf. Ser. ( (2004).
[29] C. Cachin, Entropy Measures and Unconditional Security in Cryptography, Swiss Federal Institute of Technology Zurich, 1997.
[30] G. Smith, in Found. Quant. Inf. Flow (Springer Berlin Heidelberg, 2009), pp. 28302.

Appendix

Lemma 1. Define a function \( f \) with expression below

\[
f(N, k, p) = p^k(1-p)^{N-k} + p^{N-k}(1-p)^k
\]

Then, if \( N \geq 2 \) and \( p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1) \), the equation about \( k \)

\[
f(N, k, p) = f(N, k, \frac{1}{2})
\]

has two and only two roots in \([0, N]\).

Proof. First, we simplify equation (1) below

\[
(1-p)^N \left( \frac{p}{1-p} \right)^k + p^N \left( \frac{1-p}{p} \right)^k = 2 \cdot \left( \frac{1}{2} \right)^N.
\]

The left of equation is symmetric about \( p = \frac{1}{2} \), so we only need to consider the case \( p \in (\frac{1}{2}, 1) \), and then \( \frac{1-p}{p} > 1 \). Let \( a = (p/(1-p))^k \). Then for any \( k \in [0, N] \), we have
Let \( a \in [1, (\frac{p}{1-p})^N] \). As \( k \) and \( a \) are one to one, we simplify equation (2) and obtain

\[
(1 - p)^N a + p^N \frac{1}{a} = 2 \cdot \frac{1}{2}.
\]

Define the function

\[
\varphi(a) = (1 - p)^N a + p^N \frac{1}{a} - 2 \cdot \frac{1}{2},
\]

as \( p \in (\frac{1}{2}, 1) \), we have

\[
\varphi(1) = \varphi\left((\frac{p}{1-p})^N\right) = (1 - p)^N + p^N - 2 \cdot \frac{1}{2} > 0.
\]

Next, the derivative of \( \varphi \) is

\[
\varphi'(a) = (1 - p)^N - p^N \frac{1}{a^2}.
\]

Let \( \varphi'(a_0) = 0 \), we obtain \( a_0 = \sqrt{(\frac{p}{1-p})^N} \). Furthermore, if \( a \in [1, \sqrt{(\frac{p}{1-p})^N}] \), then \( \varphi'(a) < 0 \); if \( a \in (\sqrt{(\frac{p}{1-p})^N}, (\frac{p}{1-p})^N) \), then \( \varphi'(a) > 0 \). In addition, \( \varphi\left((\frac{p}{1-p})^N\right) = 2 \sqrt{p^N (1 - p)^N} - 2 \cdot \frac{1}{2} < 0 \), so the equation \( \varphi(a) = 0 \) has two roots in \((1, \frac{p}{1-p})\).

As \( k \) and \( a \) are one to one, it is easy to prove that equation (1) has two and only two roots in \((0, N)\), which we denote as \( x_1 \) and \( x_2 \), \( x_1 < x_2 \), then if \( k \in [0, x_1) \cup (x_2, N] \), \( f(N, k, p) > f(N, k, \frac{1}{2}) \); if \( k \in (x_1, x_2) \), \( f(N, k, p) < f(N, k, \frac{1}{2}) \). In addition, as \( f(N, k, p) = f(N, N - k, p) \), we have \( x_1 + x_2 = N \).

This completes the proof of Lemma. □

Theorem 1. For any integer \( N \geq 2 \), the optimal \( p \) for normalized extracted entropy \( H(N, p) \) is \( \frac{1}{2} \), and \( H(N, p) \to S(p) \) as \( N \to \infty \), where \( S(p) \) is the Shannon entropy of a single Bernoulli trial, and \( H(N, p) \) is defined as

\[
H(N, p) = -\frac{1}{N} \sum_{k=1}^{N-1} N_k p^k (1 - p)^{N-k} (\log_2 \frac{1}{N_k}),
\]

and \( N_k = \binom{N}{k} \) is the binomial coefficient.

Proof. Let

\[
f(N, k, p) = p^k (1 - p)^{N-k} + p^{N-k} (1 - p)^k,
\]

then

\[
H(N, p) = -\frac{1}{N} \sum_{k=1}^{N-1} N_k p^k (1 - p)^{N-k} (\log_2 \frac{1}{N_k})
\]

\[
= \frac{1}{N} \sum_{k=0}^{N} N_k p^k (1 - p)^{N-k} (\log_2 \frac{1}{N_k})
\]

\[
= \frac{1}{2N} \sum_{k=0}^{N} [p^k (1 - p)^{N-k} + p^{N-k} (1 - p)^k] N_k (\log_2 N_k)
\]

\[
= \frac{1}{2N} \sum_{k=0}^{N} f(N, k, p) N_k (\log_2 N_k)
\]

