1-2012

Estimating Statistical Characteristics of Lognormal and Delta-Lognormal Distributions under Interval Uncertainty: Algorithms and Computational Complexity

Nitaya Buntao
King Mongkut's Institute of Technology North Bangkok, taltanot@hotmail.com

Sa-aat Niwitpong
King Mongkut's Institute of Technology North Bangkok, snw@kmutnb.ac.th

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-12-02

Estimating Statistical Characteristics of Lognormal and Delta-Lognormal Distributions under Interval Uncertainty: Algorithms and Computational Complexity

Nitaya Buntao[1], Sa-aat Niwitpong[1], and Vladik Kreinovich[2]

[1]Department of Applied Statistics
King Mongkut's University of Technology North Bangkok
1518 Piboonsongkhram Road, Bangsue
Bangkok 10800 Thailand
taltanot@hotmail.com, snw@kmutnb.ac.th
[2]Computer Science Department
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

Traditional statistical estimates $\widehat{S}(x_1, \ldots, x_n)$ for different statistical characteristics S (such as mean, variance, etc.) implicitly assume that we know the sample values $x_1, \ldots, x_n$ exactly. In practice, the sample values $\widetilde{x}_i$ come from measurements and are, therefore, in general, different from the actual (unknown) values x_i of the corresponding quantities. Sometimes, we know the probabilities of different values of the measurement error $\Delta x_i = \widetilde{x}_i - x_i$, but often, the only information that we have about the measurement error is the upper bound Δ_i on its absolute value – provided by the manufacturer of the corresponding measuring instrument. In this case, the only information that we have about the actual values x_i is that they belong to the intervals $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

In general, different values $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ lead to different values of the corresponding estimate $\widehat{S}(x_1, \ldots, x_n)$. In this case, it is desirable to find the range of all possible values of this characteristic.

In this paper, we consider the problem of computing the corresponding range for the cases of lognormal and delta-lognormal distributions. Interestingly, it turns out that, in contrast to the case of normal distribution for which it is feasible to compute the range of the mean, for lognormal and delta-lognormal distributions, computing the range of the mean is an NP-hard problem.

1 Introduction

Need for interval uncertainty. Traditional statistical estimates $\widehat{S}(x_1, \ldots, x_n)$ for different statistical characteristics S (such as mean, variance, etc.) implicitly assume that we know the sample values $x_1, \ldots, x_n$ exactly. In practice, the sample values $\widetilde{x}_i$ come from measurements and are, therefore, in general, different from the actual (unknown) values x_i of the corresponding quantities. Sometimes, we know the probabilities $\rho_i(\Delta x_i)$ of different values of the measurement error $\Delta x_i = \widetilde{x}_i - x_i$, but often, the only information that we have about the measurement error is the upper bound Δ_i on its absolute value – provided by the manufacturer of the corresponding measuring instrument. In this case, the only information that we have about the actual values x_i is that they belong to the intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$; see, e.g., [18].

In general, different values $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ lead to different values of the corresponding estimate $\widehat{S}(x_1, \ldots, x_n)$. In this case, it is desirable to find the range of all possible values of this estimate:

$$\widehat{\mathbf{S}} = \{\widehat{S}(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

For continuous estimates $\widehat{S}(x_1, \ldots, x_n)$, this range is an interval $\widehat{\mathbf{S}}$.

What is known. For different statistical estimates, there exist numerous efficient algorithms for computing the interval ranges of these characteristics under interval uncertainty; see, e.g., [2, 3, 6, 8, 12, 13, 15, 14, 16, 17, 21].

For example, the standard ways to estimate mean E and variance V based on the same $x_1, \ldots, x_n$ is to use the estimates $\widehat{E} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i$ and

$$\widehat{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \widehat{E}\right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\widehat{E}\right)^2.$$

For the normal distribution, these estimates correspond to the Maximum Likelihood Methods (and are, therefore, asymptotically optimal). For other distributions, these same estimates – while not necessarily asymptotically optimal – also converge to the actual mean and the actual variance when the sample size increases.

Comment. In many practical situations, we are interested in an *unbiased* estimate $\widehat{V}_u$ of the population variance V:

$$\widehat{V}_u(x_1, \ldots, x_n) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \widehat{E}\right)^2.$$

In this paper, we will describe how to estimate the range of $\widehat{V}$ under interval uncertainty; since $\widehat{V}_u = \dfrac{n}{n-1} \cdot \widehat{V}$, we can easily transform the range of the estimate $\widehat{V}$ into the range for the estimate $\widehat{V}_u$.

Computing arithmetic average under interval uncertainty. The arithmetic average $\widehat{E}$ is a monotonically increasing function of each of its n variables $x_1, \ldots, x_n$, so its smallest possible value $\underline{\widehat{E}}$ is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$) and its largest possible value is attained when $x_i = \overline{x}_i$ for all i. In other words, the range $\widehat{\mathbf{E}}$ of $\widehat{E}$ is equal to $\left[\widehat{E}(\underline{x}_1, \ldots, \underline{x}_n), \widehat{E}(\overline{x}_1, \ldots, \overline{x}_n)\right]$. In other words, $\underline{\widehat{E}} = \dfrac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{\widehat{E}} = \dfrac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)$.

Similarly, the standard estimate $\widehat{m} = \mathrm{med}(x_1, \ldots, x_n)$ for the median is a monotonic function of all its variables, so its range can be computed as

$$\left[\underline{\widehat{m}}, \overline{\widehat{m}}\right] = [\widehat{m}(\underline{x}_1, \ldots, \underline{x}_n), \widehat{m}(\overline{x}_1, \ldots, \overline{x}_n)]\,.$$

Computing sample variance under interval uncertainty. In contrast to the arithmetic average, the dependence of the sample variance $\widehat{V}$ on x_i is not monotonic, so the above simple idea does not work. Rather surprisingly, it turns out that the problem of computing the exact range for the sample variance over interval data is, in general, NP-hard [7, 16] which means, crudely speaking, that the worst-case computation time grows exponentially with n.

Comment. To be more precise, a problem $\mathcal{P}$ is NP-hard if every problem from a class NP can be reduced to $\mathcal{P}$; see, e.g., [9]. For a more detailed description of NP-hardness in relation to interval uncertainty, see, e.g., [11, 16].

Computing sample variance under interval uncertainty (cont-d). Specifically, computing the upper endpoint $\overline{\widehat{V}}$ of the range $\left[\underline{\widehat{V}}, \overline{\widehat{V}}\right]$ is NP-hard. Moreover, if we want to compute the variance range or $\overline{\widehat{V}}$ with a given accuracy ε, the problem is still NP-hard [7, 16].

Specifically, the lower endpoints $\underline{\widehat{V}}$ can be computed feasibly (i.e., in polynomial time), while computing the upper endpoint $\overline{\widehat{V}}$ is, in general, NP-hard. In the example on which NP-hardness is proven, all n intervals have a common point. In many practically important situations, it is possible to feasibly compute $\overline{\widehat{V}}$. For example, $\overline{\widehat{V}}$ can be feasibly computed when there exists a constant C for which every subset of C intervals has an empty intersection.

Maximum likelihood estimates for the lognormal distribution. In many practical situations, we encounter lognormal distributions [1], i.e., distribution of a random variable x whose logarithm $y = \ln(x)$ is normally distributed. This distribution is usually characterized by the parameters μ and σ of the corresponding normal distribution.

In principle, we can use the above formula to estimate the mean and the variance from the sample $x_1, \ldots, x_n$. However, as we have mentioned, for non-normal distributions, these estimates are not asymptotically optimal. To get asymptotically optimal estimates, we need to use the Maximum Likelihood

Method. For the lognormal distribution, once we have a sample $x_1, \ldots, x_n$, we can compute the values $y_i = \ln(x_i)$ and then use the Maximum Likelihood Method to estimate the parameters μ and σ:

$$\widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i; \quad \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \widehat{\mu})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - (\widehat{\mu})^2 .$$

Once we know the values $\widehat{\mu}$ and $\widehat{\sigma}$, we can estimate the mean E and the variance V of the lognormal distribution as

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right); \quad \widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right).$$

The coefficient of variation CV – which is defined as the ratio $\dfrac{\sqrt{V}}{E}$ of the standard deviation over mean, is therefore estimated as

$$\widehat{CV} = \sqrt{\exp\left(\widehat{\sigma}^2\right) - 1}.$$

The values of the median m and the mode m_0 can also be estimated based on μ and σ: $\widehat{m} = \exp\left(\widehat{\mu}\right)$ and $\widehat{m}_0 = \exp\left(\widehat{\mu} - \widehat{\sigma}^2\right)$.

Comment. For estimates originating from the normal distribution, we did not consider problems for estimating median and mode, since for normal distribution, median, mode, and mean coincide.

Case of the lognormal distribution: first problem that we solve in this paper. As we have mentioned, in practice, after the measurements and/or observations, instead of the exact values x_i of the sample, we often only know the intervals $[\underline{x}_i, \overline{x}_i]$ of possible values of x_i. Different values x_i from these intervals lead, in general, to different values of $\widehat{E}$, $\widehat{V}$, $\widehat{CV}$, $\widehat{m}$, and $\widehat{m}_0$. It is therefore desirable to find the ranges of these characteristics. This is the first problem with which we deal in this paper.

Maximum likelihood estimates for the delta-lognormal distribution. In many practical applications, e.g., in medical applications and in meteorology, a quantity can take any non-negative values but have a positive probability of 0 values. In many such cases, the probabilities are described by the *delta-lognomal* distribution, in which with a given probability $d > 0$, we get a value 0, and with the remaining probability $1 - d$, we get a lognormal distribution; see, e.g., [1, 4, 20].

In medical applications, in distribution of test costs, zeros correspond to the cases when a patient refused a test. In environmental applications, zeros correspond to the case when the actual concentration of the analyzed chemical is below the detection limit. In biological applications, e.g., in distribution of certain species in different geographic areas, zeros correspond to areas with are unsuitable for these species, etc.

The corresponding probability density has the form

$$\rho(x,\mu,\sigma,d) = d\cdot\delta(x) + (1-d)\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma\cdot x}\cdot\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right),$$

where $\delta(x)$ denotes Dirac's delta-function (a generalized function that describes the probability density of a random variable which is located at point 0 with probability 1).

