Representing Informational Harmoniums using Semifield-Valued Formal Concept Analysis

Francisco J. Valverde-Albacete[1] * and Carmen Peláez-Moreno[1]

Depto. Teoría de Señal y Comunicaciones, Univ. Carlos III de Madrid, Madrid, Spain, fva@tsc.uc3m.es carmen@tsc.uc3m.es

Abstract. In this paper we apply semifield-based Formal Concept Analysis (FCA) to the analysis of Information Harmoniums. These are unsupervised graphical models that, similarly to energy-based models, are written in terms of information functions, but whose expression is linear in an information semifield. In particular we concentrate in the case of the limit information semifields that are the completed max-plus and min-plus semifields for which strong representation theorems in relation to Galois connections and semifield-valued FCA have been proven. We center our contribution in the analysis of the representation spaces for visible and hidden nodes of Information Harmoniums and the lattices related to them.

1 Introduction and Motivation

Restricted Boltzmann Machines (RBM) [1] are one of the basic techniques that revolutionized Artificial Neural Networks some 10 years ago transforming them into Deep Neural Networks [2].

RMBs are, in fact, a type of *harmonium* [3] at the intersection of Boltzmann Machines [4]—a kind of Energy-Based Model [5]—and a Product-of-experts (PoE) [6]. PoE are readily trained by means of Contrastive Divergence, a better approach for Maximum Likelihood Estimation (MLE) than Gibbs sampling [7], which explains their efficiency.

In this paper we put in evidence some discrepancies in the definition of harmoniums (Section 2.1) that suggest that a more natural point of view is to consider their energy functions as based in an information semifield (Section 2.2). If this is the case, a particular instance of the information semifields are the $\overline{\mathbb{R}}_{\max,+}$ and $\overline{\mathbb{R}}_{\min,+}$ semifields over which a multi-valued generalization of FCA can be defined, $\overline{\mathbb{R}}_{\max,+}$-FCA (Section 2.3). In Section 3 we present our results and contend that FCA in general, and $\overline{\mathbb{R}}_{\max,+}$-FCA in particular, provides a framework for the visualization and understanding of information harmoniums that could also provide clues for other types of harmoniums. Finally we provide some conclusions.

* Corresponding author.

© Francisco J. Valverde-Albacete, Martin Trnecka (Eds.): CLA 2020, pp. 159–170,
ISBN 978–9949–83–557–7, Tallinn University of Technology, 2020.

2 Theory and Methods

2.1 RBM and Harmoniums

The basic model. Technically speaking a harmonium is an undirected graphical model for the generation of a joint probability distribution. Their graphical model can be seen in Fig. 1.

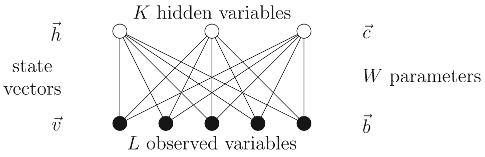

Fig. 1: The architecture of a RBM as an undirected graphical model, and its parameters.

One type of nodes tagged with v_l labels are *the set of L input or observed nodes*. The other type is the *set of K output or hidden nodes*, tagged with h_k. Conventionally we will use indices $l \in [1 \ldots L]$ and $k \in [1 \ldots K]$ to go over them.

Harmoniums are generative models: consider random vectors $\overline{V} = \{V_i \mid i \in I\}$ and $\overline{H} = \{H_j \mid j \in J\}$ with component random variables V_i and H_j associated to the visible and hidden nodes, respectively. We posit that for a particular pair of random vector (values) $(\vec{v}, \vec{h}) \in \overline{V} \times \overline{H}$ the joint probability distribution $p_{\overline{VH}}$ of Figure 1 takes the form of a Boltzmann distribution,

$$p_{\overline{VH}}(\vec{v}, \vec{h}) = \frac{e^{-\beta E(\vec{v}, \vec{h})}}{\sum_{\vec{v} \in \overline{V}} \sum_{\vec{h} \in \overline{H}} e^{-\beta E(\vec{v}, \vec{h})}} \tag{1}$$

where $\beta \in (0, \infty]$ is a formal parameter, the *coldness*, and $E(\vec{v}, \vec{h})$ is a provided *energy function*, given in terms of bias vectors $\vec{b} \in \overline{V}$, $\vec{c} \in H$ and the weight matrix W:

$$E(\vec{v}, \vec{h}) = \vec{v}^{\mathrm{T}} \vec{c} + \vec{b}^{\mathrm{T}} \vec{h} + \vec{v}^{\mathrm{T}} W \vec{h}$$

Several points are worth making here:

- The coldness parameter β is in the context of Physics often written as its inverse, the temperature $T = 1/\beta$ respecting the original form of the Boltzmann's function. It was not originally introduced in the Harmonium [3], but was always considered as part of the training procedure of Boltzmann machines as a relaxation parameter [4].
- It is easy to see, e.g. [8], that the energy function can be interpreted as a scalar product *in the standard field of reals*,

$$E(\vec{v}, \vec{h}) = \langle \vec{v}' | W' | \vec{h}' \rangle \triangleq \begin{bmatrix} 1 \; \vec{v}^{\mathrm{T}} \end{bmatrix} \cdot W' \cdot \begin{bmatrix} 1 \\ \vec{h} \end{bmatrix} \qquad W' = \begin{bmatrix} a & \vec{b}^{\mathrm{T}} \\ \vec{c} & W \end{bmatrix} \tag{2}$$

where $\vec{v}' = \left[1\ \vec{v}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\vec{h}' = \left[1\ \vec{h}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $a = 0$.

