Decision Trees with Soft Numbers

Oren Fivel
School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Be'er Sheva, Israel
fivel@post.bgu.ac.il

Moshe Klein
Dep. of Industrial Engineering
Tel-Aviv University
Tel-Aviv, Israel
mosheklein@mail.tau.ac.il

Oded Maimon
Dep. of Industrial Engineering
Tel-Aviv University
Tel-Aviv, Israel
maimon@tauex.tau.ac.il

***Abstract*—In the classical probability, in continuous random variables there is no distinguishing between the probability involving strict inequality and non-strict inequality. Moreover, a probability involves equality collapse to zero, without distinguishing among the values that we would like that the random variable will have for comparison. This work presents *Soft Probability*, by incorporating of *Soft Numbers* into probability theory. *Soft Numbers* are set of new numbers that are linear combinations of multiples of "ones" and multiples of "zeros". In this work, we develop a probability involving equality as a "soft zero" multiple of a probability density function (PDF). We also extend this notion of soft probabilities to the classical definitions of Complements, Unions, Intersections and Conditional probabilities, and also to the expectation, variance and entropy of a continuous random variable, condition being in a union of disjoint intervals and a discrete set of numbers. This extension provides information regarding to a continuous random variable being within discrete set of numbers, such that its probability does not collapse completely to zero. When we developed the notion of soft entropy, we found potentially another soft axis, multiples of 0log(0), that motivates to explore the properties of those new numbers and applications. We extend the notion of soft entropy into the definition of Cross Entropy and Kullback–Leibler-Divergence (KLD), and we found that a soft KLD is a soft number, that does not have a multiple of 0log(0). Based on a soft KLD, we defined a soft mutual information, that can be used as a splitting criteria in decision trees with data set of continuous random variables, consist of single samples and intervals.**

***Index Terms*—Probability, Continuous Random Variable, PDF, Soft Number, Soft Logic, Soft Entropy**

I. Introduction

A. Research Motivation and Direction

Probability theory is used in order to model processes and phenomenons, involving randomness of the parameters and variables (See Appendix A for a brief review and notations regarding to probability theory). A probability a continues random variable is defined by a Probability Density Function (PDF). The PDF can be used is to approximate the probability of the continuous random variable X to be adjacent to x in the following sense

$$\Pr(x < X \leq x + \Delta x) \approx f_X(x)\Delta x, \tag{1}$$

where $\Delta x > 0$ is a small value, that defines how much this probability is accurate. However, continuous random variables have the following properties:

- No distinguishing between strict inequality and non-strict in equality e.g., $\Pr(X \leq x) = \Pr(X < x)$;
- Equality collapses to zero i.e., $\Pr(X = x) = 0$. Although any value of $x \in S_X$ (S_X denotes the support of X) is possible for X, the the probability of X to be equal to any value of $x \in S_X$ is (almost surely) zero.

Because of these properties, we lose some information regarding to a continuous random variable to have an exact value. On one hand, an event "$X = x$" might be possible (if $x \in S_X$) but improbable (i.e., with zero probability), which seems to be a paradox. On the other hand, we can express the zero probability by of an event "$X = x$" by letting Δx to approach to zero in (1)

$$\Pr(X = x) = f_X(x) \cdot 0. \tag{2}$$

This equation presents the probability $\Pr(X = x)$ as a multiple of zero with a factor of the PDF $f_X(x)$ for all x. Instead of taking $\Pr(X = x)$ to be completely zero, we can assign to it a zero multiple of $f_X(x)$ and compare different probability values for different observation values x. This approach can be implemented by using *Soft Numbers* (see Appendix B and Klein and Maimon's papers e.g., [1], [2] and [3]).

In this work, we introduce the *Soft Numbers* to give a probability interpretation of a continuous random variable to have an exact value, that provides distinguishing between strict inequality and non-strict in equality in the probability function.

B. Organization of the Work

Section II incorporates Soft Numbers into probability theory to present the notion of "Soft Probability". Section III extends this notion to conditional probability. Section IV defines a Soft Expectation, a Variance and a Soft Entropy, where the last generates potentially another soft axis, multiples of $0 \cdot \log 0$. Section V presents an example for application on Decision Trees based on a Soft Mutual Information as a Splitting Criteria. Conclusions and suggestion for future research are shown on sections VI and VII respectively to summarize this work. For completion, Appendix A provides a brief review of probability theory, and Appendix B provides a presentation of Soft Numbers.

II. Soft Probability: Incorporation of Soft Number into Probability Theory

In order to incorporate the notion of (B.10) in Appendix B, we define (A.2) in Appendix A differently for a cumulative distribution function (CDF) of a continuous random variable

$$\mathrm{Ps}(X \le x) = F_X(1 \cdot \bar{0}\dot{+}x), \tag{3}$$

where $\mathrm{Ps}(\cdot)$ is a suggested type of a probability function, dented as a "Soft Probability" [instead of a regular probability notation "$\Pr(\cdot)$" or $P(\cdot)$], and $F_X(\cdot)$ is the regular CDF function of the random variable X but it is applied on a soft number $1 \cdot \bar{0}\dot{+}x$. Our motivation is to generate an alternative evaluation of the probability at the left hand side (LHS), so that we can distinguish between $\mathrm{Ps}(X < x)$ and $\mathrm{Ps}(X \le x)$ for a continuous random variable X [i.e., $\mathrm{Ps}(X < x) \neq \mathrm{Ps}(X \le x)$]. We will show that the evaluation of the soft number at the CDF in the right hand side (RHS) will create this distinction.

The RHS of (3) can be decomposed by (B.10) as follows

$$F_X(1 \cdot \bar{0}\dot{+}x) \overset{\text{def}}{=} f_X(x)\bar{0}\dot{+}F_X(x), \tag{4}$$

The LHS of (3) can be decomposed by separating the event "$X \le x$" into a disjoint union "$X = x \uplus X < x$". In a regular probability, we have the known identities

$$\begin{aligned}\Pr(X \le x) &\overset{\text{"}X=x\text{"}\cap\text{"}X<x\text{"}=\emptyset}{=} \underbrace{\Pr(X = x)}_{=0} + \Pr(X < x)\\ &= \Pr(X < x),\end{aligned}$$

So we do not have a distinction between $\Pr(X \le x)$ and $\Pr(X < x)$. We distinguish between $\mathrm{Ps}(X \le x)$ and $\mathrm{Ps}(X < x)$ by the following definition for $\mathrm{Ps}(X \le x)$

$$\mathrm{Ps}(X \le x) \overset{\text{def}}{=} \mathrm{Ps}(X = x) + \mathrm{Ps}(X < x), \tag{5}$$

so that we define the terms on the LHS as follows

$$\mathrm{Ps}(X = x) \overset{\text{def}}{=} f_X(x)\bar{0}, \tag{6}$$

$$\mathrm{Ps}(X < x) \overset{\text{def}}{=} F_X(x) \equiv \Pr(X < x). \tag{7}$$

By this setup we achieve a distinguishing between $\mathrm{Ps}(X \le x)$ and $\mathrm{Ps}(X < x)$, an also we provide an interpretation to $\mathrm{Ps}(X = x)$ be infinitesimally small but not collapse completely to zero due to the factor $\bar{0}$ of the PDF.

In the next subsection, we provide two examples of implementations on PDFs, Gaussian distribution and uniform distribution, in order to demonstrate the effect of soft numbers (and more precisely, soft zeros) on PDFs.

A. Examples

1) Gaussian distribution: Let X be a Gaussian random variable parameterized by a mean μ and a variance σ^2 [denoted $X \sim N(\mu, \sigma^2)$]. The PDF of X is well known as

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \tag{8}$$

The maximum of the PDF, $\max_x f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$, occurs at $x = \mu$. We would like to have a high probability as X is closer to μ e.g., $\mathrm{Ps}(X = \mu) > \mathrm{Ps}(X = x), \forall x \neq \mu$. By (6) and we have the following definition a soft probability in the Gaussian case

$$\mathrm{Ps}(X = x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \cdot \bar{0}, \tag{9}$$

which presents an absolute low probability of X to have an exact value x but relative high probability when X is closer to μ.

