Directed Metric Structures arising in Large Language Models

Stéphane Gaubert, Yiannis Vlassopoulos

DIRECTED METRIC STRUCTURES ARISING IN LARGE LANGUAGE MODELS

STÉPHANE GAUBERT AND YIANNIS VLASSOPOULOS

Abstract. Large Language Models are transformer neural networks which are trained to produce a probability distribution on the possible next words to given texts in a corpus, in such a way that the most likely word predicted is the actual word in the training text. In this paper we find what is the mathematical structure defined by such conditional probability distributions of text extensions. Changing the view point from probabilities to -log probabilities we observe that the subtext order is completely encoded in a metric structure defined on the space of texts $\mathcal{L}$, by -log probabilities. We then construct a metric polyhedron $P(\mathcal{L})$ and an isometric embedding (called Yoneda embedding) of $\mathcal{L}$ into $P(\mathcal{L})$ such that texts map to generators of certain special extremal rays. We explain that $P(\mathcal{L})$ is a $(\min, +)$ (tropical) linear span of these extremal ray generators. The generators also satisfy a system of $(\min +)$ linear equations. We then show that $P(\mathcal{L})$ is compatible with adding more text and from this we derive an approximation of a text vector as a Boltzmann weighted linear combination of the vectors for words in that text. We then prove a duality theorem showing that texts extensions and text restrictions give isometric polyhedra (even though they look a priory very different). Moreover we prove that $P(\mathcal{L})$ is the lattice closure of (a version of) the so called, Isbell completion of $\mathcal{L}$ which turns out to be the $(\max, +)$ span of the text extremal ray generators. All constructions have interpretations in category theory but we don't use category theory explicitly. The categorical interpretations are briefly explained in an appendix. In the final appendix we describe how the syntax to semantics problem could fit in a general well known mathematical duality.

Contents

Date: May 20, 2024.

1. Overview

Large Language Models (LLM) are transformer neural networks that are trained to compute the probability of the possible next words to a text in such a way that the most probable next word predicted by the network, is the actual next word in the training text [20, 15, 16, 2].

They are often characterized as "just statistical models". In this paper, continuing the approach introduced in [1], this time without explicitly using categories [1], we would like to make a proposal for what is the underlying mathematical structure these probability distributions actually encode and show the evidence and possible consequences. We find that a rich structure is revealed if we change the point of view from probabilities to negative log probabilities. While it is entirely equivalent, this point of view can be reinterpreted as an asymmetric metric on texts.

Indeed consider a set $\mathcal{L} := \{a_1, \dots, a_n\}$ whose elements are texts in the language. Equip $\mathcal{L}$ with a poset structure, where $a_i \leqslant a_j$ if and only if a_i is a subtext of a_j. Denote by $\Pr(a_j|a_i)$ the probability of extending a text a_i to a text a_j. The probability is 0 exactly when a_i is not a subtext of a_j. Our main assumption is that conditional probabilities of extensions of texts multiply i.e.

$$\text{if } a_i \leqslant a_j \leqslant a_k \text{ then } \Pr(a_j|a_i)\Pr(a_k|a_j) = \Pr(a_k|a_i). \tag{1}$$

We call the triple $(\mathcal{L}, \leqslant, \Pr)$ a *probabilistic language model*.

Next, recall the notion of a *directed metric* δ on a set X. It is defined to be a function $\delta : X \times X \to (-\infty, \infty]$ which satisfies the triangle inequality and $\delta(a, a) = 0$. Nevertheless, unless stated otherwise it does not have to be symmetric, $\delta(a, b) = 0$ does not necessarily imply $a = b$ and δ can take the value ∞ and can also take negative values. [2]

Now we notice that the probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$ defines a directed metric d on the poset $\mathcal{L}$ by

$$d(a_i, a_j) = \begin{cases} -\log \Pr(a_j|a_i) & \text{if } a_i \leqslant a_j, \\ \infty & \text{if } a_i \text{ and } a_j \text{ are not comparable.} \end{cases} \tag{2}$$

$(\mathcal{L}, d)$ is then a directed metric space.

Indeed

$$\text{if } a_i \leqslant a_j \leqslant a_k \text{ then } d(a_i, a_k) = d(a_i, a_j) + d(a_j, a_k) \tag{3}$$

[1] We note though that all constructions and theorems have a categorical interpretation. For those already familiar with categories Appendix A provides a brief categorical explanation of constructions and results in this paper.

[2] We will see however that restricting to positive values is natural when the directed metric comes from probability distributions.

and otherwise the triangle inequality is satisfied with at least one side being ∞ (Proposition 1). We see also that the metric d fully determines the poset structure on $\mathcal{L}$ (Corollary 1).

Although d satisfies a rather degenerate form of the triangle inequality, it is enough to define a non trivial directed metric polyhedron $P(\mathcal{L})$ in $(\mathbb{R} \cup \{\infty\})^n$, inside which $\mathcal{L}$ is isometrically embedded as a remarkable set of extremal rays.

Indeed, in Section 3 we define

$$P(\mathcal{L}) := \{x \in (\mathbb{R} \cup \{\infty\})^n \backslash (\infty, \dots, \infty) | x_i \leqslant x_j + d(a_i, a_j)\} \tag{4}$$

and thus the finite part of $P(\mathcal{L})$ is a polyhedron in $\mathbb{R}^n$. Define the Funk directed metric D on $(\mathbb{R} \cup \{\infty\})^n \backslash (\infty, \dots, \infty)$ by

$$D(x, x') := \max_i \{x'_i - x_i | x_i \neq \infty\}. \tag{5}$$

Then $(P(\mathcal{L}), D)$ becomes a directed metric space. The Funk directed metric originates from Hilbert geometry [14].

To understand the relevance of $P(\mathcal{L})$ note first (Proposition 4) that

$$Y : (\mathcal{L}, d) \hookrightarrow (P(\mathcal{L}), D), \text{ given by } Y(a_k) := d(-, a_k), \tag{6}$$

is an isometric embedding (called the Yoneda embedding). Moreover, each $Y(a_k)$ is a generator of an extremal ray of $P(\mathcal{L})$ (Theorem 1).

To be precise about the term extremal ray here, consider the image $Q(\mathcal{L})$ of $P(\mathcal{L})$, under the coordinate wise map $x_i \to z_i := e^{-x_i}$, i.e.

$$Q(\mathcal{L}) := \{z := (z_1, \dots z_n) \in [0, \infty)^n | z_i := e^{-x_i} \text{ for } x = (x_1, \dots, x_n) \in P(\mathcal{L})\} \tag{7}$$

We see that

$$Q(\mathcal{L}) := \{z = (z_1, \dots z_n) \in [0, \infty)^n \backslash (0, \dots, 0) | z_i \geqslant \Pr(a_j | a_i) z_j\} \tag{8}$$

Denote by $e^{-Y(a_k)}$, the image of $Y(a_k)$ in $Q(\mathcal{L})$. Then (Theorem 1) $e^{-Y(a_k)}$ is a generator of a usual extremal ray in the polyhedral cone $Q(\mathcal{L})$. When we speak of extremal rays of $P(\mathcal{L})$ we always mean the subsets of $P(\mathcal{L})$ that map to extremal rays in the polyhedral cone $Q(\mathcal{L})$ by the map $(x_i) \mapsto (e^{-x_i})$.

The polyhedral cone $Q(\mathcal{L})$ is a generalization of the *order polytope* studied by Stanley [18]. The order polytope corresponds to the case where $\Pr(a_j|a_i)$ takes only the values 0 or 1 and $\mathcal{L}$ is simply a poset. Moreover, Stanley does not consider a cone, but rather the intersection of this cone with the unit box. Lam and Postnikov [10] defined an *alcoved polytope* to be a bounded cell of a Coxeter arrangement (of type A_n). The definition of $P(\mathcal{L})$ is similar, but we do not require the cell to be bounded. Alcoved polytopes have been studied in tropical geometry, in relation with metric spaces, see e.g. [9, 19].

Moving on, we prove in Proposition 5 that

$$\text{if } x = (x_1, \dots, x_n) \in P(\mathcal{L}) \text{ then } x_k = D(Y(a_k), x). \tag{9}$$

We now see that the defining equations for $P(\mathcal{L})$ are exactly the triangle inequalities for D. Indeed

$$x_i \leqslant x_j + d(a_i, a_j) \iff D(Y(a_i), x) \leqslant D(Y(a_j), x) + D(Y(a_i), Y(a_j)). \tag{10}$$

Note also (Proposition 3) that we can think of the points $x \in P(\mathcal{L})$ as functions on $\mathcal{L}$ (just as we can think of usual vectors as functions on a set). Indeed, if we

denote by $d_{\mathbb{R}}$ the Funk metric on $\mathbb{R}\cup\{\infty\}$, namely $d_{\mathbb{R}}(s,t)=t-s$ and $d_{\mathbb{R}}(\infty,\infty)=\max\emptyset=-\infty$ [3] then

$$P(\mathcal{L})=\{x:(\mathcal{L},d^t)\to((-\infty,\infty],d_F)|x \text{ is non-expansive.}\}$$

We then have that $x(a_i)=x_i=D(Y(a_i),x)$ (Proposition 5).

Therefore $P(\mathcal{L})$ can also be thought of as the set of maps $x:\mathcal{L}\to(-\infty,\infty]$ which satisfy the triangle inequalities for the metric D with respect to all the maps $Y(a_k):=d(-,a_k):\mathcal{L}\to(-\infty,\infty]$ for $k=1,\dots,n$. It is therefore a kind of convex metric span of the $Y(a_k):=d(-,a_k)$.

From the point of view of *language semantics now, we consider* $Y(a_k)=d(-,a_k)$ *to represent the meaning of a text* a_k *in terms of all the texts it contains,* (Section 4.1) in accordance with the statistical semantics principal and as was advocated in [1].

Dually, we can consider the meaning of a text a_k to be given by all the texts extending a_k, namely $d(a_k,-)$. This is then encoded in

$$\widehat{P}(\mathcal{L})=\{y\in(\mathbb{R}\cup\{\infty\})^n|y_i\leqslant y_j+d(a_j,a_i)\}. \tag{11}$$

Indeed we have the isometric co-Yoneda embedding

$$\widehat{Y}:\mathcal{L}\hookrightarrow P(\mathcal{L}) \text{ given by } \widehat{Y}(a_k):=d(a_k,-) \tag{12}$$

and moreover $y_i=D(\widehat{Y}(a_i),y)$.

Remark 1. We said that $Y(a_k):=d(-,a_k)$ or $\widehat{Y}(a_k):=d(a_k,-)$ represent the meaning of a text a_k but it's also the "location" of these vectors in the whole space $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ respectively. In particular if $w:=a_k$ is a word then $Y(w):=d(-,w)$ is supported only on w so it does not seem to contain much information. However the relevant semantic information is in $D(Y(w),-)$. Moreover we will see shortly that the fact that the vector $d(-,w)$ is in $P(\mathcal{L})$ means that it satisfies a whole system of equations (Eq 19, 20, Proposition 11) with respect to other texts. We will explain this system of equations, later in this overview when we describe section 4.

We already saw that $Y(a_k)$ is an extremal ray in $P(\mathcal{L})$ but it turns out that $P(\mathcal{L})$ has generally exponentially many additional extremal rays – that are not in the image of Y. We explicitly characterize the extremal rays of $P(\mathcal{L})$ as corresponding to connected lower sets of $(\mathcal{L},\leqslant)$ in Proposition 6 and Theorem 2. In fact, if x is such an extremal ray and $l(x)$ denotes the corresponding lower set then we show that after a diagonal change of variables the new coordinates of the extremal ray give the characteristic function of the lower set $l(x)$.

What distinguishes the lower sets corresponding to elements $Y(a_i)$ is that they are principal.

Therefore $P(\mathcal{L})$ *can be considered as a space parameterizing semantics in the language. However, only the extremal rays corresponding to principal lower sets of* $\mathcal{L}$*, correspond to texts.*

Notice now that the Funk metric D on $P(\mathcal{L})$ defines a metric D_Q on the polyhedral cone $Q(\mathcal{L})$ where, if $z,z'\in Q(\mathcal{L})$, then

$$D_Q(z,z'):=\max_i\{\log(\frac{z_i}{z'_i})|z'_i\neq 0\}. \tag{13}$$

[3] We explain later in Remark 2 that it is possible and useful in some cases to extend the values of a directed metric to $[-\infty,\infty]$ and this is one of the cases we do so.

By definition we have $D_Q(z,z') = D(-\log z, -\log z')$ and $D(x,x') = D_Q(e^{-x}, e^{-x'})$. Then the fact that Y is an isometric embedding into $P(\mathcal{L})$ implies that

$$e^{-Y} : (\mathcal{L}, d) \to (Q(\mathcal{L}), D_Q) \tag{14}$$

is an isometric embedding.

In section 4 we explain the construction of $P(\mathcal{L})$ in terms of tropical or $(\min,+)$ algebra. Recall that the $(\min,+)$ semifield $\mathbb{R}_{\min}$, is defined as $((-\infty,\infty], \min, +)$. We show in Section 4 that $P(\mathcal{L})$ is generated by the vectors $Y(a_k) = d(-,a_k)$ as a $(\min,+)$ module. To see that, note first that d is a directed metric if and only if it is a $(\min,+)$ projector. This is because if we let $d_{i,j} := d(a_i,a_j)$ and define

$$d_{\min}(x)_i := \min_j\{d_{i,j} + x_j\}, \tag{15}$$

then the triangle inequality and the fact that $d_{i,i} = 0$, are equivalent to $d_{i,k} = \min\{d_{i,j} + d_{j,k}\}$ which is equivalent to $d_{\min}^2 = d_{\min}$.

We then note that $P(\mathcal{L}) = \{x | d_{\min} x = x\}$. Indeed: $x = d_{\min}x \iff x_i = \min_j\{d_{i,j} + x_j\} \iff x \in P(\mathcal{L})$.

Introduce the notation $\mathrm{Fix}(d_{\min}) := \{x | d_{\min}x = x\}$. Since $d_{\min}$ is a projection $\mathrm{Im}(d_{\min}) = \mathrm{Fix}(d_{\min})$. Therefore

$$P(\mathcal{L}) = \mathrm{Fix}(d_{\min}) = \mathrm{Im}(d_{\min}). \tag{16}$$

We see that $P(\mathcal{L})$ is the $(\min,+)$ (tropical) span of the columns of the matrix d. Analogously, if we denote by d^t the transpose of d, we see that $\widehat{P}(\mathcal{L}) = \mathrm{Im}(d^t_{\min}) = \mathrm{Fix}(d^t_{\min})$ and therefore $\widehat{P}(\mathcal{L})$ is the $(\min,+)$ row span of d. We let $(u \oplus v)_i := \min\{u_i, v_i\}$ and $(\lambda \odot v)_i := \lambda + v_i$. Then for $x \in P(\mathcal{L})$ we have

$$x = \oplus_j x_j \odot d(-,a_j) = \oplus_j D(Y(a_j), x) \odot d(-,a_j) = \oplus_j D(Y(a_j), x) \odot Y(a_j). \tag{17}$$

and an analogous formula holds for $z \in \widehat{P}(\mathcal{L}) = \mathrm{Im}(d^t_{\min})$. From these we get that (Proposition 10)

$$Y(a_k) = d(-,a_k) = \oplus_{a_j \leqslant a_k} d_{j,k} \odot Y(a_j) \tag{18}$$

and

$$\widehat{Y}(a_k) = d(a_k,-) = \oplus_{a_k \leqslant a_l} d_{k,l} \odot \widehat{Y}(a_l). \tag{19}$$

These are the systems of equations we referred to in Remark 1.

In section 5 we study how $P(\mathcal{L})$ changes when we enlarge the language corpus $\mathcal{L}$. We prove that if a probabilistic language model $(\mathcal{L}_1, d_1)$ is extended to $(\mathcal{L}_2, d_2)$, namely if there is an isometric embedding $\phi : (\mathcal{L}_1, d_1) \hookrightarrow (\mathcal{L}_2, d_2)$ then there is an isometric embedding $\widetilde{\phi} : (P(\mathcal{L}_1)), D_1) \hookrightarrow (P(\mathcal{L}_2), D_2)$ such that $\widetilde{\phi}(Y_1(a)) = Y_2(\phi(a))$. Moreover there is a non-expansive, $(\min,+)$ projection $\mathcal{R} : P(\mathcal{L}_2) \to P(\mathcal{L}_2)$ such that $\mathrm{Im}(\mathcal{R}) = \widetilde{\phi}(P(\mathcal{L}_1))$.

Using this we show that if $\mathcal{L}_1 := \{w_1, \ldots, w_l\}$ is the set of words in the language and b is a text in $\mathcal{L}$ then $\mathcal{R}(Y(b)) = \bigoplus_{w_i \leqslant b} d_2(w_i, b) \odot Y_2(w_i)$ Introducing a temperature parameter T we get

$$\mathcal{R}(Y(b)) = \lim_{T \to 0} -T \log\Big(\sum_{w_i \leqslant b} e^{-\frac{d(w_i,b)}{T}} e^{-\frac{Y(w_i)}{T}}\Big) \tag{20}$$

Therefore for small T we have $e^{-\frac{\mathcal{R}(Y(b))}{T}} \approx \sum_{w_i \leqslant b} e^{-\frac{d(w_i,b)}{T}} e^{-\frac{Y(w_i)}{T}}$.

Putting $v_i \coloneqq e^{-Y(w_i)}$ we have

$$e^{-\frac{\mathcal{R}(Y(b))}{T}} \approx \sum_i e^{-d(Y(w_i),b)/T} v_i \tag{21}$$

This approximation of text vectors is similar to the one calculated by transformer neural networks in the self attention module.

In section 6 we describe a duality between semantics via texts extensions and text restriction. Indeed we have already seen that $P(\mathcal{L}) = \{x | d_{\min} x = x\}$ is the $(\min, +)$ column span of d and $\widehat{P}(\mathcal{L}) = \{x | d^t_{\min} x = x\}$ is the $(\min, +)$ row span of d. It is easy to see that $d_{\min} x = x \iff d^t_{\min}(-x) = -x$ (Proposition 19). Indeed it follows from the fact that $x_i \leqslant d_{i,j} + x_j \iff -x_j \leqslant d_{i,j} - x_i$. However this requires extending the $(\min, +)$ semifield, as well as the values of the directed metric, to $[-\infty, \infty]$. This results in the definition of extended $(\min, +)$ modules $P^-(\mathcal{L})$ and $\widehat{P}^-(\mathcal{L})$.

We show that $(P^-(\mathcal{L}), D)$ and $(\widehat{P}^-(\mathcal{L}), D^t)$ are isometric (and in fact tropically anti-isomorphic). We interpret this as saying that *the semantic space $\widehat{P}^-(\mathcal{L})$, defined by extensions of texts is isomorphic to the semantic space $P^-(\mathcal{L})$ defined by restrictions of texts.* We note that this is quite a non-trivial isomorphism as $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ don't even have the same number of extremal rays in general. We give an example (Example 1) illustrating the polyhedra $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$. In fact we consider the corresponding polyhedral cones $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$. We then define

$$Q_0(\mathcal{L}) := Q(\mathcal{L}) \cap \Delta \tag{22}$$

to be the intersection of $Q(\mathcal{L})$ with the unit simplex Δ. Points in $Q_0(\mathcal{L})$ are normalized to probability distributions. Analogously define $\widehat{Q}_0(\mathcal{L}))$ to be the intersection of $\widehat{Q}(\mathcal{L})$ with the unit simplex. Extremal rays of $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L}))$ define vertices of $Q_0(\mathcal{L})$ and $\widehat{Q}_0(\mathcal{L})$ respectively. We show the correspondence of extremal rays to lower sets and upper sets. The example also showcases the difference between a probabilistic language model and a general directed metric space where infinite distances are approximated uniformly by a big number M.

