Two Etudes on Combining Probabilistic and Interval Uncertainty: Processing Correlations and Measuring Loss of Privacy

Martine Ceberio, Gang Xiang,
Luc Longpré, and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
Contact email: vladik@utep.edu

Hung T. Nguyen
Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA
Email: hunguyen@nmsu.edu

Daniel Berleant
Department of Information Science
University of Arkansas at Little Rock
Little Rock, Arkansas 72204, USA
email: jdberleant@ualr.edu

***Abstract*— In many practical situations, there is a need to combine interval and probabilistic uncertainty. The need for such a combination leads to two types of problems: (1) how to process the given combined uncertainty, and (2) how to gauge the amount of uncertainty and – a related question – how to best decrease this uncertainty. In our research, we concentrate on these two types of problems. In this paper, we present two examples that illustrate how the corresponding problems can be solved.**

I. Introduction: Interval Computations

Why indirect measurements? In many real-life situations, we are interested in the value of a physical quantity y that is difficult or impossible to measure directly. Examples of such quantities are the distance to a star and the amount of oil in a given well. Since we cannot measure y directly, a natural strategy is to measure y *indirectly*. Specifically, we find some easier-to-measure quantities $x_1, \ldots, x_n$ which are related to y by a known relation $y = f(x_1, \ldots, x_n)$. To estimate y, we first obtain measurements $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of the quantities $x_1, \ldots, x_n$, and then compute an estimate for y of $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

Why interval computations? Measurement are never 100% accurate, so the actual value x_i of measured quantity i can differ from the measurement result $\widetilde{x}_i$. Because of these *measurement errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$ of the desired quantity y [15].

It is desirable to describe the error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ in the result. To do that, we must have some information about the errors of direct measurements.

What do we know about the errors Δx_i of direct measurements? First, the manufacturer of the measuring instrument may supply us with an upper bound Δ_i on the measurement error. In this case, once we perform a measurement and get a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity is in the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

In many practical situations, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after performing a measurement and getting a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.[1] In such situations, the only information that we have about the (unknown) actual value of $y = f(x_1, \ldots, x_n)$ is that y belongs to the range $\mathbf{y} = [\underline{y}, \overline{y}]$ of the function f over the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$:

$$\mathbf{y} = [\underline{y}, \overline{y}] = \{f(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

The process of computing this interval range based on the input intervals $\mathbf{x}_i$ is part of *interval computations*; see, e.g., [6].

Interval computations techniques: brief reminder. Historically what is often called the "straightforward" method was the first for estimating the desired range of a function. This method is based on the fact that inside the computer, every algorithm for processing real numbers is implemented as a sequence of elementary operations $a+b$, $a-b$, $a \cdot b$, and a/b; usually, a/b is computed as $a \cdot (1/b)$, making $a+b$, $a-b$, $a \cdot b$, and $1/a$ sufficient. For each of these elementary operations $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*:

$$[\underline{a}, \overline{a}] + [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}];$$

$$[\underline{a}, \overline{a}] - [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}];$$

$$[\underline{a}, \overline{a}] \cdot [\underline{b}, \overline{b}] =$$

$$[\min(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}), \max(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b})];$$

$$1/[\underline{a}, \overline{a}] = [1/\overline{a}, 1/\underline{a}] \text{ if } 0 \notin [\underline{a}, \overline{a}].$$

In straightforward interval computations, we replace each floating point operation in the program f by the corresponding

[1] We use the convention of bold, non-italic symbols for naming intervals.

interval operation. It is known that, as a result, we get an enclosure $\mathbf{Y} \supseteq \mathbf{y}$ of the desired range.

In some cases, $\mathbf{Y} = \mathbf{y}$. In more complex cases, the enclosure has excess width ($\mathbf{Y} \supset \mathbf{y}$). There exist more sophisticated techniques for producing narrower enclosures, e.g., centered form methods [6]. However, for each of these techniques, there are cases when we still get excess width. Reason: it is known (see, e.g., [11]), that the problem of computing the exact range is NP-hard even for polynomial functions $f(x_1, \ldots, x_n)$ (indeed, even for quadratic functions f).

What we plan to do in this paper. In many practical situations, there is a need to combine interval and probabilistic uncertainty. The need for such a combination leads to two types of problems:

- how to *process* the given combined uncertainty, and
- how to *gauge* the amount of uncertainty and – a related question – how to best *decrease* this uncertainty.

In our research, we concentrate on these two types of problems. In this paper, we present two examples that illustrate how the corresponding problems can be solved.

II. Adding Probabilities and Correlations to Interval Computations: Formulation of the First Problem

Motivating practical problem. In some practical situations, in addition to lower and upper bounds on each random variable x_i, we know bounds $\mathbf{E}_i = [\underline{E}_i, \overline{E}_i]$ on its mean E_i.

Indeed, in measurement practice (e.g. [15]), the overall measurement error Δx is usually represented as a sum of two components: a *systematic* error component $\Delta_s x$ which is defined as the expected value $E[\Delta x]$, and a *random* error component $\Delta_r x$ which is defined as the difference between overall measurement error Δx and the systematic error component $\Delta_s x$: $\Delta_r x \stackrel{\text{def}}{=} \Delta x - \Delta_s x$. In addition to an upper bound Δ on the magnitude of overall measurement errors, the manufacturers of a measuring instrument often provide an upper bound Δ_s on the magnitude of the systematic error component: $|\Delta_s x| \le \Delta_s$.

