EXACT CONFIDENCE INTERVAL FOR GENERALIZED FLAJOLET-MARTIN ALGORITHMS

GIACOMO ALETTI

Abstract. This paper develop a deep mathematical-statistical approach to analyze a class of Flajolet-Martin algorithms (FMa), and provide a exact analytical confidence interval for the number F_0 of distinct elements in a stream, based on Chernoff bounds. The class of FMa has reached a significant popularity in bigdata stream learning, and the attention of the literature has mainly been based on algorithmic aspects, basically complexity optimality, while the statistical analysis of these class of algorithms has been often faced heuristically. The analysis provided here shows a deep connections with special mathematical functions and with extreme value theory. The latter connection may help in explaining heuristic considerations, while the first opens many numerical issues, faced at the end of the present paper. Finally, MonteCarlo simulations are provided to support our analytical choice in this context.

Contents

1. Introduction

Data streams [7] are sequence of objects that cannot be available for random access, but must be analyzed sequentially when they arrive and immediately discharged. Streaming algorithms process data streams, and have reached a very rich audience since the last decades. Typically, these kinds of algorithms have a limited time to complete their processes and have access to limited amount of memory, usually logarithmic in the quantity of interest.

One of the main application in streaming algorithms concerns the problem of counting distinct elements F_0 in a stream. In [13], the authors develop the first algorithm for approximating F_0based on hash functions. This algorithm was then formalized and made popular in [6], where it was presented the forefather of the class of algorithms that takes the name of *Flajolet-Marin algorithms* (here, FMa). Three extensions in FMa were presented in [8], together with a complete description of the drawback and of the strength of the previous attempts. The first optimal (in complexity) algorithm has been proposed and proved in [18] and, nowadays, the FMa covers a lot of applications. As only an example, in [16], an application with multiset framework is developed from one of the most recent versions of FMa, and it estimates the number of "elephants" in a stream of IP packets.

This class of algorithms is essentially based on the following concept. When an object arrives form the stream, one (ore more, independent) hash functions are applied to it, and then the object is immediately discharged. The results of these functions are melted with what saved in memory (that has a comparable size). The memory is updated, if necessary, with the result of this procedure, and then the process is ready for the next object. The estimate of F_0 may be queried when necessary, and it is a function of the memory content.

The key point is the fact that the central operation is made with a function which must be associative, commutative and idempotent, so that multiple evaluations on the same object do not affect the final outcome, which results in the combination of the hash values of the F_0 distinct objects. A good candidate for such a function is the max function applied to a "signature" of each object, that is the core of such streaming algorithms. The same idea has recently used for other distributed algorithms (see [5] for simulation of discrete random variables), where new entries or single changes should not make all the algorithm starts afresh.

As stated before, the main contribution in the study of FMa concerned complexity problems, and a deep mathematical-statistical approach has not yet developed, even if this class of algorithm is probabilistic. This paper is a first attempt in this direction. The main contribution here is the analytical and numerical control of FMa based on a pure mathematical statistic approach, while we leave the measure of the goodness of the FMa to other studies (see [12] for a continuously updated work). In particular, we give here an analytical exact confidence interval for the quantity F_0. More precisely, we analyze an extension of the algorithms given above, and given $\mathfrak{p} > 0$, we will find $a, b > 0$, function of the memory content, such that

$$P(a \leq f(F_0) \leq b) \geq 1 - \mathfrak{p}, \tag{1}$$

where f is a given, strictly increasing, special function. It is important to note that the approximations for F_0 as in (1) given in literature are not satisfactory. In some situations, the asymptotic behavior of the interval is calculated through a Central Limit Theorem (see [14]), but the huge skewness implicit in the algorithm variables (even in logarithmic scale) makes the Central Limit Theorem questionable. To overcome this observation, Chebichev and Markov bounds are sometimes used to compute confidence intervals, see the papers cited in [18], where the results are analyzed in terms of optimal complexity (in space and time) without exploiting possible benefits in reducing the magnitude of the interval length.

These facts suggest us to not base the confidence interval on statistical asymptotic properties, but to build an exact confidence interval, based on concentration inequalities. In particular, we use Chernoff bounds, and we give an analytical approximation of the resulting inequalities. We show with MonteCarlo simulations that the analytical approximation does not affect the result significantly. Moreover, we show that the same result derive from the use of the Chernoff bounds on the limiting distribution that would be obtained with extreme value theory.

It is not surprising that some new analytical special functions appear in the analysis of the algorithm. In particular, a p-modification of the analytical extension $\mathbb{h}(x) = \mathbb{h}_1(x)$ of the harmonic number function arises here as the mean value of a particular statistics of interest, and $\mathbb{h}_p(F_0)$ is a quantity that appears in the paper. Notably, the heuristic approximation in the classical case studied in literature (with $p = 1$) gives a value that is of the order of magnitude we give in our subsequent estimations.

In addition, we discuss here a possible numerical implementation of the confidence interval in real time. To answer to this question, we develop a algorithm to solve all the relevant nonlinear problem with a cubic rate of convergence and we provide the necessary numeric bounds to apply it. As a byproduct, we could give the algorithm that calculates the log-shortest confidence interval.

The paper is structured in the following way. In the next Section 2 we first describe how FMa works. We show how data are stored in memory and queried from it, and then we analyze these processes from a mathematical and a statical point of view.

The main result, Theorem 3.1, is given at the beginning of the Section 3, and the connection with the asymptotic results of the extreme value theory is immediately discussed in Section 3.1. The proof of the main result is based both on a analytical computations of Chernoff bounds given in Section 4, and on the expected value of a quantity of interest, given in Section 5. The Section 6 shows the goodness of the choice of the analytical approximations given in Section 4.

In Section 7 we face numerically some nonlinear equations that are necessary to for query the interval (1) in the equivalent form: $P(f^{-1}(a) \leq F_0 \leq f^{-1}(b)) \geq 1 - \mathfrak{p}$. In particular, we give some sharp upper and lower bounds to develop cubic rate algorithms together with more robust bisecting algorithms. As a byproduct, the algorithm that calculates the log-shortest confidence interval is given at the end of the section.

Finally, Appendix A defines the main properties of some special mathematical functions that are used in this paper. Appendix B concludes the paper with the technicalities needed to find lower and upper bounds contained in Section 7.

2. Description of the algorithm

The main task of FMa is to provide an estimation of F_0, the unknown number of distinct elements in a real-time stream of possible repeating objects, based on c_0 independent hash functions. Our memory data structure is a generalization of a HyperLogLog data structure (see [10, 18, 12]), and consists of two matrices $\mathbb{X}$ and $\mathbb{Z}$ with 2^{r_0} rows and c_0 columns. The use of $\mathbb{Z}$ is an addition of this paper to the classical algorithms given above, and it is used to increase the accuracy of the estimation of F_0 (see Section 3), by using z_0 bits of each hashing function. The streaming algorithm that updates $\mathbb{X}$ and $\mathbb{Z}$ in memory is given in Algorithm 1. The flow of information is as follows. From each hash function H_c, we extract the following information on a stream object o:

$$H_c(o) = \underbrace{\overset{r_0 \text{ bits}}{01\cdots101}}_{R\in\{1,\dots,2^{r_0}\}}\ \underbrace{\overset{z_0 \text{ bits}}{10\cdots01}}_{Z}\underbrace{\overset{X \text{ bits}}{00\cdots0001}}_{X\in\{1,2,\dots\}}\underbrace{01101000\cdots}_{\text{not used}} \tag{2}$$

The data are then updated according to the following procedure:

if $X < \mathbb{X}_{R\,c}$**:** do nothing;
if $X > \mathbb{X}_{R\,c}$**:** set $\mathbb{X}_{R\,c} = \bar{X}$ and $\mathbb{Z}_{R\,c} = Z$;
if $X = \mathbb{X}_{R\,c}$**:** set $\mathbb{Z}_{R\,c} = \min(\mathbb{Z}_{R\,c}, Z)$.

```
Data: Data Stream of Objects $\{o_1, o_2, \dots, \}$
Input: $c_0$ hash functions, $r_0 \geq 0$ and $z_0 \geq 0$ small integers
Output: Two matrices $\mathbb{X}$ and $\mathbb{Z}$ with $r_0 = 2^{r_0}$ rows and $c_0$ columns
Set $\mathbb{X} \equiv 0$, $\mathbb{Z} \equiv 2^{z_0} - 1$ (binary);
foreach $o$ in Stream do
    for $c \leftarrow 1$ to $c_0$ do
        /* compute the $c$-hash function on $o$, obtaining a finite sequence $(s_1, s_2, \dots)$ of 0 and
           1                                                                   */
        $(s_1, s_2, \dots) \leftarrow H_c(o)$;
        $R \leftarrow 1 + \sum_{r=1}^{r_0} 2^{s_r - 1}$ ;                     ▷ $R \in \{1, \dots, 2^{r_0}\}$
        $Z \leftarrow (s_{r_0+1} s_{r_0+2} \dots s_{r_0+z_0})$;
        $X \leftarrow \{\inf n \geq 1 \colon s_{r_0+z_0+n} = 1\}$ ;    ▷ $P(X + r_0 + z_0 >$ length of hash$) \ll 1$
        if $X > \mathbb{X}_{R\,c}$ then
            $\mathbb{X}_{R\,c} \leftarrow X$;
            $\mathbb{Z}_{R\,c} \leftarrow Z$;
        else if $X = \mathbb{X}_{R\,c}$ then
            $\mathbb{Z}_{R\,c} \leftarrow \min_2(Z, \mathbb{Z}_{R\,c})$ ;        ▷ $\min_2$ is the minimum in base 2
    end
    discharge $o$;
end
```

Algorithm 1: Streaming algorithm to store the data in memory. $\mathbb{X}$ is an integer-valued matrix, whose values are of the order of $\log_2(F_0)$, while $\mathbb{Z}$ takes values in $1, \dots, 2^{z_0}$

```
Input: $\mathbb{X}$ and $\mathbb{Z}$, output of Algorithm 1
Output: $\mathbb{Y} = \{Y_{r\,c}, r = 1, \dots, 2^{r_0}, c = 1, \dots, c_0\}$
Set $\tilde{Y} = 0$;
for $c \leftarrow 1$ to $c_0$ do
    for $r \leftarrow 1$ to $2^{r_0}$ do
        $(b_1 b_2 \dots b_{z_0}) \leftarrow \mathbb{Z}_{r\,c}$;                 ▷ made by $z_0$ bits
        $y \leftarrow \sum_{j=1}^{z_0} b_j 2^{-j}$ ;       ▷ $y \in [0, 1 - 2^{-z_0}] \ \Rightarrow\ (1 + y) \in [1, 2)$
        $Y_{r\,c} \leftarrow \mathbb{X}_{r\,c} - \log_2(1 + y)$ ;  ▷ $\mathbb{X}_{r\,c} - \log_2(1 + y) \in (\mathbb{X}_{r\,c} - 1, \mathbb{X}_{r\,c}]$
    end
end
return $\mathbb{Y} = (Y_{r\,c})_{r=1,\dots,2^{r_0}, c=1,\dots,c_0}$;
```

Algorithm 2: Querying algorithm to extract $\mathbb{Y}$, starting from the memory content $\mathbb{X}$ and $\mathbb{Z}$ given in Algorithm 1

The querying algorithm produces the value $\tilde{Y}$, which is the arithmetic mean of $a_0 = c_0 2^{r_0}$ values built with the contents of $\mathbb{X}$ and $\mathbb{Z}$ as in Algorithm 2. As an example, in Algorithm 3, we show how to compute a $1 - \mathfrak{p}_-$-confidence interval for F_0 of the form $(0, \text{upper})$, based on the Theorem 3.1. The nonlinear problems involved in this computation will be faced in Section 7.

