ON THE CORRECTNESS OF IMPLEMENTING RANDOM PERMUTATION AS SORTING RANDOM KEYS

XIANG GAO

Abstract. This documentation studies how many bits are required in random keys in order to implement random permutation of N numbers. Fewer bits means faster radix sort but poorer randomness. This documentation shows that, in order to generate a random permutaion of N numbers with correctness probability $\geq q$, the random numbers are required to have at least $\left\lceil \log_2 \left(N - \frac{6N^2+1}{12 \cdot \log q} \right) \right\rceil$ bits.

1. Introduction

A common way to implement random permutation of N numbers on GPU is by generating N random numbers as keys then sorting. This implementation is correct only when these N random numbers are different, because stable sort algorithms will not permute two identical keys, making $(0, 1)$ more preferred than $(1, 0)$. If these random keys had infinite precision, then duplicate keys was not a problem, because the probability of this case was 0. However, for finite precision random numbers, this is not the case.

Though facing correctness issue, for better performance, these N random numbers are usually generated independently and there is no effort to guarantee they are different with each other. We want to study the probability of getting duplicate keys. From this probability, we can know how many bits are required in random keys in order to get a good enough randomness.

2. The probability of getting non-duplicate keys

An m bit random number has $M = 2^m$ different values. Getting N independent random numbers from an m bit random generator is equivalent to drawing N samples from a size M pool with replacement, with uniform distribution, and order matters. Therefore, there are in total M^N different cases. Within these cases, $P(M, N)$ of them don't have duplicate samples, where the $P(M, N)$ is often written as $n\mathrm{P}r$, is defined by

$$P(M, N) \coloneqq \frac{M!}{(M-N)!}.$$

So the probability of getting non-duplicating keys is

$$\Pr[\text{unique samples}] = \frac{P(M, N)}{M^N}$$

According to a narrowed version of Stirling's formula by Robbins[1]

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n}}$$

we then have

$$\frac{M^{M+\frac{1}{2}}e^{-M}e^{\frac{1}{12M+1}}}{(M-N)^{(M-N)+\frac{1}{2}}\,e^{-(M-N)}e^{\frac{1}{12(M-N)}}} \leq P\left(M,N\right) \leq \frac{M^{M+\frac{1}{2}}e^{-M}e^{\frac{1}{12M}}}{(M-N)^{(M-N)+\frac{1}{2}}\,e^{-(M-N)}e^{\frac{1}{12(M-N)+1}}}$$

which simplify to

$$\frac{M^{M+\frac{1}{2}}\cdot e^{-N}}{(M-N)^{(M-N)+\frac{1}{2}}}\cdot e^{\frac{1}{12M+1}-\frac{1}{12(M-N)}} \leq P\left(M,N\right) \leq \frac{M^{M+\frac{1}{2}}\cdot e^{-N}}{(M-N)^{(M-N)+\frac{1}{2}}}\cdot e^{\frac{1}{12M}-\frac{1}{12(M-N)+1}}$$

so

$$\frac{M^{(M-N)+\frac{1}{2}}\cdot e^{-N}}{(M-N)^{(M-N)+\frac{1}{2}}}\cdot e^{\frac{1}{12M+1}-\frac{1}{12(M-N)}} \leq$$

$$\Pr\left[\text{unique samples}\right] \leq \frac{M^{(M-N)+\frac{1}{2}}}{(M-N)^{(M-N)+\frac{1}{2}}\,e^{N}}\cdot e^{\frac{1}{12M}-\frac{1}{12(M-N)+1}}$$

which simplify to

$$\left(\frac{M}{M-N}\right)^{(M-N)+\frac{1}{2}}\cdot e^{-N}\cdot e^{\frac{1}{12M+1}-\frac{1}{12(M-N)}} \leq$$

$$\Pr\left[\text{unique samples}\right] \leq \left(\frac{M}{M-N}\right)^{(M-N)+\frac{1}{2}}\cdot e^{-N}\cdot e^{\frac{1}{12M}-\frac{1}{12(M-N)+1}}$$

