Efficient estimation of the cardinality of large data sets

Philippe Chassaing, Lucas Gerin

Efficient estimation of the cardinality of large data sets

Philippe Chassaing and Lucas Gerin

April 22, 2011

Abstract

The estimation of the number of distinct elements in a sequence of words, under strong constraints coming from applications to database analysis and network routing, is the problem we address in this article. Giroire has recently proposed a solution, under the form of a probabilistic algorithm that uses statistical properties of uniform random variables in [0,1]. Our objective here is twofold :

First, the analysis of this algorithm within the framework of Kullback information and estimation theory allows us to reinterpret a lower bound due to Indyk & Woodruff as a consequence of well-known inequalities.

Second, we show that a slight modification of the Giroire algorithm returns an estimation whose accuracy is optimal, among algorithms based on order statistics.
NB: This paper is the extended version of [2]

1 Introduction

Let $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)$ be a sequence of elements of a finite set $\mathcal{C}$. We wish to design an algorithm that computes $\theta = \operatorname{card}\{y_1, y_2, \ldots, y_n\}$. This question is sometimes referred to as the *distinct elements problem*. Consider the following naive algorithm:

Algorithm.

```
Initialize Dictionary={}
For j=1 to n
   Look for y_j in Dictionary
   If y_j is in Dictionary, do nothing
   otherwise, add y_j to Dictionary
next j
Return the size of Dictionary
```

During the execution of this algorithm, each word must be stored on the disk, so that the memory requirement cannot be less than linear in θ. For each y_j, a query in `Dictionary` is needed, that costs at least $\mathcal{O}(\log \theta)$ elementary operations. For a number of applications, these linear-space and log-time complexity lower bounds are not satisfactory, but, for the time being, they cannot be improved (see [1] for a discussion). For instance, finding the number of distinct elements in a given sequence is a main issue in network routing, where huge data sets have to be handled : according to [5], typically, at a given node of a network, packets arrive every 60 nanoseconds, approximately, but hardware limitations make impossible to process

more than 100 elementary operations on each packet, and the size of $\mathcal{C}$ does not allow to store every word.

Problem 1.1. *Does there exist an algorithm $\mathcal{A}$ that returns the number of different elements in $\mathbf{y}_n$, while statisfying the two following constraints:*

1. *$\mathcal{A}$ uses at most M bits of memory ;*
2. *each data y_i is treated in one pass: a few operations are processed (possibly modifying the M bits), then y_i is irreversibly erased.*

These constraints are too strong to allow an exact solution of the distinct elements problem, for a memory of M bits counts at most 2^M distinct elements. Partly following [10], Flajolet and Martin [4] proposed a probabilistic algorithm that overcomes this limitation by returning only an approximation of θ.

1.1 Approximate counting

We randomize Problem 1.1 through the use of *hash functions* :

Definition 1.2. *Given a* typical *sequence of distinct words, a* hash function $h : \mathcal{C} \to [0,1]$ *returns a sequence of random numbers, i.e. a sequence of numbers that behaves as the realization of a sequence of independent random variables, uniform on* $[0,1]$.

We assume from now on that we are given this idealized version h of a hash function, and that the *set*

$$\mathbf{X} = \{X_1, \ldots, X_n\},$$

where $X_i = h(y_i)$, is distributed like a *set* of θ independent realisations of a uniform random variable on $[0,1]$. The design of *good* hash functions is discussed for example in Knuth's book [7].

The algorithm proposed by Flajolet and Martin [4] is based on the distribution of some patterns of the dyadic representation of the X_i's. Their work has revealed the following phenomenon.

> *It is possible to recover (an approximate value of) θ, using only a (small and) constant memory M.*

Expectedly, the accuracy of the approximation of θ increases with M. Indyk and Woodruff provide the following theoretical bound for the accuracy reached with the help of a M-bits memory.

Proposition 1.3 (Indyk-Woodruff [6])**.** *For fixed $\varepsilon, \delta > 0$, one says that a probabilistic algorithm $\mathcal{A}(\varepsilon,\delta)$-approximates θ if it returns a value $\hat{\theta}$ such that*

$$\mathbb{P}(|\hat{\theta} - \theta| > \theta\varepsilon) \leq \delta.$$

Let $\mathcal{A}$ a one-pass algorithm that (ε,δ)-approximates θ, and assume that $\varepsilon = \mathcal{O}(a^{-\frac{1}{9}})$, in which a is the number of elements in $\mathcal{C}$. Let M be the memory required by $\mathcal{A}$ (M is expressed in bits). Then

$$\varepsilon^{-1} = \mathcal{O}\left(\sqrt{M}\right).$$

The proof of this bound is derived from learning theory (mainly, from bounds given in terms of VC-dimensions of well-choosen sets [11]), and requires a fine study of the geometry of some ℓ^1 or ℓ^2 spaces.

