Directed Metric Spaces, Alcoved Polytopes and LLMs

Yiannis Vlassopoulos

November 22, 2024

Some general observations, questions and goals

- **Question:** Is there a mathematical structure defined by probabilities of extension of texts and if yes, what is it?
- Note that **computing text continuations** is part of a well know structure in mathematics.
- Consider **language as a poset** $\mathcal{L}$ with the subtext order, either one sided or two sided. One sided eg: red $\leq$ red rose . Two sided eg: red $\leq$ pretty red rose .
- Extensions of a text are an **upper set**. So to "red" we associate the set of all texts containing "red". It is a well known mathematical object called a **filter** or a **co-presheaf** (in categorical language)

Some general observattons, questions and goals

- The dual is to consider all texts that are subtexts of a given one. This is called the **lower set** or the **ideal** generated by it or a **presheaf** (in categorical language).
- The set of lower sets (a subset S of $\mathcal{L}$ is a lower set if it is such that, if $x \in S$ and $y \leq x$ then $y \in S$) or the set of upper sets (filters) are **lattice completions** of $\mathcal{L}$.
- In fact an ideal or a filter is determined by a collection of non comparable elements of $\mathcal{L}$, its generators.

Some general observations, questions and goals

- The same if we consider **language as a monoid** by juxtaposing texts. Associate to “red” the (one sided or two sided ideal) of all texts containing “red”.
- **Distributional semantics** says that the ideals or the filters of the language poset or the language monoid encode something about the meaning.

Some general observations, questions and goals

- **The space of ideals in commutative algebra is part of a very general duality between spaces and algebras.** To a space associate the algebra of functions on it and to a commutative algebra the space of (prime) ideals, called the spec of the algebra.
- **Questions:** Can we generalize and extend the constructions and result for posets and monoids to the case of text extensions with probabilities which LLM compute? What mathematical structure do these define? Is it something analogous to the set of ideals (co presheaves) or filters (presheaves)? What can we learn from this structure ?
- This work is about answering these questions.

References

- For current talk: joint paper with Stephane Gaubert: *"Directed metric structures arising in Large Language Models "* arXiv 2405.12264
- Previous work: *"An enriched category theory of language: from syntax to semantics"*, T. D. Bradley, J. Terilla and Y. Vlassopoulos, 2021, arXiv 2106.07890 (published in *La Matematica*)

Related Experiments

"Meaning Representations from trajectories in autoregressive models"

arXiv 2310.18348, Stefano Soatto et al.

Contents

- From probabilities of texts extensions to distances.
- Isometric embedding of the text metric space $\mathcal{L}$ to the metric polyhedron $P(\mathcal{L})$ (Yoneda embedding).
- Texts in $\mathcal{L}$ are mapped to special extremal rays
- Description of all extremal rays.
- Description of $P(\mathcal{L})$ as a $(\min, +)$ linear space.
- $(\min, +)$ system of equations satisfied by text vectors (generators of text extremal rays).

Contents

- $P(\mathcal{L})$ as a representation of a monoid algebra.
- Duality of polyhedron $P(\mathcal{L})$ generated by text extensions and polyhedron $\widehat{P}(\mathcal{L})$ generated by restrictions.
- Expression of text vectors in terms of word vectors
- Relation with Isbell completion (generalizing Dedekind Mac Neille completion of posets).
- Directions for future research and some speculations (Morita equivalence).

Probabilistic Language model (a.k.a Syntactic category)

Definition 1

A probabilistic language model is a triple $(\mathcal{L}, \leq, Pr)$ where,

$\mathcal{L} := \{a_0, a_1, \ldots, a_n\}$ is a collection of texts, $\leq$ is the subtext order and

$Pr : \mathcal{L} \times \mathcal{L} \to [0, 1]$ is a function such that

$a_i \leq a_j \leq a_k \implies \Pr(a_k|a_i) = \Pr(a_k|a_j)\Pr(a_j|a_i)$.

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

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

PLM is a special case of a directed metric space

Definition 3

Given the probabilistic language model $(\mathcal{L}, \leq, Pr)$ where $\leq$ 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

$$d(a_i, a_j) = \begin{cases} -\log \Pr(a_j|a_i) & \text{if } a_i \leq a_j, \\ \infty & \text{if } a_i \text{ and } a_j \text{ are not comparable.} \end{cases} \tag{1}$$

The map d satisfies the triangle inequality:

$d(a_i, a_k) \leq d(a_i, a_j) + d(a_j, a_k)$ and the equality holds if and only if $a_i \leq a_j \leq a_k$ or $a_i \not\leq a_k$.

