Algebraic Positional Encodings

Konstantinos Kogkalidis
Jean-Philippe Bernardy
Vikas Garg

NEURAL INFORMATION PROCESSING SYSTEMS

TL;DR

> *"syntax is an algebra,*
> *semantics is an algebra,*
> *and meaning is a homomorphism between them"*
>
> Montague's theory of meaning

We argue that:
- understanding and explicating the formation rules and rewrite properties of **positions** over different **ambient structures** (*syntax*)
- and finding appropriate structure-preserving **interpretations** (*meaning*)

is the only way to structure-faithful **positional encodings** (*semantics*).

We call these Algebraic Positional Encodings (APE). APE readily apply to:
- sequences
- trees
- grids
- ...

We show that **sequential APE theoretically subsume RoPE**. Beyond sequences, APE are a **theoretically disciplined and highly general extension of RoPE across multiple dimensions** (both metaphorical and literal).

Sequences

Let $\mathbb{P}$ be a *path* (*i.e.*, a relative offset) between two points in a sequence.

$\mathbb{P}$ admits a simple inductive definition:

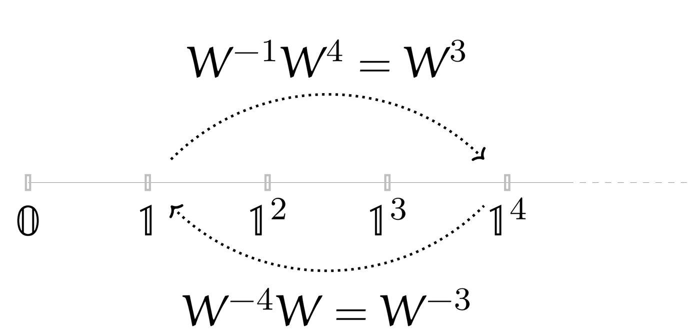
$$
\begin{aligned}
\mathbb{P} &:= \mathbb{1} && \text{\# take a step to the right}\\
&\mid \mathbb{P}+\mathbb{P} && \text{\# join two paths together}\\
&\mid \mathbb{P}^{-1} && \text{\# flip a path around}
\end{aligned}
$$

where $+$ associative and commutative with $\mathbb{0} := \mathbb{1}+\mathbb{1}^{-1}$ as its neutral element.

Remark 1. The signature coincides with that of the integers, $\mathbb{P} \equiv \mathbb{Z}$.
Remark 2. The signature corresponds to an infinite cyclic group, $\mathbb{P} \equiv \langle \mathbb{1} \rangle$.
Remark 3. The signature admits a representation in $O(n)$.
Consider the interpretation $\lceil\,\rceil : \langle \mathbb{1} \rangle \to \langle W \rangle$, such that:

$$
\begin{aligned}
\lceil \mathbb{1} \rceil &\mapsto W && \text{\# } W \text{ represents a single step}\\
\lceil p+q \rceil &\mapsto \lceil p \rceil \lceil q \rceil && \text{\# path composition} \rightsquigarrow \text{matrix multiplication}\\
\lceil p^{-1} \rceil &\mapsto \lceil p \rceil^{-1} && \text{\# path inversion} \rightsquigarrow \text{matrix transposition}
\end{aligned}
$$

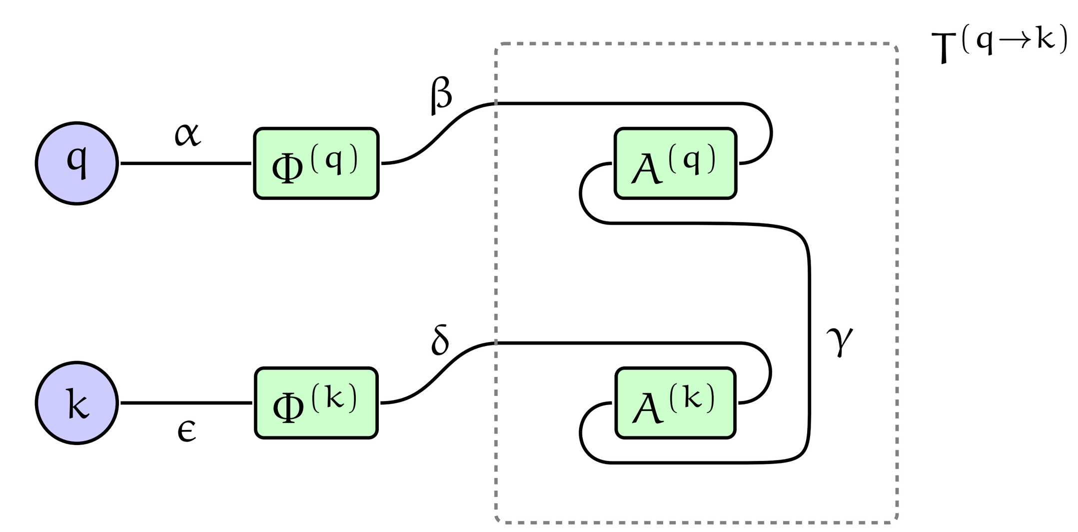
Remark 4. $A \to B = (A \to \mathbb{0}) + (\mathbb{0} \to B)$. Visually:



Remark 5. This setup offers an inductive parameterization of sequential PE using just **one trainable primitive** (a single matrix).

