Reasoning about Additive Measures

Bassem Sayrafi and Dirk Van Gucht *

{bsayrafi,vgucht}@cs.indiana.edu
Computer Science Department,Indiana University,
Bloomington, IN 47405-4101, USA

Abstract. We establish a link between measures and certain types of inference systems and we illustrate this connection on examples that occur in computing applications, especially in the areas of databases and data mining.

1 Introduction

The main contribution of our paper is the establishment of a link between set-based additive measures and certain types of inference systems. To show the applicability of our result, we apply it to particular measures, especially some that occur in the areas of of databases and data mining. Our work significantly generalizes that of Malvestuto [10], Lee [13], and Dalkilic and Robertson [5], where it was shown how Shannon's entropy measure [11] can be used to derive inference systems for functional and multivalued dependencies in relational databases [6].

Our measure framework can be used to find evidence of presence or absence of relationships (possibly causal). For example, if $\mathcal{M}$ is a measure, and X and Y are sets, then the quantity $\mathcal{M}(X \cup Y) - \mathcal{M}(X)$, i.e., *the rate of change of* $\mathcal{M}$ *in going from* X *to* $X \cup Y$, plays a crucial role in this regard. Depending on its value, this rate can capture interesting relationships. For example, when this rate is 0, it can interpreted as "X fully determines Y according to $\mathcal{M}$", and if it is $\mathcal{M}(Y)$, it can be interpreted as X *and* Y *are independent according to* $\mathcal{M}$.

As a simple, motivating example consider the cardinality measure $|.|$ defined over all subsets of some set S. The cardinality measure has some important properties: for all X, Y, and Z subsets of S, it holds that

$$\begin{array}{ll} |X| \leq |X \cup Y| & \textbf{isotonicity}, \text{ and} \\ |X \cup Y \cup Z| + |X| \leq |X \cup Y| + |X \cup Z| & \textbf{subadditivity}. \end{array}$$

From these properties follow some others. For example, we can deduce the following "transitivity" property:

$$(|X \cup Y| - |X|) + (|Y \cup Z| - |Y|) \geq (|X \cup Z| - |Z|). \tag{1}$$

Given this, we can consider constraints on cardinalities. For example, the constraint $|X| = |X \cup Y|$ states that $X \supseteq Y$. Well-known inference rules for set

* The authors were supported by NSF Grant IIS-0082407.

containment can then be derived from the rules about the cardinality measure. For example, if the constraints $|X \cup Y| = |X|$ and $|Y \cup Z| = |Z|$ are true, then, by the transitivity and the isotonicity rules, $|X \cup Z| = |Z|$. A simpler way of writing this is an inference rule about the set-inclusion relation:

$$\frac{X \supseteq Y \ \& \ Y \supseteq Z}{X \supseteq Z}$$

The paper is organized into several sections. In Section 2, we introduce additive measures and give examples. In Section 3, we introduce finite differentials for such measures and study the properties of these differentials. In Section 4, we introduce measure constraints and derive inference systems for these constraints from the rules of differentials. We illustrate our approach by deriving some specific inference systems from measures. Finally, in Section 5, we establish a duality between measures and differentials similar to the one that exists between integrals and derivatives in calculus.

2 Additive measures

In this section, we define additive measures. We then give several examples of such measures that occur in practice.

In the rest of the paper, S denotes a finite set, $\mathcal{S}$ denotes 2^S, U, V, X, Y, and Z (possibly subscripted) denote subsets of S, $\mathcal{Y}$ and $\mathcal{Z}$ denote subsets of $\mathcal{S}$ and $\mathcal{M}$ denotes a real-valued function over $\mathcal{S}$. Furthermore, we use the following abbreviations:

$$\begin{array}{ll} XY & = X \cup Y; \\ X \cdot \mathcal{Y} & = \{XY \mid Y \in \mathcal{Y}\}; \\ \sqcup\mathcal{Y} & = \bigcup_{Y \in \mathcal{Y}} Y; \\ \sqcap\mathcal{Y} & = \bigcap_{Y \in \mathcal{Y}} Y; \\ \mathcal{Y}[Y \leftarrow Z] & = \mathcal{Y} - \{Y\} \cup \{Z\}. \end{array}$$

Our definitions for measures are inspired by the inclusion-exclusion principle for counting finite sets.[1]. In light of this, we define the following function $\mathcal{D}$:

Definition 1. *Let f be a function from $\mathcal{S}$ into the reals, let $X \subseteq S$, and let $\mathcal{Y}$ be subset of $\mathcal{S}$. Then the function $\mathcal{D}_f$ at X and $\mathcal{Y}$ is defined as follows:*

$$\mathcal{D}_f(X, \mathcal{Y}) = \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \mathtt{odd}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \mathtt{even}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}). \tag{2}$$

We illustrate this definition in the following table.

X	$\mathcal{Y}$	$\mathcal{D}_f(X,\mathcal{Y})$
X	$\emptyset$	$-f(X)$
X	$\{Y\}$	$f(XY) - f(X)$
X	$\{Y_1, Y_2\}$	$f(XY_1) + f(XY_2) - f(XY_1Y_2) - f(X)$
X	$\{Y_1, Y_2, Y_3\}$	$f(XY_1) + f(XY_2) + f(XY_3) + f(XY_1Y_2Y_3)$ $-f(XY_1Y2) - f(XY_1Y_3) - f(XY_2Y_3) - f(X)$
$Y_1 \cap Y_2$	$\{Y_1, Y_2\}$	$f(Y_1) + f(Y_2) - f(Y_1Y_2) - f(Y_1 \cap Y_2)$

With the use of the function $\mathcal{D}$, we can now define subadditive and superadditive measures.

Definition 2. *Let S be a finite set, let $\mathcal{M}$ be a function from $\mathcal{S}$ into the reals, and let n be a positive natural number. $\mathcal{M}$ is called a* n-subadditive *(n-*superadditive*)* measure *if for each $X \subseteq S$, and each nonempty set $\mathcal{Y}$ of subsets of S, with $|\mathcal{Y}| \leq n$, $\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y}) \geq 0$ ($\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y}) \leq 0$, respectively).*

Example 1. Mathematical measures [3] are n-subadditive for each $n \geq 1$. For such measures, $\mathcal{M}(\emptyset) = 0$. When also $\mathcal{M}(S) = 1$ these measures are called probability measures.

The following proposition, the proof of which is straightforward, relates isotone with anti-isotone functions, and subadditive measures with superadditive measures. This proposition allows us to focus on subadditive measures.