Poset structure, categorical interpretation

- **The metric determines the poset** since we have $a_i \leq a_j \leq a_k \iff d(a_i, a_j) + d(a_j, a_k) = d(a_i, a_k)$ and $d(a_i, a_k) < \infty$
- Categorically, (X, d) directed metric space means (X, d) is a **category enriched over the monoidal closed category** $(-\infty, \infty]$ considered as poset (with the opposite of the usual order) and with monoidal structure given by addition. Indeed :

$$Hom(a_i, a_j) \otimes Hom(a_j, a_k) \to Hom(a_i, a_k) \iff$$

$$d(a_i, a_j) + d(a_j, a_k) \geq d(a_i, a_k).$$

The metric polyhedron $P(\mathcal{L})$

- We equip $\{\mathbb{R} \cup \{\infty\}\}^n \backslash \{(\infty, \dots, \infty)\}$ with the *Funk* directed metric D defined by $D(x, y) := \max_i \{y_i - x_i \mid x_i \neq \infty\}$.

Definition 4

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

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

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

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

.

- $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$ are **alcoved polytopes** defined by the root system A_n since they are given by $x \cdot (e_i - e_j) \leq d_{i,j}$ and $y \cdot (e_i - e_j) \leq d_{j,i}$.

Geometric/Categorical description of $P(\mathcal{L})$

- Equip $(-\infty, \infty]$ with the directed metric

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

- Geometrically $P(\mathcal{L})$ is a directed metric space whose points are **non-expansive maps**.
- $P(\mathcal{L}) = \{x : (\mathcal{L}, d^t) \to ((-\infty,\infty], d_{\mathbb{R}}) | d_{\mathbb{R}}(x(a_j), x(a_i)) \leq d^t(a_j, a_i)\}$
- $\widehat{P}(\mathcal{L}) = \{y : (\mathcal{L}, d) \to ((-\infty,\infty], d_{\mathbb{R}}) | d_{\mathbb{R}}(y(a_j), y(a_i)) \leq d(a_j, a_i)\}$.
- Categorically $P(\mathcal{L})$ is the category of **presheaves** and $\widehat{P}(\mathcal{L})$ is the category of **co-presheaves**.

The Yoneda isometric embedding $Y : (\mathcal{L}, d) \hookrightarrow (P(\mathcal{L}), D)$

- The map

  $Y : (\mathcal{L}, d) \hookrightarrow (P(\mathcal{L}), D)$ given by $Y(a_k) := d(-, a_k) : \mathcal{L} \to \mathbb{R}$ is called the **Yoneda embedding** and is an **isometric embedding**, namely $D(Y(a_i), Y(a_j)) = d(a_i, a_j)$

- The map

  $\widehat{Y} : (\mathcal{L}, d) \hookrightarrow (\widehat{P}(\mathcal{L}), D^t)$ given by $\widehat{Y}(a_k) := d(a_k, -) : \mathcal{L} \to \mathbb{R}$ is is called the **co-Yoneda embedding**. and is an isometric embedding, namely $D(\widehat{Y}(a_j), \widehat{Y}(a_i)) = d(a_i, a_j)$.

- The Funk metric D is the **Hom on the category of presheaves**.

Yoneda Lemma and $P(\mathcal{L})$ as a Metric span.

- If $x \in P(\mathcal{L})$ then

$$x_i = D(d(-, a_i), x) = D(Y(a_i), x).$$

- The defining inequalities, $x_i \leq x_j + d_{i,j}$ of $P(\mathcal{L})$ become

$$D(Y(a_i), x) \leq D(Y(a_i), Y(a_j)) + D(Y(a_j), x).$$

Namely they are **triangle inequalities** for maps

$x : \mathcal{L} \to ((-\infty, \infty], d_{\mathbb{R}})$ and for the maps

$Y(a_k) = d(-, a_k) : \mathcal{L} \to ((-\infty, \infty], d_{\mathbb{R}})$.

co-Yoneda Lemma and $\widehat{P}(\mathcal{L})$ as a Metric span.

