1-2010

Estimating Information Amount under Uncertainty: Algorithmic Solvability and Computational Complexity

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

Gang Xiang

Comments:

Technical Report: UTEP-CS-09-37a

Published in *International Journal of General Systems*, 2010, Vol. 39, No. 4, pp. 349-378.

International Journal of General Systems
Vol. 00, No. 00, February 2010, 1–33

Estimating Information Amount under Uncertainty: Algorithmic Solvability and Computational Complexity

Vladik Kreinovich[a*] and Gang Xiang[b,a]

[a] *Department of Computer Science, University of Texas, El Paso, TX 79968, USA*;
[b] *Philips Healthcare, Business Line RIS, 6006 N. Mesa, El Paso, TX 79912, USA*

(*Received November 2009*)

Measurement results (and, more generally, estimates) are never absolutely accurate: there is always an uncertainty, the actual value x is, in general, different from the estimate $\widetilde{x}$. Sometimes, we know the probability of different values of the estimation error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$, sometimes, we only know the interval of possible values of Δx, sometimes, we have interval bounds on the cdf of Δx. To compare different measuring instruments, it is desirable to know which of them brings more information – i.e., it is desirable to gauge the amount of information. For probabilistic uncertainty, this amount of information is described by Shannon's entropy; similar measures can be developed for interval and other types of uncertainty. In this paper, we analyze the computational complexity of the problem of estimating information amount under different types of uncertainty.

Keywords: amount of information; uncertainty; probabilistic uncertainty; interval uncertainty; entropy; computational complexity

1. Introduction

Uncertainty is inevitable. For each type of information that we are soliciting, there are several ways to acquire this information.

For example, if we are interested in measuring the value of a physical quantity x, we may use different types of sensors. No matter how accurate the sensor, the measured value $\widetilde{x}$ is, in general, different from the actual value x of the measured quantity.

Types of uncertainty: in brief. For different sensors, we have different type of information about this difference $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$:

In some cases, we know which values of Δx are possible and what is the frequency of each of the different possible values. In other words, we know a probability distribution on Δx. This type of uncertainty is usually called a *probabilistic uncertainty.* It is reasonable to describe the corresponding probability distribution by a cumulative distribution function (cdf, for short) $F(t) \stackrel{\text{def}}{=} Prob(x \le t)$.

In other cases, the only information we have is an upper bound Δ on the measurement error. In this case, after we got the measured value $\widetilde{x}$, the only information that we have about the actual (unknown) value x of the measured quantity is that x belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$. This is the case of *interval uncertainty.*

So far, we have described two extreme cases:

[*] Corresponding author. Email: vladik@utep.edu

ISSN: 0308-1079 print/ISSN 1563-5104 online

DOI: 10.1080/0308107YYxxxxxxxx
http://www.informaworld.com

- Probabilistic uncertainty describes the case when we have a *complete* information about the probability distribution.
- Interval uncertainty corresponds to the case when we have *no* information about the probabilities.

In most practical situations, we have *some* information about the probabilities.

As we have mentioned, to get a complete description of a probability distribution, we need to know the values of cdf $F(t)$ for all possible real numbers t. When we have a partial information about the probabilities, this means that we only have a partial information about the values $F(t)$. In other words, for every t, instead of the actual; (unknown) value $F(t)$, we only know the interval $[\underline{F}(t), \overline{F}(t)]$ that contains the (unknown) actual value $F(t)$. In other words, we have a *probability box* (p-box, for short) that contains the actual (unknown) cdf $F(t)$ (Ferson 2002, Ferson *et al.* 2003).

In measurements, the p-box is probably the most general description of possible uncertainty. In many practical situations, however, we cannot get all the information from measurements, we must also use human expertise. The accuracy of human expertise is rarely described solely in terms of guaranteed bounds. For expert estimates, in addition to guaranteed bounds on Δx and on $F(t)$, we also have expert estimates that provide better bounds but with limited confidence.

For example, by looking at a medical image such as an X-ray image, an expert medical doctor can guarantee that the size of the tumor is, say, between 1 and 2 cm. However, with 80% certainty, she can say that the size is between 1.2 and 1.7 cm.

To take such uncertainty into consideration, we can use fuzzy techniques. For example, a nested family of intervals corresponding to different levels of certainty forms a fuzzy number (the intervals are the α-cuts of this fuzzy number). For p-boxes, we have, similarly, a nested family of p-boxes corresponding to different levels of certainty – i.e., a fuzzy-valued cdf.

Need to compare different types of uncertainty. Often, there is a need to compare different types of uncertainty. For example, we may have two sensors: one with a smaller bound on a systematic (interval) component of the measurement error, the other with the smaller bound on the standard deviation of the random component of the measurement error. If we can only afford one of these sensors, which one should we buy? Which of the two sensors brings us more information about the measured signal?

To be able to make such decisions, we must be able to compare which of the uncertainties corresponding to the two sensors carries more information – and for that, we must be able to gauge this amount of information.

Resulting problems. To gauge the amount of information, we must have an *algorithm* for computing the corresponding amount of information. For the result of this algorithm to be meaningful, the corresponding expression for the amount of information must be *well-justified.* So, we face two important problems:

- to select (and justify) an appropriate expression for the amount of information, and
- to find efficient algorithms for computing the selected expression.

At first glance, it may sound as if these two problems are largely independent, and can be solved separately. However, because of the practical nature of the problem, these problems are actually closely related: for an expression to be meaningful, it has to be efficiently computable. In other words, efficient computability is one of the most important requirements for selecting an expression for the amount of information.

In view of this relation, in this paper, we describe both the justification of the corresponding expression(s) and the algorithms for computing these expressions.

Traditional amount of information: brief reminder. The traditional Shannon's notion of the amount of information is based on defining information as the (average) number of "yes"-"no" (binary) questions that we need to ask so that, starting with the initial uncertainty, we will be able to completely determine the object.

After each binary question, we can have 2 possible answers. So, if we ask q binary questions, then, in principle, we can have 2^q possible results. Thus, if we know that our object is one of n objects, and we want to uniquely pinpoint the object after all these questions, then we must have $2^q \geq n$. In this case, the smallest number of questions is the smallest integer q that is $\geq \log_2(n)$. This smallest number is called a *ceiling* and denoted by $\lceil \log_2(n) \rceil$.

For discrete probability distributions, we get the standard formula for the average number of questions $-\sum p_i \cdot \log_2(p_i)$. For the continuous case, we can estimate the average number of questions that are needed to find an object with a given accuracy ε – i.e., divide the whole original domain into sub-domains of radius ε and diameter 2ε.

For example, if we start with an interval $[a, b]$ of width $b - a$, then we need to subdivide it into $n \sim (b - a)/(2\varepsilon)$ sub-domains, so we must ask

$$\log_2(n) \sim \log_2(b - a) - \log_2(\varepsilon) - 1$$

questions. In the limit, the term that does not depend on ε leads to $\log_2(b - a)$. For continuous probability distributions, we get the standard Shannon's expression $\log_2(n) \sim S - \log_2(2\varepsilon)$, where $S = -\int \rho(x) \cdot \log_2 \rho(x)\, dx$.

How to extend these formulas to p-boxes etc.? Axiomatic approach. To extend the formulas for information to more general uncertainty, i.e., to come up with generalized information theory, several researchers use an axiomatic approach: they find properties of information, and look for generalizations that satisfy as many of these properties as possible; see, e.g. (Klir and Wierman 1999) and (Kosheleva 1998).

This approach has led to many interesting results, but sometimes, there are several possible generalizations, so which of them should we choose?

Our idea. A natural idea is to choose the definition that kind of coincides with the average number of binary questions that we need to ask.

Since we want to extend the information to the case when probabilities are not known exactly, the average number of questions may also depend on which exactly distribution is actually there. So, it is reasonable to consider the worst-case average number of questions – this is in line with the definition for intervals.

Comment. As we have mentioned, for this idea to be workable, we need to check that this worst-case average number of questions can be efficiently computed.

What we do in this paper. In this paper, we describe how the above idea can be transformed into a formal definition of the amount of information corresponding to different types of uncertainty, and how to compute the corresponding amounts of information.

In particular, we show that for many important types of uncertainty, this worst-case average number of questions can indeed be efficiently computed – and therefore, this measure is not only theoretically reasonable, it can be applied to practical problems.

One such application is given in this paper.

Comment. It is well known that practical applications are often more complex than the corresponding (somewhat simplified) theoretical models. Not surprisingly, our application also goes beyond simply counting the (worst-case) number of questions.

Bibliographic comment. Several of our results first appeared in (Ceberio *et al.* 2006a,b, Kreinovich *et al.* 2005, Xiang *et al.* 2006, 2007).

2. Traditional Amount of Information: Detailed Reminder

Our objective is to extend estimates of the average number of binary questions from the probability distributions to a more general case. To do that, let us recall, in detail, how this number is estimated for probability distributions. The need for such a reminder comes from the fact that while most researchers are familiar with Shannon's formula for the entropy, most researchers are not aware how this formula was (or can be) derived.

Discrete case: no information about probabilities. Let us start with the simplest situation when we know that we have n possible alternatives $A_1, \ldots, A_n$, and we have no information about the probability (frequency) of different alternatives. Let us show that in this case, the smallest number of binary questions that we need to determine the alternative is indeed $q \stackrel{\text{def}}{=} \lceil \log_2(n) \rceil$.

We have already shown that the number of questions cannot be smaller than $\lceil \log_2(n) \rceil$; so, to complete the derivation, we need to show that it is sufficient to ask q questions.

Indeed, let's enumerate all n possible alternatives (in arbitrary order) by numbers from 0 to $n-1$, and write these numbers in the binary form. Using q binary digits, one can describe numbers from 0 to $2^q - 1$. Since $2^q \geq n$, we can describe each of the n numbers by using only q binary digits. So, to uniquely determine the alternative A_i out of n given ones, we can ask the following q questions: "is the first binary digit 0?", "is the second binary digit 0?", etc, up to "is the q-th digit 0?".

Case of a discrete probability distribution. Let us now assume that we also know the probabilities $p_1, \ldots, p_n$ of different alternatives $A_1, \ldots, A_n$. If we are interested in an individual selection, then the above arguments show that we cannot determine the actual alternative by using fewer than $\log_2(n)$ questions. However, if we have many (N) similar situations in which we need to find an alternative, then we can determine all N alternatives by asking $\ll N \cdot \log_2(n)$ binary questions.

To show this, let us fix i from 1 to n, and estimate the number of events N_i in which the output is i.

This number N_i is obtained by counting all the events in which the output was i, so $N_i = n_1 + n_2 + \ldots + n_N$, where n_k equals to 1 if in k-th event the output is

i and 0 otherwise. The average $E(n_k)$ of n_k equals to $p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i$. The mean square deviation $\sigma[n_k]$ is determined by the formula

$$\sigma^2[n_k] = p_i \cdot (1 - E(n_k))^2 + (1 - p_i) \cdot (0 - E(n_k))^2.$$

If we substitute here $E(n_k) = p_i$, we get $\sigma^2[n_k] = p_i \cdot (1 - p_i)$. The outcomes of all these events are considered independent, therefore n_k are independent random variables. Hence the average value of N_i equals to the sum of the averages of n_k: $E[N_i] = E[n_1] + E[n_2] + \ldots + E[n_N] = Np_i$. The mean square deviation $\sigma[N_i]$ satisfies a likewise equation $\sigma^2[N_i] = \sigma^2[n_1] + \sigma^2[n_2] + \ldots = N \cdot p_i \cdot (1 - p_i)$, so $\sigma[N_i] = \sqrt{p_i \cdot (1 - p_i) \cdot N}$.

For big N the sum of equally distributed independent random variables tends to a Gaussian distribution (the well-known *Central Limit Theorem*), therefore for big N, we can assume that N_i is a random variable with a Gaussian distribution. Theoretically a random Gaussian variable with the average a and a standard deviation σ can take any value. However, in practice, if, e.g., one buys a voltmeter with guaranteed 0.1V standard deviation, and it gives an error 1V, it means that something is wrong with this instrument. Therefore it is assumed that only some values are practically possible. Usually a "k-sigma" rule is accepted that the real value can only take values from $a - k \cdot \sigma$ to $a + k \cdot \sigma$, where k is 2, 3, or 4. So in our case we can conclude that N_i lies between $N \cdot p_i - k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$ and $N \cdot p_i + k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$. Now we are ready for the formulation of Shannon's result.

Comment. In this quality control example the choice of k matters, but, as we'll see, in our case the results do not depend on k at all.

