The Adaptive sampling revisited

Guy Louchard* Yvik Swan †

December 14, 2017

Abstract

The problem of estimating the number n of distinct keys of a large collection of N data is well known in computer science. A classical algorithm is the adaptive sampling (AS). n can be estimated by $R2^J$, where R is the final bucket size and J is the final depth at the end of the process. Several new interesting questions can be asked about AS (some of them were suggested by P.Flajolet and popularized by J.Lumbroso). The distribution of $W = \log(R2^J/n)$ is known, we rederive this distribution in a simpler way. We provide new results on the moments of J and W. We also analyze the final cache size R distribution. We consider colored keys: assume also that among the n distinct keys, m do have color K We show how to estimate $p = \frac{m}{n}$. We study keys with some multiplicity : we provide a way to estimate the *total* number M of some color K keys among the *total* number N of keys. We consider the case where we know a priori the multiplicities but not the colors. There we want to estimate the total number of keys N. An appendix is devoted to the case where the hashing function provides bits with probability different from 1/2.

Keywords: Adaptive sampling, moments, periodic components, hashing functions, cache, colored keys, key multiplicity, Stein method, urn model, asymmetric adaptive sampling

2010 Mathematics Subject Classification: 68R05, 68W40.

1 Introduction

The problem of estimating the number n of distinct keys of a large collection of N data is well known in computer science. It arises in query optimization of data base systems. A classical algorithm is the adaptive sampling (AS) . The mean and variance of AS are considered in Flajolet [3] . Let us summarize the principal features of AS. Elements of the given set of N data are hashed into binary keys. These keys are infinitely long bit streams such that each bit has probability 1/2 of being 0 or 1. A uniformity assumption is made on the hashing function .

The algorithm keeps a bucket (or cache) B of at most b distinct keys. The depth of sampling, j which is defined below , is also saved. We start with $j = 0$ and throw only distinct keys into B . When B is full, depth j is increased by 1, the bucket is scanned, and only keys starting with 0 are kept. The scanning on the set is resumed and only distinct keys starting with 0 are considered. More generally, at step j , only distinct keys starting with 0^j are taken into account. When we have exhausted the set of N data, n can be estimated by $R2^J$, where R is the final bucket size and J is the final depth at the end of the process.

AS has some advantages in terms of processing time and of conceptual simplicity. As shown in [3], AS outperforms standard sorting methods by a factor of about 8. In terms of storage consumptions, using 100 words of memory will provide for a typical accuracy of 12%. This is to be contrasted again with sorting, where the auxiliary memory required has to be at least as large as the file itself. Finally

*Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email: louchard@ulb.ac.be

†Université de Liège, Département de Mathématique, zone polytech 1, 12, allée de la découverte, Bât. B37 pkg 33a, B-4000 Liège, Belgium, email: yswan@ulg.ac.be

AS is an unbiased estimator of cardinalities of large files that necessitates minimal auxiliary storage and processes data in a single pass.

Several new interesting questions can be asked about AS (some of them were suggested by P.Flajolet and popularized by J. Lumbroso). The distribution of $W = \log(R2^J/n)$ is known (see [7]), but in Sec.3, we rederive this distribution in a simpler way. In Sec.4 we provide new results on the moments of J and W. The final cache size R distribution is analyzed in Sec.5. Colored keys are considered in Sec.6: assume that we have a set of colors and that each key has some color. Assume also that among the n distinct keys, m do have color K and that m is large such that $\frac{m}{n} = p = \Theta(1)$. We show how to estimate p. We consider keys with some multiplicity in Sec.7: assume that, to each key κ_i, we attach a counter giving its *observed* multiplicity μ_i. Also we assume that the multiplicities of color K keys are given by iid random variables (RV), with distribution function F, mean μ, variance σ^2 (functions of K). We show how to estimate the *total* number M of color K keys among the *total* number N of keys. Sec.8 deals with the case where we know a priori the multiplicities but not the colors. We want to estimate the total number of keys N. An appendix is devoted to the case where the hashing function provides bits with probability different from 1/2.

2 Preliminaries.

Let us first give the notations we will use throughout the paper.

$$
\begin{aligned}
N &:= \text{ total number of keys, } N \text{ large , key } \kappa_i \text{ appearing with multiplicity } \mu_i \text{ say,}\\
n &:= \text{number of } \textit{distinct} \text{ keys, } n \text{ large,}\\
\sim &:= \text{ asymptotic to, for large } n,\\
b &:= \text{ cache size , } b \text{ fixed, independent of } n,\\
\overset{b}{\sim} &:= \text{ asymptotic to, for large } n \text{ and } b,\\
R &:= \text{ number of keys in the cache, at the end of the process,}\\
J &:= \text{ depth of the cache, at the end of the process,}\\
Z &:= \frac{R2^J}{n},\\
W &:= \log(Z),\\
\mathcal{P}(\lambda, u) &:= e^{-\lambda}\lambda^u/u!, \text{ (Poisson distribution),}\\
\log &:= \log_2,\\
\eta &:= j - \log n,\\
L &:= \ln 2,\\
\tilde{\alpha} &:= \alpha/L,\\
\{x\} &:= \text{ fractional part of } x,\\
\chi_l &:= \frac{2l\pi\mathbf{i}}{L},\\
p &:= \text{ In Section 6: parameter related to the color } K, q := 1-p,\\
p &:= \text{ In the Appendix: parameter related to the probability of bit 1 in the Asymmetric Adaptive Sampling,}\\
\mathbb{V}(X) &:= \text{ Variance of random variable } X,\\
w &:= \text{ (small) periodic function of } \log n.
\end{aligned}
$$

From Flajolet [3], we have the exact distribution

$$p(r,j) := \mathbb{P}(R=r, J=j) = \binom{n}{r}\Big(\frac{1}{2^j}\Big)^r\Big(1-\frac{1}{2^j}\Big)^{n-r}\Big[1-\sum_{k=0}^{b-r}\binom{n-r}{k}\Big(\frac{1}{2^j}\Big)^k\Big(1-\frac{1}{2^j}\Big)^{n-r-k}\Big], \quad (1)$$

$$p(j) := \mathbb{P}(J = j) = \sum_{r=0}^{b} p(r, j),$$
$$p_r := \mathbb{P}(R = r) = \sum_{j} p(r, j),$$
$$P(r, j) := \mathbb{P}(R = r, J \leq j).$$

We can now see Adaptive Sampling as an urn model, where balls (keys), are thrown into urn $J = j$ with probability $1/2^j$. We recall the main properties of such a model.

- Asymptotic independence. We have asymptotic independence of urns, for all events related to urn j containing $\mathcal{O}(1)$ balls. This is proved, by Poissonization-De-Poissonization, in [9], [10] and [5]. The error term is $\mathcal{O}(n^{-C})$ where C is a positive constant.

- Asymptotic distributions. We obtain asymptotic distributions of the interesting random variables as follows. The number of balls in each urn is asymptotically Poisson-distributed with parameter $n/2^j$ in urn j containing $\mathcal{O}(1)$ balls (this is the classical asymptotic for the Binomial distribution). This means that the asymptotic number ℓ of balls in urn j is given by

$$\exp\left(-n/2^j\right) \frac{\left(n/2^j\right)^\ell}{\ell!},$$

  and with $\eta = j - \log n$, this is equivalent to $\mathcal{P}\left(e^{-L\eta}, \ell\right)$. The asymptotic distributions are related to Gumbel distribution functions (given by $\exp\left(-e^{-x}\right)$) or convergent series of such. The error term is $\mathcal{O}(n^{-1})$.

- Extended summations. Some summations now go to ∞. This is justified, for example, in [9].

- Uniform Integrability. We have uniform integrability for the moments of our random variables. To show that the limiting moments are equivalent to the moments of the limiting distributions, we need a suitable rate of convergence. This is related to a uniform integrability condition (see Loève [6, Section 11.4]). For Adaptive Sampling, the rate of convergence is analyzed in detail in [8]. The error term is $\mathcal{O}(n^{-C})$.

- Mellin transform. Asymptotic expressions for the moments are obtained by Mellin transforms (for a good reference to Mellin transforms, see Flajolet et al. [4]). The error term is $\mathcal{O}(n^{-C})$. We proceed as follows (see [8] for detailed proofs): from the asymptotic properties of the urns, we have obtained the asymptotic distributions of our random variables of interest. Next we compute the Laplace transform $\phi(\alpha)$ of these distributions, from which we can derive the dominant part of probabilities and moments as well as the (tiny) periodic part in the form of a Fourier series. This connection will be detailed in the next sections.

- Fast decrease property. $\Gamma(s)$ decreases exponentially in the direction $i\infty$:

$$|\Gamma(\sigma + \mathbf{i}t)| \sim \sqrt{2\pi}|t|^{\sigma-1/2}e^{-\pi|t|/2}.$$

  Also, we this property is true for all other functions we encounter. So inverting the Mellin transforms is easily justified.

- Early approximations. If we compare the approach in this paper with other ones that appeared previously, then we can notice the following. Traditionally, one would stay with exact enumerations as long as possible, and only at a late stage move to asymptotics. Doing this, one would, in terms of asymptotics, carry many unimportant contributions around, which makes the computations quite heavy, especially when it comes to higher moments. Here, however, approximations are carried out as early as possible, and this allows for streamlined (and often automatic) computations of the higher moments.

We set $\eta = j - \log n$, (1) leads to

$$p(r,j) \sim f(r,\eta) = \exp(-2^{-\eta})\frac{2^{-r\eta}}{r!}\left[1 - \exp(-2^{-\eta})\sum_{k=0}^{b-r}\frac{2^{-k\eta}}{k!}\right], \tag{2}$$

and similar functions for $P(r,j)$. Asymptotically, the distribution will be a periodic function of the fractional part of $\log n$. The distribution $P(r,j)$ does not converge in the weak sense, it does however converge along subsequences n_m for which the fractional part of $\log n_m$ is constant. This type of convergence is not uncommon in the Analysis of Algorithms. Many examples are given in [8].