Proposition 1. *Let $\mathcal{M}$ be a function from $\mathcal{S}$ into the reals and define $\overline{\mathcal{M}}$ (also a function from $\mathcal{S}$ into the reals) as follows:*

$$\overline{\mathcal{M}}(X) = [\mathcal{M}(S) - \mathcal{M}(X)] + \mathcal{M}(\emptyset).^1 \tag{3}$$

$\mathcal{M}$ is an n-subadditive measure if and only if $\overline{\mathcal{M}}$ is an n-superadditive measure.

2.1 Frequently used measures

In this subsection, we describe a variety of application areas in databases and data mining where measures occur naturally. We identify these measures and fit them in the our measures framework. In the area of databases, we consider aggregate functions and relational data-uniformity measures. In the area of data mining, we focus on measures that occur in the context of the item sets problems.

[1] Notice that $\overline{\mathcal{M}}(S) = \mathcal{M}(\emptyset)$ and $\overline{\mathcal{M}}(\emptyset) = \mathcal{M}(S)$.

Databases - aggregation functions Computations requiring aggregate functions occur frequently in database applications such as query processing, datacubes [8], and spreadsheets. Among these, the most often used are `count`, `sum`, `min`, `max`, `avg`, `variance`, *order statistics*, and `median`. Each of these functions operates on finite sets (`count` on arbitrary finite sets, and the others on finite sets of (nonnegative)[2] numbers) and each returns a nonnegative number. Thus they are measures. We elaborate on how they fit precisely in our framework.

1. Define `count(X)` to be the cardinality of X. From the inclusion-exclusion principle, it follows that `count` is n-subadditive for each $n \geq 1$. (Similar reasoning demonstrates that `sum` is n-subadditive for all $n \geq 1$.)
2. Let S consist of positive integers. Define `max(X)` to be equal to the largest integer in X, for $X \neq \emptyset$, and `max`$(\emptyset)$ to be equal to the smallest element in S. Then `max` is an n-subadditive measure for $n \geq 1$. The key to showing that `max` is n-subadditive for $n \geq 1$ is the observation that `max`$(\mathcal{Y}) = maximum(Y)$ for some set $Y \in \mathcal{Y}$. (Similar reasoning demonstrates that `min` is n-superadditive for all $n \geq 1$.)
3. Let S consist of positive integers. Order-statistics are used to determine the i^{th} smallest element of S. For example, the 2^{nd} order statistics, denoted `min2`(X), returns the second smallest element in X. Clearly, `min2` is 1-superadditive. However, it is not 2-superadditive (e.g. let $Y_1 = \{1, 4, 5\}$, $Y_2 = \{2, 4, 5\}$ and $X = Y_1 \cap Y_2$).
4. The functions `avg`, `variance`, and `median` are neither n-subadditive nor n-superadditive for any $n \geq 1$. However, observe that in the case of `avg` both the numerator and the denominator come from n-subadditive measures (`sum` and `count`, respectively). It follows that the quotient of two subadditive measures is not necessarily a subadditive measure.

Databases - data uniformity Consider the values occurring under an attribute of a relation in a relational database. These values can occur uniformly (e.g. the values 'male' and 'female' in the gender attribute of a census), or skewed (e.g. the values for the profession attribute in the same census). Measuring these degrees of uniformity can influence how data is stored or processed. When data is numeric, a common way to measure uniformity is to use the variance statistic. This statistic computes the average of the distances between data values and their average. To measure data uniformity for categorical data we consider the Simpson measure [12], and the Shannon entropy measure [11]. Unlike variance, these measures are specified in terms of probability distributions defined over the data sets. We show that, unlike variance (Section 2.1), the Shannon measure is n-subadditive for $n \leq 2$ and the Simpson measure is n-subadditive for $n \geq 1$.

Let T be a nonempty finite relation over the relation schema S and let p be a probability distribution over T. For $X \subseteq S$, define p_X to be the marginal probability distribution of p on X. Thus if $x \in \Pi_X(T)$ then $p_X(x) = \sum \{t \in T \mid t[X] = x\} p(t)$.

[2] We restrict ourselves in this paper to nonnegative numbers, but it is straightforward to adapt our framework to include negative numbers as well.

The *Simpson* measure $\mathcal{S}$ and the *Shannon* measure $\mathcal{H}$ are defined as follows:[3]

$$\mathcal{S}(X) = \sum_{x \in \Pi_X(T)} p_X(x)(1 - p_X(x)) = 1 - \sum_{x \in \Pi_X(T)} p_X^2(x), \tag{4}$$

$$\mathcal{H}(X) = - \sum_{x \in \Pi_X(T)} p_X(x) \log p_X(x). \tag{5}$$

Proposition 2. *The Simpson measure ($\mathcal{S}$) is an n-subadditive measure for all $n \geq 1$ while the Shannon entropy measure ($\mathcal{H}$) is only 2-subadditive measure.*

Proof. We prove that the Simpson measure is n-subadditive for $n \geq 1$. To prove n-subadditivity, let $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$, and $\boldsymbol{Y}$ be a bit vector of length n such that when $\boldsymbol{Y}[i] = 1$, we have $Y_i = Y_i'$, and when $\boldsymbol{Y}[i] = 0$ we have either $Y_i = Y_i'$ or $Y_i \neq Y_i'$ (in this way the 0 acts very much like a wildcard $*$). We define $Terms(\boldsymbol{Y}) = \{(P_{Y_1 \ldots Y_n} P_{Y_1' \ldots Y_n'}) | \forall i \; Y_i = Y_i' \; when \; \boldsymbol{Y}[i] = 1\}$.

The generalized inclusion-exclusion principle [9] states

$$\begin{aligned} \bigcap_{k=1}^{n} W(\overline{Terms(\boldsymbol{Y_k})}) = & \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ even \geq 2}} \bigcap_{j \in \mathcal{Z}} W(Terms(\boldsymbol{Y_j})) + W(Terms(\boldsymbol{0})) \\ & - \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ odd}} \bigcap_{j \in \mathcal{Z}} W(Terms(\boldsymbol{Y_j})) \end{aligned} \tag{6}$$

where $W(Terms(\boldsymbol{Y}))$ is some weight function and $\boldsymbol{Y_j}$ implies only $Y_j = Y_j'$. If we choose $W(Terms(\boldsymbol{Y})) = Terms(\boldsymbol{Y})$ and rearrange terms, then we have

$$Terms(\boldsymbol{0}) = \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ odd}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) - \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ even \geq 2}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) + \overline{\bigcup_{k=1}^{n} Terms(\boldsymbol{Y_k})}$$

which after simplification yields

$$\bigcup_{k=1}^{n} Terms(\boldsymbol{Y_k}) = \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ odd}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) - \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ even \geq 2}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) \tag{7}$$

And since the more zero bits in a vector means more wildcards, then the more zero bits a vector $\boldsymbol{Y}$ has, the larger it is; that is $Terms(\boldsymbol{0}) \geq Terms(\boldsymbol{Y})$. Given this and 7, we have

$$Terms(\boldsymbol{0}) \geq \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ odd}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) - \sum_{\substack{\mathcal{Z} \subseteq \{1, \ldots, n\} \\ even \geq 2}} \bigcap_{j \in \mathcal{Z}} Terms(\boldsymbol{Y_j}) \tag{8}$$

[3] In ecology, $\mathcal{S}$ is known as the Simpson rarity function.