Definition 2.1:

- Let a real number $k > 0$ and a positive integer n be given. The number n is called *the number of outcomes.*
- By a *probability distribution*, we mean a sequence $\{p_i\}$ of n real numbers, $p_i \geq 0$, $\sum p_i = 1$. The value p_i is called a *probability* of i-th event.
- Let an integer N is given; it is called *the number of events.*
- By a *result of N events* we mean a sequence r_k, $1 \leq k \leq N$ of integers from 1 to n. The value r_k is called the *result of k-th event.*
- The total number of events that resulted in the i-th outcome will be denoted by N_i.
- We say that the result of N events is *consistent* with the probability distribution $\{p_i\}$ if for every i, we have $N \cdot p_i - k \cdot \sigma_i \leq N_i \leq N + k \cdot \sigma_i$, where $\sigma_i \overset{\text{def}}{=} \sqrt{p_i \cdot (1 - p_i) \cdot N}$.
- Let's denote the number of all consistent results by $N_{cons}(N)$.
- The number $\lceil \log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$.
- The fraction $Q(N)/N$ will be called the *average number of questions.*
- The limit of the average number of questions when $N \to \infty$ will be called the *information.*

Theorem 2.2: *(Shannon) When the number of events N tends to infinity, the average number of questions tends to $S(p) \overset{\text{def}}{=} -\sum p_i \cdot \log_2(p_i)$.*

Comments.

- Shannon's theorem says that if we know the probabilities of all the outputs, then the average number of questions that we have to ask in order to get a complete knowledge equals to the entropy of this probabilistic distribution.
- As we promised, this average number of questions does not depend on the threshold k.
- Since we somewhat modified Shannon's definitions, we cannot use the original proof. Our proof (and proof of other results) is given in the appendices.

Case of a continuous probability distribution. After a finite number of "yes"-"no" questions, we can only distinguish between finitely many alternatives. If the actual situation is described by a real number, then, since there are infinitely many different possible real numbers, after finitely many questions, we can only get an approximate value of this number.

Once we fix the accuracy $\varepsilon > 0$, we can talk about the number of questions that are necessary to determine a number x with this accuracy ε, i.e., to determine an approximate value r for which $|x - r| \le \varepsilon$.

Once an *approximate* value r is determined, possible *actual* values of x form an interval $[r - \varepsilon, r + \varepsilon]$ of width 2ε. Vice versa, if we have located x on an interval $[\underline{x}, \overline{x}]$ of width 2ε, this means that we have found x with the desired accuracy ε: indeed, as an ε-approximation to x, we can then take the midpoint $(\underline{x} + \overline{x})/2$ of the interval $[\underline{x}, \overline{x}]$.

Thus, the problem of determining x with the accuracy ε can be reformulated as follows: we divide the real line into intervals $[x_i, x_{i+1}]$ of width 2ε ($x_{i+1} = x_i + 2\varepsilon$), and by asking binary questions, find the interval that contains x. As we have shown, for this problem, the average number of binary question needed to locate x with accuracy ε is equal to $S = -\sum p_i \cdot \log_2(p_i)$, where p_i is the probability that x belongs to i-th interval $[x_i, x_{i+1}]$.

In general, this probability p_i is equal to $\int_{x_i}^{x_{i+1}} \rho(x)\,dx$, where $\rho(x)$ is the probability distribution of the unknown values x. For small ε, we have $p_i \approx 2\varepsilon \cdot \rho(x_i)$, hence $\log_2(p_i) = \log_2(\rho(x_i)) + \log_2(2\varepsilon)$. Therefore, for small ε, we have

$$S = -\sum \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot 2\varepsilon - \sum \rho(x_i) \cdot 2\varepsilon \cdot \log_2(2\varepsilon).$$

The first sum in this expression is the integral sum for the integral

$$S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(x)\,dx$$

(this integral is called the *entropy* of the probability distribution $\rho(x)$); so, for small ε, this sum is approximately equal to this integral (and tends to this integral when $\varepsilon \to 0$). The second sum is a constant $\log_2(2\varepsilon)$ multiplied by an integral sum for the interval $\int \rho(x)\,dx = 1$. Thus, for small ε, we have

$$S \approx -\int \rho(x) \cdot \log_2(x)\,dx - \log_2(2\varepsilon).$$

So, the average number of binary questions that are needed to determine x with a given accuracy ε, can be determined if we know the entropy of the probability distribution $\rho(x)$.

Our results: in brief. Of course, the abstract definition is a good idea, but the big challenge is translating this abstract definition into explicit easy-to-use analytical formulas and/or algorithms. This is what we do in this paper.

Comment. In our previous work (Chokr and Kreinovich 1994, Ramer and Kreinovich 1994a,b) we provided such formulas for fuzzy numbers and for Dempster-Shafer knowledge bases. In this paper, we provide similar analytical (or at least computable) formulas for the more general case of p-boxes and fuzzy-valued probability distributions.

3. Case of Partial Information about Probability Distribution

Partial information about probability distribution: discrete case. In many real-life situations, instead of having *complete* information about the probabilities $p = (p_1, \ldots, p_n)$ of different alternatives, we only have *partial* information about these probabilities – i.e., we only know a *set* P of possible values of p.

If it is possible to have $p \in P$ and $p' \in P$, then it is also possible that we have p with some probability α and p' with the probability $1 - \alpha$. In this case, the resulting probability distribution $\alpha \cdot p + (1 - \alpha) \cdot p'$ is a convex combination of p and p'. Thus, it it reasonable to require that the set P contains, with every two probability distributions, their convex combinations – in other words, that P is a convex set; see, e.g., (Walley 1991).

Definition 3.1:

- By a *probabilistic knowledge*, we mean a convex set P of probability distributions.
- We say that the result of N events is *consistent* with the probabilistic knowledge P if this result is consistent with one of the probability distributions $p \in P$.
- Let's denote the number of all consistent results by $N_{cons}(N)$.
- The number $\lceil \log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$.
- The fraction $Q(N)/N$ will be called the *average number of questions.*
- The limit of the average number of questions when $N \to \infty$ will be called the *information.*

Definition 3.2: By the *entropy* $S(P)$ of a probabilistic knowledge P, we mean the largest possible entropy among all distributions $p \in P$; $S(P) \stackrel{\text{def}}{=} \max\limits_{p \in P} S(p)$.

Proposition 3.3: *When the number of events N tends to infinity, the average number of questions tends to the entropy $S(P)$.*

Partial information about probability distribution: continuous case. In the continuous case, we also often encounter situations in which we only have partial information about the probability distribution; one such case is the case of p-boxes. In such situations, instead of a knowing the *exact* probability distribution $\rho(x)$, we only know a (convex) class P that contains the (unknown) distribution.

In such situations, we can similarly ask about the average number of questions that are needed to determine x with a given accuracy ε.

Once we fix an accuracy ε and a subdivision of the real line into intervals $[x_i, x_{i+1}]$ of width 2ε, we have a discrete problem of determining the interval containing x.

Due to Proposition 3.3, for this discrete problem, the average number of "yes"-"no" questions is equal to the largest entropy $S(p)$ among all the corresponding discrete distributions $p_i = \int_{x_i}^{x_{i+1}} \rho(x)\,dx$. As we have mentioned, for small ε, $S(p) \sim S(\rho) - \log_2(2\varepsilon)$, where $S(\rho) = -\int \rho(x) \cdot \log_2(\rho(x))\,dx$ is the entropy of the corresponding continuous distribution. Thus, the largest discrete entropy $S(p)$ comes from the distribution $\rho(x) \in P$ for which the corresponding (continuous) entropy $S(\rho)$ attains the largest possible value.

Computing the amount of information. According to the above results, the amount of information in p-box – or more generally, in a class of distributions P – is equal to the largest entropy among all the distributions from the given class P.

Good news is that a lot of research has gone into algorithms for finding distributions with the largest entropy among different classes P – largely as a part of the Maximum Entropy approach in which when we only know a class of distributions P, then we assume that the actual distribution is the one with the largest entropy from P; see, e.g., (Jaynes 2003).

Because of this, for many classes P, we already know the corresponding maximum entropy distribution, so we can explicitly compute the corresponding amount of information. For classes P for which the corresponding maximum entropy distribution is not known, finding such a distribution requires maximizing a convex function (entropy) over a convex set P; it is known that maximizing a convex function over a convex set is a computationally feasible problem; see, e.g., (Vavasis 1991).

Problem with our definition: we need a multi-dimensional notion of information. In our approach, we measure the information as the average number of "yes"-"no" questions that are needed to locate an object with a given accuracy.

According to our results, for a p-box, thus defined amount of information is equal to the amount of information corresponding to the distribution with the largest entropy among all the distributions from a given p-box.

So, by the above definition of the amount of information, we are not able to distinguish between this distribution and entire p-box. This is counter-intuitive. For example, it is well known that the Gaussian distribution has the largest entropy among all the distribution with the same standard deviation σ, but clearly, we have more information if we know that the distribution is Gaussian than if we simply know its standard deviation but not its shape.

To account for this difference, we must supplement the average number of questions by additional characteristics describing the desired amount of information. Thus, to describe the amount of information for general uncertainty, instead of a single number, we need several different numbers, which form a multi-dimensional measure of uncertainty.

In this paper, we explore two natural ways to implement this idea.

4. First Approach: Entropy Interval Instead of a Single Entropy Value

Idea. If we know the probability distribution ρ, then the amount of information is uniquely determined by the corresponding entropy value $S(\rho)$.

We are interested in the situations when we do not know the probability distribution ρ, we only know that the probability distribution belongs to the class P. Based only on this information, the only thing that we can guarantee about the average number of questions is that $S(P)$ questions is sufficient. Later on, as we gather more information, we may learn more about the actual probability distribution – all the way to knowing the exact distribution $\rho_0 \in P$. With this additional knowledge, we may be able to reduce the average number of questions from $S(P) = \max\limits_{\rho \in P} S(\rho)$ to $S(\rho_0)$.

So, if the only information that we have about the probability distribution ρ is that $\rho \in P$, then the only information that we have about the future average number of "yes"-"no" questions is that this number $S(\rho)$ belongs to the range of possible values $\mathbf{S}(P) = \{S(\rho) : \rho \in P\}$. Since the set P is convex – hence connected, and entropy is a continuous function, this range is an interval: $\mathbf{S}(P) = [\underline{S}(P), \overline{S}(P)]$.

The upper endpoint of this interval is the entropy $S(P) = \max\limits_{\rho \in P} S(\rho)$ of the distribution with the largest entropy. So, our idea is to supplement this "pessimistic" (worst-case) estimate $S(P)$ with the "optimistic" (best-case) estimate $\underline{S}(P) = \min\limits_{\rho \in P} S(\rho)$.

Foundationally, this sounds reasonable, but computationally, we have a problem: while computing the *maximum* of a convex function $S(\rho)$ over a convex set P is a feasible problem, computing the *minimum* of a convex function over a convex set is, in general, NP-hard; see, e.g., (Vavasis 1991). So if we compute $\underline{S}(P)$, great; otherwise we may need to look into different approaches.

Discrete case: reminder of the problem. In most practical situations, our knowledge is incomplete: there are several (n) different states which are consistent with our knowledge. How can we gauge this uncertainty? A natural measure of uncertainty is the average number of binary ("yes"-"no") questions that we need to ask to find the exact state. According to Shannon's information theory, when we know the probabilities $p_1, \ldots, p_n$ of different states (for which $\sum p_i = 1$), then this average number of questions is equal to $S = -\sum\limits_{i=1}^{n} p_i \cdot \log_2(p_i)$.

In practice, we rarely know the exact values of the probabilities p_i; these probabilities come from experiments and are, therefore, only known with uncertainty. Usually, from the experiments, we can find *confidence intervals* $\mathbf{p}_i = [\underline{p}_i, \overline{p}_i]$, i.e., intervals which contain the (unknown) values p_i. Since $p_i \geq 0$ and $\sum p_i = 1$, we must have $\underline{p}_i \geq 0$ and $\sum \underline{p}_i \leq 1 \leq \sum \overline{p}_i$. How can we estimate the amount of information under such interval uncertainty?

For different values $p_i \in \mathbf{p}_i$, we get, in general, different values of the amount of information S. Since S is a continuous function, the set of possible values of S is an interval. So, to gauge the corresponding uncertainty, we must find the range $\mathbf{S} = [\underline{S}, \overline{S}]$ of possible values of S.

Thus, we arrive at the following computational problem:

- given n intervals $\mathbf{p}_i = [\underline{p}_i, \overline{p}_i]$,
- find the range

$$\mathbf{S} = [\underline{S}, \overline{S}] = \left\{ -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) \,\middle|\, p_i \in \mathbf{p}_i \,\&\, \sum_{i=1}^{n} p_i = 1 \right\}.$$

In this section, we show:

- that we can efficiently compute $\overline{S}$;
- that the problem of computing $\underline{S}$ is, in general, NP-hard, and
- that in many practically important situations we can efficiently compute $\underline{S}$.

Comment. Shannon's entropy is not the only way to describe uncertainty. Researchers have observed that many practically useful properties of the Shannon's entropy function S do not use its specific form, they only use the fact that the expression $f(p) = -p \cdot \log_2(p)$ is equal to 0 for $p = 0$ and for $p = 1$ and that this expression is differentiable and strictly concave – i.e., that its second derivative $f''(p)$ is negative for all p.

As a result of this observation, they proposed to use *generalized entropy* measures $S = \sum\limits_{i=1}^{n} f(p_i)$ for some differentiable strictly concave function $f(p)$ for which $f(0) = f(1) = 0$. Such generalized entropy measures are indeed useful in many practical applications; see, e.g., (Klir 2005). In addition to Shannon's entropy function $f(p) = -p \cdot \log_2(p)$, several other functions are used in practice such as $f(p) = p \cdot (1 - p^\beta)$ for some $\beta > 0$ – a function that tends to Shannon's entropy function when $\beta \to 0$.