Section 7 further explores the extremal rays of $P(\mathcal{L})$.

In section 8 we explore the relation with the so called Isbell completions $I(\mathcal{L})$ and $\widehat{I}(\mathcal{L})$. This is similar to the duality of section 6. The Isbell adjunction is defined over the extended ring $[-\infty, \infty]$ and the fixed parts of the adjunction turn out to be $I(\mathcal{L}) = \mathrm{Im}(d_{\max})$ where $d_{\max}(x)_i := \max_j \{d_{i,j} + x_j\}$ and $\widehat{I}(\mathcal{L}) = \mathrm{Im}(d^t_{\max})$. In (Proposition 29) it is shown that $P(\mathcal{L})$ is the lattice completion of the so called Isbell completion.

When restricting coefficients to $[0, \infty]$, the Isbell completion has been studied by Willerton [21] where it is proven to be isomorphic to the directed tight span $DTS(\mathcal{L})$ of Hirai and Koichi [7] (inspired by the undirected tight span defined by Isbell [8] and Dress [5]). It also generalizes the Dedekind-MacNeille completion of a poset.

Informally speaking $I(\mathcal{L})$ can be though of as the minimal space in which we can isometrically embed $\mathcal{L}$. Unlike $P(\mathcal{L})$ though, it is far from being convex. A simple example is shown in section 8.

Section 9 is a collection of observations about probabilistic language models and their relation to transformers.

As mentioned already, all constructions and results in this paper have categorical interpretations and though we have avoided using categorical language in the main text we explain briefly in Appendix A, for the benefit of readers familiar with categories, these categorical interpretations.

Finally in Appendix B we present a general perspective which locates the language syntax and semantics problems in the realm of a basic duality in mathematics which in its simplest form appears as a duality between algebra and geometry. This allows us to locate future directions of research.

Experimental evidence for the semantic meaning of the (co-)Yoneda embedding vectors $\widehat{Y}(a_k)$ has been provided in [12]. There experiments based on (a slight variation of) the co-Yoneda embedding vectors, were performed using an actual Transformer LLM, by sampling over continuations of texts. Several semantic tests were conducted and the results were in general very good.

1.1. **Acknowledgements.** YV would like to thank Tai-Danae Bradley, Michael Douglas, Ioannis Emiris, Harris Papageorgiou, Alex Takeda, John Terilla, Matthew Trager, Maxim Kontsevich, Matilde Marcolli, Jack Morava, Stefano Soatto, and Elias Zafiris for useful conversations. He also would like to thank Anna Geneveaux for computing several useful examples of polyhedra for probabilistic language models during her internship. SG and YV thank Gleb Koshevoy and Panayotis Mertikopoulos for useful conversations. Finally YV would like to thank IHES for providing excellent working conditions.

2. From probabilities of text extensions to distances

Consider a language with a set of words $W := \{w_1, \dots, w_l\}$. Consider also a set of training texts from the language, $\mathcal{L} := \{a_0, a_1, \dots, a_n\}$ where $a_i := w_{i_1} \dots w_{i_{k_i}}$. We endow $\mathcal{L}$ with a poset structure where $a_i \leqslant a_j$ if and only if a_i is a subtext of a_j. We consider two possibilities for the notion of subtext. The first is

$$a_i \leqslant_1 a_j \iff \exists a_k \in \mathcal{L} \text{ such that } a_j = a_i a_k \tag{23}$$

and we refer to this as the *one sided subtext order* and the second is

$$a_i \leqslant_2 a_j \iff \exists a_{k_1}, a_{k_2} \in \mathcal{L} \text{ such that } a_j = a_{k_1} a_i a_{k_2} \tag{24}$$

and we refer to that as the *two sided subtext order.*

We define always a_0 to be the empty text and a_0 is the only text such that $a_0 \leqslant a_i \forall i$ in either order. (However see remark (1) bellow for how a_0 interacts with the probabilities we will soon add to the model.)

If $a_i \leqslant_1 a_j$ in the one sided subtext order then $a_i \leqslant_2 a_j$. The results and constructions that follow hold equally for both orders so we will simply write $a_i \leqslant a_j$ and when there is need to separate the two orders we will make a special comment.

If $a_i \leqslant a_j$ then denote by $\Pr(a_j|a_i)$ the probability of extension from a_i to a_j.

It is important to note that these probabilities are not calculated from a corpus of texts, as any probability for a sufficiently long text would be vanishingly small. Instead we are talking about the probabilities that the large language model (LLM) computes. Namely prompted with a text a_i the model outputs a probability distribution $\Pr(a_i w_{j_1}|a_i) \forall w_{j_1} \in W$ and this is the probabilities we are referring to, above. To continue extending to $a_i w_{j_1} w_{j_2}$ we simply have

$$\Pr(a_i w_{j_1} w_{j_2}|a_i) = \Pr(a_i w_{j_1}|a_i) \Pr(a_i w_{j_1} w_{j_2}|a_i w_{j_1}) \tag{25}$$

And continuing this way $\Pr(a_j|a_i)$ is computed. If a_i is not a subtext of a_j then we put $\Pr(a_j|a_i) = 0$.

Recall that the LLM is trained to produce the probability distribution of the next word to a text, in such a way that the most likely next word predicted by the model is the one in the training text.

As a consequence of Equation (25) *we make our fundamental assumption that*

$$a_i \leqslant a_j \leqslant a_k \implies \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i) \tag{26}$$

Note that the transformer LLM produces the probabilities for the one-sided subtext order. However in the attention layers of the transformer, two sided extensions are used in order to construct the text vector. We consider therefore the case of the two sided order as well. Indeed in section 5 we will see that the text vector we define, is expressed in terms of word vectors when we consider the two sided subtext order. We put these together in the following

Definition 1. *A probabilistic language model is a triple* $(\mathcal{L}, \leqslant, Pr)$ *where,* $\mathcal{L} := \{a_0, a_1, \ldots, a_n\}$ *is a collection of texts,* $\leqslant$ *is the subtext order and* $\Pr : \mathcal{L}\times\mathcal{L} \to [0,1]$ *is a function such that* $a_i \leqslant a_j \leqslant a_k \implies \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i)$.

Recall now the following

Definition 2. (X, δ) *is called a directed metric space if* X *is a set and* $\delta : X \times X \to (-\infty, \infty]$ *satisfies the triangle inequality*

$$\delta(a,c) \leqslant \delta(a,b) + \delta(b,c) \tag{27}$$

for all $a, b, c \in X$ *and* $\delta(a,a) = 0, \forall a \in X$

Note that this generalises usual metrics in that we don't require $\delta(a,b) = \delta(b,a)$, $\delta(a,b) = 0$ does not necessarily imply $a = b$ and moreover we allow negative values. This definition of a directed metric, in the special case of positive valued δ, has appeared in [21, 11] and is also known as a generalised metric or a pseudo quasi metric.

Remark 2. We need the following technical specification: In three cases (proposition 3 and the duality theorems of sections 6 and 8) we will need to extend definition 2, of a directed metric to allow the value $-\infty$ so that we will have. $\delta : X \times X \to [-\infty, \infty]$. In that case the definition is the same but we need to specify that we use the convention that $+\infty$ is the absorbing element so that $s + (+\infty) = +\infty$ for all s and in particular $-\infty + \infty = +\infty$. This will be needed in Proposition 2. In section 6 we will explain further that this is the so called $(\min, +)$ convention, as this is the only one compatible with the structure of $(\min, +)$ semiring; there is also a dual $(\max, +)$ convention.

We define now a directed metric space structure on the underlying poset of a probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$.

Definition 3. *Given the probabilistic language model* $(\mathcal{L}, \leqslant, \Pr)$ *where* $\leqslant$ *is the subtext order and* $\Pr(a_j|a_i)$ *are the probabilities of extension, define the directed metric* $d : \mathcal{L} \times \mathcal{L} \to [0, \infty]$ *by*

$$d(a_i, a_j) = \begin{cases} -\log \Pr(a_j|a_i) & \text{if } a_i \leqslant a_j, \\ \infty & \text{if } a_i \text{ and } a_j \text{ are not comparable.} \end{cases} \tag{28}$$

It is clear that $d(a_i, a_i) = 0$. To verify that d is a directed metric we have the following:

Proposition 1. *The map d satisfies the triangle inequality:*

$$d(a_i, a_k) \leqslant d(a_i, a_j) + d(a_j, a_k) \ , \tag{29}$$

and equality holds if and only if $a_i \leqslant a_j \leqslant a_k$ or $a_i \not\leqslant a_k$.

Proof. Indeed, if $a_i \leqslant a_j \leqslant a_k$ then $a_i \leqslant a_k$, and the equality holds in (29), since by our main assumption (the standard property of conditional probabilities), $\Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i)$. If $a_i \not\leqslant a_k$, then, $d(a_i, a_k) = \infty$, and either $a_i \not\leqslant a_j$ or $a_j \not\leqslant a_k$, which entails that both sides of (29) are equal to infinity. Finally, if $a_i \leqslant a_k$ but $a_i \not\leqslant a_j$ or $a_j \not\leqslant a_k$, the left-hand side of (29) is finite whereas the right-hand side is $+\infty$. □

We then have

Corollary 1. *The following statements are equivalent:*

1. $a_i \leqslant a_j \leqslant a_k$
2. $d(a_i, a_k) = d(a_i, a_j) + d(a_j, a_k)$ *and* $d(a_i, a_k) < \infty$

Remark 3. Note that from Proposition 1 and Corollary 1 it follows that the partial order $\leqslant$ on $\mathcal{L}$ can be fully recovered by the directed metric d or equivalently the conditional probabilities $\Pr(a_j|a_i)$, therefore we will also denote the probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$ as $(\mathcal{L}, \Pr)$ or $(\mathcal{L}, d)$.

Remark 4. In a Large Language model probabilities are normalized to add up to one, over all extensions of a given text by a word.

Remark 5. Note that the probabilistic language model $(\mathcal{L}, d)$ is a special case of a directed metric space. Whenever it is possible we will prove results for a general directed metric space and derive the language case as a corollary. Moreover, it is possible to imagine that even the main assumption $a_i \leqslant a_j \leqslant a_k \implies \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i)$ should be generalized to $\Pr(a_k|a_i) \geqslant \Pr(a_k|a_j)\Pr(a_j|a_i)$, namely this of a general directed metric space with $d(a_i, a_k) \leqslant d(a_i, a_j) + d(a_j, a_k)$. This is a reasonable assumption by itself and can be interpreted a saying that the shortest path to go from a_i to a_k is at least as short as a path that is forced to go from a_i to a_k but passing though a_j. As we will see in what follows, the only result that requires the main assumption of conditional probabilities multiplying is Theorem 2 and all the rest are valid for general directed metric spaces. It is a matter of experimental verification to check for a given LLM if the multiplicative assumption is best or the general case.

Remark 6. Note that we can slightly modify the definition of the Probabilistic Language model so that instead of Pr taking values in $[0, 1]$ we put $\Pr : \mathcal{L} \times \mathcal{L} \to [0, \infty)$. Then definition 3 will again produce a directed metric space. In fact we develop most of the theory using the more general extended assumption since most results are valid for general directed metric spaces as we mentioned in the previous remark.

Remark 7. Since the machine produces probabilities for all possible next words it is natural to assume it is learning probabilities of extension for the free monoid generated by words. Obviously most strings of words will have vanishing probability and only those which are part of the language should have big probability.

We can then consider $\mathcal{L}$ to contain the whole free monoid and it is natural to grade it by the word length of each text.

Remark 8. Note that if we assume that there exists a_0 such that $a_0 \leqslant a_k \forall a_k \in \mathcal{L}$ then $a_i \leqslant a_j$ implies $a_0 \leqslant a_i \leqslant a_j$ and therefore $d(a_0, a_j) = d(a_0, a_i) + d(a_i, a_j)$ and thus $d(a_i, a_j) = d(a_0, a_j) - d(a_0, a_i)$. This is equivalent to the statement that there is a globally defined probability distribution for absolute probabilities of texts, giving rise to all the conditional probabilities. Namely if $a_i \leqslant a_j$ then $\Pr(a_j|a_i) = \frac{\Pr(a_j|a_0)}{\Pr(a_i|a_0)}$. The element a_0 can be considered to be the empty text and from this point of view it is natural to assume it exists in $\mathcal{L}$. However the fact that it implies all conditional probabilities come from a global probability distribution shows that the inclusion of a_0 in the probabilistic language model, is not an entirely trivial assumption.

It would be a matter for experimental verification to see if it applies in the transformer Large Language Models. Therefore we will not assume it by default. We will specify explicitly whenever we assume $a_0 \in \mathcal{L}$.

Next, we illustrate what the main assumption implies by the following:

Proposition 2. *Consider a probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$ then on every connected component C of the Hasse diagram of $\mathcal{L}$, there is a function $P_C : C \to [0, \infty)$ such that if $a_i, a_j \in C$ and $a_i \leqslant a_j$ then $\Pr(a_j|a_i) = \frac{P_C(a_j)}{P_C(a_j)}$. The function P_C is unique up to multiplication by a positive number.*

Proof. The fact that $(\mathcal{L}, \leqslant, \Pr)$ is a probabilistic language model means that

$$a_i \leqslant a_j \leqslant a_k \text{ is equivalent to } \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i) \text{ and } \Pr(a_k|a_i) < \infty.$$

Let G denote the directed graph which is the Hasse diagram of C. We construct a new weighted graph $\tilde{G}$ as follows: If $a_i \leqslant a_j$, we draw an arrow from node a_i to node a_j with weight $\Pr(a_j|a_i)$. If $a_j \leqslant a_i$, we draw an arrow from node a_i to node a_j with weight $\Pr(a_j|a_i)^{-1}$. We now choose arbitrarily an element $c \in C$. If $a_i \in C$, we define $P_C(a_i)$ to be the weight in the graph $\tilde{G}$ of an arbitrary path from the point c to a_i. Owing to our main assumption, (1), the weight is independent of the choice of the path from c to a_i. Moreover, for all i, j such that $a_i \leqslant a_j$, we have $P_C(a_i)\Pr(a_j|a_i)P_C(a_j)^{-1} = 1$ therefore $\Pr(a_j|a_i) = \frac{P_C(a_j)}{P_C(a_i)}$.

Picking a different reference element $c' \in C$ scales $P_C(a_i)$ by $P_C(c')$ therefore the ratio stays the same. □

3. From the text metric space $\mathcal{L}$ to the polyhedra $P(\mathcal{L})$ and $Q(\mathcal{L})$

First notice that we can also equip $\mathcal{L}$ with the transpose directed metric d^t where $d^t(a_i, a_j) := d(a_j, a_i)$.

We now construct two directed metric, polyhedra $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ in which the directed metric space $(\mathcal{L}, d)$ is isometrically embedded as a special set of extremal rays.

To that end, we equip $\{\mathbb{R} \cup \{\infty\}\}^n \backslash \{(\infty, \dots, \infty)\}$ for $n \geqslant 2$, with the *Funk* metric D defined by

$$D(x, y) \coloneqq \inf\{\lambda \in \mathbb{R} \cup \{+\infty\} \mid \lambda + x \geqslant y\} = \max_i\{y_i - x_i \mid x_i \neq \infty\} \ . \tag{30}$$

This is a directed metric. Note that it takes possibly negative values, and that it can also take the value ∞.

We also denote by D^t the transpose directed metric with $D^t(x, y) := D(y, x)$.

Definition 4. *Let $(P(\mathcal{L}), D)$ be the directed metric polyhedron*

$$P(\mathcal{L}) := \{x = (x_1, \ldots, x_n) \in \{\mathbb{R} \cup \{\infty\}\}^n \backslash \{(\infty, \ldots, \infty)\} | x_i \leqslant x_j + d_{i,j}\}. \tag{31}$$

Moreover let $(\widehat{P}(\mathcal{L}), D^t)$ be the directed metric polyhedron

$$\widehat{P}(\mathcal{L}) := \{y = (y_1, \ldots, y_n) \in \{\mathbb{R} \cup \{\infty\}\}^n \backslash \{(\infty, \ldots, \infty)\} | y_i \leqslant y_j + d_{j,i}\}. \tag{32}$$

For the following proposition, we need to extend the Funk metric in eq. 30 to $n = 1$. For this we will use the fact that $\max \emptyset = -\infty$. From this it follows that for $n = 1$ the Funk metric $d_{\mathbb{R}} : (-\infty, \infty]^2 \to [-\infty, \infty]$ is given by

$$d_{\mathbb{R}}(s, t) := t - s \text{ if } s \neq \infty \text{ and } d_{\mathbb{R}}(\infty, t) = -\infty \ . \tag{33}$$

In particular $d_{\mathbb{R}}(\infty, \infty) = -\infty$. Notice that $d_{\mathbb{R}}$ can also take the value $-\infty$ and this case of a directed metric, was explained in Remark 2.

Proposition 3. *$P(\mathcal{L})$ is the set of non expansive maps $x : (\mathcal{L}, d^t) \to ((-\infty, \infty], d_{\mathbb{R}})$. Namely x satisfies*

$$d_{\mathbb{R}}(x(a_j), x(a_i)) \leqslant d^t(a_j, a_i) \tag{34}$$

Moreover $\widehat{P}(\mathcal{L})$ is the set of non expansive maps $y : (\mathcal{L}, d) \to ((-\infty, \infty], d_{\mathbb{R}})$. Namely y satisfies

$$d_F(y(a_j), y(a_i)) \leqslant d(a_j, a_i) \tag{35}$$

Proof. To see this description of $P(\mathcal{L})$, let $x_i \coloneqq x(a_i)$. Then

$$d_{\mathbb{R}}(x(a_j), x(a_i)) \leqslant d^t(a_j, a_i) \iff x(a_i) - x(a_j) \leqslant d(a_i, a_j) \iff x_i - x_j \leqslant d_{i,j}.$$

Likewise to see this description of $\widehat{P}(\mathcal{L})$, let $y_i := y(a_i)$. Then

$$d_{\mathbb{R}}(y(a_j), y(a_i)) \leqslant d(a_j, a_i) \iff y(a_i) - y(a_j) \leqslant d(a_j, a_i) \iff y_i - y_j \leqslant d_{j,i}.$$

□

Remark 9. Following Proposition 3 we see that we can view $P(\mathcal{L})$ as a space of functions on the metric space $\mathcal{L}$ and we will see in Section 4 that it is similar to considering real vectors as real valued functions on a set.

Proposition 4. *The map*

$$Y : (\mathcal{L}, d) \hookrightarrow (P(\mathcal{L}), D) \text{ given by } Y(a_k) \coloneqq d(-, a_k) \tag{36}$$

is called the Yoneda embedding [4] *and is an isometric embedding. Moreover the map*

$$\widehat{Y} : (\mathcal{L}, d) \hookrightarrow (\widehat{P}(\mathcal{L}), D^t) \text{ given by } \widehat{Y}(a_k) \coloneqq d(a_k, -) \tag{37}$$

is also an isometric embedding and is called the co-Yoneda embedding.