When this additional information is given, then, after obtaining a measurement result $\widetilde{x}$, we not only have the information that the actual value x of the measured quantity belongs to the interval $\mathbf{x} = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$, but we can also conclude that the expected value $E[x]$ of $x = \widetilde{x} - \Delta x$ (which is $E[x] = \widetilde{x} - E[\Delta x] = \widetilde{x} - \Delta_s x$) belongs to the interval $[\widetilde{x} - \Delta_s, \widetilde{x} + \Delta_s]$.

If we have this information for every x_i, then, in addition to the interval $\mathbf{y}$ of possible values of y, we can also know the interval of possible values of $E[y]$. This additional interval will, we hypothesized, provide us with information on how repeated measurements can improve the accuracy of this indirect measurement. Thus, we arrive at the following problem.

New problem in precise terms. Given an algorithm computing a function $f(x_1, \ldots, x_n)$ from $\mathbb{R}^n$ to $\mathbb{R}$, and values $\underline{x}_1$, $\overline{x}_1$, …, $\underline{x}_n$, $\overline{x}_n$, $\underline{E}_1$, $\overline{E}_1$, …, $\underline{E}_n$, $\overline{E}_n$, we want to find

$$\underline{E} \stackrel{\text{def}}{=} \min\{E[f(x_1, \ldots, x_n)] : \text{ all distributions of } (x_1, \ldots, x_n) \text{ for which } x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n],$$
$$E[x_1] \in [\underline{E}_1, \overline{E}_1], \ldots E[x_n] \in [\underline{E}_n, \overline{E}_n]\};$$

and $\overline{E}$ which is the maximum of $E[f(x_1, \ldots, x_n)]$ for all such distributions.

In addition to considering all possible distributions, we can also consider the case when all the variables x_i are independent, or, more generally, when we know the correlations among the x_i.

III. First Problem: What Is Known

Extending interval arithmetic to handle expectations. The main idea behind standard interval computations can be applied here as well. First we find out how to solve the problem when $n = 2$ and $f(x_1, x_2)$ is one of the standard arithmetic operations. Then, once we have an arbitrary algorithm $f(x_1, \ldots, x_n)$, we parse it and replace each elementary operation on real numbers with the corresponding operation on quadruples $(\underline{x}, \underline{E}, \overline{E}, \overline{x})$.

To implement this idea, we must therefore know how to solve the above problem for elementary operations.

For *addition*, the answer is straightforward: $E[x_1 + x_2] = E[x_1] + E[x_2]$. So, if $y = x_1 + x_2$, the only possible value for $E = E[y]$ is $E = E_1 + E_2$. This value does not depend on whether we have correlation or whether we have any information about the correlation. Thus, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$.

Similarly, the answer is straightforward for *subtraction:* if $y = x_1 - x_2$, there is only one possible value for $E = E[y]$: the value $E = E_1 - E_2$. Thus, $\mathbf{E} = \mathbf{E}_1 - \mathbf{E}_2$.

For *multiplication*, if the variables x_1 and x_2 are independent, then $E[x_1 \cdot x_2] = E[x_1] \cdot E[x_2]$. Hence, if $y = x_1 \cdot x_2$ and x_1 and x_2 are independent, there is only one possible value for $E = E[y]$: the value $E = E_1 \cdot E_2$; hence $\mathbf{E} = \mathbf{E}_1 \cdot \mathbf{E}_2$.

The only non-trivial case is the case of multiplication in the presence of possible correlation. When we know the exact values of E_1 and E_2, the solution to the above problem is known [9]:

Theorem 1. *If $y = x_1 \cdot x_2$, and we have no information about the correlation, then the range $[\underline{E}, \overline{E}]$ of $E[x_1 \cdot x_2]$ is $[E_{\min}, E_{\max}]$, where $p_i \stackrel{\text{def}}{=} (E_i - \underline{x}_i)/(\overline{x}_i - \underline{x}_i)$, and:*

$$E_{\min} \stackrel{\text{def}}{=} \max(p_1 + p_2 - 1, 0) \cdot \overline{x}_1 \cdot \overline{x}_2 + \min(p_1, 1 - p_2) \cdot \overline{x}_1 \cdot \underline{x}_2 + \min(1 - p_1, p_2) \cdot \underline{x}_1 \cdot \overline{x}_2 + \max(1 - p_1 - p_2, 0) \cdot \underline{x}_1 \cdot \underline{x}_2; \quad (1)$$

$$E_{\max} \stackrel{\text{def}}{=} \min(p_1, p_2) \cdot \overline{x}_1 \cdot \overline{x}_2 + \max(p_1 - p_2, 0) \cdot \overline{x}_1 \cdot \underline{x}_2 + \max(p_2 - p_1, 0) \cdot \underline{x}_1 \cdot \overline{x}_2 + \min(1 - p_1, 1 - p_2) \cdot \underline{x}_1 \cdot \underline{x}_2. \quad (2)$$

Comment. In this case, $\mathbf{E} = [E_{\min}, E_{\max}]$. In the following text, we will use the expressions (1) and (2) to describe the ranges of E for other cases, when the expression for the

range $\mathbf{E} = [\underline{E}, \overline{E}]$ is different from the above expression $[E_{\min}, E_{\max}]$.