Finally, note that the data structure becomes that of [12] when $c_0 = 1$ and $z_0 = 0$ (the content of $\mathbb{Z}$ is not significant and the update reduces to $\mathbb{X}_{R\,c} \leftarrow \max(X, \mathbb{X}_{R\,c})$, without the if-else loop). When, in addition, $r_0 = 0$ the data structure reduces to the original one [15].

2.1. **Mathematical and Statistical analysis of the algorithm.** The Algorithm 1 has the following properties. First, the multiple application of this algorithm to the same object will result in the same outcome as if we had applied it only once. Mathematically speaking, this is a idempotent algorithm and, in addition, it can be seen to be associative and commutative. A typical mathematical function with these properties is the max function that, evaluated on different, even repeated numbers, gives the same result, independently of the order and of the repetitions.

```
Input: 1) $\mathbb{Y} = \{Y_{r\,c}, r = 1, \dots, 2^{r_0}, c = 1, \dots, c_0\}$, output of Algorithm 2.
2) the confidence $\alpha \in (0,1)$ -usually $\alpha \in [0.9, 0.995]$-
Output: A $\alpha$ confidence interval for $F_0$ of the form $(0, \text{upper})$
Set $p_- = 1 - \alpha$;
Set $y = -\log(\mathfrak{p}_-)/(2^{r_0} c_0)$;
Set $x \leftarrow InvAlphaMinus(y)$ ; /* Solve (in $x$) the problem $y - ((x-\gamma)t_- - \ln(\Gamma(1+t_-))) = 0$, with
  $\psi(1+t_-) = x - \gamma$ */
Set $\hat{y} \leftarrow 0$;
for $c \leftarrow 1$ to $c_0$ do
    for $r \leftarrow 1$ to $2^{r_0}$ do
        $\hat{y} \leftarrow \hat{y} + Y_{r\,c}$.
    end
end
$\bar{y} \leftarrow \hat{y}/(2^{r_0} c_0)$;
Set $z \leftarrow \log(2)\bar{y} + x + 2^{-z_0}$;
Set $p_0 \leftarrow 2^{-r_0}$;
return upper $= invHpM(z, p_0)$ ;                    ▷ Solve (in $x$) the problem $z - \mathbb{h}_{p_0}(x) = 0$
```

Algorithm 3: Querying algorithm that builds a $1-\mathfrak{p}_-$-confidence interval for F_0 of the form $(0, \text{upper})$, based on the Theorem 3.1

This is the reason why this algorithm works and why, for what concerns the final result of the matrices, the *Algorithm 1 may be thought as applied only once to each of the F_0 different objects.*

We will assume that each hash function generates a sequence of bits that are equally distributed on the all possible outcomes. Moreover, the evaluation on different objects are assumed to be statistically independent as for the evaluation of different functions. As only an example, the SHA functions have been certified to have such a properties [20, 21, 22], and can be used for this purpose: by cutting the result of a SHA_{512} function into 4 parts, it is possible to obtain 4 independent hash functions of sufficient length for any reasonable application.

From a probabilistic point of view, the bit sequences of (2) are independent for different choice of the object o and hash function c, and are uniformly distributed on all the possible sequences.

In other words, every s_i in each sequence of the form

$$H_c(o) = \quad \underbrace{\overset{r_0 \text{ bits}}{s_1 s_2 \cdots s_{r_0}}}_{R=1+\sum_{r=1}^{r_0} 2^{s_r-1}} \quad \underbrace{\overset{z_0 \text{ bits}}{s_{r_0+1} s_{r_0+2} \cdots s_{r_0+z_0}}}_{Z=(s_{r_0+1} \dots s_{r_0+z_0})} \; \underbrace{\overset{X \text{ bits}}{s_{r_0+z_0+n} \cdots s_{r_0+z_0+X}}}_{X=\{\inf n\geq 1:\, s_{r_0+z_0+n}=1\}} \; \underbrace{s_{r_0+z_0+X+1} \cdots}_{\text{not used}}$$

is distributed as a Bernoulli of parameter $1/2$, and it is independent from the others.

If we analyze the Algorithm 2 we note that, for each index (r,c) of the matrices, the unique information that is kept after querying may be in the following manner. For each fixed hash function c and object o, just compute R, X, and Z as in (2), and $Y(o,c)$ is defined as

$$Y(o,c) = X - \log_2\Big(1 + \sum_{z=1}^{z_0} b_z 2^{-z}\Big) = X - \log_2\Big(1 + \sum_{z=1}^{z_0} s_{r_0+z} 2^{-z}\Big).$$

The successive quantity $Y_{r\,c}$ in Algorithm 2 is the result of the following procedure

$$Y_{r\,c} = \max_{\{o:\, R=r\}} (Y(o,c)). \tag{3}$$

Note that, if we complete the bit sequence $(b_1 \dots b_{z_0})$ in Z with an i.i.d. sequence of equally distributed bits $(b_{z_0+1} b_{z_0+2} \dots)$, the random variables

$$\bar{Y}(o,c) = X - \log_2\Big(1 + \sum_{z=1}^{\infty} b_z 2^{-z}\Big), \qquad o \in \{F_0 \text{ different objects}\}, c \in \{1, \dots, c_0\}, \tag{4}$$

form a family of random variables, independent and identically distributed. The fact here is that, instead of measuring $\bar{Y}(o,c)$, we can only collect $Y(o,c)$, due to computational limitations, and this introduces a further bias. We get the following result.

Lemma 2.1. *There exists a family*

$$\big\{\bar{Y}(o,c), o \in \{F_0 \textit{ different objects}\}, c \in \{1,\dots,c_0\}\big\}$$

of independent and identically distributed random variables with exponential distribution of parameter $\lambda_0 = \log 2$*, such that, if we define,*

$$\bar{Y}_{r\,c} = \max_{\{o\colon R(o,c)=r\}} (\bar{Y}(o,c)),$$

then, uniformly in r *and* c*,*

$$0 \le Y_{r\,c} - \bar{Y}_{r\,c} \le 2^{-z_0},$$

where each $Y_{r\,c}$ *is defined in* (3)*. Moreover, for any fixed* $c \in \{1,\dots,c_0\}$*, define*

$$m_{r\,c} = \#\big\{o \in \{F_0 \textit{ different objects}\}\colon R(o,c) = r\big\}.$$

Then the random vectors $\{\boldsymbol{m}_c = (m_{1\,c},\dots,m_{2^{r_0}\,c}), c = 1,\dots,c_0\}$ *are i.i.d, distributed as multinomial vectors of parameters* F_0 *and* 2^{-r_0}*. Conditioned on* $\boldsymbol{m}_c$*, the random variables* $(\bar{Y}_{r\,c})_{r,c}$ *are independent.*

Proof. Take $(\bar{Y}(o,c))_{o,c}$ as in (4). Define

$$\begin{aligned}\bar{U}(o,c) = 2^{-\bar{Y}(o,c)} &= 2^{-X}\big(1 + \sum_{z=1}^{\infty} b_z 2^{-z}\big) = 2^{-X} + 2^{-X}\sum_{z=1}^{\infty} b_z 2^{-z}\\ &= \sum_{x=1}^{X} s_{r_0+x}2^{-x} + \sum_{z=1}^{\infty} b_z 2^{-z+X}, \qquad o \in \{F_0 \text{ different objects}\}, c \in \{1,\dots,c_0\};\end{aligned}$$

we note that it forms a family of random variables, independent and uniformly distributed on $(0,1)$ (see, e.g., [26, § 4.6]). Since $\bar{Y}(o,c) = -\log_2(\bar{U}(o,c)) = -\frac{\log(\bar{U}(o,c))}{\lambda_0}$, the first part of the lemma holds. In addition,

$$Y(o,c) - \bar{Y}(o,c) = \log_2\Big(1 + \frac{2^{-z_0}\sum_{z=1}^{\infty} b_{z_0+z}2^{-z}}{1 + \sum_{z=1}^{z_0} b_z 2^{-z}}\Big).$$

Since $\sum_{z=1}^{\infty} b_{z_0+z}2^{-z} \in [0,1]$ and $1 + \sum_{z=1}^{z_0} b_z 2^{-z} \ge 1$, then $0 \le Y(o,c) - \bar{Y}(o,c) \le 2^{-z_0}$. Hence we get the second part of the thesis, since

$$\max_{\{o\colon R(o,c)=r\}} (Y(o,c) - \bar{Y}(o,c)) = Y_{r\,c} - \bar{Y}_{r\,c}, \qquad \text{for any } r, c.$$

To conclude, just note that the first r_0 bits of each hash function generate R, uniformly distributed on $1,\dots,2^{r_0}$, independently of the remaining processes. Them, for each one of the F_0 different objects and each hash function, a uniformly assignment R is made, that gives the multinomial sample. The conditional independence of the family $(\bar{Y}_{r\,c})_{r,c}$ is a consequence of the independence of the family $(\bar{Y}(o,c))_{o,c}$. □

3. Confidence interval for F_0

The main result of this paper is the construction of a confidence interval for F_0.

Theorem 3.1. *Let* $\mathbb{Y}$ *be collected as in Section 2, and define*

$$\mathcal{Y} = \frac{\sum_{r=1}^{2^{r_0}}\sum_{c=1}^{c_0} Y_{r\,c}}{2^{r_0}c_0}.$$

Then

$$\begin{aligned} \mathbb{h}_{p_0}^{-1}(\lambda_0\mathcal{Y} - h_d) <&F_0 \\ &F_0 < \mathbb{h}_{p_0}^{-1}(\lambda_0\mathcal{Y} + h_u + 2^{-z_0}) \\ \mathbb{h}_{p_0}^{-1}(\lambda_0\mathcal{Y} - h_d) <&F_0 < \mathbb{h}_{p_0}^{-1}(\lambda_0\mathcal{Y} + h_u + 2^{-z_0}) \end{aligned}$$

are confidence intervals for the unknown parameter F_0, where

- *the function $\mathbb{h}_p$ is defined in Definition A.1 and* (19)*;*
- *$p_0 = 2^{-r_0}, \lambda_0 = \log(2)$;*
- *the levels of confidence are $1 - \mathfrak{p}_+$, $1 - \mathfrak{p}_-$, and $1 - (\mathfrak{p}_+ + \mathfrak{p}_-)$ respectively, where*

  $$\mathfrak{p}_+ = \exp\Big(-2^{r_0}c_0\big[(h_d+\gamma)t_+ - \ln\Gamma(1-t_+)\big]\Big), \qquad t_+ = 1 - \psi^{-1}(-h_d-\gamma);$$
  $$\mathfrak{p}_- = \exp\Big(-2^{r_0}c_0\big[(h_u-\gamma)t_- - \ln\Gamma(1+t_-)\big]\Big), \qquad t_- = \psi^{-1}(h_u-\gamma) - 1;$$

  γ is the Euler constant and ψ is the digamma function (see Appendix A).