2) Uniform distribution: Let X be Uniformly distributed at the interval (a, b) [denoted $X \sim U(a, b)$]. The PDF of X is well known as

$$f_X(x; a, b) = \frac{1}{b-a} \mathbb{1}_{x \in (a,b)}, \tag{10}$$

where $\mathbb{1}_A$ is the indication function, indicates '1' if 'A' is true and '0' if 'A' is false. Similarly to previous example (but with maximal PDF to be trivially $\frac{1}{b-a}$) we have the following soft probability in the uniform case

$$\mathrm{Ps}(X = x; a, b) = \frac{1}{b-a} \cdot \mathbb{1}_{x \in (a,b)} \cdot \bar{0}, \tag{11}$$

which implies the following property:

$$\begin{aligned}&\forall x, y \in \mathbb{R}, x \in (a, b) \wedge y \notin (a, b)\\ &\Rightarrow \mathrm{Ps}(X = x) = \frac{1}{b-a} \cdot \bar{0} > \mathrm{Ps}(X = y) = 0 \cdot \bar{0}.\end{aligned} \tag{12}$$

This property emphasises the probability to X to have any value within (a, b) is absolutely small, but still relative greater than the probability to have any value outside of which is almost surely impossible).

B. Observations

In soft numbers development, we may consider to distinct between two options to define an absolute value of a soft number: Option 1 is by the definition in (B.10) with $|x|' = \mathrm{sign}(x)$, ignoring the fact that this derivative is not continuous, so that

$$|a\bar{0}\dot{+}x| = \alpha\bar{0} \cdot \mathrm{sign}(x)\dot{+}|x|. \tag{13}$$

Option 2 is to define a soft conjugate of $\alpha\bar{0}\dot{+}x$ to be $(-\alpha)\bar{0}\dot{+}x$ such that

$$\begin{aligned}|\alpha\bar{0}\dot{+}x| &= \sqrt{(\alpha\bar{0}\dot{+}x)((-\alpha)\bar{0}\dot{+}x)}\\ &= \sqrt{-(\alpha\bar{0})^2 + x^2}\\ &= \sqrt{-0 + x^2}\\ &= \sqrt{x^2}\\ &= |x|.\end{aligned} \tag{14}$$

If we use Option 2, then we can have the following properties for a soft probability on a continuous random variable:

1) $\mathrm{Ps}(X \le x) \neq \mathrm{Ps}(X < x)$
   but $|\mathrm{Ps}(X \le x)| = |\mathrm{Ps}(X < x)| > |\mathrm{Ps}(X = x)| = 0$;
2) $f_X(x) > f_X(y) \Rightarrow \mathrm{Ps}(X = x) > \mathrm{Ps}(X = y)$
   but $|\mathrm{Ps}(X = x)| = |\mathrm{Ps}(X = y)| = 0$;

3) $f_X(x) > f_Y(y) \Rightarrow \mathrm{Ps}(X = x) > \mathrm{Ps}(Y = y)$ but $|\mathrm{Ps}(X = x)| = |\mathrm{Ps}(Y = y)| = 0$;
4) $|\mathrm{Ps}(X \leq x)| = \mathrm{Ps}(X < x) = \mathrm{Pr}(X < x) = \mathrm{Pr}(X \leq x)$.

By taking absolute values of the soft probability term, we return to the classic probability results for continuous random variable e.g., not distinguishing between strict inequality and non-strict inequality, and equality collapse to zero.

In the next section, we extend the notion of "Soft Probability" into the events' complements, unions and intersections , and into conditional probability.

III. Complements, Union, Intersection and Conditional Soft Probability

In the following section we extend the notion of "Soft Probability" into the events' complements, unions and intersections, and into conditional probability. In the First Subsection we show the results for complements, unions and intersections corresponding to event with zero probability in the classical probability sense. In the second subsection we show the results for a conditional of soft probability, referring to Kolmogorov definition and Bayes theorem.

A. Complements, Unions and Intersections

Recall that a probability of A^c, a complement of the event A, is given by

$$\mathrm{Pr}(A^c) = 1 - \mathrm{Pr}(A). \tag{15}$$

A Soft probability of a complement is defined similarly as follows

$$\mathrm{Ps}(A^c) = 1 - \mathrm{Ps}(A). \tag{16}$$

Therefore, we have the following probability complement for a continuous random variable X:

$$\begin{aligned} \mathrm{Ps}(X \neq x) &= 1 - \mathrm{Ps}(X = x) \\ &= [-f_X(x)]\bar{0}\dot{+}1. \end{aligned} \tag{17}$$

This equation distinguishes among different values of x for the event $X \neq x$ to be with almost surely with probability 1 due the the soft zero term $[-f_X(x)]\bar{0}$. This equation is analogous to the event $X \neq x$ to have zero probability almost surely, correct by the soft zero term $[-f_X(x)]\bar{0}$.

In order to analyse unions and intersections, we need to consider two cases: unions and intersections among singletons events $X = x, X = y$ etc; unions and intersections between a singleton event $X = x$ and a range event e.g. $a \leq X \leq b$.

For all $x \neq y$ we have that the events $X = x$ and $X = y$ are disjoint, and the for a union we have

$$\begin{aligned} \mathrm{Ps}(X = x \cup X = y) &= \mathrm{Ps}(X = x) + \mathrm{Ps}(X = y) \\ &= [f_X(x) + f_X(y)]\bar{0}. \end{aligned} \tag{18}$$

For an intersection we have

$$\mathrm{Ps}(X = x \cap X = y) = \mathbb{1}_{x=y} f_X(x)\bar{0}, \tag{19}$$

where the indicator $\mathbb{1}_{x=y}$ is zero in the case that $x \neq y$. More generally, we have the following soft probabilities for the following set $\{x_i\}_{i=1}^n$ with distinct values:

$$\mathrm{Ps}\left(\bigcup_{i=1}^{n} X = x_i\right) = \sum_{i=1}^{n} \mathrm{Ps}(X = x_i) = \left[\sum_{i=1}^{n} f_X(x_i)\right]\bar{0}, \tag{20}$$

and

$$\mathrm{Ps}\left(\bigcap_{i=1}^{n} X = x_i\right) = \mathbb{1}_{x_i = x_j}^{\forall i,j \in \{1,2,\ldots,n\}} f_X(x_i)\bar{0}. \tag{21}$$

In order to analyse unions and intersections, between a singleton event $X = x$ and a range event e.g. $a \leq X \leq b$, we need to distinguish among x's values that are either between a and b or not. Moreover we need to distinguish between the strict inequality case $a < X < b$ and the non-strict inequality $a \leq X \leq b$. For simplicity, assume $a < b$ and without loss of generality (WLOG) assume $x \neq a$ and $x \neq b$.

For the strict inequality case $a < X < b$ we have the union

$$\mathrm{Ps}(X = x \cup a < X < b) = \mathbb{1}_{x \notin (a,b)} f_X(x)\bar{0}\dot{+}[F_X(b) - F_X(a)], \tag{22}$$

and for the intersection

$$\mathrm{Ps}(X = x \cap a < X < b) = \mathbb{1}_{x \in (a,b)} f_X(x)\bar{0}. \tag{23}$$

This union is a soft number when x is not in the interval (a, b) and a real number when it does. This intersection is a soft zero when x is in (a, b) and an absolute zero when it doesn't.

For the non-strict inequality case $a \leq X \leq b$ we have the union

$$\begin{aligned} &\mathrm{Ps}(X = x \cup a \leq X \leq b) = \\ &[\mathbb{1}_{x \notin [a,b]} f_X(x) + f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)], \end{aligned} \tag{24}$$

and for the intersection

$$\mathrm{Ps}(X = x \cap a \leq X \leq b) = [\mathbb{1}_{x \in [a,b]} f_X(x)]\bar{0}. \tag{25}$$

the two terms $f_X(a) + f_X(b)$ in (24) are added to the soft zero part, due to (20).