From (2), we compute the Laplace transform

$$\phi(r,\alpha) = \int_{-\infty}^{\infty} e^{\alpha\eta} f(r,\eta)d\eta = \frac{\Gamma(r-\tilde{\alpha})}{Lr!} - \sum_{k=0}^{b-r}\frac{\Gamma(r+k-\tilde{\alpha})}{Lr!k!2^{r+k-\tilde{\alpha}}}.$$

The moments of Z are already given in [7] and [8]. As shown in [7], we must have $d \le b$. For the sake of completeness, we repeat them here:

$$\begin{aligned}
&\mathbb{E}[Z^d] \sim m_{1,d} + w_{1,d},\\
&m_{1,d} = 1 + \frac{(b-d)!}{L}\sum_{k=1}^{d-1}\left\{{d \atop k}\right\}\frac{2^{d-k}-1}{(d-k)(b-k)!},\\
&w_{1,d} = \sum_{l\neq 0}\frac{1}{L}\sum_{j=1}^{d-1}\left\{{d \atop j}\right\}\left[(1-2^{d-j}\right]\Gamma(j-d+\chi_l)\binom{b-d+\chi_l}{b-j}e^{-2l\pi\mathbf{i}\log n},\\
&m_{1,1} = 1, w_{1,1} = 0, m_{1,2} = 1 + \frac{1}{(b-1)L}\\
&\mathbb{V}(Z) \sim \frac{1}{(b-1)L}.
\end{aligned}$$

Note that, in [3], Flajolet already computed $m_{1,1}, m_{1,2}, w_{1,2}, w_{1,2}$.

3 Asymptotic distribution of $W = J - \log n + \log R$

Let us recover this distribution from (2). In the sequel, we will denote by $\mathbb{E}(A; R>0)$ the expectation of event A related to positive R. We have

Theorem 3.1

$$\mathbb{P}(W \le \alpha; R>0) \sim \sum_{r=1}^{b}\sum_{l\ge 0}\exp(-2^{-\varphi})\frac{2^{-r\varphi}}{r!}\left[1 - \exp(-2^{-\varphi})\sum_{k=0}^{b-r}\frac{2^{-k\varphi}}{k!}\right],$$

with

$$\varphi := \lfloor\{\log n\} - \log r + \alpha\rfloor - \{\log n\} - \ell.$$

Proof.

$$\begin{aligned}
\mathbb{P}(W \le \alpha; R>0) &= \mathbb{P}[J \le \log n - \log R + \alpha; R>0]\\
&= \mathbb{P}[J \le \lfloor\log n\rfloor + \lfloor\{\log n\} - \log R + \alpha\rfloor; R>0]\\
&\sim \sum_{r=1}^{b}\sum_{\ell\ge 0}\exp(-2^{-(\eta-\ell)})\frac{2^{-r(\eta-\ell)}}{r!}\left[1 - \exp(-2^{-(\eta-\ell)})\sum_{k=0}^{b-r}\frac{2^{-k(\eta-\ell)}}{k!}\right],
\end{aligned}$$

with

$$\eta = \lfloor\{\log n\} - \log r + \alpha\rfloor - \{\log n\},$$

or

$$\mathbb{P}(W \le \alpha; R > 0) \sim \sum_{r=1}^{b} \sum_{l \ge 0} \exp(-2^{-\varphi}) \frac{2^{-r\varphi}}{r!} \left[1 - \exp(-2^{-\varphi}) \sum_{k=0}^{b-r} \frac{2^{-k\varphi}}{k!} \right],$$

with

$$\varphi := \lfloor \{\log n\} - \log r + \alpha \rfloor - \{\log n\} - \ell.$$

This is exactly [7, Thm 4.1] that we obtain here in a simpler way. ■

4 Moments of $J - \log n$ and W

Two interesting parameters are given by the moments of $J - \log n$ and W. Their asymptotic behaviour is given as follows

Theorem 4.1

$$\mathbb{E}[(J - \log n)^k; R = r] \sim \tilde{m}_{k,r} + w_{k,r},$$

where

$$\tilde{m}_{1,r} = -\frac{\psi(r)}{L^2 r} + \sum_{k=0}^{b-r} \frac{(\psi(r+k) - L) 2^{-(r+k)} \Gamma(r+k)}{L^2 \Gamma(r+1)\Gamma(k+1)}, \quad r > 0,$$

$$\tilde{m}_{1,0} = \frac{1}{2} + \frac{\gamma}{L} + \sum_{k=1}^{b} \frac{(\psi(k) - L) 2^{-k}}{k L^2},$$

$$w_{1,r} = \sum_{l \ne 0} \left[-\frac{\psi(r+\chi_l)\Gamma(r+\chi_l)}{L^2 \Gamma(r+1)} + \sum_{k=0}^{b-r} \frac{(\psi(r+k+\chi_l) - L) 2^{-(r+k)} \Gamma(r+k+\chi_l)}{L^2 \Gamma(r+1)\Gamma(k+1)} \right] e^{-2l\pi \mathbf{i} \log n}, \quad r > 0,$$

$$w_{1,0} = \sum_{l \ne 0} \left[-\frac{\psi(\chi_l)\Gamma(\chi_l)}{L^2} + \sum_{k=0}^{b} \frac{\Gamma(k+\chi_l)}{L^2 k! 2^k} (\psi(k+\chi_l) - L) \right].$$

$$\mathbb{E}(W; R > 0) \sim \sum_{r=1}^{b} \tilde{m}_{1,r} + \sum_{r=1}^{b} p_r \log r + \sum_{r=1}^{b} w_{1,r},$$

$$\mathbb{E}(W^2; R > 0) \sim \sum_{r=1}^{b} \tilde{m}_{2,r} + 2 \sum_{r=1}^{b} \tilde{m}_{1,r} \log r + \sum_{r=1}^{b} p_r (\log r)^2 + \sum_{r=1}^{b} w_{2,r} + 2 \sum_{r=1}^{b} w_{1,r} \log r.$$

Proof. Using the techniques developed in [8], we obtain the dominant (constant) part of the moments of J as follows:

$$\mathbb{E}[(J - \log n)^k; R = r] \sim \tilde{m}_{k,r} + w_{k,r},$$

where the non-periodic component is given by

$$\tilde{m}_{k,r} := \phi^{(k)}(r, 0),$$

and the corresponding periodic term is given by

$$w_{k,r} = \sum_{l \ne 0} \phi^{(k)}(r, \alpha)\Big|_{\alpha = -L\chi_l} e^{-2l\pi \mathbf{i} \log n}.$$

This was already computed in [8], but with some errors. The first corrected values are now provided.

As $W = J - \log n + \log R$, the rest of the Thm is immediate ■

It will be useful to obtain an asymptotic for the expectation of $J - \log n$ (non-periodic component) for large b. This is computed as follows. First of all, we rewrite $\sum_{r=1}^{b} \tilde{m}_{1,r}$ as

$$\sum_{r=1}^{b} \tilde{m}_{1,r} = -\sum_{r=1}^{b} \frac{\psi(r)}{L^2 r} + \sum_{u=1}^{b} \left[\sum_{r=1}^{u} \frac{1}{\Gamma(r+1)\Gamma(u-r+1)} \right] \frac{(\psi(u) - L)2^{-u}\Gamma(u)}{L^2}.$$

Now it is clear that the main contribution of the second term is related to large u. So we set $r = \frac{u}{2} + v$. This gives, by Stirling,

$$\Gamma(r+1) \sim e^{-(u/2+v)} e^{v+v^2/u} \left(\frac{u}{2}\right)^{u/2+v} \sqrt{\pi u},$$

and

$$\Gamma(r+1)\Gamma(u-r+1) \sim e^{-u} e^{2v^2/u} \left(\frac{u}{2}\right)^{u} \pi u.$$

By Euler-Maclaurin, we have

$$\sum_{r=1}^{u} \frac{1}{\Gamma(r+1)\Gamma(u-r+1)} \sim 2 \sum_{v=0}^{u/2} \frac{e^u}{\left(\frac{u}{2}\right)^u \pi u} e^{-2v^2/u} \sim \int_0^{\infty} 2e^{-2v^2/u} dv \frac{e^u}{\left(\frac{u}{2}\right)^u \pi u} = \frac{e^u}{\left(\frac{u}{2}\right)^u \sqrt{2\pi u}},$$

and, finally,

$$\sum_{r=0}^{b} \tilde{m}_{1,r} \overset{b}{\sim} \frac{1}{2} + \frac{\gamma}{L} + \sum_{u=1}^{b} \left[-\frac{\psi(u)}{L^2 u} + \frac{(\psi(u) - L)2^{-u}}{uL^2} + \frac{(\psi(u) - L)2^{-u}\Gamma(u)}{L^2} \frac{e^u}{\left(\frac{u}{2}\right)^u \sqrt{2\pi u}} \right],$$

and, to first order,

$$\mathbb{E}(J - \log n) \sim \sum_{r=0}^{b} \tilde{m}_{1,r} \overset{b}{\sim} \frac{1}{2} + \frac{\gamma}{L} - \sum_{u=1}^{b} \frac{1}{Lu} \sim -\log b + \mathcal{O}(1) \tag{3}$$

5 Distribution of R

The asymptotic moments and distribution of R are given as follows

Theorem 5.1 *The non-periodic components are given by*

$$\begin{aligned}
\mathbb{E}(R) &\sim \frac{b}{2L}, \\
\mathbb{E}(R^2) &\sim \frac{b(3b+1)}{8L}, \\
\mathbb{V}(R) &\sim \frac{b(3Lb - 2b + L)}{8L^2}, \\
\mathbb{P}(R = r) \sim p_r &= \frac{1}{L} \left[\frac{1}{r} - \sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}} \right], r \geq 1,
\end{aligned}$$

