Theoretical Task for Tutorial 1

Sarat Moka

Analysis of the Linear Regression Optimization Problem

Consider the standard linear regression model:

$$y = X\beta + \varepsilon$$

where y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of regressors (features). We assume X has full column rank, i.e., $\text{rank}(X) = p$, β is a $p \times 1$ vector of unknown coefficients, and ε is an $n \times 1$ vector of independent random errors.

Task

Our goal is to find the vector β that minimizes the sum of squared residuals, which is the loss function $C(\beta)$:

$$C(\beta) = \|y - X\beta\|_2^2 = (y - X\beta)^\top (y - X\beta)$$

Execute this task via the following steps:

S1 Obtain gradient expression of $C(\beta)$ with respect to β.

S2 Obtain Hessian expression.

S3 Show that Hessian is positive definite, which implies unique minimum (i.e., unique stationary point).

S4 Equate the gradient to zero to obtain the solution of the target problem.

Hint

Definition 1 (Linear independence of vectors)**.** *A set of p vectors $a^{(1)}, a^{(2)}, \ldots, a^{(p)}$ is said to be **linearly independent** if the only scalars $c_1, c_2, \ldots, c_p$ satisfying the equation*

$$c_1 a^{(1)} + c_2 a^{(2)} + \cdots + c_p a^{(p)} = 0$$

are $c_1 = c_2 = \cdots = c_p = 0$.

Hint

Definition 2 (Positive Definite Matrix)**.** *A symmetric $p \times p$ matrix A is said to be **positive definite** if for any non-zero vector $z \in \mathbb{R}^p$, the quadratic form $z^\top A z$ is strictly positive:*

$$z^\top A z > 0 \quad \text{for all } z \neq 0.$$

Hint

Definition 3 (Convex Function)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be **convex** if for any two points $u, u' \in \mathbb{R}$ and for any $\alpha \in [0, 1]$, the following inequality holds:*

$$f(\alpha u + (1 - \alpha)u') \leq \alpha f(u) + (1 - \alpha) f(u').$$

This means that the line segment connecting any two points on the graph of the function lies on or above the graph of the function.

In our case, the function is $C(\beta)$.

Hint

For a twice-differentiable function $f(u) : \mathbb{R}^p \to \mathbb{R}$, if its Hessian matrix is positive definite for all $u \in \mathbb{R}^p$, then the function f is strictly convex, guaranteeing a unique stationary point: the only global minimum.

Definition 4 (Hessian Matrix)**.** *The* ***Hessian matrix*** *H (or $\nabla^2 f(u)$) of a scalar-valued function $f(u)$ of k variables $u = (u_1, u_2, \ldots, u_k)^\top$ is the $k \times k$ matrix of second-order partial derivatives:*

$$(H)_{ij} = \frac{\partial^2 f}{\partial u_i \partial u_j}.$$

For our loss function $C(\beta)$, the Hessian matrix $\nabla^2 C(\beta)$ will be a $p \times p$ matrix where the (i, j)-th entry is $\frac{\partial^2 C(\beta)}{\partial \beta_i \partial \beta_j}$.

1 Step 1: Derive Gradient of $C(\beta)$

The loss function is $C(\beta) = (y - X\beta)^\top (y - X\beta)$. Let's expand this:

$$\begin{aligned} C(\beta) &= y^\top y - y^\top X\beta - (X\beta)^\top y + (X\beta)^\top X\beta \\ &= y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta \end{aligned}$$

Since $y^\top X\beta$ is a scalar, it is equal to its transpose: $y^\top X\beta = (y^\top X\beta)^\top = \beta^\top X^\top y$. So,

$$C(\beta) = y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta$$

Now, we compute the gradient with respect to β. We use the following matrix differentiation rules:

- $\frac{\partial (a^\top u)}{\partial u} = a$
- $\frac{\partial (u^\top A u)}{\partial u} = (A + A^\top)u$. If A is symmetric ($A = A^\top$), then this simplifies to $2Au$. Note that the matrix $X^\top X$ is symmetric.

