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 $L(\beta)$:
> $$L(\beta) = \|y - X\beta\|_2^2 = (y - X\beta)^\top (y - X\beta)$$

Execute this task via the following steps:

S1 Obtain gradient expression of $L(\beta)$ 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 $L(\beta)$.

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