Similarly, the periodic components are given by

$$\begin{aligned}
w_0(R) &= \sum_{r=0}^{b} \sum_{l \neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n}, \\
w_1(R) &= \sum_{r=0}^{b} r \sum_{l \neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n},
\end{aligned}$$

$$w_2(R) = \sum_{r=0}^{b} r^2 \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n},$$
$$w_0(r) = \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n}$$

Proof. We have

$$\mathbb{P}(R = r) \sim p_r = \phi(r, 0) = \frac{1}{L}\left[\frac{1}{r} - \sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}}\right], r \geq 1,$$

and $p_0 = 1 - \sum_1^b p_r$, with

$$\begin{aligned}\sum_1^b p_r &= \frac{1}{L}\left[H_b - \sum_{r=1}^{b}\sum_{k=0}^{b-r} \frac{\Gamma(r+k)}{r!k!2^{r+k}}\right] \\ &= \frac{1}{L}\left[H_b - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{1}{r!(u-r)!}\right].\end{aligned}$$

This quantity was already obtained in [7] after some complicated algebra! This leads to

$$p_0 = 1 - \sum_{u=1}^{b} \frac{1}{u2^u L},$$

which is also the probability of $Z = 0$. This is also easily obtained from $\lim_{r\to 0} \phi(r, 0)$. Figure 1 gives p_r for $b = 50$

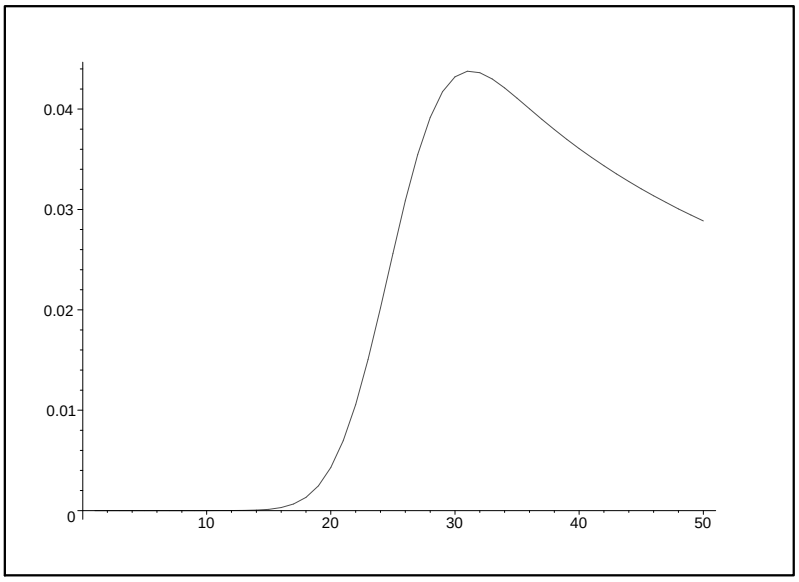

Figure 1: p_r for $b = 50$

Conditioning on $R > 0$, the expectation of event A is now given by

$$\frac{\mathbb{E}(A; R > 0)}{1 - p_0}.$$

Also

$$w_0(r) = \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n}.$$

The moments of R are computed as follows.

$$\begin{aligned}
\mathbb{E}(R) \sim \sum_{r=1}^{b} r p_r &= \frac{1}{L}\left[b - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{r}{r!(u-r)!}\right] \\
&= \frac{1}{L}\left[b - \sum_{u=1}^{b} \frac{1}{2^u}[2^{u-1}]\right] \\
&= \frac{b}{2L}.
\end{aligned}$$

More generally, the generating function of p_r is given by

$$\begin{aligned}
\sum_{r=1}^{b} z^r p_r &= \frac{1}{L}\left[\sum_{r=1}^{b} \frac{z^r}{r} - \sum_{u=1}^{b} \frac{(u-1)!}{2^u} \sum_{r=1}^{u} \frac{z^r}{r!(u-r)!}\right] \\
&= \frac{1}{L}\left[\sum_{r=1}^{b} \frac{z^r}{r} - \sum_{u=1}^{b} \frac{1}{u2^u}[(1+z)^u - 1]\right].
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathbb{E}(R^2) \sim \sum_{r=1}^{b} r^2 p_r &= \frac{b(3b+1)}{8L}, \\
\mathbb{V}(R) &\sim \frac{b(3Lb - 2b + L)}{8L^2}.
\end{aligned}$$

Similarly, the periodic components are given by

$$\begin{aligned}
w_1(R) &= \sum_{r=0}^{b} r \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n}, \\
w_2(R) &= \sum_{r=0}^{b} r^2 \sum_{l\neq 0} \phi(r, -L\chi_l) e^{-2l\pi \mathbf{i} \log n}.
\end{aligned}$$

■

6 Colors

Assume that we have a set of colors and that each key has some color. Assume also that among the n distinct keys, m do have color K and that m is large such that $\frac{m}{n} = p = \Theta(1), q := 1 - p$. In the cache, the R keys (we assume $R > 0$) contain U keys with color K with probability distribution

$$\mathbb{P}(U = u | R = r) = \frac{\binom{m}{u}\binom{m-m}{r-u}}{\binom{n}{r}},$$

and, if $r = o(n)$, this is asymptotically given by the conditioned binomial distribution $Bin(r, p)$. We want to estimate p. We are interested in the distribution of the statistic $\tilde{p} = U/R$. We have

Theorem 6.1

$$\begin{aligned}\mathbb{E}\left(\frac{U}{R};R>0\right)&\sim p,\\ \mathbb{E}\left(\left(\frac{U}{R}\right)^2;R>0\right)&\sim p^2+pq\mathbb{E}\left(\frac{1}{R};R>0\right),\\ \mathbb{V}\left(\frac{U}{R};R>0\right)&\sim pq\mathbb{E}\left(\frac{1}{R};R>0\right).\end{aligned}\tag{4}$$

Proof. We have

$$\mathbb{P}\left[\left(\frac{U}{R}\right)\leq\alpha;R>0\right]=\mathbb{P}(U\leq\alpha R;R>0]\sim\sum_{r=1}^{b}p_r\sum_{u=0}^{\lfloor\alpha r\rfloor}\binom{r}{u}p^uq^{r-u}.\tag{5}$$

Now, conditioned on R, we have

$$\mathbb{E}(U|R)\sim Rp,\mathbb{E}(U^2|R)\sim Rpq+R^2p^2.$$

So, conditioned on R,

$$\begin{aligned}\mathbb{E}\left(\frac{U}{R}\right)&\sim p,\\ \mathbb{E}\left(\left(\frac{U}{R}\right)^2\right)&\sim p^2+\frac{pq}{R},\end{aligned}$$

and, unconditionning leads to the theorem. ■

Intuitively, if the cache size b is large, we should have an asymptotic Gaussian distribution for U/R. Actually, the fit is quite good, even for $b=30$ (and $p=0.2$).

This is proved as follows.

6.1 The distribution of U/R for large b .

Let R be a (possibly degenerate) random variable taking values on the (strict) positive integers. Conditionning on R, let $U\sim Bin(R,p)$ for some known $0<p<1$, and set $Y=U/R$. It appears that, as R grows large, the distribution of Y becomes asymptotically Gaussian. This claim can be made precise as follows.

Theorem 6.2 *Let $V\sim\mathcal{N}(0,1)$ and write $\Psi=\sqrt{R}\frac{Y-p}{\sqrt{pq}}$. Then there exists an absolute constant $\kappa\in\mathbb{R}$ such that*

$$d_{\mathcal{W}}(\Psi,V)\leq\kappa\mathbb{E}\left\{\frac{1}{\sqrt{R}}\right\}$$

for $d_{\mathcal{W}}(\Psi,V)$ the Wasserstein distance between the law of Ψ and that of V; moreover this constant is such that

$$\kappa\leq\frac{q^2-p^2}{\sqrt{pq}}+4\left[\frac{p^3+q^3}{pq}-1\right]^{1/2}.$$

Proof. We will prove this theorem using the Stein methodology which, for $h\in\mathcal{H}$, ($\mathcal{H}$ is a nice class of test functions), suggests to write

$$\mathbb{E}h(\Psi)-\mathbb{E}h(V)=\mathbb{E}\left(\Psi f(\Psi)-f'(\Psi)\right)$$

with $f := f_h$ such that

$$xf(x) - f'(x) = h(x) - \mathbb{E}h(V). \tag{6}$$

(this is known as the *Stein equation* for the Gaussian distribution) so that

$$d_{\mathcal{W}}(\Psi, V) = \sup_{h\in\mathcal{H}} \mathbb{E}\left|h(\Psi) - h(V)\right| \leq \sup_{f_h} \left|\mathbb{E}\left(\Psi f(\Psi) - f'(\Psi)\right)\right|. \tag{7}$$

The reason why (7) is interesting is that properties of the solutions f of (6) are well-known – see, e.g., [1, Lemma 2.3] and [2, Lemma 2.3] – and quite good so that they can be used with quite some efficiency to tackle the rhs of (7). In the present configuration we know that f_h is continuous and bounded on $\mathbb{R}$, with

$$\|f_h'\| \leq \min\left(2\|h - \mathbb{E}h(V)\|, 4\|h'\|\right) \tag{8}$$

and

$$\|f_h''\| \leq 2\|h'\|. \tag{9}$$

In particular, if H is the class of Lipschitz-1 functions with $\|h'\| \leq 1$ (this class generates the Wasserstein distance) then

$$\|f_h'\| \leq 4 \text{ and } \|f_h''\| \leq 2.$$

These will suffice to our purpose.