Proof. First note that for any $a_k \in \mathcal{L}$ the function $Y(a_k) := d(-, a_k)$ is in $P(\mathcal{L})$ and the function $\widehat{Y} := d(a_k, -) : \mathcal{L} \to [0, \infty]$ is in $\widehat{P}(\mathcal{L})$.

Indeed by the triangle inequality, $d(a_i, a_k) \leqslant d(a_i, a_j) + d(a_j, a_k)$, in other words if $x := d(-, a_k)$ and $x_i := d(a_i, a_k)$ then $x_i \leqslant x_j + d_{i,j}$ proving that $Y(a_k) \in P(\mathcal{L})$.

Analogously $d(a_k, a_i) \leqslant d(a_k, aj) + d(a_j, a_i)$ and therefore if $y := d(a_k, -)$ then $y_i \leqslant y_j + d_{j,i}$ proving that $\widehat{Y}(a_k) \in \widehat{P}(\mathcal{L})$.

[4]The reason for the name Yoneda embedding comes from its appearance in category theory and was explained in [1]. It is similar to the so called, Kuratowski embedding of a metric space.

Moreover, the inequality Equation (29) entails that $d(-,a_i)+d(a_i,a_j) \geqslant d(-,a_j)$, and so, $D(d(-,a_i),d(-,a_j)) \leqslant d(a_i,a_j)$. On the other hand, if $x,y \in P(\mathcal{L})$ then $D(x,y) \geqslant y_l - x_l$ for all l and thus $D(d(-,a_i),d(-,a_j) \geqslant d(a_i,a_j) - d(a_i,a_i) = d(a_i,a_j)$. Consequently $D(d(-,a_i),d(-,a_j)) = d(a_i,a_j)$ i.e. $a \mapsto d(-,a)$ is an isometry.

Likewise we have $d(a_j,a_i) + d(a_i,-) \geqslant d(a_j,-)$ and $d(a_j,-) - d(a_i,-) \leqslant d_{j,i}$ which implies that $D(d(a_i,-),d(a_j,-)) \leqslant d_{j,i}$.

Moreover $D(d(a_i,-),d(a_j,-)) \geqslant d(a_j,a_i) - d(a_i,a_i) = d(a_j,a_i)$. Thus $D(d(a_i,-),d(a_j,-)) = d_{j,i}$.

□

To further understand the polyhedron $P(\mathcal{L})$ we consider the change of variables $z_i := e^{-x_i}$ and introduce the following:

Definition 5. *Let $Q(\mathcal{L})$ be the polyhedral cone*

$$Q(\mathcal{L}) := \{z = (z_1, \dots z_n) \in [0,\infty)^n \backslash \{(0,\dots,0)\} | z_i \geqslant \Pr(a_j|a_i) z_j\} \tag{38}$$

Moreover let $\widehat{Q}(\mathcal{L})$ be the polyhedral cone

$$\widehat{Q}(\mathcal{L}) := \{u = (u_1, \dots u_n) \in [0,\infty)^n \backslash \{(0,\dots,0)\} | u_i \geqslant \Pr(a_i|a_j) u_j\} \tag{39}$$

Note that if z is in $Q(\mathcal{L})$ then $\lambda z \in Q(\mathcal{L})$ for $\lambda \in [0,\infty)$ therefore $Q(\mathcal{L})$ is indeed a polyhedral cone in the positive orthant and so is $\widehat{Q}(\mathcal{L})$.

To simplify notation we introduce the convention that if $v = (v_1,\dots,v_n) \in \mathbb{R}^n$ then

$$e^v := (e^{v_1},\dots,e^{v_n}) \text{ and } \log(v) := (\log(v_1),\dots,\log(v_n)) \tag{40}$$

We see that

$$Q(\mathcal{L}) = \{z \in [0,\infty)^n | z := e^{-x} \text{ for } x \in P(\mathcal{L})\} \tag{41}$$

and

$$\widehat{Q}(\mathcal{L}) = \{u \in [0,\infty)^n | u := e^{-y} \text{ for } y \in \widehat{P}(\mathcal{L})\} \tag{42}$$

And vice versa

$$P(\mathcal{L}) = \{x \in (-\infty,\infty]^n | x := -\log(z) \text{ for } z =\in Q(\mathcal{L})\} \tag{43}$$

and

$$\widehat{P}(\mathcal{L}) = \{y \in (-\infty,\infty]^n | y := -\log(u) \text{ for } u \in \widehat{Q}(\mathcal{L})\}. \tag{44}$$

Using the map $-\log : Q(\mathcal{L}) \to P(\mathcal{L})$ we can define a directed metric D_Q on $Q(\mathcal{L})$ using the Funk metric D on $P(\mathcal{L})$. We put

$$D_Q(z,z') := \max_i \{\log(\frac{z_i}{z'_i}) | z'_i \neq 0\}. \tag{45}$$

By definition we have

$$D_Q(z,z') = D(-\log z, -\log z') \text{ and } D(x,x') = D_Q(e^{-x}, e^{-x'}). \tag{46}$$

Clearly the transpose D_Q^t defines a directed metric on $\widehat{Q}(\mathcal{L})$.

Then Proposition 4 implies that

Corollary 2. *The maps*

$$e^{-Y} : (\mathcal{L},d) \to (Q(\mathcal{L}), D_Q) \text{ and } e^{-\widehat{Y}} : (\mathcal{L},d) \to (\widehat{Q}(\mathcal{L}), D_Q^t) \tag{47}$$

are isometric embeddings

Proof. It follows from Proposition 2 since

$$d(a_k, a_l) = D(Y(a_k), Y(a_l)) = D_Q(e^{-Y(a_k)}, e^{-Y(a_l)}) \ . \qquad \square$$

We define the *unit simplex*

$$\Delta := \{z \in [0,1]^n | \sum_i z_i = 1\} \ .$$

Definition 6. *Define the polyhedron $Q_0(\mathcal{L})$ by*

$$Q_0(\mathcal{L}) := Q(\mathcal{L}) \cap \Delta. \tag{48}$$

Then $(Q_0(\mathcal{L}), D_Q)$ is a directed metric polyhedron and points in $Q_0(\mathcal{L})$ are probability distributions.

Analogously we define the polyhedron $\widehat{Q}_0(\mathcal{L})$ as the intersection of $\widehat{Q}(\mathcal{L})$ with the unit simplex and D_Q^t is a directed metric on it.

Remark 10. The polyhedra $Q_0(\mathcal{L})$ and $\widehat{Q}_0(\mathcal{L})$ define a normalization to probability distributions of our probabilistic language model $\mathcal{L}$. Indeed in the definition of $(\mathcal{L}, \leqslant, \Pr)$ we only ask for the conditional probabilities multiplicative property but there is no normalization to a probability distribution.

Now if we consider the vertex of $Q_0(\mathcal{L})$, corresponding to the ray generated by $e^{-Y(a_k)}$, it will be $\frac{1}{n(a_k)} e^{-Y(a_k)}$ where

$$n(a_k) := \sum_{a_j \leqslant a_k} (e^{-Y(a_k)})_j = \sum_{a_j \leqslant a_k} \Pr(a_k | a_j)$$

is the normalization factor.

While the vertex of $\widehat{Q}_0(\mathcal{L})$corresponding to the ray generated by $e^{-\widehat{Y}(a_k)}$ will be $\frac{1}{\widehat{n}(a_k)} e^{-Y(a_k)}$ where

$$\widehat{n}(a_k) := \sum_{a_k \leqslant a_l} (e^{-\widehat{Y}(a_k)})_j = \sum_{a_k \leqslant a_l} \Pr(a_l | a_k)$$

is the normalization factor.

Remark 11. The polyhedral cone $Q(\mathcal{L})$ is a generalization of the *order polytope* defined by Stanley [18]. The order polytope corresponds to the case where $\Pr(a_j|a_i)$ takes only the values 0 or 1, moreover, Stanley adds the "box constraint" $z_i \in [0,1]$ which translates to $x_i \in [0, \infty]$. Up to the box constraint, $Q(\mathcal{L})$ corresponds to the order polytope of the poset $(\mathcal{L}^{\mathrm{op}}, \leqslant)$ where $\mathcal{L}^{\mathrm{op}}$ is the opposite poset.

Stanley [18] proves that vertices of an order polytope correspond to upper sets of the poset. We will prove a generalization of that result in Theorem 2 in section 3.2.

We now explain what is the geometric meaning of the coordinates of a point $x = (x_1, x_2, \ldots, x_n) \in P(\mathcal{L})$ and a point $y = (y_1, y_2, \ldots, y_n) \in \widehat{P}(\mathcal{L})$.

Proposition 5. *If $x \in P(\mathcal{L})$ then*

$$x_i = D(d(-, a_i), x) = D(Y(a_i), x). \tag{49}$$

Moreover if $y \in \widehat{P}(\mathcal{L})$ then

$$y_i = D^t(y, d(a_i, -)) = D(\widehat{Y}(a_i), y). \tag{50}$$

Proof. We have $D(d(-,a_i),x)) \geqslant x_i - d(a_i,a_i) = x_i$.

On the other hand $D(d(-,a_i),x)) = \max_j\{x_j - d_{j,i}\} \leqslant x_i$ where the last inequality follows because $x \in P(\mathcal{L}) \iff x_j \leqslant x_i + d_{j,i}$ which implies $x_j - d_{j,i} \leqslant x_i$.

Moreover if $y \in \widehat{P}(\mathcal{L})$ then $D^t(y, d(a_i,-)) \geqslant y_i - d(a_i,a_i) = y_i$

On the other hand

$D^t(y,d(a_i,-)) = \max_j\{y_j - d_{i,j}\} \leqslant y_i$ since $y \in \widehat{P}(\mathcal{L}) \iff y_j \leqslant y_i + d_{i,j}$ which implies $y_j - d_{i,j} \leqslant y_i$. □

Remark 12. (1) We note that using the previous proposition, the defining inequalities, $x_i \leqslant x_j + d_{i,j}$ of $P(\mathcal{L})$ become

$$D(Y(a_i),x) \leqslant D(Y(a_i),Y(a_j)) + D(Y(a_j),x). \tag{51}$$

Namely they are triangle inequalities for maps $x : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$ and for the maps $Y(a_k) = d(-,a_k) : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$. Therefore we can think of $P(\mathcal{L})$ as the space of all maps $x = \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$ that satisfy the triangle inequalities for the metric D, with respect to all the maps $Y(a_k) = d(-,a_k) : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$. Thus $P(\mathcal{L})$ is a kind of convex metric span of the maps $Y(a_k) = d(-,a_k) : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$.

(2) Analogously, the defining inequalities, $y_i \leqslant y_j + d_{j,i}$ of $\widehat{P}(\mathcal{L})$ become

$$D^t(y,\widehat{Y}(a_i)) \leqslant D^t(y,\widehat{Y}(a_j)) + D^t(\widehat{Y}(a_j),\widehat{Y}(a_i)), \tag{52}$$

namely the triangle inequalities for maps $y : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$. This implies that $\widehat{P}(L)$ is the space of all maps $y = \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$ that satisfy the triangle inequalities for the metric D^t, with respect to all the maps $\widehat{Y}(a_k) = d(a_k,-) : \mathcal{L} \to ((-\infty,\infty], d_{\mathbb{R}})$. Again we see $\widehat{P}(L)$ as a kind of convex metric span of $\widehat{Y}(a_k) = d(a_k,-)$

(3) A restatement of (51) is to say that the shortest path that connects $Y(a_i)$ and x is at most as long as the shortest path that connects them but has to also go though a_j. Analogously for (52).

(4) Note that all constructions and results in this section work for a general directed metric space and not just for the special case of a probabilistic Language model

3.1. Texts define special Extremal rays of $P(\mathcal{L})$ and $Q(\mathcal{L})$.

Definition 7. *An extremal ray of a polyhedral cone in $\mathbb{R}^n$ is a ray generated by a vector that cannot be expressed as a positive linear combination of two non-proportional vectors in the polyhedral cone.*

Recall that a vector in a polyhedral cone in $\mathbb{R}^n$ generates an extremal ray if and only if it saturates $n-1$ linearly independent inequalities [17].

Definition 8. *An additive extremal ray of $P(\mathcal{L})$ (respectively $\widehat{P}(\mathcal{L})$) is defined to be the image under $-\log$ of a usual extremal ray of the polyhedral cone $Q(\mathcal{L})$ (respectively $\widehat{Q}(\mathcal{L})$).*

Note that the name additive extremal ray is chosen since a usual extremal ray in $Q(\mathcal{L})$ is invariant under scaling by λ and therefore its image under $-\log$ is invariant under translation by $-\log\lambda$. Note that the extremal rays of $Q(\mathcal{L})$ and of $\widehat{Q}(\mathcal{L})$ have generators which have in general some zero coordinates. Then, their $-\log$-images have vectors such that some of their coordinates are ∞, which is why we speak of

"additive extremal rays" of $P(\mathcal{L})$ and $\widehat{P}(L)$ (they are not extremal rays in the usual sense).

From now on though we will simply refer to the additive extremal rays of $P(\mathcal{L})$ as extremal rays*, when there is no chance of confusion.*

We now define a directed graph associated to any point $x \in P(\mathcal{L})$ which encodes the saturated inequalities satisfied by the point x and which we call its *saturation graph* $S(x)$.

Definition 9. *Let $x \in P(\mathcal{L})$. Define $S(x)$, the* saturation graph *of x, to be the graph whose vertices are the elements of $\mathcal{L}$ and whose set of directed edges $E(x)$ is the set of saturated inequalities that coordinates of x satisfy, namely $E(x) := \{(a_i, a_j) : x_i = x_j + d_{i,j}\}$. When $(a_i, a_j) \in E(x)$ we introduce a directed edge from a_i to a_j.*

The graph always contains trivial arcs (a_i, a_i) (loops), for $i \in [n]$, since $d_{i,i} = 0$. It contains non-trivial arcs if and only if x is on the boundary of P.

The graph $S(x)$ and in particular its support $\mathrm{Supp}(x)$ encodes all the hyperplanes on which x lies.

Note that the graph can be disconnected.

Theorem 1. *The isometric embedding $Y : \mathcal{L} \hookrightarrow P(\mathcal{L})$, maps points of $\mathcal{L}$ to extremal rays of the polyhedron $P(\mathcal{L})$ namely $Y(a_k) = d(-, a_k)$ is an extremal ray in $P(\mathcal{L})$. Moreover the isometric embedding $\widehat{Y} : \mathcal{L} \hookrightarrow \widehat{P}(\mathcal{L})$, maps points of $\mathcal{L}$ to extremal rays of the polyhedron $\widehat{P}(\mathcal{L})$ namely $\widehat{Y}(a_k) = d(a_k, -)$ is an extremal ray in $\widehat{P}(\mathcal{L})$.*

Proof. We have $Y(a_k) := d(-, a_k)$, and therefore $Y(a_k)_i = d(a_i, a_k), i = 1 \dots |\mathcal{L}|$. Define the support of $Y(a_k)$, $\mathrm{Supp}(Y(a_k))$, to be the set of texts a_i such that $Y(a_k)_i$ is finite. We recall that a vector in a cone in $\mathbb{R}^n$ defined by finitely many linear constraints generates an extreme ray of the cone if, and only if, the family of gradients of active constraints at this point is of rank $n - 1$.

Let $x := Y(a_k)$, and $y \in Q(\mathcal{L})$ denote the image of x by the map which applies $\exp(-\cdot)$ entrywise. Each edge (a_i, a_j) of the saturation graph $S(Y(a_k))$ yields $x_i = d(a_i, a_j) + x_j$, and so the vector y induces the active inequality $y_i = \Pr(a_j|a_i)y_j$ with gradient $e_i - \Pr(a_j|a_i)e_j$ where e_i denotes the ith vector of the canonical basis of $\mathbb{R}^n$. Moreover, each text a_i in $\mathcal{L} \setminus \mathrm{Supp}(Y(a_k))$ yields the active inequality $y_i = 0$, with gradient e_i.

The saturation graph $S(Y(a_k))$ has a connected component which is a directed tree with a_k as its root since $Y(a_k)_i = Y(a_k)_j + d_{i,j} \iff d(a_i, a_k) = d(a_i, a_j) + d(a_j, a_k)$ and from corollary 1 it follows that $a_i \leqslant a_j \leqslant a_k$, namely a_j extends a_i and a_k extends a_j. It has also trivial connected components, reduced to loops at the vertices a_i such that $a_i \notin \mathrm{Supp}(Y(a_k))$. Using the fact that the non-trivial connected component of $S(Y(A_k)$ is a tree, we see that any vector z satisfying the saturated equalities is uniquely defined by its value on the root of the tree. Hence, the space orthogonal to the family $e_i - \Pr(a_j|a_i)e_i$ with $(i, j) \in S(y_k)$ and e_l with $a_l \in \mathcal{L} \setminus \mathrm{Supp}(Y(a_k))$ is of dimension one, which entails that this family is of rank $|\mathcal{L}| - 1$, showing that y is an extreme ray of $Q(\mathcal{L})$.

Likewise let $\widehat{Y}(a_k) := d(a_k, -)$ and $\widehat{Y}(a_k)_i = d(a_k, a_i)$. If $a_k \leqslant a_i \leqslant a_j$ then $d(a_k, a_j) = d(a_k, a_i) + d(a_i, a_j)$, i.e. $\widehat{Y}(a_k)_j = \widehat{Y}(a_k)_i + d_{i,j}$.

The saturation graph for $\widehat{Y}(a_k)$ is the same as for $Y(a_k)$ with all arrow reversed. Therefore the same proof applies. □

It follows from Theorem 1, that we can identify the texts in $\mathcal{L}$ with some of the extremal rays of $P(\mathcal{L})$ and also with some of the extremal rays in $\widehat{P}(\mathcal{L})$.

However there are many other extremal rays of $P(\mathcal{L})$, which we next characterize.

3.2. **All Extremal rays correspond to connected lower sets of $\mathcal{L}$.** Consider the equations $y_i \geqslant \Pr(a_j|a_i)y_j$ which define $Q(\mathcal{L})$. We denote $P_{i,j} := \Pr(a_j|a_i)$. If we assume that $a_0 \in \mathcal{L}$ is the empty text, we have $a_0 \leqslant a_i \leqslant a_j$ for any $a_i \leqslant a_j$, and then $\Pr(a_j|a_0) = \Pr(a_i|a_0)\Pr(a_j|a_i)$.

$$\text{Define } P_i := \Pr(a_i|a_0) \text{ then } P_{i,j} = \frac{P_j}{P_i}. \tag{53}$$

Therefore $y_i \geqslant P_{i,j}y_j$ becomes $P_i y_i \geqslant P_j y_j$.

$$\text{Define } \tilde{y} = (\tilde{y}_1, \dots \tilde{y}_n) \text{ where } \tilde{y}_i := P_i y_i. \tag{54}$$

Notice that this change of coordinates maps extremal rays to extremal rays. We get then a new polyhedral cone

$$\tilde{Q}(\mathcal{L}) := \{\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n) \in [0,\infty)^n \backslash \{(0,\dots,0)\} | \tilde{y}_i \geqslant \tilde{y}_j \text{ whenever } a_i \leqslant a_j\}. \tag{55}$$

Therefore $\tilde{Q}(\mathcal{L}) := \{\tilde{y} | y \in Q(\mathcal{L})\}$.

$\tilde{Q}(\mathcal{L})$ is a polyhedral cone variant of Stanley's order polytope for the opposite poset $\mathcal{L}^{\mathrm{op}}$ – the latter is the intersection of $\tilde{Q}(\mathcal{L})$ with the box $[0,1]^n$ [18].