For such generalized information measures, we have a similar problem:

- given n intervals $\mathbf{p}_i = [\underline{p}_i, \overline{p}_i]$,
- find the range

$$\mathbf{S} = [\underline{S}, \overline{S}] = \left\{ \sum_{i=1}^{n} f(p_i) \,\middle|\, p_i \in \mathbf{p}_i \ \&\ \sum_{i=1}^{n} p_i = 1 \right\}.$$

Our results will be described for this general case.

An $O(n \log_2(n))$ algorithm for computing $\overline{S}$.

- First, we sort $2n$ endpoints $\underline{p}_i$ and $\overline{p}_i$ into a sequence

  $$0 = p_{(0)} < p_{(1)} < p_{(2)} < \ldots < p_{(m)} < p_{(m+1)} = 1.$$

  In the process of this sorting, for each k from 1 to m, we form the sets $A_k^- = \{i : \underline{p}_i = p_{(k)}\}$ and $A_k^+ = \{i : \overline{p}_i = p_{(k)}\}$.
- Then, for each k from 0 to m, we compute the values M_k, P_k, and n_k as follows.
  - We start with $M_0 = \sum\limits_{i=1}^{n} f(\underline{p}_i)$, $P_0 = \sum\limits_{i=1}^{n} \underline{p}_i$, and $n_0 = n$.
  - Once we know M_k, P_k, and n_k, we compute the next values of these quantities as follows:

    $$M_{k+1} = M_k - \sum_{j \in A_{k+1}^-} f(\underline{p}_j) + \sum_{j \in A_{k+1}^+} f(\overline{p}_j); \;\; P_{k+1} = P_k - \sum_{j \in A_{k+1}^-} \underline{p}_j + \sum_{j \in A_{k+1}^+} \overline{p}_j;$$

    $$n_{k+1} = n_k - \#(A_{k+1}^-) + \#(A_{k+1}^+).$$
- If $n_k = n$, we take $S_k = M_k$.
- If $n_k < n$, then we compute $p = \dfrac{1 - P_k}{n - n_k}$.
  - If $p \in [p_{(k)}, p_{(k+1)}]$, then we compute $S_k = M_k + (n - n_k) \cdot f(p)$.
  - Otherwise, we ignore this k.
- Finally, we find the largest of these values S_k as the desired bound $\overline{S}$.

Towards a linear-time algorithm for computing $\overline{S}$. In the previous text, we described a $O(n \cdot \log_2(n))$ algorithm for computing $\overline{S}$.

In this algorithm, most stages require linear time $O(n)$. The only stage that requires time $O(n \cdot \log_2(n))$ is sorting. It turns out that instead of using sorting, we can use the median – and the median of n elements can be computed in linear time $O(n)$; see, e.g., (Cormen *et al.* 2009).

Linear-time algorithm for computing $\overline{S}$. This algorithm is iterative. At each iteration of this algorithm we have three sets:

- the set J^- of all the endpoints $\underline{p}_i$ and $\overline{p}_j$ for which we already know that for the optimal vector p we have, correspondingly, $p_i \neq \underline{p}_i$ (for $\underline{p}_i$) or $p_j = \overline{p}_j$ (for $\overline{p}_j$);
- the set J^+ of all the endpoints $\underline{p}_i$ and $\overline{p}_j$ for which we already know that for the optimal vector p we have, correspondingly, $p_i = \underline{p}_i$ (for $\underline{p}_i$) or $p_j \neq \overline{p}_j$ (for $\overline{p}_j$);
- the set J of the endpoints $\underline{p}_i$ and $\overline{p}_j$ for which we have not yet decided whether these endpoints appear in the optimal vector p.

In the beginning, $J^- = J^+ = \emptyset$ and J is the set of all $2n$ endpoints. At each iteration we also update the values $N^- = \#(J^-)$, $N^+ = \#(J^+)$, $E^- = \sum\limits_{\overline{p}_j \in J^-} \overline{p}_j$, and $E^+ = \sum\limits_{\underline{p}_i \in J^+} \underline{p}_i$. Initially, $N^- = N^+ = E^- = E^+ = 0$.

At each iteration we do the following.

- First we compute the median m of the set J.
- Then, by analyzing the elements of the undecided set J one by one, we divide them into two subsets $Q^- = \{p \in J : p \leq m\}$ and $Q^+ = \{p \in J : p > m\}$. We also compute $m^+ = \min\{p : p \in Q^+\}$.
- We compute $e^- = E^- + \sum\limits_{\overline{p}_j \in Q^-} \overline{p}_j$, $e^+ = E^+ + \sum\limits_{\underline{p}_i \in Q^+} \underline{p}_i$,
$$n^- = N^- + \#\{\overline{p}_j \in Q^-\}, \quad n^+ = N^+ + \#\{\underline{p}_i \in Q^+\},$$
and $r = \dfrac{1 - e^- - e^+}{N - n^- - n^+}$.
- If $r < m$, then we replace J^- with $J^- \cup Q^-$, E^- with e^-, J with Q^+, and N^- with n^-.
- If $r > m^+$, then we replace J^+ with $J^+ \cup Q^+$, E^+ with e^+, J with P^-, and N^+ with n^+.
- If $m \leq r \leq m^+$, then we replace J^- with $J^- \cup Q^-$, J^+ with $J^+ \cup Q^+$, J with $\emptyset$, E^- with e^-, E^+ with e^+, N^- with n^-, and N^+ with n^+.

At each iteration the set of undecided indices is divided in half. Iterations continue until all indices are decided. After this we return, as $\overline{S}$, the value of the entropy for the vector x for which:

- $p_j = \overline{p}_j$ for indices j for which $\overline{p}_j \in J^-$,
- $p_i = \underline{p}_i$ for indices i for which $\underline{p}_i \in J^+$, and
- $p_i = r$ for all other indices i.

Comment.. This algorithm was, in effect, first presented in our 2007 paper (Xiang *et al.* 2007), in which we first introduced linear-time algorithms for computing population variance and entropy under interval uncertainty. However, in our 2007 paper, we described, in detail, algorithms for computing population *variance* (and their detailed justifications), while the algorithms and justifications for the *entropy* case were only briefly outlined. In this paper, we present, in detail, linear-time algorithms for entropy and their justifications (in the appendices).

Computing $\underline{S}$ is, in general, NP-hard. Several algorithms for computing $\underline{S}$ are known; see, e.g., (Abellan and Moral 2000, 2003, 2004, 2005, 2006). In the worst case, these algorithms require time that grows exponentially with n.

The following result shows that this exponential time is caused by the complexity of the problem.

Proposition 4.1: *The problem of computing $\underline{S}$ is NP-hard.*

Effective algorithm for computing $\underline{S}$ when intervals are not contained in each other. Usually, when we know p_i with some uncertainty, we know the approximate values $\widetilde{p}_i$ and the accuracy Δ of this approximation. In this case, we know that the actual (unknown) value of p_i belongs to the interval $[\widetilde{p}_i - \Delta, \widetilde{p}_i + \Delta]$. Since these intervals all have the same width 2Δ, none of them can be a proper subset of the other. It turns out that if we restrict ourselves to intervals that satisfy this condition, then it is possible to compute $\underline{S}$ efficiently.

Definition 4.2: We say that intervals $[\underline{p}_i, \overline{p}_i]$ satisfy the *no-subset property* if $[\underline{p}_i, \overline{p}_i] \not\subset (\underline{p}_j, \overline{p}_j)$ for all i and j (for which the intervals $\mathbf{p}_i$ and $\mathbf{p}_j$ are non-degenerate).

An $O(n \cdot \log_2(n))$ algorithm that computes $\underline{S}$ for all cases when the no-subset property holds.

- First, we sort n intervals $\mathbf{p}_i$ in lexicographic order:

$$\mathbf{p}_1 \le_{\text{lex}} \mathbf{p}_2 \le_{\text{lex}} \cdots \le_{\text{lex}} \mathbf{p}_n$$

  where $[\underline{a}, \overline{a}] \le_{\text{lex}} [\underline{b}, \overline{b}]$ if and only if either $\underline{a} < \overline{b}$, or $\underline{a} = \underline{b}$ and $\overline{a} \le \overline{b}$.
- Second, for each i from 1 to n, we compute

$$M_i = \sum_{j:j<i} f\left(\underline{p}_j\right) + \sum_{m:m>i} f\left(\overline{p}_m\right); \quad P_i = \sum_{j:j<i} \underline{p}_j + \sum_{m:m>i} \overline{p}_m.$$

  First, we compute $M_1 = \sum\limits_{j=2}^{n} f(\overline{p}_j)$ and $P_1 = \sum\limits_{j=2}^{n} \overline{p}_j$; then, we sequentially compute other values as

$$M_i = M_{i-1} + f\left(\underline{p}_{i-1}\right) - f\left(\overline{p}_i\right); \quad P_i = P_{i-1} + \underline{p}_{i-1} - \overline{p}_i.$$

- For every i, we compute $p_i = \dfrac{1 - P_i}{n-1}$. If $p_i \in [\underline{p}_i, \overline{p}_i]$, we compute

$$S_i = M_i + f(p_i).$$

- Finally, we return the smallest of these values S_i as $\underline{S}$.

Linear-time algorithm for computing $\underline{S}$ for the case when narrowed intervals satisfy the no-subset property. For simplicity, let us consider the case when all the intervals are non-degenerate, i.e., when $\Delta_i > 0$ for all i.

The proposed algorithm is iterative. At each iteration of this algorithm we have three sets:

- the set I^- of all the indices i from 1 to n for which we already know that for the optimal vector p, we have $p_i = \underline{p}_i$;
- the set I^+ of all the indices j for which we already know that for the optimal vector p, we have $p_j = \overline{p}_j$;
- the set $I = \{1, \ldots, n\} \setminus (I^- \cup I^+)$ of the indices i for which we are still undecided.

In the beginning, $I^- = I^+ = \emptyset$ and $I = \{1, \ldots, n\}$. At each iteration we also update the values of two auxiliary quantities $E^- \overset{\text{def}}{=} \sum\limits_{i \in I^-} \underline{p}_i$ and $E^+ \overset{\text{def}}{=} \sum\limits_{j \in I^+} \overline{p}_j$. In principle, we could compute these values by computing these sums. However, to speed up computations on each iteration, we update these two auxiliary values in a way that is faster than re-computing the corresponding two sums. Initially, since $I^- = I^+ = \emptyset$, we take $E^- = E^+ = 0$.

At each iteration we do the following:

- first, we compute the median m of the set I (median in terms of sorting by $\widetilde{p}_i$);
- then, by analyzing the elements of the undecided set I one by one, we divide them into two subsets $P^- = \{i : \widetilde{p}_i \le \widetilde{p}_m\}$ and $P^+ = \{j : \widetilde{p}_j > \widetilde{p}_m\}$;
- we compute $e^- = E^- + \sum\limits_{i \in P^-} \underline{p}_i$ and $e^+ = E^+ + \sum\limits_{j \in P^+} \overline{p}_j$;
- If $e^- + e^+ > 1$, then we replace I^- with $I^- \cup P^-$, E^- with e^-, and I with P^+.
- If $e^- + e^+ + 2\Delta_m < 1$, then we replace I^+ with $I^+ \cup P^+$, E^+ with e^+, and I with P^-.
- Finally, if $e^- + e^+ \le 1 \le e^- + e^+ + 2\Delta_m$, then we replace I^- with
$$I^- \cup (P^- - \{m\}),$$
I^+ with $I^+ \cup P^+$, I with $\{m\}$, E^- with $e^- - \underline{p}_m$, and E^+ with e^+.

At each iteration the set of undecided indices is divided in half. Iterations continue until we have only one undecided index $I = \{k\}$. After this we return, as $\underline{S}$, the value of the entropy for the vector p for which $p_i = \underline{p}_i$ for $i \in I^-$, $p_j = \overline{p}_j$ for $j \in I^+$, and $p_k = 1 - e^- - e^+$ for the remaining value k.

5. Continuous Case: p-Box

Formulation of the problem and a seemingly natural solution. As we have mentioned, in the traditional statistical approach, the uncertainty in a probability distribution is usually described by Shannon's entropy

$$S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx,$$

where $\rho(x) = F'(x)$ is the probability density function of this distribution.

In the situations when we have partial information about the probability distribution $F(x)$ – e.g., when we only know that $F(x)$ belongs to a non-degenerate p-box $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$, a reasonable estimate for an arbitrary statistical characteristic S is the range of possible values of S over all possible distributions $F(x) \in \mathbf{F}(x)$.

It therefore seems natural to apply this approach to entropy as well – and return the range of entropy as a gauge of uncertainty of a p-box; see, e.g., (Klir 2005, Xiang *et al.* 2006).

Limitations of the above (seemingly natural) solution. The problem with the above approach is that every non-degenerate p-box includes discrete distributions, i.e., distributions which take discrete values $x_1, \ldots, x_n$ with finite probabilities. For such distributions, Shannon's entropy is $-\infty$.