- If $y \in \widehat{P}(\mathcal{L})$ then

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

- The defining inequalities, $y_i \leq y_j + d_{j,i}$ of $\widehat{P}(\mathcal{L})$ become

$$D(y, \widehat{Y}(a_i)) \leq D(y, \widehat{Y}(a_j)) + D(\widehat{Y}(a_j), \widehat{Y}(a_i))$$

namely the triangle inequalities for maps $y : \mathcal{L} \to ((-\infty, \infty], d_{\mathbb{R}})$.

$Q(\mathcal{L})$: The multiplicative versions of $P(\mathcal{L})$

- To further understand the polyhedron $P(\mathcal{L})$ we consider the **change of variables** $z_i := e^{-x_i}$ and introduce the following:
- Let $Q(\mathcal{L})$ be the **polyhedral cone**
  $Q(\mathcal{L}) := \{z = (z_1, \ldots z_n) \in [0, \infty)^n \backslash \{(0, \ldots, 0)\} | z_i \geq \Pr(a_j|a_i) z_j\}$.
- We see that
  $Q(\mathcal{L}) = \{z := (z_1, \ldots z_n) \in [0, \infty)^n | z_i := e^{-x_i}$ for $x = (x_1, \ldots, x_n) \in P(\mathcal{L})\}$
- For $Pr(a_j|a_l)$ taking values only 0 or 1 it is a **cone version of Stanley's order polytope**.

The Probabilistic language model as enriched category

- The Probabilistic language model $\mathcal{L}$ is a category enriched over the monoidal category $[0, \infty)$ considered as a poset with the usual order and monoidal structure given by multiplication. Indeed put $\mathcal{L}(a_i, a_j) := Pr(a_j|a_i)$ then

$$\mathcal{L}(a_i, a_k) \geq \mathcal{L}(a_i, a_j)\mathcal{L}(a_j, a_k)$$

  with equality if $a_i \leq a_j \leq a_K$.

- $Q(\mathcal{L})$ is the category of presheaves on $\mathcal{L}$ and $\widehat{Q}(\mathcal{L})$ is the category of copresheaves.

The metric D_Q on $Q(\mathcal{L})$

- 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

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

- By definition we have

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

$\widehat{Q}(\mathcal{L})$: The multiplicative version of $\widehat{P}(\mathcal{L})$

- Moreover let $\widehat{Q}(\mathcal{L})$ be the polyhedral cone $\widehat{Q}(\mathcal{L}) := \{u = (u_1, \dots u_n) \in [0, \infty)^n \backslash \{(0, \dots, 0)\} | u_i \geq \Pr(a_i|a_j)u_j\}$
- $\widehat{Q}(\mathcal{L}) = \{u := (u_1, \dots u_n) \in [0, \infty)^n | u_i := e^{-y_i}$ for $y = (y_1, \dots, y_n) \in \widehat{P}(\mathcal{L})\}$
- Clearly the transpose D_Q^t defines a directed metric on $\widehat{Q}(\mathcal{L})$.
- We have **isometric embeddings**

$$e^{-Y} : \mathcal{L} \to Q(L) \text{ and } e^{-\widehat{Y}} : \mathcal{L} \to \widehat{Q}(L)$$

Extremal Rays of $P(\mathcal{L})$ and $Q(\mathcal{L})$

- 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.
- A vector in a polyhedral cone in $\mathbb{R}^n$ generates an extremal ray if and only if it **satisfies** $n-1$ **linearly independent conditions**.
- 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 $\widehat{Q}(\mathcal{L})$).

Texts define special Extremal Rays

Theorem 5

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 $Y(a_k) = d(-, a_k)$ generates 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 $\widehat{Y}(a_k) = d(a_k, -)$ generates an extremal ray in $\widehat{P}(\mathcal{L})$.

- The reason is that, if $a_i \leq a_j \leq a_k$ then $d(a_i, a_k) = d(a_i, a_j) + d(a_j, a_k)$ i.e. $Y(a_k)_i = d_{i,j} + Y(a_k)_j$.

