Problem H-29. Assume that nucleotide frequencies of A and T are $p_A = 0.2, p_T = 0.3$.
(a) Draw the graph of a Markov chain (analogous to $M_1$ on the Markov chain slides) for analyzing overlapping occurrences of ATA in a nucleotide sequence.
(b) Determine the transition matrix of the Markov chain in (a).
(c) Determine the 2-step and 3-step transition matrices for this Markov chain.
(d) Determine the stationary distribution for this Markov chain.

Note: This problem is easy to do by setting up and solving an appropriate system of equations in the manner shown on the slides. In principle, this could also be done by solving for the eigenvalues and eigenvectors. However, eig in Matlab and eigen in R return incorrect results since it turns out the transition matrix has repeated eigenvalues and is not diagonalizable; rather than detect it is not diagonalizable, they just return incorrect results without warning. Since it’s not diagonalizeable, one would have to use the more general concepts of “generalized eigenvectors” and the Jordan Canonical Form. This is available in Matlab via jordan but is not available in R.
(e) Draw the graph of the Markov chain for non-overlapping occurrences of ATA (analogous to $M_2$ on the Markov chain slides) and determine its transition matrix.
(f) Consider the Markov chain constructed in step (a). Number the states 1 for “∅” through 4 for “ATA”. For each state $i = 1, 2, 3, 4$, if you start in state $i$, determine the expected number of steps until the next time in state 4.

Problem H-30. Get the file H30data.txt. It contains 100 $(x, y)$ points in a $2 \times 100$ matrix.
(a) Compute the covariance matrix (make sure it’s $2 \times 2$, not $100 \times 100$).
(b) Compute the principal components, the variance along each one, and the fraction of variance explained by each one.

Part (c) is not required and will not be graded. Linear regression may be covered too late for the homework, but you can do it on your own later for practice.
(c) Compute the least squares regression line $y = \beta_0 + \beta_1 x$. 