Proof of Theorem 3.1. We first note that, by Lemma 2.1, if we define

$$\bar{\mathcal{Y}} = \frac{\sum_{r=1}^{2^{r_0}}\sum_{c=1}^{c_0}\bar{Y}_{r\,c}}{2^{r_0}c_0}, \tag{5}$$

then it is sufficient to prove that

$$\begin{aligned} \mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} - h_d) <&F_0 \\ &F_0 < \mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} + h_u) \\ \mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} - h_d) <&F_0 < \mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} + h_u) \end{aligned}$$

are confidence intervals for the unknown parameter F_0 at the same levels given in the theorem. To prove this last assertion, we prove the following conditions that result sufficient:

$$P\Big(\mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} - h_d) \geq F_0\Big) \leq \mathfrak{p}_+\,;$$
$$P\Big(\mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} + h_u) \leq F_0\Big) \leq \mathfrak{p}_-\,.$$

Observe that, since the function $\mathbb{h}_{p_0}$ is invertible with continuous inverse (see Section A), we get

$$P\Big(\mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} - h_d) \geq F_0\Big) = P\Big(\bar{\mathcal{Y}} \geq \frac{\mathbb{h}_{p_0}(F_0)+h_d}{\lambda_0}\Big)\,;$$
$$P\Big(\mathbb{h}_{p_0}^{-1}(\lambda_0\bar{\mathcal{Y}} + h_u) \leq F_0\Big) = P\Big(\bar{\mathcal{Y}} \leq \frac{\mathbb{h}_{p_0}(F_0)-h_u}{\lambda_0}\Big)\,;$$

and hence the proof will be based on the following step:

- in Lemma 5.1 in Section 5 we prove that $E(\bar{Y}_{r\,c}) = \frac{\mathbb{h}_{p_0}(F_0)}{\lambda_0}$, for any $\bar{Y}_{r\,c}$. As an imediate consequence the following equality holds

  $$E(\bar{\mathcal{Y}}) = \frac{\mathbb{h}_{p_0}(F_0)}{\lambda_0};$$

- the inequalities

  $$P\Big(\bar{\mathcal{Y}} \geq E(\bar{\mathcal{Y}}) + \frac{h_d}{\lambda_0}\Big) \leq \mathfrak{p}_+;$$
  $$P\Big(\bar{\mathcal{Y}} \leq E(\bar{\mathcal{Y}}) - \frac{h_u}{\lambda_0}\Big) \leq \mathfrak{p}_-$$

  are Chernoff bound inequalities and will be proved in Corollary 5.2. □

3.1. **Connection with extreme value theory.** As stated in Lemma 2.1, the main result of this paper is based on the mean of the random variables $(\bar{Y}_{r\,c})_{r,c}$, which are independent, conditioned on $\boldsymbol{m}_c$. As discussed in the introduction, this variables are given through a commutative, associative and idempotent function, that is the max function in this context:

$$\bar{Y}_{r\,c} = \max_{\{o\colon R(o,c)=r\}} (\bar{Y}(o,c)).$$

A natural question is the relation of such a consideration with the extreme value theory. The well-known Fisher–Tippett–Gnedenko theorem [17] provides an asymptotic result, and shows that, when $F_0 \to \infty$, if there are sequences a_{F_0} and b_{F_0} such that $(\bar{Y}_{r\,c} - a_{F_0})/b_{F_0}$ converges in law to a random variables Z, then Z must be Gumbel, Fréchet or Weibull (Type 1,2 or 3). As in the proof of Lemma 4.1, we have that

$$E(e^{s(\bar{Y}_{r\,c}-E(\bar{Y}_{r\,c}))}) \underset{F_0\to\infty}{\longrightarrow} \left(\Gamma(1-\tfrac{s}{\lambda_0})e^{-\gamma\frac{s}{\lambda_0}}\right) = E(e^Z),$$

from which we can recognize that Z has a Gumbell law. Since the Chernoff bounds on the mean of such variables gives the same concentration inequalities as in Theorem 3.1, our result gives also the confidence interval based on the Chernoff bounds of the asymptotic distribution based on the extreme value theory.

Our result underlines the fact that the analytical approximation gives an exact *upper bound* for the concentration inequality, based on the monotonicity of the limit $E(e^{s(\bar{Y}_{r\,c}-E(\bar{Y}_{r\,c}))}) \nearrow E(e^Z)$, that is the key point in the proof of Lemma 4.1.

Finally, the accuracy of such a bound is discussed in Section 6.

4. Chernoff bounds: auxiliary results for the maximum of exponential random variables

We recall the Chernoff bound of a sum $X = X_1 + \cdots + X_{a_0}$ of independendent random variables $X_1, \ldots, X_{a_0}$: for any $s \in \mathbb{R}$,

$$P(X \geq s) \leq \min_{t>0} e^{-ts} \prod_{i=1}^{a_0} E(e^{tX_i}), \tag{6}$$

which is one of the most powerful concentration inequality in probability theory, since it involves the entire moment generating functions $E(e^{tX_i})$ instead of only some moments of each X_i.

Lemma 4.1. *Let A be a finite set of cardinality a_0, and let $(m_a)_{a\in A}$ be a collection of nonnegative integer numbers. Let $\{X_{a\,j}, a \in A, j \leq m_j\}$ be an array of i.i.d. exponential random variables with parameter λ. Define, for any $a \in A$, $Y_a = \max_j X_{a\,j}$ and $Y = \frac{\sum_{a\in A} Y_a}{a_0}$. Then*

$$\sup_{(m_a)_{a\in A}} P\big(Y \geq E(Y) + \tfrac{h_d}{\lambda}\big) \leq \mathfrak{p}_+; \qquad \sup_{(m_a)_{a\in A}} P\big(Y \leq E(Y) - \tfrac{h_u}{\lambda}\big) \leq \mathfrak{p}_-; \tag{7}$$

where $\mathfrak{p}_\pm$ are defined in Theorem 3.1.

Proof. To apply (6) with $X = Y$ and $X_a = \frac{Y_a}{a_0}$, it is possible in principle to compute $E(e^{tY_a})$ by noticing that the density of Y_a may be expressed as the density of the maximum of m_a independent exponential random variables:

$$f_{Y_a}(y) = \frac{d}{dy}(1-(1-e^{-\lambda y})^{m_a}) = \sum_{j=1}^{m_a} \binom{m_a}{j} (-1)^{j-1} \lambda j e^{-\lambda j y}.$$

A more interesting interpretation leads to simpler computations. Denote by $X_{a(j)}$ is the jth-order statistic of $(X_{a\,1}, \ldots, X_{a\,m_a})$, set $X_{a(0)} = 0$ for consistency. As noted for example recently in [11, Eq. (2)], for any $i = 1, \dots, K$, the random variables

$$X_{a(1)} - X_{a(0)}, X_{a(2)} - X_{a(1)}, \dots, X_{a(m_a)} - X_{a(m_a-1)},$$

are independent exponential random variables with parameter $\{\lambda, 2\lambda, \dots, m_a\lambda\}$. Since

$$\begin{aligned} Y_a = \max(X_{a\,1}, \dots, X_{a\,m_a}) &= X_{a(m_a)} \\ &= (X_{a(1)} - X_{a(0)}) + (X_{a(2)} - X_{a(1)}) + \cdots + (X_{a(m_a)} - X_{a(m_a-1)}), \end{aligned}$$

then each Y_a may be seen as a sum of m_a independent exponential random variables with parameter λj, $j = 1, \dots, m_a$. As a direct consequence,

$$E(Y_a) = \sum_{j=1}^{m_a} \frac{1}{\lambda j} = \frac{\mathbb{h}(m_a)}{\lambda}, \tag{8}$$

where $\mathbb{h}(m_a)$ is the m_a-th harmonic number defined in (18). More remarkable, it is possible to calculate the moment-generating function of Y. In fact, since

$$E(e^{s(X_{a(j)} - X_{a(j-1)})}) = (1 - \tfrac{s}{\lambda j})^{-1} \implies E(e^{sY_a}) = \prod_{j=1}^{m_a} (1 - \tfrac{s}{\lambda j})^{-1}, \qquad 0 < s < \lambda,$$

we get, for $0 < s < a_0\lambda$,

$$E(e^{sY}) = E\Big(e^{s\sum_{a\in A}\frac{Y_a}{a_0}}\Big) = \prod_{a\in A}\prod_{j=1}^{m_a} (1 - \tfrac{s}{\lambda a_0 j})^{-1}.$$

Thus, since $E(Y) = \sum_{a\in A} \frac{\sum_{j=1}^{m_a} \frac{1}{\lambda j}}{a_0} = \sum_{a\in A}\sum_{j=1}^{m_a} \frac{1}{\lambda a_0 j}$, the Chernoff bound (6) becomes

$$\begin{aligned} P(Y \geq E(Y) + \tfrac{h_d}{\lambda}) &\leq \min_{s>0} e^{-s(E(Y)+\frac{h_d}{\lambda})} \prod_{a\in A}\prod_{j=1}^{m_a} (1 - \tfrac{s}{\lambda a_0 j})^{-1} \\ &= \min_{s>0} e^{-\frac{h_d}{\lambda}s} \prod_{a\in A}\prod_{j=1}^{m_a} \frac{e^{-\frac{s}{\lambda a_0 j}}}{1 - \frac{s}{\lambda a_0 j}}. \end{aligned}$$

Since $\frac{\exp^{-x}}{1-x} \geq 1$ for any $x < 1$, then for $t = \frac{s}{\lambda a_0} \in (0,1)$, by (21)

$$\prod_{a\in A}\prod_{j=1}^{m_a} \frac{e^{-\frac{s}{\lambda a_0 j}}}{1 - \frac{s}{\lambda a_0 j}} = \prod_{a\in A}\prod_{j=1}^{m_a} \frac{e^{-\frac{t}{j}}}{1 - \frac{t}{j}} \leq \prod_{a\in A}\Big(\prod_{j=1}^{\infty} \frac{e^{-\frac{t}{j}}}{1 - \frac{t}{j}}\Big) = \Big(\Gamma(1-t)e^{-\gamma t}\Big)^{a_0}.$$

Combining the two expressions above, we get

$$\begin{aligned} P(Y \geq E(Y) + \tfrac{h_d}{\lambda}) \leq \min_{t\in(0,1)} e^{-h_d t a_0}\Big(\Gamma(1-t)e^{-\gamma t}\Big)^{a_0} &= \min_{t\in(0,1)} \Big(\Gamma(1-t)e^{-(\gamma+h_d)t}\Big)^{a_0} \\ &= \exp\Big(-a_0 \max_{t\in(0,1)} \big[(h_d+\gamma)t - \ln\Gamma(1-t)\big]\Big). \end{aligned}$$

The part of the proof that concerns $\mathfrak{p}_+$ is hence proved by Lemma A.2.