For the delta-lognormal distribution, Maximum Likelihood leads to the following estimates:

$$\widehat{d} = \frac{\#\{i: x_i = 0\}}{n};\ \ \widehat{\mu} = \frac{1}{\#\{i: x_i > 0\}}\cdot\sum_{i:x_i>0}\ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i: x_i > 0\}}\cdot\sum_{i:x_i>0}\left(\ln(x_i)-\widehat{\mu}\right)^2;\ \ \widehat{E} = \left(1-\widehat{d}\right)\cdot\exp\left(\widehat{\mu}+\frac{\widehat{\sigma}^2}{2}\right);$$

$$\widehat{V} = \left(1-\widehat{d}\right)\cdot\exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot\left(\exp\left(\widehat{\sigma}^2\right)+\widehat{d}-1\right);$$

$$\widehat{CV} = \sqrt{\frac{\exp\left(\widehat{\sigma}^2\right)+\widehat{d}-1}{1-\widehat{d}}}.$$

Case of the delta-lognormal distribution: second problem that we solve in this paper. In practice, instead of the exact values x_i of the sample, we often only know the intervals $[\underline{x}_i, \overline{x}_i]$ of possible values of x_i. Different values x_i from these intervals lead, in general, to different values of $\widehat{E}$, $\widehat{V}$, and $\widehat{CV}$. It is therefore desirable to find the ranges of these characteristics. This is the second problem with which we deal in this paper.

2 Case of the Lognormal Distribution: Estimating Median under Interval Uncertainty

Problem: reminder. Let us start with computing the range $[\underline{\widehat{m}}, \overline{\widehat{m}}]$ of the estimate for the median $\widehat{m} = \exp(\widehat{\mu})$ of the lognormal distribution, where $\widehat{\mu} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i$ and $y_i = \ln(x_i)$, under interval uncertainty $x_i \in [\underline{x}_i, \overline{x}_i]$.

In this problem, we are given n intervals $[\underline{x}_i, \overline{x}_i]$, and we need to find the range of possible values of $\widehat{m} = \exp(\widehat{\mu})$.

Analysis of the problem. The median estimate $\widehat{m} = \exp(\widehat{\mu})$ is an increasing function of the parameter estimate $\widehat{\mu}$. The parameter estimate $\widehat{\mu}$, in its turn, is an increasing function of all its variables y_i, and each $y_i = \ln(x_i)$ is an increasing function of x_i. Thus, the median estimate $\widehat{m}$ is a monotonically increasing function of all its variables $x_1, \ldots, x_n$. Due to this monotonicity, we can make the same conclusion as when we estimated the mean of the normal distribution:

- the median estimate $\widehat{m}$ attains its smallest value when each variable x_i takes its smallest possible value $\underline{x}_i$, and
- the median estimate $\widehat{m}$ attains its largest value when each variable x_i takes its largest possible value $\overline{x}_i$.

Thus, we arrive at the following range $\left[\underline{\widehat{m}}, \overline{\widehat{m}}\right]$:

Resulting estimate. When the inputs x_i are in the intervals $[\underline{x}_i, \overline{x}_i]$, the range range $\left[\underline{\widehat{m}}, \overline{\widehat{m}}\right]$ of the possible values of the median estimate is

$$\left[\underline{\widehat{m}}, \overline{\widehat{m}}\right] = \left[\exp\left(\underline{\widehat{\mu}}\right), \exp\left(\overline{\widehat{\mu}}\right)\right],$$

where

$$\underline{\widehat{\mu}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(\underline{x}_i); \quad \overline{\widehat{\mu}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(\overline{x}_i).$$

3 Case of the Lognormal Distribution: Estimating Mean under Interval Uncertainty

Observation: for lognormal distribution, the Maximum Likelihood estimate for the mean is not monotonic. For the normal distribution, the Maximum Likelihood estimate for the mean is the arithmetic average $\widehat{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$, which is an increasing function of all its variables $x_1, \ldots, x_n$. Our first observation is that for the lognormal distribution, the corresponding Maximum Likelihood estimate for the mean $\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right)$ is not always an increasing function of all its inputs.

Proposition 1. *For the lognormal distribution, the Maximum Likelihood estimate for the mean*

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2,$$

is not always an increasing function of all its inputs.

Comment. For reader's convenience, all the proofs are placed in a special Proofs section.

In general, computing $\overline{\widehat{E}}$ is NP-hard. It turns out that in general, the problem of computing the upper endpoint $\overline{\widehat{E}}$ of the interval of possible values of the mean estimate $\widehat{E}$ is NP-hard:

Proposition 2. *Let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the lognormal distribution:*

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$.*

Comment. As we have mentioned earlier, for the lognormal distribution, there are two possible estimates for the mean based on the sample $x_1, \ldots, x_n$:

- the general estimate $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ which is applicable to all possible distributions but which is not necessarily asymptotically optimal, and
- the estimate $\widehat{E}$ based on the Maximum Likelihood Method for the lognormal distribution, an estimate which is asymptotically optimal.

For the first estimate, results are not optimal, but the range can be easily computed, while for the second estimate, the results are asymptotically optimal, but the problem is computationally difficult (NP-hard).

It is worth mentioning that the situation with estimating the variance V in the normal distribution case is somewhat similar:

- computing the maximum likelihood (asymptotically optimal) estimate is, as we have mentioned, NP-hard, while
- the average mean deviation $\frac{1}{n} \cdot \sum_{i=1}^{n} \left|x_i - \widehat{E}\right|$ (where $\widehat{E} = \frac{1}{n} \cdot \sum x_i$), a quantity which can also be used to provide an estimate for V, can be computed in polynomial time [10].

Approximate computation of $\overline{\widehat{E}}$ is also NP-hard. The above problem remains NP-hard if we are only interested in computing $\overline{\widehat{E}}$ with a given accuracy.

Proposition 3. *Let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the lognormal distribution:*

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Let $\varepsilon > 0$ be a real number. Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute a value r whose difference from the upper endpoint $\overline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$ does not exceed ε: $\left|r - \overline{\widehat{E}}\right| \le \varepsilon$.*

Feasible algorithms for computing $\underline{\widehat{E}}$ and $\overline{\widehat{E}}$. We have shown that computing the upper endpoint $\overline{\widehat{E}}$ is, in general, NP-hard. Let us show how to feasibly compute the lower endpoint $\underline{\widehat{E}}$ and how to feasibly compute the upper endpoint $\overline{\widehat{E}}$ when the intervals do not intersect much.

Proposition 4. *Let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the lognormal distribution:*

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the lower endpoint $\underline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$.*

Comment. As we will see, this algorithm takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n + 1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$. For each zone, we do the following:

- First, we compute the values

$$s_k^- = \sum_{i:\overline{y}_i \le r_k} \overline{y}_i; \quad s_k^+ = \sum_{j:r_{k+1} \le \underline{y}_i} \underline{y}_j;$$

$$M_k^- = \sum_{i:\overline{y}_i \le r_k} (\overline{y}_i)^2; \quad M_k^+ = \sum_{j:r_{k+1} \le \underline{y}_j} \left(\underline{y}_i\right)^2;$$

  and the number n_k of all the indices i for which $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

- Then, we compute the value $\widehat{\mu}_k$ as

$$\widehat{y}_k = \frac{s_k^- + s_k^+ - n_k}{n - n_k}.$$

- If $r_k \le \widehat{\mu}_k - 1 \le r_{k+1}$, we then compute

$$M_k = \frac{M_k^- + M_k^+ + n_k \cdot (\widehat{\mu} - 1)^2}{n}; \quad \sigma_k^2 = M_k - (\widehat{\mu}_k)^2; \quad A_k = \widehat{\mu}_k + \frac{\sigma_k^2}{2}.$$

We then take the smallest $A = \min\limits_k A_k$ of all the values A_k, and return $\exp(A)$ as the desired value $\underline{\widehat{E}}$.

Proposition 5. *Let C be a positive integer, and let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the lognormal distribution:*

$$\widehat{E} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\ln(x_i) - \widehat{\mu})^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the upper endpoint $\overline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$.*

Comment. This algorithm also takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n+1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$. For each zone, we do the following:

- For each value i, we select:
  - the value $y_i = \underline{y}_i$ if $\overline{y}_i \le r_k$;
  - the value $y_i = \overline{y}_i$ if $r_{k+1} \le \underline{y}_i$; and
  - both values in all other cases, i.e., when $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

  For each of the resulting tuples, we compute the value $A = \widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}$.

Then, we compute the largest $A_{\max}$ of these values A and finally, the desired bound $\overline{\widehat{E}} = \exp(A_{\max})$.

4 Case of the Lognormal Distribution: Estimating Mode under Interval Uncertainty

In general, computing $\underline{\widehat{m}_0}$ is NP-hard. It turns out that in general, the problem of computing the lower endpoint $\underline{\widehat{m}_0}$ of the interval of possible values of the mode estimate $\widehat{m}_0$ is NP-hard:

Proposition 6. *Let $\widehat{m}_0$ be the Maximum Likelihood estimate for the mode m_0 corresponding to the lognormal distribution:*

$$\widehat{m}_0 = \exp(\widehat{\mu} - \widehat{\sigma}^2), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the lower endpoint $\underline{\widehat{m}_0}$ of the range of corresponding values of $\widehat{m}_0$.*

Proposition 7. *Let $\widehat{m}_0$ be the Maximum Likelihood estimate for the mode m_0 corresponding to the lognormal distribution:*

$$\widehat{m}_0 = \exp(\widehat{\mu} - \widehat{\sigma}^2), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Let $\varepsilon > 0$ be a real number. Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute a value r whose difference from the lower endpoint $\underline{\widehat{m}_0}$ of the range of corresponding values of $\widehat{m}_0$ does not exceed ε:*

$$\left| r - \underline{\widehat{m}_0} \right| \le \varepsilon.$$

Feasible algorithms for computing $\underline{\widehat{m}_0}$ and $\overline{\widehat{m}_0}$. We have shown that computing the lower endpoint $\underline{\widehat{m}_0}$ is, in general, NP-hard. Let us show how to feasibly compute the upper endpoint $\overline{\widehat{m}_0}$ and how to feasibly compute the lower endpoint $\underline{\widehat{m}_0}$ when the intervals do not intersect much.