For the *inverse* $y = 1/x_1$, a finite range is possible only when $0 \notin \mathbf{x}_1$. Without loss of generality, we can consider the case when $0 < \underline{x}_1$. In this case, we have the following bound [9]:

Theorem 2. *For the inverse $y = 1/x_1$, the range of possible values of E is $\mathbf{E} = [1/E_1, p_1/\overline{x}_1 + (1-p_1)/\underline{x}_1]$.*

(Here p_1 denotes the same value as in Theorem 1.)

Taking correlation into account. As we have seen, for elementary arithmetic operations other than multiplication, the range of the result's expectation is uniquely determined by the ranges of the input expectations. For multiplication, the range of $E[x_1 \cdot x_2]$ depends on both the ranges of $E[x_i]$ and the correlation between the x_i.

For multiplication, we know the bounds on $E[x_1 \cdot x_2]$ for two cases: when x_1 and x_2 are independent, and when we have no information about their correlation. It reality, we may have partial information about the correlation. For example, we may know the exact value ρ of the correlation

$$\rho(x_1, x_2) \stackrel{\text{def}}{=} \frac{E[x_1 \cdot x_2] - E_1 \cdot E_2}{\sigma_1 \cdot \sigma_2} \tag{3}$$

(where σ_i is the standard deviation of x_i). Or more generally we might have an interval $[\underline{\rho}, \overline{\rho}]$ of possible values of ρ.

Analytical expressions are desirable. In [1], a linear programming-based numerical method is described for computing the ranges of binary functions under constraints on the correlation of its arguments. For example, this method can be applied to the problem of estimating the range of $E[x_1 \cdot x_2]$ under known correlation.

In the cases of independence and unknown correlation, there are explicit analytical expressions for the range of $E[x_1 \cdot x_2]$. In general, analytical expressions are much faster to compute than numerical methods. In this paper, we provide analytical expressions for the correlation case as well.

IV. First Problem: Main Results

Preliminaries. Our objective is, given the intervals $[\underline{x}_1, \overline{x}_1]$, $[\underline{x}_2, \overline{x}_2]$, the values $E_1 = E[x_1]$, $E_2 = E[x_2]$, and $\rho = \rho(x_1, x_2)$, to find the range $[\underline{E}, \overline{E}]$ of possible values of $E[x_1 \cdot x_2]$.

Before we derive an expression for the general situation, let us identify the quantitative values for Pearson correlation coefficient ρ corresponding to the known cases – independence and unknown correlation. For the former case, $\rho = 0$. For the latter, according to [9] both $E_{\min}$ and $E_{\max}$ are attained when each of the variables x_i has a 2-point (2-impulse) marginal distribution: $p(x_i = \overline{x}_i) = p_i$ and $p(x_i = \underline{x}_i) = 1 - p_i$. (Probability p_i is uniquely determined by expected value $E[x_i]$.) For this marginal distribution,

$$\sigma^2[x_i] = E[(x_i - E_i)^2] = p_i \cdot (\overline{x}_i - E_i)^2 + (1-p_i) \cdot (E_i - \underline{x}_i)^2.$$

Since $p_i = (E_i - \underline{x}_i)/(\overline{x}_i - \underline{x}_i)$, algebraic manipulation yields

$$\sigma^2[x_i] = (\overline{x}_i - E_i) \cdot (E_i - \underline{x}_i).$$

Thus, using eq. (3), the correlation coefficients $\rho_{\min}$ and $\rho_{\max}$ corresponding to these extreme distributions are equal to $\rho_{\min} = \dfrac{E_{\min} - E_1 \cdot E_2}{\sigma}$ and $\rho_{\max} = \dfrac{E_{\max} - E_1 \cdot E_2}{\sigma}$, where

$$\sigma \stackrel{\text{def}}{=} \sigma_1 \cdot \sigma_2 = \sigma[x_1] \cdot \sigma[x_2] =$$
$$\sqrt{(\overline{x}_1 - E_1) \cdot (E_1 - \underline{x}_1)} \cdot \sqrt{(\overline{x}_2 - E_2) \cdot (E_2 - \underline{x}_2)}.$$

Case of exactly known non-zero correlation. The negative value $\rho_{\min}$ corresponds to the smallest possible value $E_{\min}$ of $E[x_1 \cdot x_2]$, and the positive value $\rho_{\max}$ corresponds to the largest possible value $E_{\max}$. Because the corresponding analyses are limited to the extremes, it is therefore desirable to extend results to include intermediate values of ρ.

Theorem 3. *Let $[\underline{x}_1, \overline{x}_1]$ and $[\underline{x}_2, \overline{x}_2]$ be given intervals, $E_1 \in [\underline{x}_1, \overline{x}_1]$ and $E_2 \in [\underline{x}_1, \overline{x}_1]$ be given numbers, and ρ be a number from the interval $[\rho_{\min}, \rho_{\max}]$. Then the closure $[\underline{E}, \overline{E}]$ of the range of possible values $E[x_1, x_2]$ for all possible distributions for which:*

- *x_1 is located in $[\underline{x}_1, \overline{x}_1]$, and x_2 is located in $[\underline{x}_2, \overline{x}_2]$;*
- *$E[x_1] = E_1$, and $E[x_2] = E_2$; and*
- *$\rho[x_1, x_2] = \rho$,*

is

- *for $\rho \geq 0$: $[E_1 \cdot E_2, E_1 \cdot E_2 + \rho \cdot \sigma]$;*
- *for $\rho \leq 0$: $[E_1 \cdot E_2 + \rho \cdot \sigma, E_1 \cdot E_2]$.*

Comment. The need for closure comes from the fact that ρ is only defined when $\sigma_i > 0$. Thus, e.g., for $\rho > 0$, eq. (3) implies $E[x_1 \cdot x_2] > E[x_1] \cdot E[x_2]$. So, under the standard definition of (Pearson) correlation, the lower endpoint $E_1 \cdot E_2$ might be unattainable.