Thus, for every non-degenerate p-box, the resulting interval $[\underline{S}, \overline{S}]$ has the form $[-\infty, \overline{S}]$. Thus, once the distribution with the largest entropy $\overline{S}$ is fixed, we cannot distinguish between a very narrow p-box or a very thick p-box – in both case, we end up with the same interval $[-\infty, \overline{S}]$.

It is therefore desirable to develop a new approach that would enable us to distinguish between these two cases.

Case of p-boxes: description of the situation. The traditional approach of interval-valued entropy does not allow us to distinguish between narrow and wide p-boxes. For a wide p-box, it is OK to make a wide interval like $[-\infty, \overline{S}]$, but for narrow p-boxes, we would like to have narrower estimates. Let us therefore consider narrow p-boxes.

Since entropy is defined for smooth (differentiable) cdfs $F(x)$, it is reasonable to start with the case when the central function of a p-box is also smooth. In other words, we consider p-boxes of the type

$$\mathbf{F}(x) = [F_0(x) - \Delta F(x), F_0(x) + \Delta F(x)],$$

where $F_0(x)$ is differentiable, with derivative $\rho_0(x) \stackrel{\text{def}}{=} F_0'(x)$, and $\Delta F(x)$ is small.

Formulation of the problem. For each $\varepsilon > 0$ and for each distribution $F(x) \in \mathbf{F}(x)$, we can use the above formulas to estimate the average number $S_\varepsilon(F)$ of "yes"-"no" question that we need to ask to determine the actual value with accuracy ε. Our objective is to compute the range $[\underline{S}, \overline{S}] = \{S_\varepsilon(F) : F \in \mathbf{F}\}$.

Estimates. We have mentioned earlier that asymptotically,

$$\overline{S} \sim -\int \rho_0(x) \cdot \log_2(\rho_0(x))\, dx - \log_2(2\varepsilon).$$

It turns out that for the lower bound, we have the following asymptotics:

$$\underline{S} \sim -\int \rho_0(x) \cdot \log_2(\max(2\Delta F(x), 2\varepsilon \cdot \rho_0(x)))\, dx.$$

(The derivation of this formula is given in Appendix H.)

Comment. This result holds when ε and the width of ΔF both tends to 0. If instead we fix the width ΔF and let $\varepsilon \to 0$, then $\overline{S} \to \infty$ but $\underline{S}$ remains finite.

6. Alternative Approach: An Entropy of Determining the Probability Distribution

We started with the situation when we do not know the object, we only know the probabilities of different objects, and we wanted to find out how many "yes"-"no" questions we need to find the object x.

In the new situation, in addition to not knowing the object x, we also do not know the exact probability distribution $\rho(x)$. It is therefore reasonable, in addition to finding out how many binary questions we need to find x, to also find out how many "yes"-"no" questions we need to find the exact probability distribution $\rho(x)$.

Of course, just like we cannot determine the real number x after finitely many "yes"-"no" questions, we are not able to determine $\rho(x)$ exactly after finitely many questions, we can only obtain an approximate value of a probability distribution.

A natural way to describe a probability distribution is via its cdf $F(x)$. There are two reasons why the approximate cdf may be different from the actual one: we may get the probabilities only approximately, and we may get the values at which these probabilities are attained only approximately. It is therefore reasonable to fix two accuracy values ε (accuracy with which we approximate probabilities) and δ (accuracy with which we approximate x) and try to find an approximation $\widetilde{F}(x)$ to $F(x)$ in which, for every x, we have $|\widetilde{F}(\widetilde{x}) - F(x)| \le \varepsilon$ for some $\widetilde{x}$ for which $|\widetilde{x} - x| \le \delta$.

When P is a p-box, then, for every number x_0, we have the interval $[\underline{F}(x_0), \overline{F}(x_0)]$ of possible values of the probability $F(x_0) = Prob(X \le x_0)$. We want to find the actual value of ε with the accuracy ε. We have already mentioned that this is equivalent to localizing $F(x_0)$ within an interval of width 2ε. Within the original interval of width $w(x_0) \stackrel{\text{def}}{=} \overline{F}(x_0) - \underline{F}(x_0)$, there are $n(x_0) \stackrel{\text{def}}{=} w(x_0)/(2\varepsilon)$ such subintervals, so, to localize $F(x_0)$, we need $\sim \log_2(n(x_0)) = \log_2(w(x_0)) - \log_2(2\varepsilon)$ questions.

To get the spatial accuracy δ, we need to repeat this procedure for the values x_1, $x_2 = x_1 + 2\delta$, etc. Overall, we thus need $\sum \log_2(w(x_i)) - \sum \log_2(2\varepsilon)$ questions. If we multiply the first sum by 2δ, then we get the integral sum for $\int \log_2(w(x))\,dx$; so, the first sum is $\sim \int \log_2(w(x))\,dx/(2\delta)$. The second sum is a constant that does not depend on the p-box at all.

Thus, for a p-box $[\underline{F}(x), \overline{F}(x)]$, the overall number of questions that we need to ask to determine the probability distribution $F(x)$ with a given accuracy is determined by the integral $\int \log_2(\overline{F}(x) - \underline{F}(x))\,dx$. This easy-to-compute integral can thus serve as an additional information measure for p-boxes.

7. Adding Fuzzy Uncertainty

The main idea behind fuzzy uncertainty is that, instead of just describing which objects are possible, we also describe, for each object, the degree to which this object is possible. For each degree of possibility α, we can determine the set of objects that are possible with at least this degree of possibility – the *α-cut* of the original fuzzy set. Vice versa, if we know α-cuts for every α, then, for each object x, we can determine the degree of possibility that x belongs to the original fuzzy set.

A fuzzy set can be thus viewed as a nested family of its α-cuts.

Thus, if instead of a (crisp) set P of possible probability distributions (e.g., a p-box), we have a fuzzy set $\mathcal{P}$ of possible probability distributions, then we can view this information as a family of nested crisp sets $\mathcal{P}(\alpha)$ – α-cuts of the given fuzzy set.

In this case, once we fix a measure of information $I(P)$ for crisp sets of distributions – e.g., the maximum entropy, we can then extend this measure to fuzzy sets $\mathcal{P}$ – by defining $I(\mathcal{P})$ as a fuzzy number whose α-cut coincides with $I(\mathcal{P}(\alpha))$.

Comment. Instead of describing the information in a fuzzy set by a fuzzy number, we can, alternatively, interpret degree of possibility in probabilistic terms and compute the corresponding information by using probability formulas; see, e.g., (Ramer and Kreinovich 1994a,b).

8. Application: How to Measure Loss of Privacy

Need to take into account that not all information is equally important. In the main text, we estimated the amount of information by the number of "yes"-"no" questions that we need to ask so that, starting with the initial uncertainty, we will be able to completely determine the object (or at least determine it with a given accuracy ε).

The very fact that we are simply counting the number of questions means that we implicitly assume that all these questions are (in some reasonable sense) equally important – i.e., in other words, that all pieces of information about the objects are (in some sense) equally important.

In many practical applications, this assumption is very reasonable – e.g., when we are estimating how much computer memory we need to store this information or how much computation time we need to process it.

However, in some applications, different pieces of information are of drastically different importance. In such applications, it is desirable to modify the above definition so as to take into account relative importance of different questions. In this paper, we provide one example of such an application: to measuring the loss of privacy.

Measuring loss of privacy is important. Before explaining why the Shannon-type amount of information is not always a very good measure of privacy loss, let us first explain why it is important to measure loss of privacy in the first place.

Privacy means, in particular, that we do not disclose all information about ourselves. If some of the originally un-disclosed information is disclosed, some privacy is lost. To compare different privacy protection schemes, we must be able to gauge the resulting loss of privacy.

Seemingly natural idea: measuring loss of privacy by the acquired amount of information. Since privacy means that we do not have complete information about a person, a seemingly natural idea is to gauge the loss of privacy by the amount of new information that we gained about this person.

Often, this idea is in good accordance with our intuition. In some cases, the above definition is in good accordance with the intuitive notion of a loss of privacy. As an example, let us consider the case when our only information about some parameter x is that the (unknown) actual value of this parameter x belongs to the (unknown) interval $[L, U]$. In this case, the amount of information is proportional to $\log_2(U - L)$. If we learn a narrower interval containing x, e.g., if we learn that the actual value of x belongs to the left half $[u, l] \stackrel{\text{def}}{=} [L, (L + U)/2]$ of the original interval, then the resulting amount of information is reduced to

$$\log_2((L+U)/2-L)=\log_2((U-L)/2)=\log_2(U-L)-1.$$

Thus, by learning the narrower interval for x, we gained

$$\log_2(U-L)-(\log_2(U-L)-1)=1$$

bit of new information.

The narrower the new interval, the smaller the resulting new amount of information, so the larger the information gain.

The above definition is not always perfect. In some other situations, however, the above idea is not in perfect accordance with our intuition.

Indeed, when we originally knew that a person's salary is between \$10,000 and \$20,000 and later learn that the salary is between \$10,000 and \$15,000, we gained one bit of information. On the other hand, if the only new information that we learned is that the salary is an even number, we also learn exactly one bit of new information. However, intuitively:

- in the first case, we have a substantial privacy loss, while
- in the second case, the direct privacy loss is minimal.

Comment. It is worth mentioning that while the direct privacy loss is small, the information about evenness may indirect lead to a huge privacy loss. The fact that the salary is even means that we know its remainder modulo 2. If, in addition, we learn the remainder of the salary modulo 3, 5, etc., then we can can combine these seemingly minor pieces of information and use the Chinese remainder theorem (see, e.g., (Cormen *et al.* 2009)) to uniquely reconstruct the salary.

What we plan to do. The main objective of this section is to describe an alternative definition of privacy loss which is in better accordance with our intuition.

Why information is not always a perfect measure of loss of privacy. The amount of new information is not always a good measure of the loss of privacy because it does not distinguish between:

- crucial information that may seriously affect a person, and
- irrelevant information – that may not affect a person at all.

To make a distinction between these two types of information, let us estimate potential financial losses caused by the loss of privacy.

Example when loss of privacy can lead to a financial loss. As an example, let us consider how a person's blood pressure x affects the premium that this person pays for his or her health insurance.

From the previous experience, insurance companies can deduce, for each value of blood pressure x, the expected (average) value of the medical expenses $f(x)$ of all individuals with this particular value of blood pressure. So, when the insurance company knows the exact value x of a person's blood pressure, it can offer this person an insurance rate $F(x)\overset{\text{def}}{=}f(x)\cdot(1+\alpha)$, where α is the general investment profit. Indeed:

- If an insurance company offers higher rates, then its competitor will be able to offer lower rates and still make a profit.
- On the other hand, if the insurance company is selling insurance at a lower rate, then it will not earn enough profit, and investors will pull their money out and invest somewhere else.

To preserve privacy, we only keep the information that the blood pressure of all individuals from a certain group is between two bounds L and U, and we do not know have any additional information about the blood pressure of different individuals. Under this information, how much will the insurance company charge to insure people from this group?

Based on the past experience, the insurance company is able to deduce the relative frequency of different values $x \in [L, U]$ – e.g., in the form of the corresponding probability density $\rho(x)$. In this case, the expected medical expenses of an average person from this group are equal to $E[f(x)] \stackrel{\text{def}}{=} \int \rho(x) \cdot f(x)\, dx$. Thus, the insurance company will insure the person for a cost of $E[F(x)] = \int \rho(x) \cdot F(x)\, dx$.

Let us now assume that for some individual, the privacy is lost, and for this individual, we know the exact value x_0 of his or her blood pressure. For this individual, the company can now better predict its medical expenses as $f(x_0)$ and thus, offer a new rate $F(x_0) = f(x_0) \cdot (1 + \alpha)$. When $F(x_0) > E[F(x)]$, the person whose privacy is lost also experiences a financial loss $F(x_0) - E[F(x)]$. We will use this financial loss to gauge the loss of privacy.

Need for a worst-case comparison. In the above example, there is a financial loss only if the person's blood pressure x_0 is worse than average. A person whose blood pressure is lower than average will only benefit from reduced insurance rates.

However, in a somewhat different situation, if the person's blood pressure is smaller (better) than average, this person's loss or privacy can also lead to a financial loss. For example, an insurance company may, in general, pay for a preventive medication that lowers the risk of heart attacks – and of the resulting huge medical expenses. The higher the blood pressure, the larger the risk of a heart attack. So, if the insurance company learns that a certain individual has a lower-than-average blood pressure and thus, a lower-than-average risk of a heart attack, this risk may not justify the expenses on the preventive medication. Thus, due to a privacy loss, the individual will have to pay for this potentially beneficial medication from his/her own pocket – and thus, also experience a financial loss.

So, to gauge a privacy loss, we must consider not just a single situation, but several different situations, and gauge the loss of privacy by the worst-case financial loss caused by this loss of privacy.

Which functions $F(x)$ should we consider. In different situations, we may have different functions $F(x)$ that describe the dependence of a (predicted) financial gain on the (unknown) actual value of a parameter x.