Finally we expand this equation to the original probabilities to get.

$$[\sum_{Y_1} \cdots \sum_{Y_n} P_{Y_1 \ldots Y_n}]^2 \geq \sum_{\substack{\mathcal{Z} \subseteq \{Y_1, \ldots Y_n\} \\ odd}} \sum_{Y_{i_1}} \cdots \sum_{Y_{i_{|Z|}}} [P_{Y_{i_1} \ldots Y_{i_{|Z|}}}]^2 - \sum_{\substack{\mathcal{Z} \subseteq \{Y_1, \ldots Y_n\} \\ even \geq 2}} \sum_{Y_{i_1}} \cdots \sum_{Y_{i_{|Z|}}} [P_{Y_{i_1} \ldots Y_{i_{|Z|}}}]^2$$, which

can be further simplified to yield

$$\sum_{\substack{\mathcal{Z} \subseteq \{Y_1, \ldots Y_n\} \\ even}} \sum_{Y_{i_1}} \cdots \sum_{Y_{i_{|Z|}}} [P_{Y_{i_1} \ldots Y_{i_{|Z|}}}]^2 \geq \sum_{\substack{\mathcal{Z} \subseteq \{Y_1, \ldots Y_n\} \\ odd}} \sum_{Y_{i_1}} \cdots \sum_{Y_{i_{|Z|}}} [P_{Y_{i_1} \ldots Y_{i_{|Z|}}}]^2$$, which yields

$$\sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \texttt{odd}(\mathcal{Z})}} \mathcal{S}(\sqcap \mathcal{Y} \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \texttt{even}(\mathcal{Z})}} \mathcal{S}(\sqcap \mathcal{Y} \sqcup \mathcal{Z}) \geq 0.$$

The Shannon Entropy measure ($\mathcal{H}$) is a 2-subadditive measure [4]. However, $\mathcal{H}$ is not a 3-subadditive measure. Indeed, for the following relation over attributes A, B, C, $\mathcal{D}_{\mathcal{H}}(\emptyset, \{\{A\}, \{B\}, \{C\}\}) < 0$.

A	B	C
1	1	1
1	1	2
1	2	1
2	1	1

□

Data mining - frequent item sets

An important problem in data mining is discovering frequent item sets. In this problem, a set of baskets is given. Each basket contains a set of items. In practice, the items may be products sold at a grocery store, and baskets correspond to items bought together by customers. The frequent items sets problem is to find the item sets that occur frequently within the baskets.

More formally, let S be a set of items and let $\mathcal{B}$ be a subset of $\mathcal{S}$ consisting of the baskets. Define $\mathcal{B}(X) = \{B \mid X \subseteq B \text{ and } B \in \mathcal{B}\}$ and define the *frequency* measure `freq` as $\texttt{freq}(X) = \frac{|\mathcal{B}(X)|}{|\mathcal{B}|}$. It can be shown that `freq` is an n-superadditive measure for $n \geq 1$ [2].

3 Measure Differentials

Some natural issues that arise for measures is (1) to calculate their rate of change and (2) to determine where these rate changes reach optima. Typically, these issues are considered for functions over continuous domains by using traditional calculus techniques, in particular *derivatives*. In our framework for additive measures, we have discrete, set-based functions, and thus reasoning about derivatives must be done with the methods of finite differences and finite difference equations [7].

Definition 3. *Let f be a function from $\mathcal{S}$ into the reals, let X be a subset of S, let $\mathcal{Y}$ be a subset of $\mathcal{S}$, and let Y be in $\mathcal{Y}$. We define the* finite difference *of f at X relative to $\mathcal{Y}$ as follows:*

$$\Delta_f(X,\mathcal{Y}) = f(X) \text{ if } \mathcal{Y} = \emptyset, \tag{9}$$

and

$$\Delta_f(X,\mathcal{Y}) = \Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\}) \text{ otherwise.} \tag{10}$$

Notice that the definition is dependent on the choice for Y in $\mathcal{Y}$. We will show however that each possible choice of Y leads to the same result, i.e., $\Delta_f(X,\mathcal{Y})$ is well defined.

Proposition 3. *Let f be a function from $\mathcal{S}$ into the reals. Then, for each $X \subseteq S$ and for each nonempty set $\mathcal{Y} \subseteq \mathcal{S}$, $\Delta_f(X,\mathcal{Y})$ is well-defined.*

Proof. Trivially, $\Delta_f(X,\mathcal{Y})$ is well-defined when $0 \leq |\mathcal{Y}| \leq 1$. When $|\mathcal{Y}| \geq 2$, $\mathcal{Y}$ contains two different sets Y and Y'. We need to show $\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\}) = \Delta_f(XY',\mathcal{Y}-\{Y'\}) - \Delta_f(X,\mathcal{Y}-\{Y'\})$. We show this by induction on $|\mathcal{Y}|$.

1. When $|\mathcal{Y}| = 2$ this equation becomes

   $$\Delta_f(XY,\{Y'\}) - \Delta_f(X,\{Y'\}) = \Delta_f(XY',\{Y\}) - \Delta_f(X,\{Y\}).$$

   Further expansion leads to the equation $f(XYY') - f(XY) - f(XY') + f(X) = f(XY'Y) - f(XY') - f(XY) + f(X)$ which is clearly true.
2. When $|\mathcal{Y}| \geq 3$, by induction, we are allowed to expand the left hand side of the equation, i.e., the expression $\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\})$, into the expression $\Delta_f(XYY',\mathcal{Y}-\{Y,Y'\}) - \Delta_f(XY,\mathcal{Y}-\{Y,Y'\}) - \Delta_f(XY',\mathcal{Y}-\{Y,Y'\}) + \Delta_f(X,\mathcal{Y}-\{Y,Y'\})$. Similarly, the righthand of the equation can be expanded to expression $\Delta_f(XY'Y,\mathcal{Y}-\{Y',Y\}) - \Delta_f(XY',\mathcal{Y}-\{Y',Y\}) - \Delta_f(XY,\mathcal{Y}-\{Y',Y\}) + \Delta_f(X,\mathcal{Y}-\{Y',Y\})$. Clearly both expressions are equal.