Our proof follows closely the standard one for independent summands (see, e.g., [11, Section 3]). First we remark that, given $R \geq 1$, we can write Ψ as

$$\Psi = \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i$$

where, taking X_i i.i.d. $Bin(1, p)$, we let $\xi_i = (X_i - p)/\sqrt{pq}$ (which are centered and have variance 1). Next, for $r \geq 1$ and $1 \leq i \leq r$, define

$$\Psi_i^r = \Psi - \frac{1}{\sqrt{r}}\xi_i = \frac{1}{\sqrt{r}} \sum_{j\neq i} \xi_j.$$

Next take f solution of (6) with h some Lipschitz-1 function. Then note that $\mathbb{E}\left\{\xi_i f\left(\Psi_i^r\right)\right\} = 0$ for all $1 \leq i \leq r$. We abuse notations and, given R, write $\Psi_i^R = \Psi_i$. Then

$$\begin{aligned}
\mathbb{E}\left\{\Psi f(\Psi) \,|\, R\right\} &= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i f(\Psi) \,\middle|\, R \right\} \\
&= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i \left(f(\Psi) - f(\Psi_i)\right) \,\middle|\, R \right\} \\
&= \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i \left(f(\Psi) - f(\Psi_i) - (\Psi - \Psi_i) f'(\Psi)\right) \,\middle|\, R \right\} \\
&\quad + \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i (\Psi - \Psi_i) f'(\Psi) \,\middle|\, R \right\}
\end{aligned}$$

so that

$$\begin{aligned}
\left|\mathbb{E}\left\{\Psi f(\Psi) - f'(\Psi) \,|\, R\right\}\right| &\leq \mathbb{E}\left\{ \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \left|\xi_i \left(f(\Psi) - f(\Psi_i) - (\Psi - \Psi_i) f'(\Psi)\right)\right| \,\middle|\, R \right\} \\
&\quad + \left| \mathbb{E}\left\{ f'(\Psi) \left(1 - \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \xi_i (\Psi - \Psi_i)\right) \,\middle|\, R \right\} \right|
\end{aligned}$$

$$=: |\chi_1(R)| + |\chi_2(R)|.$$

Recall that $\Psi - \Psi_i = \frac{1}{\sqrt{R}}\xi_i$. Then (by Taylor expansion) we can easily deal with the first term to obtain

$$|\chi_1(R)| = \frac{\|f''\|}{2}\frac{1}{\sqrt{R}}\mathbb{E}\left|\xi_1\right|^3.$$

Taking expectations with respect to R and using (9) we conclude

$$E|\chi_1(R)| \leq \mathbb{E}\left|\xi_1\right|^3 E\left(\frac{1}{\sqrt{R}}\right). \tag{10}$$

For the second term note how

$$\begin{aligned}
|\chi_2(R)| &= \left|\mathbb{E}\left\{f'(\Psi)\left(1 - \frac{1}{R}\sum_{i=1}^{R}\xi_i^2\right) \mid R\right\}\right| \\
&= \left|\mathbb{E}\left\{\frac{f'(\Psi)}{R}\sum_{i=1}^{R}\left(1-\xi_i^2\right) \;\middle|\; R\right\}\right| \\
&\leq \frac{\|f'\|}{R}\mathbb{E}\left\{\left|\sum_{i=1}^{R}\left(1-\xi_i^2\right)\right| \;\middle|\; R\right\}.
\end{aligned}$$

Since $\mathbb{E}\left\{\sum_{i=1}^{R}\left(1-\xi_i^2\right) \mid R\right\} = 0$ we can pursue to obtain

$$\begin{aligned}
|\chi_2(R)| &\leq \frac{\|f'\|}{R}\sqrt{\mathbb{V}\left(\sum_{i=1}^{R}\left(1-\xi_i^2\right) \;\middle|\; R\right)} \\
&= \frac{\|f'\|}{\sqrt{R}}\sqrt{\mathbb{V}\left(\xi_1^2\right)}
\end{aligned}$$

where we used (conditional) independence of the ξ_i. Taking expectations with respect to R and using (8) we deduce (recall $\mathbb{V}\left(\xi_1^2\right) = \mathbb{E}\xi_1^4 - 1$)

$$\mathbb{E}|\chi_2(R)| \leq 4\sqrt{\mathbb{E}\xi_1^4 - 1}\,\mathbb{E}\left(\frac{1}{\sqrt{R}}\right). \tag{11}$$

Combining (10) and (11) we can conclude

$$d_{\mathcal{W}}(\Psi, V) \leq \left(\mathbb{E}|\xi_1^3| + 4\sqrt{\mathbb{E}\xi_1^4 - 1}\right)\mathbb{E}\left(\frac{1}{\sqrt{R}}\right).$$

The claim follows. ■

So we need the moments of $1/R$ for large b (we limit ourselves to the dominant term).

6.2 Moments of $1/R, R > 0$ for large b

We have the following property

Theorem 6.3

$$\mathbb{E}\left(\frac{1}{R^\alpha}; R > 0\right) \overset{b}{\sim} \frac{1}{L\alpha b^\alpha}(2^\alpha - 1), \alpha > 0.$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{R};R>0\right) &\sim \sum_{r=1}^{b} p_r/r = \frac{1}{L}\left[\sum_{r=1}^{b}\frac{1}{r^2} - \sum_{u=1}^{b}\frac{(u-1)!}{2^u}\sum_{r=1}^{u}\frac{1}{rr!(u-r)!}\right]\\ &= \frac{1}{L}\left[H_b^{(2)} - \sum_{u=1}^{b}\frac{1}{2^u}{}_2F_2[1,-u+1;2,2;-1]\right]\\ &= \frac{1}{L}\left[-\psi(1,b+1)+\frac{\pi^2}{6} - \sum_{u=1}^{\infty}\frac{1}{2^u}{}_2F_2[[1,-u+1],[2,2],-1]\right]\\ &+ \frac{1}{L}\sum_{u=b+1}^{\infty}\frac{1}{2^u}{}_2F_2[[,1-u+1],[2,2],-1].\end{aligned}$$

But[1]

$$\begin{aligned}&\sum_{u=1}^{\infty}\frac{1}{2^u}{}_2F_2[[1,-u+1],[2,2],-1]\\ &= \sum_{r=1}^{\infty}\sum_{u=r}^{\infty}\frac{(u-1)!}{2^u rr!(u-r)!}\\ &= \sum_{r=1}^{\infty}\frac{1}{r^2}\sum_{u=r}^{\infty}\frac{(u-1)!}{2^u(r-1)!(u-r)!}\\ &= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{u=v+1}^{\infty}\frac{(u-1)!}{2^u v!(u-v-1)!}\\ &= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{w=v}^{\infty}\frac{w!}{2^{w+1}v!(w-v)!}\\ &= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{w=v}^{\infty}\binom{w}{v}2^{-w}\\ &= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\sum_{s=0}^{\infty}\binom{s+v}{v}2^{-(s+v)}\\ &= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}2^{-v}\sum_{s=0}^{\infty}\binom{-v-1}{s}(-2)^{-s}\\ &= \frac{1}{2}\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}2^{-v}\left(1-\frac{1}{2}\right)^{-(v+1)}\\ &= \sum_{v=0}^{\infty}\frac{1}{(v+1)^2}\\ &= \zeta(2) = \frac{\pi^2}{6}.\end{aligned} \tag{12}$$

Now

$$\psi(1,b+1) \overset{b}{\sim} \frac{1}{b} + \mathcal{O}\left(\frac{1}{b^2}\right),$$

and

$$\sum_{u=b+1}^{\infty}\frac{1}{2^u}{}_2F_2[[1,-u+1],[2,2],-1]$$

[1]We are indebted to H.Prodinger for this identity

$$
\begin{aligned}
&= T_1 + T_2, \\
T_1 &= \sum_{r=1}^{b+1} \sum_{u=b+1}^{\infty} \frac{(u-1)!}{2^u r!(u-r)!} \\
&= \frac{1}{2} \sum_{v=0}^{b} \frac{1}{(v+1)^2} \sum_{w=b}^{\infty} \binom{w}{v} 2^{-w}, \\
T_2 &= \sum_{r=b+1}^{\infty} \sum_{u=r}^{\infty} \frac{(u-1)!}{2^u r!(u-r)!} \\
&= \frac{1}{2} \sum_{v=b}^{\infty} \frac{1}{(v+1)^2} \sum_{w=v}^{\infty} \binom{w}{v} 2^{-w} \\
&= \frac{1}{2} \sum_{v=b}^{\infty} \frac{1}{(v+1)^2} 2.
\end{aligned}
$$

In order to compute T_1, we now turn to the asymptotics of $\binom{w}{v}2^{-w}$ for large w. We obtain, by Stirling and setting $w = 2v + \alpha$,

$$
\begin{aligned}
\binom{w}{v} 2^{-w} &\sim \frac{e^{-w} w^w \sqrt{2\pi w}}{e^{-(w-v)}(w-v)^{w-v}\sqrt{2\pi(w-v)} 2^w e^{-v} v^v \sqrt{2\pi v}} \\
&= \frac{e^{-(2v+\alpha)}(2v+\alpha)^{2v+\alpha}\sqrt{2\pi(2v+\alpha)}}{e^{-(v+\alpha)}(v+\alpha)^{v+\alpha}\sqrt{2\pi(v+\alpha)} 2^{2v+\alpha} e^{-v} v^v \sqrt{2\pi v}} \\
&\sim \frac{e^{-v}(2v)^{2v+\alpha}\left(1+\frac{\alpha}{2v}\right)^{2v+\alpha}\sqrt{2}}{v^{v+\alpha}\left(1+\frac{\alpha}{v}\right)^{v+\alpha} 2^{2v+\alpha} e^{-v} v^v \sqrt{2\pi v}} \\
&\sim \frac{e^{\alpha+\frac{\alpha^2}{4v}}\sqrt{2}}{e^{\alpha+\frac{\alpha^2}{2v}}\sqrt{2\pi v}} \\
&\sim \frac{e^{-\frac{\alpha^2}{4v}}}{\sqrt{\pi v}} \\
&= 2\frac{e^{-\frac{\alpha^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}},
\end{aligned}
$$

with $\sigma^2 = 2v$. This is a Gaussian function, centered at $2v$ with variance $\sigma^2 = 2v$. So, by Euler-Maclaurin, replacing sums by integrals, we obtain