Proposition 8. *Let $\widehat{m}_0$ be the Maximum Likelihood estimate for the mode m_0 corresponding to the lognormal distribution:*

$$\widehat{m}_0 = \exp\left(\widehat{\mu} - \widehat{\sigma}^2\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the upper endpoint $\overline{\widehat{m}_0}$ of the range of corresponding values of $\widehat{m}_0$.*

Comment. As we will see, this algorithm takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n+1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$. For each zone, we do the following:

- First, we compute the values

$$s_k^- = \sum_{i:\overline{y}_i \le r_k} \overline{y}_i; \quad s_k^+ = \sum_{j:r_{k+1} \le \underline{y}_i} \underline{y}_j;$$

$$M_k^- = \sum_{i:\overline{y}_i \le r_k} (\overline{y}_i)^2; \quad M_k^+ = \sum_{j:r_{k+1} \le \underline{y}_j} \left(\underline{y}_i\right)^2;$$

  and the number n_k of all the indices i for which $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

- Then, we compute the value $\widehat{\mu}_k$ as

$$\widehat{\mu}_k = \frac{s_k^- + s_k^+ + \dfrac{n_k}{2}}{n - n_k}.$$

- If $r_k \le \widehat{\mu}_k + \dfrac{1}{2} \le r_{k+1}$, we then compute

$$M_k = \frac{M_k^- + M_k^+ + n_k \cdot \left(\widehat{\mu} + \dfrac{1}{2}\right)^2}{n}; \quad \sigma_k^2 = M_k - \left(\widehat{\mu}_k\right)^2; \quad A_k = \widehat{\mu}_k - \sigma_k^2.$$

We then take the largest $A = \max\limits_k A_k$ of all the values A_k, and return $\exp(A)$ as the desired value $\overline{\widehat{m}_0}$.

Proposition 9. *Let C be a positive integer, and let $\widehat{m}_0$ be the Maximum Likelihood estimate for the mode m_0 corresponding to the lognormal distribution:*

$$\widehat{m}_0 = \exp\left(\widehat{\mu} - \widehat{\sigma}^2\right), \text{ where } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i) \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the lower endpoint $\underline{\widehat{m}_0}$ of the range of corresponding values of $\widehat{m}_0$.*

Comment. This algorithm also takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n+1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$. For each zone, we do the following:

- For each value i, we select:
  - the value $y_i = \underline{y}_i$ if $\overline{y}_i \le r_k$;
  - the value $y_i = \overline{y}_i$ if $r_{k+1} \le \underline{y}_i$; and
  - both values in all other cases, i.e., when $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

  For each of the resulting tuples, we compute the value $A = \widehat{\mu} - \widehat{\sigma}^2$.

Then, we compute the smallest $A_{\min}$ of these values A and finally, the desired bound $\underline{\widehat{m}_0} = \exp(A_{\min})$.

5 Case of the Lognormal Distribution: Estimating Variance under Interval Uncertainty

Proposition 10. *Let $\widehat{V}$ be the Maximum Likelihood estimate for the variance V corresponding to the lognormal distribution:*

$$\widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right),$$

where $\widehat{\mu} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)$ and $\widehat{\sigma}^2 = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2$. Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{V}}$ of the range of corresponding values of $\widehat{V}$.*

Proposition 11. *Let $\widehat{V}$ be the Maximum Likelihood estimate for the variance V corresponding to the lognormal distribution:*

$$\widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right),$$

where $\widehat{\mu} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)$ *and* $\widehat{\sigma}^2 = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2$. *Then, there exists a feasible (polynomial time) algorithm that:*

- *given* n *intervals* $[\underline{x}_i, \overline{x}_i]$,
- *computes the lower endpoint* $\underline{\widehat{V}}$ *of the range of corresponding values of* $\widehat{V}$.

Comment. As we will see, this algorithm takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln\left(\underline{x}_i\right)$ and $\overline{y}_i = \ln\left(\overline{x}_i\right)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n+1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$. For each zone, we do the following:

- First, we compute the values

$$s_k^- = \sum_{i:\overline{y}_i \le r_k} \overline{y}_i; \quad s_k^+ = \sum_{j:r_{k+1} \le \underline{y}_i} \underline{y}_j;$$

$$M_k^- = \sum_{i:\overline{y}_i \le r_k} (\overline{y}_i)^2; \quad M_k^+ = \sum_{j:r_{k+1} \le \underline{y}_j} \left(\underline{y}_i\right)^2;$$

  and the number n_k of all the indices i for which $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

- Then, we compute the value r from the equation

$$r = \widehat{\mu} - \frac{\exp\left(\widehat{\sigma}^2\right) - 1}{2\exp\left(\widehat{\sigma}^2\right) - 1}, \text{ where}$$

$$\widehat{\mu} = \frac{s_k^- + s_k^+ + n_k \cdot r}{n - p}; \quad \widehat{\sigma}^2 = \frac{M_k^- + M_k^+ + n_k \cdot r^2 - \left(\widehat{\mu}\right)^2}{n - p}.$$

- If $r_k \le r \le r_{k+1}$, we then compute the value

$$V = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right).$$

We then take the smallest of all the values V as the desired value $\underline{\widehat{V}}$.

Proposition 12. *Let* C *be a positive integer, and let* $\widehat{V}$ *be the Maximum Likelihood estimate for the variance* V *corresponding to the lognormal distribution:*

$$\widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right),$$

where $\widehat{\mu} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i)$ *and* $\widehat{\sigma}^2 = \dfrac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2$. *Then, there exists a feasible (polynomial time) algorithm that:*

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the upper endpoint $\overline{\widehat{V}}$ of the range of corresponding values of $\widehat{V}$.*

Comment. This algorithm also takes time $O(n \cdot \log(n))$.

Description of the corresponding algorithm. First, we sort all $2n$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n}.$$

Let us add $r_0 = -\infty$ and $r_{2n+1} = +\infty$. This divides the real line into $2n+1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n$.

For each zone, for each value i, we select:

- the value $y_i = \underline{y}_i$ if $\overline{y}_i \le r_k$;
- the value $y_i = \overline{y}_i$ if $r_{k+1} \le \underline{y}_i$; and
- for each other index i, i.e., when $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$, select one of the three possible cases: $y_i = \underline{y}_i$, $y_i = \overline{y}_i$, and $y_i = r$, for a constant r to be determined later.

For each of the resulting tuples, we find r from the equation

$$r = \widehat{\mu} - \frac{\exp\left(\widehat{\sigma}^2\right) - 1}{2\exp\left(\widehat{\sigma}^2\right) - 1},$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are sample mean and sample variance of the selected values y_i – thus, depending on r as well. Once we find r, we thus know the values of all selected y_i, so we can compute the values $\widehat{\mu}$, $\widehat{\sigma}^2$, and

$$\widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right).$$

Then, we compute the largest of these values $\widehat{V}$ and return it as the desired bound $\overline{\widehat{V}}$.

6 Case of the Lognormal Distribution: Estimating Coefficient of Variation under Interval Uncertainty

The Maximum Likelihood estimate $\widehat{CV}$ for the coefficient of variation CV – which is defined as the ratio $\dfrac{\sqrt{V}}{E}$ of the standard deviation over mean – is equal to

$$\widehat{CV} = \sqrt{\exp\left(\widehat{\sigma}^2\right) - 1}.$$

This estimate is a monotonic function of the sample variance $\widehat{\sigma}^2$; thus:

- its largest value $\overline{\widehat{CV}}$ is attained when the sample variance $\widehat{\sigma}^2$ attains its largest possible value $\overline{\widehat{\sigma}^2}$, and
- its smallest value $\underline{\widehat{CV}}$ is attained when the sample variance $\widehat{\sigma}^2$ attains its smallest possible value $\underline{\widehat{\sigma}^2}$:

$$\left[\underline{\widehat{CV}}, \overline{\widehat{CV}}\right] = \left[\sqrt{\exp\left(\underline{\widehat{\sigma}^2}\right) - 1}, \sqrt{\exp\left(\overline{\widehat{\sigma}^2}\right) - 1}\right].$$

Thus, from the known results about computational complexity of computing the endpoints $\overline{\widehat{\sigma}^2}$ and $\underline{\widehat{\sigma}^2}$ [7, 16, 21], we make the following conclusions:

Proposition 13. *Let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation $CV = \sqrt{VE}$ corresponding to the lognormal distribution:*

$$\widehat{CV} = \sqrt{\exp\left(\widehat{\sigma}^2\right) - 1}, \text{ where } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2 \text{ and } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i).$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Proposition 14. *Let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation $CV = \sqrt{VE}$ corresponding to the lognormal distribution:*

$$\widehat{CV} = \sqrt{\exp\left(\widehat{\sigma}^2\right) - 1}, \text{ where } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2 \text{ and } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i).$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the lower endpoint $\underline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Comment. This algorithm takes linear time $O(n)$.

Proposition 15. *Let C be a positive integer, and let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation CV corresponding to the lognormal distribution:*

$$\widehat{CV} = \sqrt{\exp\left(\widehat{\sigma}^2\right) - 1}, \text{ where } \widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\ln(x_i) - \widehat{\mu}\right)^2 \text{ and } \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} \ln(x_i).$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the upper endpoint $\overline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Comment. This algorithm takes linear time $O(n)$.

7 Case of the Delta-Lognormal Distribution: Estimating Mean under Interval Uncertainty

Proposition 16. *Let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the delta-lognormal distribution:*

$$\widehat{E} = \left(1 - \widehat{d}\right) \cdot \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{d} = \frac{\#\{i : x_i = 0\}}{n},$$

$$\widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} \ln(x_i); \text{ and } \widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} (\ln(x_i) - \widehat{\mu})^2.$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$.*

Proposition 17. *Let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the delta-lognormal distribution:*

$$\widehat{E} = \left(1 - \widehat{d}\right) \cdot \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{d} = \frac{\#\{i : x_i = 0\}}{n},$$

$$\widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} \ln(x_i); \text{ and } \widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} (\ln(x_i) - \widehat{\mu})^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the lower endpoint $\underline{\widehat{E}}$ of the range of corresponding values of $\widehat{E}$.*

Comment. This algorithm takes time $O(n^2 \cdot \log(n))$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n - p$ intervals, we sort the corresponding $2 \cdot (n-p)$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n-2p}.$$

Let us add $r_0 = -\infty$ and $r_{2n-2p+1} = +\infty$. This divides the real line into $2n - 2p + 1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n - 2p$. For each zone, we do the following:

- First, we compute the values

$$s_k^- = \sum_{i:\overline{y}_i \le r_k} \overline{y}_i; \quad s_k^+ = \sum_{j:r_{k+1} \le \underline{y}_i} \underline{y}_j;$$

$$M_k^- = \sum_{i:\overline{y}_i \le r_k} (\overline{y}_i)^2; \quad M_k^+ = \sum_{j:r_{k+1} \le \underline{y}_j} \left(\underline{y}_i\right)^2;$$

  and the number n_k of all the indices i for which $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

- Then, we compute the value $\widehat{\mu}_k$ as

$$\widehat{\mu}_k = \frac{s_k^- + s_k^+ - n_k}{n - p - n_k}.$$

- If $r_k \le \widehat{\mu}_k - 1 \le r_{k+1}$, we then compute

$$M_k = \frac{1}{n-p} \cdot \left(M_k^- + M_k^+ + n_k \cdot (\widehat{\mu}_k - 1)^2\right); \quad \sigma_k^2 = M_k - (\widehat{\mu}_k)^2;$$

$$\widehat{E} = \left(1 - \frac{p}{n}\right) \cdot \exp\left(\widehat{\mu}_k + \frac{\sigma_k^2}{2}\right).$$

We then compute the smallest of all the resulting values $\widehat{E}$ (corresponding to all values p and to all zones), and return this smallest value as the desired value $\underline{\widehat{E}}$.

