STAT 37601 HW4

Albert Chu

April 2025

1. (a) We calculate the gradient as

$$\nabla L = \sum_{i=1}^{n} -Y_i X_i^t \frac{e^{-Y_i X_i^t \theta}}{1 + e^{-Y_i X_i^t \theta}} = \sum_i X_i (p_i - y_i)$$

for

$$p_i = \frac{1}{1 + e^{-X_i^t \theta}}, y_i = \frac{Y_i + 1}{2}$$

and the hessian as

$$\nabla^2 L = \sum_i^n X_i X_i^T \frac{e^{-Y_i X_i^t \theta}}{(1 + e^{-Y_i X_i^t \theta}} = \sum_i X_i X_i^T p_i (1 - p_i) = X^T W X$$

for

$$W = diag((p_i(1 - p_i))_{i=1}^n$$

We can then plug in

$$\theta_{new} = \theta_{old} - H^{-1} g = \theta_{old} - (X^T W X)^{-1} X^T (p - y)$$

$$(X^T W X)\theta_{new} = X^T W X \theta_{old} - X^T (p - y)$$

Observe that if we define

$$Z = X\theta_{old} + W^{-1}(y - p)$$

then $X^T(p - y) = -X^T W (Z - X\theta_{old})$,

$$X^T W X \theta_{new} = X^T W X \theta_{old} + X^T W (Z - X\theta_{old}) = X^T W Z$$

This gives us each Newton update as the solution of the WLS normal equation above. We have explicit formulas as well,

$$W_{ii} = p_i(1 - p_i), p_i = \frac{1}{1 + e^{-X_i^t \theta_{old}}}$$

$$Z_i = X_i^t \theta_{old} + \frac{y_i - p_i}{p_i(1-p_i)}, y_i = \frac{Y_i + 1}{2}$$

(b) Let $\theta = \alpha\theta^*$ for scalar α. Then we have that

- If $Y_i = 1$, $X_i^T(\alpha\theta^*) = \alpha(X_i^T\theta^*) \to \infty$ as $\alpha \to \infty$, so $p_i = \frac{1}{1+e^{-X_i^T\theta}} \to 1$
- If $Y_i = 0$, $X_i^T(\alpha\theta^*) \to -\infty$, so

$$1 - p_i \to 1$$

Therefore every factor in the product $L(\alpha\theta^*)$ tends to 1, implying $L(\alpha\theta^*) \to 1$ as $\alpha \to \infty$. But we also know that for any finite θ, $L(\theta) < 1$, Therefore, there is no finite maximizer of the likelihood.

In terms of Newton's iteratively reweighted least squares algorithm, we see that with perfectly linearly separable data, the probabilities p_i approach either 0 or 1, so each weight $p_i(1-p_i) \to 0$. This implies the diagonal entries of W will shrink towards zero and $|\theta|$ will be driven larger, yet the algorithm won't be able to converge to a finite solution.

(c) In the hinge loss case, we scale $\theta = \alpha\theta^*$ as usual with $\alpha \geq \max_i \frac{1}{Y_i X_i^T \theta^*}$, so each margin $Y_i X_i^T \theta \geq 1$ and thus $[1 - Y_i X_i^t \theta]_+ = 0$. This implies that the infimum of the hinge loss is indeed 0 for any θ with $Y_i X_i^t \theta \geq 1$. We also see that the set of minimizers is unbounded – we can always scale any separating θ up and still have zero hinge loss.

In the quadratic loss case, as we send $\alpha \to \infty$, each $[1 - Y_i X_i^t(\alpha\theta^*)]^2 \approx \alpha^2(Y_i X_i^t \theta^*)^2 \to +\infty$. If we consider the hessian and gradient, we see that

$$\nabla^2 L(\theta) = 2\sum_i X_i X_i^t = 2X^tX, \nabla L(\theta) = 0 \implies X^tX\theta = X^ty$$

and so the hessian is positive definite, and we thus have a unique finite minimizer at $\theta = (X^tX)^{-1}X^ty$. Therefore, any gradient-based method will converge to the unique Ls solution.

2. (a) We see the log-likelihood as

$$\sum_i^n \sum_c^C Z_{ic} \log\left(\frac{\exp(\theta_c^t X_i}{\sum_k^C \exp(\theta_k^t X_i}\right)$$

$$= \sum_i^n \sum_c^C Z_{ic} \left(\theta_c^t X_i - \log \sum_k^C \exp(\theta_k^t X_i) \right)$$

$$= \sum_i^n \left(\theta_{Y_i}^t X_i - \log \sum_k^C \exp(\theta_k^t X_i) \right)$$

(b)

$$\nabla_{\theta_c} \log L = \sum_i^n \left(Z_{ic} - \frac{\exp(\theta_k^t X_i)}{\sum_k^C \exp(\theta_k^t X_i)} \right) X_i$$

$$= \sum_i^n (Z_{ic} - \pi_{ic}) X_i$$

$$= X^t (Z_c - \pi_c)$$

(c) Using the above, we get

$$G = \begin{bmatrix} \nabla_{\theta_1} \log L & \ldots & \nabla_{\theta_C} \log L \end{bmatrix} = X^t (Z - \pi)$$