Extremal Rays of $P(\mathcal{L})$

- Let $\tilde{Q}(\mathcal{L}) := \{\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n) \in [0, \infty)^n \backslash \{(0, \ldots, 0)\} | \tilde{y}_i \geq \tilde{y}_j$ whenever $a_i \leq a_j\}$.
- Let $(\mathcal{L}, \leq, \mathrm{Pr})$ be a probabilistic language model then there is a **diagonal change of variables mapping** $Q(\mathcal{L})$ **to** $\widetilde{Q}(\mathcal{L})$.
- Easy case if the empty text a_0 is included. Then if $a_0 \leq a_i \leq a_j$, we have $Pr(a_j|a_0)) = Pr(a_i|a_0)Pr(a_j|a_i)$. Then $y_i \geq P(a_j|a_i)y_j$ becomes $y_i \geq \frac{P(a_j|a_0)}{Pr(a_i|a_0)}y_j$. Setting $\tilde{y}_i := Pr(a_i|a_0)y_i$ we get $\tilde{y}_i \geq \tilde{y}_j$. **However can also prove without assuming** a_0.

Extremal Rays of $P(\mathcal{L})$ correspond to connected lower sets of $\mathcal{L}$

Theorem 6

The vector $\tilde{y} := (\tilde{y}_1, \ldots \tilde{y}_n) \in \tilde{Q}(\mathcal{L})$ 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.

- Therefore **extremal rays of $Q(\mathcal{L})$ correspond to connected lower sets of $\mathcal{L}$ and the ones in the image of the Yoneda embedding correspond to principle lower sets**.

Extremal Rays of $P(\mathcal{L})$ correspond to collections of texts in $\mathcal{L}$.

- Note that a connected lower set is generated by its maximal elements.
- Analogously extremal rays of $\widetilde{P}(\mathcal{L})$ correspond to connected upper sets of $\mathcal{L}$.

Directed metric d is $(\min, +)$ idempotent

- **Recall the** $(\min, +)$ **semifield**: On $(-\infty, \infty]$ consider operations $s \oplus t := \min\{s, t\}$ and $\lambda \odot s := \lambda + s$. Think of it as log algebra.
- Important identity using T **temperature**:

$$\lim_{T \to 0} -Tlog(e^{-\frac{s}{T}} + e^{-\frac{t}{T}}) = \min\{s, t\}$$

- $(\mathcal{L}, d)$ directed metric space means $d_{i,k} = \min_j\{d_{i,j} + d_{j,k}\}$. Define $d_{\min} : \mathbb{R}^n \to \mathbb{R}^n$ by $d_{\min}(x)_i := \min_j\{d_{i,j} + x_j\}$. We have $d_{\min}^2 = d_{\min}$ so $d_{\min}$ **is a projection**.

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

- Let $\mathrm{Fix}(d_{\min}) := \{x : d_{\min}(x) = x\}$.

  We have

  $$P(\mathcal{L}) = Fix(d_{\min}) = Im(d_{\min})$$

  i.e $P(\mathcal{L})$ is the $(\min, +)$ column span of d.

- Proof: But $d_{\min}x = x \iff x_i = \min_j\{d_{i,j} + x_j\} \iff x_i \leq x_j + d_{i,j} \iff x \in P(\mathcal{L})$.

  $d_{\min}^2 = d_{\min} \iff \mathrm{Im}(d_{\min}) = \mathrm{Fix}(d_{\min})$.

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

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

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

  This is a $(\min, +)$ linear system of equations defining $P(\mathcal{L})$.

- $\widehat{P}(\mathcal{L}) = Im(d^t)$, is the $(\min, +)$ span of the rows of d.

Categorical interpretation

- The categorical interpretation of the fact that $P(\mathcal{L})$ is the $(\min, +)$ column span of d is that **any presheaf can be expressed as a weigted colimit of representable presheaves**, namely the Yoneda images $Y(a_k) := d(-, a_k)$.
- Analogously the fact that $\widehat{P}(\mathcal{L})$ is the $(\min, +)$ row span of d means that **any co-presheaf can be expressed as a weighted colimit of representable co-presheaves,** namely the Yoneda images $Y(a_k) := d(a_k, -)$.
- We will see that not every presheaf can be expressed as a weighted limit of representables. .

Equations for $Y(a_k)$, $\widehat{Y}(a_k)$.