□

It turns out that the functions $\mathcal{D}$ and Δ are closely related:

Proposition 4. *Let f be a function from $\mathcal{S}$ into the reals. Then for each $X \subseteq S$ and for each nonempty set $\mathcal{Y} \subseteq \mathcal{S}$*

$$\mathcal{D}_f(X,\mathcal{Y}) = (-1)^{|\mathcal{Y}|-1}\Delta_f(X,\mathcal{Y}). \tag{11}$$

Proof. The proof is by induction on $|\mathcal{Y}|$. For $\mathcal{Y} = \emptyset$, we have $\mathcal{D}_f(X,\emptyset) = -f(X) = -\Delta_f(X,\emptyset)$. For $\mathcal{Y} = \{Y\}$, we have $\mathcal{D}_f(X,\{Y\}) = f(XY) - f(X) = \Delta_f(X,\mathcal{Y})$, and the claim follows.
For $|\mathcal{Y}| \geq 2$, and $Y \in \mathcal{Y}$, we have by the definition of Δ

$$(-1)^{|\mathcal{Y}|-1}\Delta_f(X,\mathcal{Y}) = (-1)(-1)^{|\mathcal{Y}|-2}(\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\})),$$

which, by induction, is equal to

$$(-1)(\mathcal{D}_f(XY, \mathcal{Y} - \{Y\}) - \mathcal{D}_f(X, \mathcal{Y} - \{Y\})).$$

By the definition of $\mathcal{D}$, we have that $\mathcal{D}_f(X, \mathcal{Y}-\{Y\})-\mathcal{D}_f(XY, \mathcal{Y}-\{Y\})$ is equal to
$\sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{odd}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{even}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - (\sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{odd}(\mathcal{Z})}} f(XY \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{even}(\mathcal{Z})}} f(XY \sqcup \mathcal{Z}))$ which, after rearranging terms and realizing that $|\mathcal{Z}|$ is even if and only if $|\mathcal{Z} \cup \{Y\}|$ is odd, is equal to
$\sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{odd}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) + \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{even}(\mathcal{Z})}} f(XY \sqcup \mathcal{Z})$
$- \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{even}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} - \{Y\} \\ \mathtt{odd}(\mathcal{Z})}} f(XY \sqcup \mathcal{Z})$
This is equal to $\sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \mathtt{odd}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \mathtt{even}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) = \mathcal{D}_f(X, \mathcal{Y})$. □

In the following proposition we summarize some important properties of $\mathcal{D}$. These properties are specified as equalities and inequalties, but it is more useful here to view them as inference rules.

Proposition 5. *Let $\mathcal{M}$ be an n-subadditive measure ($n \geq 1$). Let $\mathcal{Y}$ be a subset of $\mathcal{S}$. Then $\mathcal{D}_{\mathcal{M}}$ satisfies following properties:*

$$\frac{1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq 0} \quad \textbf{sign rule;}$$

$$\frac{Y \in \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y} - \{Y\})} \quad \textbf{reduction.}$$

When $\mathcal{M}$ is an n-superadditive measure, the reduction rule remain valid. The sign rule however needs to be altered by replacing $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq 0$ with $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \leq 0$.

Proof. The sign rule follows from the fact we always define $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq 0$ for $1 \leq |\mathcal{Y}| \leq n$. Reduction follows from (10) and (11). □

Using Proposition 5, we derive interesting rules about measure differentials in the next proposition.

Proposition 6. *Let $\mathcal{M}$ be an n-subadditive measure (when $\mathcal{M}$ is n-superadditive, the inequalities change direction) and $n \geq 1$. Let $\mathcal{Y}$ be a set of subsets of $\mathcal{S}$ such that $0 \leq |\mathcal{Y}| \leq n$. Then the rules display in Figure 1 follow from Proposition 5.*

$$\frac{Y \in \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow YZ]) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y}[Y \leftarrow Z])}$$ **chain rule;**

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0}$$ **triviality;**

$$\frac{Y \in \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow YZ]) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y})}$$ **decomposition;**

$$\frac{Y \in \mathcal{Y} \qquad U \subseteq X}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow UY])}$$ **right augmentation;**

$$\frac{(Y \in \mathcal{Y} \ \wedge \ U \subseteq XY) \ or \ (U \subseteq X)}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})}$$ **weak left augmentation;**

$$\frac{0 \leq |\mathcal{Y}| < n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})}$$ **left augmentation;**

$$\frac{X \subseteq Z \qquad |\mathcal{Y}| = 1}{\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \{Z\}) \geq \mathcal{D}_{\mathcal{M}}(X, \{Z\})}$$ **weak transitivity (a);**

$$\frac{Y \in \mathcal{Y} \qquad Y' \in \mathcal{Y}}{\begin{array}{l}\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y' \leftarrow XY', Y \leftarrow Z]) \\ \quad \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])\end{array}}$$ **weak transitivity (b);**

$$\frac{Y \in \mathcal{Y} \qquad 1 \leq |\mathcal{Y}| < n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y \leftarrow Z]) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])}$$ **transitivity**;

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y})}$$ **replication**;

$$\frac{Y \in \mathcal{Y} \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\})}$$ **coalescence**.

Fig. 1. Additional rules for $\mathcal{D}$

Proof. We provide a sketch of how these rules can be proved as follows. The chain rule follows directly from reduction and then simplifying. Triviality follows directly from reduction on $Y \subseteq X$. Decomposition follows directly from the chain rule (on $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow YZ])$) and the sign rule.

Right augmentation follows by applying reduction on Y for the first term, UY for the second term, and then using the fact $U \subseteq X$ to simplify. Weak left augmentation follows by applying the general chain rule (identify Y with U, Z with Y) and the sign rule. For $\mathcal{Y} = \emptyset$, weak left augmentation follows from the fact $U \subseteq X$ (thus $\mathcal{D}_{\mathcal{M}}(X, \emptyset) - \mathcal{D}_{\mathcal{M}}(XU, \emptyset) = 0$). Left augmentation follows by using reduction to combine terms and finally using the sign rule at level n.

