CSCI 145 Problem 19: The Matrix That Commutes with Shift

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 19: The Matrix That Commutes with Shift

Shift a photograph two pixels to the left and it is still the same cat. Lecture claimed convolutional networks respect this automatically: shift the input and the output shifts along with it. Claims about neural networks are usually checked empirically, on benchmarks, with error bars. This one you can *prove*. The whole argument lives in the wrap-around world lecture mentioned in passing: signals arranged in a circle, so that sliding off the right edge re-enters on the left and every position is treated identically.

Throughout, signals are vectors $\mathbf{x} \in \mathbb{R}^6$ with indices read cyclically (x_7 means x_1, and x_8 means x_2), and the **circular convolution** of $\mathbf{x}$ with the 3-tap kernel $\mathbf{w} = (w_1, w_2, w_3)$ is the vector $\mathbf{y} \in \mathbb{R}^6$ with

$$y_i = w_1 x_i + w_2 x_{i+1} + w_3 x_{i+2}, \qquad i = 1, \ldots, 6.$$

This is exactly lecture's convolution, except that the window never falls off the end, so the output has all six entries instead of four.

Part A: Convolution Is a Matrix

Write the circular convolution as a matrix equation $\mathbf{y} = \mathbf{C}\mathbf{x}$: give all six rows of $\mathbf{C} \in \mathbb{R}^{6\times 6}$ in terms of w_1, w_2, w_3, and observe that each row is the row above it shifted one place to the right (with wrap-around). A matrix with this structure is called **circulant**: thirty-six entries, three free numbers, locality and weight sharing baked into its shape exactly as in lecture's 4×6 version.

Now take lecture's kernel $\mathbf{w} = (-1, 0, 1)$ and signal $\mathbf{x} = (1, 1, 1, 5, 5, 5)$, and compute $\mathbf{C}\mathbf{x}$ by hand. Check that the first four entries reproduce the in-class answer $(0, 4, 4, 0)$, and explain in one sentence what *edge* the two new entries have detected.

Part B: Equivariance Is a Two-Line Proof

Define the **shift matrix** $\mathbf{S} \in \mathbb{R}^{6\times 6}$ by $(\mathbf{S}\mathbf{x})_i = x_{i+1}$ (cyclically, so $\mathbf{S}$ rotates the signal one position). Write out $\mathbf{S}$ as a matrix, and show that

$$\mathbf{C} = w_1\mathbf{I} + w_2\mathbf{S} + w_3\mathbf{S}^2$$

by comparing what each side does to the i-th entry of a signal. The convolution matrix is a *polynomial in the shift*.

Use this to prove $\mathbf{C}\mathbf{S} = \mathbf{S}\mathbf{C}$ in two lines (multiply the polynomial by $\mathbf{S}$ on each side and compare). Then interpret the equation you just proved: applied to a signal, $\mathbf{C}(\mathbf{S}\mathbf{x}) = \mathbf{S}(\mathbf{C}\mathbf{x})$ says that convolving a shifted signal and shifting a convolved signal are *the same computation*. That is translation equivariance, and it holds for every kernel at once. Finally, explain in one sentence why a generic dense matrix $\mathbf{W}$ has no reason to satisfy $\mathbf{W}\mathbf{S} = \mathbf{S}\mathbf{W}$.

Part C: Equivariance Survives Depth

A convolutional *network* is a stack of convolutions with nonlinearities in between, so equivariance is only useful if it survives the stack. Let σ denote the ReLU applied entrywise. Show that $\sigma(\mathbf{S}\mathbf{x}) = \mathbf{S}\,\sigma(\mathbf{x})$, so that shifting and then applying ReLU is the same as applying ReLU and then shifting, and explain in one sentence why the same argument works for *any* entrywise nonlinearity. (Hint: $\mathbf{S}$ only reorders coordinates; it never mixes them.)

Now let $\mathbf{C}_1$ and $\mathbf{C}_2$ be circulant matrices for two (possibly different) kernels, and consider the two-layer convolutional network $N(\mathbf{x}) = \mathbf{C}_2\,\sigma(\mathbf{C}_1\mathbf{x})$. Chain your facts from Parts B and C to prove

$$N(\mathbf{S}\mathbf{x}) = \mathbf{S}\,N(\mathbf{x}),$$

and conclude in one sentence: weight sharing is a symmetry constraint on the weight matrix, and the network's equivariance is a theorem about that constraint rather than an empirical observation that might fail on the next benchmark.

Part D*: One Eigenbasis for Every Kernel

Part B showed every 3-tap circulant is a polynomial in $\mathbf{S}$. By the logic of the Decompositions lecture, once we know $\mathbf{S}$'s eigenvectors we know every circulant's eigenvectors, *before the kernel is even chosen*. Let $\omega = e^{2\pi i/6}$ (here $i = \sqrt{-1}$, not an index, and $\omega^6 = 1$), and for $m = 0, 1, \ldots, 5$ define the **Fourier vector** $\mathbf{f}_m \in \mathbb{C}^6$ with entries $(\mathbf{f}_m)_j = \omega^{m(j-1)}$ for $j = 1, \ldots, 6$.

First show $\mathbf{S}\mathbf{f}_m = \omega^m \mathbf{f}_m$: shifting a sampled complex exponential just multiplies it by a phase, so each $\mathbf{f}_m$ is an eigenvector of the shift. Then apply $\mathbf{C} = w_1\mathbf{I} + w_2\mathbf{S} + w_3\mathbf{S}^2$ to $\mathbf{f}_m$ and conclude

$$\mathbf{C}\mathbf{f}_m = \left(w_1 + w_2\,\omega^m + w_3\,\omega^{2m}\right)\mathbf{f}_m.$$

The eigenvalue is the *discrete Fourier transform* of the kernel, evaluated at frequency m. State the punchline in one sentence: every circulant, meaning every kernel anyone will ever learn, is diagonalized by the same six Fourier vectors, so in the Fourier basis convolution is entrywise multiplication. (This is the *convolution theorem*, and it is why convolutions can be computed in $O(n \log n)$ time via the fast Fourier transform.)

Where this goes next. This problem constrains a weight matrix to respect a symmetry ($\mathbf{CS} = \mathbf{SC}$), and that constraint yields a guarantee no amount of training data could provide, in two lines of algebra. At the end of this unit the same move reappears: rotary positional embeddings force attention scores to depend only on the *difference* of two positions, making the score matrix Toeplitz, which is today's circulant with the circle cut open. And Part D repeats the lesson of the spectral theorem: find the right basis once, and a whole family of matrices is diagonal in it.