- Since $d(a_i, a_k) = \min_j\{d(a_i, a_j) + d(a_j, a_k)\}$ we have
- $d(-, a_k) = \min_j\{d(-, a_j) + d(a_j, a_k)\}$ namely

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

- $d(a_i, -) = \min_j\{d(a_i, a_j) + d(a_j, -)\}$ namely

$$\widehat{Y}(a_i) = \oplus_{a_i \leq a_j} d_{i,j} \odot \widehat{Y}(a_j)$$

- Can consider that the neural net is finding a solution to these systems of equations.

D as tropical inner product.

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

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

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

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

$\widehat{Q}(\mathcal{L})$ as Monoid algebra representation

- We now make the assumption that a_0, the empty text, is in $\mathcal{L} = \mathcal{A}^*$, the free monoid. Then if $a_i \leq a_j$ we have $a_0 \leq a_i \leq a_j$ and therefore $P(a_j|a_0) = Pr(a_i|a_0)Pr(a_j|a_i)$
- For $a_i \in \mathcal{L}$, let $\widehat{\mathbf{Y}}(a_i) : \mathcal{L} \to [0, \infty)$ be the Yoneda embedding of a_i, namely

$$\widehat{\mathbf{Y}}(a_i) := \mathcal{L}(a_i, -) = e^{\widehat{Y}(a_i)} = Pr(-|a_i).$$

  The $(\max, \cdot)$ span of $\widehat{\mathbf{Y}}(a_i)$ is the polyhedral cone $\widehat{Q}(\mathcal{L})$.
- Consider the function $L : \mathcal{A}^* \to [0, 1]$, defined by

$$L(x) := Pr(x|a_0) = \mathcal{L}(a_0, x) = \widehat{\mathbf{Y}}(a_0)(x)$$

$\widehat{Q}(\mathcal{L})$ as Monoid algebra representation

- Denote by S the semiring $([0,\infty), \max, \cdot)$ and by $S[\mathcal{A}^*]$ the monoid algebra generated by the free monoid $\mathcal{A}^*$ over the semiring S.
- Recall that an element of $S[\mathcal{A}^*]$ can be considered equivalently as a formal sum of elements in $\mathcal{A}^*$ or as a function $F : \mathcal{A}^* \to [0,\infty)$.
- Indeed given F we can construct the formal sum $f := \sum_{a_i} F(a_i)a_i$.

$\widehat{Q}(\mathcal{L})$ as Monoid algebra representation

- Note that the function $L : \mathcal{A}^* \to [0, \infty)$ defines an element of $S[\mathcal{A}^*]$.
- Consider the left regular representation of $\mathcal{A}^*$. Namely denote $a_i L$ the action of $a_i \in \mathcal{A}^*$ on L given by

$$a_i L(x) := L(a_i x).$$

$\widehat{Q}(\mathcal{L})$ as Monoid algebra representation

- We see that

$$a_i L(x) = L(a_i x) = \mathcal{L}(a_0, a_i x) = P(a_i x|a_0) = Pr(a_i|a_0)Pr(a_i x|a_i) =$$

$$= Pr(a_i|a_0)\widehat{\mathbf{Y}}(a_i)(a_i x).$$

- Recall that $\widehat{Q}(\mathcal{L})$ is the $(\max, \cdot)$ span of the representable copresheaves $\widehat{\mathbf{Y}}(a_i)$. Therefore **the orbit of L under the left regular representation of $S[\mathcal{A}^*]$ generates** $\widehat{Q}(\mathcal{L})$ over S. This means that the category of copresheaves $\hat{Q}(\mathcal{L})$ is a representation of $S[\mathcal{A}^*]$.

Duality between text extensions $\widehat{P}(\mathcal{L})$ and text restrictions $P(\mathcal{L})$

- Easy way to see how they are related:
  $x_i \leq d_{i,j} + x_j \iff -x_j \leq d_{i,j} + (-x_i)$.
- Namely
  $$dx = x \iff d^t(-x) = -x.$$
- To use this for our duality we need to use the completed $(\min, +)$ semiring $[-\infty, \infty]$ where $+\infty$ is absorbing element so, $-\infty + (+\infty) = +\infty$.