This prediction only makes sense only if we can predict $F(x)$ for each person with a reasonable accuracy, e.g., with an accuracy $\varepsilon > 0$. Measurements are never 100% accurate, and measurement of x are not exception. Let us denote by δ the accuracy with which we measure x, i.e., the upper bound on the (absolute value of) the difference $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ between the measured value $\widetilde{x}$ and the (unknown) actual value x. Due to this difference, the estimated value $F(\widetilde{x})$ is different from the ideal prediction $F(x)$. Usually, measurement errors Δx are small, so we can expand the prediction inaccuracy $\Delta F \stackrel{\text{def}}{=} F(\widetilde{x}) - F(x) = F(x + \Delta x) - F(x)$ in Taylor series in Δx and ignore quadratic and higher order terms in this expansion, leading to $\Delta F \approx F'(x) \cdot \Delta x$. Since the largest possible value of Δx is δ, the largest possible value for ΔF is thus $|F'(x)| \cdot \delta$. Since this value should not exceed ε, we thus conclude that $|F'(x)| \cdot \delta \le \varepsilon$, i.e., that $|F'(x)| \le M \stackrel{\text{def}}{=} \varepsilon/\delta$.

Resulting definitions. Thus, we arrive at the following definition:

Definition 8.1: Let P be a class of probability distributions on a real line, and let $M > 0$ be a real number. By the *amount of privacy* $A(P)$ related to P, we mean the largest possible value of the difference $F(x_0) - \int \rho(x) \cdot F(x)\,dx$ over:

- all possible values x_0,
- all possible probability distributions $\rho \in P$, and
- all possible functions $F(x)$ for which $|F'(x)| \le M$ for all x.

The above definition involves taking a maximum over all distributions $\rho \in P$ which are consistent with the known information about the group to which a given individual belongs. In some cases, we know the exact probability distribution, so the family P consists of only one distribution. In other situations, we may not know this distribution. For example, we may only know that the value of x is within the interval $[L, U]$, and we do not know the probabilities of different values within this interval. In this case, the class P consists of all distributions which are located on this interval (with probability 1).

When we learn new information about this individual, we thus reduce the group and hence, change from the original class P to a new class Q. This change, in general, decreases the amount of privacy.

In particular, when we learn the exact value x_0 of the parameter, then the resulting class of distribution reduces to a single distribution concentrated on this x_0 with probability 1 – for which $F(x_0) - \int \rho(x) \cdot F(x)\,dx = 0$ and thus, the privacy is 0. In this case, we have a 100% loss of privacy – from the original value $A(P)$ to 0. In other cases, we may have a partial loss of privacy.

In general, it is reasonable to define the *relative loss of privacy* as a ratio

$$\frac{A(P) - A(Q)}{A(P)}. \tag{1}$$

In other words, it is reasonable to use the following definition.

Definition 8.2:

- By a *privacy loss*, we mean a pair $\langle P, Q\rangle$ of classes of probability distributions.
- For each privacy loss $\langle P, Q\rangle$, by the measure of a privacy loss, we mean the ratio (1).

Comment. At first glance, it may sound as if these definitions depend on an (unknown) value of the parameter M. However, it is easy to see that the actual measure of the privacy loss does not depend on M:

Proposition 8.3: *For each pair $\langle P, Q\rangle$, the measure of the privacy loss is the same for all $M > 0$.*

The new definition of privacy loss is in good agreement with intuition. Let us show that the new definition adequately describes the difference between learning that the parameter is in the lower half of the original interval and that the parameter is even.

Proposition 8.4: *Let $[l,u] \subseteq [L,U]$ be intervals, let P be the class of all probability distributions located on the interval $[L,U]$, and let Q be the class of all probability distributions located on the interval $[l,u]$. For this pair $\langle P,Q\rangle$, the measure of the privacy loss is equal to* $1 - \dfrac{u-l}{U-L}$.

Comment. In particular, if we start with an interval $[L,U]$, and then we learn that the actual value x is in the lower half $[L,(L+U)/2]$ of this interval, then we get a 50% privacy loss.

What about the case when we assume that x is even? Similarly to the proof of the above proposition, one can prove that if both L and U are even, and Q is the class of all distributions $\rho(x)$ which are located, with probability 1, on even values x, we get $A(Q) = A(P)$. Thus, the even-values restriction lead to a 0% privacy loss.

Thus, the new definition of the privacy loss is indeed in good agreement with our intuition.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721, by Grant 1 T36 GM078000-01 from the National Institutes of Health, by Grant MSM 6198898701 from MŠMT of Czech Republic, and by Grant 5015 “Application of fuzzy logic with operators in the knowledge based systems” from the Science and Technology Centre in Ukraine (STCU), funded by European Union.

The authors are grateful to George Klir and Radim Belohlavek for their encouragement and valuable suggestions, and to the anonymous referees for their thorough analysis and important advice.

References

Abellan, J., and Moral, S, 2000. A non-specificity measure for convex sets of probability distributions. *Intern. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8 (3), 357–367.

Abellan, J., and Moral, S., 2003. Maximum of entropy for credal sets”, *Intern. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 11 (5), 587–597.

Abellan, J., and Moral, S., 2004. Range of entropy for credal sets”, In: López-Diaz, M., et al. (eds.), *Soft Methodology and Random Information Systems*, Springer, Berlin and Heidelberg, 157–164.

Abellan, J., and Moral, S., 2005. Difference of entropies as a nonspecificity function on credal sets, *Intern. J. of General Systems*, 34 (3), 201–214.

Abellan, J., and Moral, S., 2005a. An algorithm to compute the upper entropy for order-2 capacities, *Intern. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 14 (2), 141–154.

Ceberio, M., Kreinovich, V., Xiang, G., Ferson, S., and Joslyn, C., 2006a. Adding constraints to situations when, in addition to intervals, we also have partial information about probabilities. *In:* IEEE Proceedings of the 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics, Duisburg, Germany, September 26–29, 2006.

Ceberio, M., Xiang, G., Longpré, L., Kreinovich, V., Nguyen, H.T., and Berleant, D., 2006b. Two etudes on combining probabilistic and interval uncertainty: processing correlations and measuring loss of privacy. *Proceedings of the 7th International Conference on Intelligent Technologies InTech'06*, Taipei, Taiwan, December 13–15, 2006, 8–17.

Chokr, B. and Kreinovich, V., 1994. How far are we from the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach. *In*: R.R. Yager, J. Kacprzyk, and M. Pedrizzi, eds., *Advances in the Dempster-Shafer Theory of Evidence.* New York: Wiley, 555–576.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. *Introduction to Algorithms.* Cambridge, Massachusetts: MIT Press.

Ferson, S., 2002, *RAMAS Risk Calc 4.0: Risk assessment with uncertain numbers.* Boca Raton, Florida: CRC Press.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D.S., and Sentz, K., 2003. *Constructing probability boxes and Dempster-Shafer structures*, Sandia National Laboratories, Report SAND2002-40.

Jaynes, E.T., 2003. *Probability theory: the logic of science*, Cambridge, Massachusetts: Cambridge University Press.

Klir, G.J., 2005. *Uncertainty and Information: Foundations of Generalized Information Theory.* Hoboken, New Jersey: Wiley.

Klir, G.J. and Wierman, M.J., 1999. *Uncertainty-based information: elements of Generalized Information Theory.* Heidelberg: Springer Verlag.

Kosheleva, O., 1998. Symmetry-group justification of maximum entropy method and generalized maximum entropy methods in image processing. In: Erickson, G.J., Rychert, J.T. and Smith, C. R. (eds.), *Maximum Entropy and Bayesian Methods*, Dordrecht: Kluwer, 101–113.

Kreinovich, G., 1996. Maximum entropy and interval computations. *Reliable Computing*, 2 (1), 63–79.

Kreinovich, V., Xiang, G., and Ferson, S., 2005. How the concept of information as average number of 'yes-no' questions (bits) can be extended to intervals, p-boxes, and more general uncertainty. *Proceedings of the 24th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, 80–85.

Papadimitriou, C.H., 1994. *Computational Complexity.* Reading, Massachusetts: Addison-Wesley.

Ramer, A. and Kreinovich, V., 1994a. Information complexity and fuzzy control. *In*: A. Kandel and G. Langholtz, eds., *Fuzzy control systems*, Boca Raton, Florida: CRC Press, 75–97.

Ramer, A. and Kreinovich, V., 1994b. Maximum entropy approach to fuzzy control. *Information Sciences*, 81 (3–4), 235–260.

Vavasis, S.A., 1991. Nonlinear optimization: complexity issues. New York: Oxford University Press.

Walley, P., 1991. *Statistical reasoning with imprecise probabilities.* New York: Chapman & Hall.

Xiang, G., Ceberio, M., and Kreinovich, V., 2007. Computing population variance and entropy under interval uncertainty: linear-time algorithms. *Reliable Computing*, 13 (6), 467–488.

Xiang, G., Kosheleva, O., and Klir, G.J., 2006. Estimating information amount under interval uncertainty: algorithmic solvability and computational complexity. *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006, 840–847.

Appendices: Proofs

Appendix A. Proof of Shannon's theorem

Let's first fix some values N_i, that are consistent with the given probabilistic distribution. Due to the inequalities that express the consistency demand, the ratio $f_i = N_i/N$ tends to p_i as $N \to \infty$. Let's count the total number C of results, for which for every i the number of events with outcome i is equal to this N_i. Once we know C, we will be able to compute N_{cons} by adding these C's.

Actually we are interested not in N_{cons} itself, but in $Q(N) = \lceil \log_2(N_{cons}) \rceil$, and moreover, in $\lim(Q(N)/N)$. So we'll try to estimate not only C, but also $\log_2(C)$ and $\lim\ \log_2(C)/N$.

To estimate C means to count the total number of sequences of length N, in which there are N_1 elements, equal to 1, N_2 elements, equal to 2, etc. It is known that this number is equal to

$$C = \frac{N!}{N_1! \cdot N_2! \cdot \ldots \cdot N_n!}$$

To simplify computations, we can use the well-known Stirling formula

$$k! \sim (k/e)^k \cdot \sqrt{2\pi \cdot k}.$$

Then, we get

$$C \approx \frac{\left(\dfrac{N}{e}\right)^N \sqrt{2\pi \cdot N}}{\left(\dfrac{N_1}{e}\right)^{N_1} \cdot \sqrt{2\pi \cdot N_1} \cdot \ldots \cdot \left(\dfrac{N_n}{e}\right)^{N_n} \cdot \sqrt{2\pi \cdot N_n}}$$

Since $\sum N_i = N$, terms e^N and e^{N_i} cancel each other.

To get further simplification, we substitute $N_i = N \cdot f_i$, and correspondingly $N_i^{N_i}$ as $(N \cdot f_i)^{N \cdot f_i} = N^{N \cdot f_i} \cdot {f_i}^{N \cdot f_i}$. Terms N^N is the numerator and

$$N^{N \cdot f_1} \cdot N^{N \cdot f_2} \cdot \ldots \cdot N^{N \cdot f_n} = N^{N \cdot f_1 + N \cdot f_2 + \ldots + N \cdot f_n} = N^N$$

in the denominator cancel each other. Terms with $\sqrt{N}$ lead to a term that depends on N as $c \cdot N^{-(n-1)/2}$. So, we conclude that

$$\log_2(C) \approx -N \cdot f_1 \cdot \log_2(f_1) - \ldots - N \cdot f_n \log_2(f_n) -$$

$$\frac{n-1}{2} \cdot \log_2(N) - \text{const}.$$

When $N \to \infty$, we have $1/N \to 0$, $\log_2(N)/N \to 0$, and $f_i \to p_i$, therefore

$$\frac{\log_2(C)}{N} \to -p_1 \cdot \log_2(p_1) - \ldots - p_n \cdot \log_2(p_n),$$

i.e., $\log_2(C)/N$ tends to the entropy of the probabilistic distribution.

Now, that we have found an asymptotics for C, let's compute N_{cons} and $Q(N)/N$. For a given probabilistic distribution $\{p_i\}$ and every i, possible values of N_i form an interval of length $L_i \stackrel{\text{def}}{=} 2k \cdot \sqrt{p_i \cdot (1-p_i)} \cdot \sqrt{N}$. So there are no more than L_i possible values of N_i. The maximum value for $p_i \cdot (1-p_i)$ is attained when $p_i = 1/2$, therefore $p_i \cdot (1-p_i) \le 1/4$, and hence $L_i \le 2k \cdot \sqrt{N/4} = (k/2) \cdot \sqrt{N}$. For every i from 1 to n there are at most $(k/2) \cdot \sqrt{N}$ possible values of N_i, so the total number of possible combinations of $N_1, \ldots, N_n$ is smaller than $((k/2) \cdot \sqrt{N})^n$. Let us denote this number of combinations by $N(p)$.