Weak transitivity can be proved by using weak left augmentation and the chain rule. By applying right augmentation on the first term, and using weak left augmentation on the second term, we get for $|\mathcal{Y}| \geq 2$, $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y \leftarrow Z, Y' \leftarrow XY']) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y' \leftarrow XY']) + \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y}[Y \leftarrow Z, Y' \leftarrow XY'])$. Which when we apply the chain rule is equal to $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow ZY, Y' \leftarrow Y'X)$ which is greater or equal than $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])$ by decomposition. For $|\mathcal{Y}| = 1$, weak transitivity can be proved by weak left augmentation, chain rule and decomposition. By weak left augmentation we have $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \{Z\}) \geq \mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(XY, \{Z\})$. Using the chain rule, this is equal to $\mathcal{D}_{\mathcal{M}}(X, \{YZ\})$ which by decomposition is greater or equal to $\mathcal{D}_{\mathcal{M}}(X, \{Z\})$.

Transitivity can be proved along the lines of weak transitivity by using left augmentation and the general chain rule. Replication can be proved directly from reduction (which imposes the restriction $|\mathcal{Y} - \{Y\}| \geq 1$) and the sign rule (to remove the last term). Coalescence can be proved by using reduction and left augmentation. $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\}) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) - \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y} - \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\})$. Since $\mathcal{M}$ is n-subadditive, then by left augmentation we have $\mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y} - \{Y\}) \leq \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\})$ which implies coalescence. □

An interesting special case to consider is when $\mathcal{S} = X \sqcup \mathcal{Y}$. In this case, some of the rules of Proposition 6 collapse and we obtain the following proposition.

Proposition 7. *Let $\mathcal{M}$ be an n-subadditive measure (when $\mathcal{M}$ is n-superadditive, the inequalities change direction) and $n \geq 1$. Let $\mathcal{S} = X \sqcup \mathcal{Y}$ or $\mathcal{S} = Z \sqcup \mathcal{Y}$. Then the rules display in Figure 2 follow from Proposition 6.*

Proof. $R1$ derived directly from the chain rule and triviality. $R2$ is the triviality rule proved in Proposition 6. $R3$ can be proved by weak augmentation of Proposition 6 by dividing U into partitions such that $U = (U \cap XY_1) \cup \ldots \cup (U \cap XY_n)$. Using weak augmentation, $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(X \cup (U \cap XY_1), \mathcal{Y}) \geq \ldots \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})$. $R4$, $R5$ are nothing but the weak transitivity rule proved in Proposition 6. $R6$, $R7$ are the replication and coalescence rules derived in Proposition 6. □

4 Measure Constraints

In this section, we consider the situations wherein measure differentials are minimized. In particular, for context-subadditive (context-superadditive) measures,

$$\frac{Y \in \mathcal{Y} \qquad \mathcal{S} = X \sqcup \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow YZ]) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y})} \quad R1 \text{ - } \textbf{chain rule;}$$

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0} \quad R2 \text{ - } \textbf{triviality;}$$

$$\frac{U \subseteq \mathcal{S} = X \sqcup \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})} \quad R3 \text{ - } \textbf{augmentation;}$$

$$\frac{Y \in \mathcal{Y} \qquad X \subseteq \mathcal{S} = Z \sqcup \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y \leftarrow Z]) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])} \quad R4 \text{ - } \textbf{transitivity (a);}$$

$$\frac{Y, Y' \in \mathcal{Y} \qquad \mathcal{S} = XZ \sqcup \mathcal{Y}}{\begin{array}{l}\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y' \leftarrow XY', Y \leftarrow Z]) \\ \quad \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])\end{array}} \quad R5 \text{ - } \textbf{transitivity (b);}$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y})} \quad R6 \text{ - } \textbf{replication;}$$

$$\frac{Y \in \mathcal{Y} \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\})} \quad R7 \text{ - } \textbf{coalescence.}$$

Fig. 2. Additional rules for $\mathcal{D}$ when $\mathcal{S} = X \sqcup \mathcal{Y}$

we consider when $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$ for $0 \leq |\mathcal{Y}| \leq n$. This leads us to introduce *level-n constraints* and to derive inference rules for them. By applying these results to particular measures, we uncover certain classes of constraints in databases and data mining, as well as corresponding inference systems.

Definition 4. Let $\mathcal{M}$ be an n-subadditive (n-superadditive) measure. We call $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$ for $0 \leq |\mathcal{Y}| \leq n$ a *level-n constraint* and we say that $\mathcal{M}$ satisfies $X \Rightarrow \mathcal{Y}$ if $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$.

It turns out that Definition 4 and Propositions 5,6 yield the inference rules for level-n constraints. These rules are a direct consequence of the rules in Propositions 5-6 although care must be taken regarding rules when $\mathcal{Y} = \emptyset$.

Proposition 8. *Let $\mathcal{M}$ be an n-subadditive (n-superadditive) measure. Let $\mathcal{Y}$ be a set of subsets of S such that $0 \leq |\mathcal{Y}| \leq n$, and $Z, U \subseteq S$. Then the level-n constraint of $\mathcal{M}$ satisfies the inequalities in Figure 3.*

Proof. The proof of these rules follows directly from Propositions 5,6.

As a side note, in instances where $\mathcal{Y} = \emptyset$ is possible, care must be taken to deduce these inference rules. We show for example how this applies to the reduction rule and coalescence when $\mathcal{Y} - \{Y\} = \emptyset$. For reduction, we have $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(XY, \emptyset) = \mathcal{D}_{\mathcal{M}}(X, \emptyset)$. Since the left hand side of the equation

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y}[Y \leftarrow Z]}{X \Rightarrow \mathcal{Y}[Y \leftarrow ZY]}$$ **chain rule (a)**;

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}[Y \leftarrow ZY]}{X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y}[Y \leftarrow Z]}$$ **chain rule (b)**;

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}}$$ **reduction.**

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{X \Rightarrow \mathcal{Y}}$$ **triviality**;

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}[Y \leftarrow YZ]}{X \Rightarrow \mathcal{Y}}$$ **decomposition**;

$$\frac{Y \in \mathcal{Y} \qquad U \subseteq X \qquad X \Rightarrow \mathcal{Y}}{X \Rightarrow \mathcal{Y}[Y \leftarrow UY]}$$ **right augmentation (a);**

$$\frac{Y \in \mathcal{Y} \qquad U \subseteq X \qquad X \Rightarrow \mathcal{Y}[Y \leftarrow UY]}{X \Rightarrow \mathcal{Y}}$$ **right augmentation (b);**

$$\frac{(U \subseteq X \cup Y \wedge Y \in \mathcal{Y}) \; or \; (U \subseteq X) \qquad X \Rightarrow \mathcal{Y}}{XU \Rightarrow \mathcal{Y}}$$ **weak left augmentation**;

$$\frac{1 \leq |\mathcal{Y}| < n \qquad X \Rightarrow \mathcal{Y}}{XU \Rightarrow \mathcal{Y}}$$ **left augmentation**;

$$\frac{X \subseteq Z \qquad X \Rightarrow \{Y\} \qquad Y \Rightarrow \{Z\}}{X \Rightarrow \{Z\}}$$ **weak transitivity (a)**;

$$\frac{Y, Y' \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y}[Y \leftarrow Z, Y' \leftarrow XY']}{X \Rightarrow \mathcal{Y}[Y \leftarrow Z]}$$ **weak transitivity (b)**;

$$\frac{Y \in \mathcal{Y} \qquad 1 \leq |\mathcal{Y}| < n \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y}[Y \leftarrow Z]}{X \Rightarrow \mathcal{Y}[Y \leftarrow Z]}$$ **transitivity**;

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y}}$$ **replication**;

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}}$$ **coalescence**.