The second inequality in (7) may be proved with the same spirit. To find the Chernoff bound with the second equality of (21), note that we get, for any $t = \frac{s}{\lambda a_0} > 0$

$$\begin{aligned} P(Y \leq E(Y) - \tfrac{h_u}{\lambda}) &= P(-Y \geq -E(Y) + \tfrac{h_u}{\lambda}) \\ &\leq \min_{s>0} e^{s(E(Y) - \frac{h_u}{\lambda})} \prod_{a\in A} \prod_{j=1}^{m_a} (1 + \tfrac{s}{\lambda a_0 j})^{-1} \\ &= \min_{s>0} e^{-\frac{h_u}{\lambda} s} \prod_{a\in A} \prod_{j=1}^{m_a} \frac{e^{\frac{s}{\lambda a_0 j}}}{1 + \frac{s}{\lambda a_0 j}} \\ &\leq \min_{t>0} \Big(\Gamma(1+t) e^{(\gamma - h_u)t} \Big)^{a_0}, \end{aligned}$$

and then we apply again Lemma A.2 to $g_-(t) = (x-\gamma)t - \ln \Gamma(1+t)$. □

5. Computation of $E(Y)$

To complete the computation of the confidence interval, we give the following result, which connects the expectation of the core variables with the special functions we have introduced in this paper.

Lemma 5.1. *For any $r = 1, \ldots, 2^{r_0}$ and $c = 1, \ldots, c_0$, we have that*

$$E(\bar{Y}_{r\,c}) = \frac{\mathbb{h}_{p_0}(F_0)}{\lambda_0}$$

Proof of Lemma 5.1. Let $\{\boldsymbol{m}_c = (m_{1\,c}, \ldots, m_{2^{r_0}\,c}), c = 1, \ldots, c_0\}$ as in Lemma 2.1. Combining Lemma 2.1 and (8), we know that

$$\lambda_0 E(Y_{r\,c} | \{\boldsymbol{m}_c, c = 1, \ldots, c_0\}) = \mathbb{h}(m_{r\,c}).$$

Again, as stated in Lemma 2.1, the random variable $m_{r\,c}$ is distributed as a binomial distribution, with F_0 trials and probability $p_0 = 2^{-r_0}$. Then, by (18),

$$\begin{aligned} \lambda_0 E(Y_{r\,c}) &= \lambda_0 E(E(Y_{r\,c} | \{\boldsymbol{m}_c, c = 1, \ldots, c_0\})) = E(\mathbb{h}(m_{r\,c})) \\ &= \sum_{m=0}^{F_0} \mathbb{h}(m) \binom{F_0}{m} {p_0}^m (1-p_0)^{F_0 - m} \\ &= \sum_{m=0}^{F_0} \Big(\int_0^1 \frac{1-x^m}{1-x}\, dx \Big) \binom{F_0}{m} {p_0}^m (1-p_0)^{F_0 - m} \\ &= \int_0^1 \frac{1}{1-x} \Big(\sum_{m=0}^{F_0} (1-x^m) \binom{F_0}{m} {p_0}^m (1-p_0)^{F_0-m} \Big) dx \\ &= \int_0^1 \frac{1}{1-x} \Big(\sum_{m=0}^{F_0} \binom{F_0}{m} {p_0}^m (1-p_0)^{F_0-m} \\ &\qquad - \sum_{m=0}^{F_0} \binom{F_0}{m} (p_0 x)^m (1-p_0)^{F_0-m} \Big) dx \\ &= \int_0^1 \frac{1 - (1 - p_0 + p_0 x)^{F_0}}{1-x}\, dx = \mathbb{h}_{p_0}(F_0), \end{aligned}$$

the last equality being the Definition A.1. □

Corollary 5.2. *Let $\bar{\mathcal{Y}}$ as in* (5)*. The following inequalities hold*

$$P\Big(\bar{\mathcal{Y}} \geq E(\bar{\mathcal{Y}}) + \frac{h_d}{\lambda_0}\Big) \leq \mathfrak{p}_+;$$
$$P\Big(\bar{\mathcal{Y}} \leq E(\bar{\mathcal{Y}}) - \frac{h_u}{\lambda_0}\Big) \leq \mathfrak{p}_-.$$

Proof. To prove the assertion, we apply Lemma 4.1 at the objects given in Lemma 2.1. We begin by setting $A = \{(r,c), r = 1,\dots,2^{r_0}, c = 1,\dots,c_0\}$, which implies $a_0 = 2^{r_0}c_0$. Moreover, for $a = (r,c)$, we have $m_a = m_{r\,c}$ and

$$\{X_{a\,j}, j \leq m_a\} = \{Y(o,c) \colon R(o,c) = r\}.$$

With this setting, λ in Lemma 4.1 is $\lambda_0 = \log(2)$ and Y is exactly $\bar{\mathcal{Y}}$. The thesis follows. □

6. Analytical asymptotic discussion

In this section we discuss the accuracy of the analytical approximation given in the main result to show the appropriateness in this context.

We could find a (m_a)-uniform bound in Lemma 4.1 with the following inequalities:

$$\begin{aligned}
&\text{for } \mathfrak{p}_+ : \quad \prod_{j=1}^{m_a} \frac{e^{-\frac{t}{j}}}{1-\frac{t}{j}} \leq \Big(\prod_{j=1}^{\infty} \frac{e^{-\frac{t}{j}}}{1-\frac{t}{j}}\Big) = \Gamma(1-t)e^{-\gamma t}, \quad t \in (0,1);\\
&\text{for } \mathfrak{p}_- : \quad \prod_{j=1}^{m_a} \frac{e^{\frac{t}{j}}}{1+\frac{t}{j}} \leq \Big(\prod_{j=1}^{\infty} \frac{e^{\frac{t}{j}}}{1+\frac{t}{j}}\Big) = \Gamma(1+t)e^{\gamma t}, \qquad t > 0.
\end{aligned} \tag{9}$$

We recall that, in our context,

$$m_a = m_{r\,c} = \#\{Y(o,c) \colon R(o,c) = r\},$$

is the (random) number of object assigned to register r by the hash function c. In Figure 1 we underline that this approximation is good for small values of t and big m_a. To show that the

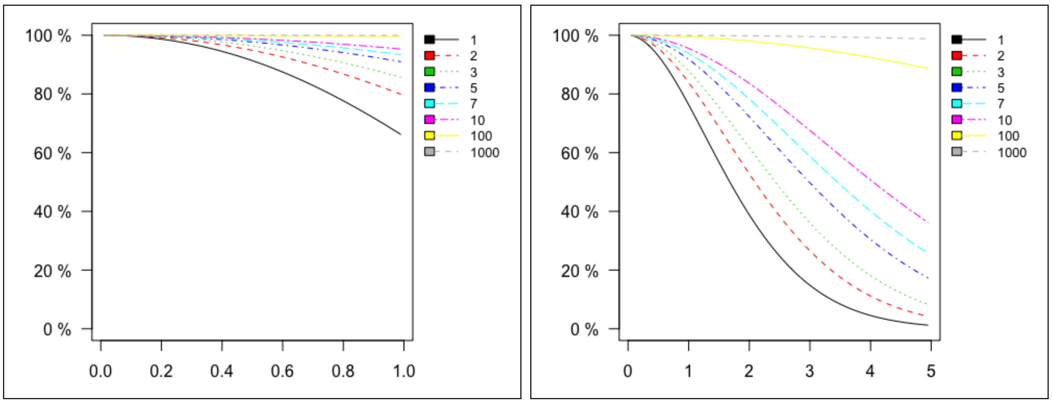

Figure 1. Ratio between the finite products and the series quantities given in (9), for different values of m_a and t, expressed as percentage of $\Gamma(1 \mp t)e^{\mp\gamma t}$ given by $\prod_1^{m_a} \frac{e^{\mp\frac{t}{j}}}{1\mp\frac{t}{j}}$. The different lines refer to different values of m_a, given in the legend. Left: percentage of approximation for $\Gamma(1-t)e^{-\gamma t}$, $t \in (0,1)$. Right: percentage of approximation for $\Gamma(1+t)e^{+\gamma t}$, $t \in (0,5)$.

uniform bound in this paper does not affect the Chernoff bounds of the $\mathfrak{p}$-values, we compare for different values of x_u and x_d:

$$\begin{aligned}
&\text{for } \mathfrak{p}_+ : \quad \min_{t\in(0,1)} \Big(\prod_{c=1}^{c_0} \prod_{r=1}^{2^{r_0}} e^{-tx_d} \prod_{j=1}^{m_{rc}} \frac{e^{-\frac{t}{j}}}{1-\frac{t}{j}} \Big) \quad \text{vs.} \quad \Big(\Gamma(1-t_+) e^{-(\gamma+x_d)t_+} \Big)^{c_0 2^{r_0}} ; \\
(10) \qquad &\text{for } \mathfrak{p}_- : \quad \min_{t>0} \Big(\prod_{c=1}^{c_0} \prod_{r=1}^{2^{r_0}} e^{-tx_u} \prod_{j=1}^{m_{rc}} \frac{e^{\frac{t}{j}}}{1+\frac{t}{j}} \Big) \quad \text{vs.} \quad \Big(\Gamma(1+t_-) e^{(\gamma-x_u)t_-} \Big)^{c_0 2^{r_0}} .
\end{aligned}$$

For $r_0 \in \{0, \dots, 4\}$, $c_0 \in \{1, \dots, 4\}$, and $\alpha \in \{.9, .95, .975, .99\}$, we choose the values of x_u and x_d for which

$$\Big(\Gamma(1-t_+) e^{-(\gamma+x_d)t_+} \Big)^{c_0 2^{r_0}} = \mathfrak{p}_+ = 1-\alpha = \mathfrak{p}_- = \Big(\Gamma(1+t_-) e^{(\gamma-x_u)t_-} \Big)^{c_0 2^{r_0}} .$$

Then, for any $F_0 \in \{50, 100, 500, 1000, 5000, 10000, 50000, 100000\}$, with a MonteCarlo procedure, we estimate the mean value and the standard deviation of the quantities

$$\mathfrak{P}_- = \min_{t\in(0,1)} \Big(\prod_{c=1}^{c_0} \prod_{r=1}^{2^{r_0}} e^{-tx_d} \prod_{j=1}^{m_{rc}} \frac{e^{-\frac{t}{j}}}{1-\frac{t}{j}} \Big) \qquad \text{and} \qquad \mathfrak{P}_+ = \min_{t>0} \Big(\prod_{c=1}^{c_0} \prod_{r=1}^{2^{r_0}} e^{-tx_u} \prod_{j=1}^{m_{rc}} \frac{e^{\frac{t}{j}}}{1+\frac{t}{j}} \Big)$$

by simulating different values of the multinomial vectors $\{\boldsymbol{m}_c, c = 1, \dots, c_0\}$, and, as expected, all the simulated quantities above results smaller than $\mathfrak{p} = 1 - \alpha$. Then, for each r_0, c_0, α, F_0 we have built a 3σ confidence interval $[\mathfrak{p}_-^l, \mathfrak{p}_-^u]$ and $[\mathfrak{p}_+^l, \mathfrak{p}_+^u]$ for $\mathfrak{P}_-$ and $\mathfrak{P}_+$, respectively. All the data are presented in Figure 2. On the left-hand side , it is drawn the scatter-plot of

$$\begin{aligned}
x &= \text{range of confidence interval} && = \mathfrak{p}_+^u - \mathfrak{p}_+^l && (\mathfrak{p}_-^u - \mathfrak{p}_-^l, \text{ respectively}); \\
y &= \text{maximum imprecision} && = \mathfrak{p}_+ - \mathfrak{p}_+^l && (\mathfrak{p}_- - \mathfrak{p}_-^l, \text{ respectively});
\end{aligned}$$

which shows a good linear dependence in a log-log scale. As the linear coefficient is close to 2, on the right -hand side, the scatterplot of y/x^2 vs. x confirms this scale of dependence, and it suggests that the variability of the constant depends mainly on $\mathfrak{p}_\pm$, firstly on the choice of the sign, and then on the $\mathfrak{p}$ value.