If we instead define a distribution with correlation ρ as a distribution for which

$$E[x_1 \cdot x_2] = E[x_1] \cdot E[x_2] + \rho \cdot \sigma[x_1] \cdot \sigma[x_2],$$

then the degenerate distribution $x_1 \equiv E_1$, $x_2 \equiv E_2$, with $\sigma[x_1] = \sigma[x_2] = 0$, is a distribution with a given ρ for which $E[x_1 \cdot x_2] = E_1 \cdot E_2$. Under this alternative definition, closure is not needed.

Proof. When $\rho = 0$, then, by definition of the correlation, $E[x_1 \cdot x_2] = E_1 \cdot E_2$. So, it is sufficient to consider values of $\rho \neq 0$. In this proof, we will only consider the case $\rho > 0$; the case $\rho < 0$ is similar.

We first prove that the value $E[x_1 \cdot x_2]$ always belongs to the interval $[E_1 \cdot E_2, E_1 \cdot E_2 + \rho \cdot \sigma]$. $E_1 \cdot E_2$ is the lower bound because, since $\rho > 0$, we have $E[x_1 \cdot x_2] = E_1 \cdot E_2 + \rho \cdot \sigma[x_1] \cdot \sigma[x_2] > E_1 \cdot E_2$.

To prove the upper bound, we show that for each x_i, $\sigma^2[x_i] \leq (E_i - \underline{x}_i) \cdot (\overline{x}_i - E_i)$. Let us first consider discrete distributions that take values $x_i^{(j)} \in [\underline{x}_i, \overline{x}_i]$ $(1 \leq j \leq N)$ with probabilities $p^{(j)} \geq 0$ such that $\sum\limits_{j=1}^{N} p^{(j)} = 1$. For

such distributions, the constraint $E[x_i] = E_i$ takes the form $\sum_{j=1}^{N} p^{(j)} \cdot x_i^{(j)} = E_i$. Under these constraints, let us find the largest possible value of

$$\sigma^2[x_i] = E[x_i^2] - E_i^2 = \sum_{j=1}^{N} p^{(j)} \cdot \left(x_i^{(j)}\right)^2 - E_i^2.$$

In terms of the unknown probabilities $p_i^{(j)}$, we are minimizing a linear function under linear constraints (equalities and inequalities). Geometrically, the set of all points that satisfy several linear constraints is a polytope. It is well known that to find the minimum of a linear function on a polytope, it is sufficient to consider its vertices (this is the idea behind linear programming). In algebraic terms, a vertex can be characterized by the fact that for N variables, N of the original constrains are equalities. Thus, in our case, all but two probabilities $p_i^{(j)}$ must be equal to 0, i.e., the distribution must be located at two points x_i^- and x_i^+. Since the mean is E_i, we these values must be on different sides of E_i. Without losing generality, we can thus assume that $x_i^- \le E_i \le x_i^+$.

We have already mentioned that for 2-point distributions, once the points x_i^- and x_i^+ are fixed, the condition that the mean equals E_i uniquely determines the probabilities, and the resulting variance is $(x_i^+ - E_i) \cdot (E_i - x_i^-)$. When $x_i^+ \le \overline{x}_i$ and $x_i^- \ge \underline{x}_i$, the largest value of this product is attained when x_i^+ attains its largest possible value $\overline{x}_i$, and x_i^- attains its smallest possible value $\underline{x}_i$. Thus, for discrete distributions, $\sigma^2[x_i] \le (\overline{x}_i - E_i) \cdot (E_i - \underline{x}_i)$.

An arbitrary distribution can be approximated by discrete ones to arbitrary accuracy (in weak topology), so this inequality is true for all distributions. Thus, $\sigma[x_1] \cdot \sigma[x_2] \le \sigma$, and the equality $E[x_1 \cdot x_2] = E_1 \cdot E_2 + \rho \cdot \sigma[x_2] \cdot \sigma[x_2]$ implies that $E[x_1 \cdot x_2] \le E_1 \cdot E_2 + \rho \cdot \sigma$.

We now prove that both endpoints are exact. For every $\varepsilon > 0$, if we take a distribution in which each x_i is located in the ε-vicinity of E_i, then $x_1 \cdot x_2$ (and hence $E[x_1 \cdot x_2]$) is located in the close vicinity of $E_1 \cdot E_2$. When $\varepsilon \to 0$, we conclude that $E[x_1 \cdot x_2]$ can be arbitrarily close to $E_1 \cdot E_2$, so the lower endpoint is indeed exact.