Let's write the above inequality as

$$f\left(M,N\right) \leq \Pr\left[\text{unique samples}\right] \leq g\left(M,N\right)$$

3. Bound

We want to make sure the probability of getting non-duplicate samples above a certain threshold:

$$\Pr\left[\text{unique samples}\right] \geq q \tag{3.1}$$

As long as $f\left(M,N\right) \geq q$, equation (3.1) will be automatically satisfied. Note that

$$f\left(M,N\right) = \left(\frac{M}{M-N}\right)^{(M-N)+\frac{1}{2}}\cdot e^{-N}\cdot e^{\frac{1}{12M+1}-\frac{1}{12(M-N)}}$$

$$= \left(1+\frac{N}{M-N}\right)^{(M-N)+\frac{1}{2}}\cdot e^{-N}\cdot e^{\frac{1}{12M+1}-\frac{1}{12(M-N)}}$$

then

$$\log f\left(M,N\right) = \left(M-N+\frac{1}{2}\right)\log\left(1+\frac{N}{M-N}\right) - N + \frac{1}{12M+1} - \frac{1}{12\left(M-N\right)}$$

since

$$\log\left(1+\frac{N}{M-N}\right) \geq \frac{N}{M-N} - \frac{1}{2}\left(\frac{N}{M-N}\right)^{2}$$

$$M-N+\frac{1}{2} \geq M-N$$

$$\frac{1}{12M+1} \geq 0$$

we have

$$\log f(M,N) \geq (M-N)\left[\frac{N}{M-N} - \frac{1}{2}\left(\frac{N}{M-N}\right)^2\right] - N - \frac{1}{12(M-N)}$$
$$= -\frac{1}{12}\cdot\frac{6N^2+1}{M-N}$$

So, as long as

$$-\frac{1}{12}\cdot\frac{6N^2+1}{M-N} \geq \log q$$

that is

$$M \geq N - \frac{6N^2+1}{12\cdot\log q}$$

there is a guarantee that the probability of getting unique keys is above q.

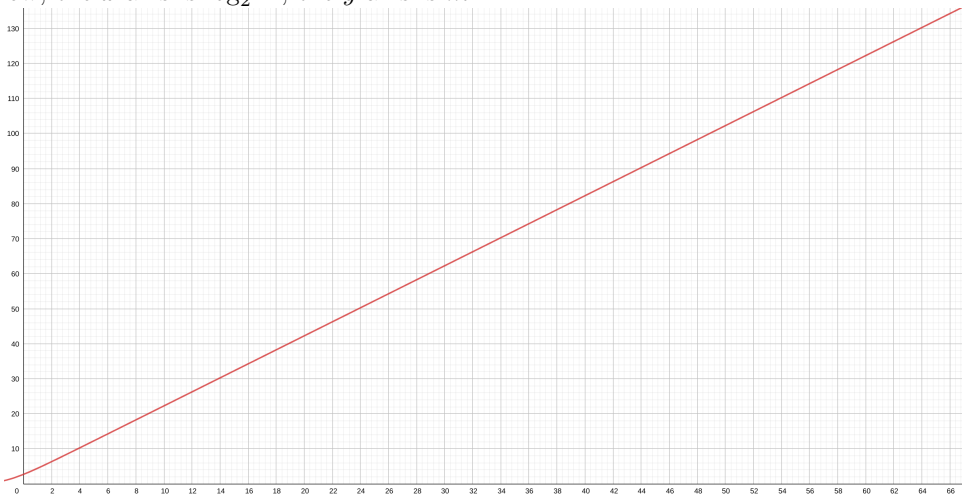
4. Summary

So, in order to have N different random numbers at probability $\geq q$, you will need an

$$m = \left\lceil \log_2 \left(N - \frac{6N^2 + 1}{12 \cdot \log q} \right) \right\rceil$$

bit random number generator.

Plotting the above equation for q= 0.9. The plot of the above function is shown below, the x-axis is $\log_2 N$, the y-axis is m



References

[1] Herbert Robbins. A remark on stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955.