- if $b/2 < v \leq b$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} \overset{b}{\sim} 2,$$
- if $0 \leq v < b/2$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} \text{ is exponentially negligible ,}$$
- if $v \geq b$,
$$\sum_{w=b}^{\infty} \binom{w}{v} 2^{-w} = 2 \text{ by (12),}$$
but this will not be used in the sequel,

and finally

$$T_1 + T_2 \overset{b}{\sim} \frac{1}{2}\left[2\sum_{v=b/2}^{b}\frac{1}{(v+1)^2} + 2\sum_{v=b}^{\infty}\frac{1}{(v+1)^2}\right] \overset{b}{\sim} \frac{2}{b}.$$

This leads to

$$\mathbb{E}\left(\frac{1}{R}; R>0\right) \overset{b}{\sim} \frac{1}{L}\left[\frac{2}{b} - \frac{1}{b}\right] = \frac{1}{Lb}. \tag{13}$$

In the neighbourhood of $v = b/2$, only part of the Gaussian is integrated. But if we choose an interval $\Delta := [b/2 - b^{5/8}, b/2 + b^{5/8}]$, $(b^{5/8} \gg \sigma)$, this contributes to

$$\mathcal{O}\left(\int_{\Delta}\frac{1}{v^2}dv\right) = \mathcal{O}(b^{-5/8}) = o(1/b).$$

Similarly, we derive (we omit the details)

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{R^2}; R>0\right) &\overset{b}{\sim} \frac{3}{2Lb^2},\\
\mathbb{V}\left(\frac{1}{R^2}; R>0\right) &\overset{b}{\sim} \frac{1}{b^2}\left[\frac{3}{2L} - \frac{1}{L^2}\right],\\
\mathbb{E}\left(\frac{1}{R^{1/2}}; R>0\right) &\overset{b}{\sim} 2(\sqrt{2}-1)/(L\sqrt{b}).
\end{aligned}$$

More generally,

$$\mathbb{E}\left(\frac{1}{R^{\alpha}}; R>0\right) \overset{b}{\sim} \frac{1}{L\alpha b^{\alpha}}(2^{\alpha}-1), \alpha > 0$$

■

Now we obtain, by (5) and Thm 6.2 the following Thm

Theorem 6.4 *The limiting distribution of U/R for large b is Gaussian.*

Note that, by (4) and (13), we obtain

$$\begin{aligned}
\mathbb{E}\left(\left(\frac{U}{R}\right)^2; R>0\right) &\overset{b}{\sim} p^2 + \frac{pq}{Lb},\\
\mathbb{V}\left(\frac{U}{R}; R>0\right) &\overset{b}{\sim} \frac{pq}{Lb}.
\end{aligned} \tag{14}$$

This provides a confidence interval for p. With a confidence level of 5% for instance, we have

$$\left[\frac{U}{R} - 2\sqrt{\frac{pq}{Lb}} \le p \le \frac{U}{R} + 2\sqrt{\frac{pq}{Lb}}\right],$$

and, as we can estimate p by $\frac{U}{R}$, this leads to

$$\left[\frac{U}{R} - 2\sqrt{\frac{\frac{U}{R}\left(1-\frac{U}{R}\right)}{Lb}} \le p \le \frac{U}{R} + 2\sqrt{\frac{\frac{U}{R}\left(1-\frac{U}{R}\right)}{Lb}}\right].$$

6.3 Several Colors

If we are interested in the joint distribution of the statistic $U_1/R, \dots U_k/r$, which correspond to k different colors among the present colors, we have an asymptotic conditional multinomial distribution. For instance, for $k = 2$, this leads to

$$\binom{r}{u_1, u_2, r - u_1 - u_2} p_1^{u_1} p_2^{u_2} (1 - p_1 - p_2)^{r-u_1-u_2},$$

with mean rp_1, rp_2. So

$$\mathbb{E}\left(\frac{U_1}{R}; R > 0\right) = p_1, \mathbb{E}\left(\frac{U_2}{R}; R > 0\right) = p_2,$$

and we obtain similarly, conditioned on R

$$\mathbb{E}(U_1 U_2) = R(R-1)p_1 p_2,$$
$$\mathbb{E}\left(\left(\frac{U_1}{R}\right)\left(\frac{U_2}{R}\right)\right) = p_1 p_2 - \frac{p_1 p_2}{R},$$

and, unconditionning,

$$\mathbb{E}\left(\left(\frac{U_1}{R}\right)\left(\frac{U_2}{R}\right); R > 0\right) \overset{b}{\sim} p_1 p_2 - \frac{p_1 p_2}{Lb},$$

or

$$Cov\left(\frac{U_1}{R}, \frac{U_2}{R}; R > 0\right) \overset{b}{\sim} -\frac{p_1 p_2}{Lb}.$$

7 Multiplicities of colored keys

Assume that, to each key κ_i, we attach a counter giving its *observed* multiplicity μ_i. Also we assume that the multiplicities of color K keys are given by iid random variables (RV), with distribution function F, mean μ, variance σ^2 (functions of K). We can estimate the *total* number M of color K keys among the *total* number N of keys as follows.

Let m be the number of *distinct* color K keys among the n *distinct* keys. We recall that U is the number of color K keys among the R keys in the cache. From Section 6 (see(14)), we can estimate $p := m/n$ by $\tilde{p} = (U/R; R > 0)$. We have

$$\mathbb{E}(\tilde{p}; R > 0) \overset{b}{\sim} p,$$
$$\mathbb{V}(\tilde{p}; R > 0) \overset{b}{\sim} \frac{pq}{Lb}.$$

Also, we can estimate mean μ and variance σ^2 by $\tilde{\mu}$ and $\tilde{\sigma}^2$ as given by

$$\tilde{\mu} := \frac{V}{U}, \quad V := \sum_1^U \mu_i,$$
$$\tilde{\sigma}^2 := \frac{\sum_1^U (\mu_i - \tilde{\mu})^2}{U}.$$

Next we estimate n by $\tilde{n} = R2^J$ (see Sec. 2). We have, conditioned on U,

Theorem 7.1

$$\mathbb{E}(\tilde{\mu}) = \mu,$$
$$\mathbb{E}(\tilde{\mu}^2) = \mu^2 + \sigma^2 \mathbb{E}\left(\frac{1}{U}\right),$$
$$\mathbb{V}(\tilde{\mu}) = \sigma^2 \mathbb{E}\left(\frac{1}{U}\right).$$

$$\mathbb{E}(\tilde{n}) \sim n,$$
$$\mathbb{E}(\tilde{n}^2) \sim n^2\left(1+\frac{1}{(b-1)L}\right),$$
$$\mathbb{V}(\tilde{n}) \sim \frac{n^2}{(b-1)L}.$$

Proof. We only need

$$\mathbb{E}\left[\frac{V^2}{U^2}\middle| U\right] = \left[\frac{U\sigma^2+U^2\mu^2}{U^2}\middle| U\right]$$

. ■

Now we estimate m by $\tilde{m} = \tilde{n}\tilde{p} = 2^J U$ and M by $\tilde{M} = \tilde{m}\tilde{\mu}$. . But if we have two independent RV, X, Y, with mean and variance respectively $m_X, m_Y, \sigma_X^2, \sigma_Y^2$, it is easy to see that

$$\begin{aligned}\mathbb{E}(XY) &= m_X m_Y, \\ \mathbb{V}(XY) &= \sigma_X^2 m_Y^2 + \sigma_Y^2 m_X^2 + \sigma_X^2\sigma_Y^2.\end{aligned} \tag{15}$$

Here, our RV are not independent, but we can check that (15) is correct. The relation for the variances gives us a useful approximation. For instance

$$\mathbb{E}(\tilde{m}) \overset{b}{\sim} np = m,$$

and the approximation

$$\mathbb{V}(\tilde{m}) \sim \mathbb{V}(\tilde{n})p^2 + \mathbb{V}(\tilde{p})n^2 + \mathbb{V}(\tilde{n})\mathbb{V}(\tilde{p}) \overset{b}{\sim} \frac{n^2p(Lb-L+pL+1-p)}{L^2(b-1)b} \overset{b}{\sim} \frac{n^2p}{Lb} \text{ for large } b.$$

It remains to estimate $\mathbb{E}\left(\frac{1}{U}\right)$ in order to complete $\mathbb{E}(\tilde{\mu}^2), \mathbb{V}(\tilde{\mu})$. Using the binomial distribution $Bin(r,p)$ does not lead to a tractable expression. But, as R is large whp, we can use the Gaussian approximation for U as follows: conditioned on $R = r$, we have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{U}\right) &\sim \int_1^r \frac{\exp\left(-\frac{(u-rp)^2}{2rpq}\right)}{\sqrt{2\pi rpqu}}du \\ &\sim \int_{-rp}^{rq} \frac{\exp\left(-\frac{v^2}{2rpq}\right)}{\sqrt{2\pi rpq}}\frac{1}{rp}\left(1-\frac{v}{rp}+\frac{v^2}{r^2p^2}+\dots\right) \\ &\sim \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{v^2}{2rpq}\right)}{\sqrt{2\pi rpq}}\frac{1}{rp}\left(1-\frac{v}{rp}+\frac{v^2}{r^2p^2}+\dots\right) \\ &\sim \frac{1}{rp}\left(1+\frac{q}{rp}\right).\end{aligned}$$

Unconditionning, this gives

$$\mathbb{E}\left(\frac{1}{U}; R>0\right) \overset{b}{\sim} \frac{1}{Lbp} + \frac{3q}{4p^2Lb^2} \overset{b}{\sim} \frac{1}{Lbp}$$