To complete the proof, we next show that the upper endpoint $E_1 \cdot E_2 + \rho \cdot \sigma$ is attainable, and thus also exact. Indeed, as we have mentioned, the largest possible value $E_{\max}$ is attained for a joint distribution in which both marginal distributions are 2-point ones, located on the endpoints of the corresponding interval $[\underline{x}_i, \overline{x}_i]$, and that for such distributions, $\sigma^2[x_i] = (\overline{x}_i - E_i) \cdot (E_i - \underline{x}_i)$. In general, distributions with such marginals are located at 4 vertices of the rectangle $[\underline{x}_1, \overline{x}_1] \times [\underline{x}_2, \overline{x}_2]$. The set of such distributions is determined by linear constraints and is, thus, connected. Along this set, the correlation ranges from 0 to the value $\rho_{\max}$. Since $\rho \in [0, \rho_{\max}]$ and correlation continuously depends on the probabilities, these exists an intermediate value of these probabilities where the correlation exactly equals the given value ρ.

The theorem is proven.

Case of correlation known with interval uncertainty. We can handle the case of an interval $[\underline{\rho}, \overline{\rho}]$ of possible values for ρ instead of an exact value of ρ by simply combining the intervals from Theorem 3 and using the fact that the corresponding formulas monotonically depend on ρ.

Theorem 4. *Let $[\underline{x}_1, \overline{x}_1]$ and $[\underline{x}_2, \overline{x}_2]$ be given intervals, $E_1 \in [\underline{x}_1, \overline{x}_1]$ and $E_2 \in [\underline{x}_1, \overline{x}_1]$ be given numbers, and $[\underline{\rho}, \overline{\rho}]$ be a subinterval of the interval $[\rho_{\min}, \rho_{\max}]$. Then the closure $[\underline{E}, \overline{E}]$ of the range of possible values $E[x_1, x_2]$ for all possible distributions for which:*

- *x_1 is located in $[\underline{x}_1, \overline{x}_1]$, and x_2 is located in $[\underline{x}_2, \overline{x}_2]$;*
- *$E[x_1] = E_1$, and $E[x_2] = E_2$; and*
- *$\rho[x_1, x_2] \in [\underline{\rho}, \overline{\rho}]$*

equals

- *for $0 \le \underline{\rho}$: $[E_1 \cdot E_2, E_1 \cdot E_2 + \overline{\rho} \cdot \sigma]$;*
- *for $\overline{\rho} \le 0$: $[E_1 \cdot E_2 + \underline{\rho} \cdot \sigma, E_1 \cdot E_2]$;*
- *for $\underline{\rho} \le 0 \le \overline{\rho}$: $[E_1 \cdot E_2 + \underline{\rho} \cdot \sigma, E_1 \cdot E_2 + \overline{\rho} \cdot \sigma]$.*

V. First Problem: Auxiliary Results

Computationally efficient expressions for $E_{\min}$ and $E_{\max}$.

Proposition 1.

$$E_{\max} = E_1 \cdot E_2 + \min((E_1 - \underline{x}_1) \cdot (\overline{x}_2 - E_2), (\overline{x}_1 - E_1) \cdot (E_2 - \underline{x}_2));$$

$$E_{\min} = E_1 \cdot E_2 - \min((E_1 - \underline{x}_1) \cdot (E_2 - \underline{x}_2), (\overline{x}_1 - E_1) \cdot (\overline{x}_2 - E_2)).$$

Proof. Let us first simplify the expression for $E_{\max}$ from Theorem 1. When $p_1 \le p_2$, we get

$$E_{\max} = p_1 \cdot \overline{x}_1 \cdot \overline{x}_2 + (p_2 - p_1) \cdot \underline{x}_1 \cdot \overline{x}_2 + (1 - p_2) \cdot \underline{x}_1 \cdot \underline{x}_2 =$$

$$p_1 \cdot (\overline{x}_1 - \underline{x}_1) \cdot \overline{x}_2 + p_2 \cdot \underline{x}_1 \cdot (\overline{x}_2 - \underline{x}_2) + \underline{x}_1 \cdot \underline{x}_2.$$

Substituting the definitions of p_i, we conclude that

$$E_{\max} = (E_1 - \underline{x}_1) \cdot \overline{x}_2 + (E_2 - \underline{x}_2) \cdot \underline{x}_1 + \underline{x}_1 \cdot \underline{x}_2.$$

Opening parentheses, we get

$$E_{\max} = E^{(1)} \stackrel{\text{def}}{=} E_1 \cdot \overline{x}_2 - \underline{x}_1 \cdot \overline{x}_2 + E_2 \cdot \underline{x}_1.$$

By using the symmetry between x_1 and x_2, we can now conclude that when $p_1 \ge p_2$,

$$E_{\max} = E^{(2)} \stackrel{\text{def}}{=} E_2 \cdot \overline{x}_1 - \overline{x}_1 \cdot \underline{x}_2 + E_1 \cdot \underline{x}_2.$$

The condition $p_1 \le p_2$ is equivalent to

$$(E_1 - \underline{x}_1) \cdot (\overline{x}_2 - \underline{x}_2) \le (E_2 - \underline{x}_2) \cdot (\overline{x}_1 - \underline{x}_1),$$

i.e.,

$$E_1 \cdot \overline{x}_2 - E_1 \cdot \underline{x}_2 - \underline{x}_1 \cdot \overline{x}_2 + \underline{x}_1 \cdot \underline{x}_2 \le E_2 \cdot \overline{x}_1 - E_2 \cdot \underline{x}_1 - \overline{x}_1 \cdot \underline{x}_2 + \underline{x}_1 \cdot \underline{x}_2.$$

