Soft Decision Trees

Oren Fivel[1], Moshe Klein[2] and Oded Maimon[2]

[1]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev,
Be'er Sheva, Israel
[2]Department of Industrial Engineering, Tel-Aviv University,
Tel-Aviv, Israel
fivel@post.bgu.ac.il, mosheklein@mail.tau.ac.il,
maimon@tauex.tau.ac.il

January 2022

Abstract

In this paper we develop the foundation of a new theory for decision trees based on new modeling of phenomena with soft numbers. Soft numbers represent the theory of soft logic that addresses the need to combine real processes and cognitive ones in the same framework. At the same time soft logic develops a new concept of modeling and dealing with uncertainty: the uncertainty of time and space. It is a language that can talk in two reference frames, and also suggest a way to combine them. 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). Based on Soft Probability, we introduced an approach to implement C4.5 algorithm as an example for a Soft Decision Tree.

Keywords— Bridge Number, Continuous Random Variable, Decision Trees, Information Theory, PDF, Probability, Soft Logic, Soft Number, Soft Probability, Zero Axis.

1 Introduction

In this paper we develop the foundation of a new theory for decision trees based on new modeling of phenomena with soft numbers. This calls for major concept change of probability, which is developed in this paper, so that decision trees can be modeled. Soft numbers represent the theory of soft logic that addresses the need to combine real processes and cognitive ones in the same framework. At the same time soft logic develops a new concept of modeling and dealing with uncertainty: the uncertainty of time and space. It is a language that can talk in two reference frames, and also suggest a way to combine them.

1.1 Research Motivation and Direction

Probability theory is used in order to model processes and phenomenons, involving randomness of the parameters and variables. A probability of a continuous 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 Eq. (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., [5], [6] and [7]).

In addition there is an approach to represent a discrete distribution as a continuous distribution by a linear combination of Dirac delta functions $\delta(x - x_i)$, or by any approximations of Dirac delta functions e.g., Gaussian functions (also known as *Gaussian Mixture Model* or GMM) or rectangular functions (based on a *"Uniformly" Mixture Model* or UMM) etc (see Eq. (14) for more details). Our approach it to establish the opposite in some sense, i.e., to represent a continuous random variable with a possibility to have a discrete values with probability that will not collapse absolutely to zero.

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. This probability interpretation is implemented by "Soft Probability" [3].

1.2 Organization of the Work

Section 2 incorporates Soft Numbers into probability theory to present the notion of "Soft Probability". Section 3 presents an example for application on Decision Trees based C4.5 algorithm. Conclusions and suggestion for future research are shown in sections 4 and 5 respectively to summarize this work. For completion, Appendix A provides an extension of Soft Probability for Complements, Union, Intersection and Conditional Probability. Appendix B provides a presentation of Soft Numbers.

2 Soft Probability: Incorporation of Soft Number into Probability Theory

In order to incorporate the notion of Eq. (B.5) in Appendix B, we define a cumulative distribution function (CDF) of a continuous random variable

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

where $\mathrm{Ps}(\cdot)$ is a suggested type of a probability function, dented as a "Soft Probability" [3] instead of a regular probability notation "$\mathrm{Pr}(\cdot)$" 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 $\mathrm{Ps}(X \leq x)$ for a continuous random variable X [i.e., $\mathrm{Ps}(X < x) \neq \mathrm{Ps}(X \leq x)$]. 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 Eq. (3) can be decomposed by Eq. (B.5) as follows

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

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

$$\begin{aligned} \mathrm{Pr}(X \leq x) &\stackrel{"X=x" \cap "X<x" = \emptyset}{=} \underbrace{\mathrm{Pr}(X = x)}_{=0} + \mathrm{Pr}(X < x) \\ &= \mathrm{Pr}(X < x), \end{aligned}$$

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

$$\mathrm{Ps}(X \leq x) \stackrel{\mathrm{def}}{=} \mathrm{Ps}(X = x) + \mathrm{Ps}(X < x), \tag{5}$$

so that we define the terms on the LHS as follows

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

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

By this setup we achieve a distinguishing between $\mathrm{Ps}(X \leq x)$ 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 implementation on mixture models PDFs, GMM and UMM, in order to demonstrate the effect of soft numbers (and more precisely, soft zeros) on PDFs.

2.1 Examples on Mixture Models

In order to demonstrate the effect of soft numbers (and more precisely, soft zeros) on PDFs, we provide two examples of implementation on mixture models:

- Gaussian Mixture Model (GMM); and
- Uniformly Mixture Model (UMM).

A PDF of a continuous random variable X with mixture model is given by

$$f_X\left(x; \{(\theta_i, w_i)\}_{i=1}^n\right) = \sum_{i=1}^n w_i f_i(x; \theta_i), \tag{8}$$

where $f_i(\cdot;\theta_i)$ is a PDF with a set of parameter θ_i, and $w_i \in [0, 1]$ is a weight that multiplies the i'th PDF f_i and sum to 1 i.e., $\sum_{i=1}^n w_i = 1$. In the GMM case the i'th PDF f_i is parameterized with mean μ_i and variance σ_i^2 such that

$$f_i(x; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(x-\mu_i)^2}. \tag{9}$$

In the UMM case the i'th PDF f_i is parameterized with the open interval (a_i, b_i) such that

$$f_i(x; a_i, b_i) = \frac{1}{b_i - a_i} \mathbb{1}_{x \in (a_i, b_i)}, \tag{10}$$

where $\mathbb{1}_A$ is the indication function, indicates '1' if 'A' is true and '0' if 'A' is false. The corresponding Soft probability is obtained by

$$\text{Ps}\left(X = x; \{(\theta_i, w_i)\}_{i=1}^n\right) = \sum_{i=1}^n w_i f_i(x; \theta_i) \cdot \bar{0}. \tag{11}$$

For illustration we chose following parameters for the mixture model with 2 components each and plot them in Figure 1:

- GMM: $(\mu_1, \sigma_1, w_1) = (0, 0.2, 0.33)$, $(\mu_2, \sigma_2, w_2) = (4, 1, 0.67)$;
- UMM: $((a_1, b_1), w_1) = ((-0.3, 0.3), 0.33)$, $((a_2, b_2), w_2) = ((2, 6), 0.67)$.