Fig. 3. Constraint rules for $\mathcal{D}$

is zero, then we must have that $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$. The converse is not true however, i.e. $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$ does not imply $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) = 0$ and $\mathcal{D}_{\mathcal{M}}(XY, \emptyset) = 0$, since one of the terms maybe positive and other negative. Futhermore, coalescence for $|\mathcal{Y}| = 1$ does not hold, that is $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) = 0$ and $\mathcal{D}_{\mathcal{M}}(Y, \emptyset) = 0$ does not imply $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$. Even though $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \emptyset) = 0 \geq \mathcal{D}_{\mathcal{M}}(X, \emptyset)$, yet $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = -\mathcal{M}(X)$ which can be less than zero.

Note in the special case when $\mathcal{S} = X \sqcup \mathcal{Y}$, some of the rules of Proposition 8 collapse and we obtain the following proposition.

Proposition 9. *Let $\mathcal{M}$ be an n-subadditive (n-superadditive) measure. Let $\mathcal{S} = X \sqcup \mathcal{Y}$ or $\mathcal{S} = Z \sqcup \mathcal{Y}$. Then the level-n constraint of $\mathcal{M}$ satisfies the rules displayed in Figure 4 follow from Proposition 8.*

$$\frac{Y \in \mathcal{Y} \qquad \mathcal{S} = X \sqcup \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}[Y \leftarrow YZ]}{X \Rightarrow \mathcal{Y}} \qquad R1 \text{ - } \textbf{chain rule (a);}$$

$$\frac{Y \in \mathcal{Y} \qquad \mathcal{S} = X \sqcup \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}}{X \Rightarrow \mathcal{Y}[Y \leftarrow YZ]} \qquad R1 \text{ - } \textbf{chain rule (b);}$$

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}} \qquad R2 \text{ - } \textbf{reduction.}$$

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{X \Rightarrow \mathcal{Y}} \qquad R3 \text{ - } \textbf{triviality;}$$

$$\frac{U \subseteq \mathcal{S} = X \sqcup \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}}{XU \Rightarrow \mathcal{Y}} \qquad R4 \text{ - } \textbf{augmentation;}$$

$$\frac{Y \in \mathcal{Y} \qquad X \subseteq \mathcal{S} = Z \sqcup \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y}[Y \leftarrow Z]}{X \Rightarrow \mathcal{Y}[Y \leftarrow Z]} \qquad R5 \text{ - } \textbf{transitivity (a);}$$

$$\frac{\begin{array}{c} Y, Y' \in \mathcal{Y} \qquad \mathcal{S} = XZ \sqcup \mathcal{Y} \\ X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y}[Y' \leftarrow XY', Y \leftarrow Z] \end{array}}{X \Rightarrow \mathcal{Y}[Y \leftarrow Z]} \qquad R5 \text{ - } \textbf{transitivity (b);}$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y}} \qquad R6 \text{ - } \textbf{replication;}$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}} \qquad R7 \text{ - } \textbf{coalescence.}$$

Fig. 4. Additional inference rules when $\mathcal{S} = X \sqcup \mathcal{Y}$

4.1 Case studies

It turns out that when we apply Definition 4, and Propositions 5 and 6 to specific measures we uncover useful inference systems that can be used to reason about the relationships between the sets invlolved. Here we briefly cover the inference systems that can be uncovered when we use the measures `count` for counting sets, the Shannon entropy and the Simpson measure for data uniformity in databases, and finally `freq` in data mining.

1. The level-n constraint $\mathcal{D}_{count}(X, \mathcal{Y}) = 0$ holds when $\sqcap\mathcal{Y} \subseteq X$ for $|\mathcal{Y}| \geq 1$. This is a direct consequence of the inclusion-exclusion principle for counting finite sets. The resulting inference system for count follow directly from Proposition 8. Given that $X \Rightarrow \mathcal{Y}$ holds when $\sqcap\mathcal{Y} \subseteq X$, we have the rules below (derived from Subsection 2.1). The case where $\mathcal{Y} = \emptyset$ deserves special consideration as it implies $count(X) = 0$ which implies that $X = \emptyset$. For example, when $\mathcal{Y} - \{Y\} = \emptyset$, the reduction rule becomes $Y \subseteq X$ and $count(XY) = 0$ imply $count(X) = 0$.

chain rule a ($Y \in \mathcal{Y}$)
$$\frac{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq XY}{\sqcap\mathcal{Y}[Y \leftarrow ZY] \subseteq X}$$

chain rule b
$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y}[Y \leftarrow ZY] \subseteq X}{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq XY}$$

triviality
$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

decomposition
$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y}[Y \leftarrow YZ] \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

right augmentation a
$$\frac{Y \in \mathcal{Y} \qquad U \subseteq X \qquad \sqcap\mathcal{Y} \subseteq X}{\sqcap\mathcal{Y}[Y \leftarrow UY] \subseteq X}$$

right augmentation b
$$\frac{Y \in \mathcal{Y} \qquad U \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow UY] \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

left augmentation
$$\frac{1 \leq |\mathcal{Y}| < n \qquad \sqcap\mathcal{Y} \subseteq X}{\sqcap\mathcal{Y} \subseteq XU}$$

weak left augmentation
$$\frac{(U \subseteq X \cup Y \wedge Y \in \mathcal{Y}) \text{ or } (U \subseteq X) \qquad \sqcap\mathcal{Y} \subseteq X}{\sqcap\mathcal{Y} \subseteq XU}$$

weak transitivity
$$\frac{X \subseteq Z \qquad Y \subseteq X \qquad Z \subseteq Y}{Z \subseteq X}$$

weak transitivity ($Y, Y' \in \mathcal{Y}$)
$$\frac{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z, Y' \leftarrow XY'] \subseteq Y}{\sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq X}$$

transitivity ($Y \in \mathcal{Y}, |\mathcal{Y}| < n$)
$$\frac{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq Y}{\sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq X}$$

reduction
$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq XY}{\sqcap\mathcal{Y} - \{Y\} \subseteq X}$$

replication ($Y \in \mathcal{Y}$)
$$\frac{2 \leq |\mathcal{Y}| \leq n \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

coalescence ($2 \leq |\mathcal{Y}| \leq n$)
$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq Y}{\sqcap\mathcal{Y} - \{Y\} \subseteq X}$$