Proposition 18. *Let C be a positive integer, and let $\widehat{E}$ be the Maximum Likelihood estimate for the mean E corresponding to the delta-lognormal distribution:*

$$\widehat{E} = \left(1 - \widehat{d}\right) \cdot \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right), \text{ where } \widehat{d} = \frac{\#\{i : x_i = 0\}}{n},$$

$$\widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i:x_i>0} \ln(x_i); \text{ and } \widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i:x_i>0} (\ln(x_i) - \widehat{\mu})^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*

- *computes the upper endpoint* $\overline{\widehat{E}}$ *of the range of corresponding values of* $\widehat{E}$.

Comment. This algorithm also takes time $O(n^2 \cdot \log(n))$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n - p$ intervals, we sort the corresponding $2 \cdot (n - p)$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n-2p}.$$

Let us add $r_0 = -\infty$ and $r_{2n-2p+1} = +\infty$. This divides the real line into $2n - 2p + 1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n - 2p$. For each zone, we do the following:

- For each value i, we select:
  - the value $y_i = \underline{y}_i$ if $\overline{y}_i \le r_k$;
  - the value $y_i = \overline{y}_i$ if $r_{k+1} \le \underline{y}_i$; and
  - both values in all other cases, i.e., when $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.
- For each of the resulting tuples, we compute the value

$$\widehat{E} = \left(1 - \frac{p}{n}\right) \cdot \exp\left(\widehat{\mu} + \frac{\sigma^2}{2}\right).$$

We then compute the largest of all the resulting values $\widehat{E}$ (corresponding to all values p and to all zones), and return this largest value as the desired value $\overline{\widehat{E}}$.

8 Case of the Delta-Lognormal Distribution: Estimating Variance under Interval Uncertainty

Proposition 19. *Let* $\widehat{V}$ *be the Maximum Likelihood estimate for the variance* V *corresponding to the delta-lognormal distribution:*

$$\widehat{V} = \left(1 - \widehat{d}\right) \cdot \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1\right),$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{V}}$ of the range of corresponding values of $\widehat{V}$.*

Proposition 20. *Let $\widehat{V}$ be the Maximum Likelihood estimate for the variance V corresponding to the delta-lognormal distribution:*

$$\widehat{V} = \left(1 - \widehat{d}\right) \cdot \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1\right),$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i: x_i > 0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the lower endpoint $\underline{\widehat{V}}$ of the range of corresponding values of $\widehat{V}$.*

Comment. As we will see, this algorithm takes time $O(n^2 \cdot \log(n))$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n - p$ intervals, we sort the corresponding $2 \cdot (n - p)$ endpoints $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n-2p}.$$

Let us add $r_0 = -\infty$ and $r_{2n-2p+1} = +\infty$. This divides the real line into $2n - 2p + 1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n - 2p$. For each zone, we do the following:

- First, we compute the values

$$s_k^- = \sum_{i: \overline{y}_i \le r_k} \overline{y}_i; \quad s_k^+ = \sum_{j: r_{k+1} \le \underline{y}_i} \underline{y}_j;$$

$$M_k^- = \sum_{i: \overline{y}_i \le r_k} (\overline{y}_i)^2; \quad M_k^+ = \sum_{j: r_{k+1} \le \underline{y}_j} \left(\underline{y}_i\right)^2;$$

and the number n_k of all the indices i for which $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$.

- Then, we compute the value r from the equation

$$r = \widehat{\mu} - \frac{\exp\left(\widehat{\sigma}^2\right) + \dfrac{p}{n} - 1}{2\exp\left(\widehat{\sigma}^2\right) + \dfrac{p}{n} - 1}, \text{ where}$$

$$\widehat{\mu} = \frac{s_k^- + s_k^+ + n_k \cdot r}{n-p}; \quad \widehat{\sigma}^2 = \frac{M_k^- + M_k^+ + n_k \cdot r^2 - \left(\widehat{\mu}\right)^2}{n-p}.$$

- If $r_k \le r \le r_{k+1}$, we then compute the value

$$\widehat{V} = \left(1 - \frac{p}{n}\right) \cdot \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) + \frac{p}{n} - 1\right).$$

We then compute the smallest of all the resulting values $\widehat{V}$ (corresponding to all values p and to all zones), and return this smallest value as the desired value $\underline{\widehat{V}}$.

Proposition 21. *Let C be a positive integer, and let $\widehat{V}$ be the Maximum Likelihood estimate for the variance V corresponding to the delta-lognormal distribution:*

$$\widehat{V} = \left(1 - \widehat{d}\right) \cdot \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1\right),$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the upper endpoint $\overline{\widehat{V}}$ of the range of corresponding values of $\widehat{V}$.*

Comment. This algorithm also takes time $O(n^2 \cdot \log(n))$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n - p$ intervals, we sort the corresponding $2 \cdot (n-p)$ endpoints $\underline{y}_i = \ln\left(\underline{x}_i\right)$ and $\overline{y}_i = \ln\left(\overline{x}_i\right)$ into a non-decreasing sequence

$$r_1 \le r_2 \le \ldots \le r_{2n-2p}.$$

Let us add $r_0 = -\infty$ and $r_{2n-2p+1} = +\infty$. This divides the real line into $2n - 2p + 1$ *zones* $[r_k, r_{k+1}]$, $k = 0, 1, \ldots, 2n - 2p$. For each zone, for each value i, we select:

- the value $y_i = \underline{y}_i$ if $\overline{y}_i \le r_k$;
- the value $y_i = \overline{y}_i$ if $r_{k+1} \le \underline{y}_i$; and
- for each other index i, i.e., when $\underline{y}_i \le r_k \le r_{k+1} \le \overline{y}_i$, select one of the three possible cases: $y_i = \underline{y}_i$, $y_i = \overline{y}_i$, and $y_i = r$, for a constant r to be determined later.

For each of the resulting tuples, we find r from the equation

$$r = \widehat{\mu} - \frac{\exp\left(\widehat{\sigma}^2\right) + \dfrac{p}{n} - 1}{2\exp\left(\widehat{\sigma}^2\right) + \dfrac{p}{n} - 1},$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are sample mean and sample variance of the selected values y_i – thus, depending on r as well. Once we find r, we thus know the values of all selected y_i, so we can compute the values $\widehat{\mu}$, $\widehat{\sigma}^2$, and

$$\widehat{V} = \left(1 - \frac{p}{n}\right) \cdot \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) + \frac{p}{n} - 1\right).$$

Then, we compute the largest of all the resulting values $\widehat{V}$ (corresponding to all values p and to all zones), and return this largest value as the desired value $\overline{\widehat{V}}$.

9 Case of the Delta-Lognormal Distribution: Estimating Coefficient of Variation under Interval Uncertainty

Proposition 22. *Let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation $CV = \sqrt{V}E$ corresponding to the delta-lognormal distribution:*

$$\widehat{CV} = \sqrt{\frac{\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1}{1 - \widehat{d}}},$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, the following problem is NP-hard:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *compute the upper endpoint $\overline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Proposition 23. *Let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation $CV = \sqrt{V}E$ corresponding to the delta-lognormal distribution:*

$$\widehat{CV} = \sqrt{\frac{\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1}{1 - \widehat{d}}},$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given n intervals $[\underline{x}_i, \overline{x}_i]$,*
- *computes the lower endpoint $\underline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Comment. This algorithm takes quadratic time $O(n^2)$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n - p$ indices, we compute the intervals $\left[\underline{y}_i, \overline{y}_i\right]$, where $\underline{y}_i = \ln\left(\underline{x}_i\right)$ and $\overline{y}_i = \ln\left(\overline{x}_i\right)$. To these $n - p$ intervals, we then apply the known linear-time algorithm for computing the lower endpoint $\widehat{\sigma}^2$ for the sample variance, and compute the value

$$\widehat{CV} = \sqrt{\frac{\exp\left(\widehat{\sigma}^2\right) + \frac{p}{n} - 1}{1 - \frac{p}{n}}}.$$

The smallest of the resulting values $\widehat{CV}$ (corresponding to all possible values $p = 0, 1 \ldots, n_0$) is then returned as the desired value $\underline{\widehat{CV}}$.

Proposition 24. *Let C be a positive integer, and let $\widehat{CV}$ be the Maximum Likelihood estimate for the coefficient of variation CV corresponding to the delta-lognormal distribution:*

$$\widehat{CV} = \sqrt{\frac{\exp\left(\widehat{\sigma}^2\right) + \widehat{d} - 1}{1 - \widehat{d}}},$$

where

$$\widehat{d} = \frac{\#\{i : x_i = 0\}}{n}, \quad \widehat{\mu} = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i : x_i > 0} \ln(x_i);$$

$$\widehat{\sigma}^2 = \frac{1}{\#\{i : x_i > 0\}} \cdot \sum_{i:x_i>0} \left(\ln(x_i) - \widehat{\mu}\right)^2.$$

Then, there exists a feasible (polynomial time) algorithm that:

- *given a list of n intervals $[\underline{x}_i, \overline{x}_i]$ for which every sublist of C intervals has an empty intersection,*
- *computes the upper endpoint $\overline{\widehat{CV}}$ of the range of corresponding values of $\widehat{CV}$.*

Comment. This algorithm takes quadratic time $O(n^2)$.

Description of the corresponding algorithm. Let $n_0 \le n$ be the total number of intervals $[\underline{x}_i, \overline{x}_i]$ for which $\underline{x}_i = 0$. Let us sort all these intervals in the increasing order of the upper endpoints $\overline{x}_i$, so that we have $\underline{x}_1 = \ldots = \underline{x}_{n_0} = 0$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_{n_0}$.

For each $p = 0, 1, 2, \ldots, n_0$, we set $x_1 = \ldots = x_p = 0$. For the remaining $n-p$ indices, we compute the intervals $\left[\underline{y}_i, \overline{y}_i\right]$, where $\underline{y}_i = \ln(\underline{x}_i)$ and $\overline{y}_i = \ln(\overline{x}_i)$. To these $n - p$ intervals, we then apply the known linear-time algorithm for computing the upper endpoint $\widehat{\sigma}^2$ for the sample variance, and compute the value

$$\widehat{CV} = \sqrt{\frac{\exp(\widehat{\sigma}^2) + \dfrac{p}{n} - 1}{1 - \dfrac{p}{n}}}.$$

The largest of the resulting values $\widehat{CV}$ (corresponding to all possible values $p = 0, 1 \ldots, n_0$) is then returned as the desired value $\overline{\widehat{CV}}$.