Applying these rules:

$$\begin{aligned} \nabla C(\beta) = \frac{\partial C(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta}(y^\top y) - \frac{\partial}{\partial \beta}(2\beta^\top X^\top y) + \frac{\partial}{\partial \beta}(\beta^\top X^\top X\beta) \\ &= 0 - 2X^\top y + 2X^\top X\beta \\ &= -2X^\top y + 2X^\top X\beta \end{aligned}$$

So, the gradient is:

$$\nabla C(\beta) = 2X^\top X\beta - 2X^\top y$$

2 Step 2: Compute the Hessian Matrix of $C(\beta)$

The Hessian matrix is the derivative of the gradient $\nabla C(\beta)$ with respect to $\beta^\top$ (or, equivalently, the matrix of second partial derivatives $\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}$). Recall that

$$\nabla C(\beta) = 2X^\top X\beta - 2X^\top y$$

Differentiating again with respect to $\beta^\top$: The term $-2X^\top y$ is constant with respect to β, so its derivative is 0. For the term $2X^\top X\beta$, using the rule $\frac{\partial (Au)}{\partial u^\top} = A$:

$$\nabla^2 C(\beta) = \frac{\partial (\nabla C(\beta))}{\partial \beta^\top} = \frac{\partial}{\partial \beta^\top}(2X^\top X\beta - 2X^\top y) = 2X^\top X$$

Thus, the Hessian matrix is

$$H = \nabla^2 C(\beta) = 2X^\top X$$

Note that the Hessian is constant and does not depend on β.

3 Step 3: Show the Hessian is Positive Semi-Definite

The Hessian matrix is $H = 2X^\top X$. This is a $p \times p$ matrix. To show that H is positive definite, we need to show that for any non-zero vector $z \in \mathbb{R}^p$, $z^\top Hz > 0$. Towards that,

$$\begin{aligned} z^\top Hz &= z^\top (2X^\top X)z \\ &= 2z^\top X^\top Xz \\ &= 2(Xz)^\top (Xz) \\ &= 2\|Xz\|_2^2 \end{aligned}$$

Since the squared L2 norm $\|Xz\|_2^2$ is always greater than or equal to zero, we have:

$$z^\top Hz = 2\|Xz\|_2^2 \geq 0.$$

Furthermore, we are given that X has full column rank (p). This means that the columns of X are linearly independent (see the hint). Thus, $Xz = 0$ if and only if $z = 0$. Therefore, for any $z \neq 0$, $Xz \neq 0$, which implies $\|Xz\|_2^2 > 0$. Consequently, for $z \neq 0$:

$$z^\top Hz = 2\|Xz\|_2^2 > 0$$

This means that if X has full column rank, the Hessian $H = 2X^\top X$ is **positive definite**. Since the Hessian is positive definite, the loss function $C(\beta)$ is strictly convex, which guarantees a unique minimum.

4 Step 4: Obtain OLS Solution using the Gradient

For a strictly convex function, the minimum occurs where the gradient is equal to the zero vector. We set the gradient $\nabla C(\beta)$ to 0, all-zero vector to get

$$\nabla C(\beta) = 2X^\top X\beta - 2X^\top y = 0$$

$$2X^\top X\beta = 2X^\top y$$

$$X^\top X\beta = X^\top y.$$

Since X has full column rank, the matrix $X^\top X$ (which is $p \times p$) is invertible. We can pre-multiply both sides by $(X^\top X)^{-1}$ to solve for β:

$$(X^\top X)^{-1}(X^\top X)\beta = (X^\top X)^{-1}X^\top y$$

$$I_p\beta = (X^\top X)^{-1}X^\top y$$

The OLS estimator for β, denoted as $\hat{\beta}_{\text{OLS}}$, is:

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1}X^\top y$$

This is the unique solution that minimizes the sum of squared residuals $C(\beta)$, due to the strict convexity of the loss function.