Assume that $\mathcal{A}$ returns an unbiased estimation of θ (*i.e.* $\mathbb{E}[\hat{\theta}] = \theta$). From the Bienaimé-Tchebychev inequality and Proposition 1.3, we obtain

$$\begin{aligned}\mathbb{P}(|\hat{\theta} - \theta| > \theta\varepsilon) &\leq \frac{\mathrm{Var}(\hat{\theta})}{(\theta\varepsilon)^2} \\ &\leq \mathrm{c}^{\mathrm{ste}} \ \frac{\mathrm{Var}(\hat{\theta})}{\theta^2} \ M,\end{aligned}$$

leading to the following lower bound :

$$\mathrm{Var}(\hat{\theta}) \geq \mathrm{c}^{\mathrm{ste}} \ \frac{\theta^2 \, \mathbb{P}(|\hat{\theta} - \theta| > \theta\varepsilon)}{M}.$$

Thus, an interpretation of Proposition 1.3 is that the variance of $\hat{\theta}$ cannot decrease faster than $\frac{\theta^2 \, \mathbb{P}(|\hat{\theta}-\theta|>\theta\varepsilon)}{M}$. The main goal of the present article is to analyse the algorithm MinCount, proposed by Giroire [5], within the framework of estimation theory. For this algorithm, and other algorithms based on *order statistics*, we shall show that a more precise lower bound of θ^2/M appears. It does not follow from geometric considerations but as a consequence of well-known inequalities in estimation theory. We show furthermore that a slightly improved variant of MinCount achieves this bound.

1.2 The MinCount algorithm

First, we describe a simplified version of MinCount [5], with parameter k (k is a given integer, not smaller than 2). Each word y_i is hashed, the corresponding value is compared to the k smallest values already observed. As usual, $X_{(1)} = \min_{j \leq n} X_j$ stands for the smallest X_j, and $X_{(k)}$ for the k-th smallest.

Algorithm.

```
Set MIN[1]=MIN[2]=...=MIN[k]=1
For j=1 to n,
    Compare X_j := h(y_j) with MIN[1:k]
    Update the vector (MIN[1]≤MIN[2]≤ ...≤ MIN[k]) of the k smallest
    values of the sequence (X_k)_{1≤k≤j-1} according to the value of X_j
next j
Return (k-1)/MIN[k].
```

This simplified algorithm satisfies the constraints of Problem 1.1, provided that $k = \mathcal{O}(M)$ (we are more precise below). Indeed, MinCount processes each word y_i using a single pass, and throughout the execution, only k real numbers are kept in memory. To see why $(k-1)/X_{(k)}$ gives a good estimation of θ, recall from [3, pages 8-13] the density of probability of the k-th order statistic:

$$\mathbb{P}(X_{(k)} \in [t, t+dt)) = \theta \binom{\theta - 1}{k - 1} t^{k-1} (1-t)^{\theta - k} dt.$$

It follows that $\mathbb{E}[1/X_{(1)}] = +\infty$, but as soon as $k \geq 2$,

$$\mathbb{E}[1/X_{(k)}] = \theta \binom{\theta - 1}{k - 1} \mathrm{B}(k - 1, \theta - k + 1) = \frac{\theta}{k - 1}, \tag{1}$$

where B is the Euler beta function.

If a number of the unit interval is stored with a precision of 2^{-s}, its storage requires s bits, and the available number of bits, M, allows to store $k = M/s$ numbers. In [5], MinCount is tuned in two directions :

1. the interval $[0, 1]$ is split into m sub-intervals $[0/m, 1/m)$, $[1/m, 2/m)$, ... ; the algorithm returns the k-th smallest value lying in each of these intervals ;
2. rather than $x \mapsto (k - 1)/x$, three different functions are proposed (depending on the parameters k, m).

The division of $[0, 1]$ in m intervals, called *stochastic averaging* in [4], allows to obtain a sample of m copies of $X_{(k)}$ almost at the same cost as one copy : when $m = 1$, each X_j has to be compared with the $k = M/s$ smallest values stored at time j, while, when $m \neq 1$, one has first to find i such that

$$\frac{i - 1}{m} \leq X_j < \frac{i}{m},$$

which can be done at almost no cost[1], then X_i is compared to the $\tilde{k} = M/(m\tilde{s})$ smallest values lying in $[(j - 1)/m, j/m)$, stored at time j.

Given the number M of bits, we see that a trade-off has to be made between m large (fewer comparisons, and a larger sample, but $\tilde{k}$ or $\tilde{s}$ smaller) and m small (k and s are larger). In what follows, we shall study the impact of k and m on the accuracy of the estimation of θ, and we shall consider that

$$M = k \times m,$$

thus not taking the impact of s into account. Here is the algorithm MinCount, as given in [5].

Algorithm.
Fix two integer parameters $k \geq 2, m \geq 1$.
Set $Z_{(p),i} = \frac{i}{m}$ for each $i \leq m, p \leq k$.
For j=1 to N
  $Z_j = h(y_j)$.
  Let i such that $Z_j \in [\frac{i-1}{m}, \frac{i}{m})$.
  Update the vector $(Z_{(1),i}, \ldots, Z_{(k),i})$ of the k smallest values in $[\frac{i-1}{m}, \frac{i}{m})$.
next j.
For each p, i, set $X_{(p),i} = Z_{(p),i} - \frac{i-1}{m}$.

Return a function $\hat{\xi} = \hat{\xi}(X_{(l),i}; 1 \leq i \leq m; 1 \leq l \leq k)$.

This algorithm fulfills the constraints of Problem 1.1, and the memory needed, when expressed in bits, is linear in $M := k \times m$.

The distinct elements problem is now reduced to a statistical problem : given a $k \times m$-*sample*

$$\Xi_{k,m} = \left(X_{(1),1}, \ldots, X_{(k),1}, \ldots, X_{(1),m}, \ldots, X_{(k),m}\right)$$

[1]It is a simple truncature if m is a power of 2.

