Position Embedding

Karl Stratos

1 Sinusoidal Embedding

An **absolute position embedding** $E : \mathbb{N}_0 \rightarrow \mathbb{R}^d$ maps any position $t \geq 0$ to a vector $E(t) \in \mathbb{R}^d$. A "good" embedding should satisfy properties like

1. Well-behaved output values (e.g., have bounded variance)
2. $E(t) \neq E(t')$ for all $t \neq t'$
3. Holding k fixed, $E(t + k)$ has the same relationship with $E(t)$ for all t.

Property 3 is useful because if a model only uses pairwise positions, it allows for generalization to unseen positions.

1.1 Case $d = 2$

The transformer paper [2] proposed a **sinusoidal embedding** defined as

$$E(t) := \begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix} \tag{1}$$

See Appendix A for a review of trigonometry. We see that

1. The output values oscillate between -1 and 1 with frequency $\frac{1}{2\pi}$.
2. Sine and cosine have period 2π which is irrational, so $E(0), E(1), \ldots$ will never repeat.
3. Holding k fixed, $E(t + k) = A_k E(t)$ for all t where $A_k \in \mathbb{R}^{2\times 2}$ is some matrix (8).

1.2 Case $d = 2M$

Assuming M frequency multipliers $\sigma_1 \ldots \sigma_M$, define $E : \mathbb{N}_0 \rightarrow [-1, 1]^{2M}$ by $E(t) = (E_1(t) \ldots E_M(t))$ where

$$E_l(t) = \begin{bmatrix} \sin(\sigma_l t) \\ \cos(\sigma_l t) \end{bmatrix} \tag{2}$$

which is an independent sinusoidal embedding like (1) but with frequency $\frac{\sigma_l}{2\pi}$. Clearly $E(t) \in \mathbb{R}^d$ is still a "good" embedding, with $E(t + k) = B_k E(t)$ for some block diagonal matrix $B_k \in \mathbb{R}^{d\times d}$. The multipliers are typically chosen to decay geometrically as $\beta^{\frac{l-1}{M}}$ for some $0 < \beta < 1$, that is

$$\sigma_1 = 1 > \sigma_2 = \beta^{\frac{1}{M}} > \sigma_3 = \beta^{\frac{2}{M}} > \cdots > \sigma_M = \beta^{\frac{M-1}{M}} > \beta \tag{3}$$

(popularly $\beta = 0.0001$). Intuitively, the different frequencies can capture different types of relative distance (e.g., long vs short).

2 Rotary Position Embeddings (RoPE)

A **relative position embedding** $E : \mathbb{R}^d \times \mathbb{N}_0 \rightarrow \mathbb{R}^d$ maps vectors $q, k \in \mathbb{R}^d$ at positions $t, t' \geq 0$ to new vectors $E(q, t), E(k, t') \in \mathbb{R}^d$ satisfying

$$\langle E(q, t), E(k, t')\rangle = F(t' - t) \tag{4}$$

where $F : \mathbb{Z} \rightarrow \mathbb{R}$ is some function. The goal is to make pairwise inner products sensitive to relative distances.

2.1 Case $d = 2$

RoPE [1] proposes to rotate $q, k \in \mathbb{R}^2$ by $t, t' \in \mathbb{N}_0$ radians. This achieves (4) since the resulting inner product corresponds to a rotation by $t' - t$ radians. Mathematically, recall the Θ-radian counterclockwise rotation matrix

$$R_\Theta = \begin{bmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{bmatrix} \tag{5}$$

Define $E(u, t) := R_t u$. Since the rotation matrix is orthogonal,

$$\begin{aligned} \langle E(q,t), E(k,t') \rangle &= q^\top R_t^\top R_{t'} k \\ &= q^\top R_{t'-t} k \\ &= \langle q, k \rangle \cos(t' - t) + (q_2 k_1 - q_1 k_2) \sin(t' - t) \end{aligned}$$

which is a periodic function of $t' - t$. Note that the rotation cancels if $t = t'$.

