1. Problem 9.1 from the text book.

2. Let \( J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 \)
   (a) Compute \( \nabla_{\theta} J(\theta) \) and \( \nabla^2_{\theta} J(\theta) \).
   (b) Show that \( J(\theta) \) is convex.
   (c) Under what conditions on input samples, \( J(\theta) \) is strictly convex?

3. Let \( \{x^{(1)}, \ldots, x^{(m)}\} \) be \( m \) i.i.d samples drawn from a Gaussian distribution \( \mathcal{N}(\mu_{\text{true}}, \sigma_{\text{true}}^2 I) \) where parameters \( \mu_{\text{true}} \) and \( \sigma_{\text{true}} \) are unknown. A common approach to estimate model parameters is maximum likelihood estimation (MLE).
   (a) The likelihood function \( L(\mu, \sigma) \) is defined as the probability of observing samples \( \{x^{(1)}, \ldots, x^{(m)}\} \) from the distribution \( \mathcal{N}(\mu, \sigma^2 I) \). Write down the likelihood function in this case.
   (b) Argue that argmax_{\mu,\sigma} L(\mu, \sigma) = argmax_{\mu,\sigma} \log L(\mu, \sigma).
   (c) By maximizing the log likelihood function, compute MLE estimates of model parameters.

4. Compute \( \nabla_X \text{Tr}[AXBX^T CXD] =? \)
   - **Hint 1**: if \( dy = \text{Tr}[A(dX)] \), then the \( \frac{dy}{dX} = A^T \).
   - **Hint 2**: The trace is invariant to cyclic permutations. For example, \( \text{Tr}[A_1 A_2 A_3] = \text{Tr}[A_3 A_1 A_2] = \text{Tr}[A_2 A_3 A_1] \). In general, \( \text{Tr}[A_1 A_2 \ldots A_n] = \text{Tr}[A_k A_{k+1} \ldots A_n A_1 \ldots A_{k-1}] \), \( 1 \leq k \leq n \).
   - **Hint 3**: \( \text{Tr}[A] = \text{Tr}[A^T] \).
   - **Hint 4**: The following is a solution for a simplified version of the problem. To compute \( \nabla_X \text{Tr}[AXBX^T] \), we can write
     \[
     d \text{Tr}[AXBX^T] = \text{Tr}[d(AXBX^T)]
     \]
     \[
     = \text{Tr}[A d(X) BX^T] + \text{Tr}[AXB d(X^T)] \quad \text{(using the product rule of derivatives)}
     \]
     \[
     = \text{Tr}[BX^T A d(X)] + \text{Tr}[AXB d(X^T)] \quad \text{(using the cyclic permutation property for the first term)}
     \]
     \[
     = \text{Tr}[BX^T A d(X)] + \text{Tr}[d(X) B^T X^T A^T] \quad \text{(using the transpose invariance property for the second term)}
     \]
     \[
     = \text{Tr}[BX^T A d(X)] + \text{Tr}[B^T X^T A^T d(X)] \quad \text{(using the cyclic permutation property for the second term)}
     \]
     \[
     = \text{Tr}[BX^T A + B^T X^T A^T] d(X).
     \]
     Therefore, using **Hint 1**, we have \( \nabla_X \text{Tr}[AXBX^T] = A^T XB^T + AXB \).
5. (Programming Assignment) Let $x \in \mathbb{R}^n$ and $z \in \mathbb{R}$ be zero-mean independent Gaussian random variables with covariance matrices $\mathbf{I}$ and $\sigma^2$, respectively. That is, $x \sim \mathcal{N}(0, \mathbf{I})$ and $z \sim \mathcal{N}(0, \sigma^2)$. Define $y = \theta^T x + \theta_0 + z$. In this assignment, we want to use stochastic gradient descent (SGD) to compute a linear regression model between $x$ and $y$. Write a Python code to do the following:

(a) Let $n = 4$, $\sigma^2 = 1/4$, $\theta = [1, 1/2, 1/4, 1/8]^T$ and $\theta_0 = 2$. Generate $m = 10,000$ i.i.d. training samples from $P_{X,Y}$. That is $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$.

(b) Use SGD with a batch size of 10 to estimate model parameters. Plot the Mean-Squared Error (MSE) vs. the number of iterations.

(c) Generate $m$ new i.i.d. test samples from $P_{X,Y}$. Use estimated parameters to compute the MSE on the test set.

(d) Repeat parts (a)-(c) using $m = 10$. How do training and test errors change? Why?