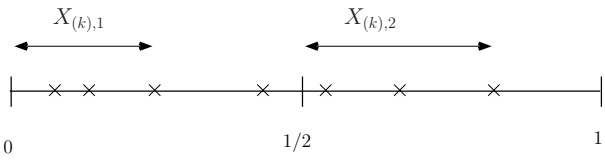

Figure 1: An example of the sample $(X_{(k),1}, \ldots, X_{(k),m})$, with $\theta = 7, m = 2, k = 3$. The crosses represent the 7 hashed values.

the distribution of which depends on an unknown parameter θ, one has to find the best *estimation* of θ. Giroire [5] proposes three different estimators ξ_1, ξ_2, ξ_3, each depending on $\Xi_{k,m}$:

PC: y'a un bleme avec la taille des captions

$$\begin{aligned}
\xi_1 &= \frac{k-1}{m} \sum_{i=1}^{m} \frac{1}{X_{(k),i}}, \\
\xi_2 &= C(k,m) \left(\sum_{i=1}^{m} \frac{1}{\sqrt{X_{(k),i}}} \right)^2, \\
\xi_3 &= \left(\frac{\Gamma(k-1/m)}{\Gamma(k)} \right)^{-m} \exp\left(-\frac{1}{m} \sum_{i=1}^{m} \log X_{(k),i} \right).
\end{aligned}$$

According to [5], the ξ_i's are asymptotically unbiased: $\mathbb{E}[\xi_i] \sim \theta$ as $\theta \to +\infty$. To our knowledge, for the distinct elements problem, Giroire's algorithms leading to the ξ_i were the best one-pass algorithms, i.e. producing the estimation with the lowest quadratic error, with ξ_3 ahead of the 2 others. One of the goals of this paper is to show that the best (with respect to quadratic error) estimation of θ based on $\Xi_{k,m}$, is given by

$$\hat{\xi} = \frac{km-1}{\sum_{i=1}^{m} X_{(k),i}}.$$

Remark 1.4. *Given $X_{(k),i} = x < \frac{1}{m}$, the random variables $(X_{(1),i}, \ldots, X_{(k-1),i})$ are uniformly distributed on $[0, x]$, i.e. their conditional distribution does not depend on θ. In other words, given $X_{(k),i}$, the knowledge of the $k-1$ observations $(X_{(1),i}, \ldots, X_{(k-1),i})$ does not bring any additional information on θ, so it comes as no surprise that $\hat{\xi}$, or even ξ_1, ξ_2, ξ_3, depend only on $X_{(k),i}$.*

Structure of the paper. The number of values falling in each of the m subintervals follows a multinomial distribution, and, conditionally, given the number of values falling in each of the m subintervals, the distribution of the $X_{(k),i}$'s is the distribution of the k-th order statistic of a uniform random sample with random (binomial) size. The reader surely admits that this description does not sound very tractable. In the next section, we shall thus study the asymptotic distribution of the $X_{(k),i}$'s for θ large, for this asymptotic distribution is much simpler than the true distribution. Then we shall prove that $\hat{\xi}$ is the best estimator, provided that the sample $\Xi_{k,m}$ follows the asymptotic distribution, and we shall explain the lower bound given by [6]. In the last Section, we discuss the optimality of $\hat{\xi}$ when $\Xi_{k,m}$ follows its actual distribution.

2 The best estimation in the asymptotic model

2.1 The asymptotic model

First we describe the asymptotic behavior of the $X_{(k),i}$'s. Recall that a random variable $\Gamma_{k,\theta}$ is said to follow the Gamma distribution with parameters k and θ if

$$\mathbb{P}(\Gamma_{k,\theta} \in [t, t+dt)) = \frac{t^{k-1}}{\Gamma(k)} \theta^k e^{-\theta t} \mathbf{1}_{t \geq 0}\, dt.$$

Furthermore, the vector $(m_{\theta,i})_{1 \leq i \leq k}$ of the k smallest among θ i.i.d. random variables, uniform on $[0,1]$, satisfies

$$\theta\, (m_{\theta,i})_{1 \leq i \leq k} \xrightarrow[\theta \to \infty]{\text{(law)}} T_k Y,$$

in which the k components of the column vector Y are i.i.d. and follow the exponential distribution with expectation 1, i.e. the Gamma distribution with parameters 1 and 1, and T_k is the $k \times k$ matrix with ones on and below the diagonal, and zeroes above the diagonal. As a consequence,

$$\theta\, m_{\theta,k} \xrightarrow[\theta \to \infty]{\text{(law)}} \Gamma_{k,1}.$$

Since there are approximately θ/m elements in each subinterval, and they are distributed as i.i.d. random variables, uniform on $[0, 1/m]$, we can expect that

$$(\theta X_{(k),1}, \ldots, \theta X_{(k),m}) \xrightarrow[\theta \to \infty]{\text{(law)}} (\Gamma^{(1)}, \ldots, \Gamma^{(m)}), \tag{2}$$

in which the $\Gamma^{(i)}$'s are m i.i.d. r.v. with common distribution Gamma $(k,1)$. Also, we can expect $\theta\, \Xi_{k,m}$ to be distributed, asymptotically, as m independent copies of $T_k Y$. Since

$$\frac{1}{\theta} \Gamma_{k,1} \overset{\text{(law)}}{=} \Gamma_{k,\theta},$$

equation (2) roughly says that, when θ goes large, the $X_{(k),m}$ behave as m independent random variables with distribution Gamma(k,θ). We shall not prove (2), for we need only a more specific result : the quadratic error is asymptotically the same when the $X_{(k),i}$'s are replaced by m independent Gamma random variables. As we shall see later,

$$\mathbb{E}[(\hat{\xi} - \theta)^2] = \mathcal{O}(\theta^2).$$

Thus we only need to compare $\hat{\xi}$ to functions $f(\Xi_{k,m})$ such that $\mathbb{E}[(f(\Xi_{k,m}) - \theta)^2]$ is not too large.