The change of variables mapping $Q(\mathcal{L})$ to $\widetilde{Q}(\mathcal{L})$ can be also done under our more general assumption of a Probabilistic language model without assuming the existence of a global minimum $a_0 \in \mathcal{L}$. Indeed

Proposition 6. *Let $(\mathcal{L}, \leqslant, \Pr)$ be a probabilistic language model then there is a diagonal change of variables mapping $Q(\mathcal{L})$ to $\widetilde{Q}(\mathcal{L})$.*

Proof. The fact that $(\mathcal{L}, \leqslant, \Pr)$ is a probabilistic language model means that

$$a_i \leqslant a_j \leqslant a_k \text{ is equivalent to } \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i) \text{ and } \Pr(a_k|a_i) < \infty.$$

We define the directed graph G whose nodes are the texts $a_1, \dots, a_n$. If $a_i \leqslant a_j$, we draw an arrow from node a_i to node a_j with weight $\Pr(a_j|a_i)$. If $a_j \leqslant a_i$, we draw an arrow from node a_i to node a_j with weight $\Pr(a_j|a_i)^{-1}$. Consider the Hasse diagram of $\mathcal{L}$. We make use of the observation in Proposition 2. For every connected component C_m, let us select arbitrarily an element c_m. If $a_i \in C_m$, we define w_i to be the weight in the graph G of an arbitrary path from the point c_m to a_i. Owing to our main assumption, (1), the weight is independent of the choice of the path from c_m to a_i. Moreover, for all i, j such that $a_i \leqslant a_j$, we have $w_i \Pr(a_j|a_i) w_j^{-1} = 1$. Setting $\tilde{y}_i = w_i y_i$, we rewrite the constraint $y_i \geqslant \Pr(a_j|a_i)y_j$ as $\tilde{y}_i \geqslant \tilde{y}_j$.

In this way, we transformed $Q(\mathcal{L})$ to $\tilde{Q}(\mathcal{L})$ by a diagonal scaling. □

We then have the following

Theorem 2. *The vector $\tilde{y} := (\tilde{y}_1, \dots \tilde{y}_n) \in \tilde{Q}(\mathcal{L})$ generates an extremal ray of $\tilde{Q}(\mathcal{L})$ if and only if the function $a_i \mapsto \tilde{y}(a_i) := y_i$ is a positive scalar multiple of the characteristic function of a lower set in $\mathcal{L}$ whose Hasse diagram is connected.*

Proof. Let $\{\lambda_1, \dots, \lambda_s\}$ be the distinct values taken by $\tilde{y}_i$, ordered so that $0 \leqslant \lambda_1 < \lambda_2 < \cdots < \lambda_s$. Let $\mathcal{L}_m = \{i | \tilde{y}_i = \lambda_m\}$.

We define the rank of a family of affine inequalities to be the rank of the family of gradients of the affine forms defining these inequalities. For $\tilde{y}$ to be an extremal ray it has to saturate a family of inequalities of rank $n-1$, where $|\mathcal{L}| = n$, see [17].

Let us first assume $\lambda_1 = 0$. The rank of the family of saturated constraints given by $\mathcal{L}_1$ is then $r_1 = |\mathcal{L}_1|$ since we get equations of the form $\tilde{y}_i = 0$ which is a hyperplane normal to e_i for $i \in \mathcal{L}_1$.

Moreover, we claim that the rank r_k of the family of saturated constraints given by any $\mathcal{L}_k$ for $k \neq 1$ given by $r_k = |\mathcal{L}_k| - c_k$ where c_k is the number of connected components of the Hasse diagram of $(\mathcal{L}_k, \leqslant)$. To see this, it suffices to observe that a solution $h \in \mathbb{R}^{\mathcal{L}_k}$ of the system of saturated inequalities $h_i = h_j$ for $a_i \leqslant a_j$ and $a_i, a_j \in \mathcal{L}_k$ is uniquely determined by fixing precisely one coordinate of h on every connected component of the Hasse diagram (in other words, we have c_k degrees of freedom for the choice of h).

Therefore the rank of the family of saturated constraints at the point $\tilde{y}$ is less than or equal to $|\mathcal{L}_1|+|\mathcal{L}_2|-1+\cdots+|\mathcal{L}_s|-1$. We also have that $|\mathcal{L}_1|+|\mathcal{L}_2|+\cdots+|\mathcal{L}_s| = n$. We know though that $\tilde{y}$ is an extremal ray if and only if the rank of the family of saturated constraints is $n-1$. Therefore we must have $n-1 \leqslant |\mathcal{L}_1| + (|\mathcal{L}_2|-1) + \cdots + (|\mathcal{L}_s|-1) = n - s + 1$.

This is only possible if $s \leqslant 2$ but, for $\tilde{y}$ to generate an extremal ray, not all coordinates of $\tilde{y}$ can be 0, and then our assumption $\lambda_1 = 0$ excludes the case $s = 1$. This entails that $s \geqslant 2$ and therefore $s = 2$. We then have $n - 1 = |\mathcal{L}_1| + |\mathcal{L}_2| - 1$. In that case $\tilde{y}_i = 0$ for $i \in \mathcal{L}_1$ and $\tilde{y}_j = \lambda_2$ for $j \in \mathcal{L}_2$. We then scale $\tilde{y}$ by $\frac{1}{\lambda_2}$ so as to get a representative vector of the same the ray with $\tilde{y}_i = 0$ for $i \in \mathcal{L}_1$ and $\tilde{y}_j = 1$ for $j \in \mathcal{L}_2$. Therefore $\tilde{y}$ is the characteristic function of $\mathcal{L}_2$.

Moreover $\mathcal{L}_2$ is a lower set. Indeed if $a_j \in \mathcal{L}_2$ and $a_i \leqslant a_j$ then $a_i \in \mathcal{L}_2$. This holds because $a_j \in \mathcal{L}_2$ implies $\tilde{y}(a_j) = 1$ and $a_i \leqslant a_j$ implies $\tilde{y}(a_i) = \tilde{y}(a_j) = 1$ therefore $a_i \in \mathcal{L}_2$.

If now $\lambda_1 > 0$ then $n - 1 \leqslant (|\mathcal{L}_1| - 1) + (|\mathcal{L}_2| - 1) + \cdots + (|\mathcal{L}_s| - 1) = n - s$ therefore $s \leqslant 1$ which implies $s = 1$ and $|\mathcal{L}_1| = n$. In that case we have a single extremal ray $\tilde{y} = (1, 1, \ldots, 1)$ which is the characteristic function of the maximal lower set $\mathcal{L}_1 = \mathcal{L}$.

Conversely let C be a lower set in $\mathcal{L}$ and let $\tilde{y} : \mathcal{L} \to \{0, 1\}$ be the characteristic function of U. Consider $a_j \in U$ then $\tilde{y}(a_j) = 1$. Now if $a_i \leqslant a_j$ then $a_i \in U$ and therefore $\tilde{y}(a_i) = 1$ which means $\tilde{y}(a_j) = \tilde{y}(a_i)$.

□

Remark 13. Note that if $\mathcal{L}$ admits a bottom element a_0 then any lower set is connected since it must include a_0, and the Hasse diagram of $\mathcal{L}$ contains a path from a_0 to every element of $\mathcal{L}$.

Notice that Stanley in [18] has proven that vertices of his order polytope correspond to upper sets. In contrast, rays of $Q(\mathcal{L})$ correspond only to *connected* lower sets. Notwithstanding the order reversal, there is a discrepancy which arises because Stanley considers the intersection of $Q(\mathcal{L})$ with a box, with creates additional vertices, not associated to rays of $Q(\mathcal{L})$.

Remark 14. Note that a vector $Y(a_k) \in P(\mathcal{L})$ corresponds to the principal lower set generated by a_k. We will therefore call the extremal rays generated by images of the Yoneda embedding, *principal extremal rays*.

Corollary 3. *Assume $\mathcal{L}$ includes the bottom element a_0 and recall from* (53) *that $P_i := \Pr(a_i|a_0)$. If C is a lower set, the extremal ray corresponding to C is generated by y in $Q(\mathcal{L})$ with coordinates*

$$y_i = \begin{cases} \frac{1}{P_i} & \text{if } a_i \in C, \\ 0 & \text{if } a_i \text{ not in } C. \end{cases} \tag{56}$$

Proof. It follows from Theorem 2 and the change of coordinates $\tilde{y}_i = P_i y_i$ in eq 53. □

Remark 15. Notice that Corollary 3 is consistent with the coordinates of an extremal ray y in $Q(\mathcal{L})$, corresponding to a text $a_j \in \mathcal{L}$. Indeed according to corollary 2, any element y on the extremal ray has $y_i = \frac{\lambda}{P_i}$ for $\lambda \in [0, \infty)$. We also have $y_j = \Pr(a_j|a_j) = 1$, therefore $\lambda = P_j$. This implies $y_i = \frac{P_j}{P_i} = \Pr(a_j|a_i)$

$$y_i = \begin{cases} \Pr(a_j|a_i) & \text{if } a_i \leqslant a_j, \\ 0 & \text{if } a_i \text{ not a subtext of } a_j. \end{cases} \tag{57}$$

Now we want a general version of the Corollary 3.

For any subset C of $\mathcal{L}$, selecting an element $c \in C$, for every element a_i in the connected component of C in the graph induced by the Hasse diagram of $\mathcal{L}$, we denote by w_i^c the weight of any path from c to a_i in the directed graph constructed in the proof of Proposition 6.

Proposition 7. *Let C be a connected lower set of the Hasse diagram of $(\mathcal{L}, \leqslant)$. Let c denote any element of C. Then, the vector*

$$y_i = \frac{1}{w_i^c}, \qquad \text{for } a_i \in C, \qquad y_i = 0 \text{ for } a_i \in \mathcal{L} \setminus \{C\} \tag{58}$$

generates an extreme ray of $Q(\mathcal{L})$, and all the extreme rays arise in this way.

Proof. We showed in Theorem 2 that $\tilde{y}$ is a positive scalar multiple of the characteristic function of C. If a_i belongs to C, we have $y_i = (w_i^c)^{-1}\tilde{y}_i$, from which (58) follows. We note that a change of the reference point c in C only modifies the vector w^c by a positive scalar multiple. Indeed, for all c and $c' \in C$, we have $w^c = \mu w^{c'}$ where μ is the weight of any path from c to c' in the directed graph G. □

Proposition 8. *If y generates an extremal ray of $Q(\mathcal{L})$ corresponding to a lower set C in $\mathcal{L}$ then the saturation graph of y has an edge from a_i to a_j if and only if $a_i \leqslant a_j$ for $a_i, a_j \in C$.*

Proof. This follows from (58), using the main assumption (1). Indeed if $a_i \leqslant a_j$ in C then we have $y_i = \frac{1}{w_i^c}$ and $y_j = \frac{1}{w_j^c}$. Therefore $y_i = \frac{w_j^c}{w_i} y_j$ and therefore $y_i = \Pr(a_j|a_i) y_j$ which means that there is an edge from a_i to a_j in the saturation graph of y. □

Corollary 4. *Extremal rays of $\widehat{Q}(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ correspond to connected upper sets of $\mathcal{L}$.*

Proof. We have $\widehat{Q}(\mathcal{L}) = Q(\mathcal{L}^{\mathrm{op}})$ and upper sets of $\mathcal{L}$ correspond to lower sets of $\mathcal{L}^{\mathrm{op}}$ therefore the result follows from Proposition 7. □

Remark 16. Note that Proposition 6, and Theorem 2 are only valid for a probabilistic language model and not for general directed metric space. In the latter case there will still be exponentially many extremal rays not coming from the Yoneda embedding, but we the characterization in terms of connected lower sets no longer holds.

Corollary 5. *If the empty text a_0 is in $\mathcal{L}$ then the set $P^0(\mathcal{L})$ of extremal rays of $P(\mathcal{L})$ is identified with the lower set completion of the poset $\mathcal{L}$.*

Proof. Since $a_0 \in \mathcal{L}$, every lower set of $\mathcal{L}$ is connected so $P^0(\mathcal{L})$ is identified with the set of lower sets of $\mathcal{L}$.

□

Remark 17. Note that having explicit equations for the polyhedral cone $Q(\mathcal{L})$, the extremal rays of $Q(\mathcal{L})$ can be computed, for instance by the double description method [6].

4. The polyhedron $P(\mathcal{L})$ as a $(\min, +)$ linear space

To further understand the polyhedra $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ we need to consider their description in terms of tropical algebra.

Consider the metric space $(\mathcal{L}, d)$. Recall the $(\min, +)$ (tropical) semifield $\mathbb{R}_{\min}$ defined as $\mathbb{R}_{\min} := ((-\infty, \infty], \oplus_{\min}, \odot)$ where for $s, t \in (-\infty, \infty]$,

$$s \oplus_{\min} t := \min\{s, t\} \text{ and } s \odot t := s + t. \tag{59}$$

We denote by $d_{\min} : \mathbb{R}^n \to \mathbb{R}^n$ the $(\min, +)$ linear operator defined by

$$d_{\min}(x)_i := \min_j \{d_{i,j} + x_j\} \tag{60}$$

Proposition 9. *$(\mathcal{L}, d)$ is a directed metric if and only if $d_{\min}^2 = d_{\min}$, namely $d_{\min}$ is a $(\min, +)$ projector.*

Proof. We have $d_{\min}^2 = d_{\min} \iff d_{i,k} = \min_j\{d_{i,j} + d_{j,k}\}$ which is the same as the triangle inequality $d_{i,k} \leqslant d_{i,j} + d_{j,k}$. □

Let us denote $\mathrm{Im}(d_{\min})$, the image of $d_{\min}$, namely the $(\min, +)$ column span of d.

Lemma 1. *We have $\mathrm{Im}(d_{\min}) = \mathrm{Fix}(d_{\min})$, where $\mathrm{Fix}(d_{\min})$ is the $(\min, +)$ module $\mathrm{Fix}(d_{\min}) := \{x : d_{\min}(x) = x\}$.*

Proof. It follows from $d_{\min}^2 = d_{\min}$. □

We note now that there is a very natural description of our polyhedra as follows:

Proposition 10. *The polyhedron $P(\mathcal{L})$ is equal to $\mathrm{Im}(d_{min}) = \mathrm{Fix}(d_{\min})$ and the polyhedron $\widehat{P}(\mathcal{L})$ is equal to $\mathrm{Im}(d_{min}^t) = \mathrm{Fix}(d_{\min}^t)$.*

Proof. Since $d_{\min}^2 = d_{\min}$ we have that $x \in \mathrm{Im}(d_{min}) \iff d_{\min} x = x$ which means that $x_i = \min_j\{d_{i,j} + x_j\}$ and thus $x_i \leqslant x_j + d_{i,j}$. Likewise for $\mathrm{Im}(d_{min}^t)$ we get $x_i \leqslant x_j + d_{j,i}$. □

Since we use much more often the $(\min, +)$ semifield than the $(\max, +)$ that will appear later on, *to simplify notation we denote $\oplus_{\min}$ by $\oplus$.*

In particular we introduce the notation, for $u, v \in \mathbb{R}^n$, $(u \oplus v)_i := \min\{u_i, v_i\}$. We then have

Corollary 6. *If $x \in P(\mathcal{L})$ then*

$$x = \oplus_j D(Y(a_j), x) \odot Y(a_j). \tag{61}$$

Proof. From Proposition 10, $x \in P(\mathcal{L}) = \mathrm{Im}(d_{\min}) = \mathrm{Fix}(d_{\min}) \iff d_{\min}(x) = x$. Therefore we have the $(\min, +)$ linear expression for x in terms of the columns of d:

$x = \oplus_j x_j \odot d(-, a_j) = \oplus_j x_j \odot Y(a_j) = \oplus_j D(Y(a_j), x) \odot Y(a_j)$. □

It is known that an order polytope, and more generally, an alcoved polytope (of A_n type) is closed under min and max, (61) expresses this fact for our metric case for min. In Proposition 29 we will see that $P(\mathcal{L})$ is also closed under max.

Proposition 11. *We have*

$$Y(a_k) = \oplus_{a_j \leqslant a_k} d_{j,k} \odot Y(a_j) \tag{62}$$

and

$$\widehat{Y}(a_k) = \oplus_{a_k \leqslant a_l} d_{k,l} \odot \widehat{Y}(a_l) \tag{63}$$

Proof. The fact that $d^2_{\min} = d_{\min}$ is equivalent to

$$d_{i,k} = \min_j \{d_{i,j} + d_{j,k}\}. \tag{64}$$

We have $Y(a_k) := d(-, a_k)$ and $Y(a_j) := d(-, a_j)$ therefore eq. 64 implies

$$Y(a_k)_i = \oplus_j d_{j,k} \odot Y(a_j)_i$$

which means

$$Y(a_k) = \oplus_j d_{j,k} \odot Y(a_j)$$

. Since $d_{j,k} = \infty$ unless $a_j \leqslant a_k$ we have

$$Y(a_k) = \oplus_{a_j \leqslant a_k} d_{j,k} \odot Y(a_j)$$

.

Analogously for d^t we have $d^t_{i,k} = \min_l \{d^t_{i,l} + d^t_{l,k}\} \iff d_{k,i} = \min_l \{d_{k,l} + d_{l.i}\}$. Recall that $\widehat{Y}(a_k) := d(a_k, -)$ and $\widehat{Y}(a_l) := d(a_l, -)$. This implies

$$\widehat{Y}(a_k) = \oplus_l d_{k,l} \odot \widehat{Y}(a_l).$$

Since $d_{k,l} = \infty$ unless $a_k \leqslant a_l$ we have

$$\widehat{Y}(a_k) = \oplus_{a_k \leqslant a_l} d_{k,l} \odot \widehat{Y}(a_l) \ . \qquad \square$$

Finally we have the following

Proposition 12. *The Funk metric $D(x, y) := \max_i \{y_i - x_i | x_i \neq \infty\}$ has the property that $D(-, w)$ is tropically antilinear, namely*

$$D(\lambda_1 \odot x \oplus_{\min} \lambda_2 \odot y, z) = -\lambda_1 \odot D(x, z) \oplus_{\max} -\lambda_2 \odot D(y, z) \tag{65}$$

while $D(w, -)$ is linear, namely

$$D(x, \lambda_1 \odot y \oplus_{\max} \lambda_2 \odot z) = \lambda_1 \odot D(x, z) \oplus_{\max} \lambda_2 \odot D(y, z). \tag{66}$$

Proof. We have $D(\lambda \odot x, y) = \max_i\{y_i - \lambda - x_i\} = D(x,y) - \lambda$.

We calculate $D(x \oplus_{\min} y, z) = \max_i\{z_i - \min\{x_i, y_i\}\} = \max_i\{z_i + \max_i\{-x_i, -y_i\}\} =$
$= \max\{\max_i\{z_i - x_i\}, \max\{z_i - y_i\}\} = D(x,z) \oplus_{\max} D(y,z)$.

Moreover $D(x, \lambda \odot y) = \max_i\{\lambda + y_i - x_i\} = \lambda - D(x,y)$.

Finally, $D(x, y \oplus_{\max} z) = D(x, \max\{y,z\}) = \max_i\{\max\{y_i, z_i\} - x_i\} =$

$= \max_i\{\max\{y_i - x_i\}, \max\{z_i - x_i\}\} = \max\{\max_i\{y_i - x_i\}, \max_i\{z_i - x_i\}\} =$
$= D(x,y) \oplus_{\max} D(x,z)$ □

This means that we can think of D as a tropical inner product.