2. The level-1 constraint for the Shannon entropy measure holds when we have a functional dependency $X \rightarrow Y$. Let $V = XY$. The equality holds if and

only for each $x \in X$, $p_X(x) \log p(x) = \sum_{v_x \in V_x} p_V(v_x) \log p_V(v_x)$ if and only if for each $x \in X$, $\sum_{v_x \in V_x} p_V(v_x) \log p_X(v_x) = \sum_{v_x \in V_x} p_V(v_x) \log p_V(v_x)$ if and only if for each $x \in X$, $|V_x| = 1$ if and only if the relation T satisfies the functional dependency $X \rightarrow Y$. This was shown in [10][13][5]. The corresponding inference system rules that can be derived correspond to the well known rules of functional dependencies. Some of the inference system rules are shown in Figure 5.

The level-2 constraint for the Shannon entropy measure holds when we have a multivalued dependency $X \twoheadrightarrow Y$. In this case, $X \twoheadrightarrow Y$ holds if and only if $\mathcal{H}(X \cup Y) + \mathcal{H}(X \cup Z) = \mathcal{H}(X \cup Y \cup Z) + \mathcal{H}(X)$ and $Z = R - Y$ [5]. The corresponding inference system rules that can be derived using our measure framework correspond directly to the well known rules of multivalued dependencies. Some of the inference system rules are shown in Figure 6.

$$\textbf{reflexivity}\ \frac{Y \subset X}{X \rightarrow Y} \qquad \textbf{augmentation}\ \frac{X \rightarrow Y}{XU \rightarrow Y} \qquad \textbf{transitivity}\ \frac{X \rightarrow Y \qquad Y \rightarrow Z}{X \rightarrow Z}$$

Fig. 5. Inference system for function dependencies derived from the Shannon entropy measure.

$$\textbf{reflexivity}\ \frac{Y \subset X}{X \twoheadrightarrow Y \mid \mathcal{S} - XY} \qquad \textbf{augmentation}\ \frac{X \twoheadrightarrow Y \mid \mathcal{S} - XY}{XU \twoheadrightarrow Y \mid \mathcal{S} - XUY} \qquad \textbf{transitivity}\ \frac{X \twoheadrightarrow Y \mid \mathcal{S} - XY \qquad Y \twoheadrightarrow Z \mid \mathcal{S} - ZY}{X \twoheadrightarrow Z \mid \mathcal{S} - XZ}$$

$$\textbf{replication}\ \frac{X \rightarrow Y_2}{X \twoheadrightarrow Y_1 \mid S - XY_1} \qquad \textbf{coalescence}\ \frac{X \twoheadrightarrow Y | \mathcal{S} - XY \qquad Y \rightarrow Y'}{X \rightarrow Y'}$$

Fig. 6. Inference system for multivalued dependencies derived from the Shannon entropy measure.

3. The level-1 constraint for the Simpson measure holds when we have a functional dependency $X \rightarrow Y$. Let $V = X \cup Y$. The equality holds if and only for each $x \in X$, $p_X^2(x) = \sum_{v_x \in V_x} p_V^2(v_x)$. if and only if for each $x \in X$, $(\sum_{v_x \in V_x} p_V(v_x))^2 = \sum_{v_x \in V_x} p_V^2(v_x)$ if and only if for each $x \in X$, $|V_x| = 1$ if and only if the relation T satisfies the functional dependency $X \rightarrow Y$. The corresponding inference system rules that can be derived correspond to the well known rules of functional dependencies. Some of the inference system rules are shown in Figure 5.

The level-2 constraint for the Simpson measure $X \twoheadrightarrow Y$ holds when we have a special multivalued dependency $X \twoheadrightarrow Y$ such that $|Y_x| = 1$ or $|Z_x| = 1$ (where $Z = S - XY$, $Y_x = \Pi_Y(\sigma_{X=x}(T))$ and $Z_x = \Pi_Z(\sigma_{X=x}(T))$). This can be shown by expanding $X \rightarrow Y|Z = 0$ for Simpson's measure, which works out to be

$$S(X \cup Y) + S(X \cup Z) = S(X \cup Y \cup Z) + S(X)$$

Which can be expanded and simplified to be $\sum_{x \in \Pi_x} \sum_{y \in Y_x} P^2_{XY}(xy)$ $+ \sum_{x \in \Pi_x} \sum_{z \in Z_x} P^2_{XZ}(xz) = \sum_{x \in \Pi_x} \sum_{y \in Y_x} \sum_{z \in Z_{xy}} P^2_{XYZ}(xyz) + \sum_{x \in \Pi_x} P^2_X(x)$. This equation is true when $|Y_x| = 1$ or $|Z_x| = 1$ which implies a special multivalued dependency where one of independent columns has one distinct value only.The corresponding inference system rules that can be derived using our measure framework correspond directly to the well known rules of multivalued dependencies. Some of the inference system rules are shown in Figure 6. The level-n constraint of the Simpson measure can be generalized accordingly. A level-n constraint holds when we have two tuples such that $\forall t_1 t_2, R(t_1) \wedge R(t_2) \Rightarrow \bigcup_{i=1}^{n} t_1[i] = t_2[i]$.