8 The Black-Green Sampling

Assume that there are n distinct keys, among which np $(0 < p < 1)$ are Black (B), with known multiplicity μ_B and $nq, q := 1-p$ are Green (G), with known multiplicity $\mu_G > \mu_B$. We consider a case in some sense opposite to the one of Sec. 6: here we do not observe the color of each key, but we know a priori the multiplicities. We want to estimate the total number of keys: $N = np\mu_B + qn\mu_G$. (Here, we consider only the mean of our estimates). For instance, assume that each B key is unique

and each G key is present in triplicate. So we have a total of $N = np + 3qn = n(3-2p)$ keys. In the cache, we affect each key with an integer ν, representing the number of times this key has been observed. At the end, each key with $\nu = 1$ is obviously B. At each step $j, j = 0..J$, each time a key obtains the value $\nu = 3$, it is obviously G, and it is extracted from the cache. We have a vector counter C such that, each time a G key is extracted at step j from the cache, $C[j]$ is increased by 1. As all N keys are assumed to be distributed according to the uniform permutation distribution, we can consider the effect of each key on the cache as a Markov process: with probability $\frac{p}{3-2p}$, the key is B and it is inserted, with probability $\frac{3(1-p)}{3-2p}$, the key is G and three cases can occur: assume that the observed key appears in position $v, 1 \leq v \leq N$. Set $\tau := v/N$. Then

- With probability τ^2, the key was the third one among the three G keys with the same value, so it is deleted from the cache
- With probability $2\tau(1-\tau)$, the key was the second one, and it remains in the cache
- With probability $(1-\tau)^2$, the key is the first one, and it is inserted in the cache.

This can be seen as a Random walk on the cache. So the mean effect (on the cache size) of a G key at position v is given by

$$-\tau^2 + 0 \times 2\tau(1-\tau) + (1-\tau)^2 = 1 - 2\tau.$$

Finally, the mean effect of a key at position v is given by

$$\pi(\tau) = \frac{p}{3-2p} + \frac{3(1-p)}{3-2p}(1-2\tau) = \frac{3-6\tau-2p+6p\tau}{3-2p}.$$

Consider now the first step ($j = 0$). How many keys (in the mean) must be read in order to fill up the b positions in the cache? This is given by v_0, where $b = V(0, v_0)$ and

$$V(u_1, u_2) = \int_{u_1}^{u_2} \pi(\tau)dv = N\int_{u_1/N}^{u_2/N} \pi(\tau)d\tau = \frac{3(u_1^2-u_2^2)+3p(u_2^2-u_1^2)-3N(u_1-u_2)+2Np(u_1-u_2)}{N(3-2p)}.$$

This leads to

$$v_0 = \frac{-3N+2Np+[N(-3+2p)(2Np-12bp-3N+12b)]^{1/2}}{6(p-1)}.$$

An average of $b/2$ keys (starting with bit 1) are killed for the next step $j = 1$. But the mean number of available keys is also divided by 2. So the mean number of keys necessary to fill up the $b/2$ remaining positions in the cache is given by $v_1 - v_0$, where $b/2 = \frac{1}{2}V(v_0, v_1)$. This leads to

$$v_1 = \frac{-3N+2Np+[N(-3+2p)(2Np-24bp-3N+24b)]^{1/2}}{6(p-1)}.$$

More generally, the mean number of keys necessary to fill up the $b/2$ remaining positions in the cache at step j is given by $v_j - v_{j-1}$, where $b/2 = 2^{-j}V(v_{j-1}, v_j)$. This leads to

$$v_j = \frac{-3N+2Np+\left[N(-3+2p)(2Np-2^j12bp-3N+2^j12b)\right]^{1/2}}{6(p-1)},$$

and finally, the mean total number J of steps is given by $J = \lceil J^* \rceil$, where J^* is the solution of

$$N = \frac{-3N+2Np+\left[N(-3+2p)(2Np-2^{J^*}12bp-3N+2^{J^*}12b)\right]^{1/2}}{6(p-1)}.$$

This gives

$$J^* = \log\left(\frac{Np}{(3-2p)b}\right) = \log N - \log b + \log p - \log((3-2p)).$$

This is significally better than (3) *only* if $p \ll 1$.

Note that, at the end, the number of B keys in the sample is estimated by

$$2^J \times \text{ number of B keys in the cache,}$$

obviously only B keys (with $\nu = 1$) remain in the cache and the number of G keys in the sample is estimated by

$$3. \sum_{j=0}^{J} 2^j C[j].$$

Indeed, imagine that we mark a key with a $*$ as soon as it is decided to be G (because it is the third time we observe it). At step 0, $v \in [0, v_0)]$, all marked keys are counted in $C[0]$. At step 1, $v \in [v_0, v_1]$, all marked keys (staring with 0) are counted in $C[1]$, this corresponds in the mean, to $2C[1]$ G keys, etc. Actually, the vector counter C could be replaced by a single counter C into which, at each step j, we add the number of extracted R keys $\times 2^j$.

9 Appendix: Asymmetric Adaptive Sampling

For the sake of completeness, we analyze in this section the Asymmetric Adaptive Sampling. Assume that the hashing function gives asymmetric distributed bits. Let p denote the probability of bit 1 ($q := 1 - p$). Now, the number of keys in the cache is asymptotically Poisson with parameter nq^j and the number of keys in the twin bucket is asymptotically Poisson with parameter $npq^{j-1} = n\frac{p}{q}q^j$. So we set here

$$\begin{aligned} Q &:= 1/q, \\ Z &:= \frac{RQ^J}{n}, \\ \log &:= \log_Q, \\ L &:= \ln Q, \\ \tilde{\alpha} &:= \alpha/L, \\ \{x\} &:= \text{ fractional part of } x, \\ \chi_l &:= \frac{2l\pi\mathbf{i}}{L}. \end{aligned}$$

So the asymptotic distribution is now given, with $\eta := j - \log n$, by

$$p(r, j) \sim f(r, \eta) = \exp(-e^{-L\eta})\frac{e^{-Lr\eta}}{r!}\left[1 - \exp(-e^{-L\eta}p/q)\sum_{k=0}^{b-r}\frac{e^{-Lk\eta}(p/q)^k}{k!}\right], \tag{16}$$

and

$$p(j) := \mathbb{P}(J = j) = \sum_{r=0}^{b} p(r, j).$$

This leads to

$$\phi(r, \alpha) = \int_{-\infty}^{\infty} e^{\alpha\eta} f(r, \eta) d\eta = \frac{\Gamma(r - \tilde{\alpha})}{Lr!} - \sum_{k=0}^{b-r}\frac{\Gamma(r + k - \tilde{\alpha})q^{r+k-\tilde{\alpha}}(p/q)^k}{Lr!k!}.$$

9.1 Moments of $J - \log n$

Now we have

Theorem 9.1

$$\tilde{m}_{1,r} = -\frac{\psi(r)}{L^2 r} + \sum_{k=0}^{b-r} \frac{(\psi(r+k) - L) q^r p^k \Gamma(r+k)}{L^2 \Gamma(r+1)\Gamma(k+1)}, \quad r > 0,$$

$$\tilde{m}_{1,0} = \frac{1}{2} + \frac{\gamma}{L} + \sum_{k=1}^{b} \frac{(\psi(k) - L) p^k}{k L^2},$$

$$\tilde{m}_{2,r} = \frac{\psi(1,r) + \psi(r)^2}{L^3 r} + \sum_{k=0}^{b-r} -\frac{(-2\psi(r+k)L + L^2 + \psi(1,r+k) + \psi(r+k)^2) q^r p^k \Gamma(r+k)}{L^3 \Gamma(r+1)\Gamma(k+1)}, \quad r > 0,$$

$$\tilde{m}_{1,0} = \frac{1}{3} + \frac{\gamma}{L} + \frac{\pi^2}{6L^2} + \frac{\gamma^2}{L^2} + \sum_{k=1}^{b} -\frac{(-2\psi(k)L + L^2 + \psi(1,k) + \psi(k)^2) p^k}{k L^3},$$

$$w_{1,r} = \sum_{l \neq 0} \left[-\frac{\psi(r+\chi_l)\Gamma(r+\chi_l)}{L^2 \Gamma(r+1)} + \sum_{k=0}^{b-r} \frac{(\psi(r+k+\chi_l) - L)\Gamma(r+k+\chi_l) q^{r+k}}{L^2 \Gamma(r+1)\Gamma(k+1)} \right] e^{-2l\pi \mathbf{i} \log n}, \quad r > 0,$$

$$w_{1,0} = \sum_{l \neq 0} \left[-\frac{\psi(\chi_l)\Gamma(\chi_l)}{L^2} + \sum_{k=0}^{b} \frac{(\psi(k+\chi_l) - L)\Gamma(k+\chi_l) q^k}{L^2 \Gamma(k+1)} \right] e^{-2l\pi \mathbf{i} \log n}, \quad r > 0.$$

Proof. This is classical computer algebra (using Maple, with human guidance as usual) from $\phi(r, \alpha)$. ■

9.2 Moments of Z

Theorem 9.2 *The non-periodic components are given by*

$$m_{1,d} = 1 + \frac{(b-d)!}{L} \sum_{k=1}^{d-1} \left\{ {d \atop k} \right\} \frac{q^{k-d} - 1}{(d-k)(b-k)!},$$

$$\mathbb{V}(Z) \sim \frac{p}{(b-1)qL}.$$

The periodic component is obtained as follows

$$w_{1,d} = \sum_{l \neq 0} \frac{1}{L} \sum_{j=1}^{d-1} \left\{ {d \atop j} \right\} \left[(1 - q^{j-d} \right] \Gamma(j - d + \chi_l) \binom{b - d + \chi_l}{b - j} e^{-2l\pi \mathbf{i} \log n}.$$

Proof. We follow now the lines of [7] and [8], with suitable modifications.

$$\mathbb{E}[Z^d] \sim m_{1,d} = \sum_{r=1}^{d} \lim_{\alpha \to Ld} \phi(r, \alpha).$$

Let us compute $L m_{1,d}$.