The total number N_{cons} of consistent results is the sum of $N(p)$ different values of C (values that correspond to $N(p)$ different combinations of $N_1, N_2, \ldots, N_n$). Let's denote the biggest of these values C by $C_{\max}$. Since N_{cons} is the sum of $N(p)$ terms, and each of these terms is not larger than the largest of them $C_{\max}$, we conclude that $N_{cons} \le N(p) \cdot C_{\max}$. On the other hand, the sum N_{cons} of non-negative integers is not smaller than the largest of them, i.e., $C_{\max} \le N_{cons}$. Combining these two inequalities, we conclude that $C_{\max} \le N_{cons} \le N(p) \cdot C_{\max}$. Since $N(p) \le ((k/2) \cdot \sqrt{N})^n$, we conclude that $C_{\max} \le N_{cons} \le ((k/2) \cdot \sqrt{N})^n \cdot C_{\max}$. Turning to logarithms, we find that $\log_2(C_{\max}) \le \log_2(N_{cons}) \le \log_2(C_{\max}) + (n/2) \cdot \log_2(N) + \text{const}$. Dividing by N, tending to the limit $N \to \infty$ and using the fact that $\log_2(N)/N \to 0$ and the (already proved) fact that $\log_2(C_{\max})/N$ tends to the entropy S, we conclude that $\lim Q(N)/N = S$. The proposition is proven.

Appendix B. Proof of Proposition 3.3

By definition, a result is consistent with the probabilistic knowledge P if and only if it is consistent with one of the distributions $p \in P$. Thus, the set of all the results which are consistent with P can be represented as a union of the sets of all the results consistent with different probability distributions $p \in P$. In the proof of Shannon's theorem, we have shown that for each $p \in P$, the corresponding number is asymptotically equal to $\exp(N \cdot S(p))$.

To be more precise, for every N, the number C of results with given frequencies $\{f_j\}$ ($f_j \approx p_j$) has already been computed in the proof of Shannon's theorem: $\lim\ (\log_2(C))/N = -\sum f_j \log_2(f_j)$.

The total number of the results N_{cons} which are consistent with a given probabilistic knowledge P is equal to the sum of N_{co} different values of C that correspond to different f_j. For a given N, there are at most $N+1$ different values of $N_1 = N \cdot f_1$ $(0,1,\ldots,N)$, at most $N+1$ different values of N_2, etc., totally at most $(N+1)^n$ different sets of $\{f_j\}$. So, we get an inequality $C_{\max} \le N_{cons} \le (N+1)^n \cdot C_{\max}$, from which we conclude that $\lim Q(N)/N = \lim \log_2(C_{\max})/N$.

Appendix C. Justification of the $O(n \cdot \log_2(n))$ Algorithm for Computing $\overline{S}$

Computing $\overline{S}$: analysis of the problem. Let $(p_1, \ldots, p_n)$ be the values of probabilities at which the entropy S attains its maximum. The fact that S attains its maximum means that if we change the values p_i, then the corresponding change ΔS in S is non-positive: $\Delta S \le 0$. We will use this condition for different changes in p_i.

For each value of p_i, we have three possibilities:

- this value can be strictly inside the corresponding interval $[\underline{p}_i, \overline{p}_i]$;
- this value can be at the left end of this interval, i.e., $p_i = \underline{p}_i$; and
- this value can be at the right end of this interval, i.e., $p_i = \overline{p}_i$.

Let us consider these possibilities one by one.

Let us first consider the values p_j which are strictly inside the corresponding intervals. If for some j and k, the corresponding probabilities are strictly inside the corresponding intervals, i.e., if we have $p_j \in (\underline{p}_j, \overline{p}_j)$ and $p_k \in (\underline{p}_k, \overline{p}_k)$, then for a sufficiently small real number Δ, we can replace p_j with $p_j + \Delta$ and p_k with $p_k - \Delta$ and still get a sequence of probabilities for which $p_i \in [\underline{p}_i, \overline{p}_i]$ for all i and $\sum p_i = 1$. For small Δ, the corresponding change ΔS in entropy is equal to

$$\left(\frac{\partial S}{\partial p_j} - \frac{\partial S}{\partial p_k}\right) \cdot \Delta + o(\Delta) = (f'(p_j) - f'(p_k)) \cdot \Delta + o(\Delta).$$

Since Δ can be positive or negative, the only way to have $\Delta S \le 0$ for all small Δ is to make sure that the coefficient at Δ is equal to 0, i.e., that $f'(p_j) - f'(p_k) = 0$. Since $f(p)$ is a strictly concave function, i.e., $f''(p) < 0$, the derivative $f''(p) = (f'(p))'$ of the function $f'(p)$ is always negative – which means that this derivative is a strictly decreasing function. Thus, $f'(p_j) = f'(p_k)$ implies that $p_j = p_k$ – i.e., that all the values p_j which are inside the corresponding intervals coincide. Let us denote this common value of p_j by p.

Let us now consider the situation when p_j is at the left end of the corresponding interval, i.e., when $p_j = \underline{p}_j$. If for some other k, the corresponding value p_k is at the right end or strictly inside the corresponding interval, then $p_k > \underline{p}_k$. In this case, we can only make a similar change $p_j \to p_j + \Delta$ and $p_k \to p_k - \Delta$ when $\Delta > 0$. Then, the requirement that $\Delta S \le 0$ means that the coefficient at Δ should be non-positive, i.e., that $f'(p_j) - f'(p_k) \le 0$. Since the derivative $f'(p)$ is a strictly decreasing function, we conclude that $p_k \le p_j$. In particular, for the case when p_k is inside the corresponding interval – and is, thus, equal to p – we conclude that $p \le p_j$.

Similarly, if p_j is at the right end of the corresponding interval, i.e., if $p_j = \overline{p}_j$, then, for every k for which $p_k > \underline{p}_k$, we conclude that $p_k \ge p_j$. In particular, we can conclude that $p_j \le p$.

Let us now consider the case when there are some values p_i strictly inside the corresponding interval, so there is a value p. Let us show that is we know where p is located in comparison with all the endpoints $[\underline{p}_i, \overline{p}_i]$, then we can uniquely determine all the values p_i.

Indeed, if the entire interval $[\underline{p}_i, \overline{p}_i]$ is located to the left of p, i.e., if $\overline{p}_i < p$, then:

- the minimum cannot be attained strictly inside the interval – because it would have been attained at the point $p_i = p$, and we are considering the case when the entire interval $[\underline{p}_i, \overline{p}_i]$ is located to the left of p;
- similarly, the minimum cannot be attained for $p_i = \overline{p}_i$, because then, as we have proven, we should have $p \le p_i$, and the entire interval $[\underline{p}_i, \overline{p}_i]$ is located to the left of p.

Thus, in this case, the only remaining possibility is $p_i = \overline{p}_i$.

Similarly, if the entire interval $[\underline{p}_i, \overline{p}_i]$ is located to the right of p, i.e., if $p < \underline{p}_i$, then $p_i = \underline{p}_i$.

If $\underline{p}_i < p < \overline{p}_i$, then, similarly, we cannot have $p_i = \underline{p}_i$ and $p_i = \overline{p}_i$, so we must have p_i inside and hence, $p_i = p$.

To exploit this conclusion, let us formalize how we can describe the location of p in relation to $2n$ endpoints. If we sort these endpoints $\underline{p}_i$ and $\overline{p}_i$ into a sequence $p_{(1)} \le p_{(2)} \le \ldots \le p_{(2n)}$, then we divide the entire real line into $2n + 1$ "zones" $[p_{(k)}, p_{(k+1)}]$, where we denoted $p_{(0)} \overset{\text{def}}{=} 0$ and $p_{(2n+1)} \overset{\text{def}}{=} 1$.

Let us pick a zone $[p_{(k)}, p_{(k+1)}]$, and show how we can find the possibly optimal values p_i (and the corresponding value of the entropy) under the assumption that the (unknown) value p belongs to the this zone.

If $\overline{p}_i < p$, then we must have $\overline{p}_i \le p_{(k)}$ – otherwise, if $\overline{p}_i > p_{(k)}$, then, since $p_{(k)}$

describe all the endpoints, we would have $\overline{p}_i \ge p_{(k+1)}$ and hence $\overline{p}_i > p$. Thus, in the optimal arrangement of probabilities, we have $p_i = \overline{p}_i$.

Similarly, if $\underline{p}_i > p$, then we have $p_i = \underline{p}_i$. For all other i, we have $p_i = p$. This value p can be computed based on the fact that $\sum p_i = 1$.

For each of $2n+1$ zones, we need to analyze n values p_i; thus, for each of the zones, we need $O(n)$ computation steps. Overall, we get a quadratic algorithm for computing $\overline{S}$.

Before we describe this algorithm, we should mention that the above description only works when we actually have an index i for which p_i is strictly inside the corresponding interval. If no such index exists, then we can still conclude that every value $p_j = \overline{p}_j$ is smaller than or equal than every value $p_k = \underline{p}_k$. Thus, there exists a value p that is greater than or equal than all j for which $p_j = \overline{p}_j$ and less than or equal than all k for which $p_k = \underline{p}_k$. By using this p, we arrive at the same conclusion about the values p_i.

Thus, in general, we arrive at the following algorithm (first described in (Kreinovich 1996)).

Quadratic-time algorithm for computing $\overline{S}$.

- First, we sort $2n$ endpoints of n intervals $\mathbf{p}_i$ into an increasing sequence $p_{(0)} = 0 < p_{(1)} < p_{(2)} < \ldots < p_{(m)} < p_{(m+1)} = 1$. (If all the endpoints are different, then $m = 2n$, but since some endpoints may coincide, we may have $m < 2n$; in general, $m \le 2n$.)
- Second, for every k from 0 to $m-1$, we compute the following three values:

$$M_k = \sum_{i:\overline{p}_i \le p_{(k)}} f(\overline{p}_i) + \sum_{j:\underline{p}_j \ge p_{(k+1)}} f(\underline{p}_j); \quad P_k = \sum_{i:\overline{p}_i \le p_{(k)}} \overline{p}_i + \sum_{j:\underline{p}_j \ge p_{(k+1)}} \underline{p}_j;$$

$$n_k = \#\{i : \overline{p}_i \le p_{(k)} \vee \underline{p}_i \ge p_{(k+1)}\}.$$

- If $n_k = n$, we take $S_k = M_k$.
- If $n_k < n$, then we compute $p = \dfrac{1 - P_k}{n - n_k}$.
  - If $p \in [p_{(k)}, p_{(k+1)}]$, then we compute $S_k = M_k + (n - n_k) \cdot f(p)$.
  - Otherwise, we ignore this k.
- Finally, we find the largest of these values S_k as the desired bound $\overline{S}$.

How to reduce the computation time to $O(n \cdot \log_2(n))$. Let us show that the computation time for this algorithm can be reduced to $O(n \cdot \log_2(n))$. Indeed, sorting requires $O(n \cdot \log_2(n))$ steps; see, e.g., (Cormen *et al.* 2009). Once we have a sorted list, we can find, for each of the $2n$ endpoints $\underline{p}_i$ and $\overline{p}_i$, where they are in this sorting. We can thus, for each of the values $p_{(j)}$, mark which endpoints coincide with this value.

The initial computation of the values M_0, P_0, and n_0 requires $O(n)$ steps. Once we go from M_k to M_{k+1} (or from P_k to P_{k+1}), we only need to update the values corresponding to the endpoints of this zone. Overall, for all the updates, we thus need as much time as there are updated values p_i overall.

Each endpoint in this arrangement changes only once, so overall, we need a linear number of steps ($2n$) to update all the values M_k, all the values P_k, and all the values n_k. Thus, overall, we need time $O(n \cdot \log_2(n)) + O(n) + O(n) = O(n \cdot \log_2(n))$.

Appendix D. Proof that the Proposed Fast Algorithm Always Computes $\overline{S}$ in Linear Time

Let us first prove that the fast algorithm described in the main text always computes the desired bound $\overline{S}$. Indeed, in the previous appendices, we have shown that if we sort all $2n$ endpoints into a sequence $p_{(1)} \le p_{(2)} \le \ldots \le p_{(2n)}$, then for some $k = k_{\max}$ the maximum $\overline{S}$ is attained for the vector p for which the following holds:

- For all indices j for which $\overline{p}_j \le p_{(k)}$, we have $p_j = \overline{p}_j$.
- For all indices i for which $\underline{p}_i \ge x_{(k+1)}$, we have $p_i = \underline{p}_i$.
- For all other indices, we have $p_i = \text{const}$. Since $\sum\limits_{i=1}^{n} p_i = 1$, we conclude that this constant is equal to $r_k \stackrel{\text{def}}{=} \dfrac{1 - E_k}{n - N_k}$, where

$$E_k = \sum_{j:\overline{p}_j \le p_{(k)}} \overline{p}_j + \sum_{i:\underline{p}_i \ge p_{(k+1)}} \underline{p}_i;$$

$$N_k = \#\{j : \overline{p}_j \le p_{(k)}\} + \#\{i : \underline{p}_i \ge p_{(k+1)}\}.$$

It can also be proven that for the optimal k we have $r_k \in [p_{(k)}, p_{(k+1)}]$. These facts can proven by the same analysis (adding Δp to one value p_j and subtracting Δp from another value p_k) as in our above analysis of $\underline{S}$.