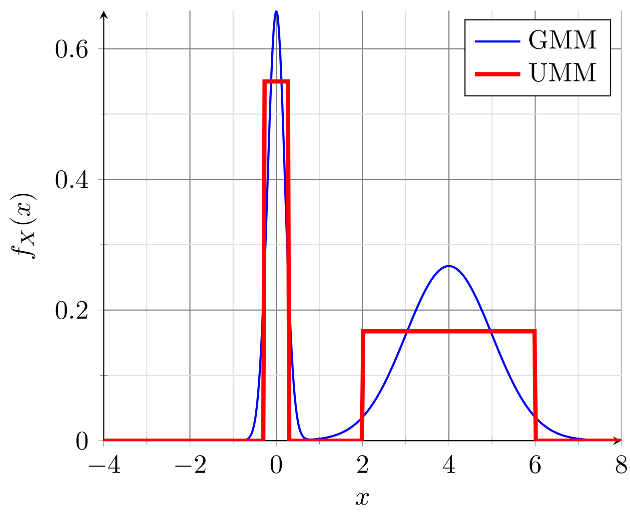

Figure 1: Examples of Gaussian Mixture Model and Uniformly Mixture Model's PDFs

In the GMM case, we have two local maximums at $x = \mu_1$ (a global maximum) and $x = \mu_2$. Hence, if we take three points $x_1 \approx \mu_1$, $x_2 \approx \mu_2$ and x_3 such that $|x_1|, |x_2| \ll |x_3|$, we have the following order of Soft Probabilities:

$$\text{Ps}(X = x_1) > \text{Ps}(X = x_2) > \text{Ps}(X = x_3) > 0 \cdot \bar{0}.$$

We have a strict inequality in the RHS, because the support of the GMM is infinite. The Soft Probability of the GMM presents an absolute low probability of X to have an exact value x but relative high probability when X is closer to μ_1 (where the GMM is maximal in this case). In other words, in contrast to classic probability that would collapse absolutely to have exact value $X = x$, with Soft probability we are able to distinguish (and order) among events $X = x$ that will have a positive soft probability (due to infinite support in GMM).

In the UMM case, if we take three point $x_1 \in (a_1, b_1)$ [global maximum], $x_2 \in (a_2, b_2)$ and $x_3 \notin (a_1, b_1) \cup (a_2, b_2)$, we have the following order of Soft Probabilities:

$$\mathrm{Ps}(X = x_1) > \mathrm{Ps}(X = x_2) > \mathrm{Ps}(X = x_3) = 0 \cdot \bar{0}.$$

We have equality to absolute zero in the RHS because x_3 is outside of the support $(a_1, b_1) \cup (a_b, b_2)$, while for $i = 1, 2$ we assign some probability value by the soft zero $\mathrm{Ps}(X = x_i) = \frac{w_i}{b_i - a_i} \cdot \bar{0} > 0 \cdot \bar{0}$. Hence, with soft probability, we succeed to distinguish between "zero probabilities" of "impossible" events ($X = x$ with x is out side of the support) and "possible" events ($X = x$ with x is in the support), that would collapse to zero in the classical probability sense, so that the "impossible" events will be absolute zero and the "possible" events will be some soft zero and ordered accordingly.

2.2 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 Eq. (B.5) with $|x|' = \mathrm{sign}(x)$, ignoring the fact that this derivative is not continuous, so that

$$|a\bar{0} \dot{+} x| = \alpha\bar{0} \cdot \mathrm{sign}(x) \dot{+} |x|. \tag{12}$$

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

$$\begin{aligned} |\alpha\bar{0} \dot{+} x| &= \sqrt{(\alpha\bar{0} \dot{+} x)((-\alpha)\bar{0} \dot{+} x)} \\ &= \sqrt{-(\alpha\bar{0})^2 + x^2} \\ &= \sqrt{-0 + x^2} \\ &= \sqrt{x^2} \\ &= |x|. \end{aligned} \tag{13}$$

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

1. $\mathrm{Ps}(X \leq x) \neq \mathrm{Ps}(X < x)$
   but $|\mathrm{Ps}(X \leq x)| = |\mathrm{Ps}(X < x)| > |\mathrm{Ps}(X = x)| = 0$;

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 literature (e.g., [8], [4] and [13]), there is an approach to represent a discrete distribution as a continuous distribution by a linear combination of Dirac delta functions $\delta(x - x_i)$, or by any approximations of Dirac delta functions based on a mixture model e.g., GMM or UMM. Suppose X is a discrete random variable with the probability $\Pr(X = x_i) = p_i$. Then X can be represented with a continuous distribution as follows

$$\begin{aligned} f_X(x) &= \sum_i p_i \delta(x - x_i) \\ &\approx \sum_i p_i \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-x_i)^2}, \sigma^2 \ll 1 \\ &\approx \sum_i p_i \cdot \frac{1}{2a} \mathbb{1}_{x-x_i \in (-a,a)}, a \ll 1. \end{aligned} \tag{14}$$

recall for Dirac delta function properties,

$$\begin{aligned} \delta(x) &= \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases} \\ \int_{-\infty}^{\infty} &\delta(x) dx = 1, \end{aligned} \tag{15}$$

and also $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2} \xrightarrow{\sigma^2 \to 0} \delta(x)$, $\frac{1}{2a} \mathbb{1}_{x \in (-a,a)} \xrightarrow{a \to 0} \delta(x)$, i.e., Gaussian distribution and uniformly distribution converge to Dirac delta function (degenerative distribution) when the variance of the Gaussian distribution and the length of the interval in the uniformly distribution approach to zero respectively (cf. Figure 1). Our approach it to establish the opposite in some sense, i.e., to represent a continuous random variable with a possibility to have a discrete values with probability that will not collapse absolutely to zero. See Appendix A for the extension of the soft probability's notion to complement, union intersection and conditional probability

In this work, we introduce the *Soft Numbers* (see Klein and Maimon's papers e.g., [5], [6] and [7]) 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.

In the next section, we use the notion of "Soft Probability" into Decision Trees.

3 Soft Decision Trees

3.1 Overview

Decision trees are simple yet successful techniques for predicting and explaining the relationship between some measurements about an item and its target value (see e.g., [12]-[11]). 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. There are various top-down decision trees inducers such as: **ID3** (Quinlan (1986) Ref. [10]), the basic algorithm; **C4.5** (Quinlan (1993) Ref. [9]), extension of ID3 which can handle continuous variables; **CART** (Breiman et al. (1984) Ref. [1]), Classification And Regression Tree; etc.

