Article

On Defining Expressions for Entropy and Cross-Entropy: The Entropic Transreals and Their Fracterm Calculus

Jan A. Bergstra [1,*] and **John V. Tucker** [2]

1 Informatics Institute, University of Amsterdam, Science Park 900, 1098 XH Amsterdam, The Netherlands
2 Department of Computer Science, Bay Campus, Fabian Way, Swansea University, Swansea SA1 8EN, UK; j.v.tucker@swan.ac.uk
* Correspondence: j.a.bergstra@uva.nl or janaldertb@gmail.com

Abstract: Classic formulae for entropy and cross-entropy contain operations $\frac{x}{0}$ and $\log_2 x$ that are not defined on all inputs. This can lead to calculations with problematic subexpressions such as $0 \log_2 0$ and uncertainties in large scale calculations; partiality also introduces complications in logical analysis. Instead of adding conventions or splitting formulae into cases, we create a new algebra of real numbers with two symbols $\pm\infty$ for signed infinite values and a symbol named $\perp$ for the undefined. In this resulting arithmetic, entropy, cross-entropy, Kullback–Leibler divergence, and Shannon divergence can be expressed without concerning any further conventions. The algebra may form a basis for probability theory more generally.

Keywords: partial formulae; fracterm calculus; transreals; entropic transreals; peripheral numbers; entropy; cross-entropy

1. Introduction

Consider a probability function, or more precisely, a probability mass function, P on a finite sample space S for which it is assumed that

$$\forall_{v \in S} P(v) \geq 0 \text{ and } \sum_{v \in S} P(v) = 1.$$

The definition of entropy for P is often formulated as follows:

$$H(P) = -\sum_{s \in S} (P(s) \cdot \log_2 P(s)).$$

Alternately, it can be formulated as

$$H(P) = \sum_{s \in S} (P(s) \cdot \log_2 \frac{1}{P(s)}).$$

A closely related concept is cross-entropy defined for two probability mass functions, say P and Q, by

$$H(P,Q) = -\sum_{s \in S} (P(s) \cdot \log_2 Q(s)).$$

Alternately, it can be formulated as

$$H(P,Q) = \sum_{s \in S} (P(s) \cdot \log_2 \frac{1}{Q(s)}).$$

In the formulae, there are partial functions, i.e., functions that are not defined for all the values required: neither $\log_2(x)$ nor $\frac{1}{x}$ are defined for $x = 0$. However, a mass function with $P(s) = 0$ is a valid argument in both formulae.

Correct mathematical writing practice typically guards against partiality by expressing conditions that rule out arguments or break up formulae into different cases. Another technique, one which aims to preserve uniformity, is to invent conventions for applying the formulae, which may or may not have plausible justifications. Consider the first formula. Commonly, an additional convention often prescribed is that

$$0 \cdot \log_2 0 = 0.$$

This has an underlying argument that

$$\lim_{x \downarrow 0} x \cdot (\log_2 x) = 0$$

to which we will return to shortly.

The convention to adopt $0 \cdot \log_2 0 = 0$ may seem unproblematic [1]. But, from an algebraic and logical perspective in pure mathematics, and especially from the precisely formalised logical perspectives of computer science and software engineering, introducing a convention that allows one to calculate *as if* $0 \cdot \log_2 0 = 0$ is not at all straightforward as it raises questions about the effects such an identification. There are two issues in need of attention: first,

(i) *What is the scope of the assumption $0 \cdot \log_2 0 = 0$? Is it only local to the definitions of entropy and cross-entropy or is the scope more extensive.*
If the assumption is made more generally, for calculations in theoretical work on entropy, one may wonder about the value of other expressions, such as $0 \cdot \log_2^2 0, 0 \cdot \log_2^3 0$, etc.

(ii) *If $0 \cdot \log_2 0 = 0$, then what is the status of the log term $\log_2 0$?*
The question applies to the equally relevant term $\frac{1}{0}$ just as well, of course. Now, these formulae have long been very widely used in computing; however, of theoretical interest for programming is the question:

(iii) *What are the effects of the inherent partiality of the formulae being disguised?*
Partial operators on inputs that sometimes return no output are to be avoided in programming, and logical reasoning about programs becomes hugely more complicated than with total operators.

In this paper, we will adopt a foundational approach and explore the partiality of these entropy formulae—and of several derived information-theoretic formulae—in order to develop some new algebraic structures for real number arithmetic that can provide single uniform expressions that are well founded because they are based not on ad hoc conventions but on the algebra of the number systems that underpin the formulae.

To accomplish this, we will need to examine technique(s) for making partial functions total and detecting their effect on calculations. In pursuing one technique, that of using infinities, we develop an arithmetic algebra of real numbers customised to our problem and hence called *entropic transreals*. In the entropic transreals, the convention $0 \cdot \log_2 0 = 0$ becomes an authentic algebraic property, derivable from the algebra of entropic transreals.

1.1. Gauging the Problem

In addition to addition, subtraction and multiplication (which are total operations), the algebra of real numbers we aim to construct will also have division $\frac{x}{y}$ and $\log_2 x$ (which are partial operations that will need to be made total). What factors shape the design of the new algebra? For example, the ubiquitous presence of summation over a sample space, as in the definitions of entropy, suggests the requirement that addition is associative.