4. The level-1 constraint of the `freq` measure holds if and only if $\texttt{freq}(X \cup Y) = \texttt{freq}(X)$ if and only if $\mathcal{B}(X \cup Y) = \mathcal{B}(X)$ if and only if there is a pure association rule from X to Y, denoted $X \rightarrow Y$, in $\mathcal{B}$. (A pure association rule is an association rule with confidence 100%.) The inference rules of our framework hold for association rules.
   The level-n constraint for the `freq` measure can be interpreted to yield weaker forms of association rules. Using the inclusion-exclusion principle, then $\mathcal{D}_{freq}(X, \mathcal{Y}) = 0$ implies that $X - \sqcup\mathcal{Y} = \emptyset$ (or alternatively $X \Rightarrow \sqcup\mathcal{Y}$). This means that item X cannot be bought alone; that is it is bought with at least one item $Y \in \mathcal{Y}$. Given that $\mathcal{D}_{freq}(X, \mathcal{Y}) = 0$, then this yields a weak association rule that can be interpreted that item X can only be bought with elements of $\mathcal{Y}$. For example, $\mathcal{D}_{freq}(X, \{Y_1, Y_2\}) = 0$ implies $freq(XY_1) + freq(XY_2) = freq(X) + freq(XY_1Y_2)$. To illustrate the use of the inclusion-exclusion principle in this interpretation, refering to Figure 7, we have $freq(X) = |a| + |b| + |c| + |d|$, $freq(XY_1Y_2) = |c|$, $freq(XY_1) = |b| + |c|$, and $freq(XY_2) = |c| + d|$. Putting everything together, we must have $|a| = 0$. This implies that X can only be bought with Y_1 or Y_2. The inference rules in Figure 3 also hold for these rules. The case where $\mathcal{Y} = \emptyset$ deserves special consideration as it implies $freq(X) = 0$ which implies that item X is not bought. For example, when $\mathcal{Y} - \{Y\} = \emptyset$, the reduction rule becomes $freq(XY) = freq(X)$ and $freq(XY) = 0$ imply $freq(X) = 0$.

5 Duality

In this section we will establish a duality between measures and differentials. This duality is similar to the one that exists between derivatives and integrals in calculus:

$$\int_x^{x+y} F'(u)du = F(x+y) - F(x).$$

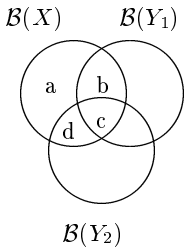

Fig. 7. Frequency constraints example

In our setting this duality is captured by the expression

$$\mathcal{D}_{\mathcal{M}}(X, \{X \cup Y\}) = \mathcal{M}(X \cup Y) - \mathcal{M}(X).$$

In other words, one can reasonably think about the expression $\mathcal{D}_{\mathcal{M}}(X, \{X \cup Y\})$ as stating the integration of the function $\mathcal{D}_{\mathcal{M}}$ "from" X "to" $X \cup Y$.

We wish to explore this duality in more depth. To do so, we consider functions satisfying the properties of measure differentials (Proposition 5) and "integrate" them. We can show that the resulting functions are measures and that their measure differentials are the original functions. These results establish that it is possible go back and forth between measures and differentials.

Definition 5. *Let $\mathcal{D}$ be a function from $2^S \times 2^{2^S}$ into the reals and let $n \geq 1$. We call $\mathcal{D}$ an n-differential if it has the following property:*

$$\frac{Y \in \mathcal{Y} \qquad |\mathcal{Y}| \geq 2}{\mathcal{D}(X, \mathcal{Y} - \{Y\}) = \mathcal{D}(X, \mathcal{Y}) + \mathcal{D}(XY, \mathcal{Y} - \{Y\})}\ \textbf{reduction}$$

We call $\mathcal{D}$ a positive n-differential *if $\mathcal{D}$ is an n-differential and $\mathcal{D}$ satisfies the property:*

$$\frac{X \subseteq S \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}(X, \mathcal{Y}) \geq 0}\ \textbf{positive}.$$

We call $\mathcal{D}$ a negative n-differential if $\mathcal{D}$ is an n-differential and $\mathcal{D}$ satisfies the following property:

$$\frac{X \subseteq S \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}(X, \mathcal{Y}) \leq 0}\ \textbf{negative}.$$

The following proposition formulates the duality between measures and differentials.

Proposition 10. *Let $\mathcal{D}$ be a n-differential ($n \geq 1$) and let $\mathcal{M}$ be the function from 2^S into the reals defined as follows:*

$$\mathcal{M}(X) = -\mathcal{D}(X, \emptyset). \tag{12}$$

Then for each $X \subseteq S$ and for each nonempty set $\mathcal{Y}$ of subsets of S such that $|\mathcal{Y}| \leq n$

$$\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = \mathcal{D}(X, \mathcal{Y}). \tag{13}$$

If $\mathcal{D}$ is a positive (negative) n-differential then $\mathcal{M}$ is an n-subadditive (an n-superadditive) measure.

Proof. We prove this by induction on $|\mathcal{Y}|$. For $|\mathcal{Y}| = 0$, $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = -\mathcal{M}(X) = \mathcal{D}(X, \emptyset)$. When $|\mathcal{Y}| \geq 1$, by the properties of $\mathcal{D}_{\mathcal{M}}$ (Proposition 5), $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) - \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y} - \{Y\}) = \mathcal{D}(X, \mathcal{Y} - \{Y\}) - \mathcal{D}(XY, \mathcal{Y} - \{Y\})$, by the induction hypothesis. By the reduction rule for $\mathcal{D}$ this is equal to $\mathcal{D}(X, \mathcal{Y})$.

It immediately follows from the the definition of measure that when $\mathcal{D}$ is a positive (negative) n-differential, $\mathcal{M}$ is an n-subadditive (an n-superadditive) measure. □

Acknowledgments: We thank Marc Gyssens, Paul Purdom, and Edward Robertson for helpful discussions on topics covered in this paper.

References

1. R.A. Brualdi. *Introductory Combinatorics (3rd edition).* Prentice-Hall, 1999.
2. T. Calders. *Axiomatization and Deduction Rules for the Frequency of Itemsets.* PhD dissertation- University of Antwerp, 2003.
3. D. Cohn. *Measure Theory.* Birkhäuser-Boston, 1980.
4. T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley-Interscience Publication, 1991.
5. M. Dalkilic and E. Robertson. Information dependencies. In *Symposium on Principles of Database Systems*, pages 245–253, 2000.
6. R. Fagin. Multivalued dependencies and a new normal form for relational databases. *ACM Trans. Database Syst.*, 2(3):262–278, 1977.
7. S. Goldberg. *Introduction to Difference Equations.* Dover, 1986.
8. J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *J. Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.
9. V. Krishnamurthy. *Combinatorics: Theory and Applications.* Ellis Horwood Limited, 1986.
10. F. M. Malvestuto. Statistical treatment of the information content of a database. *Information Systems*, 11:211–223, 1986.
11. C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.
12. E. Simpson. Measurement of diversity. In *Nature*, volume 163, page 688, 1949.
13. T. Lee. An Information-Theoretic Analysis of Relational Databases – Part I: Data Dependencies and Information Metric. *IEEE Transactions on Software Engineering*, SE-13:1049–1061, 1987.