Recall the relation between a union and an intersection of two events A, B, according to De Morgan's Law, we have

$$\mathrm{Pr}(A \cup B) = \mathrm{Pr}(A) + \mathrm{Pr}(B) - \mathrm{Pr}(A \cap B). \tag{26}$$

It can be shown that the soft probabilities in (22)-(25) hold for De Morgan's Law (26). For example $A = \{X = x\}$, $B = \{a \leq X \leq b\}$ and $x \notin [a, b]$, we have

$$\begin{aligned} &\mathrm{Ps}(X = x \cup a \leq X \leq b) = \\ &\mathrm{Ps}(X = x) + \mathrm{Pr}(a \leq X \leq b) - \mathrm{Pr}(X = x \cap a \leq X \leq b). \end{aligned} \tag{27}$$

The LHS is

$$[f_X(x) + f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]$$

and the RHS is

$$f_X(x)\bar{0} + \left[\{f_X(a) + f_X(b)\}\,\bar{0}\dot{+}\{F_X(b) - F_X(a)\}\right] - 0,$$

so that we obtain the LHS to be equal to the RHS, and thus we have a "Soft De Morgan's Law"

$$\mathrm{Ps}(A \cup B) = \mathrm{Ps}(A) + \mathrm{Ps}(B) - \mathrm{Ps}(A \cap B). \tag{28}$$

In the next subsection, we show the results for a conditional of soft probability, referring to Kolmogorov definition and Bayes theorem.

B. Conditional Probability

Recall Kolmogorov definition for conditional probability

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \tag{29}$$

and for Bayes theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}, \tag{30}$$

We define a "Soft Conditional Probability" similarly, e.g., for $x, y \in S_X$, let $A = \{X = x\}$, $B = \{X = y\}$, and at the LHS of Kolmogorov definition (29) we have

$$\frac{\mathrm{Ps}(X = x \cap X = y)}{\mathrm{Ps}(X = y)} = \frac{\mathbb{1}_{x=y} f_X(x)\bar{0}}{f_X(y)\bar{0}} = \frac{\mathbb{1}_{x=y} \cdot \bar{0}}{1 \cdot \bar{0}}. \tag{31}$$

With a definition of $\frac{1\cdot\bar{0}}{1\cdot\bar{0}} = 1$ and $\frac{0\cdot\bar{0}}{1\cdot\bar{0}} = 0$, the conditional soft probability is given by

$$\mathrm{Ps}(X = x|X = y) = \mathbb{1}_{x=y}. \tag{32}$$

In this case we have a trivial equality with optional real values 0 or 1. For comparison with Bayes theorem (30)

$$\frac{\mathrm{Ps}(X = y|X = x)\mathrm{Ps}(X = x)}{\mathrm{Ps}(X = y)} = \frac{\mathbb{1}_{y=x} f_X(x)\bar{0}}{f_X(y)\bar{0}} = \mathbb{1}_{x=y}. \tag{33}$$

Now we consider $x, y \in S_X$, let $A = \{X = x\}$, $B = \{a \le X \le b\}$, with $x, a, b \in S_X$ such that $a < b$,$x \neq a$ and $x \neq b$. At the LHS of Kolmogorov definition (29) we have

$$\frac{\mathrm{Ps}(X = x \cap a \le X \le b)}{\mathrm{Ps}(a \le X \le b)} = \frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]}. \tag{34}$$

When applying Bayes theorem (30), we have

$$\frac{\mathrm{Ps}(a \le X \le b|X = x)\mathrm{Ps}(X = x)}{\mathrm{Ps}(a \le X \le b)} = \frac{[\mathrm{Ps}(a \le x \le b|X = x)] f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]}, \tag{35}$$

where $\mathrm{Ps}(a \le x \le b|X = x) = \mathrm{Ps}(a \le x \le b) = \mathbb{1}_{x\in[a,b]}$. Both Kolmogorov theorem form and Bayes theorem form are equal, and therefore

$$\mathrm{Ps}(X = x|a \le X \le b) = \frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]}. \tag{36}$$

We can simplify the RHS by the property

$$\frac{A\bar{0}}{B\dot{+}C\bar{0}} = \frac{A\bar{0}}{B\dot{+}C\bar{0}} \cdot \frac{B\dot{+}(-C)\bar{0}}{B\dot{+}(-C)\bar{0}} = \frac{AB\bar{0}}{B^2} = \frac{A\bar{0}}{B},$$

and we have the following conditional soft probability with a given non-strict inequality condition:

$$\mathrm{Ps}(X = x|a \le X \le b) = \frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{F_X(b) - F_X(a)}, \tag{37}$$

and for a given strict inequality condition, we have.

$$\mathrm{Ps}(X = x|a < X < b) = \frac{\mathbb{1}_{x\in(a,b)} f_X(x)\bar{0}}{F_X(b) - F_X(a)}. \tag{38}$$

The meaning of these last two equation is that we have a soft zero when the observation x makes sense (i.e. between a and b), and it is an absolute zero if x makes no sense (i.e. not between a and b), due to the indicator in the numerator. In addition, division by the denominator $F_X(b) - F_X(a) \in (0, 1)$ makes higher probability than the unconditional probability, which make sense since we have an additional information regarding to the random variable X to be between a and b. In the next subsection, we extend the notion of soft probability for 2 continuous random variables, based on a Soft De Morgan's Law (28).

C. Extension of Soft Probability for 2 Dimensions

Suppose that X and Y are two continuous random variables. By the regular De Morgan's Law (26), we can decompose the regular probability object $\Pr(X \le x, Y \le y)$ into a sum of the following probabilities

$$\Pr(X \le x, Y \le y) = \overbrace{[\Pr(X < x, Y = y) + \Pr(X = x, Y < y) + \Pr(X = x, Y = y)]}^{0} + \Pr(X < x, Y < y), \tag{39}$$

such that each of the first three terms in the bracket collapses to zero in the classical probability. We define the soft probability object $\mathrm{Ps}(X \le x, Y \le y)$ in 2 random variables based on a Soft De Morgan's Law (28) as follows

$$\mathrm{Ps}(X \le x, Y \le y) = [\mathrm{Ps}(X < x, Y = y) + \mathrm{Ps}(X = x, Y < y) + \mathrm{Ps}(X = x, Y = y)] + \mathrm{Ps}(X < x, Y < y). \tag{40}$$

In this case, we define the first three terms in the bracket as the following soft zero objects in terms of the CDF $F_{X,Y}(x, y)$ and the PDF $f_{X,Y}(x, y)$:

$$\mathrm{Ps}(X < x, Y = y) = \frac{\partial F_{X,Y}(x, y)}{\partial y} \cdot \bar{0}, \tag{41}$$

$$\mathrm{Ps}(X = x, Y < y) = \frac{\partial F_{X,Y}(x, y)}{\partial x} \cdot \bar{0}, \tag{42}$$

$$\mathrm{Ps}(X = x, Y = y) = \frac{\partial F_{X,Y}(x, y)}{\partial x \partial y} \cdot \bar{0} = f_{X,Y}(x, y) \cdot \bar{0}. \tag{43}$$

the last term is a regular probability along the 1-axis i.e.,

$$\mathrm{Ps}(X<x,Y<y)=\mathrm{Pr}(X<x,Y<y)=F_{X,Y}(x,y), \tag{44}$$

so that $\mathrm{Ps}(X\leq x,Y\leq y)$ equals to the following soft number

$$\begin{aligned}&\mathrm{Ps}(X\leq x,Y\leq y)=\\&\left[\frac{\partial F_{X,Y}(x,y)}{\partial x}+\frac{\partial F_{X,Y}(x,y)}{\partial y}+f_{X,Y}(x,y)\right]\cdot\bar{0}\\&\dot{+}F_{X,Y}(x,y).\end{aligned} \tag{45}$$

Now, we want to construct the soft probability objects $\mathrm{Ps}(X\leq x,Y<y)$ and $\mathrm{Ps}(X\leq x,Y=y)$ [by symmetry, we can construct $\mathrm{Ps}(X<x,Y\leq y)$ and $\mathrm{Ps}(X=x,Y\leq y)$ accordingly]. Based on a Soft De Morgan's Law (28), we construct the soft probability $\mathrm{Ps}(X\leq x,Y<y)$ similarly as follows:

$$\mathrm{Ps}(X\leq x,Y<y)=\frac{\partial F_{X,Y}(x,y)}{\partial x}\cdot\bar{0}\dot{+}F_{X,Y}(x,y). \tag{46}$$

Therefore, we can distinguish among the soft probabilities: $\mathrm{Ps}(X\leq x,Y\leq y)$, $\mathrm{Ps}(X<x,Y<y)$, $\mathrm{Ps}(X\leq x,Y<y)$ and $\mathrm{Ps}(X<x,Y\leq y)$. Similarly, we have

$$\mathrm{Ps}(X\leq x,Y=y)=\left[\frac{\partial F_{X,Y}(x,y)}{\partial y}+f_{X,Y}(x,y)\right]\cdot\bar{0}, \tag{47}$$

that is a soft zero. In the next section, we define soft expectation, soft variance and soft entropy.