Proposition 2.1. *Let f be a continuous function: $\mathbb{R}_+^m \backslash \{0\} \to \mathbb{R}_+$. We assume that outside some neighbourhood of 0, f is bounded, while, in the neighbourhood of 0, there exists $r > 0$ such that*

$$|f(x)| = \mathcal{O}\left(\frac{1}{\| x \|^r}\right). \tag{3}$$

Then, for any $\varepsilon > 0$, there exists $c_\varepsilon > 0$ such that, for θ large enough,

LG: Terme exponentiellement petit en θ^2 pour être tranquille

$$\left|\mathbb{E}\left[\left(f(X_{(k),1},\dots,X_{(k),m})-\theta\right)^2\right]-\mathbb{E}\left[\left(f(\Gamma^{(1)},\dots,\Gamma^{(m)})-\theta\right)^2\right]\right|$$
$$\leq c_\varepsilon\theta^{\varepsilon-1}\mathbb{E}\left[\left(f(X_{(k),1},\dots,X_{(k),m})-\theta\right)^2\right]+4m\theta^2\exp(-\theta/(2m)^2). \quad (4)$$

in which the $\Gamma^{(i)}$'s are i.i.d. random variables with law Gamma(k,θ).

Consequently, we shall first assume that θ is large enough, and we shall replace the $\left\{X_{(k),i}, i=1,\dots,m\right\}$ with their limits $\left\{\Gamma^{(i)}, i=1,\dots,m\right\}$. We know that the $X_{(k),i}$'s are large only if θ is small, thus the assumption of boundedness of f outside a neighbourhood of 0 is not really binding for an estimator of θ. Similarly, a good estimation has to be moderately large when the $X_{(k),i}$'s are small, thus a good estimation fulfills necessarily (3). The proof of Proposition 2.1 is postponed to Section 2.3.

2.2 Lehmann-Scheffé and Cramér-Rao inequalities

We are now given a *sample* $(\Gamma^{(1)},\dots,\Gamma^{(m)})$ of m independent Gamma r.v. with parameters (k,θ). We assume that k is known, and we want to estimate the unknown parameter θ. We proceed by maximum likehood estimation: let $f_\theta:\mathbb{R}_+^m\to\mathbb{R}_+$ be the density of the m-sample :

$$f_\theta(x_1,\dots,x_m)= \theta^{km}\ \exp(-\theta\textstyle\sum_i x_i)\ \prod_i \frac{{x_i}^{k-1}}{\Gamma(k)}.$$

We are to compute

$$\hat\theta := \mathrm{argmax}_{\theta>0}\ \ln\left(f_\theta(\Gamma^{(1)},\dots,\Gamma^{(m)})\right),$$

thus we have to solve :

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\ \ln\left(f_\theta(\Gamma^{(1)},\dots,\Gamma^{(m)})\right)\\ &= \frac{km}{\theta}\ -\ \sum_i \Gamma^{(i)}, \end{aligned}$$

leading to $\tilde\theta=\frac{km}{\sum_i\Gamma^{(i)}}$. It turns out that this estimator is biased:

$$\mathbb{E}[\tilde\theta]=\theta\frac{km}{km-1}.$$

The next Proposition fixes the problem :

Proposition 2.2. *Set*

$$\hat\xi(x_1,\dots,x_m)=\frac{km-1}{x_1+\dots+x_m}.$$

For any θ,

$$\mathbb{E}[\hat\xi(\Gamma^{(1)},\dots,\Gamma^{(m)})]=\theta,$$
$$Var(\hat\xi(\Gamma^{(1)},\dots,\Gamma^{(m)}))=\frac{\theta^2}{km-2}.$$

Proof. By stability of the class of Gamma distributions under convolution,

$$\gamma := \Gamma^{(1)} + \cdots + \Gamma^{(m)}$$

is Gamma distributed with parameters km and θ. Thus,

$$\begin{aligned}\mathbb{E}\left[\frac{km-1}{\gamma}\right] &= \frac{(km-1)\theta^{km}}{\Gamma(km)} \int_0^{+\infty} x^{km-2}e^{-\theta x}dx \\ &= \frac{(km-1)\Gamma(km-1)}{\Gamma(km)}\,\theta \;=\; \theta.\end{aligned}$$

The variance is obtained through similar computations. □

Combined with Proposition 2.1, Proposition 2.2 entails that

Corollary 2.3.

$$\mathbb{E}\left[\left(\hat{\xi}(X_{(k),1},\ldots,X_{(k),m})-\theta\right)^2\right] = \frac{\theta^2}{km-2} + \mathrm{o}(\theta^2).$$

Corollary 2.3 calls for two remarks:

1. the asymptotic variance $\theta^2/(km-2)$ is indeed smaller than the variance of the estimators proposed in [5], and sets a new record for the lower quadratic error for a one-pass algorithm requiring bounded memory ;
2. since the memory required is linear in km, the quadratic error obtained in Corollary 2.3 is consistent with the theoretical result given in [6].

Let us now recall a few definitions from estimation theory. Given a m-sample $(X_1,\ldots,X_m)$, whose law is denoted P_θ, any random variable $S = S(X_1,\ldots,X_m)$ is called a *statistic*. Here, $(X_1,\ldots,X_m) = (\Gamma^{(1)},\ldots,\Gamma^{(m)})$ and we shall focus on the statistic $S = \sum_i \Gamma^{(i)}$. First, we check that S fulfills two conditions with deep connections with accuracy : *sufficiency* and *completeness*.