Let us first prove that if $r_k = \dfrac{1 - E_k}{n - N_k} \le p_{(k+1)}$ then the similar inequality $r_{k+1} = \dfrac{1 - E_{k+1}}{n - N_{k+1}} \le p_{(k+2)}$ holds for the next value k. Indeed, the given inequality $\dfrac{1 - E_k}{n - N_k} \le p_{(k+1)}$ is equivalent to $1 - E_k \le (n - N_k) \cdot p_{(k+1)}$.

The only difference between the sums $E_k = \sum\limits_{j:\overline{p}_j \le p_{(k)}} \overline{p}_j + \sum\limits_{i:\underline{p}_i \ge p_{(k+1)}} \underline{p}_i$ and $E_{k+1} = \sum\limits_{j:\overline{p}_j \le p_{(k+1)}} \overline{p}_j + \sum\limits_{i:\underline{p}_i \ge p_{(k+2)}} \underline{p}_i$ is that:

- some terms equal to $p^{(k+1)}$ may be added (if there are j for which $\overline{p}_j = p_{(k+1)}$), and
- some other terms equal to to $p^{(k+1)}$ may be subtracted (if there are i for which $\underline{p}_i = p_{(k+1)}$).

In general, $E_{k+1} = E_k + c_k \cdot p_{(k+1)}$ for some integer c_k (positive, negative, or zero), and $N_{k+1} = N_k + c_k$. Subtracting $c_k \cdot p_{(k+1)}$ from both sides of the given inequality $1 - E_k \le (n - N_k) \cdot p_{(k+1)}$, we conclude that $1 - E_{k+1} \le (n - N_{k+1}) \cdot p_{(k+1)}$, i.e. that $r_{k+1} = \dfrac{1 - E_{k+1}}{n - N_{k+1}} \le p_{(k+1)}$. Since the sequence $p_{(k)}$ is sorted, we thus conclude that $p_{(k+1)} \le p_{(k+2)}$ and hence $r_{k+1} \le p_{(k+2)}$.

So if the inequality $r_k \le p_{(k+1)}$ holds for some k, it holds for all larger values of k as well. Thus this inequality holds for all k after a certain value l_0.

Similarly, we can prove that if the inequality $r_k \ge p_{(k)}$ holds for some k, then it holds for $k - 1$ as well – since the only difference between E_k and E_{k-1} consists of adding and/or subtracting some values $p_{(k)}$. So if the inequality $r_k \ge p_{(k)}$ holds for some k, it holds for all smaller values of k as well. Thus, this inequality holds for all k until a certain value k_0.

Similarly to the proof about $\underline{V}$, we can prove that if there are several values $k = l_0, l_0 + 1, \ldots, k_0$ for which both inequalities hold $p_{(k)} \le r_k \le p_{(k+1)}$, then for these k, the entropy has exactly the same value.

So:

- for $k < k_{\max}$, we have $r_k > p_{(k+1)}$,
- for $k > k_{\max}$, we have $r_k < p_{(k)}$, and
- for $k = k_{\max}$ (or, to be more precise, for $l_0 \le k \le k_0$), we have $p_{(k)} \le r_k \le p_{(k+1)}$.

Hence:

- if $r_k < p_{(k)}$, then we cannot have $k < k_{\max}$ and $k = k_{\max}$, hence $k > k_{\max}$;
- if $r_k > p_{(k+1)}$, then we cannot have $k > k_{\max}$ and $k = k_{\max}$, hence $k < k_{\max}$;
- if $p_{(k)} \le r_k \le p_{(k+1)}$, then we cannot have $k < k_{\min}$ and $k > k_{\min}$, hence $k = k_{\max}$.

Thus, the above algorithm finds the correct value of $k_{\max}$ and thence, the correct value of $\overline{S}$.

To complete our proof, we must show that the proposed algorithm for computing $\overline{S}$ requires linear time. Indeed, at each iteration, computing median requires linear time, and all other operations with J require time t linear in the number of elements $|J|$ of J: $t \le C \cdot |J|$ for some C. We start with the set J of size $2n$. On the next iteration, we have a set of size $2n/2 = n$, then $n/2$, etc. Thus, the overall computation time is $\le C \cdot (2n + n + n/2 + \ldots) \le C \cdot 4n$, i.e. linear in n.

Appendix E. Proof that Computing $\underline{S}$ is NP-Hard

By definition, a problem is called NP-hard if every problem from the class NP can be reduced to it; see, e.g., (Papadimitriou 1994). To prove that a problem $\mathcal{P}$ is NP-hard, it is sufficient to reduce one of the known NP-hard problems $\mathcal{P}_0$ to $\mathcal{P}$. The reason for this is as follows: since $\mathcal{P}_0$ is known to be NP-hard, it means that every problem from the class NP can be reduced to $\mathcal{P}_0$, and since $\mathcal{P}_0$ can be reduced to $\mathcal{P}$, thus, we can deduce that every problem from the class NP can be reduced to $\mathcal{P}$.

1°. For our proof, we will select the following *subset* problem as the known NP-hard problem $\mathcal{P}_0$: given n positive integers $s_1, \ldots, s_n$, check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum\limits_{i=1}^{n} \eta_i \cdot s_i$ equals to 0.

We will eventually prove that this problem can be reduced to the problem of computing $\underline{S}$; this computational problem will be denoted by $\mathcal{P}$. However, directly proving that $\mathcal{P}_0$ can be reduced to $\mathcal{P}$ seems to be difficult. Therefore, we introduce the following auxiliary problem, denoted as $\mathcal{P}_1$: given a real number $a > 0$ and n intervals $\mathbf{q}_1 = [\underline{q}_1, \overline{q}_1], \mathbf{q}_2 = [\underline{q}_2, \overline{q}_2], \ldots, \mathbf{q}_n = [\underline{q}_n, \overline{q}_n]$, where $\sum\limits_{i=1}^{n} \underline{q}_i \le a \le \sum\limits_{i=1}^{n} \overline{q}_i$ and $0 \le \underline{q}_i$ for all i, find the lower endpoint $\underline{L}$ of the range

$$\mathbf{L} = [\underline{L}, \overline{L}] = \left\{ -\sum_{i=1}^{n} q_i \cdot \log_2(q_i) \,\middle|\, q_i \in \mathbf{q}_i \;\&\; \sum_{i=1}^{n} q_i = a \right\}$$

Comment. Similarly to our problem $\mathcal{P}$, the new problem $\mathcal{P}_1$ is also about minimizing entropy S: the only difference is that instead of the restriction $\sum\limits_{i=1}^{n} p_i = 1$, we have a new restriction $\sum\limits_{i=1}^{n} q_i = a$.

2°. To reduce $\mathcal{P}_0$ to $\mathcal{P}_1$ means that for every instance $(s_1, \ldots, s_n)$ of the problem $\mathcal{P}_0$, we can find a corresponding instance of the problem $\mathcal{P}_1$ from whose solution, we can easily check whether the desired signs η_i in $\mathcal{P}_0$ exist.

In order to select an appropriate instance, let us first analyze the function $-q \cdot \log_2(q)$. This function is equal to 0 for $q = 0$ and for $q = 1$. It attains its maximum when

$$\frac{\partial}{\partial q}(-q \cdot \log_2(q)) = -\log_2(e) \cdot (1 + \ln(q)) = 0,$$

i.e., when $q = \dfrac{1}{e}$. The corresponding maximum is equal to $-\dfrac{1}{e} \cdot \log_2\left(\dfrac{1}{e}\right) = \dfrac{\log_2(e)}{e}$. We have already mentioned that the function $-q \cdot \log_2(q)$ is concave; therefore, for every real number r between 0 and the maximum – i.e., for which $0 < r < \dfrac{\log_2(e)}{e}$, there exist exactly two different values q for which $-q \cdot \log_2(q) = r$. Let us denote the smaller of these two values by $q^-(r)$, and the larger one by $q^+(r)$. We can check that that $0 < q^-(r) < q^+(r) < 1$ and $0 < q^+(r) - q^-(r) < 1$. As r grows from 0 to its largest value, the difference $q^+(r) - q^-(r)$ decreases from 1 to 0.

Now, for each instance $(s_1, \ldots, s_n)$ of the problem $\mathcal{P}_0$, we select the corresponding instance of the problem $\mathcal{P}_1$, i.e., the intervals $[\underline{q}_i, \overline{q}_i]$ and the real number a, as follows:

- First, we select a positive real number z for which $z \cdot \max(s_i) < 1$.
- Next, for each i from 1 to n, we find r_i for which $q^+(r_i) - q^-(r_i) = z \cdot s_i$, and take $\underline{q}_i = q^-(r_i)$ and $\overline{q}_i = q^+(r_i)$.
- Finally, we select $a = \sum\limits_{i=1}^{n} \dfrac{\underline{q}_i + \overline{q}_i}{2}$.

It is easy to check that for thus selected values, $\underline{q}_i \geq 0$ and $\sum\limits_{i=1}^{n} \underline{q}_i \leq a \leq \sum\limits_{i=1}^{n} \overline{q}_i$.

Let $L_0 \stackrel{\text{def}}{=} -\sum\limits_{i=1}^{n} \underline{q}_i \cdot \log_2(\underline{q}_i)$. We will show that $\underline{L} = L_0$ if and only if there exist signs η_i for which $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$.

3°. Let us first prove that $\underline{L} \geq L_o$.

Indeed, due to our choice of $\underline{q}_i$ and $\overline{q}_i$, the function $-q \cdot \log_2(q)$ attains the same value at the two endpoints of the interval $[\underline{q}_i, \overline{q}_i]$ and is larger everywhere inside this interval. Thus, for every i and for every $q_i \in [\underline{q}_i, \overline{q}_i]$, we have $-q_i \cdot \log_2(q_i) \geq -\underline{q}_i \cdot \log_2(\underline{q}_i)$. By adding these inequalities, we conclude that

$$L = -\sum_{i=1}^{n} q_i \cdot \log_2(q_i) \geq -\sum_{i=1}^{n} -\underline{q}_i \cdot \log_2(\underline{q}_i) = L_0.$$

Since all the values of L are larger than or equal to L_0, the smallest possible value

$\underline{L}$ of the function L also satisfies the inequality $\underline{L} = L_0$.

4°. Let us first prove that if the desired signs η_i exist, then $\underline{L} = L_0$.

Indeed, in this case, we can select $q_i = \underline{q}_i$ when $\eta_i = -1$ and $q_i = \overline{q}_i$ when $\eta_i = 1$. Both cases can be described by a single formula

$$q_i = \frac{\underline{q}_i + \overline{q}_i}{2} + \frac{\eta_i \cdot (\overline{q}_i - \underline{q}_i)}{2} = \frac{\underline{q}_i + \overline{q}_i}{2} + \frac{\eta_i \cdot z \cdot s_i}{2}.$$

Since $-\underline{q}_i \cdot \log_2(\underline{q}_i) = -\overline{q}_i \cdot \log_2(\overline{q}_i)$, for this choice of q_i, we have

$$L = -\sum_{i=1}^n q_i \cdot \log_2(q_i) = -\sum_{i=1}^n \underline{q}_i \cdot \log_2(\underline{q}_i) = L_0.$$

In this case,

$$\sum_{i=1}^n q_i = \sum_{i=1}^n \left(\frac{\underline{q}_i + \overline{q}_i}{2} + \frac{\eta_i \cdot z \cdot s_i}{2} \right) =$$

$$\sum_{i=1}^n \frac{\underline{q}_i + \overline{q}_i}{2} + \frac{z}{2} \cdot \sum_{i=1}^n \eta_i \cdot s_i = \sum_{i=1}^n \frac{\underline{q}_i + \overline{q}_i}{2} = a.$$

Since for this choice of q_i, we have $L = L_0$, we can thus onclude that the smallest possible value $\underline{L}$ of L cannot exceed L_0: $\underline{L} \le L_0$.

We have already proven that $\underline{L} \ge L_0$, so we can conclude that $\underline{L} = L_0$.

5°. Now let us prove that if $\underline{L} = L_0$, then the desired signs η_i exists.

Let $q_1, \ldots, q_n$ be the values that minimize L, i.e., for which $L = \underline{L}$. From the equality $\underline{L} = L_0$, we will conclude that for every i, we have either $q_i = \underline{q}_i$ or $q_i = \overline{q}_i$. This can be proven by reduction to a contradiction: if for some j, we have $q_j \ne \underline{q}_j$ and $q_j \ne \overline{q}_j$, then we will get $-q_j \cdot \log_2(q_j) > -\underline{q}_j \cdot \log_2(\underline{q}_j)$. For every other i, we have $-q_i \cdot \log_2(q_i) \ge -\underline{q}_i \cdot \log_2(\underline{q}_i) = -\overline{q}_i \cdot \log_2(\overline{q}_i)$. By adding all these inequalities, we can conclude that

$$\underline{L} = L = -\sum_{i=1}^n q_i \cdot \log_2(q_i) > -\sum_{i=1}^n \underline{q}_i \cdot \log_2(\underline{q}_i) = L_0,$$

which contradicts to our assumption that $\underline{L} = L_0$. This contradiction shows that indeed, for every i, we have either $q_i = \underline{q}_i$ or $q_i = \overline{q}_i$.