Consider how the convention $0 \cdot \log_2 0 = 0$ could be a valid equation in such an algebra. If one relies on an justification of the value of $0 \cdot \log_2 0$ by way of limits, as mentioned earlier, then applying the multiplication rule for limits we deduce

$$0 = \lim_{x \downarrow 0}(x \cdot (\log_2 x)) = \lim_{x \downarrow 0}(x) \cdot \lim_{x \downarrow 0}(\log_2 x) = 0 \cdot -\infty.$$

Thus, we have another rather basic identity, $0 = 0 \cdot -\infty$, to consider and the idea that $\log_2 0 = -\infty$.

Turning to cross-entropy, in some cases one expects it to take a positive infinite $+\infty$ value, and so an arithmetic equipped with both signed infinities $\pm\infty$ is needed.

Example 1. *Consider a sample space S with elements a and b and probability mass functions P and Q over S such that*

$$P(a) = 1, P(b) = 0, Q(a) = 0, Q(b) = 1.$$

Now, for the cross-entropy $H(P,Q)$, as it was defined above, we expect to find value $+\infty$. Calculation of $H(P,Q)$ yields:

$$H(P,Q) = P(a) \cdot \log_2 \frac{1}{Q(a)} + P(b) \cdot \log_2 \frac{1}{Q(b)} = 1 \cdot \log_2 \frac{1}{0} + 1 \cdot \log_2 \frac{1}{1} = \log_2 \frac{1}{0}.$$

So, to obtain the expected value $+\infty$ in this case, we can adopt the identity $\frac{1}{0} = +\infty$, in combination with $\log_2(+\infty) = +\infty$; this seems plausible, if not necessary.

Thus, with these and other requirements in mind, in the course of the paper, the field of real numbers will be enriched with new operations and elements to make a new algebra. In particular, we will extend it with a suitable pair of signed infinite 'values' $\pm\infty$—elements outside the conventional range of numbers. On adopting $\log_2 0 = -\infty$ and enabling $0 \cdot \infty = 0 \cdot (-\infty) = 0$ to hold, the desired convention $0 \cdot \log_2 0$ can derive an algebraic property of the underlying arithmetical data type.

1.2. Structure of the Paper

In Section 2, we prepare the ground with some background and methods for treating partiality that have been developed for division. In Section 3, we continue to explore and select algebraic properties customised to the task of rebuilding the entropic formulae. In Section 4, we apply the new entropic transreals to a series of formulae for entropy, cross-entropy, Kullback–Leibler divergence, and Shannon divergence. In Section 5, we summarise the construction of the algebra. In Section 6, we reflect on the exercise and point out some next steps and problems.

2. Peripheral Numbers, Fracterms, Fracterm Calculus

Technically, we are concerned with the use of expressions $\frac{1}{0}$ and $\log_2 0$ that have no values and finding ways to give them meaning for the purpose of improving calculating and reasoning. The tools we employ are made from a variety of logical concepts and methods concerning equations and related formulae that make up the theory of abstract data types in computer science [2]. However, to keep focussed on entropy, we will limit the use of this background knowledge that informs our investigation.

2.1. Peripherals

Our methods benefit from an elementary knowledge of syntax, namely *signatures*, which list names for constants and operators of an algebra, and the *terms* that are made by composing operators and applying them to constants and variables.

The constants $\perp$ and ∞ are new syntax, from a conventional point of view, though ∞ is used quite often in an informal manner, and these constants at the same time represent values outside the conventional number system, so-called *peripheral numbers*. We will consider arithmetical structures, or arithmetics for short, which feature three peripheral values: ∞, $-\infty$ and $\perp$. We will sometimes write $+\infty$ for ∞ to emphasize that positive infinity is meant.

In introducing new infinity constants ∞, $-\infty$ we generate the need for meaning for infinitely many new expressions:

$$\infty + \infty, \; \infty \cdot \infty, \; \infty - \infty, \; \frac{\infty}{\infty}, \; \ldots, \; \log_2 +\infty, \; \log_2 -\infty, \; \ldots$$

Some seem easy to resolve with identities, such as

$$\infty + \infty = \infty, \; \infty \cdot \infty = \infty,$$

while others suggest options, such as

$$\infty - \infty =?, \; \frac{\infty}{\infty} =?,$$

and the choices and the algebras they determine ramify. In our case, we will use $\perp$ in identities to resolve the matter.

Following [3], we use the word *fracterm* for a fractional expression. We avoid the noun "fraction" because its meaning is rather ambiguous, ranging between an expression and its number value. Whereas for constants making a distinction between expression and value is rather uninformative, for fractional expressions it matters a lot. We adopt $\frac{1}{0}$ as a fracterm without hesitation. We use *fracterm calculus* loosely for "how to calculate with fracterms". Different fracterm cacluli may be distinguished and axiomatised using formulae based on equations, for instance.

2.2. Models for Division by Zero

In this paper, we start from a series of thorough studies of the case of division: what can be done about $\frac{x}{0}$? There are several options that have been analysed.

(i) *Suppes-Ono fracterm calculus.* This is based on the assumption $\frac{x}{0} = 0$ and makes no use of $\perp$. For information on Suppes-Ono fracterm calculus, we refer to [4–8].

(ii) *Common meadows fracterm calculus.* This uses $\frac{x}{0} = \perp$ and makes no use of ∞ and $-\infty$. We refer to [9–11] for common meadows.

(iii) *Transreal fracterm calculus.* This uses Φ (instead of $\perp$ and named nullity), ∞ and $-\infty$. See [12–17].