How-To

Simply substitute dot-product for the tensor contraction:



where:
- $q, k \in \mathbb{R}^n$
- $\Phi^{(q,k)} \in \mathbb{R}^{n\times n}$
- $A^{(q,k)} \in O(n)$ the representations of the positions of q and k

Note: $T^{(q\to k)} = A^{(q)\top} A^{(k)}$ the **path** representation from q to k

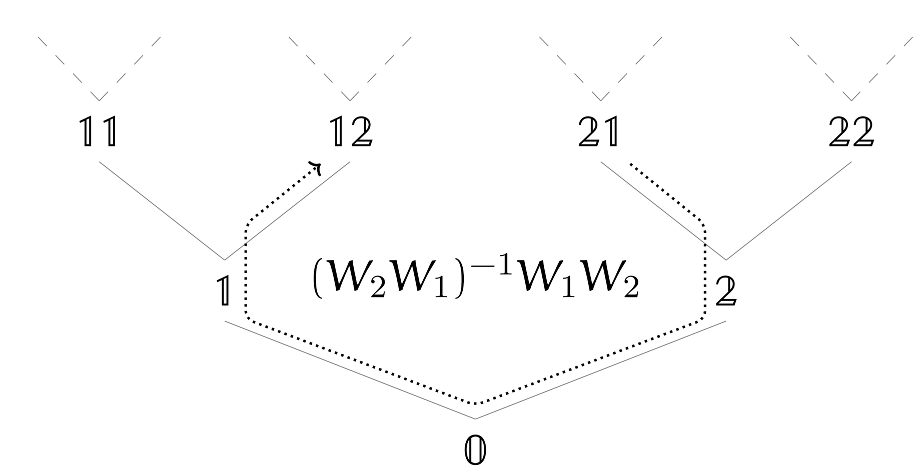
In the sequential setup RoPE $\equiv$ APE, except with a fixed W. Why?
Hint: $W = QRQ^\top$ (where $Q \in O(n)$ and R a block-diagonal rotation).

Trees

Extend the definition of $\mathbb{P}$ with **options**, to arrive at a definition of paths $\mathbb{P}_\kappa$ over κ-ary branching trees:

$$
\begin{aligned}
\mathbb{P}_\kappa &:= \mathbb{1} && \text{\# take the first branch}\\
&\mid \mathbb{2} && \text{\# take the second branch}\\
&\mid \ldots && \\
&\mid \kappa && \text{\# take the } \kappa\text{-th branch}\\
&\mid \mathbb{P}+\mathbb{P} && \text{\# join two paths together}\\
&\mid \mathbb{P}^{-1} && \text{\# flip a path around}
\end{aligned}
$$

Remark 5. This is now a generic group with κ generators.
Remark 6. Unlike sequences, the structure is not commutative.
Remark 7. All else remains the same – just extend the interpretation to: $\langle \mathbb{1}, \mathbb{2}, \ldots, \kappa \rangle \to \langle W_1, W_2, \ldots, W_\kappa \rangle$. Visually:



Grids

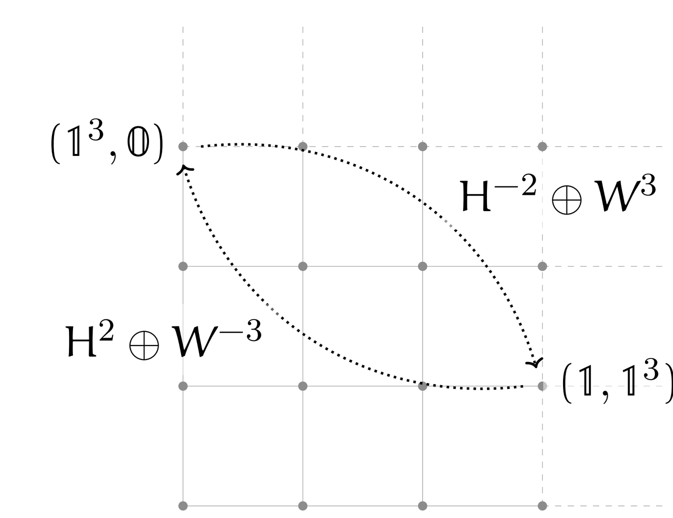
Rather than add options, we can glue two (or more) sequences together by means of the **group direct sum**, $\oplus$. Consider the composite group $\mathbb{P}^2 := \mathbb{P} \oplus \mathbb{P}$, with the group operation and inversion defined as:

$$
\begin{aligned}
(x, y) + (z, w) &= (x+z, y+w)\\
(x, y)^{-1} &= (x^{-1}, y^{-1})
\end{aligned}
$$

Remark 8. The structure is commutative once more.
Remark 9. Elements of $\mathbb{P}^2$ are still to be interpreted as (orthogonal) matrices, except now block-structured, by virtue of the **matrix direct sum**:

$$\lceil p \oplus q \rceil \mapsto \lceil p \rceil \oplus \lceil q \rceil = \begin{bmatrix} \lceil p \rceil & 0 \\ 0 & \lceil q \rceil \end{bmatrix}$$

Visually:



Remark 10. The same interpretation strategy can be applied to construct **any other composition** of established structures and their representations.

Results

We get really good results in many different setups (sequence transduction/tree manipulation/image recognition).

Details omitted for suspense (and space economy).

Learn More

- arxiv.org/abs/2312.16045
  prose, tables with numbers, references, etc.
- github.com/konstantinosKokos/APE
  reference implementation, experiment scripts, practical how-tos, etc.