- By the properties of Boltzmann distributions, the constant in the upper left hand corner of the matrix, say a, is arbitrary, since it would appear as a multiplying factor $e^{-\beta a}$ in both numerator and denominator of (1). It amounts to a minimum level attainable by the energy function,

$$E'(\vec{v}, \vec{h} \mid W') = a + E(\vec{v}, \vec{h} \mid W|_{a=0})$$

and it may be used, for instance, to ensure that the joint distribution is *positive*, that is, non-null for any $\vec{v} \in \mathcal{V}, \vec{h} \in \mathcal{H}$. We will not distinguish between these two forms and consider the $W = W'|_{a=0}$, including $\vec{b}$ and $\vec{c}$, but not a, to be the set of parameters of the model $p_{\overline{VH}}(\vec{v}, \vec{h} \mid W'|_{a=0})$.

- $Z(W) = \sum_{\vec{v}\in\overline{V}} \sum_{\vec{h}\in\overline{H}} e^{-E(\vec{v},\vec{h}|W)}$ is the well-known *partition function* that ensures the normalization of $p_{\overline{VH}|W}$. As argued, we will use sometimes $Z(W') = e^{-\beta a} \cdot Z(W)$ to make it explicit that we include all possible parameters, including the minimal offset. In fact, the partition function can be included in the energy function by defining $a = F_\beta(W|_{k=0}) = \frac{1}{\beta} \log_e Z(W|_{k=0})$ in which case

$$Z(W') = e^{-\beta F_\beta(W)} \cdot Z(W|_{k=0}) = Z(W|_{k=0})^{-1} \cdot Z(W|_{k=0}) = 1$$

so that $p_{\overline{VH}}(\vec{v}, \vec{h}) = e^{-\beta E(\vec{v},\vec{h}|W')}$ with an explicit normalization.

- With this encoding, the number of free parameters of this model is $(L+1)(K+1)-1$ [9].

Inference in harmoniums. As mentioned, RBM are harmoniums that can also be seen as PoE [1, 10] where certain joint distributions are product of individual components. In the case of the harmonium as a PoE, the components are the conditional distribution of each of the hidden nodes given the input, or the conditional distribution of each of the input nodes given the output:

$$p_{\overline{H}|\overline{V}}(\vec{h} \mid \vec{v}_0, W) = \prod_j p_{H_j|\overline{V}}(h_j \mid \vec{v}_0, W) \quad p_{\overline{V}|\overline{H}}(\vec{v} \mid \vec{h}_0, W) = \prod_i p_{V_i|\overline{H}}(v_i \mid \vec{h}_0, W) \tag{3}$$

In fact, RBMs are harmoniums with Bernouilli-distributed binary visible and hidden variables (see below) [1]. But note that harmoniums have been generalized to visible and hidden variables with distributions in the exponential family, which include the most used distributions in data models [11].

2.2 Information Semifields

Positive semifields. A technical requisite for the energy function is that it is always positive [8]. This suggests investigating energy functions with positive semifields [12].

Example 1 (Multiplicative-product real semifields [13]). Consider a free parameter $r \in [-\infty, 0) \bigcup (0, \infty]$ in the following operations

$$u \oplus_r v = \left(u^r \underset{\cdot}{+} v^r\right)^{\frac{1}{r}} \qquad u \otimes_r v = \left(u^r \underset{\cdot}{\times} v^r\right)^{\frac{1}{r}} \qquad u^* = \left(\frac{1}{u^r}\right)^{\frac{1}{r}} = u^{-1} \tag{4}$$

where the basic operations are to be interpreted in $\mathbb{R}_{\geq 0}$ and the dotted notation is adopted from that used by Moreau for convex analysis [14], where $0 \underset{\cdot}{\times} \infty = 0$ but $0 \overset{\cdot}{\times} \infty = \infty$. Notice the following properties:

- if $r \in (0, \infty]$ then $u \otimes_r v = u \underset{\cdot}{\otimes}_r v = u \underset{\cdot}{\times} v$, $\bot_r = 0$, $e_r = 1$, and $\top_r = \infty$, and the complete positive semifield generated, order-aligned with $\mathbb{R}_{\geq 0}$, is:

$$\left(\mathbb{R}_{\geq 0}\right)_r = \langle [0, \infty], \underset{\cdot}{\oplus}_r, \underset{\cdot}{\otimes}_r, \cdot^*, \bot_r = 0, e, \top_r = \infty \rangle \tag{5}$$

- if $r \in [-\infty, 0)$ then $u \otimes_r v = u \overset{\cdot}{\otimes}_r v = u \overset{\cdot}{\times} v$, $\bot_r = \infty$, $e_r = 1$, and $\top_r = 0$, and the complete positive semifield generated, dually order-aligned with $\mathbb{R}_{\geq 0}$, is:

$$\left(\mathbb{R}_{\geq 0}\right)_{-r} = \langle [0, \infty], \overset{\cdot}{\oplus}_r, \overset{\cdot}{\otimes}_r, \cdot^*, \bot_r^* = \infty, e, \top_r^* = 0 \rangle \tag{6}$$

