Exercise 1: Hessian of Logistic Regression

Binary logistic regression with training data $(x_i, y_i)_{i=1}^N$, where $x_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$. The model is:

$$p(y_i = 1 \mid x_i, w) = \sigma(w^\top x_i), \quad p(y_i = 0 \mid x_i, w) = 1 - \sigma(w^\top x_i),$$

where the logistic sigmoid function is:

$$\sigma(t) = \frac{1}{1 + e^{-t}}.$$

The negative log-likelihood (NLL) is:

$$L(w; X, y) = -\sum_{i=1}^N \Big[y_i \log \sigma(w^\top x_i) + (1 - y_i) \log(1 - \sigma(w^\top x_i)) \Big].$$

Gradient

Define:

$$p_i(w) = \sigma(w^\top x_i).$$

Then the gradient is:

$$\nabla_w L(w) = \sum_{i=1}^N (p_i(w) - y_i) x_i.$$

In matrix form:

$$\nabla_w L(w) = X^\top (p - y),$$

where $X \in \mathbb{R}^{N \times D}$, $p = (\sigma(w^\top x_1), \ldots, \sigma(w^\top x_N))^\top$.

Hessian

Differentiate again:

$$\frac{\partial^2 L}{\partial w^2} = \sum_{i=1}^N p_i(w)(1 - p_i(w)) x_i x_i^\top.$$

Define the diagonal matrix:

$$S(w) = \operatorname{diag}\big(p_1(1 - p_1), \ldots, p_N(1 - p_N)\big) \in \mathbb{R}^{N \times N}.$$

Then the Hessian in compact form is:

$$\boxed{\nabla_w^2 L(w; X, y) = X^\top S(w) X.}$$

Final Thought

The Hessian is positive semi-definite, which implies that the logistic regression loss is convex. Each diagonal entry of $S(w)$ corresponds to the variance of the Bernoulli distribution for sample i.

Exercise 2: Kullback–Leibler Divergence

The KL divergence of q from p is defined as:

$$D_{\mathrm{KL}}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

Part 1: Non-negativity

By Jensen's inequality and the concavity of log:

$$\sum_x p(x)\log\frac{p(x)}{q(x)} \geq \log\sum_x p(x)\frac{p(x)}{q(x)} = \log 1 = 0.$$

Equality holds if and only if $p(x) = q(x)$ for all x. Thus:

$$D_{\mathrm{KL}}(p\|q) \geq 0, \quad D_{\mathrm{KL}}(p\|q) = 0 \iff p = q.$$

Part 2: KL Divergence is not symmetric

Consider:

$$p = (0.9, 0.1), \quad q = (0.5, 0.5).$$

Compute:

$$D_{\mathrm{KL}}(p\|q) = 0.9\log\frac{0.9}{0.5} + 0.1\log\frac{0.1}{0.5} \approx 0.368,$$

and:

$$D_{\mathrm{KL}}(q\|p) = 0.5\log\frac{0.5}{0.9} + 0.5\log\frac{0.5}{0.1} \approx 0.511.$$

Thus:

$$D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p),$$

showing that KL divergence is not symmetric.

Final Thought

- $D_{\mathrm{KL}}(p\|q) \geq 0$ and equals zero iff $p = q$.
- KL divergence is not symmetric in general.