(iv) *Fracterm calculus for symmetric transreals.* This involves peripherals for signed infinitesimals, as well as for signed infinities. See [18].

(v) *Fracterm calculus for wheels.* This makes use of $\perp$ while identifying ∞ and $-\infty$ and maintaining $\infty + \infty = \perp$. See [19].

Below, we will propose an adaptation of transreal fracterm calculus by introducing $\perp$ besides Φ and adopting $\log_2 p = \perp$ rather than $\log_2 p = \Phi$ for negative p, in order to find a better alignment between different fracterm calculi. This leads us to our *Fracterm calculus for entropic transreals.*

Entropic transreals are an approach to enlargement of arithmetic with peripheral numbers (which will be introduced below) that is designed specifically to meet the objective to provide a precise meaning for defining expressions for entropy and cross-entropy. We will speak of *entropic* transreals—to emphasise their motivation. Arguably, the entropic transreals embody somewhat arbitrary assumptions, e.g., the combination of $0 \cdot \infty = 0$ and $\infty - \infty = \bot$ may be called rather ad hoc.

2.3. Indicating Partiality

We will adopt these conventions: instead of "$f(a_1, \ldots, a_n)$ is undefined" we write $f(a_1, \ldots, a_n) = \bot$. Thus, $\bot$ is considered an element of the domain of values. So $\bot$ plays a conventional role as an element of the domain in the setting of equational logic with the effect that, for instance, $\bot = \bot$ and $\bot \neq 0$. We often refer to $\bot$ as an "error value", but its role as a token for partiality need not be understood as signalling an error.

We assume that $\log_2$ is undefined for negative arguments, though following the approach of [9] and adopting the mechanism of quasi-partiality we prefer to work with total functions writing $f(a) = \bot$ in case one thinks of $f(-)$ as being undefined for argument a, a consideration which leads to $\log_2 p = \bot$ for any real number $p < 0$, as well as to $\log_2(-\infty) = \bot$.

When making use of a square root function $\sqrt{x}$ we will adopt $\sqrt{p} = \bot$ for negative real p as well as $\sqrt{-\infty} = \bot$. Adopting $f(a) = \bot$ only indicates that no proper value is assigned to $f(a)$ in the setting at hand, while it may be the case that in a larger structure (such as a field of complex numbers) such values may be easily found. With the use of $\bot$ in these cases we deviate from transreals (as in [12]) where $\log_2(-1) = \Phi$ is assumed.

3. Entropic Transreals

In various floating point systems for computer arithmetic one finds $\frac{1}{0} = +\infty$ and $\frac{-1}{0} = -\infty$. Remarkably, however, within theoretical computer science these ubiquitous conventions have not lead to any systematic research on versions of arithmetic with peripheral numbers for signed infinities. We are unaware of any occurrence of ∞ with the properties of entropic transreals (perhaps with another name or symbol) in the literature. The design and elaboration of transreal arithmetic stands out as a rather singular example.

3.1. The Fracterm Calculus of Transreals Fails to Match Our Requirements

The best-known instance of a version of arithmetic providing peripheral numbers $\pm\infty$ is the system of transreals as defined in [13]. Transreals contain an absorptive constant Φ, named nullity, that satisfies

$$0 \cdot \infty = 0 \cdot (-\infty) = x + \Phi = x \cdot (-\infty) = \Phi.$$

It follows that, upon adopting $\log_2 0 = -\infty$, one obtains for the convention

$$0 \cdot \log_2 0 = 0 \cdot \Phi = \Phi$$

instead of the desired $0 \cdot \log_2 0 = 0$. Thus, transreal arithmetic will not support the definition of entropy in its conventional form.

Moreover, for any probability mass function P on S which vanishes on at least one sample $s \in S$, one finds $H(P) = \Phi$ when adopting the conventional definition of entropy in combination with the conventions of transreal arithmetic.

Transreal arithmetic was designed with the IEEE 754 standard in mind (see also [14]), and maintaining definitions from probability theory without any modification has been no requirement on the design of the fracterm calculus of transreals.

3.2. The Fracterm Calculus of Entropic Transreals

We will adopt entropic transreals, an enlargement of reals with different periperhal numbers $\pm\infty$ such that

$$0 \cdot \infty = 0 \cdot (-\infty) = 0.$$

The simplification with respect to transreals lies in the fact that the familiar identity $0 \cdot x = 0$ is maintained to a greater extent. In other words, the role of nullity (Φ) is reduced, and in fact so much reduced that its remaining role is played by the partiality indicator $\bot$. Notice that with the error element $\bot$, we will be using $0 \cdot \bot = \bot$ rather than $0 \cdot \bot = 0$, so that the familiar equation $0 \cdot x = 0$ is again compromised in entropic transreals, though to a lesser extent than in transreals where $0 \cdot \infty = 0 \cdot (-\infty) = \Phi$ is adopted.

With $\pm\infty$ available, we will follow the design of transreals as in [12] and adopt the following equations: $\log_2 0 = -\infty$ and $\log_2 \infty = \infty$. Finding a value for $\log_2 -1$ and for $\log_2 -\infty$ is another matter, however.

In Section 5, we will summarise in full detail the domain, the constants, and the various operators of entropic transreals.