A finer analysis shows that, when $F_0 \geq 500$, the maximum imprecision is less than 0.00683 (with $r_0 = 4$, $c = 1$, $p_- = 0.1$, $N_0 = 500$), becoming less than $6.7 \cdot 10^{-5}$ for $F_0 \geq 50000$ (again, $r_0 = 4$, $c = 1$, $p_- = 0.1$ but $N_0 = 50000$). In other words, the uniform bounds given in (10) appear adequate in this context.

7. Computational aspects

As a consequence of Theorem 3.1, we may build confidence intervals for F_0 based on the output of of Algorithm 2. As an example, Algorithm 3 shows how to compute the confidence interval of the form $(0, \text{upper})$. Analogous procedures can be used to compute confidence intervals of other forms.

As underlined in the Algorithm 3, it is necessary to solve numerically some nonlinear equations of the form $f(x) = 0$ to find the final solution. In the following sections, we state the relevant inequalities that can be used to find the root of $f(x) = 0$ in our context, with the Halley's method [25]. This iterative method is given by

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2\big(f'(x_n)\big)^2 - f(x_n)f''(x_n)},$$

it is essentially the Newton method applied to the function $g(x) = \frac{f(x)}{\sqrt{|f'(x)|}}$, and it achieves a cubic rate of convergence in the neighborhood of the solution, see [4].

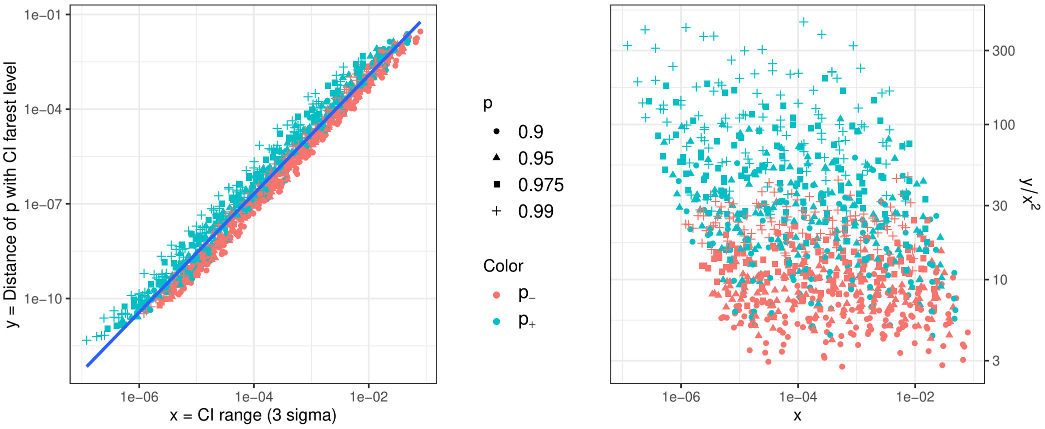

Figure 2. MonteCarlo simulation of the accuracy of the use of the analytical approximation in calculating the Chernoff bounds from $\mathfrak{P}$, with gamma function, as discussed in (10). Each point relates to a different choice of $\mathfrak{p}_+$ (light blue) or $\mathfrak{p}_-$ (light red), $r_0 \in \{0, \dots, 4\}$, $c_0 \in \{1, \dots, 4\}$ and $F_0 \in \{50, 100, 500, 1000, 5000, 10000, 50000, 100000\}$. Left: linear dependence in log-log scale ($y = 1.91 + 1.88x$) between the precision in using the exact formula (x is the length of the 3σ confidence interval of $\mathfrak{P}$) and the accuracy of the estimation of $\mathfrak{p}$ with gamma function instead of the exact formula (y is the distance between $\mathfrak{p}$ calculated with the gamma function and the farthest endpoint of the 3σ exact confidence interval). Rigth: dependence in log-log scale of y/x^2 with respect to x as function of different $\mathfrak{p}_\pm$

In addition, we give accurate lower and upper bounds for the solution, that can be shown to be contained in the basin of attraction of the solution. Note that these bounds can be used also with a much simpler and robust bisection method, which has on the counterpart only a linear rate of convergence.

7.1. **The problem $\psi(x) - y = 0$.** We recall here that the *digamma function* $\psi : (0, \infty) \to \mathbb{R}$ is defined as the logarithmic derivative of the Γ function, see [1, §6.3], and it satisfies the relation

$$\psi(x+1) = \psi(x) + \tfrac{1}{x}. \tag{11}$$

In addition ψ is a strictly monotone, concave function, with $\lim_{t \to 0^+} \psi(t) = -\infty$, $\psi(1) = -\gamma$ and $\psi(t) = \log(t) + o(1)$ when $t \to \infty$ (see, for example, [9]). Finally, it is implemented in all the recent math packages together with its first and second derivative functions ψ_1 and ψ_2.

As shown in Section B.1, we have

$$\ln(x - \tfrac{1}{2}) < y < \ln(x), \qquad e^y < x < e^y + \tfrac{1}{2}, \qquad \forall x > \tfrac{1}{2}, \forall y. \tag{12}$$

7.2. **The problem $\mathbb{h}_p(x) - y = 0$.** First note that $\mathbb{h}_p(x), \mathbb{h}'_p(x)$ and $\mathbb{h}''_p(x)$ may be computed with with arbitrary precision, because of (19) and (20) and the fact that a quad-double precision algorithm to calculate Lerch's transcendent of real arguments have been already developed, see [3].

For $p \in (0,1)$, as shown in Section B.2, we have

$$\frac{e^{y-\gamma}}{p} - \frac{1}{2} \geq x \geq \begin{cases} \frac{e^{y-\gamma}}{p} - e + \frac{1}{\ln(1-p)} & \text{if } y > \log\left(\gamma + p(\frac{1}{2} - \frac{1}{(e-1)\ln(1-p)})\right); \\ e^{y-\gamma} - 1 & \text{otherwise.} \end{cases} \tag{13}$$

7.3. **The problem $y = (x-\gamma)t(x) - \ln\Gamma(1+t(x))$, where $t(x) = \psi^{-1}(x-\gamma) - 1$.** Note that, if $g(x) = (x-\gamma)t(x) - \ln\Gamma(1+t(x))$, then

$$g'(x) = t(x) + t'(x)(x - \gamma - \psi(1+t(x))) = t(x), \tag{14}$$

since, by definition of $t(x)$, $\psi(1+t(x)) = x-\gamma$. Then the formula of the derivative of the inverse function gives

$$g''(x) = t'(x) = \frac{1}{\psi_1(\psi^{-1}(x-\gamma))} = \frac{1}{\psi_1(1+t(x))}.$$

As shown in Section B.3, we have

$$\begin{aligned} &\sqrt{\frac{1}{50}y} < x < \pi\sqrt{\frac{2}{3}y}, && \text{if } y < 3; \\ &\frac{2}{3}\Big(\log\Big(y+\frac{1}{2}\Big)+\gamma\Big) < x < 2\Big(\log\Big(\frac{4}{3}y+1\Big)+\gamma\Big), && \text{if } y \geq 3. \end{aligned} \tag{15}$$

7.4. **The problem $y = (x+\gamma)t(x) - \ln\Gamma(1-t(x))$, where $t(x) = 1-\psi^{-1}(-x-\gamma)$.** Note that, if $g(x) = (x+\gamma)t(x) - \ln\Gamma(1-t(x))$, then

$$g'(x) = t(x) + t'(x)(x + \gamma - \psi(1-t(x))) = t(x), \tag{16}$$

since, by definition of $t(x)$, $\psi(1-t(x)) = -x-\gamma$. Then the formula of the derivative of the inverse function gives

$$g''(x) = t'(x) = \frac{1}{\psi_1(\psi^{-1}(-x-\gamma))} = \frac{1}{\psi_1(1-t(x))}.$$

As shown in Section B.4, we have

$$\max\Big(-\ln(1-C)-\gamma, \frac{\pi^2}{6}C\Big) < x < 2\sqrt{(y+1)^2-1}, \tag{17}$$

where

$$C = \sqrt{1 - \frac{-(\frac{y}{2} - \frac{6+\pi^2}{12}) + \sqrt{(\frac{y}{2} - \frac{6+\pi^2}{12})^2 + 4\frac{18-\pi^2}{12}}}{2\frac{18-\pi^2}{12}}} \in (0,1).$$

7.5. **Minimum** log**-length interval.** In this section, we show how to numerically compute the minimum length interval, in log-scale, for a given confidence α, based on the inequalities given in the main result Theorem 3.1. The probem is set as follows: given $\alpha \in (0,1)$, $r_0 \geq 0$, $c_0 \geq 1$, we want to solve the nonlinear minimization problem:

$$\min(h_d + h_u)$$

subject to

$$\begin{cases} \mathfrak{p}_+ = \exp\Big(-2^{r_0}c_0\big[(h_d+\gamma)t_+ - \ln\Gamma(1-t_+)\big]\Big), & t_+ = 1-\psi^{-1}(-h_d-\gamma); \\ \mathfrak{p}_- = \exp\Big(-2^{r_0}c_0\big[(h_u-\gamma)t_- - \ln\Gamma(1+t_-)\big]\Big), & t_- = \psi^{-1}(h_u-\gamma)-1; \\ \mathfrak{p}_+ + \mathfrak{p}_- \leq (1-\alpha); \\ h_d, h_u \geq 0. \end{cases}$$

The two values $\mathfrak{p}_+$ and $\mathfrak{p}_-$ are monotone functions of h_d and h_u, respectively, as a consequence of (16) and (14). As a consequence, the minimum is attained when $\mathfrak{p}_+ + \mathfrak{p}_- = (1-\alpha)$. Then, denoting with $\mathfrak{p} = (1-\alpha)$, if we set $x = \mathfrak{p}_+$, we have $\mathfrak{p}_- = \mathfrak{p} - x$, and the problem above may be rewritten in terms of x: given $\mathfrak{p} \in (0,1)$, $a_0 = 2^{r_0}c_0 \in \{1,2,\dots\}$, find

$$\min(g(x)) = \min\big(y_+^{-1}(-\tfrac{\log x}{a_0}) + y_-^{-1}(-\tfrac{\log(\mathfrak{p}-x)}{a_0})\big)$$

subject to

$$\begin{cases} y_+(h) = (h+\gamma)t_+ - \ln\Gamma(1-t_+), & t_+ = 1 - \psi^{-1}(-h-\gamma); \\ y_-(h) = (h-\gamma)t_- - \ln\Gamma(1+t_-), & t_- = \psi^{-1}(h-\gamma) - 1; \\ 0 \le x \le \mathfrak{p}. \end{cases}$$