Therefore, $\left(\mathbb{R}_{\geq 0}\right)_r$ and $\left(\mathbb{R}_{\geq 0}\right)_{-r}$ are inverse, completed positive semifields, and $\left(\mathbb{R}_{\geq 0}\right)_r^{-1} = \left(\mathbb{R}_{\geq 0}^{-1}\right)_r$. In particular:

$$\left(\mathbb{R}_{\geq 0}\right)_1 = \mathbb{R}_{\geq 0} \qquad \left(\mathbb{R}_{\geq 0}\right)_{-1} = \mathbb{R}_{\geq 0}^{-1} \tag{7}$$

$$\lim_{r \to \infty} \left(\mathbb{R}_{\geq 0}\right)_r = \overline{\mathbb{R}}_{\max, \times} \qquad \lim_{r \to -\infty} \left(\mathbb{R}_{\geq 0}\right)_r^{-1} = \overline{\mathbb{R}}_{\min, \times} \tag{8}$$

□

Note that:

- these semifields are able to capture the usual concept of a "positive quantity", like a mass, length, etc.
- All these semifields have the same product, and the same "extreme" points, $\{0, 1, \infty\}$. Their only difference lies in the addition. Sometimes, when only the product is important in an application, the addition remains in the background and we are not really sure in which algebra we are working on.
- Instead of using the abstract notation for the inversion $\cdot^*$, since $\mathbb{R}_{\geq 0}$ is the paragon originating all other behaviours, we have decided to use the original notation for the inversion in the (incomplete) semifield.

The need for information functions. If we were to use the semifields $\left(\mathbb{R}_{\geq 0}\right)_r$ to build the energy function of the harmonium model (1), despite the fact that these semifields are positive, and they generate models that are products of positive terms, from a physical point of view there is a mismatch between them and the numbers required in the model: these semifields describe natural measurements of mass, length, etc., but model (1) demands that we use logarithmic magnitudes.

A first step is to use an *information function* to transform normalized masses, that is, something akin to probabilities, into *(quantity of) information.* Hartley's information function $h(p) = -\log_2 p, p \in [0, 1]$ was extended by Rényi to include a free parameter $\alpha \in \mathbb{R}_{\pm\infty}$, the Rényi order [15], and this has later been shifted into $r = \alpha - 1$ [12, § 3.1]—where $r \in [-\infty, \infty]$ is the *(shifted) Rényi order*—as:

$$\varphi'(h) = b^{-rh} \qquad\qquad \varphi'^{-1}(p) = \frac{-1}{r} \log_b p \tag{9}$$

Informational or Entropy semifields. The shifted Rényi order allows us to obtain new semifields from the $\mathbb{R}_{\pm\infty}$ using Rényi's information function [12]:

Example 2 (Entropy semifields [12]). Let $r \in [-\infty, \infty] \backslash \{0\}$ and $b \in (1, \infty)$. Then the algebra $\langle [-\infty, \infty], \oplus_r, \otimes_r, \cdot^{-1}, \bot = \infty, e = 0 \rangle$ obtained from the semifield of positive reals by Rényi's information function and whose basic operations are:

$$u \oplus_r v = -\frac{1}{r} \log_b \left(b^{-ru} + b^{-rv} \right) \qquad u \otimes_r v = u + v \qquad u^* = -u \tag{10}$$

can be completed to two dually-ordered positive semifields

$$\mathbb{H}_r = \langle [-\infty, \infty], \underset{\cdot}{\oplus}_r, \underset{\cdot}{\otimes}_r, -\cdot, \bot = -\infty, e = 0, \top = \infty \rangle \tag{11}$$

$$-\mathbb{H}_r = \langle [-\infty, \infty], \dot{\oplus}_r, \dot{\otimes}_r, -\cdot, -\bot = \infty, e = 0, -\top = -\infty \rangle \tag{12}$$

whose elements can be considered as generalized information values and operated accordingly. We will typically consider $b = e^1$ so that $\log_e(m) = \log(m)$ and the informations and entropies are measured in *nats.*

It is easy to see that addition is a very complicated operation in information semifields in general: For $r \in \mathbb{R}/\{0\}$ we expand the notation for addition as:

$$\sum_i{}^r h_i \triangleq h_i \dot{\oplus}_r h_2 \dot{\oplus}_r \ldots = -\frac{1}{r} \log(\sum_i e^{-rh_i}) \qquad r \in (0, \infty] \tag{13}$$

$$\sum_i{}_r h_i \triangleq h_i \underset{\cdot}{\oplus}_r h_2 \underset{\cdot}{\oplus}_r \ldots = -\frac{1}{r} \log(\sum_i e^{-rh_i}) \qquad r \in [-\infty, 0)$$

It is easy to see that $\sum^r{}_i h_i$ is smooth approximation to the minimum while $\sum_{ri} h_i$ smoothly approximates the maximum, since:

$$\sum_i{}^{\bullet} h_i \triangleq \lim_{r \to \infty} \sum_i{}^r h_i = \min_i h_i \qquad \sum_i{}_{\bullet} h_i \triangleq \sum_i{}_r h_i = \max_i h_i \tag{14}$$

In this case, the dually ordered complete positive semifields have an idempotent addition, and are normally called the (completed) max-plus $\overline{\mathbb{R}}_{\max,+}$ and min-plus $\overline{\mathbb{R}}_{\min,+}$ semifields, or *tropical semirings.*