Remark 18. All the results in this section hold for a general directed metric space.

4.1. $P(\mathcal{L})$ **and** $\widehat{P}(\mathcal{L})$ **as Semantic spaces.** We already mentioned in the overview that we consider $Y(a_k) := d(-, a_k)$ as well as $\widehat{Y}(a_k) := d(a_k, -)$ as encoding the meanings of text a_k in accordance with the statistical semantics principal namely that texts that appear in similar contexts have similar meaning. The function $d(a_k, -)$ is supported on extensions of a_k while $d(-, a_k)$ is supported on restrictions of a_k.

However it is also the position of these vectors in $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ that contains semantic information since for example, $Y(a_k) = d(-, a_k)$ for a_k a word is supported only on that word while $D(Y(a_k), -)$ is supported on all extensions of a_k.

Therefore more generally, we think of $(P(\mathcal{L}), D)$ and $(\widehat{P}(\mathcal{L}), D^t)$ as "semantic spaces" giving mathematical substance to the statistical semantics hypothesis. This point if view was advocated in [1].

We further explain our view about the syntax to semantics problem in Appendix B and show that it is located in the realm of a deep and general duality in mathematics which in some cases appears as a duality between algebra and geometry.

It is interesting that even though the whole space $P(\mathcal{L})$ (or $Q(\mathcal{L})$) and $\widehat{P}(\mathcal{L})$ (or $\widehat{Q}(\mathcal{L})$) appear as a spaces of meanings, texts appear only as special extremal rays. They are the "observable" variables while other points of $P(\mathcal{L})$ are like "hidden" variables.

The systems of equations Proposition 11 Equation (62), Equation (63) express the $(\min, +)$ linear relations satisfied by the Yoneda and co-Yoneda embedding text vectors. They are reminiscent of vector equations between word vectors as appeared first in [13].

Another way to think about them is as equations that implement the constraints imposed by the probabilities of extension. It is common to consider constraints giving rise to equations defining a geometric object and here we have something analogous but in $(\min, +)$ algebra.

Moreover we note that any $(\min, +)$ linear combination can be transformed into a Boltzmann weighted usual linear combination using a small temperature parameter and the identity Equation (91)

$$\lim_{T \to 0} -T \log(e^{-y/T} + e^{-z/T}) = \min\{y, z\}.$$

Using this we will also show in Corollary 9, Equation (90) that $e^{-Y(a_k)}$ can be approximated by a Boltzmann weighted linear combinations of word vectors for the words that make up that text. We note the similarity of this with the expression of a value vector for a text in terms of word vectors, in the attention layer of a transformer.

Notice also that from the formulation of probabilistic language models, vectors arise naturally, first in the $(\min, +)$ context but later in Boltzmann weighted usual linear combinations (section 5.1).

Moreover we will show in Section 6 that there is a duality relating $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ as well as $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$ (they are isometric and tropically anti-isomorphic). This shows that given a corpus, the semantic information given by extensions of texts is equivalent to that given by restrictions.

We note that if the transformer is computing an approximation to $\widehat{P}(\mathcal{L})$ then the fact that it is a convex space could explain why the gradient descent during training converges nicely.

Since the transformer computes probabilities for all possible next words to a text it is natural to think that the corresponding probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$ contains the whole free monoid generated by words and all texts appear as extremal rays of $P(\mathcal{L})$ corresponding to principal upper sets. Wrong texts are very far away from correct texts as they are very unlikely.

In that case the neural network should then learn an effective representation of $\widehat{P}(\mathcal{L})$, which a priori has a huge dimension. How the neural network is able to construct an effective approximation of such a huge dimensional space is not clear to us.

From another point of view we see that if we consider that the transformer neural network is learning $\widehat{P}(\mathcal{L})$ then we can think of training the transformer as finding a solution to the huge $(\min, +)$ system of Equation (63), Proposition 11, given the coefficients $d_{j,k}$.

We will see a small example of the polyhedra $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$ as well as the dualities, in section 6.

Further evidence for these spaces as semantic spaces is provided by the fact that they have a Heyting algebra structure (which is a generalization of a Boolean algebra) as explained in [1].

As already mentioned, experiments using (a slight variant of) the co-Yoneda vectors $d(-, a_k)$ were performed in [12] where an actual transformer neural network was used to sample continuations of texts and construct the co Yoneda vectors. The authors tested these vectors on several semantic tasks and obtained very good results.

4.2. **From one word text extensions to longer extensions.** We now explain how to go from one word extension probabilities to the metric d.

However d constructed in this way does not satisfy the main assumption of probabilistic language model $\Pr(a_k|a_i) = \Pr(a_j|a_i)\Pr(a_k|a_j)$. It does satisfy that d is a directed metric and therefore d is a $(\min, +)$ projector.

Indeed, let C be the matrix of one word extensions. Namely for texts a_i and a_j we put

(67)
$$C(a_i, a_j) = \begin{cases} -\log \Pr(a_j|a_i) & \text{if } a_i \leqslant a_j \text{ and } a_j \text{ extends } a_i \text{ by a single word,} \\ \infty & \text{if } a_i \leqslant a_j \text{ and } a_j \text{ extends } a_i \text{ by more than a single word,} \\ \infty & \text{if } a_i \text{ and } a_j \text{ are not comparable.} \end{cases}$$

Let Id denote the matrix with $\mathrm{Id}_{i,i} = 0$ and $\mathrm{Id}_{i,j} = \infty$ for $i \neq j$. Id is the identity matrix in the $(\min, +)$ matrix semiring. Indeed $C\,\mathrm{Id} = \mathrm{Id}\,C = C$. We note that $C_{i,i} = 0$ and therefore $C \oplus \mathrm{Id} = C$.

In that case the tropical power C^l computes distances for up to l word extensions.

If we bound the number of words in the extension to say k then $C^k = C^{k+1}$ and $d = C^k$ is our metric.

Proposition 13. *Let C be such that $C_{i,i} = 0$. If $d = C^k = C^{k+1}$ then*

$$x = dx \iff x = Cx \tag{68}$$

Proof. If $x = dx = C^k x$ then $Cx = C^{k+1}x = C^k x = dx = x$ therefore solutions of $x = dx$ are also solutions of $x = Cx$.

On the other hand if $x = Cx$ then $C^k x = x$ i.e. $dx = x$. □

Since the diagonal entries of d, and C, are equal to 0, the equations $x = dx$ and $x = Cx$ are equivalent to $x \geqslant dx$ and $x \geqslant Cx$, respectively. These two systems of inequalities describe the same polyhedron.

5. Compatibility of $P(\mathcal{L})$ with adding more texts

When training the neural network to learn by predicting continuations of texts we add more and more text. Moreover we have already mentioned in Remark 7 that it is natural to grade $\mathcal{L}$ by word length of texts. It is therefore important to understand how $P(\mathcal{L})$ changes as we add more and more text.

We have the following:

Proposition 14. *If a probabilistic language model $(\mathcal{L}_1, d_1)$ is extended to $(\mathcal{L}_2, d_2)$, namely if there is an isometric embedding $\phi : (\mathcal{L}_1, d_1) \hookrightarrow (\mathcal{L}_2, d_2)$ then there is an isometric embedding $\widetilde{\phi} : (P(\mathcal{L}_1)), D_1) \hookrightarrow (P(\mathcal{L}_2), D_2)$ such that $\widetilde{\phi}(Y_1(a)) = Y_2(\phi(a))$. Moreover $\widetilde{\phi}(P(\mathcal{L}_1))$ is a retraction (i.e. a non-expansive $(\min, +)$ projection) of $P(\mathcal{L}_2)$*

We will prove this in full generality and derive the probabilistic language model case as a special case.

Theorem 3. *Let $\phi : (X_1, \delta_1) \hookrightarrow (X_2, \delta_2)$ be an isometric embedding of discrete, finite, directed metric spaces, then there is an isometric embedding $\widetilde{\phi} : (P(X_1), \Delta_1) \hookrightarrow (P(X_2), \Delta_2)$ compatible with the Yoneda isometric embeddings $Y_1 : X_1 \to P(X_1)$ and $Y_2 : X_2 \to P(X_2)$, namely $\widetilde{\phi}(Y_1(a)) = Y_2(\phi(a))$. Moreover $\widetilde{\phi}(P(X_1))$ is a retraction (i.e. a non-expansive $(\min, +)$ projection) of $P(X_2)$.*

Proof. Say $X_1 := \{a_1 \dots a_n\}$ and $X_2 := \{b_1 \dots b_n, b_{n+1}, \dots b_{n+k}\}$, where $b_j = \phi(a_j)$ for $j = 1 \dots n$. Recall that we have $P(X_1) = \mathrm{Im}(\delta_1)$ is the span of $Y_1(a_j) := \delta_1(-, a_j)$ and $P(X_2) = \mathrm{Im}(\delta_2)$ is the span of $Y_2(b_j) := \delta_2(-, b_j)$.

Let $e_m := (\infty, \dots, 0, \dots, \infty)$ for $m = 1, \dots n$, so that $e_1, \dots, e_n$ is a basis (free and generating family) of the module $(\mathbb{R}_{\min})^n$ of the (min,+) semifield $\mathbb{R}_{\min}$. We define

$$\widetilde{\phi}(\oplus_{m=1}^n x_m \odot e_m) := \oplus_{m=1}^n x_m \odot \delta_2(-, b_m)$$

We now show that

$$\widetilde{\phi}(\delta_1(-, a_i)) = \delta_2(-, b_i).$$

for $i = 1, \dots, n$. Indeed,

$$\widetilde{\phi}(\delta_1(-, a_i)) = \tilde{\phi}(\oplus_{j=1}^n \delta_1(a_j, a_i) \odot e_j) = \oplus_{j=1}^n \delta_1(a_j, a_i)\delta_2(-, b_j) = \delta_2(-, b_i).$$

Indeed, the last equality holds since

$$\oplus_{j=1}^n \delta_1(a_j,a_i)\odot\delta_2(b_l,b_j)=\oplus_{j=1}^n\delta_2(b_j,b_i)\odot\delta_2(b_l,b_j)=\delta_2(b_l,b_i),$$

in which the last equality follows from the fact that δ_2 is a (min,+) idempotent.

Note that $\widetilde{\phi}$ is well defined since any $x\in(\mathbb{R}_{\min})^n$ has a unique expression in the basis $e_k, k=1,\dots,n$. If we attempted to define it directly on the $(\min,+)$ module spanned by the vectors $\delta_1(-,a_i)$ we would have to deal with the complication that $x\in\mathbb{R}^n$ does not always have a unique expression as a $(\min,+)$ combination of these vectors. In fact, one can show that only vectors in the interior of $P(\mathcal{L}_1)$ would have such unique expressions.

We now check that $\tilde{\phi}$ is an isometric embedding. We want to check that

$$\Delta_2(\widetilde{\phi}(x),\widetilde{\phi}(y))=\Delta_1(x,y) \tag{69}$$

Recall that $\Delta_1(x,y)=\max_{j=1}^n\{y_j-x_j|x_j\neq\infty\}$. Moreover $\Delta_2(\widetilde{\phi}(x),\widetilde{\phi}(y))=\max_{j=1}^{n+k}\{\widetilde{y}_j-\widetilde{x}_j|x_j\neq\infty\}$.

From the definition of the $\widetilde{x}_j$ the result follows.

Finally we define the retraction $\mathcal{R}:P(X_2)\to P(X_2)$ by

$$\mathcal{R}:=\bigoplus_{j=1}^n\Delta_2(-,Y(b_j))\odot\Delta_2(Y(b_j),-) \tag{70}$$

Note that as a matrix

$$\mathcal{R}_{i,k}=R(Y(b_i),Y(b_k)=\bigoplus_{j=1}^n\delta_2(b_i,b_j)\odot\delta_2(b_j,b_k)\ . \tag{71}$$

We need to check that $\mathcal{R}^2=\mathcal{R}$, $\mathrm{Im}(\mathcal{R})=\widetilde{\phi}(P(X_1))$ and $\mathcal{R}$ is non-expansive. Let us check first that $\mathcal{R}^2=\mathcal{R}$:

$$\mathcal{R}^2(Y_2(b_k),Y_2(b_l))=\oplus_{m=1}^n\mathcal{R}(Y_2(b_k),Y_2(b_m))\odot\mathcal{R}(Y_2(b_m),Y_2(b_l))=$$
$$\oplus_{m,j_1,j_2=1}^n\delta_2(b_k.b_{j_1})+\delta_2(b_{j_1}.b_m)+\delta_2(b_m.b_{j_2})+\delta_2(b_m,b_{j_2})+\delta_2(b_{j_2},b_l)=$$
$$\delta_2(b_k,b_l)=\mathcal{R}(b_k,b_l).$$

Where we have used the fact that

$$\oplus_{l=1}^n\delta_2(b_k,b_l)+\delta_2(b_l,b_m)=\oplus_{l=1}^n\delta_1(a_k,a_l)+\delta_1(a_l,a_m)=\delta_1(a_k,a_m)=\delta_2(b_k,b_m).$$

Next notice that clearly $\mathrm{Im}(\mathcal{R})\subset\mathrm{Span}_{j=1}^n\{\delta_2(-,b_j)\}=\widetilde{\phi}(P(\mathcal{L}_1)$. Moreover we claim that

$$\mathcal{R}(\delta_2(-,b_k))=\delta_2(-,b_k)$$

Indeed

$$\mathcal{R}(\delta_1(-,b_k))=\oplus_{j=1}^n\delta_2(b_j,b_k)\odot\delta_2(-,b_j).$$

and thus

$$\mathcal{R}(\delta_1(b_l,b_k))=\oplus_{j=1}^n\delta_2(b_j,b_k)+\delta_2(b_l,b_j)=\delta_2(b_l,b_k),$$

proving the claim.

Therefore $\mathrm{Span}_{j=1}^n\{\delta_2(-,b_j)\}\subset\widetilde{\phi}(P(\mathcal{L}_1))\subset\mathrm{Im}(\mathcal{R})$ showing that $\mathrm{Im}(\mathcal{R})=\widetilde{\phi}(P(\mathcal{L}_1)$.

Finally we check that $\mathcal{R}$ is non-expansive, namely that

$$\Delta_2(\mathcal{R}(x),\mathcal{R}(y))\leqslant\Delta_2(x,y)$$

To that end note that $\mathcal{R}$ is order preserving and also $\mathcal{R}(\alpha \odot x) = \alpha \odot \mathcal{R}(x)$. Indeed both of these statements follow from the $(\min, +)$ linearity of $\mathcal{R}$.

In particular $x \leqslant y \iff x \oplus y = x$ which implies that $\mathcal{R}(x \oplus y) = \mathcal{R}(x)$ and therefore $\mathcal{R}(x) \oplus \mathcal{R}(y) = \mathcal{R}(x)$ which means $\mathcal{R}(x) \leqslant \mathcal{R}(y)$.

Recall now that $\Delta_2(x, y) = \inf\{\lambda : x \leqslant \lambda \odot y\} = \max_i\{y_i - x_i | x_i \neq \infty\}$. Then

$$x \leqslant \Delta_2(x, y) \odot y \implies \mathcal{R}(x) \leqslant \mathcal{R}(\Delta_2(x, y) \odot y) \implies \mathcal{R}(x) \leqslant \Delta_2(x, y) \odot \mathcal{R}(y)$$

therefore $\Delta_2(\mathcal{R}(x), \mathcal{R}(y) \leqslant \Delta_2(x, y)$

□

Remark 19. Since $\tilde{\phi}(Y_1(a_j)) := Y_2(\phi(a_j)$ we have, if $x := \oplus_{j=1}^n x_j \odot Y_1(a_j)$,

$$\widetilde{\phi}(x) := \oplus_{j=1}^n x_j \odot Y_2(\phi(a_j)) = \oplus_{j=1}^n x_j \odot Y_2(b_j) = \oplus_{j=1}^{n+k} \tilde{x}_j \odot Y_2(b_j),$$

where $\widetilde{x}_j := x_j$ for $j = 1, \ldots, n$ and $\widetilde{x}_j = \infty$ for $j = n+1, \ldots, n+k$. So in these coordinates $P(X_1)$ is cut out inside $P(X_2)$ by the equations $\tilde{x}_j = \infty$ for $j = n+1, \ldots, n+k$. In this sense, it constitutes a "face" of $P(X_2)$ of (projective) dimension $|X_1| - 1$.

Remark 20. Note that if $x \in P(\mathcal{L}_2)$ where $x : \mathcal{L}_2 \to (-\infty, \infty]$ and $x_i := x(b_i)$ then

$$\mathcal{R}(x) = \bigoplus_{j=1}^n \Delta_2(Y_2(b_j), x) \odot \Delta_2(-, Y(b_j)) : \mathcal{L}_2 \to (-\infty, \infty] \tag{72}$$

and for $i = 1 \ldots, n+l$

(73)

$$\mathcal{R}(x)_i := \mathcal{R}(x)(b_i) = \bigoplus_{j=1}^n \Delta_2(Y_2(b_i), Y_2(b_j)) \odot \Delta_2(Y_2(b_j), x) = \bigoplus_{j=1}^n d_2(b_i, b_j) \odot x_j$$

Therefore

$$\mathcal{R}(x) = \bigoplus_{i=1}^{n+l} \mathcal{R}(x)_i \odot Y_2(b_i) = \bigoplus_{i=1}^{n+l} \bigoplus_{j=1}^n d_2(b_i, b_j) \odot x_j \odot Y_2(b_i)$$

$$= \bigoplus_{i=1}^{n+l} \bigoplus_{j=1}^n d_2(b_i, b_j) \odot \Delta_2(Y(b_j), x) \odot Y_2(b_i) \tag{74}$$

5.1. **Approximation of a text vector in terms of word vectors.** Let us see how Theorem 3 applies to the probabilistic language model case.