IV. Soft Expectation, Variance and Entropy

In this section, we define soft expectation, soft variance and soft entropy. First, we focus on expectation and variance's definitions, recalling their original and known definition and then generalizing then to soft numbers. Second, we do this original definition's recall and soft numbers' generalization to the entropy.

A. Soft Expectation and Variance

Recall for the definition of the Expectation of a random variable X with support S_X

$$\mathrm{E}(X)=\int_{S_X}xdF_x(x)=\mu_X, \tag{48}$$

where $\mathrm{E}(\cdot)$ is the expectation operator defined by the *Lebesgue integral* above, and we denote its result by μ_X (sometime we call it *mean*). For a continuous random variable the Expectation is defined by

$$\mathrm{E}(X)=\int_{S_X}xf_X(x)dx, \tag{49}$$

and for a discrete random variable

$$\mathrm{E}(X)=\sum_{x\in S_X}x\mathrm{Pr}(X=x). \tag{50}$$

The Variance of a random variable is the expectation of the square error from its mean, that is

$$\mathrm{Var}(X)=\mathrm{E}[(X-\mu)^2]=\int_{S_X}(x-\mu_X)^2dF_x(x)=\sigma_X^2, \tag{51}$$

$$\mathrm{Var}(X)=\int_{S_X}(x-\mu_X)^2f_X(x)dx,\text{for continuous case,} \tag{52}$$

$$\mathrm{Var}(X)=\sum_{x\in S_X}(x-\mu_X)^2\mathrm{Pr}(X=x),\text{for discrete case.} \tag{53}$$

Suppose that X is a continuous random variable and $\{x_i\}_{i=1}^n$ and $\{(a_j,b_j)\}_{j=1}^m$ are set of numbers and set of disjoint intervals in the support S_X. Assume also that $\{x_i\}_{i=1}^n$ and $\{(a_j,b_j)\}_{j=1}^m$ are disjoint. WLOG we consider open intervals, otherwise we can exclude the end point a_j,b_j from the interval (a_j,b_j) and include then into the set of numbers $\{x_i\}_{i=1}^n$. Under the above assumption we define a soft expectation of X as the expectation of X conditioned by being within the union of $\{x_i\}_{i=1}^n$ and $\{(a_j,b_j)\}_{j=1}^m$, i.e.,

$$\begin{aligned}&\mathrm{Es}(X|X\in\{x_i\}_{i=1}^n\cup\{(a_j,b_j)\}_{j=1}^m)\\&=\sum_{i=1}^n x_i\mathrm{Ps}(X=x_i)\dot{+}\sum_{j=1}^m\int_{a_j}^{b_j}xf_X(x)dx\\&=\sum_{i=1}^n x_if_X(x_i)\cdot\bar{0}\dot{+}\sum_{j=1}^m\int_{a_j}^{b_j}xf_X(x)dx\\&=\nu_X\bar{0}\dot{+}\kappa_X,\end{aligned} \tag{54}$$

where $\mathrm{Es}(\cdot)$ is a new notation for a Soft Expectation operator. Recall that X is a random variable with a real value, and also all the single point $\{x_i\}_{i=1}^n$ and all the intervals $\{(a_j,b_j)\}_{j=1}^m$ are real. However, due to the soft probability terms $\mathrm{Ps}(X=x_i)$ the result of the LHS of (54) is a soft number. For simplicity we denote the real part of the soft expectation by κ_X, and the soft part by ν_X. With this definition, the soft part ν_X adds some new information regarding to the mean of the continuous random variable X given being within discrete points $\{x_i\}_{i=1}^n$. This value had been collapsed to zero without this soft expectation definition.

Next, we define the soft variance (denoted by Vs), related to X conditionally being within union of $\{x_i\}_{i=1}^n$ and $\{(a_j,b_j)\}_{j=1}^m$. Using the nullity of $\bar{0}$ (**Axiom 3**) and differentiation property (B.10), we have

$$\begin{aligned}&\mathrm{Vs}(X|X\in\{x_i\}_{i=1}^n\cup\{(a_j,b_j)\}_{j=1}^m)=\\&\sum_{i=1}^n[\nu_X\bar{0}\dot{+}(\kappa_X-x_i)]^2f_X(x_i)\cdot\bar{0}\\&\dot{+}\sum_{j=1}^m\int_{a_j}^{b_j}[\nu_x\bar{0}\dot{+}(\kappa_X-x)]^2f_X(x)dx=\\&\left[\sum_{i=1}^n(\kappa_X-x_i)^2f_X(x_i)+2\nu_X\sum_{j=1}^m\int_{a_j}^{b_j}(\kappa_X-x)f_X(x)dx\right]\bar{0}\\&\dot{+}\sum_{j=1}^m\int_{a_j}^{b_j}(\kappa_X-x)^2f_X(x)dx.\end{aligned} \tag{55}$$

We would like to simplify last equation, especially the soft part. Denote

$$\gamma_{1_X}^2 = \sum_{i=1}^{n}(\kappa_X - x_i)^2 f_X(x_i) \geq 0,$$

$$\gamma_{2_X} = \sum_{j=1}^{m}\int_{a_j}^{b_j}(\kappa_X - x) f_X(x)dx$$

and

$$\lambda_X^2 = \sum_{j=1}^{m}\int_{a_j}^{b_j}(\kappa_X - x)^2 f_X(x)dx \geq 0.$$

By the linearity of the integral, we can simplify s_{2_X} as follows:

$$\begin{aligned}\gamma_{2_X} &= \kappa_X \sum_{j=1}^{m}\int_{a_j}^{b_j} f_X(x)dx - \sum_{j=1}^{m}\int_{a_j}^{b_j} x f_X(x)dx \\ &= \kappa_X \sum_{j=1}^{m}[F_X(b_j) - F_X(a_j)] - \kappa_X \\ &= -\kappa_X \left[1 - \sum_{j=1}^{m}[F_X(b_j) - F_X(a_j)]\right]\end{aligned}$$

Observe that $1 - \sum_{j=1}^{m}[F_X(b_j) - F_X(a_j) > 0$. Now we can simplify the definition for soft variance in (55) as follows

$$\begin{aligned}&\mathrm{Vs}(X|X \in \{x_i\}_{i=1}^n \cup \{(a_j,b_j)\}_{j=1}^m) = \left[\sum_{i=1}^{n}(\kappa_X - x_i)^2 f_X(x_i)\right. \\ &\left. - 2\nu_X\kappa_X\left\{1 - \sum_{j=1}^{m}[F_X(b_j) - F_X(a_j)]\right\}\right]\bar{0} \\ &\dot{+}\sum_{j=1}^{m}\int_{a_j}^{b_j}(\kappa_X - x)^2 f_X(x)dx \\ &= [\gamma_{1_X}^2 - 2\nu_X\gamma_{2_X}]\bar{0}\dot{+}\lambda_X^2 \\ &= \gamma_X\bar{0}\dot{+}\lambda_X^2.\end{aligned} \tag{56}$$

The real part λ_X^2 is non-negative (equals zero iif $x \equiv \kappa_X, \forall x \in (a_j, b_j), j = 1, 2..., m$ e.g., X is deterministic), which makes sense in terms of the original definition for variance. However, in the soft part $\gamma_X = \gamma_{1_X}^2 - 2\nu_X\gamma_{2_X}$ we some interesting phenomena: On one hand, we have a non-negative term $\gamma_{1_X}^2$ (equals zero iif $x_i \equiv \kappa_X, \forall i = 1, 2...n$). On the other hand the sign of the term $-2\nu_X\gamma_{2_X}$ in the linear combination of s_X depends on the sign of ν_X and the sign of γ_{2_X} (that depends on the sign of κ_X), so that potentially the soft part may have a negative sign. Eventually, the soft part adds more information regarding to the variance of the random variable. Applications of soft variance's with negative soft max is required to be checked.

In the next subsection, we continue to define a soft entropy, inspired by the notions in this subsections.