Note that when $\vec{h}$ is a vector of generalized informations, the additions in (13) are clearly entropies—more precisely, cross entropies—hence the name assigned to this type of semifield. □

It is difficult to ascertain in what algebra the computations inside the log-sum-exp functions in (13) are carried out. The following result was proven in [12]:

Proposition 1. *The generalized Rényi information function is an isomorphism of positive semifields between $\mathbb{R}_{\geq 0}$ and $\mathbb{H}_r$. In particular, when $r = 1$ (the case of Hartley's function), it is an isomorphism between $\mathbb{R}_{\geq 0}$ and $\mathbb{H}$.*

Therefore, all computations can actually be carried out in $\mathbb{R}_{\geq 0}$. In expressions we maintain, however, the more abstract notation, so that we write, for scalar products over an informational semifield, in general

$$\langle \vec{x}|W|\vec{y}\rangle^r \triangleq \vec{x}^{\mathrm{T}} \dot{\otimes}_r W \dot{\otimes}_r \vec{y} = \sum_{ij}^{r} x_i \dot{\oplus}_r w_{ij} \dot{\oplus}_r y_j = \frac{-1}{r} \log \left(\sum_{i,j} e^{-r(x_i \underset{\cdot}{+} w_{ij} \underset{\cdot}{+} y_j)} \right)$$

$$\langle \vec{x}|W|\vec{y}\rangle_r \triangleq \vec{x}^{\mathrm{T}} \underset{\cdot}{\otimes}_r W \underset{\cdot}{\otimes}_r \vec{y} = \sum_{ij}{}_r\, x_i \underset{\cdot}{\oplus}_r w_{ij} \underset{\cdot}{\oplus}_r y_j = \frac{1}{r} \log \left(\sum_{i,j} e^{r(x_i \underset{\cdot}{+} w_{ij} \underset{\cdot}{+} y_j)} \right)$$

2.3 $\overline{\mathbb{R}}_{\mathbf{max},+}$-FCA

In a complete positive semifield $\overline{\mathcal{K}}$, an element φ is *invertible* if and only if it is not extremal, $\varphi \in \overline{\mathcal{K}} \setminus \{\bot, \top\}$. We have then the four possible types of Galois connections due to scalar products in the semifield $\overline{\mathcal{K}}$ [16].

Theorem 1 (The four-fold connection). *Let G and M be sets of* formal objects *and* formal attributes*, respectively, with $|G| = g$ and $|M| = m$. Let $(G, M, R)_{\overline{\mathcal{K}}}$ be a formal context whose incidence $R \in K^{g \times m}$ takes values in a complete idempotent semifield $\overline{\mathcal{K}}$. Consider the vector spaces $\mathcal{X} = \overline{\mathcal{K}}^g$ and $\mathcal{Y} = \overline{\mathcal{K}}^m$, an invertible element $\varphi = \gamma \dot{\otimes} \mu \in K$ and the scaled spaces $\widetilde{\mathcal{X}}^\gamma = \gamma^{-1} \underset{\cdot}{\otimes} \mathcal{X}$ and $\widetilde{\mathcal{Y}}^\mu = \mu^{-1} \underset{\cdot}{\otimes} \mathcal{Y}$. Then*

1. *The bracket $\langle x \mid R \mid y\rangle^{\mathrm{OI}} = x^* \dot{\otimes} R \dot{\otimes} y^{-1}$ induces a Galois connection $(\cdot^{\uparrow}_{R}, \cdot^{\downarrow}_{R}) : \widetilde{\mathcal{X}}^\gamma \leftharpoonup\!\!\!\rightharpoonup \widetilde{\mathcal{Y}}^\mu$ between the scaled spaces through the polars*

   $$x^{\uparrow}_{R} = R^{\mathrm{T}} \dot{\otimes} x^{-1} \qquad\qquad y^{\downarrow}_{R} = R \dot{\otimes} y^{-1} \tag{15}$$

   which define two bijective sets, the system of extents $\mathfrak{B}^{\gamma}_{G}$ *and the* system of intents $\mathfrak{B}^{\mu}_{M}$

   $$\mathfrak{B}^{\gamma}_{G} = (\widetilde{Y}^{\mu})^{\downarrow}_{R} \qquad\qquad \mathfrak{B}^{\mu}_{M} = (\widetilde{X}^{\gamma})^{\uparrow}_{R} \tag{16}$$

and whose composition generate closure operators:

$$\pi_R(x) = \left(x^{\uparrow}_{R}\right)^{\downarrow}_{R} = R \mathbin{\dot{\otimes}} (R^* \mathbin{\underset{\cdot}{\otimes}} x) \tag{17}$$

$$\pi_{R^{\mathrm{T}}}(y) = \left(y^{\downarrow}_{R}\right)^{\uparrow}_{R} = R^{\mathrm{T}} \mathbin{\dot{\otimes}} (R^{-1} \mathbin{\underset{\cdot}{\otimes}} y)$$

which are the identities on $\mathfrak{B}^{\gamma}_{G}$ *and* $\mathfrak{B}^{\mu}_{M}$*, respectively.*

