How to Measure Loss of Privacy

Luc Longpré and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
{longpre,vladik}@utep.edu

Abstract

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 propose 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.

1 Introduction

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., [1, 7].

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 $[L, U]$ of width $U - L$, then we need to subdivide it into $n \sim (U - L)/(2\varepsilon)$ sub-domains, so we must ask $\log_2(n) \sim \log_2(U - L) - \log_2(\varepsilon) - 1$ questions. In the limit, the term that does not depend on ε leads to $\log_2(U - L)$. For continuous probability distributions, we get the standard Shannon's expression $\log_2(n) \sim S - \log_2(2\varepsilon)$, where $S = -\int \rho(x) \cdot \log_2 \rho(x)\, dx$ [4, 5, 6]; see Appendix for the derivation of this Shannon's formula.

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., [3]) to uniquely reconstruct the salary.

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

2 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

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

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 (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 1. *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

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

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 $A(\mathcal{P})$ 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 (1), we can therefore conclude that

$$\frac{A(\mathcal{P}) - A(\mathcal{Q})}{A(\mathcal{P})} = \frac{A_0(\mathcal{P}) - A_0(\mathcal{Q})}{A_0(\mathcal{P})},$$

i.e., that the measure of privacy is indeed the same for M and $M_0 = 1$. The proposition is proven.

3 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 2. *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*

$$1 - \frac{u-l}{U-L}.$$

Proof. Due to Proposition 1, 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.

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.

References

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] 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.

Appendix: How to Measure the Amount of Information

In the main text, we use Shannon's estimates for the average number of binary questions. 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 amount of information, 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, 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 A.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 $\sigma_i \stackrel{\text{def}}{=} \sqrt{p_i \cdot (1-p_i) \cdot N}$.*
- *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*

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

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 at the end of this Appendix.

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 $\rho(x)$.

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. If 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 $Q(N) \approx \log_2(N_{cons})$, 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. The total number C_1 of ways to choose N_1 elements out of N is well-known in combinatorics, and is equal to $\binom{N_1}{N} = \frac{N!}{(N_1)! \cdot (N - N_1)!}$. When we choose these N_1 elements, we have a problem in choosing N_2 out of the remaining $N - N_1$ elements, where the outcome is 2; so for every choice of 1's we have $C_2 = \binom{N_2}{N - N_1}$ possibilities to choose 2's. Therefore in order to get the total number of possibilities to choose 1's and 2's, we must multiply C_2 by C_1. Adding 3's, 4's, ..., n's, we get finally the following formula for C:

$$C = C_1 \cdot C_2 \cdot \ldots \cdot C_{n-1} =$$

$$\frac{N!}{N_1! \cdot (N - N_1)!} \cdot \frac{(N - N_1)!}{N_2! \cdot (N - N_1 - N_2)!} \cdot \ldots = \frac{N!}{N_1! \cdot N_2! \cdot \ldots \cdot N_n!}$$

To simplify computations let's use the well-known Stirling formula $k! \sim (k/e)^k \cdot \sqrt{2\pi \cdot k}$. Then, we get

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

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

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

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.

Arguments given in [2] show that the ratio $Q(N)/N = S$ also tends to this entropy. The theorem is proven.