1. Let $f(x_1, x_2) = (x_1^2 x_2, x_1 x_2^3)$, $g(y_1, y_2) = (\sin(y_1 y_2), y_1 + y_2)$, and $h = g \circ f$. Use the matrix chain rule to compute the matrix derivative Dh. Your answer must be in terms of x_1, x_2 and contain no y_1 or y_2 and be a single matrix.

2. Consider the net diagram shown in Figure 1 on the next page. The hidden layer has activation function σ (unspecified) and no activation function on the output layer.
   (a) Write the formula for the input-output function $F(\vec{x}, \eta)$.
   (b) Recall that the parameters are $\eta = (w_{11}, w_{21}, w_{12}, w_{22}, b_1, b_2, \gamma_1, \gamma_2)$. Write the formula for the derivative (or gradient) of F with respect to the parameters η. Your answer will have terms containing η'.

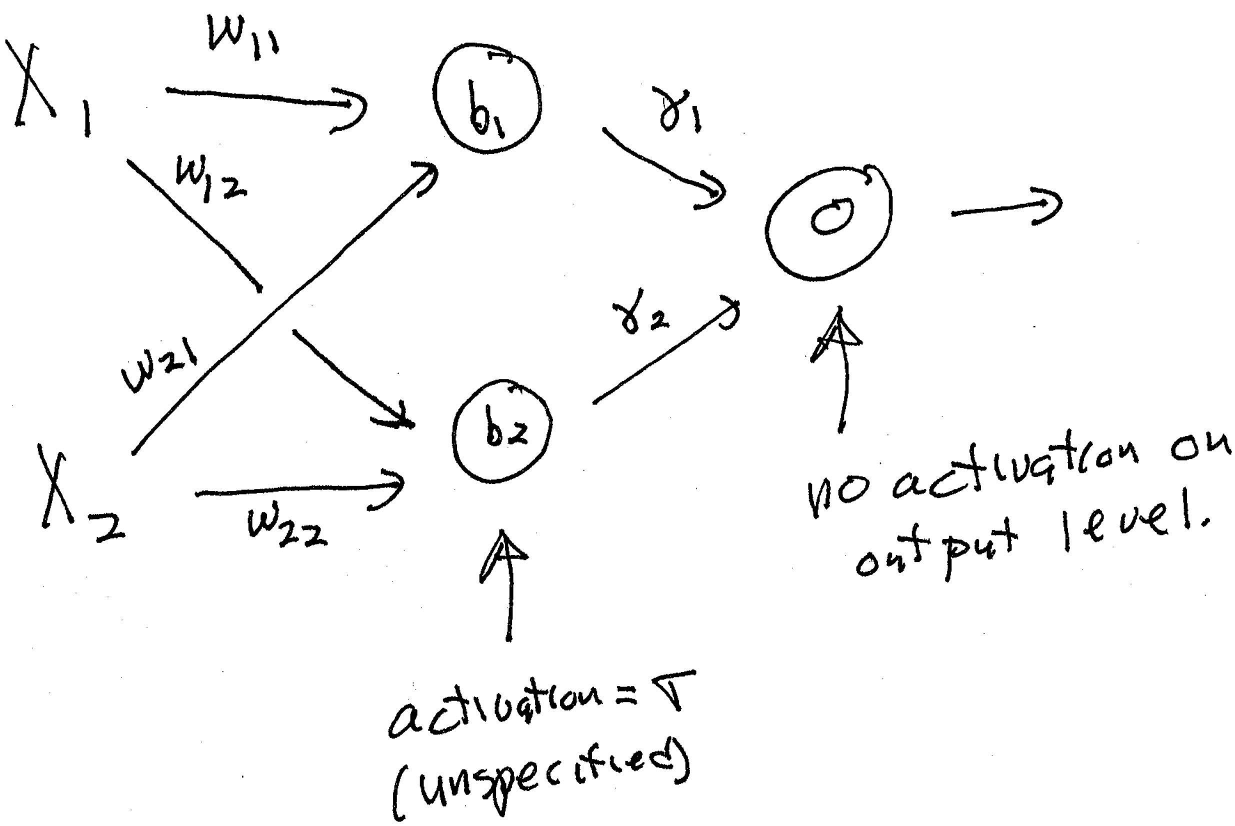
3. This problem concerns finding the least squares solution to $A\vec{x} = \vec{b}$. There are both theory and computational questions. For the computational questions your answer must include your code and the results of running it.
   (a) Let $\Phi(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2$. Compute $\nabla\Phi$ and $H\Phi$.
   (b) Let A be a full rank $(m \times n)$-matrix with $m \geq n$. Show (for example using the SVD) that $A^T A$ is symmetric, positive definite. (We also know it is invertible).
   (c) Using (a) and (b), show that under the conditions on A given in (b) there is a unique critical point for $\nabla\Phi$ and that critical point is a local minimum and thus a global minimum.
   (d) As on the exam, let

$$A = \begin{pmatrix} 2 & 2 \\ 2 & 4 \\ 2 & 2 \\ 2 & 4 \end{pmatrix} \text{ and } \vec{b} = \begin{pmatrix} 2 \\ 4 \\ 4 \\ -2 \end{pmatrix}$$

   Write a program that implements gradient descent for $\Phi(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2$ with initial vector $\vec{x}_0 = (1, 1)^T$ and a given step size h. It should include a condition so that it does not compute more than 500 iterates for reasons that will become clear.
   (e) We know that the correct solution is

$$\vec{x}_* = \begin{pmatrix} 2.5 \\ -1 \end{pmatrix}$$

   Using $\|\vec{x}_n - \vec{x}_*\|_2 < 10^{-3}$ as a halting condition and $h = .01$, what is the final value of n and $\|\nabla\Phi(x_n)\|_2$?
   (f) Repeat part (e) for $h = .013, .015, .017$, and $.019$.
   (g) Explain your results and what they say about using gradient descent.



X1
X2
w11
w12
w21
w22
b1
b2
γ1
γ2
O
no activation on output level.
activation = σ (unspecified)