Corollary 7. *Let $\mathcal{L}_1 := \{a_1 \ldots a_n\}$ and $\mathcal{L}_2 := \{b_1 \ldots b_n, b_{n+1}, \ldots b_{n+l}\}$, be probabilistic language models and $\phi : \mathcal{L}_1 \to \mathcal{L}_2$ an isometric embedding where $b_j = \phi(a_j)$ for $j = 1 \ldots n$. Let $Y_1 : (\mathcal{L}_1, d_1) \to (P(\mathcal{L}_1), D_1)$ and $Y_2 : (\mathcal{L}_2, d_2) \to (P(\mathcal{L}_2), D_2)$ be the Yoneda isometric embeddings.*

Let $\mathcal{R} : P(\mathcal{L}_2) \to P(\mathcal{L}_2)$ be the non-expansive projection of Theorem 3 given by

$$\mathcal{R} := \bigoplus_{j=1}^n D_2(-, Y_2(b_j)) \odot D_2(Y_2(b_j), -). \tag{75}$$

Then for $i, k = 1, \ldots, n+l$

$$\mathcal{R}(Y_2(b_k))_i = \bigoplus_{j=1}^n d_2(b_i, b_j) \odot d_2(b_j, b_k) \tag{76}$$

and

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{i=1}^{n+l} \mathcal{R}(Y_2(b_k))_i \odot Y_2(b_i) = \bigoplus_{i=1}^{n+l} \bigoplus_{j=1}^{n} d_2(b_i, b_j) \odot d_2(b_j, b_k) \odot Y_2(b_i). \tag{77}$$

or equivalently

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{b_i \leqslant b_j \leqslant b_k} d_2(b_i, b_j) \odot d_2(b_j, b_k) \odot Y_2(b_i). \tag{78}$$

Proof. We have

$$\mathcal{R} = \bigoplus_{j=1}^{n} D_2(-, Y_2(b_j)) \odot D_2(Y_2(b_j), -),$$

Applying to $Y_2(b_k)$ for $k = 1, \dots n+l$ we get

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{j=1}^{n} D_2(-, Y_2(b_j)) \odot D_2(Y_2(b_j), Y_2(b_k)).$$

Since Y_2 is an isometric embedding we have

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{j=1}^{n} D_2(-, Y_2(b_j)) \odot d_2(b_j, b_k) : \mathcal{L}_2 \to (-\infty, \infty].$$

Therefore for $i = 1, \dots, n+l$

$$\begin{aligned} \mathcal{R}(Y_2(b_k))_i = \mathcal{R}(Y_2(b_k))(b_i) &= \bigoplus_{j=1}^{n} D_2(Y_2(b_i), Y(b_j)) \odot d_2(b_j, b_k) \\ &= \bigoplus_{j=1}^{n} d_2(b_i, b_j) \odot d_2(b_j, b_k) \end{aligned} \tag{79}$$

Consequently

$$\mathcal{R}(Y_2(b_k)) = \oplus_{i=1}^{n+l} \mathcal{R}(Y_2(b_k))_j \odot Y(b_i) = \bigoplus_{i=1}^{n+l} \bigoplus_{j=1}^{n} d_2(b_i, b_j) \odot d_2(b_j, b_k) \odot Y_2(b_i). \tag{80}$$

or equivalently

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{b_i \leqslant b_j \leqslant b_k} d_2(b_i, b_j) \odot d_2(b_j, b_k) \odot Y_2(b_i). \tag{81}$$

□

Remark 21. We see from eq 78, $\mathcal{R}(Y_2(b_k))_i = \bigoplus_{j=1}^{n} d_2(b_i, b_j) \odot d_2(b_j, b_k)$, that only summands such that $b_i \leqslant b_j \leqslant b_k$, will be finite.

Remark 22. Recall that, according to Theorem 3, $\mathcal{R}$ is a non-expansive map therefore

$$D(\mathcal{R}(Y(b_k)), \mathcal{R}(Y(b_l))) \leqslant D(Y(b_k), Y(b_l)). \tag{82}$$

We can use the previous proposition in order to approximate a text vector by the vectors corresponding to words making up that text.

Corollary 8. *Let $\mathcal{L} := \{b_1, \ldots, b_N\}$ be a probabilistic language model and let $W := \{w_1 \ldots, w_m\}$ be the set of words identified with $b_1, \ldots, b_m$ and considered as a probabilistic language model with all pairwise distances equal to infinity. Let $Y : \mathcal{L} \to P(\mathcal{L})$ be the Yoneda embedding. Let $\mathcal{R} : P(\mathcal{L}) \to P(\mathcal{L})$ be the non-expansive projection given by*

$$\mathcal{R} := \bigoplus_{j=1}^{m} D(-, Y(w_j)) \odot D(Y(w_j), -). \tag{83}$$

Consider $Y(b_k) \in P(\mathcal{L})$, then for $i, k = 1, \ldots, N$

$$\mathcal{R}(Y_2(b_k))_i = d_2(w_i, b_k) \tag{84}$$

and

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{i=1}^{N} d_2(w_i, b_k) \odot Y_2(w_i) = \bigoplus_{w_i \leqslant b_k} d_2(w_i, b_k) \odot Y_2(w_i) \tag{85}$$

Proof. Consider the projection $\mathcal{R} : P(\mathcal{L}) \to P(\mathcal{L})$ given

$$\mathcal{R} = \bigoplus_{j=1}^{m} D(-, Y(w_j)) \odot D(Y(w_j), -),$$

where $\mathrm{Im}(R) = \widetilde{\phi}(P(W))$.

We have identified b_j with w_j for $j = 1, \ldots, m$ therefore from corollary 5 we have for $i, k = 1, \ldots, N$

$$\mathcal{R}(Y_2(b_k))_i = \bigoplus_{j=1}^{l} d_2(b_i, w_j) \odot d_2(w_j, b_k). \tag{86}$$

However $d_2(b_i, w_j)$ is finite only of j=i and $w_j = b_i$. In that case $d_2(b_i, w_i) = 0$. Therefore for $i = 1, \ldots, N$

$$\mathcal{R}(Y_2(b_k))_i = d(w_i, b_k). \tag{87}$$

Consequently

$$\mathcal{R}(Y_2(b_k)) = \bigoplus_{i=1}^{N} d_2(w_i, b_k) \odot Y_2(w_i) = \bigoplus_{w_i \leqslant b_k} d_2(w_i, b_k) \odot Y_2(w_i) \ . \tag{88}$$ □

Corollary 9. *Let $\mathcal{L} := \{b_1, \ldots, b_N\}$ be a probabilistic language model and let $W := \{w_1 \ldots, w_m\}$ be the set of words identified with $b_1, \ldots, b_m$ and considered as a probabilistic language model with all pairwise distances equal to infinity. Let $Y : \mathcal{L} \to P(\mathcal{L})$ be the Yoneda embedding. Let $T \geqslant 0$ be a parameter (which is usually called temperature), then we have*

$$\mathcal{R}(Y(b_k)) = \lim_{T \to 0} -T \log\Big(\sum_{w_i \leqslant b_k} e^{-\frac{d(w_i, b_k)}{T}} e^{-\frac{Y(w_i)}{T}} \Big) \tag{89}$$

Therefore for small T we have

$$e^{-\frac{\mathcal{R}(Y(b_k))}{T}} \approx \sum_{w_i \leqslant b_k} e^{-\frac{d(w_i,b_k)}{T}} e^{-\frac{Y(w_i)}{T}} \tag{90}$$

Proof. Recall the identity

$$\lim_{T\to 0} -T\log(e^{-y/T}+e^{-z/T}) = \min\{y,z\}. \tag{91}$$

Then eq. (88) implies the result. □

Remark 23. Equation (90) is similar to the expression for a text value vector in terms of word value vectors as computed in the attention module of a transformer.

Remark 24. As already mentioned, it is natural to filter the probabilistic language $\mathcal{L}$ by the word length of texts. Define $\mathcal{L}_k$ to be the set of texts on $\mathcal{L}$ that have word length up to k. $\mathcal{L}_1$ will be the set of words. Each $\mathcal{L}_k$ inherits the structure of a probabilistic language model from $\mathcal{L}$. The inclusions define isometric embeddings $\phi_k : \mathcal{L}_k \to \mathcal{L}_{k+1}$. Then we can consider the non-expansive projections $\mathcal{R}_k : P(\mathcal{L}_{k+1}) \to P(\mathcal{L}_{k+1})$ where $\mathrm{Im}(\mathcal{R}_{k+1}) = \widetilde{\phi}_k(P(\mathcal{L}_k))$.

6. Duality between text extensions and restrictions

We have already considered the $(\min,+)$ semifield $\mathbb{R}_{\min} := ((-\infty,+\infty], \oplus_{\min}, \odot)$. To express duality results though, it will be convenient to work with the completed $(\min,+)$ semiring $\bar{\mathbb{R}}_{\min} := ([-\infty,+\infty], \oplus_{\min}, \odot)$ where as before $s \oplus_{\min} t := \min\{s,t\}$ and $s \odot t := s+t$ but we need to further determine how $-\infty$ and ∞ interact.

Indeed we specify that the element $+\infty$ remains absorbing, so $+\infty + s = +\infty$ holds for all element s, and in particular $(+\infty)+(-\infty) = +\infty$. The definition of $d_{\min}$ in (60) extends to this semiring. We also need to extend definitions of $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$:

Definition 10. *Let $P^-(\mathcal{L}), D)$ be the directed metric polyhedron*

$$P^-(\mathcal{L}) := \{x = (x_1,\dots,x_n) \in \{\mathbb{R}\cup\{\infty,-\infty\}\}^n \backslash \{(\infty,\dots,\infty)\} | x_i \leqslant x_j + d_{i,j}\}. \tag{92}$$

Moreover let $\widehat{P}^-(\mathcal{L}), D^t)$ be the directed metric polyhedron

$$\widehat{P}^-(\mathcal{L}) := \{y = (y_1,\dots,y_n) \in \{\mathbb{R}\cup\{\infty,-\infty\}^n \backslash \{(\infty,\dots,\infty)\} | y_i \leqslant y_j + d_{j,i}\}. \tag{93}$$

Remark 25. Recall that we added the case of a directed metric which can also take the value $-\infty$ in Remark 1 and now D is such a metric. Moreover we specified that $+\infty$ is absorbing in $\mathbb{R}_{\min}$. However the Funk metric D is defined using max so we have to specify further our convention to cover expressions that contain both min and max. For that, we simply use the relation $\max(s,t) = -\min\{-s,-t\}$ to transform any max in the expression to min so that we end up with an expression containing only min. Then we compute using the $(\min,+)$ convention that $+\infty$ is absorbing.

Equivalently we can use the same relation to transform any expression to one that contains only max. Then using $-\infty$ as the absorbing element gives the same answer.

Now, analogously to Proposition 10, if we consider $d_{\min}$ and $d^t_{\min}$ acting on $\{\mathbb{R}\cup\{\infty,-\infty\}\}^n\backslash\{(\infty,\dots,\infty)\}$ then we have

Proposition 15. *The polyhedron $P^-(\mathcal{L})$ is equal to $\mathrm{Im}(d_{min}) = \mathrm{Fix}(d_{\min})$ and the polyhedron $\widehat{P}^-(\mathcal{L})$ is equal to $\mathrm{Im}(d^t_{min}) = \mathrm{Fix}(d^t_{\min})$.*

Definition 11. *Define the pair of maps (A, B) as follows. If $y : \mathcal{L} \to [-\infty, \infty]$ and $x : \mathcal{L} \to [-\infty, \infty]$ then*

$$A(y) := d_{\min}(-y), B(x) := d^t_{\min}(-x) \tag{94}$$

Or in coordinates

$$A(y)_i := \min_j\{d_{i,j} - y_j\}, B(x)_j := \min_i\{d_{i,j} - x_i\} \tag{95}$$

We also denote by D^t the transpose metric with $D^t(x, y) := D(y, x)$.

In fact we will see that A and B on non-expansive maps with respect to these metrics.

The pair (A, B) forms an adjunction in the categorical or metric sense:

Proposition 16. *If $x : \mathcal{L} \to [-\infty, \infty]$ and $y : \mathcal{L} \to [-\infty, \infty]$ then we have*

$$D(Ay, x) = D^t(y, Bx) \tag{96}$$

Proof. $D(Ay, x) = \max_i\{x_i - \min_j\{d_{i,j} - y_j\}\} = -\min_i\{\min_j\{d_{i,j} - y_j\} - x_i\} = -\min_j\{\min_i\{d_{i,j} - x_i\} - y_j\} = \max\{y_j - \min_i\{d_{i,j} - x_i\}\} = D(Bx, y) = D^t(y, Bx)$. □

Remark 26.

(1) Note the resemblance of the pair of adjoint maps (A, B) with the Legendre-Fenchel transform where the metric is replaced by the inner product of a vector space.

(2) We note, for purposes of developing intuition, that the pair of adjoint maps (A, B) is similar to a pair of adjoint linear maps (A, A^*) on a vector space with inner product $\langle -, - \rangle$. Indeed in the usual linear algebra case $\langle Au, v\rangle = \langle u, A^*v\rangle$. Moreover we have already seen in Proposition 24 that D is a kind of tropical inner product. There is a crucial difference though that $\langle v, v\rangle = |v|^2$ while $D(x, x) = 0$. This reflects the fact that to go from usual algebra to tropical algebra we apply $-\log$.

We now have the following

Proposition 17. *We have $ABA = A$ and $BAB = B$ which implies that AB and BA are idempotent.*

Proof. This follows from the fact that $D(Ay, x) = D^t(y, Bx)$.

Indeed $D(ABAy, Ay) = D^t(BAy, BAy) = 0$

and $D(Ay, ABAy) = D^t(BAy, BAy) = 0$. Therefore $ABAx = Ax$. The equality $BAB = B$ is shown analogously. □

Let us now compute the fixed parts of the adjunction $\mathrm{Fix}(AB)$ and $\mathrm{Fix}(BA)$.

Proposition 18. *We have* $\mathrm{Fix}(AB) = \mathrm{Im}(A) = \mathrm{Im}(d_{\min})$ *and* $\mathrm{Fix}(BA) = \mathrm{Im}(B) = \mathrm{Im}(d^t_{\min})$.

Proof. This follows from the fact that $ABA = A$. Indeed clearly $\mathrm{Im}(A) \subset \mathrm{Fix}(AB)$. Moreover $\mathrm{Fix}(AB) \subset \mathrm{Im}(A)$ since $AB(x) = x$ says that $x \in \mathrm{Im}(A)$. Analogously for BA. □

In this case, due to the fact that $d_{\min}$ is an idempotent we can more explicitly compute the maps A and B

Proposition 19. *We have that*

$$A : \mathrm{Im}(d^t_{\min}) = \widehat{P}^-(\mathcal{L}) \to \mathrm{Im}(d_{\min}) = P^-(\mathcal{L}) \text{ is given by } A(y) = -y \tag{97}$$

and

$$B : \mathrm{Im}(d_{\min}) = P^-(\mathcal{L}) \to \mathrm{Im}(d^t_{\min}) = \widehat{P}^-(\mathcal{L}) \text{ is given by } B(x) = -x. \tag{98}$$

Proof. Consider $x \in \mathrm{Im}(d_{\min}) = Fix(d_{\min})$. We have $d_{\min}(x) = x \iff x_i = \min_j\{d_{i,j} + x_j\} \iff -x_j = \min_j\{d_{i,j} - x_i\} \iff d^t_{\min}(-x) = -x$. Therefore $B(y) = d^t_{\min}(-x) = -x$.

Analogously consider $y \in \mathrm{Im}(d^t_{\min})$. We have $d^t_{\min}(y) = y \iff d_{\min}(-y) = -y$. Therefore $A(y) = d_{\min}(-y) = -y$. □

Remark 27. Note that we can directly check the adjunction of Proposition 16 using our explicit formula from Proposition 19. Indeed $D(Ax, y) = D(-x, y) = \max_i\{y_i + x_i | x_i \neq -\infty\}$. Moreover $D^t(x, By) = D(-y, x) = \max_i\{x_i + y_i | y_i \neq -\infty\}$. Since we have a max expression, $-\infty$ is absorbing (see Remark 25) and consequently if $x_i = -\infty$ of if $y_i = -\infty$ then $x_i + y_i = -\infty$ therefore both these conditions can be ignored for taking the max and we get $D(-x, y) = D(-y, x)$.

The following theorem has been proved in [4] and [3] from different points of view and in different generalities. Another approach using category theory was used in Willerton [21].

Here we take advantage of the explicit computation in Proposition 19 which is true because $d^2_{\min} = d_{\min}$.

Theorem 4. *We have that*

$$A : \mathrm{Fix}(BA) = \mathrm{Im}(B) = \mathrm{Im}(d^t_{\min}) = \widehat{P}^-(\mathcal{L}) \to \mathrm{Fix}(AB) = \mathrm{Im}(A) = \mathrm{Im}(d_{\min}) = P(\mathcal{L})$$

and

$$B : \mathrm{Fix}(AB) = \mathrm{Im}(A) = \mathrm{Im}(d_{\min}) = P(\mathcal{L}) \to \mathrm{Fix}(BA) = \mathrm{Im}(B) = \mathrm{Im}(d^t_{\min}) = \widehat{P}^-(\mathcal{L})$$

are anti-isomorphisms. In other words they are one to one and onto and inverses. They are isometries, namely $D(Ay, Ay') = D^t(y, y')$. Finally we have

$$A(\lambda \odot y) = -\lambda \odot A(y), \tag{99}$$

$$A(y \oplus_{\min} y') = A(y) \oplus_{\max} A(y') \tag{100}$$

and

$$A(y \oplus_{\max} y') = A(y) \oplus_{\min} A(y') \tag{101}$$

and similarly for B.

Proof. From Proposition 16

$$A : \mathrm{Im}(d^t_{\min}) = \widehat{P}(\mathcal{L}) \to \mathrm{Im}(d_{\min}) = P(\mathcal{L}) \text{ is given by } A(y) = -y \tag{102}$$

and

$$B : \mathrm{Im}(d_{\min}) = P(\mathcal{L}) \to \mathrm{Im}(d^t_{\min}) = \widehat{P}(\mathcal{L}) \text{ is given by } B(x) = -x. \tag{103}$$

therefore A and B are one on one and onto and inverses.

Moreover $D(Ay, Ay') = D(-y, -y') = \max_i\{y'_i - y_i\} = D((y', y) = D^t(y, y')$. Furthermore, $A(\lambda \odot y)_i = -(\lambda + y_i) = -\lambda \odot A(y)_i$

$A(y \oplus_{\max} y')_i = -\max\{y_i, y'_i\} = \min\{-y_i, -y'_i\} = A(y)_i \oplus_{\min} A(y')_i$.

$A(y \oplus_{\min} y')_i = -\min\{y_i, y'_i\} = \max\{-y_i, -y'_i\} = A(y)_i \oplus_{\max} A(y')_i$.

□

(Note that $\mathrm{Im}(d_{\min})$ is $(\max,+)$ closed; this follows from Proposition 29.)

We have then that the $(\min+)$ column span $P^-(\mathcal{L})$, of $d_{\min}$ is anti isomorphic to the $(\min,+)$ row span $\widehat{P}^-(\mathcal{L})$ of $d_{\min}$ (as $\bar{\mathbb{R}}_{\min}$ modules) by the two inverse maps A and B and moreover they are isometric when considered with the directed metrics D and D^t respectively. (Recall also that in Proposition 12 we saw that D can be considered as tropical inner product.)

We will see an example of this bellow. First though we would like to make this map more explicit with respect to the rows and columns of the matrix d.

Proposition 20. *Consider $x \in \mathrm{Im}(d_{\min})$. We have that*

$$x = \oplus_j x_j \odot d(-,a_j) \text{ and then } B(x) = -x = \oplus_j -x_j \odot d(a_j,-). \tag{104}$$

In particular if $x = d(-,a_k)$ then

$$d(-,a_k) = \oplus_{a_j \leqslant a_k} d(a_j,a_k) \odot d(-,a_j) \tag{105}$$

and

$$-d(-,a_k) = \oplus_{a_j \leqslant a_k} -d(a_j,a_k) \odot d(a_j,-) \tag{106}$$

Analogously for $y \in \mathrm{Im}(d^t)$ we have

$$y = \oplus_i y_i \odot d(a_i,-) \text{ and then } A(y) = -y = \oplus_i -y_i \odot d(-,a_i). \tag{107}$$

In particular if $y = d(a_k,-)$ then

$$d(a_k,-) = \oplus_{a_k \leqslant a_i} d(a_k,a_i) \odot d(a_i,-) \tag{108}$$

and

$$-d(a_k,-) = \oplus_{a_k \leqslant a_i} -d(a_k,a_i) \odot d(-,a_i). \tag{109}$$

Proof. We have $x \in \mathrm{Im}(d_{\min}) \iff d_{\min} x = x \iff x = \oplus_j x_j \odot d(-,a_j)$. From Proposition 19 we then have $d^t_{\min}(-x) = -x$ which is equivalent to $-x = \oplus_j -x_i \odot d(a_j,-)$. This proves (104).