10 Proofs

Proof of Proposition 1. By definition, $y_i = \ln(x_i)$, so $x_i = \exp(y_i)$. Thus, each x_i is a monotonic function of y_i and vice versa. So, $\widehat{E}$ is an increasing function of x_i if and only if it is an increasing function of y_i. Thus, to prove that $\widehat{E}$ is not always increasing in x_i, it is sufficient to prove that it is not always increasing in y_i.

A function is always increasing if its derivative is always non-negative. Thus, to prove that $\widehat{E}$ is not always an increasing function of y_i, it is sufficient to find situations where the derivative $\dfrac{\partial \widehat{E}}{\partial y_i}$ is negative.

We know that $\widehat{E} = \exp\left(\widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}\right)$. Due to the chain rule and to the fact that the derivative of $\exp(x)$ is this same function $\exp(x)$, we conclude that

$$\frac{\partial \widehat{E}}{\partial y_i} = \exp\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right) \cdot \frac{\partial}{\partial y_i}\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right).$$

Since $\exp(x) > 0$ for all real values x, we conclude that the sign of the derivative $\dfrac{\partial \widehat{E}}{\partial y_i}$ coincides with the sign of the second factor. Thus, to prove our proposition, it is sufficient to find the values $y_1, \ldots, y_n$ for which

$$\frac{\partial}{\partial y_i}\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right) < 0.$$

The derivative of the sum is equal to the sum of the derivatives, so

$$\frac{\partial}{\partial y_i}\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right) = \frac{\partial \widehat{\mu}}{\partial y_i} + \frac{1}{2} \cdot \frac{\partial \widehat{\sigma}^2}{\partial y_i}.$$

By definition, $\widehat{\mu} = \dfrac{1}{n} \cdot \sum_{i=1}^{n} y_i$, so $\dfrac{\partial \widehat{\mu}}{\partial y_i} = \dfrac{1}{n}$. Similarly, $\widehat{\sigma}^2 = \dfrac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - (\widehat{\mu})^2$, hence

$$\frac{\partial \widehat{\sigma}^2}{\partial y_i} = \frac{2y_i}{n} - 2\widehat{\mu} \cdot \frac{\partial \widehat{\mu}}{\partial y_i} = \frac{2y_i}{n} - \frac{2\widehat{\mu}}{n}.$$

Thus,

$$\frac{\partial}{\partial y_i}\left(\widehat{\mu} + \frac{\widehat{\sigma}^2}{2}\right) = \frac{1}{n} \cdot (1 + y_i - \widehat{\mu}).$$

So, if $y_i < \widehat{\mu} - 1$, this derivative is negative.

The values $y_i = \ln(x_i)$ are normally distributed with the mean μ. Thus, the values y_i for which $y_i < \widehat{\mu} - 1$ are possible. The proposition is proven.

Proof of Proposition 2. The function $y = \ln(x)$ is increasing. Thus, once we know the intervals $[\underline{x}_i, \overline{x}_i]$ of possible values of x_i, we can feasibly compute the intervals $\left[\underline{y}_i, \overline{y}_i\right] = [\ln(\underline{x}_i), \ln(\overline{x}_i)]$ – and vice versa, once we know the intervals $\left[\underline{y}_i, \overline{y}_i\right]$, we can compute the intervals $[\underline{x}_i, \overline{x}_i] = \left[\exp(\underline{y}_i), \exp(\overline{y}_i)\right]$. So, to prove that the problem of computing $\overline{\widehat{E}}$ from the intervals $[\underline{x}_i, \overline{x}_i]$ is NP-hard, it is sufficient to prove that the problem of computing $\overline{\widehat{E}}$ from the intervals $\left[\underline{y}_i, \overline{y}_i\right]$ is NP-hard.

Similarly, since the function $y = \ln(x)$ is increasing, once we find the range $\left[\underline{\widehat{E}}, \overline{\widehat{E}}\right]$, we can feasibly compute the range $[\underline{a}, \overline{a}]$ of the auxiliary expression $a = \ln(\widehat{E}) = \widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}$ as $[\underline{a}, \overline{a}] = \left[\ln(\underline{\widehat{E}}), \ln(\overline{\widehat{E}})\right]$. Vice versa, once we know the interval $[\underline{a}, \overline{a}]$, we can compute the interval $\left[\underline{\widehat{E}}, \overline{\widehat{E}}\right] = [\exp(\underline{a}), \exp(\overline{a})]$. So, to prove that the problem of computing $\overline{\widehat{E}}$ is NP-hard, it is sufficient to prove that the problem of computing $\overline{a}$ is NP-hard. Since computing the range of $a = \widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}$ is feasibly equivalent to computing the range of $a = \widehat{\sigma}^2 + 2\widehat{\mu}$, this NP-hardness follows from the following Lemma:

Lemma. *Let α be an arbitrary real number, and let $A \stackrel{\text{def}}{=} \widehat{\sigma}^2 + \alpha \cdot \widehat{\mu}$, where $\widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$ and $\widehat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \widehat{\mu})^2$. Then, the following problem in NP-hard:*

- *given n intervals $\left[\underline{y}_i, \overline{y}_i\right]$,*
- *compute the upper endpoint $\overline{A}$ of the range*

$$[\underline{A}, \overline{A}] = \left\{ A(y_1, \ldots, y_n) : y_i \in \left[\underline{y}_i, \overline{y}_i\right] \right\}.$$

So, once we prove the lemma, we will therefore prove the desired NP-hardness.

Proof of the Lemma.

1°. By definition, a problem $\mathcal{P}$ is NP-hard if every problem $\mathcal{P}'$ from the class NP can be reduced to $\mathcal{P}$. Thus, to prove that a problem $\mathcal{P}$ is NP-hard, it is sufficient to prove that a known NP-hard problem $\mathcal{P}_0$ can be reduced to $\mathcal{P}$. Indeed, in this case, every problem $\mathcal{P}'$ from the class NP can be reduced to $\mathcal{P}_0$, and since $\mathcal{P}_0$ can be reduced to $\mathcal{P}$, we can therefore conclude – by using transitivity of problem reduction – that $\mathcal{P}'$ can be reduced to $\mathcal{P}$.

2°. As a known NP-hard problem $\mathcal{P}_0$, we will take the following *partition* problem which is known to be NP-hard [9, 11]: given q positive integers $s_1, \ldots, s_q$, check whether exist q signs $\varepsilon_i \in \{-1, 1\}$ for which $\varepsilon_1 \cdot s_1 + \ldots + \varepsilon_q \cdot s_q = 0$.

We will use the following reduction of this problem $\mathcal{P}_0$ to our problem. For every instance $s_1, \ldots, s_q$ of the partition problem, we form the following tuple of $n = q + 1$ intervals: $\left[\underline{y}_1, \overline{y}_1\right] = [-s_1, s_1]$, $\ldots$, $\left[\underline{y}_q, \overline{y}_q\right] = [-s_q, s_q]$, and $\left[\underline{y}_n, \overline{y}_n\right] = \left[\frac{\alpha \cdot n}{2}, \frac{\alpha \cdot n}{2}\right]$.

3°. Let us show that for the resulting instance of our problem, always $\overline{A} \leq A_0 \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2 + \frac{\alpha^2 \cdot n}{4} + \frac{\alpha^2}{4}$, and $\overline{A} = A_0$ if and only if the original partition problem has a solution.

3.1°. Let us first prove that always $\overline{A} \leq A_0$.

Indeed, by definitions of the quantities A and $\widehat{\sigma}^2$, we have

$$A = \widehat{\sigma}^2 + \alpha \cdot \widehat{\mu} = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - (\widehat{\mu})^2 \right) + \alpha \cdot \widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - (\widehat{\mu})^2 + \alpha \cdot \widehat{\mu} =$$

$$\frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 + \frac{1}{n} \cdot y_n^2 - \left(\widehat{\mu} - \frac{\alpha}{2}\right)^2 + \frac{\alpha^2}{4}.$$

Since $0 \le |y_i| \le s_i$ for all $i \le q$, we get $y_i^2 = |x_i|^2 \le s_i^2$ and thus, $\frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 \le \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2$. We know that $x_n = \frac{\alpha \cdot n}{2}$, so $\frac{1}{n} \cdot x_n^2 = \frac{\alpha^2 \cdot n}{4}$. Since the square $\left(\widehat{\mu} - \frac{\alpha}{2}\right)^2$ is always non-negative, this implies that

$$A \le \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2 + \frac{\alpha^2 \cdot n}{4} + \frac{\alpha^2}{4},$$

i.e., that always $A \le A_0$. Since this inequality holds for every A, it also holds for the largest possible value $\overline{A}$ of the quantity A. Thus, we proved that $\overline{A} \le A_0$.

3.2°. If the original instance of the partition problem has a solution ε_i, then we can take $y_i = \varepsilon_i \cdot s_i$ for $i \le q$ and $y_n = \frac{\alpha \cdot n}{2}$. In this case, for every i, we have $y_i^2 = s_i^2$ and therefore, $\frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 \le \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2$ and thus,

$$\frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 = \frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 + \frac{1}{n} \cdot y_n^2 \le \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2 + \frac{\alpha^2 \cdot n}{4}.$$

Here, $\sum_{i=1}^{q} y_i = \sum_{i=1}^{q} \varepsilon_i \cdot s_i = 0$ and thus,

$$\widehat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{1}{n} \cdot \left(\sum_{i=1}^{q} y_i + y_n \right) = \frac{1}{n} \cdot y_n = \frac{1}{n} \cdot \frac{\alpha \cdot n}{2} = \frac{\alpha}{2}.$$

Hence, $\widehat{\mu} - \frac{\alpha}{2} = 0$, and so,

$$A = \frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 + \frac{\alpha^2 \cdot n}{4} - \left(\widehat{\mu} - \frac{\alpha}{2}\right)^2 + \frac{\alpha^2}{4} = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 + \frac{\alpha^2 \cdot n}{4} + \frac{\alpha^2}{4} = A_0.$$

3.3°. Vice versa, let us assume that $\overline{A} = A_0$. Since $\overline{A}$ is the maximum of a continuous function A on a closed bounded box $\left[\underline{y}_1, \overline{y}_1\right] \times \ldots \times \left[\underline{y}_n, \overline{y}_n\right]$, this maximum is attained for some point within the box, i.e., there exist values $y_i \in \left[\underline{y}_i, \overline{y}_i\right]$ for which $A = A_0$, i.e.,

$$\frac{1}{n} \cdot \sum_{i=1}^{q} y_i^2 + \frac{\alpha^2 \cdot n}{4} - \left(\widehat{\mu} - \frac{\alpha}{2}\right)^2 + \frac{\alpha^2}{4} = A_0 = \frac{1}{n} \cdot \sum_{i=1}^{q} s_i^2 + \frac{\alpha^2 \cdot n}{4} + \frac{\alpha^2}{4}.$$

Here, $y_i^2 \le s_i^2$ for all $i \le q$, so if $y_i^2 < s_i^2$ or $\widehat{\mu} - \frac{\alpha}{2} \ne 0$, we would get $A < A_0$. Thus, the fact that $A = A_0$ means that $y_1^2 = s_1^2$, $\ldots$, $y_q^2 = s_q^2$, and $\widehat{\mu} = \frac{\alpha}{2}$.