2.1.1 Complex plane

By the usual correspondence between $\mathbb{R}^2$ and $\mathbb{C}$ (Appendix B), we can compute RoPE using complex numbers. Let $z = \mathcal{C}(q) \in \mathbb{C}$ and $z' = \mathcal{C}(k) \in \mathbb{C}$ denote the complex identities of $q, k \in \mathbb{R}^2$, specifically

$$\begin{aligned} z &= q_1 + q_2 i = ||q||\, e^{\theta i} && (\theta \text{ is the angle of } q \in \mathbb{R}^2) \\ z' &= k_1 + k_2 i = ||k||\, e^{\phi i} && (\phi \text{ is the angle of } k \in \mathbb{R}^2) \end{aligned}$$

Rotating by t radians in $\mathbb{R}^2$ is the same as multiplying with cis e^{ti} in $\mathbb{C}$, so we define:

$$E_{\mathbb{C}}(z, t) := z e^{ti}$$

We can recover the real-valued RoPE embedding and their inner product as

$$\begin{aligned} E(q,t) &= \mathcal{C}^{-1}(E_{\mathbb{C}}(z,t)) \\ E(k,t') &= \mathcal{C}^{-1}(E_{\mathbb{C}}(z',t')) \\ \langle E(q,t), E(k,t') \rangle &= \left\langle \mathcal{C}^{-1}(E_{\mathbb{C}}(z,t)), \mathcal{C}^{-1}(E_{\mathbb{C}}(z',t')) \right\rangle = \mathbf{Re}(\langle E_{\mathbb{C}}(z,t), E_{\mathbb{C}}(z',t') \rangle) \end{aligned}$$

where $\langle E_{\mathbb{C}}(z,t), E_{\mathbb{C}}(z',t') \rangle = \langle z, z' \rangle\, e^{(t'-t)i}$ (Lemma C.1). The following code verifies the claim numerically:

```python
from torch import manual_seed, randn, rand, pi, tensor, view_as_complex, view_as_real, polar, dot, conj

def rotate_R2(x, angle): return tensor([[angle.cos(), -angle.sin()], [angle.sin(), angle.cos()]]) @ x
def rotate_C1(z, angle): return z * polar(tensor(1.), angle)[0]
def inner(z1, z2): return z1 * conj(z2)  # == torch.vdot(tensor([z2]), tensor([z1]))

manual_seed(0)
q, k = randn(2), randn(2)  # [1.5410, -0.2934], [-2.1788,  0.5684]
zq, zk = view_as_complex(q), view_as_complex(k)  # 1.5410-0.2934j, -2.1788+0.5684j
t1, t2 = rand(1) * pi, rand(1) * pi  # [1.4314], [1.9864]
q_t1, k_t2 = rotate_R2(q, t1), rotate_R2(k, t2)  # [0.5047, 1.4853], [0.3597, -2.2228]
zq_t1, zk_t2 = rotate_C1(zq, t1), rotate_C1(zk, t2)  # 0.5047+1.4853j, 0.3597-2.2228j

print(dot(q_t1, k_t2))  # -3.1199
print(dot(view_as_real(zq_t1), view_as_real(zk_t2)))  # -3.1199
print(inner(zq_t1, zk_t2))  # -3.1199+1.6562j
print(rotate_C1(inner(zq, zk), t1 - t2))  # -3.1199+1.6562j
```