- for $r \geq d+1$, this gives

$$T_1 = \sum_{r=d+1}^{b} r^d \frac{(r+d-1)!}{r!},$$

$$T_2 = -\sum_{r=d+1}^{b} r^d \sum_{k=0}^{b-r} \frac{(r+k-d-1)! q^{r+k-d} (p/q)^k}{r! k!},$$

- for $r \leq d$, we first obtain, for $k > d - r$,

$$T_3 = -\sum_{r=1}^{d} \frac{r^d}{r!} \left[\sum_{k=d-r+1}^{b-r} \frac{(r+k-d-1)! q^{r+k-d} (p/q)^k}{k!} \right],$$

- for $r \leq d$ and $k \leq d - r$, we must return to (16), in order to avoid singularities. This gives

$$\begin{aligned} T_4 &= \sum_{r=1}^{d} \frac{r^d}{r!} \int_0^\infty e^{-u} \left[1 - e^{-up/q} \sum_{k=0}^{d-r} \frac{u^k (p/q)^k}{k!} \right] \frac{du}{u^{d-r+1}} \\ &= \sum_{r=1}^{d} \frac{r^d}{r!} \Pi(d-r), \\ \Pi(j) &= \int_0^\infty e^{-u} \left[1 - e^{-up/q} \sum_{k=0}^{j} \frac{u^k (p/q)^k}{k!} \right] \frac{du}{u^{j+1}}. \end{aligned} \tag{17}$$

We have

$$\Pi(0) = \int_0^\infty e^{-u} \left[1 - e^{-up/q} \right] \frac{du}{u} = L,$$

and, by parts,

$$\begin{aligned} \Pi(j) &= \int_0^\infty \left[-e^{-u} + \frac{1}{q} e^{-u/q} \sum_{k=0}^{j} \frac{u^k (p/q)^k}{k!} - e^{-u/q} \sum_{k=0}^{j} \frac{k u^{k-1} (p/q)^k}{k!} \right] \frac{du}{ju^j} \\ &= \int_0^\infty \left[-e^{-u} + \frac{1}{qj!} e^{-u/q} u^j (p/q)^j + e^{-u/q} \sum_{k=0}^{j-1} \frac{u^k (p/q)^k}{k!} \right] \frac{du}{ju^j} \\ &= -\frac{\Pi(j-1)}{j} + \frac{(p/q)^j}{jj!}. \end{aligned}$$

Set

$$\Pi(j) = \frac{(-1)^j L}{j!} + \frac{(-1)^j B(j)}{j!}.$$

This leads to

$$B(j)(-1)^j = B(j-1)(-1)^j + (p/q)^j \frac{1}{j},$$

or

$$B(j) = \sum_{1}^{j} \frac{(-1)^k}{k} (p/q)^k, B(0) = 0.$$

Returning to (17), this gives (with an identity proved in [7]),

$$T_4 = \sum_{r=1}^{d} \frac{r^d}{r!} \left[\frac{(-1)^{d-r} L}{(d-r)!} + \frac{(-1)^{d-r} B(d-r)}{(d-r)!} \right] = L + T_5,$$

with

$$T_5 = \sum_{r=1}^{d} r^d \frac{(-1)^{d-r}}{r!(d-r)!} B(d-r).$$

Now T_2 and T_3 can be grouped, by setting $u = r + k$. This gives

$$T_6 = -\sum_{u=d+1}^{b} \sum_{r=1}^{u} \frac{r^d}{r!} \frac{(u-d-1)! q^{-d} p^u (q/p)^r}{(u-r!)}.$$

Let us try to simplify $T_7 := T_6 + T_1$

$$\begin{aligned}
T_6 &= -\sum_{u=d+1}^{b}\sum_{r=0}^{u}\sum_{k=0}^{d}\left\{{d \atop k}\right\}\frac{r^{\underline{k}}u!q^{-d}p^u(q/p)^r}{r!(u-r!)u^{\underline{d+1}}} \\
&= \sum_{u=d+1}^{b} q^{-d}p^u\sum_{k=0}^{d}\left\{{d \atop k}\right\}\frac{p^{-u}q^k u^{\underline{k}}}{u^{\underline{d+1}}}, \\
T_1 &= \sum_{r=d+1}^{b}\sum_{k=0}^{d}\left\{{d \atop k}\right\}\frac{r^{\underline{k}}}{r^{\underline{d+1}}}, \\
T_7 = T_6 + T_1 &= \sum_{u=d+1}^{b}\sum_{k=0}^{d-1}\left\{{d \atop k}\right\}\left[1-q^{k-d}\right]\frac{1}{(u-k)\dots(u-d)} \\
&= \sum_{k=0}^{d-1}\frac{\left\{{d \atop k}\right\}}{d-k}\left[\frac{1}{(d-k)!}-\frac{(b-d)!}{(b-k)!}\right]\left[1-q^{k-d}\right].
\end{aligned}$$

Putting everything together, this finally leads to

$$m_{1,d} = \frac{1}{L}\left[L + T_5 + T_7\right].$$

But even that could again be simplified!

$$\begin{aligned}
T_5 &= \sum_{r=1}^{d} r^d\frac{(-1)^{d-r}}{r!(d-r)!}B(d-r) \\
&= \sum_{i=0}^{d-1}\frac{(d-i)^d(-1)^i}{i!(d-i)!}\sum_{1\le j\le i}\frac{(-1)^j}{j}(p/q)^j \\
&= \sum_{\lambda=0}^{d}\left\{{d \atop \lambda}\right\}\sum_{i=0}^{d-\lambda}\frac{(d-i)!(-1)^i}{(d-i-\lambda)!i!(d-i)!}\sum_{1\le j\le i}\frac{(-1)^j}{j}(p/q)^j \\
&= \sum_{\lambda=0}^{d}\left\{{d \atop \lambda}\right\}\sum_{i=0}^{d-\lambda}\frac{(-1)^i}{(d-i-\lambda)!i!}\sum_{1\le j\le i}\frac{(-1)^j}{j}(p/q)^j \\
&= \sum_{\lambda=0}^{d}\left\{{d \atop \lambda}\right\}\sum_{1\le j\le d-\lambda}\frac{(-1)^j}{j}(p/q)^j\frac{1}{(d-\lambda)!}\sum_{i=j}^{d-\lambda}(-1)^i\binom{d-\lambda}{i} \\
&= \sum_{\lambda=0}^{d-1}\left\{{d \atop \lambda}\right\}\sum_{1\le j\le d-\lambda}\frac{(-1)^j}{j}(p/q)^j\frac{1}{(d-\lambda)!}(-1)^j\binom{d-\lambda-1}{j-1} \\
&= \sum_{\lambda=0}^{d-1}\left\{{d \atop \lambda}\right\}\frac{1}{(d-\lambda)(d-\lambda)!}\sum_{1\le j<d-\lambda}\binom{d-\lambda}{j}(p/q)^j \\
&= \sum_{\lambda=0}^{d-1}\left\{{d \atop \lambda}\right\}\frac{q^{\lambda-d}-1}{(d-\lambda)(d-\lambda)!}.
\end{aligned}$$

Now

$$m_{1,d} = 1 + \frac{1}{L}\sum_{k=1}^{d-1}\left\{{d \atop k}\right\}\frac{q^{k-d}-1}{(d-k)(d-k)!} + \frac{1}{L}\sum_{k=1}^{d-1}\frac{\left\{{d \atop k}\right\}}{d-k}\left[\frac{1}{(d-k)!}-\frac{(b-d)!}{(b-k)!}\right]\left[1-q^{k-d}\right]$$

$$= 1 + \frac{(b-d)!}{L}\sum_{k=1}^{d-1}\left\{{d \atop k}\right\}\frac{q^{k-d}-1}{(d-k)(b-k)!},$$

$$\mathbb{V}(Z) \sim \frac{p}{(b-1)qL}.$$

The periodic component is obtained as follows

$$\begin{aligned} w_{1,d} &= \sum_{l\neq 0}\sum_{r=1}^{d} r^d \lim_{\alpha\to L(d-\chi_l)} \phi(r,\alpha)e^{-2l\pi\mathbf{i}\log n} \\ &= \sum_{l\neq 0}\sum_{r=1}^{d} r^d \lim_{\alpha\to L(d-\chi_l)} \left[\frac{\Gamma(r-\tilde{\alpha})}{Lr!} - \sum_{k=0}^{b-r}\frac{\Gamma(r+k-\tilde{\alpha})q^{r+k-\tilde{\alpha}}(p/q)^k}{Lr!k!}\right] e^{-2l\pi\mathbf{i}\log n} \\ &= \sum_{l\neq 0}\frac{1}{L}\sum_{j=1}^{d-1}\left\{{d \atop j}\right\}\left[(1-q^{j-d}\right]\Gamma(j-d+\chi_l)\binom{b-d+\chi_l}{b-j}e^{-2l\pi\mathbf{i}\log n}. \end{aligned}$$

after all simplifications (we omit the details). Again, $m_{1,1} = 1, w_{1,1} = 0$. The moments of W can be similarly computed. We leave this to the interested reader. ■

9.3 Distribution of R

Theorem 9.3

$$\mathbb{P}(R=r) \sim p_r = \phi(r,0) = \frac{1}{L}\left[\frac{1}{r} - \sum_{u=r}^{b}\frac{(u-1)!}{r!(u-r)!}p^u(q/p)^r\right], r \geq 1,$$

$$\begin{aligned} \mathbb{E}(R) &\sim \frac{1}{L}\left[b - qb\right] = \frac{pb}{L}, \\ \mathbb{E}(R^2) &\sim \frac{1}{L}\left[b(b+1)/2 - q^2\frac{b(b-1)}{2} - qb\right]. \end{aligned}$$

Proof. We have

$$\mathbb{P}(R=r) \sim p_r = \phi(r,0) = \frac{1}{L}\left[\frac{1}{r} - \sum_{u=r}^{b}\frac{(u-1)!}{r!(u-r)!}p^u(q/p)^r\right], r \geq 1$$

and

$$\begin{aligned} p_0 &= 1 - \sum_{1}^{b} p_r \\ &= 1 - \frac{1}{L}\left[H_b - \sum_{r=1}^{b}\sum_{u=r}^{b}\frac{(u-1)!}{r!(u-r)!}p^u(q/p)^r\right] \\ &= 1 - \frac{1}{L}\left[H_b - \sum_{u=1}^{b}(u-1)!p^u\sum_{r=1}^{u}\frac{1}{r!(u-r)!}(q/p)^r\right] \\ &= 1 - \frac{1}{L}\left[H_b - \sum_{u=1}^{b}\frac{1}{u}p^u\sum_{r=1}^{u}\binom{u}{r}(q/p)^r\right] \\ &= 1 - \frac{1}{L}\sum_{u=1}^{b}\frac{p^u}{u}. \end{aligned} \tag{18}$$

Again, this can obtained from $\lim_{r\to 0}\phi(r,0)$.