It is clear that when \( p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1) \),

\[
\sum_{k=0}^{N} f(N, k, p) N_k = \sum_{k=0}^{N} f(N, k, \frac{1}{2}) N_k = 2,
\]

and \( f(0, 0, p) = (1 - p)^N + p^N > 2 \cdot (\frac{1}{2})^N = f(N, 0, \frac{1}{2}) \). Thus, there exists an integer \( k_0 \to (0, N) \) such that \( f(N, k_0, p) < f(N, k_0, \frac{1}{2}) \). Additionally, from the proof of Lemma 1 we know there exist \( x_1, x_2 \to (0, N) \), \( x_1 < x_2 \) and \( x_1 + x_2 = N \) such that \( f(N, x_1, p) - f(N, x_1, \frac{1}{2}) = f(N, x_2, p) - f(N, x_2, \frac{1}{2}) = 0 \), and if and only if \( k \to (x_1, x_2) \), \( f(N, k, p) < f(N, k, \frac{1}{2}) \); thus, we have \( k_0 \in [x_1, x_2) \). In addition, if \( x_1, x_2 \) are integers, we have \( k_0 \in [x_1 + 1, x_2 - 1] \); otherwise \( k_0 \in [x_1 + 1, \lfloor x_2 \rfloor] \), where \( \lfloor x_1 \rfloor \) represents the largest integer that is not larger than \( x_1 \).

With the conclusion above, we will show when \( N \geq 2 \) and \( p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1) \), we have \( H(N, p) < H(N, 1/2) \). There are two cases.

Case 1: \( x_1, x_2 \) are not integers, then

\[
2 \left\lfloor \frac{x_1 + 1}{2} \right\rfloor [f(N, k, p) - f(N, k, \frac{1}{2})] N_k
\]

\[
= \left( \sum_{k=0}^{\lfloor x_1 \rfloor} + \sum_{k=\lceil x_2 \rceil}^{N} \right) [f(N, k, p) - f(N, k, \frac{1}{2})] N_k
\]

\[
= [2 - \sum_{k=\lfloor x_1 \rfloor + 1}^{\lfloor x_2 \rfloor}] f(N, k, p) N_k
\]

\[
- [2 - \sum_{k=\lfloor x_1 \rfloor + 1}^{\lfloor x_2 \rfloor}] f(N, k, p) N_k
\]

\[
= \sum_{k=\lfloor x_1 \rfloor + 1}^{\lfloor x_2 \rfloor} [f(N, k, \frac{1}{2}) - f(N, k, p)] N_k.
\]
and

\[ H(N, p) - H(N, \frac{1}{2}) \]

\[ = \sum_{k=0}^{N} [f(N, K, P) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ = \sum_{k=0}^{N} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ + \sum_{k=|x_1|+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ = 2 \sum_{k=0}^{N} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ + \sum_{k=|x_1|+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ \leq 2 \sum_{k=0}^{N} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_{\lfloor x_1 \rfloor+1} \]

\[ + \sum_{k=|x_1|+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_{\lfloor x_1 \rfloor+1} \]

\[ = \sum_{k=|x_1|+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \cdot \log_2 N_{\lfloor x_1 \rfloor+1} - \log_2 N_{\lfloor x_1 \rfloor} < 0. \]

The last inequality holds, because there exists integer

\[ k_0 \in [x_1 + 1, x_2 - 1], \] and \[ x_1 + 1 \leq x_2 - 1 = N - x_1 - 1, \]

thus \[ x_1 + 1 \leq \frac{N}{2}, \] so \[ N_{\lfloor x_1 \rfloor+1} \geq N_{\lfloor x_1 \rfloor}. \]

**Case 2:** \( x_1, x_2 \) are integers, then analogously,

\[ H(N, p) - H(N, \frac{1}{2}) \]

\[ = 2 \sum_{k=0}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_k \]

\[ + \sum_{k=x_1+1}^{x_2-1} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_{\lfloor x_1 \rfloor+1} \]

\[ \leq 2 \sum_{k=0}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_{\lfloor x_1 \rfloor+1} \]

\[ + \sum_{k=x_1+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \log_2 N_{\lfloor x_1 \rfloor+1} \]

\[ = \sum_{k=x_1+1}^{x_2} [f(N, k, p) - f(N, k, \frac{1}{2})] \frac{N_k}{2N} \cdot \log_2 N_{\lfloor x_1 \rfloor+1} - \log_2 N_{\lfloor x_1 \rfloor} < 0. \]

The last inequality holds, because there exists integer

\[ k_0 \in [x_1 + 1, x_2 - 1], \] and \[ x_1 + 1 \leq x_2 - 1 = N - x_1 - 1, \]

thus \[ x_1 + 1 \leq \frac{N}{2}, \] so \[ N_{\lfloor x_1 \rfloor+1} \geq N_{\lfloor x_1 \rfloor}. \]

We have therefore proved that the optimal \( p \) for normalized extracted entropy \( H(N, p) \) is \( \frac{1}{2} \), and we will next show the remaining part.