2.2 Case $d = 2M$

We can treat $q, k \in \mathbb{R}^{2M}$ at positions $t, t' \geq 0$ as M pairs of 2-dimensional vectors $q^{(l)}, k^{(l)} \in \mathbb{R}^2$ where $q_1^{(l)} = q_l$ and $q_2^{(l)} = q_{M+l}$ (similarly for $k^{(l)}$). Each pair is rotated independently by $\sigma_l t$ and $\sigma_l t'$ radians where σ_l is a "base degree". Intuitively, $\sigma_1 \ldots \sigma_M$ can capture different "speeds" at which the relative distance changes the inner product. RoPE uses the same geometric decay in the sinusoidal embedding (3), thus rotating $q^{(l)} \in \mathbb{R}^2$ by $(0.00001)^{\frac{l-1}{M}} t$ and $k^{(l)} \in \mathbb{R}^2$ by $(0.00001)^{\frac{l-1}{M}} t'$ radians (progressively smaller). Since

$$R_\Theta u = \begin{bmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} (\cos\Theta)u_1 - (\sin\Theta)u_2 \\ (\cos\Theta)u_2 + (\sin\Theta)u_1 \end{bmatrix} = \cos\Theta \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \sin\Theta \begin{bmatrix} -u_2 \\ u_1 \end{bmatrix}$$

with $\sigma = (1, (0.00001)^{\frac{1}{M}}, \ldots, (0.00001)^{\frac{M-1}{M}}) \in (0,1]^M$ we can compute the M rotations efficiently as

$$\begin{aligned} E(q,t) &= (\cos(t\sigma) \oplus \cos(t\sigma)) \odot q + (\sin(t\sigma) \oplus \sin(t\sigma)) \odot (-q[M:] \oplus q[:M]) \\ E(k,t') &= (\cos(t'\sigma) \oplus \cos(t'\sigma)) \odot k + (\sin(t'\sigma) \oplus \sin(t'\sigma)) \odot (-k[M:] \oplus k[:M]) \end{aligned} \tag{6}$$

where $\oplus, \odot$ are vector concatenation and elementwise multiplication. Even more simply, we can compute

$$\begin{aligned} E(q,t) &= \mathcal{C}^{-1}(\mathcal{C}(q) \odot \text{cis}(\sigma, t)) \\ E(k,t') &= \mathcal{C}^{-1}(\mathcal{C}(k) \odot \text{cis}(\sigma, t')) \end{aligned} \tag{7}$$

where $\text{cis}_l(\sigma, t) := e^{\sigma_l t i}$ and $\mathcal{C} : \mathbb{C}^M \to \mathbb{R}^{2M}$ is a bijection like (18). RoFormer (the RoPE paper), OpenLM, and OLMo use (6) whereas Llama uses (7).

RoPE is considered one of the few modifications of the original transformer architecture (e.g., along with the layer norm reordering and more nonlinear activation functions) whose improvement generalizes well across tasks.

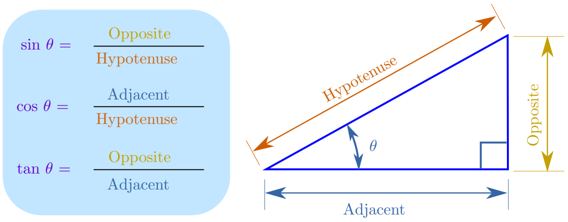
References

[1] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, **568**, 127063.

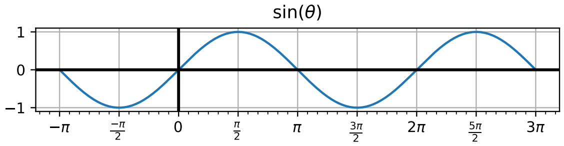
[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

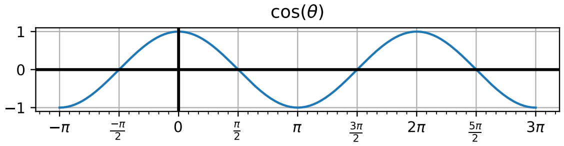
A Trigonometry

A function f is called periodic if it repeats after every P inputs. We call the smallest positive such P the **period** (or wavelength) of f and $\frac{1}{P}$ as the **frequency** of f. Trigonometric functions, mapping the angle of a right triangle θ (in radian) to a ratio of two sides, are naturally periodic. In particular, sine and cosine are bounded between -1 and 1, related as $\cos(\theta) = \sin(\frac{\pi}{2} - \theta)$, and have the period of 2π and frequency of $\frac{1}{2\pi}$.

