CSCI 145 Problem 22: A Clock with Many Hands

July 25, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 22: A Clock with Many Hands

A clock with only a minute hand is precise about the minute and silent about the hour: at 12:30 and at 1:30 it looks *exactly* the same. In lecture, RoPE stamped position m onto a query by rotating each 2D slice of it by the angle $m\theta_j$, using the frequency ladder $\theta_j = 10000^{-2j/d}$ for $j = 0, \dots, d/2 - 1$. We proved the identity $\mathbf{R}_m^\top \mathbf{R}_n = \mathbf{R}_{n-m}$, so scores depend only on relative position. But we never asked why RoPE needs *many* frequencies rather than one good one. This problem answers that: each rotating slice is a hand of a clock, one hand aliases, and the ladder tells time at every scale. Throughout, $\mathbf{R}_m$ is the 2×2 rotation by the angle $m\theta$ from lecture, and $\mathbf{\Theta}_m \in \mathbb{R}^{d \times d}$ is the block-diagonal matrix whose jth block rotates by $m\theta_j$. This is the last question of the course's structure arc, which ran from circulant convolution matrices to Toeplitz attention scores.

Part A: One Hand Aliases

Fix a single frequency $\theta > 0$, so position m is represented by the rotation $\mathbf{R}_m$ alone. Show that $\mathbf{R}_{m+P} = \mathbf{R}_m$ for every position m exactly when $P\theta$ is an integer multiple of 2π, and conclude that the smallest positive such period is

$$P = \frac{2\pi}{\theta}.$$

Verify your period by substituting $m + 2\pi/\theta$ directly into the cosine and sine entries of $\mathbf{R}_m$.

Now put in numbers for the ladder's fastest hand, $\theta_0 = 1$: compute P, and use $7 \times 2\pi \approx 43.98$ to show that position 44 is rotated to within about one degree of position 0. Conclude in one sentence: with a single fast hand, two tokens tens of positions apart are nearly indistinguishable, so one frequency cannot label a long context.

Part B: The Ladder Tells Time at Every Scale

Take lecture's typical head size $d = 64$, so the hands are $j = 0, \dots, 31$ with $\theta_j = 10000^{-j/32}$.

First, the two ends of the ladder. Compute the slowest frequency θ_{31} and its period $2\pi/\theta_{31}$, and check that this period is longer than any context you have ever pasted into a chatbot (it is roughly 47,000 positions).

Next, *nearby* positions: between positions m and $m + 1$, by what angle does the fastest hand advance, and by what angle does the slowest? (One of these is about 57° and the other is under a hundredth of a degree. Which hands can tell neighbors apart?)

Then, *faraway* positions: between positions m and $m + 10{,}000$, how many full laps has the fastest hand made, and why does its final angle carry no usable ordering information? Show the slowest hand has advanced by about 1.3 radians, less than one lap, so it reads the separation unambiguously.

Finally, put the ladder together. Show that for any two positions with $0 < n - m < 2\pi/\theta_{31}$, the slowest hand assigns them different angles, and conclude that $\mathbf{\Theta}_m \neq \mathbf{\Theta}_n$: across a 47,000-token context, distinct positions get distinct embeddings, with fast hands resolving neighbors and slow hands separating chapters.

Part C: Advancing the Clock Is One Linear Map

A relative encoding should make "k positions later" mean the same thing everywhere in the sequence. Prove it does: show that

$$\mathbf{R}_k \, \mathbf{R}_m = \mathbf{R}_{m+k},$$

either by multiplying the matrices and using the angle-addition formulas, or in two lines from the class identity $\mathbf{R}_m^\top \mathbf{R}_n = \mathbf{R}_{n-m}$ together with the observation that $\mathbf{R}_k^\top = \mathbf{R}_{-k}$. Extend the statement block by block to the full clock, $\mathbf{\Theta}_k \mathbf{\Theta}_m = \mathbf{\Theta}_{m+k}$, and conclude: the map sending every position-m query $\mathbf{\Theta}_m \mathbf{q}$ to its position-$(m+k)$ counterpart is multiplication by the *fixed* matrix $\mathbf{\Theta}_k$, independent of m. Close with one sentence connecting this to convolution: which property from the Convolutional Networks lecture gave "the same edge detector works at every pixel," and what is its analogue here?

Where this goes next. A positional embedding is a clock: one hand aliases, and a ladder of hands tells time at every scale. Part C is the algebra behind relative position: "shift by k" is a single linear map applied uniformly, the same property translation equivariance gave convolution. With it the course's structure arc closes, from circulant convolution matrices to Toeplitz attention scores. Your Part A and B estimates are used in practice. When labs stretch a model's context window past the length it was trained on, they re-tune the base 10000 so that the slow hands' periods outlast the new context, using exactly the period formula you derived. This lecture completes the Architectures unit, and the course turns to its final question: how to train a model when nobody hands it the right answer. That is reinforcement learning, where the Monte Carlo estimators of the first unit return as the engine of policy gradients.