Definition 2.4. *Assume that P_θ admits a density f_θ with respect to the Lebesgue measure :*

1. *a statistic S is said to be* sufficient *for θ if f_θ can be written*

$$f_\theta(x_1,\ldots,x_m) \;=\; g(S(x_1,\ldots,x_m),\theta)\; h(x_1,\ldots,x_m),$$

in which g, h are two non-negative measurable functions, h not depending on θ.

2. *a statistic S is said to be* complete *if, for any measurable function ϕ,*

$$\{\forall\theta,\ \mathbb{E}[\phi(S)]=0\} \Rightarrow \{\forall\theta,\ \{\phi(S)\equiv 0, P_\theta - p.s.\}\}\,.$$

The next criterion ensures sufficiency and completeness :

Proposition 2.5 (see [9],Th.16 Chap.7)**.** *Assume that f_θ can be written*

$$f_\theta(x) = h(\mathbf{x})B(\theta)\exp\left(Q(\theta)R(x)\right),$$

where h, B, Q, R are measurable functions, h, B being positive. The statistic $S(x_1,\ldots,x_m) = \sum R(x_i)$ is complete and sufficient.

We apply the criterion with $h(x) = \frac{x^{k-1}}{\Gamma(k)}$, $Q(\theta) = -\theta$, $R(\mathbf{x}) = x$, $B(\theta) = 1$, we obtain:

Corollary 2.6. *Under the asymptotic model, $S = \sum_i \Gamma^{(i)}$ is a sufficient and complete statistic.*

Proposition 2.7 (Lehmann-Scheffé's Theorem)**.** *[see [9] Th.4, Chap.8] Let S be a complete and sufficient statistic, and let ξ^* be an unbiased estimator. The estimator $\mathbb{E}[\xi^*|S]$ is the unbiased estimator with the lowest variance. It is said to be* efficient.

Corollary 2.8. *Let $\tilde{\xi}$ be an unbiased estimator of θ.*

$$\mathbb{E}[(\tilde{\xi} - \theta)^2] \geq \mathbb{E}[(\hat{\xi} - \theta)^2] = \frac{\theta^2}{km - 2}. \tag{5}$$

This is a consequence of Lehmann-Scheffé's Theorem with $S = \sum_{i=1}^m \Gamma^{(i)}$ and $\xi^* = \hat{\xi}$, since $\mathbb{E}[\hat{\xi}|S] = \hat{\xi}$. This inequality is sharp for our model, but it is valid only for unbiased estimators. Cramér-Rao inequality [8, Th. 6.4, page 122] gives a more general lower bound :

Proposition 2.9 (Cramér-Rao inequality)**.** *Let g_θ be the density of $\Gamma^{(1)}$. Assume that $\theta \mapsto \log g_\theta(x)$ is continuously differentiable, and that the quantity*

$$I(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \log g_\theta(\Gamma^{(1)})\right]$$

is finite and positive. Let ξ^* be a square-integrable function such that $b(\theta) := \mathbb{E}[\xi^*] - \theta$ is continuously differentiable. then

$$\mathbb{E}\left[(\xi^* - \theta)^2\right] \geq \frac{(1 + b'(\theta))^2}{mI(\theta)} + b(\theta)^2.$$

In particular, if ξ^* is unbiased, its variance is bounded from below by $1/mI(\theta)$.

The quantity $I(\theta)$ is the *Fisher information*. In our case, it reduces to

$$\begin{aligned} I(\theta) &= -\mathbb{E}\left[\frac{d^2}{d\theta^2} \log\left(\frac{(\Gamma^{(1)})^{k-1}\ \theta^k}{\Gamma(k)}\ e^{-\theta\Gamma^{(1)}}\right)\right], \\ &= -\mathbb{E}\left[\frac{d^2}{d\theta^2}\left(k \log\theta - \theta\Gamma^{(1)}\right)\right], \\ &= \frac{k}{\theta^2}. \end{aligned}$$

Under the asymptotic model, Cramér-Rao inequality reads

$$\mathbb{E}[(\xi^* - \theta)^2] \geq (1 + b'(\theta))^2 \frac{\theta^2}{km} + b(\theta)^2, \tag{6}$$

for any estimator ξ^* such that $\theta \mapsto \mathbb{E}[\xi^*]$ is continuously differentiable. Cramér-Rao inequality confirms that quadratic error cannot decrease faster than θ^2/M. The estimator $\hat{\xi}$ achieves this lower bound, up to a factor $km/(km-2)$.

2.3 Proof of Proposition 2.1

Let A be the event

$$\begin{aligned} A = A_{k,m,\theta} &= \{\text{For any } 1 \le i \le m, \text{ at least } k \text{ hashed values lie in the } i\text{-th interval}\} \\ &= \{\text{For any } 1 \le i \le m, X_{(k),i} < \tfrac{1}{m}\} \end{aligned}$$

When $\theta \gg 2m^2$, A occurs with a high probability :

$$\begin{aligned} 1 - \mathbb{P}(A) &= \mathbb{P}\left(\cup_{i=1}^m \{\text{less than } k \text{ hashed values lie in } [\tfrac{i-1}{m}; \tfrac{i}{m})\}\right) \\ &\le m\ \mathbb{P}\left(\{\text{less than } k \text{ values lie in } [0; \tfrac{1}{m})\}\right) \\ &\le m\ \mathbb{P}(\mathcal{B}_{\theta,1/m} < k) \\ &\le m\ \mathbb{P}(\mathcal{B}_{\theta,1/m} - \tfrac{\theta}{m} < -\tfrac{\theta}{2m}) \\ &\le m\ \exp(-\theta/(2m^2)), \end{aligned} \tag{7}$$

in which $\mathcal{B}_{\theta,1/m}$ follows the binomial distribution with parameters $(\theta, 1/m)$, and in which (7) follows from Hoeffding's inequality. Now, write

$$\mathbb{E}\left[\left(f(X_{(k),1}, \ldots, X_{(k),m}) - \theta\right)^2\right] = \mathbb{E}[(f-\theta)^2 \mathbf{1}_A] + \mathbb{E}[(f-\theta)^2 \mathbf{1}_{\overline{A}}].$$