In the next subsection we introduce an implementation of *Decision Trees with Soft Numbers*, based on C4.5 Algorithm, in order to generate a *Soft Decision Trees.*

3.2 Soft Decision Trees, Based C4.5 Algorithm

In this subsection we introduce an implementation of *Decision Trees with Soft Numbers*, based on C4.5 Algorithm, in order to generate a *Soft Decision Trees*, based on a discrete label S and continuous features $X_0, X_1, ..., X_n$.

3.2.1 Preferring a Thesis in Equilibrium State

On each stage in the C4.5 algorithm we search the splitting which will minimize the classes entropy:

$$\min\{\mathrm{H}(S|X_0), \mathrm{H}(S|X_1), \mathrm{H}(S|X_2)....\mathrm{H}(S|X_n)\} \tag{16}$$

We will split our data according to the feature which has the most information gain (IG), in some cases when calculating the information gain we may find that several features have the same information gain, and in the case that the IG is the maximum of all feature's IG we will not be able to tell by which feature we should split our data.

For describing the problem we are trying to solve, we will define the following example:
Let $\{X_0, X_1...X_n\}$ be continuous features with known distributions $f_{X_i}(x)$, and S be a binary class such that $S \in \{s_1, s_2\}$. For finding the best splitting features we will calculate the IG for each feature and for the optimum threshold of that feature.

$$\min\{\mathrm{IG}(X_0 = x_{0_{\mathrm{th}}}), \mathrm{IG}(X_1 = x_{1_{\mathrm{th}}})...\mathrm{IG}(X_n = x_{n_{\mathrm{th}}}\}. \tag{17}$$

In some cases we can have an equality between two or more IG of different features, and in that case there is no preferred thesis, so we can pick one of the suggested features randomly, for example:

$$\mathrm{IG}(X_1 = x_{1_{\mathrm{th}}}) = \mathrm{IG}(X_4 = x_{4_{\mathrm{th}}}) = \mathrm{IG}(X_8 = x_{8_{\mathrm{th}}}).$$

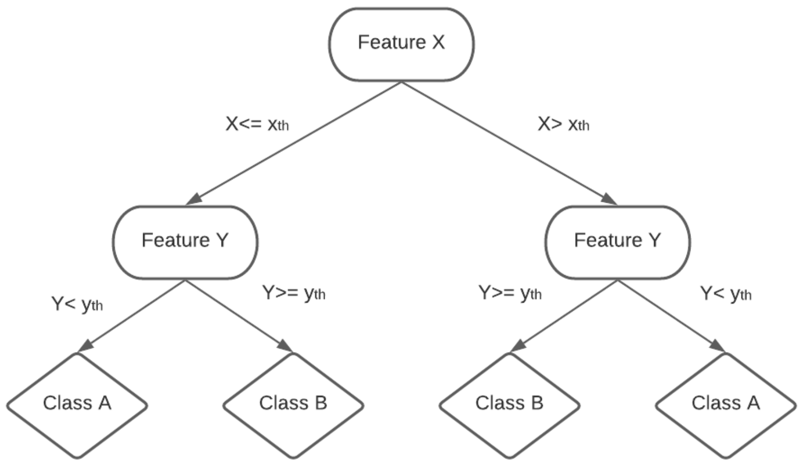

Figure 2: Classic Decision Tree splitting

Using the soft probability we suggest a method to derive the information gain of those features and also the ability to choose the preferred one. Recall the definition of information gain:

$$\begin{aligned}\mathrm{IG}(X_i) =&\Pr(X_i \geq x_i)\mathrm{H}(S|X_i \geq x_i) + \Pr(X_i < x_i)\mathrm{H}(S|X_i < x_i)\\ =&\Pr(X_i > x_i)\mathrm{H}(S|X_i > x_i) + \Pr(X < x_i)\mathrm{H}(S|X_i < x_i)\\ &+ \mathbf{Pr}(\boldsymbol{X_i = x_i})\mathbf{H}(\boldsymbol{S|X_i = x_i}).\end{aligned} \tag{18}$$

The part $\Pr(X_i = x_i)\mathrm{H}(S|X_i = x_i)$ is the IG on the threshold point. It is infinitesimal and treated as 0, so it has no impact on the total IG of each feature. $P(X_i = x) = 0$ in the classical probability sense, so in features equality we get the following in the case of $\mathrm{IG}(X_i) = \mathrm{IG}(X_j)$:

$$\begin{aligned}&\Pr(X_i > x_i)\mathrm{H}(S|X_i \geq x_i) + \Pr(X_i < x_i)\mathrm{H}(S|X_i < x)\\ =&\Pr(X_j > x_j)\mathrm{H}(S|X_j > x_j) + \mathrm{H}(X_j < x_j)\mathrm{H}(S|X_j < x_j).\end{aligned} \tag{19}$$

On the other hand using soft probability the part $\mathrm{Ps}(X_i = x)$ does have defined value by Eq. (6), $\mathrm{Ps}(X_i = x) = f_{X_i}(x) \cdot \bar{0}$, so we can express soft information gain as follows:

$$\begin{aligned}\mathrm{IGs}(X_i) =&\mathrm{Ps}(X_i = x_i) \cdot \mathrm{H}(S|X_i = x_i)\\ &+ \mathrm{Ps}(X_i > x_i)\mathrm{H}(S|X_i > x_i) + \mathrm{Ps}(X < x_i)\mathrm{H}(S|X_i < x_i)\\ =&f_{X_i}(x_i)\mathrm{Hs}(S|X_i = x_i) \cdot \bar{0}\\ &\dot{+}\Pr(X_i > x_i)\mathrm{H}(S|X_i > x_i) + \Pr(X < x_i)\mathrm{H}(S|X_i < x_i).\end{aligned} \tag{20}$$

When comparing the soft IGs we realize that we can distinguish among them by comparing the middle branch - threshold branch:

$$f_{X_i}(x_i)\mathrm{H}(S|X_i = x_i) \cdot \bar{0} \underset{<}{\overset{>}{?}} f_{X_j}(x_j)\mathrm{H}(S|X_j = x_j) \cdot \bar{0} \tag{21}$$