Subtracting the common term $\underline{x}_1 \cdot \underline{x}_2$ from both sides and moving terms to other sides, we get an equivalent form of this inequality:

$$E_1 \cdot \overline{x}_2 - \underline{x}_1 \cdot \overline{x}_2 + E_2 \cdot \underline{x}_1 \le E_2 \cdot \overline{x}_1 - \overline{x}_1 \cdot \underline{x}_2 + E_1 \cdot \underline{x}_2,$$

i.e., $E^{(1)} \le E^{(2)}$. So, if $p_1 \le p_2$, i.e., if $E^{(1)} \le E^{(2)}$, we get $E_{\max} = E^{(1)}$; otherwise, we get $E_{\max} = E^{(2)}$. These two cases can be combined into a single formula $E_{\max} = \min(E^{(1)}, E^{(2)})$, i.e.,

$$E_{\max} = \min(E_1 \cdot \overline{x}_2 - \underline{x}_1 \cdot \overline{x}_2 + E_2 \cdot \underline{x}_1, E_2 \cdot \overline{x}_1 - \overline{x}_1 \cdot \underline{x}_2 + E_1 \cdot \underline{x}_2).$$

By adding $-E_1 \cdot E_2$ to both expressions $E^{(1)}$ and $E^{(2)}$, we get the desired expression for $E_{\max}$.

Since $E[x_1 \cdot x_2] = -E[(-x_1) \cdot x_2]$, where $-x_1 \in [-\overline{x}_1, \underline{x}_1]$ with $E[-x_1] = -E_1$, we have

$$E_{\min} \stackrel{\text{def}}{=} \min E[x_1 \cdot x_2] = -\max E[(-x_1) \cdot x_2].$$

Hence, the new expression for $E_{\max}$ leads to the desired expression for $E_{\min}$. The proposition is proven.

Can we propagate correlations through computations? In straightforward interval computations, we propagate intervals through computations; can we similarly propagate correlations? The following result shows that it is not easy even for addition:

Proposition 2. *If we know that $\rho[x_1, x_2] = \rho$, then the only possible conclusion about $\rho' = \rho[x_1, x_1 + x_2]$ is that $\rho' \in [\rho, 1]$.*

Proof. If we take $x_1 \ll x_2$, we get $\rho' \approx \rho$, and if we take $x_2 \ll x_1$, we get $\rho' \approx 1$. The smaller the corresponding ratio x_1/x_2 or x_2/x_1, the closer we are, correspondingly, to ρ and to 1.

Let us prove that ρ' cannot be smaller than ρ. Since correlation can be defined in terms of the differences $x_i - E[x_i]$, we can shift both variables to $E[x_i] = 0$ without changing the correlations $\rho[x_1, x_2]$ and $\rho[x_1, x_1 + x_2]$; thus, is it sufficient to prove the desired inequality $\rho' \ge \rho$ for the case when $E[x_i] = 0$. In this case, if we denote $\sigma_i \stackrel{\text{def}}{=} \sigma[x_i]$, we get

$$\rho' = \frac{E[x_1 \cdot (x_1 + x_2)]}{\sigma_1 \cdot \sigma[x_1 + x_2]} = \frac{\sigma_1^2 + E[x_1 \cdot x_2]}{\sigma_1 \cdot \sigma[x_1 + x_2]}.$$

Here, since $E_i = 0$, we have $E[x_1 \cdot x_2] = \rho \cdot \sigma_1 \cdot \sigma_2$. Similarly,

$$\sigma^2[x_1 + x_2] = E[(x_1 + x_2)^2] = E[x_1^2] + E[x_2^2] + 2 \cdot E[x_1 \cdot x_2] =$$

$$\sigma_1^2 + \sigma_2^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2,$$

so the above expression for ρ' takes the form: $\rho' = \dfrac{\sigma_1 + \rho \cdot \sigma_1 \cdot \sigma_2}{\sigma_1 \cdot \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2}}$, and the desired inequality $\rho' \ge \rho$ takes the form $\dfrac{\sigma_1^2 + \rho \cdot \sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2}} \ge \rho$. Multiplying both sides by the denominator, we get the equivalent inequality

$$\sigma_1 + \rho \cdot \sigma_2 \ge \rho \cdot \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2}. \qquad (4)$$

If $\rho \ge 0$, then we can square both sides and get an equivalent inequality

$$\sigma_1^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2 + \rho^2 \cdot \sigma_2^2 \ge \rho^2 \cdot (\sigma_1^2 + \sigma_2^2 + 2\rho \cdot \sigma_1 \cdot \sigma_2).$$

Subtracting $\rho^2 \cdot \sigma_2^2$ from both sides, and moving all the terms to the right-hand side, we get an equivalent inequality

$$\sigma_1^2 \cdot (1 - \rho^2) + 2\rho \cdot \sigma_1 \cdot \sigma_2 \cdot (1 - \rho^2) \ge 0,$$

which is always true for $\rho \ge 0$ (since $\rho \le 1$).

If $\rho < 0$, the right-hand side of (4) is negative, so we consider two possible cases. The first case is when

$$\sigma_1 + \rho \cdot \sigma_2 \ge 0.$$

Then inequality (4) is automatically true.