The restriction to A of the distribution of $(X_{(k),1}, \ldots, X_{(k),m})$ admits a density on $\mathbb{R}^m$, that can be computed along the lines of [3, pages 8-13], leading to :

$$\mathbb{E}[\left(f(X_{(k),1}, \ldots, X_{(k),m}) - \theta\right)^2 \mathbf{1}_A] = \int_{[0,1/m]^m} (f(x)-\theta)^2\ \frac{\theta!\,(1-\sum_i x_i)^{\theta-mk}}{(\theta-km)!}\ \prod_{i=1}^m \frac{{x_i}^{k-1}}{\Gamma(k)}\ dx.$$

On the set $\overline{A}$, at least one of the $X_{(k),i}$'s is equal to $1/m$, so that, using (7) and the fact that f is bounded outside a neighbourhood of 0, we find

$$\mathbb{E}[(f-\theta)^2 \mathbf{1}_{\overline{A}}] \le 2\left(\sup_{\|x\| \ge 1/m} |f(x)|^2 + \theta^2\right)(1 - \mathbb{P}(A)) \le 4m\theta^2 \exp(-\theta/(2m)^2), \tag{8}$$

which gives the last term in (4), provided θ is large enough. Thus the proof reduces to show that for any $\varepsilon > 0$,

$$I = \mathcal{O}\left(\theta^{\varepsilon-1}\mathbb{E}\left[\left(f(X_{(k),1}, \ldots, X_{(k),m}) - \theta\right)^2\right]\right),$$

in which I is defined below :

$$\begin{aligned} I &= I_1 - I_2, \\ I_1 &= \int_{[0,1/m]^m} (f(x)-\theta)^2\ \frac{\theta!\,(1-\sum_i x_i)^{\theta-mk}}{(\theta-km)!}\ \prod_{i=1}^m \frac{{x_i}^{k-1}}{\Gamma(k)}\ dx, \\ I_2 &= \int_{\mathbb{R}_+^m} (f(x)-\theta)^2\ \theta^{km} \exp(-\theta \sum_i x_i)\ \prod_{i=1}^m \frac{{x_i}^{k-1}}{\Gamma(k)}\ dx. \end{aligned}$$

By the substitution $y_i = \theta x_i$, and with the notation $s = \sum_i y_i$, we get

$$I_1 = \int_{[0,\theta/m]^m} (f(y/\theta)-\theta)^2\ \tfrac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\ \prod_{i \le m} \frac{{y_i}^{k-1}}{\Gamma(k)}\ dy, \tag{9}$$

$$I_2 = \int_{\mathbb{R}_+^m} (f(y/\theta)-\theta)^2\ \exp(-s)\ \prod_{i=1}^m \frac{{y_i}^{k-1}}{\Gamma(k)}\ dy. \tag{10}$$

Set

$$F(y,\theta) \;=\; 1 \;-\; \frac{(\theta-km)!\;\theta^{km}\;\exp(-s)}{\theta!\,\left(1-\frac{s}{\theta}\right)^{\theta-mk}}.$$

We write $I = J - K$, with

$$J = \int_{[0,\theta/m]^m} F(y,\theta)\;\left(f\left(y/\theta\right)-\theta\right)^2\;\tfrac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;dy,$$

$$K = \int_{\mathbb{R}_+^m\setminus[0,\theta/m]^m} \left(f\left(y/\theta\right)-\theta\right)^2\;\exp(-s)\;\prod_{i=1}^{m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;dy.$$

We will show that these two integrals are $\mathrm{o}(\theta^2)$. Since f is continuous and bounded away from zero outside some neighbourhood of 0, there exist $c, c' > 0$ such that in $\mathbb{R}_+^m\setminus\left[0,\frac{\theta}{m}\right]^m$, when θ is large enough,

$$|f(y/\theta)-\theta|^2 \le c+\theta^2 \le c'\theta^2.$$

Then

$$|K| \le\; c'\theta^2\;\int_{\mathbb{R}_+^m\setminus\left[0,\frac{\theta}{m}\right]^m}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;e^{-s}\;dy$$

that vanishes exponentially fast, the integral on the right hand side being the probability that the maximum of an m-sample of Gamma distributed random variables is larger than θ/m. Let us write $J = J_1 + J_2$ in which

$$J_1 = \int_{[0,\theta^\alpha/m]^m} F(y,\theta)\;\left(f\left(y/\theta\right)-\theta\right)^2\;\tfrac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;dy,$$

$$J_2 = \int_{\left[0,\frac{\theta}{m}\right]^m\setminus\left[0,\frac{\theta\alpha}{m}\right]^m} F(y,\theta)\;\left(f\left(y/\theta\right)-\theta\right)^2\;\tfrac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;dy.$$

and assume that $\alpha \in (0,1/2)$. Let

$$F(y,\theta) = 1 - e^{\psi(s,\theta)},$$

in which

$$\begin{aligned}\psi(s,\theta) &= -s-\theta\ln\left(1-\tfrac{s}{\theta}\right)+mk\ln\left(1-\tfrac{s}{\theta}\right)-\sum_{\ell=1}^{km-1}\ln\left(1-\tfrac{\ell}{\theta}\right)\\ &= \mathcal{O}\left(\theta^{2\alpha-1}\right)+\mathcal{O}\left(\theta^{\alpha-1}\right)+\mathcal{O}\left(\theta^{-1}\right),\end{aligned}$$

as long as $y \in [0,\theta^\alpha/m]^m$. Using (9), we conclude that there exists $c_\alpha > 0$ such that for, θ large enough,

$$J_1 \le c_\alpha\theta^{2\alpha-1}I_1 \le c_\alpha\theta^{2\alpha-1}\mathbb{E}\left[\left(f(X_{(k),1},\ldots,X_{(k),m})-\theta\right)^2\right].$$