2. *The bracket* $\langle x \mid R \mid y\rangle^{\mathrm{IO}} = x^{\mathrm{T}} \mathbin{\dot{\otimes}} R \mathbin{\dot{\otimes}} y$ *induces a co-Galois connection* $(\cdot^{\uparrow}_{R^{-1}}, \cdot^{\downarrow}_{R^{-1}}) : \widetilde{\mathcal{X}}^{\gamma} \mathrel{\leftharpoondown\!\!\rightharpoondown} \widetilde{\mathcal{Y}}^{\mu}$ *between the scaled spaces through the maps:*

$$x^{\uparrow}_{R^{-1}} = R^* \mathbin{\underset{\cdot}{\otimes}} x^{-1} \qquad\qquad y^{\downarrow}_{R^{-1}} = R^{-1} \mathbin{\underset{\cdot}{\otimes}} y^{-1} \tag{18}$$

which define two bijective sets the systems of neighbourhoods of objects $\mathfrak{N}^{\gamma}_{G}$ *and* neighbourhoods of attributes $\mathfrak{N}^{\mu}_{M}$

$$\mathfrak{N}^{\gamma}_{G} = (\widetilde{Y}^{\mu})^{\downarrow}_{R^{-1}} \qquad\qquad \mathfrak{N}^{\mu}_{M} = (\widetilde{X}^{\gamma})^{\uparrow}_{R^{-1}} \tag{19}$$

and whose composition generate interior operators:

$$\kappa_{R^{-1}}(x) = \left(x^{\uparrow}_{R^{-1}}\right)^{\downarrow}_{R^{-1}} = R^{-1} \mathbin{\underset{\cdot}{\otimes}} (R^{\mathrm{T}} \mathbin{\dot{\otimes}} x) \tag{20}$$

$$\kappa_{R^*}(y) = \left(y^{\downarrow}_{R^{-1}}\right)^{\uparrow}_{R^{-1}} = R^* \mathbin{\underset{\cdot}{\otimes}} (R \mathbin{\dot{\otimes}} y)$$

which are the identities on $\mathfrak{N}^{\gamma}_{G}$ *and* $\mathfrak{N}^{\mu}_{M}$*, respectively.*

3. *The bracket* $\langle x \mid R \mid y\rangle^{\mathrm{OO}} = x^{*} \mathbin{\dot{\otimes}} R \mathbin{\dot{\otimes}} y$ *induces a left adjunction* $(\cdot^{\exists}_{R}, \cdot^{\forall}_{R}) : \widetilde{\mathcal{X}}^{\gamma} \leftrightharpoons \widetilde{\mathcal{Y}}^{\mu}$ *between the scaled spaces through the left adjunct pair of maps:*

$$x^{\exists}_{R} = R^* \mathbin{\underset{\cdot}{\otimes}} x \qquad\qquad y^{\forall}_{R} = R \mathbin{\dot{\otimes}} y \tag{21}$$

which define another bijection between the systems of extents $\mathfrak{B}^{\gamma}_{G}$ *and neighbourhoods of attributes* $\mathfrak{N}^{\mu}_{M}$

$$\mathfrak{B}^{\gamma}_{G} = (\widetilde{Y}^{\mu})^{\forall}_{R} \qquad\qquad \mathfrak{N}^{\mu}_{M} = (\widetilde{X}^{\gamma})^{\exists}_{R} \tag{22}$$

and whose compositions are the closure of extents and interior of attributes:

$$\pi_R(x) = \left(x^{\exists}_{R}\right)^{\forall}_{R} \qquad\qquad \kappa_{R^*}(y) = \left(y^{\forall}_{R}\right)^{\exists}_{R} \tag{23}$$

4. *The bracket* $\langle x \mid R \mid y\rangle^{\mathrm{II}} = x^{\mathrm{T}} \mathbin{\dot{\otimes}} R \mathbin{\dot{\otimes}} y^{-1}$ *induces an adjunction on the right* $(\cdot^{\forall}_{R^{\mathrm{T}}}, \cdot^{\exists}_{R^{\mathrm{T}}}) : \widetilde{\mathcal{X}}^{\gamma} \rightleftharpoons \widetilde{\mathcal{Y}}^{\mu}$ *between the scaled spaces through the pair of adjunct maps:*

$$x^{\forall}_{R^{\mathrm{T}}} = R^{\mathrm{T}} \mathbin{\dot{\otimes}} x \qquad\qquad y^{\exists}_{R^{\mathrm{T}}} = R^{-1} \mathbin{\underset{\cdot}{\otimes}} y \tag{24}$$

which define anothers bijection between the systems of neighbourhood of objects $\mathfrak{N}_G^\gamma$ and intents $\mathfrak{B}_M^\mu$

$$\mathfrak{N}_G^\gamma = (\widetilde{Y}^\mu)_{R^\mathrm{T}}^{\exists} \qquad \mathfrak{B}_M^\mu = (\widetilde{X}^\gamma)_{R^\mathrm{T}}^{\forall} \tag{25}$$

and whose compositions are the interior of objects and the closure of attributes:

$$\kappa_{R^{-1}}(x) = (x_{R^\mathrm{T}}^{\forall})_{R^\mathrm{T}}^{\exists} \qquad \pi_{R^\mathrm{T}}(y) = (y_{R^\mathrm{T}}^{\exists})_{R^\mathrm{T}}^{\forall} \tag{26}$$

Therefore, it makes sense to define the following (meta) concept:

Definition 1. *For a formal context $(G, M, R)_{\overline{\mathcal{K}}}$, the* 4-formal concept (a, b, c, d) *is a 4-tuple such that $a \in \underline{\mathfrak{B}}_G^\gamma$, $b \in \underline{\mathfrak{B}}_M^\mu$, $c \in \underline{\mathfrak{N}}_G^\gamma$, and $d \in \underline{\mathfrak{N}}_M^\mu$ and all the following relations hold:*

$$\begin{aligned} a &= (b)_R^{\downarrow} = (d)_R^{\forall} & \qquad b &= (a)_R^{\uparrow} = (c)_{R^\mathrm{T}}^{\forall} \\ d &= (c)_{R^{-1}}^{\uparrow} = (a)_R^{\exists} & \qquad c &= (d)_{R^{-1}}^{\downarrow} = (b)_{R^\mathrm{T}}^{\exists} \end{aligned} \tag{27}$$

3 Results

Given the distinction between probability-based and information-based semifields in Section 2.2, and considering that many sensorial magnitudes are "logarithmically perceived"—despite the criticism to the Weber-Fenchner law [17]—we would like to "fix" the magnitudes used in harmoniums while maintaining their main design considerations.

3.1 Information Measures from Mass Measures

We may further generalize Rényi's information function $\varphi'^{-1}(\cdot)$ and its inverse $\varphi'(\cdot)$ to vectors, so if we consider a mass measure $\overline{m} \in [0, \infty]^I$, we have its related information measure:

$$\begin{aligned} &h_r(\cdot)\colon [0, \infty]^I \to [-\infty, \infty]^I \\ &\overline{m} = \{m_i\}_{i \in I} \mapsto h_r(\overline{m}) = \{\varphi'^{-1}(m_i)\}_{i \in I} = \{\frac{-1}{r} \log_b m_i\}_{i \in I} \end{aligned} \tag{28}$$

whereas if we consider an information measure $\overline{h} \in [-\infty, \infty]^n$, then we have its *related mass measure*:

$$\begin{aligned} &m_r(\cdot)\colon [-\infty, \infty]^I \to [0, \infty]^I \\ &\overline{h} = \{h_i\}_{i \in I} \mapsto m_r(\overline{h}) = \{e^{-r h_i}\}_{i \in I} \end{aligned} \tag{29}$$

Note that the $m_r(\cdot)$ and $h_r(\cdot)$ thus defined are mutually inverse bijections of tuple spaces between the semifields. In particular we state without proving:

Proposition 2. *Let $r \in [-\infty, \infty] \setminus \{0\}$. Then $h_r(\cdot)\colon (\mathbb{R}_{\geq 0})^n \to (\mathbb{H}_r)^n$ is a dual isomorphism of semivector spaces over their corresponding semifields, with inverse $m_r(\cdot)\colon (\mathbb{H}_r)^n \to (\mathbb{R}_{\geq 0})^n$.*

The duality mentioned in Proposition 2 concerns their natural orders. The Hartley information function, e.g. Rényi's function with $r = 1$ is a very special case:

Corollary 1. *The Hartley information function is a dual isomorphism of semivectors spaces $h_1(\cdot) : (\mathbb{R}_{\geq 0})^n \to (\mathbb{H})^n$.*

The following is much more interesting and admits the previous one as a particular case:

Proposition 3. *The Hartley information function is a dual isomorphism of semivectors spaces $h_r(\cdot)\colon ({(\mathbb{R}_{\geq 0})}_r)^n \to (\mathbb{H}_r)^n$.*

Proof. Let $\{\alpha_j \in \mathbb{R}_{\geq 0} \mid j \in J\}$ and $\{\vec{v}_j \in {(\mathbb{R}_{\geq 0})_r}^n \mid j \in J\}$ with $a_j = -\log \alpha_j$ and $\vec{x}_j = -\log \vec{v}_j$. Then, a linear combination of vectors in ${(\mathbb{R}_{\geq 0})_r}^n$, $\sum_{r j} \alpha_j \underset{\cdot}{\otimes}_r \vec{v}_j$ is transformed into a linear combination of vectors in $(\mathbb{H}_r)^n$ as

$$-\log\left(\sum_{j}{}_r\, \alpha_j \underset{\cdot}{\otimes}_r \vec{v}_j\right) = -\log\left(\sum_j \alpha_j^r \underset{\cdot}{\times} \vec{v}_i^r\right)^{1/r} = \frac{-1}{r}\log\left(\sum_j e^{-ra_j} \underset{\cdot}{\times} e^{-r\vec{x}_j}\right) =$$

$$= \frac{-1}{r}\log\left(\sum_j e^{-r(a_j \dot{+} \vec{x}_j)}\right) = \frac{-1}{r}\log\left(\sum_j e^{-r(a_j \dot{\otimes} \vec{x}_j)}\right) =$$

$$= \sum_j^r a_j \dot{\otimes}_r \vec{x}_j$$

which is a linear combination of vectors with modified scalars. Since the transformations are equalities and biunivocal it describes an isomorphism which is inverted by taking the negative exponential function on each of the terms. □

In fact, we have also the following corollary:

Corollary 2. *Let $\{\vec{\beta}, \vec{v}\} \subset ({(\mathbb{R}_{\geq 0})}_r)^n$ and $\{\vec{b}, \vec{x}\} \subset (\mathbb{H}_r)^n$ where $\vec{\beta} = \exp(-\vec{b})$ and $\vec{v} = \exp(-\vec{x})$. Then:*

$$\vec{b}^* \dot{\otimes} \vec{x} = -\log(\vec{\beta}^* \underset{\cdot}{\otimes} \vec{v}) \qquad\qquad \vec{\beta}^* \underset{\cdot}{\otimes} \vec{v} = \exp(-\vec{b}^* \dot{\otimes} \vec{x}) \tag{30}$$