Now if $x := d(-,a_k)$ then $x_j = x(a_j) = d(a_j,a_k) = d_{j,k}$. Then from Proposition 11 we have $d(-,a_k) = \oplus_{a_j \leqslant a_k} d_{j,k} \odot d(-,a_j)$ therefore from (104) it follows that $-d(-,a_k) = \oplus_{a_j \leqslant a_k} -d(a_j,a_k) \odot d(a_j,-)$. The proof for $y \in \mathrm{Im}(d^t_{\min})$ and for $y := d(a_k,-)$ is analogous. □

Remark 28. Note that all results in this section hold for a general directed metric space.

Example 1. We now show a simple example of a probabilistic language model $(\mathcal{L}, d_1)$ along with $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$. We will also see the correspondence of extremal rays with connected lower sets for $P(\mathcal{L})$ and connected upper sets for $\widehat{P}(\mathcal{L})$ as described in Theorem 2.

We will actually consider the corresponding polyhedral cones $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$ and show in the figures the polyhedra $Q_0(\mathcal{L})$ and $\widehat{Q}^0(\mathcal{L})$ (Definition 6) which are their intersections with the unit simplex.

We will further illustrate the duality between completions $P^-(\mathcal{L})$ and $\widehat{P}^-(\mathcal{L})$ by making a uniform approximation of infinities in d with a big number M.

Indeed consider the corpus to be $\mathcal{L} := \{\text{red, colour, red colour}\}$. Denote "red" by "r", "colour" by "c" and "red colour" by "rc".

Let the metric d be given by eq. (110):

$$d = \begin{array}{c} \\ r \\ c \\ rc \end{array} \begin{array}{c} \begin{array}{ccc} r & c & rc \end{array} \\ \begin{pmatrix} 0 & \infty & \log 2 \\ \infty & 0 & \log 3 \\ \infty & \infty & 0 \end{pmatrix} \end{array} \tag{110}$$

Recall that in general $e^{-d_{i,j}} = \Pr(a_j|a_i)$ and thus the corresponding matrix of probabilities of extensions is

$$\Pr = \begin{array}{c} \\ r \\ c \\ rc \end{array} \begin{array}{c} \begin{array}{ccc} r & c & rc \end{array} \\ \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{3} \\ 0 & 0 & 1 \end{pmatrix} \end{array} \tag{111}$$

This means for example that $\Pr(rc|r) = \frac{1}{2}$ and $\Pr(c|r) = 0$, while $P(r|r) = 1$.

Recall that the equations for $P(\mathcal{L}) = \mathrm{Im}(d_{\min})$ are (Definition 4): $x_i \leqslant d_{i,j} + x_j$.

Letting $z_i := e^{-x_i}$ we have that the equations for $Q(\mathcal{L})$ are (Definition 5): $z_i \geqslant e^{-d_{i,j}} z_j$.

Therefore in our case we get that the polyhedral cone $Q(\mathcal{L})$ is defined by inequalities

$$z_1 \geqslant \frac{1}{2} z_3, z_2 \geqslant \frac{1}{3} z_3, z_1 \geqslant 0, z_2 \geqslant 0, z_3 \geqslant 0. \tag{112}$$

The intersection $Q_0(\mathcal{L})$ of $Q(\mathcal{L})$ with the unit simplex is shown on the right in Figure 1. Notice that it has three vertices.

Analogously, the equations for $\widehat{P}(\mathcal{L}) = \mathrm{Im}(d^t_{\min})$ (Definition 4) are $y_j \leqslant d_{i,j} + y_i$.

Letting $u_i := e^{-y_i}$ we have that the equations for $\widehat{Q}(\mathcal{L})$ are (Definition 4) $u_j \geqslant e^{-d_{i,j}} u_i$.

Therefore in our case we get that the polyhedral cone $\widehat{Q}(\mathcal{L})$ is defined by inequalities

$$u_3 \geqslant \frac{1}{2} u_1, u_3 \geqslant \frac{1}{3} u_2, u_1 \geqslant 0, u_2 \geqslant 0, u_3 \geqslant 0. \tag{113}$$

The intersection $\widehat{Q}_0(\mathcal{L})$ of $\widehat{Q}(\mathcal{L})$ with the unit simplex is shown on the left in Figure 1. Notice that it has four vertices.

Denote the lower set generated by "a" by $(a)_l$ and the upper set generated by a by $(a)_u$.

From Theorem 2, extremal rays of $Q(\mathcal{L})$ correspond to *connected lower sets* of $\mathcal{L}$. There are three and they are all principal: $(r)_l = \{r\}, (c)_l = \{c\}, (rc)_l = \{r, c, rc\}$. These give rise to the three vertices of $Q_0(\mathcal{L})$ as we can see in Fig 1. (Note that $(r, c)_l$ is not connected so it does not correspond to an extremal ray of $Q(\mathcal{L})$.

From Corollary 4, extremal rays of $\widehat{Q}(\mathcal{L})$ correspond to *connected upper sets* of $\mathcal{L}$. The principal ones are $(r)_u = \{r, rc\}, (c)_u = \{c, rc\}, (rc)_u = \{rc\}$ and a non-principal one $(r, c)_u = \{r, c, rc\}$. This extremal ray is not in the image of the Yoneda embedding. The corresponding four vertices of $\widehat{Q}_0(\mathcal{L})$ are shown on the left in Figure 1.

Notice that the number of extremal rays of $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ are actually different.

Now note that in general the $\mathbb{R}_{\min}$ module $P(\mathcal{L}) = \mathrm{Im}(d_{\min})$ is a geometric object. In fact $Q(\mathcal{L})$ is a polyhedral cone and $Q_0(\mathcal{L})$ is a polyhedron. However the

$\bar{\mathbb{R}}_{\min}$ module $P^-(\mathcal{L})$ is not obviously geometric. In order to approximate with a geometric object and be able to visualize the duality between the $P^-(\mathcal{L})$ and $\widehat{P}^-(\mathcal{L})$ it is natural to "truncate" the matrix d, replacing the $+\infty$ entries by a sufficiently large number M, leading to the new matrix:

$$d^M = \begin{array}{c} \\ r \\ c \\ rc \end{array} \begin{array}{c} \begin{array}{ccc} r & c & rc \end{array} \\ \begin{pmatrix} 0 & M & \log 2 \\ M & 0 & \log 3 \\ M & M & 0 \end{pmatrix} \end{array} \tag{114}$$

This matrix is sill a directed metric, satisfying $(d^M_{\min})^2 = d^M_{\min}$ *but it does not any more represent a probabilistic language model.*

We can consider $P_M(\mathcal{L}) := \mathrm{Im}(d^M_{\min})$ as in (Definition 4) and $Q_M(\mathcal{L})$ as in (Definition 5). Then the intersection $Q_{M,0}(\mathcal{L})$ of $Q_M(\mathcal{L})$ with the unit simplex is depicted in Figure 2 on the right.

Moreover we consider $\widehat{P}_M(\mathcal{L}) := \mathrm{Im}((d^M_{\min})^t)$ as in (Definition 4) and $\widehat{Q}_M(\mathcal{L})$ as in (Definition 5). Then the intersection $\widehat{Q}_{M,0}(\mathcal{L})$ of $\widehat{Q}_M(\mathcal{L})$ with the unit simplex is depicted in Figure 2 on the right.

Observe that the duality preserves the number of extreme points inside the interior of the simplex, and that the sets of Figure 2 converge to the sets of Figure 1 as $M \to \infty$.

Also note that the duality map between $P(\mathcal{L})$ an $\widehat{P}(\mathcal{L})$ is $x_i \to y_i = -x_i$ for $i = 1, 2, 3$. We also have $z_i := e^{-x_i}$ and $u_i := e^{-y_i}$ Therefore the map between the polyhedra $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$ is $z_i \to u_i = \frac{1}{z_i}$ for $i = 1, 2, 3$.

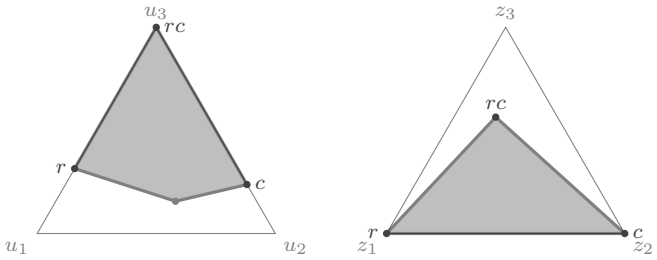

FIGURE 1. The cross section $\widehat{Q}_0(\mathcal{L})$ of the polyhedral cone $\hat{Q}(\mathcal{L})$ arising from the metric of d (left). Every vector $d(r,-), d(c,-), d(rc,-)$ determines an extreme point of the cross section, denoted by r, c, or rc. There is a fourth extreme point (shown in gray) corresponding to a non-principal upper set. The cross section $Q_0(\mathcal{L})$ (right). There are three extreme points, which correspond to the vectors $d(-,r), d(-,c), d(-,rc)$.

Remark 29. Note that approximating uniformly infinities in the matrix d with a big number M can be done in general. This is helpful since the duality theorem is easier to illustrate for matrices with finite entries. Of course, the proof of the duality theorem in section 6 goes through with coefficients in $(-\infty, \infty)$. (Note also that the Develin-Sturmfels version [4] of the adjunction between the tropical

column span and the tropical row span is exactly about matrices with finite entries, whereas the version of [3] deals with matrices with possibly infinite entries.)

However as we already saw in the example replacing ∞ with M in a directed metric d that defines a probabilistic language model gives a metric that is no longer a language model. The limit of polyhedra for $M \to \infty$ will give the polyhedra for the original metric.

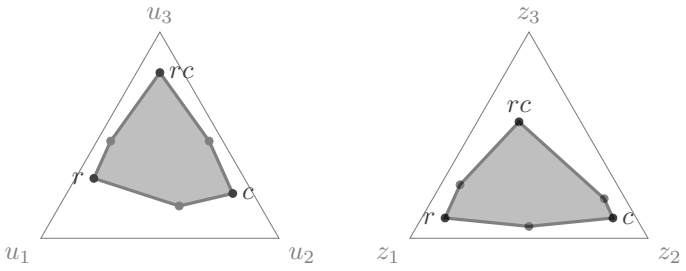

FIGURE 2. The duality between the columns and row spaces of metric matrices (Proposition 19 and Theorem 4) illustrated. On the right $\mathrm{Im}(d^M_{\min})$ and on the left $\mathrm{Im}((d^M_{\min})^t)$

Remark 30. We have said that we can encode the meaning of a_k by $d(-, a_k)$. If a_k is a word this contains very little information. We already addressed the solution to this problem in section Section 4.1. Another solution is, using the duality between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ explained previously and in particular Equation (105), Equation (106).

7. Extremal Rays in terms of text vectors

We have seen in Theorem 1 that the original texts in $\mathcal{L}$, mapped by the Yoneda isometric embedding $Y : \mathcal{L} \to P(\mathcal{L})$, appear as extremal rays (corresponding to principal lower sets) in the polyhedron $P(\mathcal{L})$ and the polyhedral cone $Q(\mathcal{L})$. As proven in Proposition 7, there are in general many other extremal rays of $P(\mathcal{L})$ corresponding to connected lower sets of $\mathcal{L}$. Nevertheless extremal rays in the image of Y, $(\min, +)$ generate $P(\mathcal{L})$ as we have already seen in Corollary 6 where we showed that

$$x \in P(\mathcal{L}) \iff x = \oplus_j D(Y(a_j), x) \odot Y(a_j). \tag{115}$$

Recall that we think of $Y(a_k) := d(-, a_k) \in P(\mathcal{L})$ as encoding the meaning of text a_k according to the statistical semantics principal.

Recall Definition 9, where we introduced the saturation graph $S(x)$ for $x \in P(\mathcal{L})$. We shall also consider the *undirected* saturation graph, obtained by forgetting the orientation of the edges in $S(x)$.

Proposition 21. *A vector $x \in P(\mathcal{L})$ can be written as a tropical linear combination of terminal elements of its saturation graph $S(x)$. Specifically if $b_1, \dots b_k$ are the terminal elements in $S(x)$, then*

$$x = \oplus_j D(Y(b_j), x) \odot Y(b_j). \tag{116}$$

Moreover, if the saturation graph of x has s undirected connected components, this vector belongs to a face of $P(\mathcal{L})$ of dimension s.

Proof. Let $x \in P(\mathcal{L})$. If i is not terminal in the saturation graph $S(x)$, then, there is a $j_1 \neq i$ such that $x_i = d_{i,j_1} + x_{j_1}$. Similarly, if j_1 is not terminal, there is $j_2 \neq j_1$ such that $x_{j_1} = d_{j_1,j_2} + x_{j_2}$ and thus $x_i = d_{i,j_1} + d_{j_1,j_2} + x_{j_2}$. Continuing this way we get $x_i = d_{i,j_1} + d_{j_1,j_2} + \dots d_{j_{n-1},j_n} + x_{j_n}$ where a_{j_n} is a terminal element of the saturation graph $S(x)$.

We know that this stops at a terminal element because there are no cycles of positive weight in the digraph of d. (This is the graph that has an edge from i to j with weight $d_{i,j}$ when $d_{i,j}$ is not infinity.)

Using the triangular inequality, we deduce that $x_i \geqslant \oplus_{j\in T} d_{i,j} \odot x_j$, where the sum is taken over the set T of terminal nodes of $S(x)$, and so, $x \geqslant \oplus_{j\in T} D(Y(a_i), Y(a_j)) + D(Y(a_j), x)$. Conversely, by definition of $P(\mathcal{L})$, $x \leqslant \oplus_k d_{ik} \odot x_k = \oplus_k D(Y(a_i), Y(a_k)) \odot D(Y(a_k), x)$ where now the sum is taken over all the indices k (possibly non terminal). This entails that (116) holds.

Finally, arguing as in the proof of Theorem 2, we get the rank of the family of active constraints at point x is given by the number s of connected component of the undirected saturation graph of x. Hence, x belongs to a face of dimension s. □

Remark 31. Note that Proposition 21 holds for a general directed metric space $\mathcal{L}$ and not just for $(\mathcal{L}, d)$ a probabilistic language model.

We can now find explicit $(min, +)$ expressions for generators of extremal rays corresponding to non principal lower sets.

Proposition 22. *Let $\mathcal{L}$ be a probabilistic language model with the empty text a_0 included. Let x denote an extremal ray corresponding to the lower set generated by $\{b_1, \dots b_n\}$. Then*

$$x = \oplus_i \log \Pr(b_i) \odot Y(b_i). \tag{117}$$

Proof. From Proposition 21 we have

$$x = \oplus_j x(b_j) \odot Y(b_j).$$

From corollary 3 we have $x(b_j) = -\log \frac{1}{\Pr(b_j)} = \log \Pr(b_j)$. This proves the result. □

Remark 32. We point out that the terminal elements $b_1, \dots b_k$ of $S(x)$ function like an orthonormal basis with respect to D, namely

$$D(Y(b_i), Y(b_j)) = d(b_i, b_j) = \infty \text{ if } i \neq j. \tag{118}$$

So for example if we know that there are λ_j such that
$x = \oplus_j \lambda_j \odot Y(b_j)$ Then $D(Y(b_i), x) = \oplus_j \lambda_j \odot D(Y(b_i), Y(b_j)) = \lambda_i$.

Proposition 23. *Let $T \in [0, \infty)$ be a parameter which will be called temperature. Consider $x \in P(\mathcal{L})$ an extremal ray and let $x = \oplus_j D(Y(b_j), x) \odot Y(b_j)$ where b_j are the terminal elements of the saturation graph S(x). Let $v_j := e^{Y(b_j)}$. Then we have*

$$x = \lim_{T\to 0} -T \log\Big(\sum_j e^{-D(Y(b_j),x)/T} e^{Y(b_j)}\Big) \tag{119}$$

and therefore, for small T

$$e^{-x/T} \approx \sum_j e^{-D(Y(b_j),x)/T} v_j \tag{120}$$

Proof. Recall the identity

$$\lim_{T\to 0} -T\log(e^{-y/T}+e^{-z/T}) = \min\{y,z\}. \tag{121}$$

If $x \in P(\mathcal{L})$, by the previous proposition $x = \oplus_j D(Y(b_j),x) \odot Y(b_j)$ where b_j are terminal elements. Then we have

$$x = \lim_{T\to 0} -T\log \sum_j e^{-D(Y(b_j),x)/T} e^{Y(b_j)} \tag{122}$$

and if we put $v_j \coloneqq e^{Y(b_j)}$, then for small T, we get

$$e^{-x/T} \approx \sum_j e^{-D(Y(b_j),x)/T} v_j \ . \tag{123}$$

□

8. $P^-(\mathcal{L})$ as the lattice completion of the Isbell completion

We have seen that $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ generalize the lower set and upper set completions respectively from the poset $\mathcal{L}$ to the directed metric space $(\mathcal{L}, d)$, at least in the case where $\mathcal{L}$ contains the empty text a_0 which is the bottom element.

However there is another completion of a poset, called the Dedekind-MacNeille completion (which also generalizes the so called notion of formal concepts).

It is known that the generalization of the Dedekind MacNeille completion from posets to directed metric spaces is the so called Isbell completion, which is the fixed part of the Isbell adjunction.

This is also relevant to our situation as it turns out to be defined by $d_{\max}$.

This was studied in [11, 21] with $[0,\infty]$ coefficients. In that case the Isbell completion is identified with the directed tight span of Hirai and Koichi [7].

We will instead define the Isbell adjunction using the extended semi ring $[-\infty,\infty]$ as we did with the $d_{\min}$ adjunction in section 6.

Recall that in section 6 Remark 25 we explained the conventions for working with $(\min,+)$ and $(\max,+)$ on $[-\infty,\infty]$. We use the same here.

Given $x : \mathcal{L} \to [-\infty,\infty]$ and $y : \mathcal{L} \to [-\infty,\infty]$ define $d_{\max}$ and $d^t_{\max}$ by

$$d_{\max}(x)_i \coloneqq \max_j\{d_{i,j}+x_j\} \text{ and } d^t_{\max}(y)_j \coloneqq \max_i\{d_{i,j}+y_i\} \tag{124}$$

Extending the definition in [11, 21] by using $[-\infty,\infty]$ coefficients we have that

Definition 12. *The Isbell adjunction is the pair of maps (L,R) defined as follows. If $x : \mathcal{L} \to [-\infty,\infty]$ and $y : \mathcal{L} \to [-\infty,\infty]$ then*

$$L(x) \coloneqq d_{\max}(-x) \text{ and } R(x) \coloneqq d^t_{\max}(-y) \tag{125}$$

Or in coordinates

$$L(x)_i \coloneqq \max_j\{d_{i,j}-x_j\} \text{ and } R(x)_j \coloneqq \max_i\{d_{i,j}-y_i\} \tag{126}$$

Recall from section 6 that the Funk metric D, is still well defined by $D(x,y) \coloneqq \max_i\{y_i - x_i | x_i \neq \infty\}$. We also denoted by D^t the transpose metric with $D^t(x,y) \coloneqq D(y,x)$.