The moments of R are computed as follows

$$\begin{aligned}
\mathbb{E}(R) &\sim \sum_1^b rp_r = \frac{1}{L}\left[b - \sum_{u=1}^b (u-1)!p^u \sum_{r=1}^u \frac{r}{r!(u-r)!}(q/p)^r\right] \\
&= \frac{1}{L}\left[b - \sum_{u=1}^b p^u (1/p)^{u-1}\frac{q}{p}\right] \\
&\sim \frac{1}{L}\left[b - qb\right] = \frac{pb}{L}, \\
\mathbb{E}(R^2) &\sim \sum_1^b r^2 p_r = \frac{1}{L}\left[b(b+1)/2 - \sum_{u=1}^b p^u (u-1)(q/p)^2 (1/p)^{u-2} - qb)\right] \\
&= \frac{1}{L}\left[b(b+1)/2 - \sum_{u=1}^b (u-1)q^2 - qb\right] \\
&= \frac{1}{L}\left[b(b+1)/2 - q^2\frac{b(b-1)}{2} - qb\right].
\end{aligned}$$

■

9.4 Moments of $1/R, R>0$ for large b

Theorem 9.4

$$\mathbb{E}\left(\frac{1}{R}; R>0\right) \overset{b}{\sim} \frac{1}{Lb}\frac{p}{q}.$$

Proof. Using (18), we have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{R}; R>0\right) &\sim \sum_{r=1}^b p_r/r = \frac{1}{L}\left[\sum_{r=1}^b \frac{1}{r^2} - \sum_{u=1}^b (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r\right] \\
&= \frac{1}{L}\left[H_b^{(2)} - \sum_{u=1}^\infty (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r + \sum_{u=b+1}^\infty (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r\right] \\
&= \frac{1}{L}\left[-\psi(1,b+1) + \frac{\pi^2}{6} - \sum_{u=1}^\infty (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r \right. \\
&\left. + \sum_{u=b+1}^\infty (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r\right]
\end{aligned}$$

Now

$$\begin{aligned}
&\sum_{u=1}^\infty (u-1)!p^u \sum_{r=1}^u \frac{1}{rr!(u-r)!}(q/p)^r \\
&= \sum_{r=1}^\infty \sum_{u=r}^\infty \frac{(u-1)!}{rr!(u-r)!}p^u (q/p)^r \\
&= \sum_{v=0}^\infty \frac{1}{(v+1)^2}(q/p)^{v+1} \sum_{w=v}^\infty \binom{w}{v} p^{w+1}
\end{aligned}$$

$$
\begin{aligned}
&= p\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}(q/p)^{v+1}p^v\sum_{s=0}^{\infty}\binom{-v-1}{v}(-p)^s\\
&= p\sum_{v=0}^{\infty}\frac{1}{(v+1)^2}(q/p)^{v+1}p^v(1-p)^{-(v+1)}\\
&= \zeta(2) = \frac{\pi^2}{6}. \qquad (19)
\end{aligned}
$$

This is the same value as in the symmetric case!

Now

$$\psi(1,b+1) \overset{b}{\sim} \frac{1}{b} + \mathcal{O}\left(\frac{1}{b^2}\right),$$

and

$$
\begin{aligned}
&\sum_{u=b+1}^{\infty}(u-1)!p^u\sum_{r=1}^{u}\frac{1}{rr!(u-r)!}(q/p)^r\\
&= T_1 + T_2,\\
T_1 &= \sum_{r=1}^{b+1}(q/p)^r\sum_{u=b+1}^{\infty}p^u\frac{(u-1)!}{rr!(u-r)!}\\
&= p\sum_{v=0}^{b}\frac{1}{(v+1)^2}(q/p)^{v+1}\sum_{w=b}^{\infty}\binom{w}{v}p^w,\\
T_2 &= \sum_{r=b+1}^{\infty}(q/p)^r\sum_{u=r}^{\infty}p^u\frac{(u-1)!}{rr!(u-r)!}\\
&= p\sum_{v=b}^{\infty}\frac{1}{(v+1)^2}(q/p)^{v+1}\sum_{w=v}^{\infty}\binom{w}{v}p^w\\
&= \sum_{v=b}^{\infty}\frac{1}{(v+1)^2}.
\end{aligned}
$$

We now turn to the asymptotics of $\binom{w}{v}p^w$ for large w. We obtain, by Stirling and setting $w = yv+\alpha$, (y will be fixed later on)

$$
\begin{aligned}
\binom{w}{v}p^w &\sim \frac{e^{-w}w^w\sqrt{2\pi w}}{e^{-(w-v)}(w-v)^{w-v}\sqrt{2\pi(w-v)}e^{-v}v^v\sqrt{2\pi v}}p^{yv+\alpha}\\
&= \frac{e^{-(yv+\alpha)}(yv+\alpha)^{yv+\alpha}\sqrt{2\pi(yv+\alpha)}}{e^{-((y-1)v+\alpha)}((y-1)v+\alpha)^{(y-1)v+\alpha}\sqrt{2\pi((y-1)v+\alpha)}e^{-v}v^v\sqrt{2\pi v}}p^{yv+\alpha}\\
&\sim \frac{e^{-v}(yv)^{yv+\alpha}\left(1+\frac{\alpha}{yv}\right)^{yv+\alpha}\sqrt{y/(y-1)}}{((y-1)v)^{(y-1)v+\alpha}\left(1+\frac{\alpha}{(y-1)v}\right)^{(y-1)v+\alpha}e^{-v}v^v\sqrt{2\pi v}}p^{yv+\alpha}\\
&\sim \frac{e^{\alpha+\frac{\alpha^2}{2yv}}\sqrt{y/(y-1)}}{e^{\alpha+\frac{\alpha^2}{2(y-1)v}}\sqrt{2\pi v}}\frac{y^{yv+\alpha}}{(y-1)^{(y-1)v+\alpha}}p^{yv+\alpha}
\end{aligned}
$$

Let us choose y such that

$$\frac{y^\alpha}{(y-1)^\alpha}p^\alpha = 1$$

This gives

$$y = \frac{1}{q}, y-1 = \frac{p}{q}$$

This finally leads to

$$\binom{w}{v} p^w \sim \frac{e^{-\frac{\alpha^2 q^2}{2vp}}}{\sqrt{2\pi vp/q^2}} \frac{p^v}{q^{v+1}}$$

This is a Gaussian function, centered at v/q with variance $\sigma^2 = pv/q^2$ and multiplied by $\frac{p^v}{q^{v+1}}$. Proceeding as in Section 6.2, and omitting the details, we finally obtain

$$T_1 \overset{b}{\sim} \sum_{v=bq}^{b} \frac{1}{(v+1)^2}$$

and

$$\mathbb{E}\left(\frac{1}{R}; R > 0\right) \overset{b}{\sim} \frac{1}{Lb} \frac{p}{q}.$$

The moments of $\left(\frac{1}{R^\alpha}; R > 0\right)$ are computed as in Section 6.2. ■

10 Conclusion

Once again, the techniques using Gumbel-like distributions and Stein methodology proved to be quite efficient in the analysis of algorithms such as Adaptive Sampling.

11 Acknowledgements

We would like to thank J. Lumbroso with whom we had many interesting discussions.

References

[1] A. D. Barbour and L. H. Y. Chen. *An introduction to Stein's method,* volume 4 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. Singapore University Press, 2005.

[2] A. D. Barbour and L. H. Y. Chen. *Stein's method and applications,* volume 5 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. Singapore University Press, 2005.

[3] P. Flajolet. On adaptive sampling. *Computing*, 34:391–400, 1990.

[4] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[5] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: a probabilistic analysis. *Random Structures and Algorithms*, 19(3,4):407–437, 2001.

[6] M. Loève. *Probability Theory, 3rd ed.* D. Van Nostrand, 1963.

[7] G. Louchard. Probabilistic analysis of adaptative sampling. *Random Structures and Algorithms*, 10:157–168, 1997.

[8] G. Louchard and H. Prodinger. Asymptotics of the moments of extreme-value related distribution functions. *Algorithmica*, 46:431–467, 2006.

[9] G. Louchard and H. Prodinger. On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discrete Mathematics*, 308,9:1538–1562, 2008. Long version: http://www.ulb.ac.be/di/mcs/louchard/gaps18.ps.

[10] G. Louchard, H. Prodinger, and M.D. Ward. The number of distinct values of some multiplicity in sequences of geometrically distributed random variables. *Discrete Mathematics and Theoretical Computer Science*, AD:231–256, 2005. 2005 International Conference on Analysis of Algorithms.

[11] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.