Proof. Just collect n of the scalars α_j on a single vector $\vec{\beta} = \{\alpha_j^{-1}\}_{i=1}^n$ and multiply by $\vec{v}$ as $\vec{\beta}^* \underset{\cdot}{\otimes} \vec{v}$. Then by the previous procedure we get $\vec{b}^* \dot{\otimes} \vec{x} = -\log(\vec{\beta}^* \underset{\cdot}{\otimes} \vec{v})$. The second equality comes from the isomorphism. □

3.2 Informational Harmoniums

The relation between mass measures and information measures suggests we define harmoniums using information:

Definition 2. *Let $\mathcal{X} \equiv {\mathbb{H}_r}^n$ and $\mathcal{Y} \equiv {\mathbb{H}_r}^m$ be semivector spaces over the informational semifield $\mathbb{H}_r$. Let $\vec{x} \in \mathcal{X}$ and $\vec{y} \in \mathcal{Y}$ be generalized information functions. We call an* informational harmonium *one whose joint information function is a scalar product between the spaces, with $W \in (\mathbb{H}_r)^{n \times m}$,*

$$\begin{aligned} &\overline{h}^{r}_{\overline{XY}}(\cdot, \cdot \mid W) \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{H}_r \\ &(\vec{x}, \vec{y}) \mapsto \overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y} \mid W) = \langle \vec{x} | W | \vec{y} \rangle_r \end{aligned} \tag{31}$$

If the information functions are related to the mass distributions $\overline{m}_{\overline{X}} = m(\vec{x})$ and $\overline{m}_{\overline{Y}} = m(\vec{y})$ by Hartley's function, then the information harmonium is the associated mass function over the whole product space $(\vec{x}, \vec{y}) \in \mathcal{X} \times \mathcal{Y}$ that is

$$\overline{m}_{\overline{XY}}(\vec{x}, \vec{y}) = m(\overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y} \mid W)) = \{e^{-\overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y}|W)} \mid \vec{x} \in \mathcal{X}, \vec{y} \in \mathcal{Y}\} \tag{32}$$

with associated partition function and distribution

$$\|\overline{m}_{\overline{XY}}\|_1 = \sum_{\vec{x} \in \mathcal{X}} \sum_{\vec{y} \in \mathcal{Y}} e^{-\overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y}|W)} \qquad q_1(\overline{m}_{\overline{XY}}) = \frac{e^{-\overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y}|W)}}{\|\overline{m}_{\overline{XY}}\|_1}$$

□

Proposition 4. *In the conditions of Definition 2, let $\vec{v} = \exp(-\vec{x}), \vec{h} = \exp(-\vec{y})$ and $U = \exp(W)$ whereby we mean the entry-wise exponentiation. Then:*

$$\overline{m}_{\overline{VH}}(\vec{v}, \vec{h} \mid U) = \exp(-\overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y} \mid W)) \quad \overline{h}^{r}_{\overline{XY}}(\vec{x}, \vec{y} \mid W) = -\log(\overline{m}_{\overline{VH}}(\vec{v}, \vec{h} \mid U)) \tag{33}$$

Proof. This is just a repeating of the proof for Proposition 3.

When $r \to \infty$ we have the following corollary:

Corollary 3. *In the conditions of Definition 2, let $\vec{v} = \exp(-\vec{x})$, $\vec{h} = \exp(-\vec{y})$ and $U = \exp(W)$ whereby we mean the entry-wise exponentiation. Then:*

$$(\vec{x})^* \dot{\otimes} W \dot{\otimes} \vec{y} = -\log((\vec{v})^* \underset{\cdot}{\times} U \underset{\cdot}{\times} \vec{h}) \qquad (\vec{v})^* \underset{\cdot}{\times} U \underset{\cdot}{\times} \vec{h} = \exp(-(\vec{x})^* \dot{\otimes} W \dot{\otimes} \vec{y})$$

where the dotted notation refers to the $\overline{\mathbb{R}}_{\min,+}$ semifield.

3.3 The FCA in Informational Harmoniums

From Theorem 1 in Section 2.3 we have:

Theorem 2. *Informational harmoniums over the $\overline{\mathbb{R}}_{\max,+}$ semifields are cryptomorphic to $\overline{\mathbb{R}}_{\max,+}$-formal contexts.*

Proof. It is clear that harmoniums are in general bipartite graphs, as shown in Fig. 1. This is one of the isomorphisms of standard formal contexts. With the provisions we made after (1), we can gather the bias information into one extra visible and one extra hidden node v_o and h_o extending the sets of nodes to L' and K'. Therefore we extend W as suggested by including the biases into the incidence W' so that (L', K', W') *is a weighted bipartite graph, that is, a formal context with entries in the carrier set of* $\overline{\mathbb{R}}_{\max,+}$ *or* $\overline{\mathbb{R}}_{\min,+}$, which is the looked-for cryptomorphism. □