As concerns J_2, we write $J_2 = J_{2,1} - J_{2,2}$, with

$$J_{2,1} = \int_{\left[0,\frac{\theta}{m}\right]^m\setminus[0,\theta^\alpha/m]^m}\;\left(f\left(y/\theta\right)-\theta\right)^2\;e^{-s}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}dy,$$

$$J_{2,2} = \int_{\left[0,\frac{\theta}{m}\right]^m\setminus[0,\theta^\alpha/m]^m}\;\left(f\left(y/\theta\right)-\theta\right)^2\;\tfrac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\;\prod_{i\le m}\frac{{y_i}^{k-1}}{\Gamma(k)}\;dy.$$

First, we bound $J_{2,1}$: by the second assumption of Proposition 2.1, there exist $c, c' > 0$ such that, for all $y \in \left[0, \frac{\theta}{m}\right]^m \setminus \left[0, \frac{\theta^\alpha}{m}\right]^m$,

$$\begin{aligned} |f(y/\theta) - \theta|^2 &\leq \theta^2 + f^2(y/\theta) \\ &\leq \theta^2 + c\, \frac{1}{1 \wedge \parallel y/\theta \parallel^{2r}} \\ &\leq \frac{c'\theta^{2r}}{\parallel y \parallel^{2r}}, \end{aligned}$$

in which r can be chosen larger than 1. Since $y \in \left[0, \frac{\theta}{m}\right]^m \setminus \left[0, \frac{\theta^\alpha}{m}\right]^m$, we have $s \geq \theta^\alpha/m$. It follows that

$$\begin{aligned} J_{2,1} &\leq \int_{\left[0,\frac{\theta}{m}\right]^m \setminus [0,\theta^\alpha/m]^m} \frac{c'\theta^{2r}\ e^{-s}}{\parallel y \parallel^{2r}} \prod_{i \leq m} \frac{{y_i}^{k-1}}{\Gamma(k)}\ dy. \\ &\leq\ c'\theta^{2r}\ e^{-\theta^\alpha/m} \int_{\left[0,\frac{\theta}{m}\right]^m \setminus [0,\theta^\alpha/m]^m} \prod_{i \leq m} \frac{{y_i}^{k-1}}{\Gamma(k)}\ \frac{dy}{\parallel y \parallel^{2r}}, \end{aligned}$$

in which the last integral is polynomial in θ. For $J_{2,2}$, we have, as soon as $\theta \geq 2mk$,

$$\frac{\theta!(1-(s/\theta))^{\theta-mk}}{(\theta-km)!\theta^{km}}\ \leq\ (1-(s/\theta))^{\theta/2}\ \leq\ \exp(-s/2).$$

Thus, by the same argument, $J_{2,2}$ vanishes exponentially fast. To finish the proof, take $\varepsilon = 2\alpha$.

3 Optimality of $\hat{\xi}$

We return to the original model, in which $X_{(k),i}$ denotes the k-th smallest value lying in $[\frac{i-1}{m}, \frac{i}{m})$. We wish to discuss the optimality of

$$\hat{\xi} = \frac{km-1}{\sum_{i=1}^m X_{(k),i}}.$$

The combination of (5) and (6) gives the following result, which is the main result of the present work.

Theorem 3.1 (Optimality of $\hat{\xi}$)**.** *Let $\tilde{\xi} = \tilde{\xi}(\mathbf{x})$ be a continuous function on $\mathbb{R}_+^m - \{0\}$. Assume that there exists $r > 0$ such that, in the neighbourhood of 0,*

$$|f(x)| = \mathcal{O}\left(\frac{1}{\parallel x \parallel^r}\right).$$

The application

$$b(\theta) := \mathbb{E}[\tilde{\xi}(\Gamma^{(1)}, \ldots, \Gamma^{(m)})] - \theta,$$

is continuously differentiable, and

$$\mathbb{E}[(\tilde{\xi}(X_{(k),1}, \ldots, X_{(k),m}) - \theta)^2] \geq \frac{\theta^2}{km}\left(1 + b'(\theta)\right)^2 + \mathrm{o}(\theta^2).$$

If we assume furthermore that $\tilde{\xi}$ is unbiased in the asymptotic model i.e.

$$\mathbb{E}[\tilde{\xi}(\Gamma_1, \ldots, \Gamma_m)] = \theta, \tag{11}$$

then

$$\mathbb{E}[(\tilde{\xi}(X_{(k),1},\dots,X_{(k),m})-\theta)^2] \geq \mathbb{E}[(\hat{\xi}-\theta)^2] + \mathrm{o}(\theta^2),$$
$$= \frac{\theta^2}{km-2} + \mathrm{o}(\theta^2).$$

Remark 3.2. *The second assumption ($\tilde{\xi}$ unbiased in the asymptotic model) was implicitly made in [5].*