Remark 33. Note that

$$L(x)_i := \max_j\{d_{i,j} - x_j\} = D(x, d(a_i, -)) = D(x, \widehat{Y}(a_i)) \tag{127}$$

and

$$R(y)_j := \max_i\{d_{i,j} - y_i\} = D(y, d(-, a_j)) = D(y, Y(a_j)) \tag{128}$$

The pair (L, R) forms an adjunction in the categorical or metric sense:

Proposition 24. *If $x : \mathcal{L} \to [-\infty, \infty]$ and $y : \mathcal{L} \to [-\infty, \infty]$ then we have $D^t(Lx, y) = D(x, Ry)$.*

Proof. $D^t(Lx, y) = D(y, Lx) = \max_i\{\max_j\{d_{i,j} - x_j\} - y_i\} = \max_j\{\max_i\{d_{i,j} - y_i\} - x_j\} = D(x, Ry)$. □

We now have the following

Proposition 25. *We have $LRL = L$ and $RLR = R$ which implies that LR and RL are idempotent.*

Proof. This follows from the fact that $D^t(Lx, y) = D(x, Ry)$. Indeed $D^t(LRLx, Lx) = D(RLx, RLx) = 0$ and $D^t(Lx, LRLx) = D(RLx, RLx) = 0$. Therefore $LRLx = Lx$. The equality $RLR = R$ is shown analogously. □

Let us now compute the fixed parts of the adjunction $\mathrm{Fix}(LR)$ and $\mathrm{Fix}(RL)$.

Proposition 26. *We have* $\mathrm{Fix}(LR) = \mathrm{Im}(L) = \mathrm{Im}(d_{\max})$ *and* $\mathrm{Fix}(RL) = \mathrm{Im}(R) = \mathrm{Im}(d^t_{\max})$.

Proof. This follows from the fact that $LRL = L$. Indeed clearly $\mathrm{Im}(L) \subset \mathrm{Fix}(LR)$. Moreover $\mathrm{Fix}(LR) \subset \mathrm{Im}(L)$ since $LR(y) = y$ says that $y \in \mathrm{Im}(L)$. □

As before we have the following

Proposition 27. *We have that*

$$L : \mathrm{Fix}(RL) = \mathrm{Im}(R) = \mathrm{Im}(d^t_{\max}) \to \mathrm{Fix}(LR) = \mathrm{Im}(L) = \mathrm{Im}(d_{\max})$$

and

$$R : \mathrm{Fix}(LR) = \mathrm{Im}(L) = \mathrm{Im}(d_{\max}) \to \mathrm{Fix}(RL) = \mathrm{Im}(R) = \mathrm{Im}(d^t_{\max})$$

are anti-isomorphisms. In other words they are one to one and onto and inverses. They are isometries, namely $D(Lx, Lx') = D^t(x, x')$. Finally we have

$$L(\lambda \odot x) = -\lambda \odot L(x) \text{ and } L(x \oplus_{\min} y) = L(x) \oplus_{\max} L(y) \tag{129}$$

and similarly for R.

Proof. First let us check that L and R are one to one and onto. Consider $x, x' \in \mathrm{Fix}(RL)$. If $L(x) = L(x')$ then $RL(x) = RL(x')$ and therefore $x = x'$. Also if $y \in \mathrm{Fix}(LR)$ then $y = L(R(y))$.

Moreover $D(Lx, Lx') = D_{op}(RLx, x') = D_{op}(x, x') = D(x', x)$

Next we check the tropical antilinearity.

$L(\lambda \odot x)_i = \max_j\{d_{i,j} - \lambda - x_j\} = \max_j\{d_{i,j} - x_j\} - \lambda = L(x)_i - \lambda = (-\lambda \odot L(x))_i$.

Moreover

$L(x \oplus_{\min} y)_i = \max_j\{d_{i,j} - \min\{x_j, y_j\}\} = \max_j\{d_{i,j} + \max\{-x_j, -y_j\} = \max_j\{\max\{d_{i,j} - x_j, d_{i,j} - y_j\}\} = \max\{\max_j\{d_{i,j} - x_j\}, \max_j\{d_{i,j} - y_j\}\} = (L(x) \oplus_{\max} L(y))_i$.

□

We have that the tropical linear space $\mathrm{Im}(d_{\max})$ is anti isomorphic to $\mathrm{Im}(d^t_{\max})$ by the two inverse maps

$$R : \mathrm{Im}(d_{\max}) \to \mathrm{Im}(d^t_{\max}) \text{ and } L : \mathrm{Im}(d^t_{\max}) \to \mathrm{Im}(d_{\max}).$$

Proposition 28. *The Yoneda isometric embedding $Y : \mathcal{L} \to P(\mathcal{L})$ given by $Y(a) := d(-, a)$ and the co-Yoneda isometric embedding $\widehat{Y} : \mathcal{L} \to \widehat{P}(\mathcal{L})$ given by $\widehat{Y}(a) := d(a, -)$, are compatible with the anti-isomorphisms L and R above, in the sense that*

$$\widehat{Y}(a) = R(Y(a)) \text{ and } Y(a) = L(\widehat{Y}(a)). \tag{130}$$

Proof. We have $d_{i,j} \leqslant d_{i,k} + d_{k,j}$, therefore

$$L(\widehat{Y}(a_k))_i = L(d(a_k, -))_i = \max_j \{d_{ij} - d_{k,j}\} = d_{i,k} = d((-, a_k)_i = Y(a_k)_i.$$

Analogously

$$R(Y(a_k))_j = R(d(-, a_k))_j = \max_i \{d_{ij} - d_{i,k}\} = d_{k,j} = d((a_k, -)_j = \widehat{Y}(a_k)_j \ . \quad \square$$

As mentioned earlier according to a theorem of Willerton [21]

Theorem 5. *The directed tight span of Hirai and Koichi* [7] *is the same as the fixed parts of the Isbell adjunction when using* $[0, \infty]$ *coefficients and the truncated* max *operations .*

We denote the Isbell completion with $[-\infty, \infty]$ coefficients by $\tilde{I}(\mathcal{L})$.

Let us finally explore the relation between $P(\mathcal{L}) = \mathrm{Im}(d_{\min})$ and $\tilde{I}(\mathcal{L}) = \mathrm{Im}(d_{\max})$.

Proposition 29. *The polyhedron $P(\mathcal{L}) = \mathrm{Im}(d_{\min})$ is the lattice completion of $\mathrm{Im}(d_{\max})$ when using $(-\infty, \infty]$ coefficients and $P^-(\mathcal{L})$ is the lattice completion of $\mathrm{Im}(d_{\max})$ when using $[-\infty, \infty]$ coefficients.*

Proof. Recall that since $d^2_{\min} = d_{\min}$ we have $\mathrm{Im}(d_{\min}) = \{x | dx = x\}$. Moreover if Id is the $(\min, +)$ identity matrix, namely $\mathrm{Id}_{i,i} = 0$ and $\mathrm{Id}_{i,j} = \infty$ for $i \neq j$, then, since $d_{i,} = 0$, we have $d \leqslant \mathrm{Id}$ and therefore we always have $dx \leqslant x$.

It follows that $x = dx \iff x \leqslant dx$.

We want to show that if $x \leqslant dx$ and $y \leqslant dy$ then

$$\max\{x, y\} \leqslant d(\max\{x, y\})$$

which will imply that $\max\{x, y\} \in \mathrm{Im}(d_{\min})$.

Indeed if $\max\{x, y\} = x$ then $d(\max\{x, y\}) = dx \geqslant x \geqslant y$ and if $\max\{x, y\} = y$ then $d(\max\{x, y\}) = dy \geqslant y \geqslant x$, therefore $x \leqslant d(\max\{x, y\})$ and $y \leqslant d(\max\{x, y\})$ which implies that $\max\{x, y\} \leqslant d(\max\{x, y\})$.

We have shown therefore that $\mathrm{Im}(d_{\min})$ is closed under the max operation. Both $\mathrm{Im}(d_{\min})$ and $\mathrm{Im}(d_{\max})$ are generated by the vectors $d(-, a_i)$ therefore the result is proved. $\square$

Example 2. To illustrate the difference between $\mathrm{Im}(d_{\min})$ and $\mathrm{Im}(d_{\max})$ (albeit for a symmetric and finite metric) we provide the following example: Consider the discrete metric on three points

$$d_2 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \tag{131}$$

The associated (min,+)- module $\mathcal{P}(d_2)$ is shown on Figure 3, left and the $(\max, +)$ module on the right.

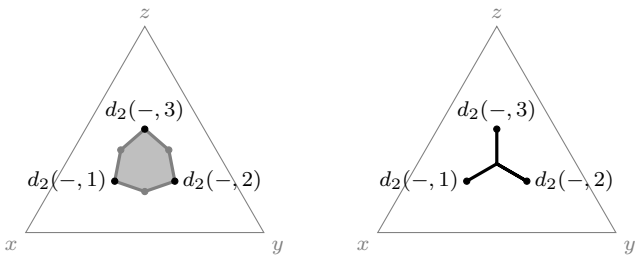

FIGURE 3. Tropical module generated by the discrete metric d_2 of Equation (131). The pseudo-vertices (vertices of the polyhedral complex that do not arise from tropical generators) are shown in gray. (left) The (max,+)-span (right).

9. Some comments about Probabilistic Language Models

We would finally like to gather some comments about how to interpret probabilistic language models $(\mathcal{L}, \leqslant, \Pr)$ and what they imply. Some of these were stated already in section Section 4.1

(1) We note that the construction of $P(\mathcal{L})$ explains why it is natural to have vectors in a problem of language. In fact we naturally get Boltzmann weighted linear combinations Equation (90), Equation (62), Equation (62) which is what is introduced by hand in the attention layers of the transformer and the final layer where the distribution over possible next words is determined
(2) If the transformer is learning $\widehat{P}(\mathcal{L})$ or equivalently $\widehat{Q}(\mathcal{L})$ it would be learning a convex body which could explain why its training is efficient in the first place.
(3) Assuming that the transformer is learning the polyhedron $\widehat{P}(\mathcal{L})$ or equivalently $\widehat{Q}(\mathcal{L})$ it would be learning an effective representation of Yoneda embeddings of texts. This can then be interpreted as solving the huge $(\min, +)$ linear systems in Equation (62), Equation (63), (Proposition 11).
(4) The duality explained in section 6 between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ shows how to resolve the paradox that both $d(-, a_k)$ and $d(a_k, -)$ should equally well encode the meaning of a text a_k, given a probabilistic language model $(\mathcal{L}, \leqslant, \Pr)$. This is most striking when a_k is a single word. In that case $d(-, a_k)$ is supported only on a_k, but we have that $-d(-, a_k) = -d(a_k, -)$. This was explained in Section 6, Proposition 20, Remark 30. It also showcases the notion that the meaning of a text a_k is not just encoded by $d(-, a_k)$ or $d(a_k, -)$ but by the whole ambient spaces $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ respectively.

Appendix A. Categorical interpretation

The metric polyhedra $(P(\mathcal{L}), D)$ and $(\widehat{P}(\mathcal{L}), D^t)$, as well as the polyhedral cones $Q(\mathcal{L})$ and $\widehat{Q}(\mathcal{L})$, arise from a categorical point of view [11, 21, 1]. In fact all

constructions have a categorical interpretations and we will briefly explain these here.

To begin with, we can consider the probabilistic language model $(\mathcal{L}, d)$ to be a category enriched over the monoidal closed category $(-\infty, \infty]$, with monoidal structure given by addition and Hom given by considering $(-\infty, , \infty]$ as a poset with the opposite of the usual order of numbers. The Hom between objects a_i and a_j in $\mathcal{L}$, is $d(a_i, a_j)$ and the triangle inequality is the composition of morphisms.

This construction (using $[0, \infty]$ instead of $(-\infty, \infty]$) was explained in [1])

Then $P(\mathcal{L})$ is the category of presheaves on $\mathcal{L}$ namely the category of enriched functors $\mathcal{L}^{\mathrm{op}} \to (-\infty, \infty]$ where $(-\infty, \infty])$ is considered as a category enriched over itself with internal Hom given by the directed metric $d_{\mathbb{R}}$ on $(-\infty, \infty]$ where $d_{\mathbb{R}}(s,t) = t - s$. This follows from the fact that we can think of the points $x \in P(\mathcal{L})$ as non-expansive functions on $\mathcal{L}$ as we have seen in Proposition 4. Indeed

$$P(\mathcal{L}) = \{x : (\mathcal{L}, d^t) \to ((-\infty, \infty], d_{\mathbb{R}}) | x \text{ is non-expansive.}\} \tag{132}$$

Moreover, the Funk directed metric D on $P(\mathcal{L})$ is the Hom on presheaves.

The isometric embedding $Y : \mathcal{L} \hookrightarrow P(\mathcal{L})$ is the Yoneda embedding, $Y(a_k)$ is a representable presheaf and the fact that $x_i = x(a_i) = D(Y(a_i), x)$, is the Yoneda lemma.

On the other hand $\widehat{P}(\mathcal{L})$ is the category of co-presheaves and $\widehat{Y}$ is the co-Yoneda embedding. The tropical anti-isomorphims between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ as already explained follows from an adjunction between $A(x) := d_{\min}(-x)$ and $B(y) = d^t_{\min}(-y)$.

Finally it was proven in [21] that the directed tight span $DTS(\mathcal{L})$ (defined in [7]) is the Isbell completion, with $[0, \infty]$ as enriching category, of the enriched category $\mathcal{L}$. Namely the fixed part of the Isbell adjunction which is given by $(L(x))_i := \max_j \{d_{i,j} - x_j\}$ and $(R(y))_j := \max_i \{d_{i,j} - y_i\}$ (where we use trancated difference so the result is always positive). We instead define the Isbell adjunction with enriching category $[-\infty, \infty]$.

The fact that the category of presheaves $P(\mathcal{L})$ is the $(\min, +)$ span of the images of the Yoneda embedding reflects the fact that colimits are given by min and every presheaf is a weighted colimit of representables.

On the other hand the Isbell completion is given by presheaves which are weighted limits of representables since limits are given by max and it is smaller that $P(\mathcal{L})$ since in general not every presheaf is such a weighted limit.

Appendix B. Syntax to Semantics and Morita equivalence

The problem of encoding allowed (with some probability) sequences of symbols, by some mathematical structure can be located in the realm of a very basic duality in mathematics.

Traditionally language has been modeled as a monoid generated by words. We can go from the monoid to a poset by considering the monoid as a category with one object and arrows corresponding to texts and constructing the factorization category (also called the twisted arrow category). This produces exactly the poset of texts with the subtext order as we have used in our probabilistic language model.

Considering the subtext poset makes it easier to add probabilities and we are led naturally to the probabilistic language model we defined which is a special case of a directed metric space. In Appendix A we saw that this is an enriched category.

In the monoid case we consider that the meaning of a text eg "red" is given by the ideal generated by red which contains all texts containg red.

In the poset case it is the same, where ideals and filters correspond to principal lower and upper sets.

This is a mathematical incarnation of the distributional semantics principle.

Now there is a very general and basic concept of duality in mathematics that in the commutative case takes the form of a duality between algebra and geometry.

The most basic case is, given a commutative algebra, to consider the space of (prime) ideals.

This is called the spec and can be thought of as a space on which the algebra of functions is the commutative algebra we started with. For example if we consider the algebra $\mathbb{C}[x, y]$ of complex polynomials in two variables then prime ideas are ideals generated by monomials $(x-a)(y-b)$ for any $a, b \in \mathbb{C}$ and therefore the space of ideals is $\mathbb{C}^2$ i.e. the complex plane. The duality then is between the commutative algebra $\mathbb{C}[x, y]$ and the space of ideals $\mathbb{C}^2$. (This so called spec construction is the cornerstone of algebraic geometry.)

We can try to extend this kind of duality for monoids, posets, and for our enriched category. Ideals in a monoid are modules over the monoid and in general we have to consider modules. Now moving to the case of an algebra, a module over the algebra (a representation) is a presheaf over the corresponding category. This is the category with one object and arrows given by the elements of the algebra. In general in a category the presheaves play the role of modules.

In our case modules i.e. presheaves are the non-expansive maps (Proposition 3) and the space $P(\mathcal{L})$ is the category of modules (the Hom is given by the metric D as already mentioned in Appendix A).

The original category defines the syntax and the presheaf category can be considered to reflect semantics (see also [1]).

In fact just like $\mathbb{C}[x, y]$ gives coordinates on the space of ideals $\mathbb{C}^2$, we could think that the language category (the syntax category) provides coordinates on the category of presheaves (modules) which can be though as the semantic category (in this particular case, for example because the Hom which is the metric D measures semantic similarity).

Now since we can translate between languages, namely the semantics of languages are in some sense the same (approximately) we expect that the categories of presheaves on different language categories, should be equivalent. This is a well known notion called *Morita equivalence*. We would then expect that enriched categories corresponding to different languages should be Morita equivalent. Moreover in that case there are associated invariants (Hochschild homology) which should be semantic invariants.

Investigating and developing this, is a future direction of research.

References

[1] Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. An enriched category theory of language: From syntax to semantics. *La Matematica*, 1:551–580, 2022. arXiv:2106.07890.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[3] G. Cohen, S. Gaubert, and J. P. Quadrat. Duality and separation theorems in idempotent semimodules. *Linear Algebra and Appl.*, 379:395–422, 2004.

[4] M. Develin and B. Sturmfels. Tropical convexity. *Documenta Mathematica*, 9:205–206, 2004.

[5] Andreas W.M Dress. Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces. *Advances in Mathematics*, 53(3):321–402, September 1984.

[6] Komei Fukuda and Alain Prodon. Double description method revisited. In *Selected papers from the 8th Franco-Japanese and 4th Franco-Chinese Conference on Combinatorics and Computer Science*, pages 91–111, London, UK, 1996. Springer-Verlag.

[7] Hiroshi Hirai and Shungo Koichi. On tight spans for directed distances. *Annals of Combinatorics*, 16(3):543–569, May 2012.

[8] J. R. Isbell. Six theorems about injective metric spaces. *Commentarii Mathematici Helvetici*, 39(1):65–76, December 1964.

[9] Michael Joswig and Katja Kulas. Tropical and ordinary convexity combined. *Adv. Geom.*, 10(2):333–352, March 2010.

[10] T. Lam and A. Postnikov. Alcoved polytopes. I. *Discrete Comput. Geom.*, 38(3):453–478, 2007.

[11] F. William Lawvere. Metric spaces, generalized logic, and closed categories. *Rendiconti del Seminario Matematico e Fisico di Milano*, 43(1):135–166, December 1973.

[12] Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. Meaning representations from trajectories in autoregressive models, 2023. arXiv:2310.18348.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. arXiv:1310.4546.

[14] Athanase Papadopoulos and Marc Troyanov. *From Funk to Hilbert geometry*, page 33–67. EMS Press, December 2014.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. https://api.semanticscholar.org/CorpusID:160025533.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Improving language understanding by genetrative pre-training, 2018. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[17] A. Schrijver. *Theory of Linear and Integer Programming*. Wiley, 1986.

[18] R. P. Stanley. Two poset polytopes. *Discrete Comput. Geom.*, 1:9–23, 1986.

[19] Ngoc Mai Tran. Enumerating polytropes. *Journal of Combinatorial Theory, Series A*, 151:1–22, October 2017.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. arXiv:1706.03762.

[21] Simon Willerton. Tight spans, Isbell completions and semi-tropical modules. *Theory and Applications of Categories*, 28(22):696–732, 2013.

SG: INRIA and CMAP, École polytechnique, IP Paris, CNRS
Email address: Stephane.Gaubert@inria.fr

YV: ATHENA Research Center, Institute for Language and Speech Processing, Athens, Greece and IHES, Bures-sur-Yvette, France
Email address: yvlassop@gmail.com, yvlassop@ihes.fr