Differentiating g with respect to x, since $y'_\pm(h) = t_\pm(h)$ by (16) and (14), we obtain,

$$g'(x) = -\frac{1}{a_0 x}\frac{1}{t_+\Big(y_+^{-1}\Big(-\frac{\log x}{a_0}\Big)\Big)} + \frac{1}{a_0(\mathfrak{p}-x)}\frac{1}{t_-\Big(y_-^{-1}\Big(-\frac{\log(\mathfrak{p}-x)}{a_0}\Big)\Big)}$$

which is null when the following equation is zero

$$f(x) = x t_+\Big(y_+^{-1}\Big(-\frac{\log x}{a_0}\Big)\Big) - (\mathfrak{p}-x)t_-\Big(y_-^{-1}\Big(-\frac{\log(\mathfrak{p}-x)}{a_0}\Big)\Big)$$

Call

$$\hat{t}_+ = \hat{t}_+(x) = t_+\Big(y_+^{-1}\Big(-\frac{\log x}{a_0}\Big)\Big), \quad \hat{t}_- = \hat{t}_-(x) = t_-\Big(y_-^{-1}\Big(-\frac{\log(\mathfrak{p}-x)}{a_0}\Big)\Big),$$

$\psi_1(x) = d\frac{\psi(x)}{dx}$ and $\psi_2(x) = d\frac{\psi_1(x)}{dx}$, then

$$d\frac{\hat{t}_+(x)}{dx} = -\frac{1}{a_0 x}\frac{1}{\hat{t}_+\psi_1(1-\hat{t}_+)}, \qquad d\frac{\hat{t}_-(x)}{dx} = +\frac{1}{a_0(\mathfrak{p}-x)}\frac{1}{\hat{t}_-\psi_1(1+\hat{t}_-)}.$$

The problem is then to find the solution for the nonlinear problem $f(x) = 0$ that may be solved with the Halley's method that involves the problems seen above, noticing that

$$\begin{aligned} f(x) &= x\hat{t}_+ - (\mathfrak{p}-x)\hat{t}_-, \\ f'(x) &= \hat{t}_+ - \frac{1}{a_0\hat{t}_+\psi_1(1-\hat{t}_+)} + \hat{t}_- - \frac{1}{a_0\hat{t}_-\psi_1(1+\hat{t}_-)} \\ f''(x) &= t'_+\Big(1 + \frac{\psi_1(1-\hat{t}_+) - \hat{t}_+\psi_2(1-\hat{t}_+)}{a_0(\hat{t}_+\psi_1(1-\hat{t}_+))^2}\Big) \\ &\qquad + t'_-\Big(1 + \frac{\psi_1(1+\hat{t}_-) + \hat{t}_+\psi_2(1+\hat{t}_-)}{a_0(\hat{t}_-\psi_1(1+\hat{t}_-))^2}\Big). \end{aligned}$$

and that a good starting point is given by $x_0 = \frac{\mathfrak{p}}{2}$.

Appendix A. Special functions used in this paper

Modification of the harmonic numbers and Lerch transcendent function. For any integer number m, we denote by $\mathbb{h}(m)$ the m-th harmonic number. We recall here that

$$\mathbb{h}(m) = \psi(m+1) + \gamma = \sum_{j=1}^{m}\frac{1}{j} = \sum_{j=0}^{m-1}\int_0^1 t^j\,dt = \int_0^1 \frac{1-t^m}{1-t}\,dt, \tag{18}$$

where ψ is the derivative of the logarithm of gamma function (also called *digamma* function). The constant γ is the Euler–Mascheroni constant throughout the whole paper. The function $\mathbb{h}$ can be extended therefore to the real non-negative numbers, by setting $\mathbb{h}(x) = \int_0^1 \frac{1-t^x}{1-t}\,dt$, which is known as the integral representation given by Euler.

Definition A.1. For $0 \le p \le 1$, $y \ge 0$, we define the *p-modification of the harmonic numbers* $\mathbb{h}_p(x)$, where

$$\mathbb{h}_p(x) = \int_0^1 \frac{1-(1-p+pt)^x}{1-t}\,dt, \qquad x > 0.$$

The function $\mathbb{h}_p(x)$ has the following properties

- $\mathbb{h}_p(0) = 0$, $\mathbb{h}_0(x) = 0$, $\mathbb{h}_p(1) = p$ and $\mathbb{h}_1(x) = \mathbb{h}(x)$ by definition;
- with two changes of integration variable $z = (1 - p(1 - t))$ and $z = (1 - p)e^{-w}$, we we may rewrite $\mathbb{h}_p(y)$ as

$$\begin{aligned} \mathbb{h}_p(x) &= \int_{1-p}^{1} \frac{1 - z^x}{1 - z}\,dz = \psi(x + 1) + \gamma - \int_0^{1-p} \frac{1 - z^x}{1 - z}\,dz \\ &= \psi(x + 1) + \gamma + \log p + \int_0^{1-p} \frac{z^x}{1 - z}\,dz \\ &= \psi(x + 1) + \gamma + \log p + (1 - p)^{x+1} \int_0^{\infty} \frac{e^{-w(x+1)}}{1 - (1 - p)e^{-w}}\,dw \\ &= \psi(x + 1) + \gamma + \log p + (1 - p)^{x+1}\, \Phi(1 - p, 1, x + 1), \end{aligned} \tag{19}$$

  where Φ is the *Lerch transcendent function*, see [23], and the last equality is a consequence of the following equation, valid for $m \in \mathbb{N}$ and $z = (1 - p)$:

$$\Phi\left(z, s, a\right) = z^m \Phi\left(z, s, a + m\right) + \sum_{n=0}^{m-1} \frac{z^n}{(a + n)^s}.$$

- By (19), $\mathbb{h}_p(x)$ is strictly increasing and continuous, both as a function of x and p. In addition, for any $p > 0$, $\lim_{x\to\infty} \mathbb{h}_p(x) = +\infty$, and hence $\mathbb{h}_p : [0, +\infty) \to [0, +\infty)$ is an isomorphism (continuous invertible function, with continuous inverse function). Its inverse function $(\mathbb{h}_p)^{-1} : [0, +\infty) \to [0, +\infty)$ is hence well-defined and it is used in the paper.

The Lerch transcendent function appears also in the derivatives of $\mathbb{h}_p$. Denote by

$$\Phi_1 = \Phi(1 - p, 1, x + 1), \qquad \Phi_2 = \Phi(1 - p, 2, x + 1), \qquad \Phi_3 = \Phi(1 - p, 3, x + 1),$$

and note that $\Phi_{n+1} = -n\partial\frac{\Phi_n}{\partial x}$; by (19) we get

$$\begin{aligned} \mathbb{h}_p'(x) &= \partial\frac{\psi(x + 1) + \gamma + \log p + (1 - p)^{x+1} \cdot \Phi(1 - p, 1, x + 1)}{\partial x} \\ &= \psi_1(x + 1) + (1 - p)^{x+1}(\log(1 - p) \cdot \Phi_1 - \Phi_2) \\ \mathbb{h}_p''(x) &= \psi_2(x + 1) + (1 - p)^{x+1}((\log(1 - p))^2 \cdot \Phi_1 - 2\log(1 - p) \cdot \Phi_2 + 2\Phi_3). \end{aligned} \tag{20}$$

Product representation and incomplete Gamma function. For what concerns the infinite product representation of the Gamma function

$$\Gamma(z) = \lim_{K\to\infty} \frac{e^{-\gamma z}}{z} \prod_{k=1}^{K} \left(1 + \frac{z}{k}\right)^{-1} e^{\frac{z}{k}}, \qquad z \neq -1, -2, \dots,$$

given by Schlömilch in 1844 and Newman in 1848, if we evaluate it in $z = \pm t$, we obtain

$$\Gamma(1 - t)e^{-\gamma t} = \prod_{j=1}^{\infty} \frac{e^{-\frac{t}{j}}}{1 - \frac{t}{j}},\ t \in (0, 1), \qquad \Gamma(1 + t)e^{\gamma t} = \prod_{j=1}^{\infty} \frac{e^{\frac{t}{j}}}{1 + \frac{t}{j}},\ t > 0. \tag{21}$$

Finally, for $x > 0$, we denote by $E_1(x)$ the *exponential integral* (or *incomplete gamma function*). As shown in [1, p. 229, 5.1.20], we have that

$$E_1(x) = \int_x^{\infty} \frac{e^{-t}}{t}\,dt < e^{-x}\, \ln\Bigl(1 + \frac{1}{x}\Bigr). \tag{22}$$

Note that, if $p \in (0,1)$ and $t = -\ln(1-p)w$,

$$E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\,dt = \int_{-\frac{x}{\ln(1-p)}}^\infty \frac{(1-p)^w}{w}\,dw.$$

We will make use of the very well known formula $-\ln(p) = \sum_{j=1}^\infty \frac{(1-p)^j}{j}$. To bound the tail of the series, we immediately obtain by (22) that, for any $x > 0$,

$$(23)\quad \sum_{j=0}^\infty \frac{(1-p)^{x+j+1}}{x+j+1} \le \int_x^\infty \frac{(1-p)^w}{w}\,dw = E_1(-x\ln(1-p))$$
$$< e^{x\ln(1-p)} \ln\Big(1 - \frac{1}{x\ln(1-p)}\Big).$$

The next representation lemma is used both in the analytical and in the numerical part of the paper.

Lemma A.2. *Let $x > 0$ be fixed. Then the functions*

$$g_+(t) = (x+\gamma)t - \ln\Gamma(1-t), \qquad t \in (0,1)$$
$$g_-(t) = (x-\gamma)t - \ln\Gamma(1+t), \qquad t > 0$$

attain their (strictly positive) maxima at the points $t_+ = 1-\psi^{-1}(-x-\gamma)$ and $t_- = \psi^{-1}(x-\gamma)-1$, respectively.