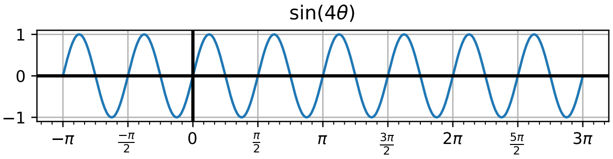

sin θ = Opposite / Hypotenuse
cos θ = Adjacent / Hypotenuse
tan θ = Opposite / Adjacent
Hypotenuse
Opposite
Adjacent
θ

(image from Math is Fun)





If $f(\theta)$ has frequency F, then $g(\theta) = f(\sigma\theta)$ has frequency σF, for instance:



$$\text{frequency} = 4 \times \frac{1}{2\pi} = \frac{2}{\pi}$$

Euler's formula allows us to write

$$\begin{bmatrix} \sin(\theta + \theta') \\ \cos(\theta + \theta') \end{bmatrix} = \begin{bmatrix} \cos(\theta') & \sin(\theta') \\ -\sin(\theta') & \cos(\theta') \end{bmatrix} \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \end{bmatrix} \tag{8}$$

B Complex Numbers

A complex number $z = x_1 + x_2 i \in \mathbb{C}$ is parameterized by $x \in \mathbb{R}^2$ where $i^2 = -1$. We can add/multiply z by another complex number $z' = y_1 + y_2 i$ as expected:

$$z + z' := x_1 + y_1 + (x_2 + y_2)i \tag{9}$$

$$zz' := x_1 y_1 - x_2 y_2 + (x_1 y_2 + x_2 y_1)i \tag{10}$$

Equipped with (9–10), $\mathbb{C}$ is a field and behaves like $\mathbb{R}$ (e.g., ensures an inverse). Let $\mathbf{Re}(a+bi) := a$ and $\mathbf{Im}(a+bi) := b$ denote extractors for the real and imaginary component.

B.1 Complex Plane

We have a bijection $\mathcal{C} : \mathbb{R}^2 \to \mathbb{C}$ where $\mathcal{C}((x_1, x_2)) = x_1 + x_2 i$ and $\mathcal{C}^{-1}(x_1 + x_2 i) = (x_1, x_2)$. This allows us to view a complex number as a point on a 2-dimensional plane called the **complex plane**. As usual, we can use either the rectangular/Cartesian or polar coordinate system to characterize this plane. In the rectangular system, $z = x_1 + x_2 i$ is identified by a pair of coordinates (x_1, x_2). In the polar system, it is identified by its length $||x|| = \sqrt{x_1^2 + x_2^2}$ and angle $\theta = \tan^{-1}(\frac{x_2}{x_1})$.[1] We can convert between the two forms by

$$\begin{aligned} z = \underbrace{x_1 + x_2 i}_{\text{rectangular form}} &= ||x|| \cos\theta + i\,||x|| \sin\theta && (\text{since } x_1 = ||x|| \cos\theta \text{ and } x_2 = ||x|| \sin\theta) \\ &= ||x|| (\cos\theta + i \sin\theta) \\ &= \underbrace{||x||\, e^{\theta i}}_{\text{polar form}} && (\text{Euler's formula: } e^{\theta i} = \cos\theta + i \sin\theta) \end{aligned}$$

[1] The inverse tangent needs to be more carefully defined to handle division by zero (e.g., atan2).

The polar form is more suited for multiplication. The product of $z = ||x||\, e^{\theta i}$ and $z' = ||y||\, e^{\phi i}$ is

$$zz' = ||x||\, ||y||\, e^{(\theta+\phi)i} \tag{11}$$

(11) and (10) have the same value, but (11) makes it clear that zz' is obtained by adding the angles and multiplying the lengths of z and z' in the polar system. In particular, we can just rotate $z \in \mathbb{C}$ by Θ radians by multipying with the complex exponential $e^{\Theta i}$ (sometimes called cis, "$\cos i \sin$"). For instance, multiplying by i corresponds to 90° counterclockwise rotation since $i = e^{\frac{\pi}{2} i}$.