Adjunction between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$

- There are two inverse maps

  $B : P(\mathcal{L}) = \mathrm{Im}(d_{\min}) \to \mathrm{Im}(d^t_{\min}) = \widehat{P}(\mathcal{L})$ given by $B(x) := -x$ and
- $A : \widehat{P}(\mathcal{L}) = \mathrm{Im}(d^t_{\min}) \to \mathrm{Im}(d_{\min}) = P(\mathcal{L})$ given by $A(y) := -y$.
- A **and** B **form an adjunction:** $D(Ax, y) = D^t(x, By)$, namely

$$D(-x, y) = D(-y, x)$$

Equivalence between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$

- A and B are **isometries**, namely

$$D(-x,-y) = D^t(x,y)$$

- Moreover they are **anti-linear**

  $A(\lambda \odot x) = -\lambda \odot A(x)$ and

  $A(x \oplus_{\max} y) = A(x) \oplus_{\min} A(y)$,

  $A(x \oplus_{\min} y) = A(x) \oplus_{\max} A(y)$ and similarly for B.

- Note that the map $e^{-B} : Q(\mathcal{L}) \to \widehat{Q}(\mathcal{L})$ is $z_i \to u_i := \frac{1}{z_i}$.

Duality in coordinates

- $$d(-, a_k) = \oplus_{a_j \leq a_k} d(a_j, a_k) \odot d(-, a_j)$$

  $$B(d(-, a_k)) = -d(-, a_k) = \oplus_{a_j \leq a_k} - d(a_j, a_k) \odot d(a_j, -)$$

- $$d(a_k, -) = \oplus_{a_k \leq a_i} d(a_k, a_i) \odot d(a_i, -)$$

  $$A(d(a_k)- = -d(a_k, -) = \oplus_{a_k \leq a_i} - d(a_k, a_i) \odot d(-, a_i).$$

Example: "red colour"

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

Example: "red colour"

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

Example: "red colour"

- We consider the corpus to be $\mathcal{L} := \{\text{red, colour, red colour}\}$. Denote lower set generated by "a" by $(a)_l$ and the upper set by $(a)_u$.
- **Extremal rays of $Q(\mathcal{L})$ correspond to connected lower sets of $\mathcal{L}$.** There are three and they are all principle: $(r)_l = \{r\}, (c)_l = \{c\}, (rc)_l = \{r, c, rc\}$. (Note that $(r, c)_l$ is not connected so it does not correspond to an extremal ray of $Q(\mathcal{L})$.)
- **Extremal rays of $\widehat{Q}(\mathcal{L})$ correspond to connected upper sets of $\mathcal{L}$.** The principle ones are $(r)_u = \{r, rc\}, (c)_u = \{c, rc\}, (rc)_u = \{rc\}$ and a non-principle one $(r, c)_u = \{r, c, rc\}$. **This extremal ray is not in the image of the Yoneda embedding.**

Example: “red colour”, figures

- Let Δ be the unit simplex. We have polyhedra $Q_0(\mathcal{L}) := Q(\mathcal{L}) \cap \Delta$ and $\widehat{Q}_0(\mathcal{L}) := \widehat{Q}(\mathcal{L}) \cap \Delta$.
- Extremal rays of $Q(\mathcal{L})$ define vertices of $Q_0(\mathcal{L})$.

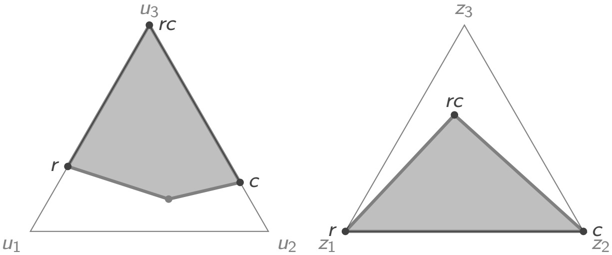

Figure: The cross section $\widehat{Q}_0(\mathcal{L})$ of the polyhedral cone $\hat{Q}(\mathcal{L})$ arising from the metric of d (left). Every vector $d(r,-), d(c,-), d(rc,-)$ determines an extreme point of the cross section, denoted by r, c, or rc. There is a fourth extreme point (shown in gray) corresponding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

Compatibility of $P(\mathcal{L})$ with extending $\mathcal{L}$

Theorem 7

If a probabilistic language model $(\mathcal{L}_1, d_1)$ *is extended to* $(\mathcal{L}_2, d_2)$ *by 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* $P(\mathcal{L}_2)$*.*