Proof. We give the proof for g_+, since the same arguments apply to g_-. We have

- $g_+(t)$ is concave, since $\ln\Gamma(1-t)$ is a convex analytic function on $(0,1)$;
- $g_+(0) = \ln\Gamma(1) = 0$, $g'_+(0) = (x+\gamma) + \psi(1) = x > 0$;
- $\lim_{t\to 1} g_+(t) = -\infty$;

and hence the maximum of g_+ on $(0,1)$ is strictly positive. The maximum point t_+ is attained when $g'_+(t_+) = 0$, that is when $(x+\gamma) + \psi(1+t_+) = 0$. □

Appendix B. Lower and upper bounds of some numerical problems

B.1. **Bounds of $y = \psi(x)$.** As shown in [9, Example 2.1], we may bound ψ from below in the following way. The Jensen inequality for $U \sim U(x-\frac{1}{2}, x+\frac{1}{2})$ shows that, for $x > \frac{1}{2}$,

$$\frac{1}{x} = \frac{1}{E[U]} < E\Big[\frac{1}{U}\Big] = \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} \frac{1}{t}\,dt = \ln(x+\tfrac{1}{2}) - \ln(x-\tfrac{1}{2}).$$

By (11), we than have that, for $x > \frac{1}{2}$,

$$\psi(x) - \ln(x-\tfrac{1}{2}) > \psi(x+1) - \ln(x+\tfrac{1}{2}) > \cdots > \liminf_{t\to\infty}(\psi(t) - \ln(t-\tfrac{1}{2})),$$

and since $\psi(t) = \log(t) + o(1) = \log(t-\frac{1}{2}) + o(1)$, the last expression is zero, and hence

$$y = \psi(x) > \ln(x-\tfrac{1}{2}), \qquad \text{for any } x > \tfrac{1}{2}.$$

With the same spirit of this example, since

$$\ln(x+1) - \ln(x) = \int_x^{x+1} \frac{1}{t}\,dt < \frac{1}{x}, \qquad \forall x > 0,$$

we obtain that

$$\psi(x) - \ln(x) < \psi(x+1) - \ln(x+1) < \cdots < \limsup_{t\to\infty}(\psi(t) - \ln(t)) = 0,$$

and hence, we may state that

$$\ln(x - \tfrac{1}{2}) < y < \ln(x), \qquad e^y < x < e^y + \tfrac{1}{2}, \qquad \forall x > \tfrac{1}{2}, \forall y.$$

B.2. **Bounds of $y = \mathbb{h}_p(x)$.** For what concerns the bounds for $\mathbb{h}_p$, by (19), we immediately get

$$\psi(x+1) + \gamma + \ln p \le \mathbb{h}_p(x) \le \psi(x+1) + \gamma,$$

and hence, by (12),

$$\frac{\exp(\mathbb{h}_p(x) - \gamma)}{p} - \frac{1}{2} \ge x \ge \exp(\mathbb{h}_p(x) - \gamma) - 1. \tag{24}$$

A better estimation for the lower bound can be found for $x > -\frac{1}{(e-1)\ln(1-p)}$. To simplify the notations, set $d_0 = -\ln(1-p)$, so that the assumption $x > -\frac{1}{(e-1)\ln(1-p)}$ becomes the more readable $xd_0 > \frac{1}{e-1}$. We are going to show that, under this hypothesis, we have

$$\frac{A}{p} - \frac{1}{2} \ge x \ge \begin{cases} \frac{A}{p} - e + \frac{1}{\ln(1-p)} & \text{if } A > p(\frac{1}{2} - \frac{1}{(e-1)\ln(1-p)}); \\ A - 1 & \text{otherwise;} \end{cases} \tag{25}$$

where $A = \exp(\mathbb{h}_p(x) - \gamma)$. To prove (25), we use the relation $\frac{1}{1-z} = \sum_{j=0}^{\infty} z^j$, valid for $|z| < 1$, in (19). We obtain

$$\begin{aligned} \mathbb{h}_p(x) &= \psi(x+1) + \gamma + \log p + \int_0^{1-p} \frac{z^x}{1-z}\, dz \\ &= \psi(x+1) + \gamma + \log p + \int_0^{1-p} \sum_{j=0}^{\infty} z^{x+j}\, dz \\ &= \psi(x+1) + \gamma + \log p + \sum_{j=0}^{\infty} \frac{(1-p)^{x+j+1}}{x+j+1}\, dz, \end{aligned}$$

which can be combined with (23), yielding

$$\begin{aligned} \mathbb{h}_p(x) - (\psi(x+1) + \gamma + \log p) &< e^{x\ln(1-p)} \ln\Big(1 - \frac{1}{x\ln(1-p)}\Big) \\ &< e^{x\ln(1-p)} \le \frac{1}{1 - x\ln(1-p)}, \end{aligned} \tag{26}$$

where the last inequality is a consequence of the fact that $\exp(x) \le \frac{1}{1-x}$ for $x < 1$.

Now, we define the positive quantity $d_1 = e - 1 + \frac{1}{d_0} > 0$ and we note that the function $g : [\frac{1}{d_0(e-1)}, \infty) \to \mathbb{R}$ so defined

$$g(x) = \frac{d_1}{d_1 + 1 + x} - \frac{1}{1 + xd_0} = \frac{x(d_0 d_1 - 1) - 1}{(d_1 + 1 + x)(1 + xd_0)}$$

is strictly positive whenever $x(d_0 d_1 - 1) - 1 > 0$, or, in other terms, when $d_1 > \frac{1+x}{d_0 x}$. We now prove that this fact implies that $g(x) > 0$ under our assumption $x > \frac{1}{d_0(e-1)}$.

In fact, since $\frac{1+y}{d_0 y}$ is decreasing in $y > 0$, then, as $x > \frac{1}{d_0(e-1)}$ we have

$$x > \tfrac{1}{d_0(e-1)} \quad \Longrightarrow \quad d_1 = \tfrac{d_0(e-1)+1}{d_0} = \frac{1 + \frac{1}{d_0(e-1)}}{d_0 \frac{1}{d_0(e-1)}} > \tfrac{1+x}{d_0 x} \quad \Longrightarrow \quad g(x) > 0,$$

or, in other terms,

$$x > \tfrac{1}{d_0(e-1)} \quad \Longrightarrow \quad \frac{d_1}{d_1 + 1 + x} > \frac{1}{1 + xd_0} = \frac{1}{1 - x\log(1-p)}.$$

Since $\frac{x}{1+x} < \ln(1+x)$ for $x > 0$, we then have that, when $x > \frac{1}{d_0(e-1)}$,

$$\frac{1}{1 - x\ln(1-p)} < \frac{d_1}{d_1 + 1 + x} = \frac{\frac{d_1}{x+1}}{1 + \frac{d_1}{x+1}} < \log\Big(1 + \frac{d_1}{x+1}\Big) = \ln\Big(\frac{x+1+d_1}{x+1}\Big) = \ln(x + e - \tfrac{1}{\ln(1-p)}) - \ln(x+1). \tag{27}$$

By combining together (26) and (27) we obtain

$$\mathbb{h}_p(x) - (\psi(x+1) + \gamma + \log p) < \ln(x + e - \tfrac{1}{\ln(1-p)}) - \ln(x+1),$$

that together with (12) yields

$$\begin{aligned} \mathbb{h}_p(x) - \gamma - \log p &< \psi(x+1) - \ln(x+1) + \ln(x + e - \tfrac{1}{\ln(1-p)}) \\ &< \ln(x + e - \tfrac{1}{\ln(1-p)}). \end{aligned}$$

Set $A = \exp(\mathbb{h}_p(x) - \gamma)$. The above inequality, exponentiated, gives

$$\frac{A}{p} - e + \tfrac{1}{\ln(1-p)} < x,$$

that, again by (24), is valid at least when

$$x > -\tfrac{1}{(e-1)\ln(1-p)} \quad \implies \quad A > p(x + \tfrac{1}{2}) > p(\tfrac{1}{2} - \tfrac{1}{(e-1)\ln(1-p)}).$$

B.3. **Bounds of $y = (x-\gamma)t(x) - \ln\Gamma(1 + t(x))$, where $t(x) = \psi^{-1}(x-\gamma) - 1$.** For what concerns the bounds in this problem, we start by recalling that, as shown in [19] (see also [24, Equation (3.112)]), for any $t > 0$,

$$-\gamma t < \ln\Gamma(1+t) < t\psi(t+1). \tag{28}$$

When this chain of inequalities is evaluated in $t = t(x)$, we obtain

$$\begin{aligned} -\gamma t(x) - \ln\Gamma(1+t(x)) < 0 \quad &\implies \quad y < xt(x) \\ \ln\Gamma(1+t(x)) < t\psi(\psi^{-1}(x-\gamma) - 1 + 1) \quad &\implies \quad y > 0. \end{aligned} \tag{29}$$

The upper bounds for x may be found in the following way. We recall that Lemma A.2 states that

$$y = \max_{t>0} \big[(x-\gamma)t - \ln\Gamma(1+t)\big].$$

Then, by (28),

$$y > \max_{t>0} \big[(x - \gamma - \psi(t+1))t\big]. \tag{30}$$

The second expression may be evaluated in $t_0 = \psi^{-1}(-\gamma + \frac{x}{2}) - 1$, so that we get

$$\begin{aligned} y &> (x - \gamma - \psi(t_0+1))t_0 \\ &= \frac{x}{2}\big(\psi^{-1}(-\gamma + \tfrac{x}{2}) - 1\big) \\ &= \frac{x}{2}\big(\psi^{-1}(-\gamma + \tfrac{x}{2}) - \psi^{-1}(-\gamma)\big). \end{aligned} \tag{31}$$

The Mean Value Theorem ensures the existence of $x_0 \in (0, \frac{x}{2})$ such that

$$\psi^{-1}(-\gamma + \tfrac{x}{2}) - \psi^{-1}(-\gamma) = \frac{x}{2} d\frac{\psi^{-1}(-\gamma + t)}{dt}\Big|_{t=x_0},$$

and by the the formula of the derivative of the inverse function, since the Trigamma function $\psi_1(t) = d\frac{\psi(t)}{dt}$ is a decreasing function with $\psi_1(1) = \frac{\pi^2}{6}$,

$$d\frac{\psi^{-1}(-\gamma+t)}{dt}\Big|_{t=x_0} = \frac{1}{\psi_1(\psi^{-1}(-\gamma+x_0))} > \frac{1}{\psi_1(\psi^{-1}(-\gamma))} = \frac{1}{\psi_1(1)} = \frac{1}{\frac{\pi^2}{6}}.$$

Summing up,

$$y > \frac{x}{2}\Big(\frac{x}{2}\frac{1}{\frac{\pi^2}{6}}\Big) = \frac{3}{2}\frac{x^2}{\pi^2} \quad\Longrightarrow\quad x < \pi\sqrt{\frac{2}{3}y}. \tag{32}$$

For $x \geq \frac{3}{2}$, which is always true if $y \geq \frac{3}{2} \cdot t(\frac{3}{2}) = 3$ by (29), a better estimates may be found if we bound the second part of (31). In fact, since $\frac{x}{2} \geq \frac{3}{4}$, by (12) we obtain

$$y > \frac{3}{4}\big(\psi^{-1}(-\gamma+\tfrac{x}{2}) - 1\big) > \frac{3}{4}\big(e^{-\gamma+\frac{x}{2}} - 1\big)$$

which completes the upper bound for x given in (32), obtaining

$$x < \begin{cases} \pi\sqrt{\frac{2}{3}y}, & \text{if } y < 3; \\ 2(\log(\frac{4}{3}y+1)+\gamma), & \text{if } y \geq 3. \end{cases} \tag{33}$$

The upper bounds for x may be found with similar ideas in both the cases $y \geq 3$ and $y < 3$. By (29), the Mean Value Theorem ensures the existence of $x_0 \in (0, x)$ such that, when $y < 3$

$$0 < y < xt(x) = x(\psi^{-1}(x-\gamma) - 1) = x^2\frac{1}{\psi_1(\psi^{-1}(x_0-\gamma))} < x^2\frac{1}{\psi_1(\psi^{-1}(\pi\sqrt{2}-\gamma))},$$

the last inequality being a consequence of (33), since, for $y < 3$, we have $x \leq \pi\sqrt{2}$. For $y \geq 3$, starting from (29), by (12), we obtain

$$0 < y < xt(x) = x(\psi^{-1}(x-\gamma) - 1) < x\Big(\exp(x-\gamma) - \frac{1}{2}\Big) < \Big(\exp(\tfrac{3}{2}x-\gamma) - \frac{1}{2}\Big),$$

which gives the lower bound for x in (15) for $y \geq 3$.