The second case is when $\sigma_1 + \rho \cdot \sigma_2 < 0$. In this case, (4) is equivalent to

$$0 < -\sigma_1 + |\rho| \cdot \sigma_2 \le |\rho| \cdot \sqrt{\sigma_1^2 + \sigma_2^2 - 2|\rho| \cdot \sigma_1 \cdot \sigma_2}.$$

By squaring both sides, we get an equivalent inequality

$$\sigma_1^2 - 2|\rho| \cdot \sigma_1 \cdot \sigma_2 + \rho^2 \cdot \sigma_2^2 \le \rho^2 \cdot (\sigma_1^2 + \sigma_2^2 - 2|\rho| \cdot \sigma_1 \cdot \sigma_2).$$

Subtracting $\rho^2 \cdot \sigma_2^2$ from both sides, and moving all the terms to the right-hand side, we get an equivalent inequality

$$\sigma_1^2 \cdot (1 - \rho^2) - 2|\rho| \cdot \sigma_1 \cdot \sigma_2 \cdot (1 - \rho^2) \le 0.$$

Dividing both sides by $\sigma_1 \cdot (1 - \rho^2) > 0$, we get an equivalent inequality $\sigma_1 - 2|\rho| \cdot \sigma_2 \le 0$. We consider the case when $\sigma_1 - |\rho| \cdot \sigma_2 < 0$, hence $\sigma_1 - 2|\rho| \cdot \sigma_2 \le \sigma_1 - |\rho| \cdot \sigma_2 < 0$. The inequality is proven.

Since $x_1 - x_2 = x_1 + (-x_2)$, and $\rho[x_1, -x_2] = -\rho[x_1, x_2]$, we have the following corollary:

Proposition 3. *If we know that $\rho[x_1, x_2] = \rho$, then:*

- *the best possible conclusion about $\rho' = \rho[x_1, x_1 - x_2]$ is that $\rho' \in [-\rho, 1]$;*
- *the best possible conclusion about $\rho'' = \rho[x_2, x_1 - x_2]$ is that $\rho'' \in [-1, \rho]$.*

For a *unary* linear function $f(x_1) = a \cdot x_1 + b$, we get $\rho[x_1, f(x_1)] = 1$ for $a > 0$ and $\rho[x_1, f(x_1)] = -1$ for $a < 0$. For non-linear unary functions $f(x_1)$, we can get different intermediate values. As an example, we take $f(x_1) = x_1^2$. Then, $\rho \approx 1$, e.g., for a 2-point distribution located at $a - \varepsilon$ and $a + \varepsilon$ (where $a > 0$ and $\varepsilon \to 0$) with probability 1/2. $\rho \approx -1$, e.g., for a similar distribution with $a < 0$. We get all possible values from -1 to 1 for intermediate distributions.

VI. First Problem: Remaining Open Questions

What if we have a multiple product? For the case of unknown correlation, analytical formulas were obtained in [10].

What if we use different correlation characteristics [16], e.g., the Spearman and Kendall correlations, or copulas [5], [14]?

What about the ranges for $E[\min(x_1, x_2)]$ and $E[\max(x_1, x_2)]$ under a given correlation (for the case of unknown correlation, such ranges were described in [9]).

VII. How to Measure Loss of Privacy: Introduction to the Second Problem

Measuring loss of privacy is important. Privacy means, in particular, that we do not disclose all the 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; see, e.g., [2], [13].

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.

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. 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 $\lceil \log_2(n) \rceil$.

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, it is let us 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 this 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 $\log(n)$ 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:

$$E[N_i] = E[n_1] + E[n_2] + \ldots + E[n_N] = N \cdot p_i.$$

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 3.

- *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*

$N \cdot p_i - k \cdot \sigma_i \le N_i \le N + k \cdot \sigma_i$, *where* $\sigma_i \stackrel{\text{def}}{=} \sqrt{p_i \cdot (1-p_i) \cdot N}$.
- *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. (Shannon) *When the number of events N tends to infinity, the average number of questions tends to*

$$S(p) \stackrel{\text{def}}{=} -\sum p_i \cdot \log_2(p_i).$$

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.

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,

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

The first sum in this expression is the integral sum for the integral $S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(x)\,dx$ (called the *entropy*), so $S \approx S(\rho) - \log_2(2\varepsilon)$. (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 $\rho(x)$. [7], [8], [12].

Often, this definition 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 definition 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., [4]) to uniquely reconstruct the salary.

What we plan to do. The main objective of this part of the paper is to propose a new definition of privacy loss which is in better accordance with our intuition.

VIII. Second Problem: Our Main Idea

Why information is not always a perfect measure of loss of privacy. In our opinion, 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 $F(x) \stackrel{\text{def}}{=} f(x) \cdot (1+\alpha)$, 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 1. *Let $\mathcal{P}$ be a class of probability distributions on a real line, and let $M > 0$ be a real number. By the* amount of privacy *$A(\mathcal{P})$ related to $\mathcal{P}$, 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 $\rho \in \mathcal{P}$, 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 $\rho \in \mathcal{P}$ 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 $\mathcal{P}$ 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 $\mathcal{P}$ 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 $\mathcal{P}$ to a new class $\mathcal{Q}$. 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 $A(\mathcal{P})$ 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

$$\frac{A(\mathcal{P}) - A(\mathcal{Q})}{A(\mathcal{P})}. \tag{5}$$

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

Definition 2.