B. Soft Entropy

Recall for the definition of the Entropy of a discrete random variable X with support S_X and a point mass function (PMF) p_X (see e.g., [4]) is defined by

$$\mathrm{H}(X) = -\mathrm{E}(\log(p_X(X))) = -\sum_{x \in S_X}\Pr(X = x)\log(\Pr(X = x)). \tag{57}$$

For a continuous case (usually referred as *differential entropy*) for a continuous random variable $X \sim f_X$

$$\mathrm{H}(X) = -\mathrm{E}(\log(f_X(X))) = -\int_{S_X} f_X(x)\log(f_X(x))dx, \tag{58}$$

where (in both definitions) the base of the logarithm operator can be chosen to be appropriate positive real number e.g. 2, e, 10 etc. depends on the application. In this work, we do not emphasize a specific base, but we consider like previously a continuous random variable X conditioned by being within the union of $\{x_i\}_{i=1}^n$ and $\{(a_j, b_j)\}_{j=1}^m$. With these definitions, we have the following definition for a soft entropy $\mathrm{Hs}(\cdot)$:

$$\begin{aligned}&\mathrm{Hs}(X|X \in \{x_i\}_{i=1}^n \cup \{(a_j,b_j)\}_{j=1}^m) \\ &= \sum_{i=1}^{n} -\mathrm{Ps}(X = x_i)\log(\mathrm{Ps}(X = x_i)) \\ &\dot{+}\sum_{j=1}^{m}\int_{a_j}^{b_j} -f_X(x)\log(f_X(x))dx \\ &= \sum_{i=1}^{n}[-f_X(x_i)\cdot\bar{0}][\log(f_X(x_i)\cdot\bar{0})] \\ &\dot{+}\sum_{j=1}^{m}\int_{a_j}^{b_j} -f_X(x)\log(f_X(x))dx \\ &= \left[\sum_{i=1}^{n} -f_X(x_i)\log(f_X(x_i))\right]\cdot\bar{0} \\ &\dot{+}\left[\sum_{i=1}^{n} -f_X(x_i))\right]\cdot[\bar{0}\log(\bar{0})] \\ &\dot{+}\sum_{j=1}^{m}\int_{a_j}^{b_j} -f_X(x)\log(f_X(x))dx \\ &= h_1\cdot\bar{0}\dot{+}h_2\cdot[\bar{0}\log(\bar{0})]\dot{+}h_3\cdot 1.\end{aligned} \tag{59}$$

The soft entropy of X is a linear combination of the objects of $\bar{0}$, $\bar{0}\log(\bar{0})$ and 1. The question is how to evaluate the object $\bar{0}\log(\bar{0})$. One option would be an absolute zero i.e., $\bar{0}\log(\bar{0})$=0

Observation:

$$\lim_{x\to 0^+} x^x = 1 \Rightarrow \lim_{x\to 0^+} x\log(x) = 0$$

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

$$1 = e^0 = \sum_{n=0}^{\infty}\frac{0^n}{n!} = \frac{0^0}{0!}+(\frac{0^1}{1!}+\frac{0^2}{2!}+...) = \frac{0^0}{0!}+(0) = \frac{0^0}{1} = 0^0.$$

Another option would be defined it as a new type of a "soft zero" object e.g., a new axis that is a continuum of multiples of $\bar{\bar{0}} = \bar{0}\log(\bar{0})$, with nullity rule $\bar{\bar{0}}^2 = 0$. With the second option, we have additional information on the Entropy of X, not only via the $\bar{0}$-axis but also via a new potential axis, $\bar{\bar{0}} = \bar{0}\log(\bar{0})$.

Similarly, we define the *soft cross entropy*, by evaluation of the soft expectation of $\log(\hat{f}_X(X))$ for some "guested" PDF (e.g. we assume incorrectly that $X \sim \hat{f}_X$) by the following

$$\begin{aligned}
&\mathrm{H}(f_X, \hat{f}_X | X \in \{x_i\}_{i=1}^n \cup \{(a_j, b_j)\}_{j=1}^m) \\
&= \left[\sum_{i=1}^{n} -f_X(x_i)\log(\hat{f}_X(x_i))\right] \cdot \bar{0} \\
&\dot{+} \left[\sum_{i=1}^{n} -f_X(x_i))\right] \cdot [\bar{0}\log(\bar{0})] \\
&\dot{+} \sum_{j=1}^{m} \int_{a_j}^{b_j} -f_X(x)\log(\hat{f}_X(x))dx.
\end{aligned} \tag{60}$$

We can notice that the term which multiplies the object $\bar{0}\log(\bar{0})$ does not depend on the guested PDF $\hat{f}_X$. Moreover, this object is identical to the coefficient of $\bar{0}\log(\bar{0})$ in the soft entropy definition (59). The *Kullback–Leibler Divergence* (KLD) is defined by $\mathrm{D}(f_X||\hat{f}_X) = \mathrm{H}(f_X, \hat{f}_X) - \mathrm{H}(X) = \mathrm{E}(\log\frac{f_X(X)}{\hat{f}_X(X)})$. Either by subtracting (59) from (60) or by evaluating soft expectation of $\log\frac{f_X(X)}{\hat{f}_X(X)}$, we define the *soft KLD* $\mathrm{Ds}(\cdot)$ by the following

$$\begin{aligned}
&\mathrm{Ds}(f_X||\hat{f}_X | X \in \{x_i\}_{i=1}^n \cup \{(a_j, b_j)\}_{j=1}^m) \\
&= \left[\sum_{i=1}^{n} f_X(x_i)\log\frac{f_X(x_i)}{\hat{f}_X(x_i)}\right] \cdot \bar{0} \\
&\dot{+} \sum_{j=1}^{m} \int_{a_j}^{b_j} f_X(x)\log\frac{f_X(x)}{\hat{f}_X(x)}dx,
\end{aligned} \tag{61}$$

that has no multiple of $\bar{0}\log(\bar{0})$ term. We can see easily that that the multiple of $\bar{\bar{0}} = \bar{0}\log(\bar{0})$ term is cancel out via subtracting (59) from (60). Another explanation for it is by observing that expectation of $\log\frac{f_X(X)}{\hat{f}_X(X)}$ consists of terms with the form $\log\frac{\mathrm{Ps}(X=x_i)}{\hat{\mathrm{Ps}}(X=x_i)}$, where $\hat{\mathrm{Ps}}(X = x_i) = \hat{f}_X(x_i) \cdot \bar{0}$ is the soft probability corresponding to the guested PDF $\hat{f}_X$. It is convenient to the perform the following cancellation of $\bar{0}$ object:

$$\log\frac{\mathrm{Ps}(X = x_i)}{\hat{\mathrm{Ps}}(X = x_i)} = \log\frac{f_X(x_i)\cdot\bar{0}}{\hat{f}_X(x_i)\cdot\bar{0}} = \log\frac{f_X(x_i)}{\hat{f}_X(x_i)},$$

so that the multiple of $\bar{0}\log(\bar{0})$ term vanishes.

In the next section, we define a soft Mutual Information, as a splitting criteria for decision trees.

V. Decision Trees Based on Soft Mutual Information

Decision trees (e.g. Id3, C4.5, J48 etc.) are simple yet successful techniques for predicting and explaining the relationship between some measurements about an item and its target value (see e.g., [5] and [6]). In most decision trees inducers, discrete splitting functions (also known as *Splitting Criteria*) are univariate, i.e. an internal node is split according to the value of a single attribute. Consequently, the inducer searches for the best attribute upon which to perform the split. A *Splitting Criteria* of a random variable X (represents the features) and a random variable Y (represents the labels) has the following structure:

$$SplittingCriteria(Y; X) = C(Y) - C(Y|X), \tag{62}$$

where $C(\cdot)$ is an expectation of some cost function. In the case when the cost function is the entropy [i.e. $C(\cdot) = \mathrm{H}(\cdot)$], we refer the splitting criteria as an *Information Gain*, that is a *Mutual Information* between X and Y, denoted by

$$\begin{aligned}
\mathrm{I}(Y; X) &= \mathrm{H}(Y) - \mathrm{H}(Y|X) \\
&= \mathrm{H}(Y) + \mathrm{H}(X) - \mathrm{H}(Y, X),
\end{aligned} \tag{63}$$

which also can be written as a KLD between the joint PDF $f_{X,Y}$ and the PDF product $f_X f_Y$ i.e.,

$$\mathrm{I}(Y; X) = \mathrm{D}(f_{X,Y}||f_X f_Y). \tag{64}$$

In this section we present the mutual information, an example of a splitting criteria, as a soft number,based on a joint PDF (and its related marginal PDFs) of two random continuous variable, but with a data set consist of single values and interval. From here the decision algorithm is clear. Suppose that X and Y are continuous random variables, such that X is within the union of $\{x_i\}_{i=1}^n$ and $\{(a_j, b_j)\}_{j=1}^m$, and Y is within the union of $\{y_i\}_{i=1}^N$ and $\{(A_j, B_j)\}_{j=1}^M$ (recall that the singles point e.g. $\{x_i\}_{i=1}^n$ and the intervals $\{(a_j, b_j)\}_{j=1}^m$ are disjoint). for simplicity denote the following sets:

$$\begin{aligned}
\mathcal{X} &= \{x_i\}_{i=1}^n \cup \{(a_i, b_i)\}_{i=1}^m \\
\mathcal{Y} &= \{y_j\}_{j=1}^N \cup \{(A_j, B_j)\}_{j=1}^M.
\end{aligned} \tag{65}$$