The condition $y_i^2 = s_i^2$ means that $y_i = \pm s_i$, i.e., that $y_i = \varepsilon_i \cdot s_i$ for some $\varepsilon_i \in \{-1, 1\}$. In this case, the condition $\widehat{\mu} = \dfrac{\alpha}{2}$ means that

$$\frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{1}{n} \cdot \sum_{i=1}^{q} \varepsilon_i \cdot s_i + \frac{1}{n} \cdot y_n = \frac{\alpha}{2}.$$

Since $y_n = \dfrac{\alpha \cdot n}{2}$, we have $\dfrac{1}{n} \cdot y_n = \dfrac{\alpha}{2}$, and thus,

$$\frac{1}{n} \cdot \sum_{i=1}^{q} \varepsilon_i \cdot s_i + \frac{\alpha}{2} = \frac{\alpha}{2}.$$

Canceling the term $\dfrac{\alpha}{2}$ in both sides and multiplying both sides of the resulting inequality by n, we conclude that $\sum\limits_{i=1}^{q} \varepsilon_i \cdot s_i = 0$, i.e., that we have a solution to the given instance of the partition problem.

The lemma is proven.

Proof of Proposition 3 can be obtained from the proof of Proposition 2 similarly to [7].

Proof of Proposition 4.

1°. To prove this proposition, let us first recall known facts from calculus: namely, when a function of one variable attains minimum and maximum on the interval.

It is known that a function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its minimum on this interval either at lone of its endpoints, or in some internal point of the interval.

If it attains is minimum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$.

If it attains its minimum at the point $x = \underline{x}$, then we cannot have $\dfrac{df}{dx} < 0$, because then, for some point $x + \Delta x \in [\underline{x}, \overline{x}]$, we would have a smaller value of $f(x)$. Thus, in this case, we must have $\dfrac{df}{dx} \geq 0$.

Similarly, if a function $f(x)$ attains its minimum at the point $x = \overline{x}$, then we must have $\dfrac{df}{dx} \leq 0$.

2°. Let us apply these known facts to our problem. For the estimate $\widehat{E}$, as we have shown in the proof of Proposition 1, the sign of the partial derivative $\dfrac{\partial \widehat{E}}{\partial y_i}$ coincided with the sign of the difference $1 + y_i - \widehat{\mu}$. Thus, for the point $(y_1, \ldots, y_n)$ at which the estimate $\widehat{E}$ attains its minimum, we can make the following conclusions:

- if $y_i = \underline{y}_i$, then $y_i \geq \widehat{\mu} - 1$;
- if $y_i = \overline{y}_i$, then $y_i \leq \widehat{\mu} - 1$;
- if $\underline{y}_i < y_i < \overline{y}_i$, then $y_i = \widehat{\mu} - 1$.

3°. In terms of the relation between the value $\widehat{\mu} - 1$ and the endpoints of the interval $\left[\underline{y}_i, \overline{y}_i\right]$ and the value $\widehat{\mu} - 1$, there are three possible cases:

- the case when the interval is fully to the left of the value $\widehat{\mu} - 1$, i.e., when $\overline{y}_i < \widehat{\mu} - 1$;
- the case when the interval is fully to the right of the value $\widehat{\mu} - 1$, i.e., when $\widehat{\mu} - 1 < \underline{y}_i$; and
- the case when the value $\widehat{\mu} - 1$ belongs to the interval.

Let us consider these three cases one by one.

3.1°. Let us first consider the case when $\overline{y}_i < \widehat{\mu} - 1$. In this case, the value $y_i \leq \overline{y}_i$ also satisfies the inequality $y_i < \widehat{\mu} - 1$. Thus, in this case:

- we cannot have $y_i = \underline{y}_i$ — because then we would have $y_i \geq \widehat{\mu} - 1$; and
- we cannot have $\underline{y}_i < y_i < \overline{y}_i$ – because then, we would have $y_i = \widehat{\mu} - 1$.

So, if $\overline{y}_i < \widehat{\mu} - 1$, the only remaining option for y_i is $y_i = \overline{y}_i$.

3.2°. Let us now consider the case when $\widehat{\mu} - 1 < \underline{y}_i$. In this case, the value $y_i \geq \underline{y}_i$ also satisfies the inequality $\widehat{\mu} - 1 < y_i$. Thus, in this case:

- we cannot have $y_i = \overline{y}_i$ — because then we would have $y_i \leq \widehat{\mu} - 1$; and
- we cannot have $\underline{y}_i < y_i < \overline{y}_i$ – because then, we would have $y_i = \widehat{\mu} - 1$.

So, if $\widehat{\mu} - 1 < \underline{y}_i$, the only remaining option for y_i is $y_i = \underline{y}_i$.

3.3°. Let us now consider the case when $\overline{y}_i \leq \widehat{\mu} - 1 \leq \overline{y}_i$. In this case, the optimizing value y_i can either coincide with the left endpoint, or with the right endpoint, or with some value in between. Let us consider these three subcases one by one.

- If $y_i = \underline{y}_i$, then $y_i = \underline{y}_i \geq \widehat{\mu} - 1$. Since $\underline{y}_i \leq \widehat{\mu} - 1$, this implies that $y_i = \widehat{\mu} - 1$.
- If $y_i = \overline{y}_i$, then $y_i = \overline{y}_i \leq \widehat{\mu} - 1$. Since $\overline{y}_i \geq \widehat{\mu} - 1$, this implies that $y_i = \widehat{\mu} - 1$.
- Finally, $\underline{y}_i < y_i < \overline{y}_i$, then $y_i = \widehat{\mu} - 1$.

In all three subcases, we have $y_i = \widehat{\mu} - 1$.

4°. So, if we know the location of $\widehat{\mu} - 1$ in relation to the endpoints of all the intervals, i.e., if we know the zone $[r_k, r_{k+1}]$ that contains this value, we can determine the optimizing values as follows:

- If $\overline{y}_i \le \widehat{\mu} - 1$, then $y_i = \overline{y}_i$.
- If $\widehat{\mu} - 1 \le \underline{y}_i$, then $y_i = \underline{y}_i$.
- In all other cases, $y_i = \widehat{\mu} - 1$.

The corresponding value $\widehat{\mu}$ can be determined from the fact that $\widehat{\mu}$ is, by definition, an arithmetic average of the corresponding values y_i, i.e., that $y_1 + \ldots + y_n = n \cdot \widehat{\mu}$. Substituting the above optimizing values y_i into this formula, we conclude that

$$\sum_{i:\overline{y}_i \le r_k} \overline{y}_i + \sum_{j:r_{k+1} \le \underline{y}_i} \underline{y}_i + n_k \cdot (\widehat{\mu} - 1) = n \cdot \widehat{\mu},$$

where n_k is the total number of indices for which the interval $\left[\underline{y}_i, \overline{y}_i\right]$ contains the zone $[r_k, r_{k+1}]$. This leads to the formula for the $\widehat{\mu}$ that we provided in the description of the algorithm.

First, we check whether this value is indeed located in the given zone. If it is, then we compute the corresponding value $\widehat{E} = \exp(A)$ where $A = \widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}$. We have already computed the value $\widehat{\mu}$. To compute $\widehat{\sigma}^2$, we use the formula $\widehat{\sigma}^2 = M - (\widehat{\mu})^2$, where $M = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} y_i^2$. Substituting the above optimizing values y_i into this formula, we conclude that $M = \dfrac{1}{n} \cdot \left(M_k^- + M_k^+ + n_k \cdot (\widehat{\mu} - 1)^2\right)$, i.e., exactly the expression used in the algorithm.

5°. The actual value $\widehat{\mu} - 1$ belongs to some zone, so to find the desired minimum $\underline{\widehat{E}}$, it is sufficient to consider all possible zones.

6°. To complete the proof of the proposition, it is sufficient to show that our algorithm indeed takes time $O(n \cdot \log(n))$.

Indeed, sorting takes time $O(n \cdot \log(n))$; see, e.g., [5]. We have $2n + 1 = O(n)$ zones, for each of which computing the values $s_k^{\pm}$, $M_k^{\pm}$, etc., takes linear time. However, we only need linear time for computing the first value. After that, we change, e.g., from s_0^- to s_1^-, from s_1^- to s_2^-, etc., by changing a new terms. Each of n terms $\overline{y}_i$ is changed only once; once it is changed, it remains. So, overall, for all $2n$ zones, we need only $O(n)$ steps. Thus, the total computation time consists of three parts:

- time $O(n \cdot \log(n))$ for sorting;
- time $O(n)$ for computing the initial values of all the parameters $s_k^{\pm}$, $M_k^{\pm}$, etc., corresponding to the first zone $k = 0$;
- time $O(n)$ for computing all the remaining values of these parameters;
- time $O(n)$ to compute all the values A_k and their minimum;

- time $O(1)$ to compute $\widehat{E} = \exp(A)$.

The total amount of computation time is

$$O(n \cdot \log(n)) + O(n) + O(n) + O(n) + O(1) = O(n \cdot \log(n)).$$

The proposition is proven.

Proof of Proposition 5.

1°. As we have mentioned in the previous proofs, computing the maximum of $\widehat{E} = \exp(A)$, where $A = \widehat{\mu} + \dfrac{\widehat{\sigma}^2}{2}$, is equivalent to computing the maximum of the auxiliary quantity A.

2°. The function A is a convex function of all its variables and therefore, its maximum on a convex set

$$\left[\underline{y}_1, \overline{y}_1\right] \times \ldots \left[\underline{y}_n, \overline{y}_n\right]$$

is attained at one of the vertices, i.e., at a point $(y_1, \ldots, y_n)$ at which each value y_i is equal either to $\underline{y}_i$ or to $\overline{y}_i$.

3°. Similarly to the previous proof, from calculus, we conclude that:

- if the maximum is attained at $y_i = \overline{y}_i$, then $\dfrac{\partial A}{\partial y_i} \geq 0$, hence $\overline{y}_i \geq \mu - 1$;
- if the maximum is attained at $y_i = \underline{y}_i$, then $\dfrac{\partial A}{\partial y_i} \leq 0$, hence $\underline{y}_i \leq \mu - 1$.