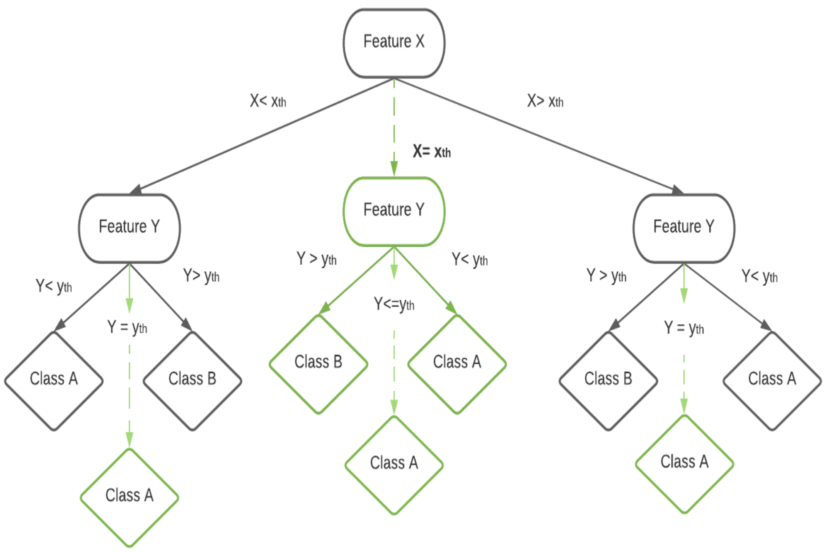

Figure 3: Soft Decision Tree splitting

To further realize the subject we will present a solution to the following problem in the next example.

3.2.2 Example: Uniformly Distribution Features and Discrete Label

Let $X \sim U(0,2)$ and $Y \sim U(0,4)$ be features of a data set and $S \in \{0,1\}$ be an event which we are trying to predict its value according to the features (X,Y). Finding the soft information gain of splitting according to each feature have the following result:

$$\begin{aligned} \mathrm{IG}(X) =& \mathrm{IG}(Y) \\ =& \Pr(X \leq x_{\mathrm{th}})\mathrm{H}(S|X \leq x_{\mathrm{th}}) + \Pr(X > x_{\mathrm{th}})\mathrm{H}(S|X > x_{\mathrm{th}}) \\ =& \Pr(Y \leq y_{\mathrm{th}})\mathrm{H}(S|Y \leq y_{\mathrm{th}}) + \Pr(Y > y_{\mathrm{th}})\mathrm{H}(S|Y > y_{\mathrm{th}}). \end{aligned} \tag{22}$$

With the classic approach we should choose one of the features X or Y randomly as they both got the same information gain. Using the soft decision approach we can try to identify which hypothesis is superior even if it is by a small margin, but it may translate to better overall decision making tool. First we will express the full equation explicitly separating the the threshold branch as follows:

$$\begin{aligned} &\Pr(X = x_{\mathrm{th}})\mathrm{H}(S|X = x_{\mathrm{th}}) + [\Pr(X < x_{\mathrm{th}})\mathrm{H}(S|X < x_{\mathrm{th}}) + \Pr(X > x_{\mathrm{th}})\mathrm{H}(S|X > x_{\mathrm{th}})] \\ &= \Pr(Y = y_{\mathrm{th}})\mathrm{H}(S|Y = y_{\mathrm{th}}) + [\mathrm{Ps}(Y < y_{\mathrm{th}})\mathrm{H}(S|Y < y_{\mathrm{th}}) + \Pr(Y > y_{\mathrm{th}})\mathrm{H}(S|Y > y_{\mathrm{th}})]. \end{aligned} \tag{23}$$

We can say as we've shown before that the last two parts of the equation [in square brackets] are equal and all it is left is to compare the threshold branch:

$$\Pr(X = x_{\mathrm{th}})\mathrm{H}(S|X = x_{\mathrm{th}}) \overset{>}{\underset{<}{=?}} \Pr(Y = y_{\mathrm{th}})\mathrm{H}(S|Y = y_{\mathrm{th}}). \tag{24}$$

In the classical probability we have equality, because $\Pr(X = x_{\mathrm{th}}) = \Pr(Y = y_{\mathrm{th}}) = 0$. In order to compare the threshold branch, we convert Eq. (24) into soft probability (replacing Pr by Ps),

$$\mathrm{Ps}(X = x_{\mathrm{th}})\mathrm{H}(S|X = x_{\mathrm{th}}) \overset{>}{\underset{<}{?}} \mathrm{Ps}(Y = y_{\mathrm{th}})\mathrm{H}(S|Y = y_{\mathrm{th}}), \tag{25}$$

and using the soft probability definition for Uniform distribution we can express Eq. (25) as follows:

$$\frac{1}{2}\mathrm{H}(S|X = x_{\mathrm{th}}) \cdot \bar{0} \overset{>}{\underset{<}{?}} \frac{1}{4}\mathrm{H}(S|Y = y_{\mathrm{th}}) \cdot \bar{0}. \tag{26}$$

Now all we need is to calculate the entropy in the threshold (which is not infinitesimal) we can conclude which has the better information gain. In the case where $\mathrm{H}(S|X = x_{\mathrm{th}}) = \mathrm{H}(S|Y = y_{\mathrm{th}})$ we can say that X is the feature we should split our data according to as $\frac{1}{2} \cdot \bar{0} > \frac{1}{4} \cdot \bar{0}$, i.e., $\mathrm{IGs}(X) > \mathrm{IGs}(Y)$.

To further understand the applicability our suggested solutions we will analyze the next example of numeric of an electric product warranty.

3.2.3 Example: Electric Product Warranty

We would like to predict if an electric product will fail before the end of its 1 year warranty or not. Suppose that we have the following data to analyze:

- Two features which are the "Average Power Consumption" A and the "Peak Power Consumption" B. Due to production process, both has uniform distributions:
  $A \sim U(0.790, 0.800)$[watt] and $B \sim U(0.900, 1.000)$[watt];

- Denote the random variable F as the failure of an electric product ('$F = 1$' or simplify $F \to$ failure, '$F = 0$' or simplify $\bar{F} \to$ no failure), and assume that we have enough samples to calculate the conditional distributions $\Pr(F|A = a)$ and $\Pr(F|B = b)$.