3.3. The Role of $\bot$

The peripheral value $\bot$ is needed for entropic transrationals in order to evaluate the sumterm $\infty + (-\infty)$. We will adopt for entropic transrationals the identity $\infty + (-\infty) = \infty - \infty = \bot$, an equation which is already valid for transrationals. This assumption is consistent with the requirement that addition and multiplication are associative. We notice that both associativity and commutativity of addition are needed for generalized addition over a finite domain to have a plausible interpretation. Generalized addition occurs in the definition of entropy and of expected value in general.

Although $\bot$ is not included in transreals, we consider entropic transreals with $\bot$ for quasi-partiality is still to be a simplification of transreals. In transreals, Φ is not supposed to play the role of $\bot$, and therefore Φ is not supposed to model partiality in general: in the design of transreals, Φ is instead the meaningful value of $0 \cdot \infty$. In our case, $\log_2(-3)$ is not supposed to have a meaningful value, so that we consider it plausible to set $\log_2(-3) = \bot$ in transreals, as well as in entropic transreals.

3.4. Dealing with Non-Distributivity

Just as with the transreals, the entropic transreals are not distributive. For transreals, assuming distributivity leads to the following inconsistency:

$$\infty = 1 \cdot \infty = (1 + 0) \cdot \infty = 1 \cdot \infty + 0 \cdot \infty = \infty + \Phi = \Phi.$$

For entropic transrationals, upon assuming distributivity, one finds an inconsistency as well:

$$\bot = \infty + (-\infty) = (1 + (-1)) \cdot \infty = 0 \cdot \infty = 0.$$

Although failure of distributivity is unpleasant, it appears not to constitute a fundamental obstacle for the use of a particular arithmetical data type. Using conditional equations, several useful versions of distiributivity can be found, for instance: $0 \cdot x = 0 \rightarrow x \cdot (y + z) = x \cdot y + x \cdot z$.

The lack of distributivity can be expressed without making use of constants for peripheral numbers (i.e., ∞ or $\bot$) and it is the following well-known rule that fails for $x = 1$, $y = -1$, $z = 0$:

$$\frac{x}{z} + \frac{y}{z} = \frac{x+y}{z}.$$

4. Application to Entropy, Cross Entropy and Other Concepts

We consider a series of formulae and make some calculations using the entropic transreals as examples.

4.1. An Expression for Entropy

The second expression for entropy in the Introduction introduces an issue of division by zero:

$$H(P) = \sum_{s \in S} (P(s) \cdot \log_2 \frac{1}{P(s)}).$$

In case for some $s \in S$, $P(s) = 0$, for this expression it is plausible to follow the conventions of transreals as follows:

$$\frac{1}{0} = +\infty, \frac{-1}{0} = -\infty, \frac{1}{+\infty} = \frac{1}{-\infty} = 0.$$

For a summand coming from a sample s with $P(s) = 0$ we find:

$$P(s) \cdot \log_2 \frac{1}{P(s)} = 0 \cdot \log_2 \frac{1}{0} = 0 \cdot \log_2 \infty = 0 \cdot \infty = 0,$$

an outcome which we consider to be adequate for the definition of entropy.

We begin a series of running examples as a simple check and illustration of calculating with the formulae.

Example 2. *Consider $S = \{a, b\}$ and $P(a) = P(b) = \frac{1}{2}$ while $Q(a) = 1$ and $Q(b) = 0$. We find for P and Q, in this case:*

$$H(P) = P(a) \cdot \log_2 \frac{1}{P(a)} + P(b) \cdot \log_2 \frac{1}{P(b)} = \frac{1}{2} \cdot \log_2 \frac{1}{(\frac{1}{2})} + \frac{1}{2} \cdot \log_2 \frac{1}{(\frac{1}{2})} =$$
$$\frac{1}{2} \cdot \log_2 2 + \frac{1}{2} \cdot \log_2 2 = 1$$

and

$$H(Q) = Q(a) \cdot \log_2 \frac{1}{Q(a)} + Q(b) \cdot \log_2 \frac{1}{Q(b)} = 1 \cdot \log_2 \frac{1}{1} + 0 \cdot \log_2 \frac{1}{0} =$$
$$1 \cdot 0 + 0 \cdot \log_2(+\infty) = 0 + 0 \cdot (+\infty) = 0.$$

Evaluating $H(P)$ and $H(Q)$ with the first definition of entropy will produce the same value because we are working with the native equality in an algebra on entropic transreals.

4.2. Cross-Entropy

Cross-entropy is defined for two probability mass functions, say P and Q, as follows:

$$H(P, Q) = \sum_{s \in S} (P(s) \cdot \log_2 \frac{1}{Q(s)})$$

Example 3. *Again, take $S = \{a, b\}$ and $P(a) = P(b) = \frac{1}{2}$ while $Q(a) = 1$ and $Q(b) = 0$. We calculate:*

For $H(P, Q)$, we find:

$$H(P, Q) = P(a) \cdot \log_2 \frac{1}{Q(a)} + P(b) \cdot \log_2 \frac{1}{Q(b)} = \frac{1}{2} \cdot \log_2 \frac{1}{1} + \frac{1}{2} \cdot \log_2 \frac{1}{0} =$$
$$\frac{1}{2} \cdot \log_2 1 + \frac{1}{2} \cdot \log_2 \infty = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \infty = 0 + \infty = \infty.$$

For $H(Q, P)$, we find:

$$H(Q,P) = Q(a) \cdot \log_2 \tfrac{1}{P(a)} + Q(b) \cdot \log_2 \tfrac{1}{P(b)} = 1 \cdot \log_2 \tfrac{1}{(\frac{1}{2})} + 0 \cdot \log_2 \tfrac{1}{(\frac{1}{2})} =$$
$$1 \cdot \log_2 2 + 0 \cdot \log_2 2 = 1 \cdot 1 + 0 \cdot 1 = 1.$$

Notice that $H(P,Q) \neq H(Q,P)$.