B.4. **Bounds of $y = (x+\gamma)t(x) - \ln\Gamma(1-t(x))$, where $t(x) = 1 - \psi^{-1}(-x-\gamma)$.** The inversion formula for the Gamma function, valid for $t \in (0,1)$, gives

$$\Gamma(1-t)\Gamma(t)t = \frac{\pi}{\sin(\pi t)}t \qquad\Longleftrightarrow\qquad \ln\Gamma(1-t) = \ln\Big(\frac{\pi t}{\sin(\pi t)}\Big) - \ln\Gamma(1+t),$$

that, together with (28), yealds

$$-t\psi(t+1) + \ln\Big(\frac{\pi t}{\sin(\pi t)}\Big) < \ln\Gamma(1-t) < \ln\Big(\frac{\pi t}{\sin(\pi t)}\Big) + \gamma t. \tag{34}$$

We recall that Lemma A.2 states that

$$y = \max_{t\in(0,1)}\big[(x+\gamma)t - \ln\Gamma(1+t)\big],$$

that, combined with the right-hand inequality of (34) gives

$$y > \max_{t\in(0,1)}\Big[xt + \ln\Big(\frac{\sin(\pi t)}{\pi t}\Big)\Big].$$

Since $\ln(y) > \frac{y-1}{y}$ and (see [2]),

$$\frac{\pi}{\sin(\pi t)} = \frac{1}{t} + \sum_{n=1}^{\infty}\frac{(-1)^n\,2t}{t^2-n^2},$$

then

$$y > \max_{t\in(0,1)} \big(xt - \frac{2t^2}{1-t^2}\big).$$

Let $t_0 = t_0(x) \in (0,1)$ be defined in the following way:

$$\frac{x}{2} = \frac{2t_0}{1-t_0^2} \qquad \Longleftrightarrow \qquad t_0 = 2\frac{\sqrt{\left(\frac{x}{2}\right)^2+1}-1}{x},$$

then

$$y > xt_0 - t_0\frac{2t_0}{1-t_0^2} = x\frac{t_0}{2} = \sqrt{\left(\frac{x}{2}\right)^2+1}-1,$$

and hence

$$x < 2\sqrt{(y+1)^2-1}. \tag{35}$$

For what concerns the lower bound for x, if we take into account the reflection formula for the digamma function

$$\psi(1-t)-\psi(t) = \pi\cot\pi t \qquad \Longrightarrow \qquad \psi(1+t) = \psi(t)+\frac{1}{t} = \psi(1-t)-\pi\cot(\pi t)+\frac{1}{t}$$

together with the left inequality in (34), we obtain

$$\begin{aligned}\ln\Gamma(1-t) &> -t\psi(t+1)+\ln\Big(\frac{\pi t}{\sin(\pi t)}\Big)\\ &= -t\Big(\psi(1-t)-\pi\cot(\pi t)+\frac{1}{t}\Big)+\ln\Big(\frac{\pi t}{\sin(\pi t)}\Big).\end{aligned}$$

We will make use of this inequality, motivated by the fact that our problem is

$$y = (x+\gamma)t(x)-\ln\Gamma(1-t(x)), \qquad -\psi(1-t(x)) = (x+\gamma),$$

which implies

$$\begin{aligned}y &= (x+\gamma)t(x)-\ln\Gamma(1-t(x))\\ &= -\psi(1-t(x))t(x)-\ln\Gamma(1-t(x))\\ &< 1-\pi t(x)\cot(\pi t(x))+\ln\Big(\frac{\sin(\pi t(x))}{\pi t(x)}\Big).\end{aligned} \tag{36}$$

Now, for $t \in (0,1)$, the following identities hold

$$\frac{\sin(\pi t)}{\pi t} = \prod_1^\infty\Big(1-\frac{t^2}{n^2}\Big), \qquad \pi\cdot\cot(\pi t) = \frac{1}{t}+\sum_{n=1}^\infty\frac{2t}{t^2-n^2},$$

(see [2]). The first identy may be used to bound the last term in (36):

$$\begin{aligned}\ln\Big(\frac{\sin(\pi t)}{\pi t}\Big) = \ln(1-t^2)+\sum_{n=2}^\infty\ln\Big(1-\frac{t^2}{n^2}\Big) &< \ln(1-t^2)+t^2-\sum_{n=1}^\infty\frac{t^2}{n^2}\\ &= \ln(1-t^2)+t^2\Big(1-\frac{\pi^2}{6}\Big).\end{aligned}$$

For what concerns the term $1-\pi t\cot(\pi t)$ in (36), we obtain

$$1-\pi t\cot(\pi t) = 2t^2\sum_{n=1}^\infty\frac{1}{n^2-t^2} = 2t^2\Big(\frac{1}{1-t^2}+\sum_{n=2}^\infty\frac{1}{n^2-t^2}\Big)$$

$$< 2t^2\Big(\frac{1}{1-t^2} + \sum_{m=1}^{\infty} \frac{1}{(m+1)^2-1}\Big) = 2t^2\Big(\frac{1}{1-t^2} + \frac{1}{2}\sum_{m=1}^{\infty} \frac{2}{m(m+2)}\Big)$$

$$= 2t^2\Big(\frac{1}{1-t^2} + \frac{1}{2}\sum_{m=1}^{\infty} \big(\frac{1}{m} - \frac{1}{m+2}\big)\Big) = \frac{2t^2}{1-t^2} + 3t^2.$$

Combining these two last inequalities in (36), since $\log y \le y-1$, we obtain

$$y < \frac{2t(x)^2}{1-t(x)^2} + 3t(x)^2 + \ln(1-t(x)^2) + t(x)^2\Big(1 - \frac{\pi^2}{6}\Big) < \frac{2}{1-t(x)^2} - 2 + t(x)^2\Big(3 - \frac{\pi^2}{6}\Big),$$

and hence, if we define

$$z = 1 - t(x)^2 \in (0,1), \qquad A = \frac{3 - \frac{\pi^2}{6}}{2} \in (0,1), \qquad B = \frac{y}{2} > 0$$

we obtain

$$Az^2 + (B + (1-A))z - 1 < 0, \qquad z \in (0,1)$$

which is solved for

$$0 < z < \frac{-(B+(1-A)) + \sqrt{(B+(1-A))^2 + 4A}}{2A}.$$

Note that, for $B \in (0,\infty)$, the right-hand side of the inequality above belongs to $(0,1)$. Then, if we define

$$C = \sqrt{1 - \frac{-(B+(1-A)) + \sqrt{(B+(1-A))^2 + 4A}}{2A}} \in (0,1),$$

we have $t(x) = \sqrt{1-z} > C$, or explicitely

(37) $$1 - \psi^{-1}(-x-\gamma) > C.$$

Two inequalities on x are consequence of (37) as follows. By (12) we imediately obtain a lower bound

$$1 - \exp(-(x+\gamma)) > 1 - \psi^{-1}(-x-\gamma) > C \qquad \Longrightarrow \qquad x > -\ln(1-C) - \gamma,$$

which is meaningful only for $C \ge 1 - \exp(-\gamma)$. For smaller C, we make use of the Mean Value Theorem, that ensures the existence of $x_0 \in (0,x)$ such that

$$t(x) = \psi^{-1}(-\gamma) - \psi^{-1}(-\gamma - x) = -x\, d\frac{\psi^{-1}(-\gamma-t)}{dt}\Big|_{t=x_0}.$$

The formula of the derivative of the inverse function gives

$$-d\frac{\psi^{-1}(-\gamma-t)}{dt}\Big|_{t=x_0} = \frac{1}{d\frac{\psi(t)}{dt}\Big|_{t=\psi^{-1}(-\gamma-x_0)}} < \frac{1}{d\frac{\psi(t)}{dt}\Big|_{t=\psi^{-1}(-\gamma)}} = \frac{1}{\frac{\pi^2}{6}},$$

so that

$$x > \frac{\pi^2}{6}C.$$

Summing up

(38) $$x > \max\Big(-\ln(1-C) - \gamma, \frac{\pi^2}{6}C\Big),$$

that completes (17) with the upper bounds for x given in (35).

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] M. Aigner and G. M. Ziegler. *Cotangent and the Herglotz trick*, pages 149–154. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[3] S. V. Aksenov, M. A. Savageau, U. D. Jentschura, J. Becher, G. Soff, and P. J. Mohr. Application of the combined nonlinear-condensation transformation to problems in statistical analysis and theoretical physics. *Computer Physics Communications*, 150(1):1 – 20, 2003.

[4] G. Alefeld. On the convergence of halley's method. *The American Mathematical Monthly*, 88(7):530–536, 1981.

[5] G. Aletti. Generation of discrete random variables in scalable frameworks. *Statist. Probab. Lett.*, 132:99–106, 2018.

[6] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.

[7] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.

[8] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In J. D. P. Rolim and S. Vadhan, editors, *Randomization and Approximation Techniques in Computer Science*, pages 1–10, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[9] H. G. Diamond and A. Straub. Bounds for the logarithm of the euler gamma function and its derivatives. *Journal of Mathematical Analysis and Applications*, 433(2):1072 – 1083, 2016.

[10] M. Durand and P. Flajolet. Loglog counting of large cardinalities (extended abstract). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2832:605–617, 2003.

[11] B. Eisenberg. On the expectation of the maximum of IID geometric random variables. *Statist. Probab. Lett.*, 78(2):135–143, 2008.

[12] O. Ertl. New cardinality estimation algorithms for hyperloglog sketches, 2017. preprint at `http://oertl.github.io/hyperloglog-sketch-estimation-paper/`.

[13] P. Flajolet. Approximate counting: A detailed analysis. *BIT Numerical Mathematics*, 25(1):113–134, Mar 1985.

[14] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *2007 Conference on Analysis of Algorithms, AofA 07*, Discrete Math. Theor. Comput. Sci. Proc., AH, pages 127–145. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2007.

[15] P. Flajolet and G. N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31(2):182–209, 1985.

[16] O. Gandouet and A. Jean-Marie. LogLog counting for the estimation of IP traffic. In *Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities*, Discrete Math. Theor. Comput. Sci. Proc., AG, pages 119–128. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2006.

[17] B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943.

[18] D. M. Kane, J. Nelson, and D. P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 41–52, New York, NY, USA, 2010. ACM.

[19] A. Laforgia and P. Natalini. On some inequalities for the gamma function. *Advances in Dynamical Systems and Applications*, 8(2):261–267, 2013.

[20] National Institute of Standards and Technology (NIST). CRYPTOGRAPHIC TOOLKIT. online at `http://csrc.nist.gov/groups/ST/toolkit/rng/`.

[21] National Institute of Standards and Technology (NIST). Guide to NIST's tests. online at `http://csrc.nist.gov/groups/ST/toolkit/rng/stats_tests.html`.

[22] National Institute of Standards and Technology (NIST). References. online at `http://csrc.nist.gov/groups/ST/toolkit/rng/references.html`.

[23] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[24] F. Qi. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010(1):493058, Mar 2010.

[25] T. R. Scavo and J. B. Thoo. On the geometry of halley's method. *The American Mathematical Monthly*, 102(5):417–426, 1995.

[26] D. Williams. *Probability with martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.

G. Aletti, ADAMSS Center, Università degli Studi di Milano, Milan, Italy
Email address: giacomo.aletti@unimi.it