Using (43), (61), (64) and (65), we can define a *soft Mutual Information* $\mathrm{Is}(\cdot)$ by the following equation after re-indexing

$$\begin{aligned}
&\mathrm{Is}(Y; X | Y \in \mathcal{Y}, X \in \mathcal{X}) = \\
&\mathrm{Ds}(f_{X,Y}||f_X f_Y | Y \in \mathcal{Y}, X \in \mathcal{X}) = \\
&\left[\sum_{j=1}^{N}\sum_{i=1}^{n} f_{X,Y}(x_i, y_j)\log\left(\frac{f_{X,Y}(x_i, y_j)}{f_X(x_i)f_Y(y_j)}\right)\right] \cdot \bar{0} \\
&\dot{+} \sum_{j=1}^{M}\sum_{i=1}^{m} \int_{A_j}^{B_j}\int_{a_i}^{b_i} f_{X,Y}(x, y)\log\left(\frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)}\right)dxdy,
\end{aligned} \tag{66}$$

So we have an example for a splitting criteria as a soft number, that can be used in decision trees algorithms in a case of data set consist of singles values and interval.

The definition for a soft Mutual Information in (66) is symmetric in X and Y [due to $\mathrm{I}(X; Y) = \mathrm{I}(Y; X)$ in the regular sense]. We can present (66) in a less symmetric form, using the Bayes Law identity $f_{X,Y} = f_{Y|X} f_X$, so that we have

$$\begin{aligned}&\mathrm{Is}(Y;X|Y\in\mathcal{Y},X\in\mathcal{X})=\\&\left[\sum_{j=1}^{N}\sum_{i=1}^{n}f_{Y|X}(y_j|x_i)f_X(x_i)\log\left(\frac{f_{Y|X}(y_j|x_i)}{f_Y(y_j)}\right)\right]\cdot\bar{0}\\&\dot{+}\sum_{j=1}^{M}\sum_{i=1}^{m}\int_{A_j}^{B_j}\int_{a_i}^{b_i}f_{Y|X}(y|x)f_X(x)\log\left(\frac{f_{Y|X}(y|x)}{f_Y(y)}\right)dxdy.\end{aligned}\tag{67}$$

This representation is applicable e.g., for emphasizing X as an input and Y as an output to some channel. An example is shown in the next subsection for a Gaussian case.

A. Gaussian Distribution Example

Consider the jointly Gaussian distributed variables X and Y as follows:

$$\begin{aligned}X\sim N(0,1), f_X(x)&=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\\Y\sim N(0,2), f_Y(y)&=\frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{2\cdot 2}y^2}\\(Y|X=x)\sim N(x,1), f_{Y|X}(y|x)&=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-x)^2}\end{aligned}\tag{68}$$

Remark 1: The above setup can be obtained by adding an uncorrelated Gaussian noise $W\sim N(0,1)$ to the Gaussian input X such that $X\perp\!\!\!\perp W$ and we have $Y=X+W$. A sketch of the proof is shown below

$$\begin{aligned}\mathrm{E}(Y)&=\mathrm{E}(X)+\mathrm{E}(W)\\&=0+0\\&=0,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y)&\overset{X\perp\!\!\!\perp W}{=}\mathrm{Var}(X)+\mathrm{Var}(W)\\&=1+1\\&=2,\end{aligned}$$

$$\begin{aligned}\mathrm{E}(Y|X=x)&=\mathrm{E}(X|X=x)+\mathrm{E}(W|=x)\\&\overset{X\perp\!\!\!\perp W}{=}x+0\\&=x,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y|X=x)&=\mathrm{Var}(Y-x|X=x)\\&=\mathrm{Var}(W|X=x)\\&\overset{X\perp\!\!\!\perp W}{=}1.\end{aligned}$$

Consider a simple case that each set $\mathcal{X}$ (input set) $\mathcal{Y}$ (output set) have one open interval and one single point

$$\begin{aligned}\mathcal{X}&=(a,b)\cup\{x_0\},\\\mathcal{Y}&=(A,B)\cup\{y_0\},\end{aligned}\tag{69}$$

so that the soft Mutual information in (67) is given by

$$\begin{aligned}&\mathrm{Is}(Y;X|Y\in\mathcal{Y},X\in\mathcal{X})=\\&\left[f_{Y|X}(y_0|x_0)f_X(x_0)\log\left(\frac{f_{Y|X}(y_0|x_0)}{f_Y(y_0)}\right)\right]\cdot\bar{0}\\&\dot{+}\int_A^B\int_a^b f_{Y|X}(y|x)f_X(x)\log\left(\frac{f_{Y|X}(y|x)}{f_Y(y)}\right)dxdy.\end{aligned}\tag{70}$$

and after plugging the Gaussian PDFs [according to (68)], we have

$$\begin{aligned}&\mathrm{Is}(Y;X|Y\in\mathcal{Y},X\in\mathcal{X})=\\&\left[\frac{1}{2\pi}e^{-\frac{1}{2}(y_0-x_0)^2}e^{-\frac{1}{2}x_0^2}\log\left(\sqrt{2}e^{\frac{1}{2\cdot 2}y_0^2}e^{-\frac{1}{2}(y_0-x_0)^2}\right)\right]\cdot\bar{0}\\&\dot{+}\int_A^B\int_a^b\frac{1}{2\pi}e^{-\frac{1}{2}(y-x)^2}e^{-\frac{1}{2}x^2}\log\left(\sqrt{2}e^{\frac{1}{2\cdot 2}y^2}e^{-\frac{1}{2}(y-x)^2}\right)dxdy.\end{aligned}\tag{71}$$

At the following Table I, we obtain some numerical results for a Soft Mutual Information (denoted by $\mathrm{Is}(Y;X)$ for simplicity) in our Gaussian Case. We used logarithm with base e:

Table I
NUMERICAL RESULTS OF A SOFT MUTUAL INFORMATION IN THE GAUSSIAN CASE

x_0	y_0	(a,b)	(A,B)	$\mathrm{Is}(Y;X)$
0	0	(1,2)	(1,2)	0.055159·$\bar{0}$ $\dot{+}$ 0.042381
0	1	(1,2)	(2,3)	0.0093225·$\bar{0}$ $\dot{+}$ 0.037941
1	0	(2,3)	(1,3)	-0.0089831·$\bar{0}$ $\dot{+}$ 0.018353
1	0	(20,30)	(10,30)	-0.0089831·$\bar{0}$ $\dot{+}$ 2.7404E-87
20	30	(2,3)	(1,3)	7.4494E-108·$\bar{0}$ $\dot{+}$ 0.018353

We can observe that, on one hand, when x_0 and y_0 are far away from the mean of X and Y (zero for both in our case), the contribution of the soft mutual information in its soft part approaches to zero. On the other hand, when the intervals (a,b) and (A,B) are away from the mean of X and Y, the contribution of the soft mutual information in its real part approaches to zero, so the soft part may have a significant value for taking a decision in a soft decision tree.

To summarize this example, we consider a case of two jointly Gaussian variables. We generate a formula for a Soft Mutual Information in a simple case of when each random variable's datum consists of one single point (to generate the soft part of the Soft Mutual Information) and one interval (to generate the real part of the Soft Mutual Information). This example can be generalized by summing the contributions of the Soft Mutual Information of any set of disjoints singles points and intervals.