4.3. Alternative Expression for Cross-Entropy

The other definition of cross entropy reads:

$$H(P,Q) = -\sum_{s \in S}(P(s) \cdot \log_2 Q(s))$$

This definition depends on the basic assumptions of entropic transreal arithmetic, though in a different manner, now making use of $\log_2 0 = -\infty$.

Example 4. *Again, take $S = \{a,b\}$ and $P(a) = P(b) = \frac{1}{2}$ while $Q(a) = 1$ and $Q(b) = 0$. We calculate:*

$$H(P,Q) = -P(a) \cdot \log_2 Q(a) - P(b) \cdot \log_2 Q(b) = -\tfrac{1}{2} \cdot \log_2 1 - \tfrac{1}{2} \cdot \log_2 0 =$$
$$-\tfrac{1}{2} \cdot \log_2 1 - \tfrac{1}{2} \cdot (-\infty) = -\tfrac{1}{2} \cdot 0 + \tfrac{1}{2} \cdot \infty = 0 + \infty = \infty,$$

and

$$H(Q,P) = -Q(a) \cdot \log_2 P(a) - Q(b) \cdot \log_2 P(b) = -1 \cdot \log_2 \tfrac{1}{2} - 0 \cdot \log_2 \tfrac{1}{2} =$$
$$-1 \cdot \log_2 -1 - 0 \cdot 1 = 1.$$

We obtain the same values.

4.4. A Modification of the Example

The running example above can be modified by adding a new sample element *c*:

Example 5. *We consider $S' = S \cup \{c\} = \{a,b,c\}$ and extend the definitions of P and Q with values on c. Define $P'(a) = P'(b) = \frac{1}{2}, P'(c) = 0$ while $Q'(a) = 1$ and $Q'(b) = Q'(c) = 0$.*
Now, we find:

$$H(P') = H(P) + P'(c) \cdot \log_2 \tfrac{1}{P'(c)} = 1 + 0 \cdot \log_2 \tfrac{1}{0} = 1 + 0 \cdot \log_2 \infty = 1 + 0 \cdot \infty = 1 + 0 = 1,$$

$$H(Q') = H(Q) + Q'(c) \cdot \log_2 \tfrac{1}{Q'(c)} = 0 + 0 \cdot \log_2 \tfrac{1}{0} = 0 \cdot \log_2 \infty = 0 \cdot \infty = 0.$$

and for cross entropy

$$H(P',Q') = H(P,Q) + P'(c) \cdot \log_2 \frac{1}{Q'(c)} = \infty + 0 = \infty.$$

4.5. Kullback–Leibler Divergence

Kullback–Leibler divergence is not symmetric on the above example for *P* and *Q*:

$$D_{\mathsf{KL}}(P||Q) = H(P,Q) - H(P) = \infty - 1 = \infty$$

and

$$D_{\mathsf{KL}}(Q||P) = H(Q,P) - H(Q) = 1 - 0 = 1.$$

We find, as is well known, that already on probability mass functions that vanish nowhere, $D_{\mathsf{KL}}(-,-)$ is asymmetric.

Example 6. *Here is a new probability mass function R given by $R(a) = \frac{1}{3}$ and $R(b) = \frac{2}{3}$. We find:*
$$H(R) = -\tfrac{1}{3} \cdot \log_2 \tfrac{1}{3} - \tfrac{2}{3} \cdot \log_2 \tfrac{2}{3} = \tfrac{1}{3} \cdot \log_2 3 + \tfrac{2}{3} \cdot \log_2 3 - \tfrac{2}{3} \log_2 2 = \log_2 3 - \tfrac{2}{3}.$$

$$H(P, R) = -P(a) \cdot \log_2 R(a) - P(b) \cdot \log_2 R(b) = -\frac{1}{2} \cdot \log_2 \frac{1}{3} - \frac{1}{2} \cdot \log_2 \frac{2}{3} = \frac{1}{2} \cdot \log_2 3 +$$
$\frac{1}{2} \cdot \log_2 3 - \frac{1}{2} \cdot \log_2 2 = \log_2 3 - \frac{1}{2}$, *and so*

$$D_{\mathsf{KL}}(P||R) = H(P, R) - H(P) = \log_2 3 - \frac{1}{2} - 1 = \log_2 3 - \frac{3}{2}.$$

Moreover, we have: $H(R, P) = -R(a) \cdot \log_2 P(a) - R(b) \cdot \log_2 P(b) = -\frac{1}{3} \cdot \log_2 \frac{1}{2} - \frac{2}{3} \cdot \log_2 \frac{1}{2} = 1$ *and so*

$$D_{\mathsf{KL}}(R||P) = H(R, P) - H(R) = 1 - (\log_2 3 - \frac{2}{3}) = \frac{5}{3} - \log_2 3.$$

4.6. Mutual Information

An instance of Kullback–Leibler divergence is so-called *mutual information*, where $S = U \times V$, and R is a probability mass function on S with marginals P and Q, i.e.,

$$P(u) = \sum_{v \in V} R(u, v) \text{ and } Q(v) = \sum_{u \in U} R(u, v).$$

Now,

$$I(R) = D_{\mathsf{KL}}(R||P \cdot Q) = \sum_{u \in U, v \in V} R(u, v) \cdot \log_2 \frac{R(u, v)}{P(u) \cdot Q(v)}$$

We notice that a probability mass function cannot have values ∞, $-\infty$ or $\perp$, and, moreover, if either $P(u) \cdot Q(v) = 0$ then necessarily also $R(u, v) = 0$ so that $I(R)$ is guaranteed to be finite, i.e., have no peripheral value.