If the informational harmonium uses the particular form in (31),

$$\langle \vec{x} \mid W \mid \vec{y} \rangle^{\text{OO}} = (\vec{x})^* \,\dot{\otimes}\, W \,\dot{\otimes}\, \vec{y})$$

this allows us to borrow the results from part 3 of Theorem 1 and we know that there is a left adjunction $(\cdot_W^{\exists}, \cdot_W^{\forall}) : \widetilde{\mathcal{X}} \leftrightharpoons \widetilde{\mathcal{Y}}$ between the spaces through the left adjunct pair of maps:

$$\vec{x}_W^{\exists} = W^* \underset{\cdot}{\otimes} \vec{x} \qquad\qquad \vec{y}_W^{\forall} = W \,\dot{\otimes}\, \vec{y}$$

which define a bijection between the systems of visible nodes $\mathfrak{B}_{L'}(L', K', W')$ and neighbourhoods of hidden nodes $\mathfrak{N}_{K'}(L', K', W')$

$$\mathfrak{B}_{L'}(L', K', W') = (\widetilde{Y})_W^{\forall} \qquad\qquad \mathfrak{N}_{K'}(L', K', W') = (\widetilde{X})_W^{\exists}$$

Regarding learning the harmonium, this type seems to resemble a heteroassociative memory, for if (a, d) is a neighbourhood concept of the pair of lattices then by the definition of the polars we have $W \geq \vec{a} \underset{\cdot}{\otimes} \vec{d}^*$, and in general the harmonium can be built (not efficiently) as the join:

$$W \geq \bigvee_{(\vec{a},\vec{d}) \in \mathfrak{N}(L',K',W')} \vec{a} \underset{\cdot}{\otimes} \vec{d}^* \tag{34}$$

4 Conclusions and Further Research

We have introduced the information harmoniums as a way to “patch” the magnitude problems of harmoniums: the energy function is not written in an algebra of logarithmic quantities, whence the “dimensions” in natural units (probabilities) of the harmonium generative model are incorrect.

By means of defining generalized information and mass functions we have been able to related the informational harmoniums to non-normalized harmoniums defined in positive semifields obtained from the basic $\mathbb{R}_{\geq 0}$ semifield.

If we further concentrate on the informational semifields of order $r = \pm\infty$ we recover the $\overline{\mathbb{R}}_{\max,\times}$, $\overline{\mathbb{R}}_{\min,\times}$, $\overline{\mathbb{R}}_{\max,+}$ and $\overline{\mathbb{R}}_{\min,+}$. When the harmoniums are described with these semirings they relate to one of the four types of Galois connection definable over the matrix of weights (W or U): they are the scalar products that defined the Galois connections, so that the posterior "mass measures" and "information measures" are simply the polars.

This opens up a number of avenues of research into the representation of the spaces associated to $r = \pm\infty$-harmoniums, and suggest that $\overline{\mathcal{K}}$-FCA has points to make relating to inference and learning of these, including the consideration of the other three types of connections between spaces. Also, more efficient ways to build harmoniums, as well as approximate building in the absence of the supervision provided by the intents, $\vec{d}$ will be explored in future work.

References

[1] Freund, Y., Haussler, D.: Unsupervised learning of distributions on binary vectors using two layer networks. In: Advances in Neural Information Processing Systems 4. (1992) 912–919

[2] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015) 436–444

[3] Smolensky, P.: Information Processing in Dynamical Systems: Foundations of Harmony Theory. In: Parallel Distributed Processing. MIT Press (1986) 194–281

[4] Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. Cognitive Science (1985)

[5] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M.A., Huang, F.J.: A tutorial on energy-based learning. In Bakir, G., Hofman, T., Schölkopf, B., Smola, A., Taskar, B., eds.: Predicting Structured Output. MIT Press (2006)

[6] Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14** (2002) 1771–1800

[7] Hinton, G.E.: A Practical Guide to Training Restricted Boltzmann Machines. In Montavon, G., Orr, G.B., Müller, K.R., eds.: Neural Networks: Tricks of the Trade. Volume 7700 of LNCS. 2nd edn. Springer, Berlin, Heidelberg (2012) 599–619

[8] Murphy, K.P.: Machine Learning. A Probabilistic Perspective. MIT Press (2012)

[9] Cueto, M.A., Morton, J., Sturmfels, B.: Geometry of the restricted Boltzmann machine. Contemporary Mathematics **516** (2010) 135–153

[10] Freund, Y., Haussler, D.: Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks. Technical Report 94 25, UCSC CRL (June 1994)

[11] Welling, M., Rosen-Zvi, M., Hinton, G.E.: Exponential Family Harmoniums with an Application to Information Retrieval. In: Advances in Neural Information Processing Systems. (2005)

[12] Valverde-Albacete, J.F., Peláez-Moreno, C.: The Rényi Entropies Operate in Positive Semifields. Entropy **21**(8) (2019)

[13] Mesiar, R., Pap, E.: Idempotent integral as limit of g-integrals. Fuzzy Sets And Systems **102**(3) (1999) 385–392

[14] Moreau, J.J.: Inf-convolution, sous-additivité, convexité des fonctions numériques. J. Math. Pures et Appl. **49** (1970) 109–154

[15] Renyi, A.: Probability Theory. Courier Dover Publications (1970)

[16] Valverde-Albacete, F.J., Peláez-Moreno, C.: The Linear Algebra in Extended Formal Concept Analysis Over Idempotent Semifields. In Bertet, K., Borchmann, D., Cellier, P., Ferré, S., eds.: Formal Concept Analysis. Springer Berlin Heidelberg, Rennes (June 2017) 211–227

[17] Mackay, D.M.: Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws. Science **139**(3560) (1963) 1213–1216