\( D_f(x_1, x_2) = \begin{bmatrix} 2x_1 x_2 & x_1^2 \\ x_2^3 & 3x_1 x_2^2 \end{bmatrix} \)

\( D_g(y_1, y_2) = \begin{bmatrix} \cos(y_1 y_2) y_2 & \cos(y_1 y_2) y_1 \\ 1 & 1 \end{bmatrix} \)

\[
\begin{align*}
h = g \circ f & \Rightarrow D_h(x) = Dg(f(x)) \cdot Df(x) \\
& = \begin{bmatrix} \cos(x_1^3 x_2^4) x_1 x_2^3 & \cos(x_1 x_2^7) x_1 x_2^2 \\ 1 & 1 \end{bmatrix} \end{align*}
\]

\[
\begin{align*}
& = \begin{bmatrix} 2x_1 x_2 & x_1^2 \\ x_2^3 & 3x_1 x_2^2 \end{bmatrix} \\
& = \begin{bmatrix} 3x_1^2 x_2 \cos(x_1^3 x_2^4) & 4x_1^3 x_2^3 \cos(x_1 x_2^7) \\ 2x_1 x_2 + x_2^3 & x_1^2 + 3x_1 x_2^2 \end{bmatrix} 
\end{align*}
\]
$2 \begin{align*}
F(x, y) &= \gamma_1 \sum \left( w_{11} x_1 + w_{21} x_2 + b_1 \right) \\
&\quad + \gamma_2 \sum \left( w_{12} x_1 + w_{22} x_2 + b_2 \right) \\
\n\nabla_{\boldsymbol{\eta}} F &= \begin{bmatrix}
\frac{dF}{dw_{11}} & \frac{dF}{dw_{12}} & \frac{dF}{dw_{21}} & \frac{dF}{dw_{22}} & \frac{dF}{db_1} & \frac{dF}{db_2} & \frac{dF}{d\eta_1} & \frac{dF}{d\eta_2}
\end{bmatrix}
\n\text{Letting} \\
\eta_1 &= w_{11} x_1 + w_{21} x_2 + b_1 \\
\eta_2 &= w_{12} x_1 + w_{22} x_2 + b_2 \\

\frac{dF}{dw_{11}} &= \gamma_1 \sum \left( \frac{d}{d\eta_1} \right) x_1 \\
\frac{dF}{dw_{12}} &= \gamma_2 \sum \left( \frac{d}{d\eta_2} \right) x_1 \\
\frac{dF}{dw_{21}} &= \gamma_1 \sum \left( \frac{d}{d\eta_1} \right) x_2 \\
\frac{dF}{dw_{22}} &= \gamma_2 \sum \left( \frac{d}{d\eta_2} \right) x_2 \\
\frac{dF}{db_1} &= \gamma_1 \sum \left( \frac{d}{db_1} \right) \\
\frac{dF}{db_2} &= \gamma_2 \sum \left( \frac{d}{db_2} \right)
\end{align*}$
\[
\frac{\partial F}{\partial \beta_1} = \gamma_1^2, \quad \frac{\partial F}{\partial \beta_2} = \gamma_2^2
\]
\[
\frac{\partial F}{\partial \gamma_1} = \gamma_1, \quad \frac{\partial F}{\partial \gamma_2} = \gamma_2
\]

3. \( \phi(x) = \|Ax-b\|_2^2 = (Ax-b)^T(Ax-b) \)

\[
= (x^TA^T-b^T)/Ax-b) = x^TA^TAx - b^TAx
\]

- \( x^TA^Tb + b^Tb \)

Now notice \( b^TAx = (x^TA^Ty)^T \) and \( x \) and \( y \) are both numbers, so \( b^TAx = x^TA^Ty \) so

\( \phi(x) = x^TA^TAx - 2b^TAx + b^Tb \)

as we computed in a previous homework.

4. \( \nabla \phi(x) = 2ATAx - 2b^TA \)

and \( H \phi(x) = 2ATA \)

5. \( A = U \Sigma V^T, \quad \Sigma = \text{diag} \{ \Sigma_1, \ldots, \Sigma_n \} \)

with \( \Sigma_n > 0 \) since \( A \) is full rank.

So \( A^TA = V \Sigma^T U^T U \Sigma V \)

\[
= V \Sigma^2 V^{-1}
\]
So the eigenvalues of $A^TA$ are $\lambda_1, \ldots, \lambda_n$ all $> 0$, so $A^TA$ is post-def.

(c) Critical points are when

$$0 = D\Phi(x) = 2A^TAx - 2b^TA$$

or when $A^TAx = b^TA$. But we know $A^TA$ is invertible, so the unique 50 and thus $A^TA$ is invertible, so the unique so is unique and thus $A^TA$ is invertible, so the unique so is unique and thus $A^TA$ is invertible, so the unique so is unique and thus $A^TA$ is invertible, so the unique so is unique and thus $A^TA$ is invertible, so the unique so is unique.

Now at that point, and every open point, $H(\Phi) = 2A^TA$ which is post-def.

So by the 2nd deriv test, $x_0$ is a local minimum. There are no other critical points and thus no other local min, so $x_0$ is the global min.

(3g) Up a a point, longer jumps get you to the minimum with less steps. However, if the magnitude $h\nabla\Phi(x)$
is too large, you jump across the critical point to a place where \( h \Delta \phi(x_{\text{min}}) \) is even larger.

And then you zig zag off to infinity.

So the choice of the value of \( h \) is crucial in using gradient descent and as mentioned in lecture there are many sophisticated tools for choosing and altering learning rate.