Thus:

- if $\overline{y}_i < \mu - 1$, then we cannot have $y_i = \overline{y}_i$, because otherwise, we would have $\overline{y}_i \geq \mu - 1$; thus, we have $y_i = \underline{y}_i$;
- if $\mu - 1 < \underline{y}_i$, then we cannot have $y_i = \underline{y}_i$ because otherwise, we would have $\underline{y}_i \leq \mu - 1$; thus, we have $y_i = \overline{y}_i$.

So, for each of n zones, only for indices i for which this zone intersects with the corresponding interval $\left[\underline{y}_i, \overline{y}_i\right]$, we have two options. Because of the condition on the intervals, there are no more than C such intervals, so we have $\leq 2^C = O(1)$ combinations of the corresponding lower and upper endpoints.

This leads us to the above algorithm.

4°. Let us estimate the computation time of this algorithm. For each of $2n+1 = O(n)$ zones, we have $O(1)$ combinations, so totally, we need $O(n)$ combinations. For each combination, we need linear time $O(1)$ to compute the corresponding value A, but in reality, we only need linear time for the first combination; after that, we make finitely many changes. Thus, similarly to the proof of Proposition 4, we need:

- time $O(n \cdot \log(n))$ to sort the endpoints;
- time $O(n)$ to compute the initial values of $\widehat{\mu}$, $\widehat{\sigma}^2$, and A:
- time $O(n)$ to compute all consequent values and to find the largest value $A_{\max}$:
- time $O(1)$ to compute $\overline{\widehat{E}} = \exp(A_{\max})$.

The total computation time is

$$O(n \cdot \log(n)) + O(n) + O(n) + O(1) = O(n \cdot \log(n)).$$

The proposition is proven.

Proof of Propositions 6 and 7. Since $\exp(x)$ is an increasing function, computing the range of $\widehat{m}_0$ is equivalent to computing the range of the difference $\widehat{\mu} - \widehat{\sigma}^2$. Thus, computing $\underline{\widehat{m}_0}$ is equivalent to computing the lower endpoint for $\widehat{\mu} - \widehat{\sigma}^2$. This, in its turn, is equivalent to computing the upper endpoint $\overline{A}$ for

$$A = -(\widehat{\mu} - \widehat{\sigma}^2) = \widehat{\sigma}^2 - \widehat{\mu}.$$

We have already proved, in the Lemma, that computing $\overline{A}$ is NP-hard. Thus, the problem of computing $\underline{\widehat{m}_0}$ is NP-hard as well. This proves Proposition 6. Similarly to [7], we can now prove Proposition 7.

Proof of Propositions 8 and 9. This proof is similar to the proof of Propositions 5 and 6, the only difference is that here, we have

$$\frac{\partial A}{\partial y_i} = \frac{1}{n} - \frac{2(y_i - \widehat{\mu})}{n} = -\frac{2}{n} \cdot \left(y_i - \widehat{\mu} - \frac{1}{2}\right),$$

and thus, the sign of the partial derivative $\dfrac{\partial \widehat{m}_0}{\partial y_i}$ is opposite to the sign of the expression $y_i - \widehat{\mu} - \dfrac{1}{2}$.

Proof of Proposition 10. To prove NP-hardness of this problem, let us reduce, to this problem, a problem which (as we have mentioned earlier) is known to be NP-hard: the problem of approximately computing the upper endpoint $\overline{\widehat{\sigma^2}}$ for the sample variance $\widehat{\sigma}^2$ under interval uncertainty. Indeed, let us assume that we know how to compute the upper bound $\overline{\widehat{V}}$. Let us show how, using these computations an "oracle", we can compute the value $\overline{\widehat{\sigma^2}}$ for a given tuple of intervals $\left[\underline{y}_i, \overline{y}_i\right]$ with any given accuracy.

First, let us note that computing the upper endpoint $\overline{\widehat{V}}$ for

$$\widehat{V} = \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right)$$

is feasibly equivalent to the problem of computing the upper endpoint $\overline{A}$ for an auxiliary quantity

$$A = \ln\left(\widehat{V}\right) = 2\widehat{\mu} + \widehat{\sigma}^2 + \ln\left(\exp\left(\widehat{\sigma}^2\right) - 1\right).$$

For an arbitrary real number $K > 0$, we can introduce new variables $z_i = K \cdot y_i$. When we multiply all the sample values y_i by K, the sample average $\widehat{\mu}$ gets multiplied by K, while the sample sample variance $\widehat{\sigma}^2$ gets multiplied by K^2. In other words, the z-based values $\widehat{\mu}_z$ and $\widehat{\sigma}_z^2$ are related to the original values $\widehat{\mu}$ and $\widehat{\sigma}^2$ by the formulas $\widehat{\mu}_z = K \cdot \widehat{\mu}$ and $\widehat{\sigma}_z^2 = K^2 \cdot \widehat{\sigma}^2$. Thus, the value $A(K)$ of the quantity A corresponding to $z_i = K \cdot y_i$ in terms of K and y_i, we get

$$A(K) = 2K \cdot \widehat{\mu} + K^2 \cdot \widehat{\sigma}^2 + \ln\left(\exp\left(K^2 \cdot \widehat{\sigma}^2\right) - 1\right).$$

Computing the upper endpoint of $A(K)$ is feasibly equivalent to computing the upper endpoint for

$$a(K) \stackrel{\text{def}}{=} K^{-2} \cdot A(K) = 2K^{-1} \cdot \widehat{\mu} + \widehat{\sigma}^2 + K^{-2} \cdot \ln\left(\exp\left(K^2 \cdot \widehat{\sigma}^2\right) - 1\right).$$

Here, we have

$$\exp\left(K^2 \cdot \widehat{\sigma}^2\right) - 1 = \exp\left(K^2 \cdot \widehat{\sigma}^2\right) \cdot \left(1 - \exp\left(-K^2 \cdot \widehat{\sigma}^2\right)\right).$$

Since logarithm of the product is equal to the sum of the logarithms, and logarithm of $\exp(t)$ is always equal to t, we get

$$\ln\left(\exp\left(K^2 \cdot \widehat{\sigma}^2\right) - 1\right) = K^2 \cdot \widehat{\sigma}^2 + \ln\left(1 - \exp\left(-K^2 \cdot \widehat{\sigma}^2\right)\right),$$

so

$$a(K) = 2K^{-1} \cdot \widehat{\mu} + 2\widehat{\sigma}^2 + \ln\left(1 - \exp\left(-K^2 \cdot \widehat{\sigma}^2\right)\right).$$

When $K \to \infty$, we have $K^{-1} \to 0$, $\exp\left(-K^2 \cdot \widehat{\sigma}^2\right) \to 0$ and hence, $\ln\left(1 - \exp\left(-K^2 \cdot \widehat{\sigma}^2\right)\right) \to \ln(1-0) = 0$. Thus, when $K \to \infty$, we get $a(K) \to 2\widehat{\sigma}^2$. So, for large K, the value $a(K)$ is approximately equal to $2\widehat{\sigma}^2$. Whatever approximation accuracy we want to achieve, we can do it by selecting appropriate K, and this appropriate K can be selected feasibly.

So, if we could compute the upper bound $\overline{\widehat{V}}$ for $\widehat{V}$, we would then be able to compute the upper bound for $A(K)$, for $a(K)$ and thus, an approximate upper endpoint $\overline{\widehat{\sigma^2}}$ for $\widehat{\sigma}^2$. We already know that approximate computation of the upper endpoint $\overline{\widehat{\sigma^2}}$ is NP-hard. Thus, we have indeed reduced a known NP-hard problem to the problem of computing $\overline{\widehat{V}}$.

This shows that the problem of computing $\overline{\widehat{V}}$ is indeed NP-hard. The proposition is proven.

Proof of Proposition 11. This proof is similar to the proof of Propositions 5 and 8, the only difference is that here, we have

$$\frac{\partial \widehat{V}}{\partial y_i} = \frac{\partial}{\partial y_i}\left[\exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right)\right] =$$

$$\exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \left(\exp\left(\widehat{\sigma}^2\right) - 1\right) \cdot \frac{2 + 2y_i - 2\widehat{\mu}}{n} + \exp\left(2\widehat{\mu} + \widehat{\sigma}^2\right) \cdot \exp\left(\widehat{\sigma}^2\right) \cdot \frac{2y_i - 2\widehat{\mu}}{n} =$$

$$\exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot\left[\left(\exp\left(\widehat{\sigma}^2\right)-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+\exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n}\right].$$

Thus, the sign of his derivative coincides with the sign of the expression

$$\left(\exp\left(\widehat{\sigma}^2\right)-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+\exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n}.$$

This expression is an increasing linear function of y_i which changes sign from negative to positive when

$$\left(\exp\left(\widehat{\sigma}^2\right)-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+\exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n}=0,$$

i.e., when

$$y_i\cdot\left(2\exp\left(\widehat{\sigma}^2\right)-1\right)=\widehat{\mu}\cdot\left(2\exp\left(\widehat{\sigma}^2\right)-1\right)-\left(\exp\left(\widehat{\sigma}^2\right)-1\right),$$

or

$$y_i=r,\ \text{where}\ r=\widehat{\mu}-\frac{\exp\left(\widehat{\sigma}^2\right)-1}{2\exp\left(\widehat{\sigma}^2\right)-1}.$$

So, similarly to the proof of Proposition 5, for the minimizing tuple $y=(y_1,\ldots,y_n)$, we have $y_i=\overline{y}_i$ when $\overline{y}_i\le r$, we have $y_i=\underline{y}_i$ when $r\le\underline{y}_i$, and $y_i=r$ for all other indices i. The proposition is proven.

Proof of Proposition 12. In the previous proof, we have found expression for the sign of the partial derivatives $\dfrac{\partial V}{\partial y_i}$: it coincides with the sign of the difference

$$y_i-r,\ \text{where}\ r=\mu+\frac{\exp\left(\widehat{\sigma}^2\right)-1}{2\exp\left(\widehat{\sigma}^2\right)-1}.$$

According to calculus, for the tuple $y=(y_1,\ldots,y_n)$ at which the maximum of V is attained, we have three possibilities:

- the first possibility is $y_i=\underline{y}_i$; in this case, $\dfrac{\partial V}{\partial y_i}\le 0$, hence $y_i\le r$;
- the second possibility is $y_i=\overline{y}_i$; in this case, $\dfrac{\partial V}{\partial y_i}\ge 0$, hence $y_i\ge r$;
- the third possibility is $\underline{y}_i<y_i<\overline{y}_i$; in this case, $\dfrac{\partial V}{\partial y_i}=0$, hence $y_i=r$.