Form the measurement we have a Heaviside probability function as follows for the Average Power Consumption:

$$\Pr(F|A = a) = \begin{cases} 0 & \text{if } a > 0.975 \\ 1 & \text{if } a \leq 0.975 \end{cases} \tag{27}$$

Same goes for the Peak Power Consumption but the threshold is in different value as follows:

$$\Pr(F|B = b) = \begin{cases} 0 & \text{if } b > 0.950 \\ 1 & \text{if } b \leq 0.950 \end{cases} \tag{28}$$

So clearly when we want to split our data by each parameter we know which value will be our splitting threshold, for the Average Power Consumption we will choose $A = 0.795$, and for the Peak Power Consumption we will choose $B = 0.950$. In this example for simplicity we will assume that the initial entropy of F is $\mathrm{H}_0(F) = 0.7$ Now we need to calculate which feature gives us the most information gain as the C4.5 algorithm defines. but if we look closely at the threshold numbers we can see that they are exactly in the middle of the possible values range of each feature, when we know they are uniformly distributed we get the following equation:

$$\Pr(A > 0.795) = \Pr(A \leq 0.795) = \Pr(B > 0.950) = \Pr(B \leq 0.950) = 0.5. \tag{29}$$

From that the conditional probability is a Heaviside function we have the following IG of for each feature:
for the Average Power Consumption

$$\begin{aligned} &\mathrm{IG}(W) \\ &= \mathrm{H}_0(F) - [\Pr(A \leq 0.795)\mathrm{H}(F|A \leq 0.795) + \Pr(A > 0.795)\mathrm{H}(F|A < 0.795)] \\ &= 0.7, \end{aligned} \tag{30}$$

and for the Peak Power Consumption

$$\begin{aligned} &\mathrm{IG}(B) \\ &= \mathrm{H}_0(F) - [\Pr(B \leq 0.950)\mathrm{H}(F|B \leq 0.950) + \Pr(B > 0.950)\mathrm{H}(F|B < 0.950)] \\ &= 0.7. \end{aligned} \tag{31}$$

As we can see we get equilibrium between the two IG, in classic approaches we can randomly choose the splitting by any feature. On the other hand using the "Soft Decision Tree" approach we suggested in this work we can get more information to get better defined decision. To calculate the soft IG we first need to calculate the soft probability in the equal "=" branch (the soft branch of the tree) as follows:

$$\begin{aligned} \mathrm{Ps}(A = 0.795) &= \frac{1}{(0.800) - (0.790)} \cdot \bar{0} = 100 \cdot \bar{0} \\ \mathrm{Ps}(B = 0.950) &= \frac{1}{(1.000) - (0.900)} \cdot \bar{0} = 10 \cdot \bar{0} \end{aligned} \tag{32}$$

For simplicity we will assume that $\mathrm{H}(F|A = 0.795) = \mathrm{H}(F|B = 0.950) = 0.5$, as we have a theoretical data set. Now we can calculate the soft information gain of both

features.
for the Average Power Consumption

$$
\begin{aligned}
\mathrm{IGs}(A) = \mathrm{H}_0(F) &- \mathrm{Ps}(A = 0.795)\mathrm{H}(F|A = 0.795) \\
&- [\mathrm{Ps}(A < 0.795)\mathrm{H}(F|A < 0.795) + \mathrm{Ps}(A > 0.795)\mathrm{H}(F|A < 0.795)] \\
= &- 50 \cdot \bar{0} \dot{+} 0.7,
\end{aligned}
\tag{33}
$$

and for the Peak Power Consumption

$$
\begin{aligned}
\mathrm{IGs}(B) = \mathrm{H}_0(F) &- \mathrm{Ps}(B = 0.950)\mathrm{H}(F|B = 0.950) \\
&- [\mathrm{Ps}(B < 0.950)\mathrm{H}(F|B < 0.950) + \mathrm{Ps}(B > 0.950)\mathrm{H}(F|B < 0.950)] \\
= &- 5 \cdot \bar{0} \dot{+} 0.7.
\end{aligned}
\tag{34}
$$

Now when we compare the results we can clearly see that $\mathrm{IGs}(B) > \mathrm{IGs}(A)$ so we will decide to split the data according to the feature, Average Power Consumption A.

4 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.

We introduced an approach to implement C4.5 algorithm as an example for a Soft Decision Tree, but with discrete labels and continuous features. We demonstrated with some numerical examples how the C4.5 algorithm would decide to split the data based on a "soft branch" created by a soft probability of event, that would collapse absolutely to zero, and without that branch, we might decide to split the data just arbitrarily.

5 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. We also suggest to explore the implementation of Decision Trees with soft Numbers into other Decision Trees algorithms, and also to consider each case whenever the labels/features are either discrete variables or continuous variable given being within intervals and single points.

We also suggest to explore the soft logic in general and soft probability in particular in additional topics in information theory, data mining, machine learning, computability, meta-verse technology, cyber-physical system (CPS) etc. We also suggest to involve the views of the theory of consciousness in the mentioned above scientific and technological topics, with the concept of the zero axis presents the inner world or virtual world, and the one axis the real world (see paragraph below Eq. (B.11) for more details). We believe that with soft logic (and soft probability) we can incorporate the spiritual concept of consciousness, that present inner/virtual world or the zero axis, into the scientific and technological topics in the real world or the one axis.

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.

We also want to that Sagy Naim for his idea to implement Soft Decision Trees based on C4.5 algorithm and for writing the related subsection 3.2.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Wadsworth Int. Group, 2017. DOI: 10.1201/9781315139470.

[2] W. Clifford, "Preliminary Sketch of Biquaternions," *Proceedings of the London Mathematical Society*, vol. s1-4, no. 1, pp. 381–395, Nov. 1871. DOI: 10.1112/plms/s1-4.1.381.

[3] O. Fivel, M. Klein, and O. Maimon, "Decision trees with soft numbers," *International Journal of Circuits, Systems and Signal Processing*, vol. 15, pp. 1803–1816, 2022. DOI: 10.46300/9106.2021.15.194.

[4] A. I. Khuri, "Applications of dirac's delta function in statistics," *International Journal of Mathematical Education in Science and Technology*, vol. 35, no. 2, pp. 185–195, 2004. DOI: 10.1080/00207390310001638313.

[5] M. Klein and O. Maimon, "Axioms of soft logic," *p-Adic Numbers, Ultrametric Analysis and Applications*, vol. 11, no. 3, pp. 205–215, 2019. DOI: 10.1134/s2070046619030038.

[6] M. Klein and O. Maimon, "The dynamics in the soft numbers coordinate system," *Journal Of Advances In Mathematics*, vol. 18, pp. 1–17, 2020. DOI: 10.24297/jam.v18i.8531.

[7] M. Klein and O. Maimon, "Fundamentals of soft logic," *New Mathematics and Natural Computation*, vol. 17, no. 03, pp. 703–737, 2021. DOI: 10.1142/s1793005721500356.

[8] J. Li and J. Chen, "Appendix A: Dirac delta function," in *Stochastic Dynamics of Structures*, J. Li and J. Chen, Eds., John Wiley & Sons, Ltd, 2009, pp. 343–348. DOI: 10.1002/9780470824269.app1.