VI. CONCLUSIONS

In the classical probability, in continuous random variables there is no distinguishing between the probability involving strict inequality and non-strict inequality. Moreover, a probability involve equality collapse to zero, without distinguishing among the values that we would like that the random variable will have for comparison. Soft numbers assist us to distinguish

between the probability involving strict inequality and non-strict inequality, and among the values that we would like that the random variable, by generating soft zeros multiples of the PDF observations.

In addition, we extended this notion of soft probabilities to the classical definitions of Complements, Unions, Intersections and Conditional probabilities under Kolmogorov definition and Bayes theorem, that makes sense with a probability of a continuous variable to be equal to an exact value does not collapse completely to zero.

We also extend the notion of soft probabilities to the expectation, variance and entropy of a continuous random variable, condition being in a union of disjoint intervals and a discrete set of numbers. with this extension, we have some information regarding to the expectation, variance and entropy of a continuous random variable being within discrete sent of numbers, but not collapse completely to zero. In addition we discover some interesting properties regarding to soft variance and soft entropy that required to be explored. In soft variance, the soft part might be a negative number. In the soft entropy, we have potentially a new zero axis with multiples of $\bar{0}\log(\bar{0})$, or alternatively we may defined $\bar{0}\log(\bar{0})$ as an absolute zero. For the first option (considering new zero axis) it may be required to define additional bridging notation in order to bridge between multiples of $\bar{0}\log(\bar{0})$ and the multiples of 0 and 1. We extended the notion of soft entropy into the definition of Cross Entropy and KLD, and we found that a soft KLD is a soft number, that does not have a multiple of $0 \cdot \log 0$. More exploration are required to be done in order to realize the consequences of this result. Based on a soft KLD, we defined a soft mutual information, that can be used as a splitting criteria in decision trees with data set of continuous random variables, consist of single samples and intervals.

VII. Suggestions for Future Research

We suggest to extend the notion of soft probability covered in this work by generalizing to the followings: continuous random vectors, mixed random variable (that has continuous and discrete distribution i.e., non piecewise constant CDF but with discontinuity), random vector with discrete, continuous and mixed random variables etc.

In addition we suggest to explore the applications of negative soft part in the soft variances, and to explore the applications of multiples of $\bar{0}\log(\bar{0})$ as an information to the soft entropy (in addition to the additional information regarding to the multiples of multiples of $\bar{0}$ in the soft entropy.). We also suggest to explore the soft probability in additional topics in Information Theory and Machine Learning.

Acknowledgment

This paper was supported by the Koret Foundation grant for Smart Cities and Digital Living 2030 bestowed upon the universities of Stanford and Tel Aviv. We are happy to express our thankfulness and gratitude for this support.

Appendix A
Probability Theory Brief Review

Probability theory is used in order to model processes and phenomenons, involving randomness of the parameters and variables. Usually, when we want to quantify a probability of an event in these processes or phenomenons, we evaluate the probability of this event by the range [0,1], e.g., '0' means the event can never (almost surely) occur and 1 means the event can always (almost surely) occur. For this quantification, a probability space is a defined by mathematical triplet $(\Omega, \mathcal{F}, P)$ defined as follows:

- Sample Space Ω: Set of all possible outcomes. An outcome is the result of a single execution of the model.
- σ-algebra $\mathcal{F}$: Collection of all the events we would like to consider. An event is a set outcomes.
- Probability Measure P: Function returning an event's probability. P maps from the σ-algebra $\mathcal{F}$ to the interval [0,1].

Random variables are used to provide outcomes numerical values. The mathematical notation for a random variable X is defined by the following:

$$X : \Omega \to S_X \tag{A.1}$$

where S_X is a set of real numbers, that the random variable S_X can have. S_X is called the *support of* X. Mainly, we distinct between two types of Random variables:

- Discrete random variables, that can have finite or countable of values; and
- Continuous random variables that can have uncountable of values.

For both discrete and continuous random variables, a cumulative distribution function (CDF) of a random variable X as follows:

$$F_X(x) = \Pr(X \leq x), \tag{A.2}$$

where the right hand side (RHS) asked what is the probability of a random variable X to be less or equal to some real number x, and the left hand side (LHS) provides the answer in terms of x by the function $F_X : \mathbb{R} \to [0, 1]$.

In a case of a discrete random variable, we can address to the question, what is the probability of a random variable X to be equal to some real number x, by the point mass function (PMF), defined as follows:

$$p_X(x) = \Pr(X = x). \tag{A.3}$$

The cumulative property is obtained by the following relation between the CDF and the PMF in the discrete case

$$F_X(b) - F_X(a) = \sum_{a < x_i \leq b} p_X(x_i) = \Pr(a < X \leq b). \tag{A.4}$$

In a case of a continuous random variable, a probability density function (PDF) is defined by:

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{A.5}$$

The function $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ denotes the PDF of X. The cumulative property is obtained by the following relation between the CDF and the PDF in the continuous case

$$F_X(b) - F_X(a) = \int_a^b f_X(x)dx = \Pr(a < X \leq b). \quad \text{(A.6)}$$

The PDF can be used is to approximate the probability of the continuous random variable X to be adjacent to x in the following sense

$$\Pr(x < X \leq x + \Delta x) \approx f_X(x)\Delta x, \quad \text{(A.7)}$$

where $\Delta x > 0$ is a small value, that defines how much this probability is accurate. However, continuous random variables have the following properties:

- No distinguishing between strict inequality and non-strict in equality e.g., $\Pr(X \leq x) = \Pr(X < x)$;
- Equality collapses to zero i.e., $\Pr(X = x) = 0$. Although any value of $x \in S_X$ is possible for X, the the probability of X to be equal to any value of $x \in S_X$ is (almost surely) zero.

Because of these properties, we lose some information regarding to a continuous random variable to have an exact value.

In this work, we introduce the *Soft Numbers* (see Klein and Maimon's papers e.g., [1], [2] and [3]) to give a probability interpretation of a continuous random variable to have an exact value, that provides distinguishing between strict inequality and non-strict in equality in the probability function.

APPENDIX B
PRESENTATION OF SOFT NUMBERS

According to traditional mathematics, the expression 0/0 is undefined, although in fact the whole set of real numbers could represent this expression, since $a \cdot 0 = 0$ for all real numbers a. This observation opens a new area for investigation, which is a part of what it is called in [1] a "Soft Logic", that refers to a new axis, "a continuum of multiples of zeros", with distinction between a positive zero "+0" and a negative zero "-0" (see also [2] and [3]).

A. Soft Number: Definitions and Axioms

A new object $\bar{0}$ is symbolized in order to generate of a continuum of multiples of zeros $a\bar{0}$ on a "$\bar{0}$" axis, where a is a real number. An object $a\bar{0}$ denotes "soft zero", while the object $\mathbf{0} = 0 \cdot \bar{0}$ denotes "absolute zero". The object $\bar{1}$ denotes the real axis (i.e., contains multuples of "ones", $b\bar{1}$), and parallel to the "$\bar{0}$" axis. For simplicity, the symbol $\bar{1}$ is omitted during computations.The following axioms and definitions are developed for soft zeros for all real numbers a and b:

Axiom 1 (Distinction): $a \neq b \Rightarrow a\bar{0} \neq b\bar{0}$.

Definition 1 (Order): $a < b \Rightarrow a\bar{0} < b\bar{0}$.

Axiom 2 (Addition): $a\bar{0} + b\bar{0} = (a + b)\bar{0}$.

Axiom 3 (Nullity): $a\bar{0} \cdot b\bar{0} = 0$, i.e., soft numbers "collapse" to zero under multiplications.

Axiom 4 (Bridging): There exists a bridge between a zero axis, and a real axis and vice versa, denoted by a pair of a bridge number and its mirror image about the bridge sign. Bridge numbers of a right type

$$b\bar{1} \perp a\bar{0}$$

and bridge numbers of a left type

$$a\bar{0} \perp b\bar{1}.$$

Axiom 5 (Non-commutativity): Bridging operator $\perp$ does not commute [3] i.e.,

$$b\bar{1} \perp a\bar{0} \neq a\bar{0} \perp b\bar{1}.$$

Definition 2 (Soft Number): A soft number is defined as a set of the of bridge numbers pair of opposite types but with the same components – the same zero axis number $a\bar{0}$ and the same real number b:

$$a\bar{0}\dot{+}b = \{a\bar{0} \perp b; b \perp a\bar{0}\}$$

We denote the set of all bridge numbers by **BN** and all soft numbers by **SN**. The coordinate system of Soft Logic is constructed, as presented in Figure 1. It starts from 0 to 1 horizontally and then it turns 90° from 1 to infinity

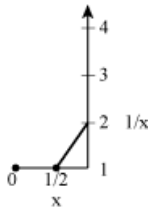

Figure 1. The Soft coordinate axis

Remark 2: There exists a one-to-one correspondence between the segment $(0, 1]$ and the segment $[1, \infty)$.