So, if $\overline{y}_i<r$, this means that we cannot have $y_i=\overline{y}_i$ and we cannot have $\underline{y}_i<y_i<\overline{y}_i$, so we must have $y_i=\underline{y}_i$.

Similarly, if $r<\underline{y}_i$, then we cannot have $y_i=\underline{y}_i$ and we cannot have $\underline{y}_i<r<\overline{y}_i$, so we must have $y_i=\overline{y}_i$.

In all other cases, we can have three options: $y_i=\underline{y}_i$, $y_i=\overline{y}_i$, and $y_i=r$. This justifies the algorithm – in which we enumerate all such tuples.

The fact that this algorithm takes time $O(n \cdot \log(n))$ is proven similarly to the proof of Proposition 5, with the only difference that since each "other" index has three choices, the number of selected tuples corresponding to each zone is now bounded not by 2^C, but by 3^C, which is still $O(1)$. The proposition is proven.

Proof of Proposition 16: for this proof, we can use the proof of Proposition 2, since in that proof, we have $\underline{x}_i > 0$ and thus, $\widehat{d} = 0$. In this case, delta-lognormal distribution turns into a lognormal one.

Proof of Proposition 17. We want to find the range of the estimate $\widehat{E}$ when $x \in [\underline{x}_i, \overline{x}_i]$. One of the quantities used in the formula for $\widehat{E}$ is the quantity $\widehat{\mu} = \dfrac{p}{n}$, where p is the total number of indices for which $x_i = 0$. For each index i, the value $x_i = 0$ is not possible when $\underline{x}_i > 0$; the value $x_i = 0$ is only possible when $\underline{x}_i = 0$. Thus, $p \le n_0$.

So, to find the desired minimum, it is sufficient to consider all possible values $p = 0, 1, \ldots, n_0$, find the minimum value of E for this p, and then find the smallest of the corresponding minima.

For each value p, we may have several possible ways of selecting p out of n_0 indices i for which we select $x_i = 0$. If we select an index i and do not select an index j for which $\overline{x}_i > \overline{x}_j$, this means that we consider all the tuples for which $x_i = 0$ and $x_j \in [0, \overline{x}_j]$. If instead we select $x_j = 0$ and $x_i \in [0, \overline{x}_i]$, then, due to $[0, \overline{x}_j] \subset [0, \overline{x}_i]$, we will get a larger set of tuples and, potentially, a smaller minimum E. Thus, to find an arrangement for which E is the smallest, it is sufficient to consider arrangements in which once we set $x_i = 0$ for some i, we select $x_j = 0$ for all j for which $\underline{x}_j = 0$ and $\overline{x}_j < \overline{x}_i$. In other words, we assign $x_i = 0$ for all the indices for which $\overline{x}_i$ is smaller than a certain threshold. This is exactly what we do in our algorithm.

For each value p, as shown in the proof of Proposition 4, we take $O(n \cdot \log(n))$ steps. There are $n_0 + 1 \le n + 1 = O(n)$ different values p, so overall, this algorithm takes time $O(n) \cdot O(n \cdot \log(n)) = O(n^2 \cdot \log(n))$. The proposition is proven.

Proof of Proposition 18 can be obtained from the proof of Proposition 5 in the same way as the proof of Proposition 17 was obtained from the proof of Proposition 4.

Proof of Proposition 19: for this proof, we can use the proof of Proposition 10, since in that proof, we have $\underline{x}_i > 0$ and thus, $\widehat{d} = 0$. In this case, delta-lognormal distribution turns into a lognormal one.

Proof of Propositions 20 and 21. Similarly to the proof of Proposition 17, we can reduce the problem of computing the desired bound to the problems of computing the bound for each fixed p, and similarly, for each p, the corresponding bound is attained when we assign $x_i = 0$ to the indices for which the upper endpoints $\overline{x}_i$ are the smallest.

For each p, we can, similarly to the proof of Proposition 11, differentiate the

delta-lognormal expression for $\widehat{V}$ by y_i. As a result, we get

$$\frac{\partial \widehat{V}}{\partial y_i} = \frac{\partial}{\partial y_i}\left[\left(1-\frac{p}{n}\right)\cdot \exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot\left(\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)\right] =$$

$$\left(1-\frac{p}{n}\right)\cdot \exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot\left(\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+$$

$$\left(1-\frac{p}{n}\right)\cdot \exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot \exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n} =$$

$$\left(1-\frac{p}{n}\right)\cdot \exp\left(2\widehat{\mu}+\widehat{\sigma}^2\right)\cdot A,$$

where

$$A = \left(\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+\exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n}.$$

Thus, the sign of his derivative coincides with the sign of the expression A. This expression is an increasing linear function of y_i which changes sign from negative to positive when

$$\left(\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)\cdot\frac{2+2y_i-2\widehat{\mu}}{n}+\exp\left(\widehat{\sigma}^2\right)\cdot\frac{2y_i-2\widehat{\mu}}{n}=0,$$

i.e., when

$$y_i\cdot\left(2\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)=\widehat{\mu}\cdot\left(2\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right)-\left(\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1\right),$$

or

$$y_i = r,\ \text{where } r=\widehat{\mu}-\frac{\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1}{2\exp\left(\widehat{\sigma}^2\right)+\frac{p}{n}-1}.$$

After that, the proof is similar to the proofs of Propositions 11 and 12.

Proof of Proposition 22: for this proof, we can use the proof of Proposition 13, since in that proof, we have $\underline{x}_i > 0$ and thus, $\widehat{d}=0$. In this case, delta-lognormal distribution turns into a lognormal one.

Proof of Propositions 23 and 24. Similarly to the proof of Proposition 17, we can reduce the problem of computing the desired bound to the problems of computing the bound for each fixed p, and similarly, for each p, the corresponding bound is attained when we assign $x_i = 0$ to the indices for which the upper endpoints $\overline{x}_i$ are the smallest.

For each p, the expression for $\widehat{CV}$ is a monotonic function of the sample variance $\widehat{\sigma}^2$; thus:

- its largest value $\overline{\widehat{CV}}$ is attained when the sample variance $\widehat{\sigma}^2$ attains its largest possible value $\overline{\widehat{\sigma}^2}$, and

- its smallest value $\underline{\widehat{CV}}$ is attained when the sample variance $\widehat{\sigma}^2$ attains its smallest possible value $\underline{\widehat{\sigma}^2}$.

Thus, for each p, we can use the known linear-time algorithms [16, 21] for computing the endpoints $\overline{\widehat{\sigma}^2}$ and $\underline{\widehat{\sigma}^2}$, and then compute the values $\widehat{CV}$ corresponding to these endpoints.

For each p, the computation takes linear time $O(n)$. As we have mentioned in the proof of Proposition 17, there are $O(n))$ different values p. Thus, totally, this algorithm takes time $O(n) \cdot O(n) = O(n^2)$.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health. The work of N. Buntao was supported by a grant from the Office of the Higher Education Commission, Thailand, under the Strategic Scholarships for Frontier Research Network, and by the Graduate College, King Mongkut's University of Technology North Bangkok.

The authors are thankful to Hung T. Nguyen for valuable discussions.

References

[1] J. Aitchison and J. Brown, *The Lognormal Distribution*, Cambridge University Press, Cambridge, UK, 1969.

[2] N. Buntao, "Estimating Parameters of Pareto Distribution under Interval and Fuzzy Uncertainty", *Proceedings of the 30th Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2011*, El Paso, Texas, March 18–20, 2011.

[3] N. Buntao and V. Kreinovich, "Measures of Deviation (and Dependence) for Heavy-Tailed Distributions and their Estimation under Interval and Fuzzy Uncertainty", In: R. R. Yager, M. Z. Reformat, S. N. Shahbazova, and S. Ovchinnikov (eds.), *Proceedings of the World Conference on Soft Computing*, San Francisco, CA, May 23–26, 2011.

[4] M. A. Cantos Rosales, *The Robustness of Confidence Intervals for the Mean of Delta Distribution*, PhD Dissertation, Department of Statistics, Western Michigan University, Kalamazoo, Michigan, 2009.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stain, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, 2009.

[6] E. Dantsin, V. Kreinovich, A. Wolpert, and G. Xiang, "Population Variance under Interval Uncertainty: A New Algorithm", *Reliable Computing*, 2006, Vol. 12, No. 4, pp. 273–280.

[7] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact bounds on finite populations of interval data, *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.

[8] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories, Report SAND2007-0939, May 2007.

[9] M. R. Garey and D. S. Johnson, *Computers and Intractability, a Guide to the Theory of NP-Completeness*, W.H. Freemand and Company, San Francisco, CA, 1979.

[10] B. Gladysz and A. Kasperski, "Computing mean absolute deviation under uncertainty", *Applied Soft Computing*, 2010, Vol. 10, pp. 361–366.

[11] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.

[12] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, "Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases", *Journal of Computational and Applied Mathematics*, 2007, Vol. 199, No. 2, pp. 418–423.

[13] V. Kreinovich, H. T. Nguyen, and S. Niwitpong, "Statistical Hypothesis Testing Under Interval Uncertainty: An Overview", *International Journal of Intelligent Technologies and Applied Statistics*, 2008, Vol. 1, No. 1, pp. 1–32.

[14] V. Kreinovich and G. Xiang, "Fast Algorithms for Computing Statistics under Interval Uncertainty: An Overview", In: V.-N. Huynh, Y. Nakamori, H. Ono, J. Lawry, V. Kreinovich, and H. T. Nguyen (eds.), *Interval/Probabilistic Uncertainty and Non-Classical Logics*, Springer-Verlag, Berlin-Heidelberg-New York, 2008, pp. 19–31.

[15] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, "Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity", *Reliable Computing*, 2006, Vol. 12, No. 6, pp. 471–501.

[16] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.

[17] S. Niwitpong, H. T. Nguyen, I. Neumann, and V. Kreinovich, "Hypothesis testing with interval data: case of regulatory constraints", *International Journal of Intelligent Technologies and Applied Statistics*, 2008, Vol. 1, No. 2, pp. 19-41.

[18] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, American Institute of Physics, New York, 2005.

[19] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[20] L. Tian and J. Wu, "Confidence Intervals for the Mean of Lognormal Data with Excess Zeros", *Biometrical Journal*, 2006, Vol. 48, No. 1, pp. 149–156.

[21] G. Xiang, M. Ceberio, and V. Kreinovich, "Computing Population Variance and Entropy under Interval Uncertainty: Linear-Time Algorithms", *Reliable Computing*, 2007, Vol. 13, No. 6, pp. 467–488.