Let us set $\eta_i = -1$ when $q_i = \underline{q}_i$ and $\eta_i = 1$ when $q_i = \overline{q}_i$. Then,

$$q_i = \frac{\underline{q}_i + \overline{q}_i}{2} + \frac{\eta_i \cdot z \cdot s_i}{2}.$$

From the condition $\sum q_i = a$, we now conclude that

$$a = \sum_{i=1}^n q_i = \sum_{i=1}^n \frac{\underline{q}_i + \overline{q}_i}{2} + \frac{\eta_i \cdot z \cdot s_i}{2} = a + z \cdot \sum_{i=1}^n \eta_i \cdot s_i,$$

hence $\sum\limits_{i=1}^n \eta_i \cdot s_i = 0$.

Therefore, we have proven that the subset problem $\mathcal{P}_0$ can be reduced to the auxiliary problem $\mathcal{P}_1$. Thus, the auxiliary problem $\mathcal{P}_1$ is also NP-hard.

6°. To complete the proof, we need to show that the auxiliary problem $\mathcal{P}_1$ can be reduced to our $\mathcal{P}$. In other words, for every instance of the auxiliary problem $\mathcal{P}_1$, we can find the corresponding instance of the original problem $\mathcal{P}$, from whose solution we can easily find the solution to the instance of $\mathcal{P}_1$.

Indeed, let us consider an instance of the auxiliary $\mathcal{P}_1$, i.e., the intervals $[\underline{q}_i, \overline{q}_i]$ and the real number a for which $\underline{q}_i \geq 0$ and $\sum\limits_{i=1}^n \underline{q}_i \leq a \leq \sum\limits_{i=1}^n \overline{q}_i$. As the corresponding instance of the original problem, we will take $\underline{p}_i = \dfrac{\underline{q}_i}{a}$ and $\overline{p}_i = \dfrac{\overline{q}_i}{a}$.

Possible values $p_i \in [\underline{p}_i, \overline{p}_i]$ and $q_i \in [\underline{q}_i, \overline{q}_i]$ can be obtained from each other by, correspondingly, multiplying or dividing by a. For each set $q_i = p_i \cdot a$, we have

$$L = -\sum_{i=1}^n q_i \cdot \log_2(q_i) = -\sum_{i=1}^n a \cdot p_i \cdot \log_2(a \cdot p_i) = -a \cdot \sum_{i=1}^n p_i \cdot \log_2(a \cdot p_i) =$$

$$-a \cdot \sum_{i=1}^n p_i \cdot \log_2(p_i) - a \cdot \log_2(a) \cdot \sum_{i=1}^n p_i =$$

$$-a \cdot \sum_{i=1}^n p_i \cdot \log_2(p_i) - a \cdot \log_2(a) = a \cdot S - a \cdot \log_2(a).$$

Thus, L is an increasing function of S, hence the minimum $\underline{L}$ is equal to

$$\underline{L} = a \cdot \underline{S} - a \cdot \log_2(a).$$

Therefore, if we get the solution $\underline{S}$ to the above instance of our original problem $\mathcal{P}$, we will thus be able to easily compute the solution $\underline{L}$ to the corresponding instance of the auxiliary problem $\mathcal{P}_1$.

Therefore, the auxiliary problem $\mathcal{P}_1$ – whose NP-hardness we have already proven – can be reduced to the original problem $\mathcal{P}$. So, we have prove that the original problem $\mathcal{P}$ of computing $\underline{S}$ is indeed NP-hard.

Appendix F. Justification of the $O(n \cdot \log_2(n))$ Algorithm for Computing $\underline{S}$ when Intervals Are Not Contained in Each Other

It is easy to show that when we sort the intervals in lexicographic order, then both their lower endpoints $\underline{p}_i$ and upper endpoints $\overline{p}_i$ are also sorted: $\underline{p}_i \leq \underline{p}_{i+1}$ and $\overline{p}_i \leq \overline{p}_{i+1}$. (Indeed, otherwise, we would get a violation of the subset property.) Let us thus assume that the intervals are thus sorted.

Let us now show that it is sufficient to consider monotonic optimal tuples $p_1, \ldots, p_n$, for which $p_i \leq p_{i+1}$ for all i. Indeed, if $p_i > p_{i+1}$, then, since $p_i \leq \overline{p}_i \leq \overline{p}_{i+1}$ and $p_i > p_{i+1} \geq \underline{p}_{i+1}$, we have $p_i \in [\underline{p}_{i+1}, \overline{p}_{i+1}]$ and similarly $p_{i+1} \in [\underline{p}_i, \overline{p}_i]$. Thus, we can swap the values p_i and p_{i+1} without changing the value of S. We can repeat this swap as many times as necessary until we get a monotonic tuple that has the exact same value $S = \underline{S}$.

Let us now show that in the optimal tuple, at most one p_i can be inside the corresponding interval. Indeed, if we have two values p_j and p_k strictly inside their intervals, then, similarly to the case of $\overline{S}$, we can conclude that $p_j = p_k$. Now, for $p_j - \Delta = p - \Delta$ and $p_k + \Delta = p + \Delta$, the function S should have a minimum at $\Delta = 0$ and thus, its second derivative relative to Δ should be non-negative. However, an explicit computation shows that this derivative is negative. Thus, our assumption is false, and at most one p_j can be inside the corresponding interval.

Similar to the case of $\overline{S}$, we can now conclude that:

- if $p_j = \underline{p}_j$ and $p_m > \underline{p}_m$, then $p_j \le p_m$; and
- if $p_m = \overline{p}_m$ and $p_j < \overline{p}_j$, then $p_m \ge p_j$.

Thus, each value $p_j = \underline{p}_j$ precede all the values $p_m = \overline{p}_m$, and the only value p_i which is strictly inside the corresponding interval lies in between these values. Thus, in a monotonic optimal tuple $p_1, \ldots, p_n$, the first elements are equal to $\underline{p}_j$, then we may have one element which is strictly inside its interval, and then we have values $p_m = \overline{p}_m$.

The above algorithm tests all such (possibly optimal) sequences and finds the one for which the entropy is the largest.

Appendix G: Proof that Under the No-Subset Property, the Fast Algorithm Always Computes $\underline{S}$ in Linear Time

In the previous appendix, we have already shown that, if we sort the intervals $\mathbf{p}_i$ by their midpoints, then the minimum $\underline{S}$ is always attained at a monotonic tuple $p_1, \ldots, p_n$ in which the first elements are equal to $\underline{p}_j$, then we may have one element which is strictly inside its interval, and then we have values $p_m = \overline{p}_m$.

For the resulting vector $p = (\underline{p}_1, \ldots, \underline{p}_{k-1}, p_k, \overline{p}_{k+1}, \ldots, \overline{p}_n)$, with $\underline{p}_k \le p_k \le \overline{p}_k$, the condition $\sum\limits_{i=1}^{n} p_i = 1$ implies that $\Sigma_k \le 1 \le \Sigma_{k-1}$, where $\Sigma_k \overset{\text{def}}{=} \sum\limits_{i=1}^{k} \underline{p}_i + \sum\limits_{j=k+1}^{n} \overline{p}_j$. When we go from Σ_k to Σ_{k+1}, we replace a larger value $\overline{p}_{k+1}$ with a smaller value $\underline{p}_{k+1}$. Hence $\Sigma_k > \Sigma_{k+1}$. Thus there has to be exactly one $k_{\max}$ for which $\Sigma_k \le 1 \le \Sigma_{k-1}$.

So if we have $\Sigma_m > 1$, this means that the value $k_{\max}$ corresponding to the minimum of S is $> m$. Hence for all the indices $i \le m$ we already know that in the optimal vector p we have $p_i = \underline{p}_i$. Thus these indices can be added to the set I^-.

If Σ_{m-1} $(= \Sigma_m + 2\Delta_m) < 1$, this means that the value $k_{\min}$ corresponding to the minimum of S is $< m$. Hence for all the indices $j \ge m$ we already know that in the optimal vector p we have $p_j = \overline{p}_j$. Thus these indices can be added to the set I^+.

Finally, if $\Sigma_m \le 1 \le \Sigma_{m-1}$ then this m is where the minimum of S is attained.

The algorithm has been justified.

The proof that the new algorithm for computing $\underline{S}$ requires linear time is similar to the proof about the linear-time algorithm for computing $\overline{S}$.

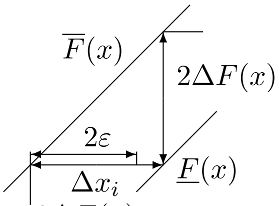
Appendix H. Proof of the Asymptotic Formula for $\underline{S}$ for the Case of p-Boxes

When we discretize the distribution, we get $p_i \approx \rho_0(x_i) \cdot \Delta x_i$, hence

$$-\sum p_i \cdot \log_2(p_i) \approx -\int \rho_0(x) \cdot \log_2(\rho_0(x) \cdot \Delta x)\, dx.$$

To minimize the entropy, we can take the discrete distribution with values

$x_1, \ldots, x_n$ as far away from each other as possible. A distribution which is located at x_i and x_{i+1} and has 0 probability to be in between is described by a cdf $F(x)$ which is horizontal on $[x_i, x_{i+1}]$. Thus, we must select a cdf $F(x) \in \mathbf{F}(x)$ for which these horizontal segments are as long as possible. The length of a horizontal segment is bounded by the geometry of the p-box:



Thus, this length cannot exceed $\dfrac{2\Delta F(x)}{\rho_0(x)}$. If this length is $> 2\varepsilon$, then we can take this interval between the sequential values x_i. If this length is $< 2\varepsilon$, then we can still take $\Delta x_i = 2\varepsilon$. Thus, in general, we take $\Delta x_i = \max\left(\dfrac{2\Delta F(x)}{\rho_0(x)}, 2\varepsilon\right)$. Substituting this expression into the above asymptotic formula, we get the desired asymptotic for $\underline{S}$.

Appendix I. Proof of Proposition 8.3

To prove this proposition, it is sufficient to show that for each $M > 0$, the measure of privacy loss is the same for this M and for $M_0 = 1$. Indeed, for each function $F(x)$ for which $|F'(x)| \le M$ for all x, for the re-scaled function $F_0(x) \stackrel{\text{def}}{=} F(x)/M$, we have $|F_0'(x)| \le 1$ for all x, and

$$F(x_0) - \int \rho(x) \cdot F(x)\,dx = M \cdot \left(F_0(x_0) - \int \rho(x) \cdot F_0(x)\,dx\right). \qquad (2)$$

Vice versa, if $|F_0'(x)| \le 1$ for all x, for the re-scaled function $F(x) \stackrel{\text{def}}{=} M \cdot F_0(x)$, we have $|F'(x)| \le M$ for all x, and (2). Thus, the maximized values corresponding to M and $M_0 = 1$ different by a factor M. Hence, the resulting amounts of privacy $A(P)$ and $A_0(P)$ corresponding to M and M_0 also differ by a factor M: $A(P) = M \cdot A_0(P)$. Substituting this expression for $A(P)$ (and a similar expression for $A(Q)$) into the definition (1), we can therefore conclude that $\dfrac{A(P) - A(Q)}{A(P)} = \dfrac{A_0(P) - A_0(Q)}{A_0(P)}$, i.e., that the measure of privacy is indeed the same for M and $M_0 = 1$. The proposition is proven.

Appendix J. Proof of Proposition 8.4

Due to Proposition 8.3, for computing the measure of the privacy loss, it is sufficient consider the case $M = 1$. Let us show that for this M, we have $A(P) = U - L$.

Let us first show that for every $x_0 \in [L, U]$, for every probability distribution $\rho(x)$ on the interval $[L, U]$, and for every function $F(x)$ for which $|F'(x)| \le 1$, the privacy loss $F(x_0) - \int \rho(x) \cdot F(x)\,dx$ does not exceed $U - L$.

Indeed, since $\int \rho(x)\,dx = 1$, we have $F(x_0) = \int \rho(x) \cdot F(x_0)\,dx$ and hence,

$$F(x_0) - \int \rho(x) \cdot F(x)\,dx = \int \rho(x) \cdot (F(x_0) - F(x))\,dx.$$

Since $|F'(x)| \le 1$, we conclude that $|F(x_0) - F(x)| \le |x_0 - x|$. Both x_0 and x are within the interval $[L, U]$, hence $|x_0 - x| \le U - L$, and $|F(x_0) - F(x)| \le U - L$. Thus, the average value $\int \rho(x) \cdot (F(x_0) - F(x))\,dx$ of this difference also cannot exceed $U - L$.

Let us now show that there exists a value $x_0 \in [L, U]$, a probability distribution $\rho(x)$ on the interval $[L, U]$, and a function $F(x)$ for which $|F'(x)| \le 1$, for which the privacy loss $F(x_0) - \int \rho(x) \cdot F(x)\,dx$ is exactly $U - L$. As such an example, we take $F(x) = x$, $x_0 = U$, and $\rho(x)$ located at a point $x = L$ with probability 1. In this case, the privacy loss is equal to $F(U) - F(L) = U - L$.

Similarly, we can prove that $A(Q) = u - l$, so we get the desired measure of the privacy loss. The proposition is proven.