Proof. Proposition 2.1 with $\varepsilon = 1/2$ implies that there exists $c > 0$ such that, for θ large enough,

$$\mathbb{E}[(\tilde{\xi}(X_{(k),1},\dots,X_{(k),m})-\theta)^2] \geq \mathbb{E}[\tilde{\xi}(\Gamma_1,\dots,\Gamma_m)] - c\theta^{-1/2}\mathbb{E}[(\tilde{\xi}(X_{(k),1},\dots,X_{(k),m})-\theta)^2] + \mathrm{o}(\theta^2),$$
$$\geq \frac{\theta^2}{km}\left(1+b'(\theta)\right)^2 - c\theta^{-1/2}\mathbb{E}[(\tilde{\xi}(X_{(k),1},\dots,X_{(k),m})-\theta)^2] + \mathrm{o}(\theta^2)$$

by (6). The conclusion of the Theorem follows : if $\mathbb{E}[(\tilde{\xi}-\theta)^2]$ is larger than, say, $\theta^{7/3}$ then there is nothing to prove ; if it is smaller then the right-hand term is

LG: Petite modif de la preuve pour être plus clair

$$\frac{\theta^2}{km}\left(1+b'(\theta)\right)^2 + \mathrm{o}(\theta^2).$$

If we assume furthermore that (11) holds, then (5) gives the second assertion of the theorem. □

3.1 The case $m = 1$

The lower bound given by Theorem 3.1 depends on the choice of (k, m) only through the product km. Thus, regardless of algorithmic considerations, the quadratic error does not benefit from the partition of $[0, 1]$ in m sub-intervals, and we may assume $m = 1$ to study the optimality of our estimator. When $m = 1$, the law of the observation $X_{(k),1}$ is easy to deal with, and we obtain a sharp and exact lower bound (*i.e.* valid for any θ).

The irrelevance, with respect to statistics, of splitting $[0, 1]$ into m subintervals, is perhaps clearer in the asymptotic model. We noted in Section 2.1 that $\Xi_{k,m}$ is distributed, asymptotically, as m independent copies of T_kY, in which Y is a k-sample of the exponential distribution with expectation $1/\theta$, and

$$T_kY = (Y_1, Y_1+Y_2, Y_1+Y_2+Y_3, \dots, Y_1+Y_2+\cdots+Y_k).$$

From a statistical perspective, since T_k is one-to-one, there is no loss of information from Y to T_kY. Thus, for the estimation of θ, $\Xi_{k,m}$ is equivalent to m independent copies of a k-sample of the exponential distribution with expectation $1/\theta$, i.e. a km-sample of the exponential distribution with expectation $1/\theta$. But this is also equivalent to $T_{km}\hat{Y}$, in which $\hat{Y}$ is a km-sample of the exponential distribution with expectation $1/\theta$: we can obtain $T_{km}\hat{Y}$ as the asymptotic distribution of the first km order statistics of a θ-sample of uniform random variables on $[0, 1]$, that is, without splitting the interval $[0, 1]$ into m subintervals. For the exponential distribution, the sum of the km elements of the sample is known to be complete and sufficient : when splitting, this corresponds to the sum of the m copies of the k-th order statistic, and when not splitting, this sum is asymptotic to the km-th order statistic.

Theorem 3.3. *Consider algorithm* MinCount*, with parameter $m = 1$. It returns*

$$\hat{\xi}(X_{(k)}) = \frac{k-1}{X_{(k)}}.$$

For any θ, $\hat{\xi}$ is an unbiased estimator. Furthermore, $\hat{\xi}$ is the unbiased estimator with the lowest variance.

Note that in this particular case $m = 1$, our estimator coincides with the estimator ξ_1 proposed by Giroire.

LG: Petite remarque ajoutée

Proof. We know from (1) that $\hat{\xi}$ is unbiased. We apply the Lehmann-Scheffé Theorem to the law Q_θ of $X_{(k)}$.

$$\begin{aligned} Q_\theta(x)dx &= \theta \binom{\theta-1}{k-1} x^{k-1}(1-x)^{\theta-k}dx \\ &= B(\theta)h(x)\exp((\theta-k)\log(1-x))dx, \end{aligned}$$

with notations of Proposition 2.5. This shows that the statistic $\log(1 - X_{(k)})$ is complete and sufficient. Thus the variance of the statistic

$$\mathbb{E}\left[\frac{k-1}{X_{(k)}} \;\middle|\; \log(1-X_{(k)}) \right] \;=\; \frac{k-1}{X_{(k)}}$$

is minimal. □

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 1999.

[2] Ph. Chassaing and L. Gerin. Efficient estimation of the cardinality of large data sets. In *DMTCS Proceedings of Fourth Colloquium on Mathematics and Computer Science*, pages 419–422, 2006.

[3] H.A. David. *Order statistics*. Wiley & Sons, 1981.

[4] Ph. Flajolet and N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

[5] F. Giroire. Order statistics and estimating cardinalities of massive data sets. *Discrete Applied Mathematics*, 157:406–427, 2009.

[6] P. Indyk and D. Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science, Boston*, 2003.

[7] D.E. Knuth. *The Art of Computer Programming, vol. 3 : Sorting and Searching*. Addison-Wesley, 1973.

[8] E.L. Lehmann. *Theory of Point Estimation*. John Wiley & sons, 1983.

[9] B.W. Lindgren. *Statistical Theory*. Chapman & Hall, 4th edition, 1993.

[10] R. Morris. Counting large numbers of events in small registers. *Communications of the ACM*, 21:840–842, 1978.

[11] V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.