[9] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Los Altos: Morgan Kaufmann, 1993.

[10] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986. DOI: 10.1007/BF00116251.

[11] L. Rokach and O. Maimon, "Decision trees," in *Data mining and knowledge discovery handbook*, O. Maimon and L. Rokach, Eds., Springer, 2005, pp. 165–192. DOI: 10.1007/0-387-25465-X_9.

[12] L. Rokach and O. Z. Maimon, *Data mining with decision trees: theory and applications*. World Scientific, 2015, vol. 81. DOI: 10.1142/9097.

[13] R. Zanetti and K. Tuggle, "A novel gaussian mixture approximation for nonlinear estimation," in *2018 21st International Conference on Information Fusion (FUSION)*, IEEE, 2018, pp. 1100–1106. DOI: 10.23919/ICIF.2018.8455485.

Appendix A Complement, Union, Intersection and Conditional Soft Probabilities

A.1 Complement, Unions and Intersections

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

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

A Soft probability of a complement is defined similarly as follows

$$\mathrm{Ps}(A^c) = 1 - \mathrm{Ps}(A). \tag{A.2}$$

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

$$\begin{aligned}\mathrm{Ps}(X \neq x) &= 1 - \mathrm{Ps}(X = x)\\ &= [-f_X(x)]\bar{0}\dot{+}1.\end{aligned} \tag{A.3}$$

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

$$\begin{aligned}\mathrm{Ps}(X = x \cup X = y) &= \mathrm{Ps}(X = x) + \mathrm{Ps}(X = y)\\ &= [f_X(x) + f_X(y)]\bar{0}.\end{aligned} \tag{A.4}$$

For an intersection we have

$$\mathrm{Ps}(X = x \cap X = y) = \mathbb{1}_{x=y} f_X(x)\bar{0}, \tag{A.5}$$

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:

$$\mathrm{Ps}\left(\bigcup_{i=1}^n X = x_i\right) = \sum_{i=1}^n \mathrm{Ps}(X = x_i) = \left[\sum_{i=1}^n f_X(x_i)\right]\bar{0}, \tag{A.6}$$

and

$$\mathrm{Ps}\left(\bigcap_{i=1}^n X = x_i\right) = \mathbb{1}_{x_i=x_j}^{\forall i,j\in\{1,2,\ldots,n\}} f_X(x_i)\bar{0}. \tag{A.7}$$

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

$$\mathrm{Ps}(X = x \cup a < X < b) = \mathbb{1}_{x \notin (a,b)} f_X(x)\bar{0} \dot{+} [F_X(b) - F_X(a)], \tag{A.8}$$

and for the intersection

$$\mathrm{Ps}(X = x \cap a < X < b) = \mathbb{1}_{x \in (a,b)} f_X(x)\bar{0}. \tag{A.9}$$

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

$$\begin{aligned} &\mathrm{Ps}(X = x \cup a \leq X \leq b) = \\ &[\mathbb{1}_{x \notin [a,b]} f_X(x) + f_X(a) + f_X(b)]\bar{0} \dot{+} [F_X(b) - F_X(a)], \end{aligned} \tag{A.10}$$

and for the intersection

$$\mathrm{Ps}(X = x \cap a \leq X \leq b) = [\mathbb{1}_{x \in [a,b]} f_X(x)]\bar{0}. \tag{A.11}$$

the two terms $f_X(a) + f_X(b)$ in Eq. (A.10) are added to the soft zero part, due to Eq. (A.6).

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

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{A.12}$$

It can be shown that the soft probabilities in Eqs. (A.8)-(A.11) hold for De Morgan's Law Eq. (A.12). For example $A = \{X = x\}$, $B = \{a \leq X \leq b\}$ and $x \notin [a, b]$, we have

$$\begin{aligned} &\mathrm{Ps}(X = x \cup a \leq X \leq b) = \\ &\mathrm{Ps}(X = x) + \Pr(a \leq X \leq b) - \Pr(X = x \cap a \leq X \leq b). \end{aligned} \tag{A.13}$$

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"

$$\mathrm{Ps}(A \cup B) = \mathrm{Ps}(A) + \mathrm{Ps}(B) - \mathrm{Ps}(A \cap B). \tag{A.14}$$

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

A.2 Conditional Probability

Recall Kolmogorov definition for conditional probability

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \tag{A.15}$$

and for Bayes theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}, \tag{A.16}$$

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 Eq. (A.15) we have

$$\frac{\mathrm{Ps}(X = x \cap X = y)}{\mathrm{Ps}(X = y)} = \frac{\mathbb{1}_{x=y} f_X(x)\bar{0}}{f_X(y)\bar{0}} = \frac{\mathbb{1}_{x=y} \cdot \bar{0}}{1 \cdot \bar{0}}. \tag{A.17}$$

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

$$\mathrm{Ps}(X = x | X = y) = \mathbb{1}_{x=y}. \tag{A.18}$$

In this case we have a trivial equality with optional real values 0 or 1. For comparison with Bayes theorem Eq. (A.16)

$$\frac{\mathrm{Ps}(X = y | X = x)\mathrm{Ps}(X = x)}{\mathrm{Ps}(X = y)} = \frac{\mathbb{1}_{y=x} f_X(x)\bar{0}}{f_X(y)\bar{0}} = \mathbb{1}_{x=y}. \tag{A.19}$$

Now we consider $x, y \in S_X$, let $A = \{X = x\}$, $B = \{a \leq X \leq b\}$, with $x, a, b \in S_X$ such that $a < b$,$x \neq a$ and $x \neq b$. At the LHS of Kolmogorov definition Eq. (A.15) we have

$$\begin{aligned}&\frac{\mathrm{Ps}(X = x \cap a \leq X \leq b)}{\mathrm{Ps}(a \leq X \leq b)} = \\ &\frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]}.\end{aligned} \tag{A.20}$$

When applying Bayes theorem Eq. (A.16), we have

$$\begin{aligned}&\frac{\mathrm{Ps}(a \leq X \leq b | X = x)\mathrm{Ps}(X = x)}{\mathrm{Ps}(a \leq X \leq b)} = \\ &\frac{[\mathrm{Ps}(a \leq x \leq b | X = x)] f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]},\end{aligned} \tag{A.21}$$

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

$$\mathrm{Ps}(X = x | a \leq X \leq b) = \frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{[f_X(a) + f_X(b)]\bar{0}\dot{+}[F_X(b) - F_X(a)]}. \tag{A.22}$$

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:

$$\mathrm{Ps}(X = x | a \leq X \leq b) = \frac{\mathbb{1}_{x\in[a,b]} f_X(x)\bar{0}}{F_X(b) - F_X(a)}, \tag{A.23}$$

and for a given strict inequality condition, we have.

$$\mathrm{Ps}(X = x | a < X < b) = \frac{\mathbb{1}_{x\in(a,b)} f_X(x)\bar{0}}{F_X(b) - F_X(a)}. \tag{A.24}$$

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 Eq. (A.14).