At this point, it is useful to define the **complex conjugate** of $z = x_1 + x_2 i$ as

$$\bar{z} = x_1 - x_2 i$$

The conjugate operation has convenient properties (e.g., distributive, self-inverse). Geometrically in the complex plane, $\bar{z}$ is obtained by flipping z across the horizontal axis. Thus if $z = ||x||\, e^{\theta i}$ in polar form, then $\bar{z} = ||x||\, e^{-\theta i}$.

B.2 Complex Inner Product

The real inner product between $x, y \in \mathbb{R}^2$ measures directional alignment:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 = ||x||\, ||y|| \cos(\theta - \phi) \tag{12}$$

where θ, ϕ are their angles on the plane. When viewed as complex numbers $z = x_1 + x_2 i$ and $z' = y_1 + y_2 i$, their (1-dimensional) **complex inner product** $\langle \cdot, \cdot \rangle : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ is

$$\begin{aligned}
\langle z, z' \rangle &:= z\bar{z'} && \text{(definition)} && (13)\\
&= (x_1 y_1 + x_2 y_2) + (x_2 y_1 - x_1 y_2) i && \text{(expaned in rectangular form)} && (14)\\
&= ||x||\, ||y||\, e^{(\theta-\phi)i} && \text{(expanded in polar form)} && (15)
\end{aligned}$$

The use of complex conjugate has several consequences.

Consequence 1. It satisfies the three axioms of an inner product:

$$\begin{aligned}
\langle z, z' \rangle = z\bar{z'} = \overline{\bar{z}z'} &= \overline{\langle z', z \rangle} && \text{(conjugate symmetry)}\\
\langle \alpha z + \beta t, z' \rangle = (\alpha z + \beta t)\bar{z'} = \alpha z \bar{z'} + \beta t \bar{z'} &= \alpha \langle z, z' \rangle + \beta \langle t, z' \rangle && \text{(linearity in the first argument)}\\
\langle z, z \rangle = x_1^2 + x_2^2 &= ||x||^2 \geq 0 && \text{(positive semi-definite)}
\end{aligned}$$

In particular, we have a real-valued induced norm $||z|| := \sqrt{\langle z, z \rangle} \in \mathbb{R}$ within $\mathbb{C}$.

Consequence 2. It provides a connection between the real inner product between $x, y \in \mathbb{R}^2$ and the complex inner product between their complex identities $z, z' \in \mathbb{C}$ as follows:

$$\langle x, y \rangle = \mathbf{Re}(\langle z, z' \rangle) \tag{16}$$

which follows immediately from (14).

Consequence 3. It gives a geometric meaning to the inner product. Expanding (15) with Euler's formula we have

$$\langle z, z' \rangle = \underbrace{||x||\, ||y|| \cos(\theta - \phi)}_{\mathbf{Re}(\langle z, z' \rangle)} + \underbrace{||x||\, ||y|| \sin(\theta - \phi)}_{\mathbf{Im}(\langle z, z' \rangle)} i \tag{17}$$

where

- $\mathbf{Re}(\langle z, z' \rangle)$: directional alignment between x, y (i.e., $\langle x, y \rangle$)
- $\mathbf{Im}(\langle z, z' \rangle)$: signed area of parallelogram formed by x, y[2]

[2] Let L denote the shortest distance between x and the span of y. Since $\sin(\theta - \phi) = \frac{L}{||x||}$, $||x||\, ||y|| \sin(\theta - \phi) = L\, ||y||$.