That $\frac{0}{0} = 0$ is the only nontrivial property needed of the underlying arithmetic for the above definition of $I(R)$ to be adequate.

4.7. Jensen–Shannon Divergence

Given probability mass functions P and Q on S, the *Jensen–Shannon divergence* is as follows:

$$D_{\mathsf{JS}}(P||Q) = \frac{D_{\mathsf{KL}}(P||M) + D_{\mathsf{KL}}(Q||M)}{2}$$

where $M = \frac{P+Q}{2}$. We notice that for all P and Q, $D_{\mathsf{JS}}(P||Q) \neq \infty$. Otherwise, for some $s \in S$, $M(s) = 0$ must hold together with either $P(s) \neq 0$ or $P(s) \neq 0$ (or both), which is impossible given the definition of M.

4.8. Expected Value

Both entropy and cross-entropy are instances of an *expected value*, obtained upon choosing a suitable function on the sample space. Defining an expected value operator, given a probability mass function and a function from the sample space to (real) numbers, requires an additional definition, however.

For a function F from a finite sample space S to reals (or rather to entropic transreals) and a probability mass function P, the following definition of an expected value $\widehat{E}_{P,F}^S$ is plausible:

$$\widehat{E}_P^S(F) = \sum_{x \in S, P(x) \neq 0} P(x) \cdot F(x)$$

The virtue of this form is that unlike $E_P^S(F) = \sum_{x \in S} P(x) \cdot F(x)$ it works well (i.e., avoids result $\perp$) in case for some $a \in S$, $P(a) = 0$ while $F(a) = \perp$.

However, in fact, we prefer that latter property of an expected value operator, i.e., whenever $F(a) = \bot$ for some $a \in S$, then the expected value of F on S equals $\bot$, independently of the probability mass function at hand. For that reason, we will adopt

$$E_P^S(F) = \sum_{x \in S} P(x) \cdot F(x)$$

as an appropriate definition of expected value and so that $E_P^S(F) = \bot$ whenever for some $a \in S, F(a) = \bot$. Adopting a convention of this kind expresses the idea that an event with probability 0 is not altogether impossible, its probability is merely extremely low. We refer to [20] for an exposition on this somewhat unconventional position regarding the status of zero-probability events.

Now, we may reformulate the defining expressions for entropy, cross-entropy and Kullback–Leibler divergence as follows:

$$H(P) = E_P^S\left(\frac{1}{P(s)}\right)$$

$$H(P,Q) = E_P^S\left(\frac{1}{Q(s)}\right)$$

$$D_{\mathsf{KL}}(P||Q) = E_P^S\left(\frac{P(s)}{Q(s)}\right)$$

5. Entropic Transreals in Detail

We will now describe entropic transreals in minute detail in order to prevent any confusion. The starting point is a field $\mathbb{R}$ of reals with constants 0 an 1 and functions addition, additive inverse, and multiplication. To this field we add operators $\frac{x}{y}$ for division and and $\log_2 x$ for logarithm. (We only discuss functions which play a role in the paper, other functions such as exponentiation and square root might be included as well.) We also assume the presence of an ordering which we handle using a sign operator $\mathsf{s}(x)$ defined for:

$$x > 0, \mathsf{s}(x) = +1, \quad x = 0, \mathsf{s}(x) = 0, \quad x < 0, \mathsf{s}(x) = -1.$$

The domain $\mathbb{R}$ of real numbers is enlarged by extending the domain with three new elements: ∞, $-\infty$ and $\bot$. Thus, the form of the algebra is:

$$(\mathbb{R} \cup \{+\infty, -\infty, \bot\} \mid 0, 1, +\infty, -\infty, \bot, +, -, \cdot, \div, \log_2, \mathsf{s}(x))$$

Let $\mathbb{R}_{\bot,\pm\infty} = \mathbb{R} \cup \{+\infty, -\infty, \bot\}$ denote the domain.

We now have to define the operations on the three peripherals.

As $\bot$ is absorptive, the value of any operation on arguments at least one of which equals $\bot$ is $\bot$. So, we need not specify in detail values of operators in case one of the arguments is $\bot$. We turn to the infinities, which can be subtle.

Sign function. This is easily extended to the larger domain $\mathbb{R}_{\bot\pm\infty}$ as follows:

$$\mathsf{s}(\infty) = 1, \mathsf{s}(-\infty) = -1, \mathsf{s}(\bot) = \bot.$$

Addition. This is extended as follows: for $p \in \mathbb{R}$:

$$\infty + p = \infty \text{ and } -\infty + p = -\infty;$$
$$\infty + \infty = \infty \text{ and } (-\infty) + (-\infty) = -\infty;$$
$$\infty + (-\infty) = (-\infty) + \infty = \bot.$$

Multiplication. This is extended as follows: for $p \in \mathbb{R}$,

$$0 \cdot \infty = 0;$$
$$p > 0, \; p \cdot \infty = \infty \text{ and } p \cdot (-\infty) = -\infty;$$
$$p < 0, \; p \cdot \infty = -\infty \text{ and } p \cdot (-\infty) = \infty.$$

Multiplication is taken to be commutative.