- If $\mathcal{L}_1 := \{a_1 \dots a_n\}$ and $\mathcal{L}_2 := \{b_1 \dots b_n, b_{n+1}, \dots b_{n+k}\}$, where $b_j = \phi(a_j)$ for $j = 1 \dots n$ then **the retraction is**

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

Text vectors in terms of word vectors

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. 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

Let $Y(b_k) \in P(\mathcal{L})$ be an extremal ray corresponding to a text $b_k \in \mathcal{L}$ then

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

Text vectors in terms of word vectors

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. Let* $Y : \mathcal{L} \to P(\mathcal{L})$ *be the Yoneda embedding. Let* $T \geq 0$ *be a parameter which will be called temperature, then we have*

$$\mathcal{R}(Y(b_k)) = \lim_{T \to 0} -T \log\left(\sum_{w_i \leq b_k} e^{-\frac{d(w_i, b_k)}{T}} e^{-\frac{Y(w_i)}{T}} \right) \tag{4}$$

Therefore for small T *we have*

$$e^{-\frac{\mathcal{R}(Y(b_k))}{T}} \approx \sum_{w_i \leq b_k} e^{-\frac{d(w_i, b_k)}{T}} e^{-\frac{Y(w_i)}{T}} \tag{5}$$

Isbell completion or directed tight span

- Consider $d_{\max}(x)_i := \max_j\{d_{i,j} + x_j\}$.
- The $(\max, +)$ span $I(\mathcal{L}) := Im(d_{\max})$ is called the Isbell completion.
- There is an adjunction $L(x) := d_{\max}(-x)$, $R(y) := d^t_{\max}(-y)$. The fixed part of the adjunction gives isomorphisms between $I(\mathcal{L}) := Im(d_{\max})$ and $\widehat{I}(\mathcal{L}) := Im(d^t_{\max})$.
- We have that $P(\mathcal{L})$ is the lattice closure of $I(\mathcal{L})$

Categorical and geometric meaning of $I(\mathcal{L})$

- $I(\mathcal{L})$ is not convex. It is a polyhedral cell complex.
- Categorically it is the set of presheaves that can be expressed as weighted limits of representables.
- If we take $[0, \infty]$ values then we get the directed tight span of Hirai and Koichi (Willerton). (Generalizes Dedekind-Mac Neille completion of poset.)

Overall picture

- **Isometrically** embed the finite, discrete, directed metric space $(\mathcal{L}, d)$ in to the continuous directed metric space $P(\mathcal{L})$. (Note that we cannot do that in general if we want to use Euclidean metric on $\mathbb{R}^n$ but here we can because we use D, the sup norm.)
- **Texts correspond to special extremal rays** of the polyhedron $P(\mathcal{L})$ (or the polyhedral cone $Q(\mathcal{L})$) which $(\min, +)$ span $P(\mathcal{L})$.
- $P(\mathcal{L})$ is **convex** and therefore easy to learn.

Overall picture

- Projection of text vectors (extremal rays in the image of the Yoneda embedding) onto the word space is a **Boltzmann weighted linear combination of word vectors**, analogously to the expression a text value vector in the attention layer.
- The transformer neural net from this point of view looks like it's learning the projection $R : P(\mathcal{L}) \to P(W)$ where P(W) is the space spanned by word vectors.
- There is a duality between $P(\mathcal{L})$ and $\widehat{P}(\mathcal{L})$, namely between the **space of texts restrictions and that of text extension**.

Some questions and speculations for next steps

- Since **we can translate between languages**, if the spaces $P(\mathcal{L})$ (of presheaves -generalizing ideals) encode meaning then they should be isomorphic (in some appropriate sense) for different languages $\mathcal{L}_1$ and $\mathcal{L}_2$.

  In math the name for such an isomorphism is **Morita equivalence**.
- Question: What is the right notion of Morita equivalence for the structure we have here?
- There are also **Morita invariants** (called Hochschild cohomology) which should be invariants of meaning. (Candidate for that is **Magnitude homology**.)

Some questions and speculations for next steps

- Following the duality between functions (coordinates) on a space and recovering the space by considering the ideals of the (commutative) algebra of functions, we could think that texts are non-commutative coordinates on some space of meanings.
  In fact this cannot be a usual geometric space but it could be a more sophisticated mathematical object. For example, In the case of non-commutive algebras the role of space is played by the category of Modules (they are presheaves).