Remark 3: All lines that connect x to $1/x$ (for all non-zero real x) intersect at a single point.

The statements in Remarks 2 and 3 were demonstrated in [1]. This "single point" denotes the beginning of the soft logic coordinate system. We call this point "the absolute zero". The distance from absolute zero to +0 is 1. An extension of this new coordinate system to the negative numbers is implemented in Figure 2.

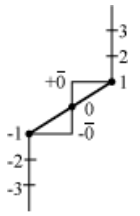

Figure 2. Distinction between -0 and +0

In Figure 2 we have, in addition to the absolute zero **0**, two additional zeros. One zero is opposite the number -1, and is not identical with the zero opposite to the number $+1$. Hence,

we suggest denoting these two different "zeros" as $+\bar{\mathbf{0}}$ and $-\bar{\mathbf{0}}$.

Figure 3 shows the extended coordinate system for positive and negative numbers with an additional line presenting the multiples of zero. The added line is called a zero line or a zero axis, and the multiples on it are called soft zeros or zero axis numbers.

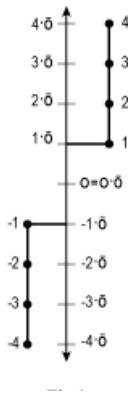

Figure 3. The extended soft coordinate system

The coordinate system in Figure 3 allows us to present all the real numbers and all the soft zeros. We now wish to construct a coordinate system for representing various Soft Numbers, which may be described as an infinite strip as shown in Figure 4. Because of the Soft Number duality, we double the strip (Figure 4). This allows us to represent both elements of a Soft Number:

$$\begin{aligned} c &= x\bar{0} \perp y, \\ c' &= y \perp x\bar{0}, \end{aligned} \tag{B.1}$$

where x and y are real numbers. Each of the elements c and c' is a mirror image of the other about the bridge sign. Note that we have expanded the coordinate system in Figure 3 to the one shown in Figure 4.

As the infinite strip, presented (partially) in Figure 4, is intended for the presentation of Soft Numbers, we call it a *'Soft Numbers Strip'* or briefly, SNS.

Definition 3 (height and width of a point on an SNS): let C be any point on the SNS.

- The **height of the point** C is the vertical distance from C to the horizontal segment with the absolute zero at its center. This distance is supplied with a plus sign if C is above this segment and with a minus sign if C is below it. The height with a sign is denoted by A.
- The **width of the point** C is the horizontal distance from C to the zero line and is denoted by B.

The definitions above provide every point C on the SNS with two parameters, $A \in \mathbb{R}$ and $B \in [0, 1]$. The condition $A > 0$ is satisfied in the positive part of the SNS, and $A < 0$ - in its negative part, or correspondingly, above and below the horizontal segment containing the absolute zero, while on this segment $A = 0$. For the second parameter B there is: $B = 0$ on the zero axis, $B = 1$ on the lines bounding the SNS, and otherwise $0 < B < 1$.

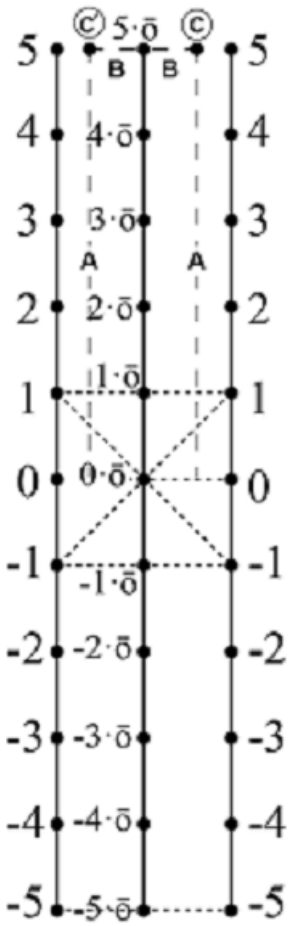

Figure 4. The complete soft coordinate system

If two points c and c' on the SNS are symmetric about the zero axis, they have the same height A and the same width B, i.e., we can symmetrically represent them by the following **BN**s:

$$\begin{aligned} c &= (1 - B)A\bar{0} \perp BA, \\ c' &= BA \perp (1 - B)A\bar{0}. \end{aligned} \tag{B.2}$$

Therefore, to define a presentation of soft numbers $x\bar{0}\dot{+}y$ by symmetric pairs (**SP**s) of points on the SNS, we have to define a correspondence between these numbers and the pairs of real numbers $(A, B) \in \mathbb{R} \times [0, 1]$ (denoted as **SP**), so that

$$\begin{aligned} x\bar{0}\dot{+}y &= \{c, c'\} \\ &= (1 - B)A\bar{0}\dot{+}BA. \end{aligned} \tag{B.3}$$

Hence, by a coefficients comparison of the real part and the soft part:

$$\begin{aligned} x &= (1 - B)A \\ y &= BA, \end{aligned} \tag{B.4}$$

or equivalently, after solving for the **SP**, (A, B)

$$\begin{aligned} A &= x + y \\ B &= \frac{y}{x+y}. \end{aligned} \tag{B.5}$$

In the next subsection, we outline some properties of mathematical operations and functions over the soft numbers.

B. Mathematical operations and Functions on Soft Numbers

In this section we outline some mathematical operations over the soft numbers. Suppose $a\bar{0}\dot{+}b, c\bar{0}\dot{+}d \in \mathbf{SN}$ are given soft numbers, then the following mathematical operations hold based on axioms 2 and 3:

- **Addition/subtraction:**

$$(a\bar{0}\dot{+}b) \pm (c\bar{0}\dot{+}d) = (a \pm c)\bar{0}\dot{+}(b \pm d); \tag{B.6}$$

- **Multiplication:**

$$(a\bar{0}\dot{+}b) \cdot (c\bar{0}\dot{+}d) = (ad + bc)\bar{0}\dot{+}bd; \tag{B.7}$$

- **Natural power:**

$$(a\bar{0}\dot{+}b)^n = nab^{n-1}\bar{0}\dot{+}b^n. \tag{B.8}$$

Based on the above equations, every polynomial $P_N(x)$ that operates on every soft number $\alpha\bar{0}\dot{+}x$ is given by

$$P_N(\alpha\bar{0}\dot{+}x) = \alpha P'_N(x)\bar{0}\dot{+}P_N(x). \tag{B.9}$$

where $P'_N(x)$ denotes the derivative of $P_N(x)$. This notion is generalized for analytic functions $f(x)$ so that

$$f(\alpha\bar{0}\dot{+}x) = \alpha f'(x)\bar{0}\dot{+}f(x). \tag{B.10}$$

REFERENCES

[1] M. Klein and O. Maimon, "Axioms of Soft Logic," *p-Adic Numbers, Ultrametric Analysis and Applications*, vol. 11, no. 3, pp. 205-215, 2019. https://doi.org/10.1134/S2070046619030038

[2] M. Klein and O. Maimon, "The Dynamics in the Soft Numbers Coordinate System," *Journal of Advances in Mathematics*, vol. 18, pp.1-17, 2020. https://doi.org/10.24297/jam.v18i.8531

[3] M. Klein and O. Maimon, "Fundamentals of Soft Logic", New Mathematics and Natural Computation, April 2021. https://doi.org/10.1142/S1793005721500356

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd ed.* Wiley, New-York, 2006

[5] O. Maimon and L. Rokach, *Data mining with decision trees: theory and applications (2nd edition).* Vol. 81. World Scientific, 2014

[6] L. Rokach and O. Maimon, "Decision Trees," in *Data Mining and Knowledge Discovery Handbook*, Rokach, L. and Maimon, O. (Eds.) Boston, MA: Springer, pp. 165–192 ,2005. https://doi.org/10.1007/0-387-25465-X_9