Division. This is defined by:

$$\frac{x}{y} = x \cdot \frac{1}{y};$$
$$\frac{1}{0} = \infty;$$
$$\frac{1}{\infty} = \frac{1}{-\infty} = 0.$$

Logarithm. This works for $p \in \mathbb{R}$:

$$p < 0, \; \log_2(p) = \bot;$$
$$\log_2 \infty = \infty;$$
$$\log_2(-\infty) = \bot.$$

Some Properties of Entropic Transreals

The algebra of entropic transreals has several properties that are worth mentioning and are easy to prove:

Proposition 1. *(i)   Addition and multiplication are associative and commutative,*
(ii) $x + 0 = x$,
(iii) $x \cdot 1 = 1$,
(iv) $x \neq \bot \to 0 \cdot x = 0$,
(v) $\frac{x}{y} = x \cdot \frac{1}{y}$,
(vi) $(x \neq \infty \wedge x \neq -\infty) \to x + (-x) = 0 \cdot x$,
(vii) $x + (-x) \neq \bot \to x + (-x) = 0$.

6. Concluding Discussion

The question we have raised is, *Can we design algebras that enlarge the real numbers so that some information theoretic formulae do not require conventions or special conditions to guard against partiality?* Such a question is particularly relevant to computing as such algebras are needed to design data types for programming. That partiality needs to be avoided or managed is essential in programming to avoid unwanted semantic behaviour and enable formal logical and automated tools for reasoning about programs.

Starting with an algebraic structure called the transreals (Section 3.1), a modification has been proposed for our purposes that we have named the entropic transreals. We have shown that the entropic transreals allows us to provide a suitable algebra for real arithmetic so that the formuale that arise in the conventional definitions of entropy and cross-entropy for probability mass functions on finite sample spaces are well defined.

6.1. On Conventions and the 'Legality' of Texts

Following the convention that we refer to an arithmetical expression with division as its leading symbol as a fracterm, we may refer to a expression with $\log_2(-)$ as its leading function symbol as a *logterm*, or when more detail is needed a *logterm with base* 2. Explanations of the definition of entropy often mention the convention that the expression $0 \cdot \log_2 0$ is understood to take value 0 to complete the formula. Upon supposing $0 \cdot \log_2 0 = 0$, we asked what to think of the logterm $\log_2 0$. There seems to be no principled impediment against writing expression that contain $\log_2 0$ as a subterm, which is in remarkable contrast

with the fracterm $\frac{1}{0}$. This textual point is the subject of an investigation in [21] of conventions for notions of legality, where a text about or involving elementary arithmetic is illegal if it makes use of division by zero. Although the logterm $\log_2 0$ makes no more sense than the fracterm $\frac{1}{0}$, both seemingly meaningless expressions are treated rather differently. We have no convincing explanation for such differences.

Entropic transreals demonstrate the consistency of the assumption $0 \cdot \log_2 = 0$ by assigning the value $-\infty$ to $\log_2 0$ and adopting $\frac{1}{-\infty}$. These assumptions hold for transreals. Providing a grounding for the definition of transreals, however, requires the assumption that $0 \cdot \infty = 0$, which leads to a deviation from the design of transreals, leading to what we call entropic transreals.

6.2. Probability Theory in the Context of Entropic Transreals

Entropric transreals are of use when dealing with definitions in connection with entropy. Perhaps the scope of the entropic transreals may extended to become a point of departure for a systematic formal logical analysis of the basics of probability. To illustrate the idea, consider a precise formulation of the well-known *Bayes–Price theorem* on inverse probability, which contains a possible division by zero in the formula.

Using the fracterm calculus of entropic transreals, the equation

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)} \quad (\star)$$

is valid under all conditions. Indeed, if $P(y) = 0$, then $P(x \wedge y) = 0$ and

$$P(x|y) = \frac{P(x \wedge y)}{P(y)} = \frac{0}{0} = 0$$

and if $P(y) \neq 0$, even including the case that $P(y) = \perp$ then $(\star)$ follows trivially.

So, how division by zero is handled generates conditions that need to be imposed on the formula. These can depend upon the details of the fracterm calculus that is used. For instance, in Suppes-Ono arithmetic (i.e., working with $\frac{x}{0} = 0$), the equation $(\star)$ can be stated without conditions such as $P(x) \neq 0$ and/or $P(y) \neq 0$. For an application of Suppes-Ono fracterm calculus to probability theory, we refer to [22]. In fact, the results of [22] can be developed almost without modification when making use of entropic transreals instead of reals with Suppes-Ono division.

6.3. Potential Applications

Entropic transreals may be useful when formalizing calculations and proofs involving entropy with an eye on subsequent automated proof checking. We feel that any systematic approach to logical reasoning about entropy and cross-entropy requires a level of precision concerning the status of problematic expressions beyond the nowadays customary conventions.