First, as

\[ 2^N = \sum_{k=0}^{N} \binom{N}{k} = \sum_{k=0}^{N} N_k. \]

we have \( \frac{\log_2 N_k}{N} < 1 \) for \( 0 \leq k \leq N \). Assuming that

\[ 0 < p < \frac{1}{2}, \] the cases that \( \frac{1}{2} < p < 1 \) and \( p = \frac{1}{2} \) are similar. so there exists \( \delta > 0 \) sufficiently small such that

\[ p + \delta < \frac{1}{2}, \]

by the weak law for a binomial distribution,

\[ \lim_{N \to \infty} \sum_{0 < p < \frac{1}{2} + \delta} N_k p^k (1-p)^{N-k} = 1. \]

Thus, given any \( \epsilon > 0 \), there is an \( N' \), such that for \( N > N' \),

\[ \sum_{\frac{1}{2} < \epsilon < \frac{1}{2} + \delta} N_n p^k (1-p)^{N-k} < \epsilon. \quad (4) \]

Thus, together with \( \frac{\log_2 N_k}{N} < 1 \), we have

\[ \sum_{\frac{1}{2} - \epsilon < p < \frac{1}{2}} N_k p^k (1-p)^{N-k} \frac{\log_2 N_k}{N} < H(N, p) \]

\[ < \sum_{\frac{1}{2} - \epsilon < p < \frac{1}{2} + \delta} N_k p^k (1-p)^{N-k} \frac{\log_2 N_k}{N} + \epsilon. \]

Because \( p + \delta < \frac{1}{2} \), when \( |\frac{1}{2} - p| < \delta \), we have

\[ \log_2 N_{\lfloor N(p-\delta) \rfloor + 1} \leq \log_2 N_k \leq \log_2 N_{\lfloor N(p-\delta) \rfloor} + 1, \]

so

\[ \sum_{\frac{1}{2} - \epsilon < p < \frac{1}{2} + \delta} N_k p^k (1-p)^{N-k} \frac{\log_2 N_{\lfloor N(p-\delta) \rfloor}}{N} \]

\[ < H(N, p) \]

\[ < \sum_{\frac{1}{2} - \epsilon < p < \frac{1}{2} + \delta} N_k p^k (1-p)^{N-k} \frac{\log_2 N_{\lfloor N(p+\delta) \rfloor + 1}}{N} + \epsilon. \]

Together with equation (4), we get

\[ (1-\epsilon) \frac{\log_2 N_{\lfloor N(p-\delta) \rfloor-1}}{N} \]

\[ < H(N, p) \]

\[ < \frac{\log_2 N_{\lfloor N(p+\delta) \rfloor + 1}}{N} + \epsilon. \]

Using Stirling’s formula on both side

\[ (1-\epsilon) S(p-\delta) \leq \lim_{N \to \infty} H(N, p) \leq S(p+\delta) + \epsilon. \]

As \( \epsilon \) and \( \delta \) are arbitrary, with the continuity of \( S(p) \), we obtain

\[ \lim_{N \to \infty} H(N, p) = S(p). \]

This completes the proof of **Theorem 1.**
Theorem 2. $H(N, \frac{1}{2})$ increases to 1 as $N$ approaches infinity.

Proof. With the conclusion of Theorem 1, we have $H(N, \frac{1}{2})$ converging to 1 as $N$ approaches infinity. It therefore remains to be proven that $H(N, \frac{1}{2})$ is an increasing function.

First, we have

$$H(N, \frac{1}{2}) = \frac{1}{(N+1) \cdot 2^{N+1}} \sum_{k=1}^{N} (N+1)_{ \frac{k}{k+1} } \log_2(N+1)_{ \frac{k}{k+1} }$$

$$= \frac{1}{(N+1) \cdot 2^{N+1}} \sum_{k=1}^{N} N_k \log_2(N_k \frac{N+1}{N+1-k})$$

$$+ \sum_{k=1}^{N} N_{k-1} \log_2(N_{k-1} \frac{N+1}{k})$$

$$= \frac{1}{(N+1) \cdot 2^{N+1}} \left[ \sum_{k=0}^{N} N_k \log_2(N_k \frac{N+1}{k+1}) \right]$$

Thus, $H(N+1, \frac{1}{2}) \geq H(N, \frac{1}{2})$ is equivalent to

$$\sum_{k=0}^{N} N_k \log_2(N_k \frac{N+1}{k+1}) \geq \frac{N}{N+1} H(N, \frac{1}{2}) + \frac{1}{(N+1)^2 N+1}$$

$$\Leftrightarrow \sum_{k=0}^{N} N_k \log_2(N_k \frac{N+1}{k+1}) \geq 2 \sum_{k=0}^{N} N_k \log_2 N_k$$

$$\Leftrightarrow \sum_{k=0}^{N} N_k \log_2 \left[ \frac{(N+1)^2}{(N+1-k)(k+1)} \right] \geq \frac{k(k+1)}{k+1} (\frac{2k}{k+1})^{(k-1)}.$$