- *By a* privacy loss*, we mean a pair $\langle \mathcal{P}, \mathcal{Q} \rangle$ of classes of probability distributions.*
- *For each privacy loss $\langle \mathcal{P}, \mathcal{Q} \rangle$, by the* measure of a privacy loss*, we mean the ratio (5).*

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 4. *For each pair $\langle \mathcal{P}, \mathcal{Q} \rangle$, the measure of the privacy loss is the same for all $M > 0$.*

Proof. 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

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

$$M \cdot \left(F_0(x_0) - \int \rho(x) \cdot F_0(x)\, dx \right). \tag{6}$$

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 (6). Thus, the maximized values corresponding to M and $M_0 = 1$ different by a factor M. Hence, the resulting amounts of privacy $A(\mathcal{P})$ and $A_0(\mathcal{P})$ corresponding to M and M_0 also differ by a factor M: $A(\mathcal{P}) = M \cdot A_0(\mathcal{P})$. Substituting this expression for $A(\mathcal{P})$ (and a similar expression for $A(\mathcal{Q})$) into the definition (5), we can therefore conclude that $\dfrac{A(\mathcal{P}) - A(\mathcal{Q})}{A(\mathcal{P})} = \dfrac{A_0(\mathcal{P}) - A_0(\mathcal{Q})}{A_0(\mathcal{P})}$, i.e., that the measure of privacy is indeed the same for M and $M_0 = 1$. The proposition is proven.

IX. 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 if even.

Proposition 5. *Let $[l, u] \subseteq [L, U]$ be intervals, let $\mathcal{P}$ be the class of all probability distributions located on the interval $[L, U]$, and let $\mathcal{Q}$ be the class of all probability distributions located on the interval $[l, u]$. For this pair $\langle \mathcal{P}, \mathcal{Q} \rangle$, the measure of the privacy loss if equal to $1 - \dfrac{u - l}{U - L}$.*

Proof. Due to Proposition 4, 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 $A(\mathcal{P}) = U - L$.

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,

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

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 $A(\mathcal{Q}) = u - l$, so we get the desired measure of the privacy loss. The proposition is proven.

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 $\mathcal{Q}$ is the class of all distributions $\rho(x)$ which are located, with probability 1, on even values x, we get $A(\mathcal{Q}) = A(\mathcal{P})$. 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.

X. Conclusion

In many practical situations, there is a need to combine interval and probabilistic uncertainty. The need for such a combination leads to two types of problems: how to *process* the given combined uncertainty, and how to *gauge* the amount of uncertainty. In this paper, we presented two examples that illustrate how the corresponding problems can be solved.

The first example is related to the fact that the traditional engineering approach to error estimation assumes that we *know the probabilities* of different values of measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. Yet in many practical situations, we only know the upper bound Δ_i for this error. Hence after the measurement, the only information that we have about x_i is that it belongs to the *interval* $\mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. In this case, we have a classic *interval computations* problem: find the narrowest possible interval $\mathbf{y}$ enclosing all possible values of the result $y = f(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i$. In this paper, we generalized the preceding case by discussing what to do when, in addition to the bounds Δ_i, we permit *partial information* about the *probabilities* of different values of Δx_i and their *correlations*.

The second example is related to the following problem. To compare different schemes for preserving privacy, it is important to be able to gauge loss of privacy. Since loss of privacy means that we gain new information about a person, it seems natural to measure the loss of privacy by the amount of information that we gained. However, this seemingly natural definition is not perfect: when we originally know 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 exactly as much information (one bit) as when we learn that the salary is an even number – however, intuitively, in the first case, we have a substantial privacy loss while in the second case, the privacy loss is minimal. In this paper, we proposed a new definition of privacy loss that is in better agreement with our intuition. This new definition is based on estimating worst-case financial losses caused by the loss of privacy.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] D. Berleant and J. Zhang, "Using Pearson correlation to improve envelopes around the distributions of functions," *Reliable Computing*, 2004, Vol. 10, No. 2, pp. 139–161.

[2] V. Chirayath, *Using entropy as a measure of privacy loss in statistical databases*, Master's thesis, University of Texas at El Paso, Computer Science Department, 2004.

[3] B. Chokr and V. Kreinovich. "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*, Wiley, N.Y., 1994, pp. 555–576.

[4] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

[5] S. Ferson, *RAMAS RiskCalc: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

[6] L. Jaulin, M. Keiffer, O. Didrit, E. Walter, *Applied Interval Analysis*, Springer-Verlag, London, 2001.

[7] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, Massachusetts, 2003.

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

[9] V. Kreinovich, "Probabilities, Intervals, What Next? Optimization Problems Related to Extension of Interval Computations to Situations with Partial Information about Probabilities", *Journal of Global Optimization*, 2004, Vol. 29, No. 3, pp. 265–280.

[10] V. Kreinovich, S. Ferson, and L. Ginzburg, "Exact Upper Bound on the Mean of the Product of Many Random Variables With Known Expectations", Reliable Computing, 2003, Vol. 9, No. 6, pp. 441–463.

[11] V. Kreinovich, A. Lakeyev, J. Rohn, P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

[12] V. Kreinovich, G. Xiang, and S. Ferson, "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 24nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pp. 80–85.

[13] L. Longpré, V. Kreinovich, E. Freudenthal, M. Ceberio, F. Modave, N. Baijal, W. Chen, V. Chirayath, G. Xiang, and J. I. Vargas "Privacy: Protecting, Processing, and Measuring Loss", *Abstracts of the 2005 South Central Information Security Symposium SCISS'05*, Austin, Texas, April 30, 2005, p. 2.

[14] R. B. Nelsen, *Introduction to Copulas*, Springer Verlag, New York, 1999.

[15] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer-Verlag, New York, 2005.

[16] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.