B.2.1 Multi-dimensional complex inner product

The general d-dimensional complex inner product $\langle \cdot, \cdot \rangle : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$ is defined as

$$\langle z, z' \rangle := \sum_{j=1}^{d} z_j \bar{z}_j' = (z')^{\mathrm{H}} z$$

where $z_j, z_j' \in \mathbb{C}$ are the j-th element of $z, z' \in \mathbb{C}^d$. The second expression alternatively uses the conjugate transpose, treating z, z' as $d \times 1$ column vectors.[3] The axioms of an inner product are obviously preserved. The connection to the real counterpart is also preserved. Define a bijection $\mathcal{C} : \mathbb{R}^{2d} \to \mathbb{C}^d$ by

$$\mathcal{C}_j(x) = x_j + x_{d+j} i \tag{18}$$

(Other bijections, such as grouping by consecutive pairs, would equally work.) For $x, y \in \mathbb{R}^{2d}$ with angles θ, ϕ in the $2d$-dimensional vector space, if $z = \mathcal{C}(x), z' = \mathcal{C}(y) \in \mathbb{C}^d$ we have

$$\underbrace{\langle z, z' \rangle}_{\text{inner product in } \mathbb{C}^d} = \underbrace{\langle x, y \rangle}_{\text{inner product in } \mathbb{R}^{2d}} + \underbrace{||x||\,||y|| \sin(\theta - \phi)}_{\text{signed area of parallelogram}} \; i \tag{19}$$

C Lemmas

Lemma C.1. Let $q, k \in \mathbb{R}^2$ denote vectors with angles $\theta, \phi \in \mathbb{R}$ in radian. Let $z = \mathcal{C}(q) = q_1 + q_2 i \in \mathbb{C}$ and $z' = \mathcal{C}(k) = k_1 + k_2 i \in \mathbb{C}$ denote their complex identities. Pick any angles $t, t' \in \mathbb{R}$ and let $q(t) = R_t q \in \mathbb{R}^2$ and $k_t = R_{t'} k \in \mathbb{R}^2$ denote the counterclockwise rotations of k, q where $R_\Theta \in \mathbb{R}^2$ is the rotation matrix (5). Let $z(t) = ze^{ti} \in \mathbb{C}$ and $z'(t') = z'e^{t'i} \in \mathbb{C}$ denote the cis multiples of z, z'. Then

$$q(t) = \mathcal{C}^{-1}(z(t)) \tag{20}$$

$$k(t) = \mathcal{C}^{-1}(z'(t)) \tag{21}$$

$$\langle q(t), k(t') \rangle = \mathbf{Re}\left(\langle z(t), z'(t') \rangle \right) \tag{22}$$

$$\langle z(t), z'(t') \rangle = \langle z, z' \rangle \, e^{(t-t')i} \tag{23}$$

Proof. (20) and (21) are immediate from the equivalence of rotation and cis multiply (Appendix B.1). It follows that (22) holds by (16). We can check (23) as

$$\begin{aligned}
\langle z(t), z'(t') \rangle &= \left\langle ze^{ti}, z'e^{t'i} \right\rangle \\
&= \left\langle ||q||\, e^{(\theta+t)i}, ||k||\, e^{(\phi+t')i} \right\rangle \\
&= \left(||q||\, e^{(\theta+t)i} \right) \left(||k||\, e^{-(\phi+t')i} \right) \\
&= ||q||\, ||k||\, e^{(\theta-\phi)i} e^{(t-t')i} \\
&= \langle z, z' \rangle \, e^{(t-t')i}
\end{aligned}$$

□

[3] Note that the order has to be switched under the column assumption. Software implementations such as torch.vdot often use this as the complex "dot product" to make it compatible with the real dot product. Specifically, vdot implements $\mathrm{vdot}(u, v) = \langle v, u \rangle = u^{\mathrm{H}} v$ which corresponds in form to the real dot product $\mathrm{dot}(x, y) = x^\top y$ assuming $u, v \in \mathbb{C}^d$ and $x, y \in \mathbb{R}^d$ are column vectors.