Outside probability theory, entropic transreals may be helpful for designing alternative approaches to floating-point arithmetic just as transreals may be helpful for that purpose. When understanding $\pm\infty$ as overflow values, the axiom/convention of entropic transreals that $0 \cdot \infty = 0 \cdot -\infty = 0$ is at least as plausible as the assumption $0 \cdot \infty = 0 \cdot -\infty = \Phi$ of transreal arithmetic which models the currently popular floating-point standard.

In the introduction, we qualify certain conventions as used in many expositions on entropy and cross-entropy as ad hoc. Undeniably, some design decisions made for the design of entropic transreals may also be qualified as being ad hoc. The difference is, however, that the latter design decisions are made in the systematic framework of the design of abstract datatypes, i.e., algebra in the framework of universal algebra.

References

1. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2005.
2. Ehrich, H.-D.; Wolf, M.; Loeckx, J. *Specification of Abstract Data Types*; Wiley: Hoboken, NJ, USA, 1997.
3. Bergstra, J.A. Arithmetical datatypes, fracterms, and the fraction definition problem. *Transmathematica* **2020**. [CrossRef]
4. Anderson, J.A.; Bergstra, J.A. Review of Suppes 1957 proposals for division by zero. *Transmathematica* **2021**. [CrossRef]
5. Bergstra, J.A.; Tucker, J.V. The rational numbers as an abstract data type. *J. ACM* **2007**, *54*, 7. [CrossRef]
6. Okumura, H.; Saitoh, S.; Matsuura, T. Relations of zero and ∞. *J. Technol. Soc. Sci.* **2017**, *1*, 70–77.
7. Ono, H. Equational theories and universal theories of fields. *J. Math. Soc. Jpn.* **1983**, *35*, 289–306. [CrossRef]
8. Suppes, P. *Introduction to Logic*; Van Nostrand Reinhold Company: New York, NY, USA, 1957.
9. Bergstra, J.A.; Ponse, A. Division by zero in common meadows. In *Software, Services, and Systems (Wirsing Festschrift)*; de Nicola, R., Hennicker, R., Eds.; Lecture Notes in Computer Science 8950; Springer: Cham, Switzerland, 2015; pp. 46–61.
10. Bergstra, J.A.; Tucker, J.V. On the axioms of common meadows: Fracterm calculus, flattening and incompleteness. *Comput. J.* **2023**, *66*, 1565–1572. [CrossRef]
11. Bergstra, J.A.; Tucker, J.V. Synthetic fracterm calculus. *J. Univers. Comput. Sci.* **2024**, *30*, 289–307. [CrossRef]
12. Anderson, J.A. Perspex Machine IX: Transreal analysis. In *Vision Geometry XV, Proceedings of SPIE, Electronic Imaging 2007, San Jose, CA, USA, 28 January–1 February 2007*; SPIE: Cergy-Pontoise, France, 2007; Volume 6499. Available online: http://www.bookofparagon.com/Mathematics/PerspexMachineIX.pdf (accessed on 10 September 2024).
13. Anderson, J.A.; Völker, N.; Adams, A.A. 2007. Perspecx Machine VIII, axioms of transreal arithmetic. In *Vision Geometry XV, Proceedings of SPIE, Electronic Imaging 2007, San Jose, CA, USA, 28 January–1 February 2007*; Latecki, J., Mount, D.M., Wu, A.Y., Eds.; SPIE: Cergy-Pontoise, France, 2007; Volume 6499.
14. Anderson, J.A. Transreal Foundation for Floating-Point Arithmetic. *Transmathematica* **2023**. . [CrossRef]
15. Bergstra, J.A.; Tucker, J.V. The transrational numbers as an abstract data type. *Transmathematica* **2020**. [CrossRef]
16. Dos Reis, T.S.; Gomide, W.; Anderson, J.A. Construction of the transreal numbers and algebraic transfields. *IAENG Int. J. Appl. Math.* **2016**, *46*, 11–23. Available online: http://www.iaeng.org/IJAM/issues_v46/issue_1/IJAM_46_1_03.pdf (accessed on 30 December 2024).
17. Dos Reis, T.S. Transreal integral. *Transmathematica* **2019**. [CrossRef]
18. Bergstra, J.A.; Tucker, J.V. Symmetric transrationals: The data type and the algorithmic degree of its equational theory. In *A Journey From Process Algebra via Timed Automata to Model Learning—A Festschrift Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday*; Jansen, N., Stoelinga, M., van den Bos, P., Eds.; Lecture Notes in Computer Science 13560; Springer: Cham, Switzerland, 2022; pp. 63–80. [CrossRef]
19. Carlström, J. Wheels—On division by zero. *Math. Struct. Comput. Sci.* **2004**, *14*, 143–184. https://10.1017/S0960129503004110. [CrossRef]
20. Taboga, M. Zero-probability events. In *Lectures on Probability Theory and Mathematical Statistics*; Online appendix; Kindle Direct Publishing: Seattle, WA, USA, 2021. Available online: https://www.statlect.com/fundamentals-of-probability/zero-probability-events (accessed on 26 November 2024).
21. Bergstra, J.A.; Tucker, J.V. Logical models of mathematical texts: The case of conventions for division by zero. *J. Logic Lang. Inf.* **2024**, *33*, 277–298. [CrossRef]
22. Bergstra, J.A. Adams conditioning and likelihood ratio transfer mediated inference. *Sci. Ann. Comput. Sci.* **2019**, *29*, 1–58. [CrossRef]