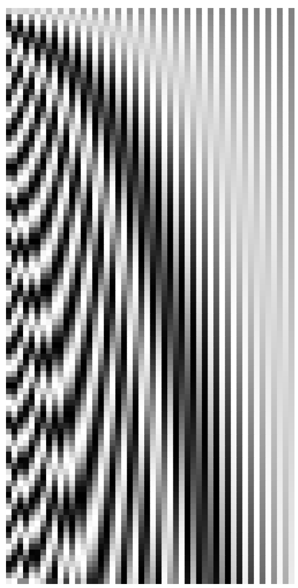
WEEKLY PARTICIPATION 12

Let T be the length of your input sequence, and d be the dimensionality of the positional encoding for each entry in your sequence.

Recall the positional encoding used in the original Transformer model: fix a frequency ω in $[0, 2\pi)$ and let $M \in \mathbb{R}^{T\times d}$ be given by

$$M_{p,i} = \begin{cases} \sin\left(p\,\omega^{i/d}\right) & \text{if } i \text{ is even} \\ \cos\left(p\,\omega^{(i-1)/d}\right) & \text{if } i \text{ is odd} \end{cases}.$$

Each row of M is a d-dimensional encoding of the corresponding position in the sequence.



Positional embedding for $\omega = \frac{1}{100}$, $T = 100$, and $d = 50$. White corresponds to -1, and black to 1.

For positional encodings to be useful, it has to be the case that there are no two distinct positions p_1 and p_2 in a sequence that have identical embeddings.

Prove that this is the case for this choice of positional embedding, regardless of T, the length of the sequence! You may assume that $d \geq 4$ for convenience[1].

[1] One approach: assume there are two positions $s \neq t$ that have the same encodings, and consider what it implies that $M_{s,2} = M_{t,2}$ *and* $M_{s,4} = M_{t,4}$ simultaneously.