A.3 Extension of Soft Probability for 2 Dimensions

Suppose that X and Y are two continuous random variables. By the regular De Morgan's Law Eq. (A.12), we can decompose the regular probability object $\Pr(X \leq x, Y \leq y)$ into a sum of the following probabilities

$$\begin{aligned} &\Pr(X \leq x, Y \leq y) = \\ &\overbrace{[\Pr(X < x, Y = y) + \Pr(X = x, Y < y) + \Pr(X = x, Y = y)]}^{0} \\ &+ \Pr(X < x, Y < y), \end{aligned} \tag{A.25}$$

such that each of the first three terms in the bracket collapses to zero in the classical probability. We define the soft probability object $\mathrm{Ps}(X \leq x, Y \leq y)$ in 2 random variables based on a Soft De Morgan's Law Eq. (A.14) as follows

$$\begin{aligned} &\mathrm{Ps}(X \leq x, Y \leq y) = \\ &[\mathrm{Ps}(X < x, Y = y) + \mathrm{Ps}(X = x, Y < y) + \mathrm{Ps}(X = x, Y = y)] \\ &+ \mathrm{Ps}(X < x, Y < y). \end{aligned} \tag{A.26}$$

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 $f_{X,Y}(x, y)$:

$$\mathrm{Ps}(X < x, Y = y) = \frac{\partial F_{X,Y}(x, y)}{\partial y} \cdot \bar{0}, \tag{A.27}$$

$$\mathrm{Ps}(X = x, Y < y) = \frac{\partial F_{X,Y}(x, y)}{\partial x} \cdot \bar{0}, \tag{A.28}$$

$$\mathrm{Ps}(X = x, Y = y) = \frac{\partial F_{X,Y}(x, y)}{\partial x \partial y} \cdot \bar{0} = f_{X,Y}(x, y) \cdot \bar{0}. \tag{A.29}$$

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

$$\mathrm{Ps}(X < x, Y < y) = \Pr(X < x, Y < y) = F_{X,Y}(x, y), \tag{A.30}$$

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

$$\begin{aligned} &\mathrm{Ps}(X \leq x, Y \leq y) = \\ &\left[\frac{\partial F_{X,Y}(x, y)}{\partial x} + \frac{\partial F_{X,Y}(x, y)}{\partial y} + f_{X,Y}(x, y)\right] \cdot \bar{0} \\ &\dot{+} F_{X,Y}(x, y). \end{aligned} \tag{A.31}$$

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 Eq. (A.14), we construct the soft probability $\mathrm{Ps}(X \leq x, Y < y)$ similarly as follows:

$$\mathrm{Ps}(X \leq x, Y < y) = \frac{\partial F_{X,Y}(x, y)}{\partial x} \cdot \bar{0} \dot{+} F_{X,Y}(x, y). \tag{A.32}$$

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

$$\mathrm{Ps}(X \leq x, Y = y) = \left[\frac{\partial F_{X,Y}(x, y)}{\partial y} + f_{X,Y}(x, y)\right] \cdot \bar{0}, \tag{A.33}$$

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

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 [5] 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 [6] and [7]).

B.1 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 [7] 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*:*

$$a\bar{0}\dotplus b = \{a\bar{0} \perp b; b \perp a\bar{0}\}$$

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

B.2 Mathematical Operations and Functions on Soft Numbers

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

- **Addition/subtraction:**

$$(a\bar{0}\dot{+}b) \pm (c\bar{0}\dot{+}d) = (a \pm c)\bar{0}\dot{+}(b \pm d); \tag{B.1}$$

- **Multiplication:**

$$(a\bar{0}\dot{+}b) \cdot (c\bar{0}\dot{+}d) = (ad + bc)\bar{0}\dot{+}bd; \tag{B.2}$$

- **Natural power:**

$$(a\bar{0}\dot{+}b)^n = nab^{n-1}\bar{0}\dot{+}b^n. \tag{B.3}$$

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

$$P_N(\alpha\bar{0}\dot{+}x) = \alpha P_N'(x)\bar{0}\dot{+}P_N(x). \tag{B.4}$$

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

$$f(\alpha\bar{0}\dot{+}x) = \alpha f'(x)\bar{0}\dot{+}f(x). \tag{B.5}$$

For a continuous random variable X with a CDF F_X and a PDF $f_X = F_X'$, we have the following soft number (cf. Eq. (6)).

$$F_X(\alpha\bar{0}\dot{+}x) = \alpha f_X(x)\bar{0}\dot{+}F_X(x). \tag{B.6}$$

In the next subsection, we discuss about the Soft Axis Coordinate System.

B.3 Soft Axis Coordinate System

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 4. It starts from 0 to 1 horizontally and then it turns 90° from 1 to infinity

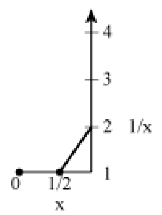

Figure 4: The Soft coordinate axis

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

Remark 2 *All lines that connect* x *to* $1/x$ *(for all non-zero real* x*) intersect at a single point.*

The statements in Remarks 1 and 2 were demonstrated in [5]. 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 5.

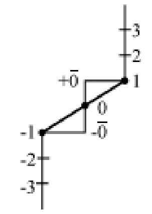

Figure 5: Distinction between -0 and +0

In Figure 5 we have, in addition to the absolute zero $\mathbf{0}$, 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 6 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.

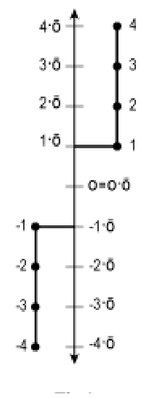

Figure 6: The extended soft coordinate system

The coordinate system in Figure 6 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 7). This allows us to represent both elements of a Soft Number:

$$\begin{aligned} c &= x\bar{0} \perp y, \\ c' &= y \perp x\bar{0}, \end{aligned} \tag{B.7}$$

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 6 to the one shown in Figure 7.

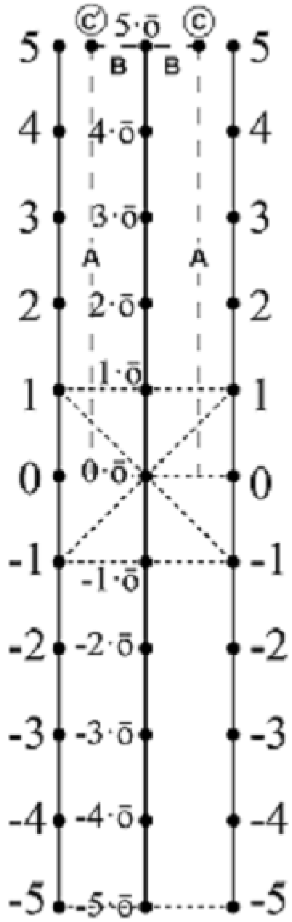

Figure 7: The complete soft coordinate system

As the infinite strip, presented (partially) in Figure 7, 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 $0 < B < 1$.

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:

$$\begin{aligned} c &= (1-B)A\bar{0} \perp BA, \\ c' &= BA \perp (1-B)A\bar{0}. \end{aligned} \tag{B.8}$$

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

$$\begin{aligned} x\bar{0}\dot{+}y &= \{c, c'\} \\ &= (1-B)A\bar{0}\dot{+}BA. \end{aligned} \tag{B.9}$$

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

$$\begin{aligned} x &= (1-B)A \\ y &= BA, \end{aligned} \tag{B.10}$$

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

$$\begin{aligned} A &= x + y \\ B &= \frac{y}{x+y}. \end{aligned} \tag{B.11}$$

It can be proven that there is an algebraic isomorphism between the bridge numbers $b\bar{0} \perp a$ and Dual numbers developed by Clifford [2] with the form $a + b\varepsilon$, where $\varepsilon^2 = 0$ but $\varepsilon \neq 0$. The main difference between ε in Dual numbers and $\bar{0}$ is the realisation and geometrical interpretation of $\bar{0}$ as an extension of the number 0 on a continuous line. This line can be a model of the inner world, while $\bar{1}$ is a model of the real world. The bridge between them enables us to treat the concept of consciousness with mathematical tools. Another difference is the possibility, in Soft logic, of developing a Soft curve [7].

One of our major topics for investigation in further research is the connection of soft numbers to Möbius strip. In order to describe the geometry of Möbius strip with soft numbers,we suggest to modify the soft coordinate system in Figure 7 by alternating the sign of left vertical line.

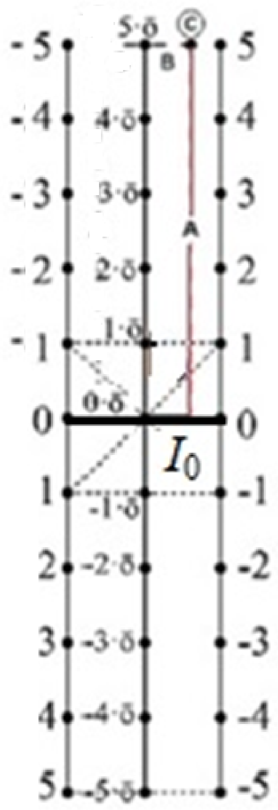

Figure 8: The alternative soft coordinate system

The horizontal line I_0 in Figure 8 can represent the connection line where the edges of a straight strip is twisted and attached together to create a Möbius strip. One of the suggestion to define a point on the Möbius strip with soft numbers is that $c = (1-B)A\bar{0} \perp BA$ is located in the front of this page, while $c' = BA \perp (1-B)A\bar{0}$ is located behind this page. This setup demonstrates locally existence of two side of Möbius strip. However, it is known that Möbius strip has globally one side. Moreover, if we start walking vertically from the point c (A units from I_0 and B units from the zero axis) on the front of this page, we will pass through the point behind this page but across the point c' and ($-A$ units from I_0 and B units from the zero axis). When we keep walking on that point, we will go back to the starting point c'. Because of this phenomenon, we are motivated to explore the possibility to represent a soft number with more than two symbols.

During exploration, it is suggested to change Definition 3 width of the point in the SNS will be between -1 and $+1$, i.e., $B \in [-1, 1]$. Given SNS height $A \in \mathbb{R}$ and SNS width $B \in [-1, 1]$, the SNS point $c = c(A, B)$ is defined as follows:

$$\begin{aligned} &\text{For } B \in [0,1], \\ &\qquad c = (1-B)A\bar{0} \perp BA, \text{ located in front of page,} \\ &\qquad c' = BA \perp (1-B)A\bar{0}, \text{ located } \textbf{behind} \text{ this page,} \\ &\qquad x\bar{0}\dotplus y = \{c, c'\}, \end{aligned} \tag{B.12}$$

and

$$\begin{aligned}
&\text{For } B \in [-1, 0],\\
&\quad c = (1 + B)A\bar{0} \perp BA, \text{ located in front of page,}\\
&\quad c' = BA \perp (1 + B)A\bar{0}, \text{ located } \textbf{behind} \text{ this page,}\\
&\quad x\bar{0}\dot{+}y = \{c, c'\},
\end{aligned} \tag{B.13}$$

or shortly in terms of absolute value of B:

$$\begin{aligned}
&c = (1 - |B|)A\bar{0} \perp BA, \text{ located in front of page,}\\
&c' = BA \perp (1 - |B|)A\bar{0}, \text{ located } \textbf{behind} \text{ this page,}\\
&x\bar{0}\dot{+}y = \{c, c'\}.
\end{aligned} \tag{B.14}$$

Hence, by a coefficients comparison of the real part and the soft part in Eq. (B.14):

$$\begin{aligned}
x &= (1 - |B|)A\\
y &= BA,
\end{aligned} \tag{B.15}$$

or equivalently, after solving for the **SP**, (A, B), using some arithmetic of absolute value and sign functions:

$$\begin{aligned}
A &= \operatorname{sign}(x) \cdot (|x| + |y|)\\
B &= \frac{y}{A} = \operatorname{sign}(x) \cdot \frac{y}{|x| + |y|}.
\end{aligned} \tag{B.16}$$

The conjecture is with Eqs. (B.15)- (B.16) and Figure 8, the geometry of Möbius strip with soft numbers can be defined more properly.