STA2311 (Fall 2023) - Practice Problems for Class 2 (Classical Optimization Methods)

1. Recall the Newton-Raphson Example 1 from Class 2: we have iid observations $\{Y_1, \ldots, Y_n\} \in \mathbb{N}^n$ from the mass function
$$f(y \mid \theta) = \frac{\theta^y}{-y \cdot \log(1-\theta)}, \quad \theta \in (0,1).$$
Derive both the Newton-Raphson and the Fisher Scoring update rules for estimating the MLE of θ. Remember that the original parameter space is constrained — you'll have to do something about that.

2. Consider standard logistic regression, in which we have $\{0,1\}$-valued observations $Y_1, \ldots, Y_n$ and covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ such that $Y_i \mid \boldsymbol{x}_i \sim \text{Bernoulli}(\pi_i)$ independently, with $\pi_i = 1/(1+e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i})$. The unknown parameter here is $\boldsymbol{\beta} \in \mathbb{R}^p$.
   (a) Derive the Newton-Raphson update for estimating the MLE of $\boldsymbol{\beta}$, and show that it's equivalent to the Fisher scoring update.
   (b) Show also that we can write the update in the form
$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)},$$
where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top & \cdots & \boldsymbol{x}_n^\top \end{bmatrix}^\top$, $\boldsymbol{W}^{(t)}$ is a diagonal matrix with i'th diagonal entry equal to $\pi_i^{(t)}(1-\pi_i^{(t)})$, and $\boldsymbol{z}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{W}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\pi}^{(t)})$. Thus, this case of Newton-Raphson is an instance of an *iteratively reweighted least squares (IRLS)* procedure.

3. Consider the locations of 10 hotels scattered around a hilly alpine village with the following geographical coordinates:

Hotel	x-coordinate	y-coordinate	z-coordinate
1	3.92	6.10	1.87
2	5.57	6.55	1.26
3	7.88	-2.48	0.05
4	-4.20	-1.02	1.73
5	-1.87	6.59	0.10
6	-0.66	6.23	0.17
7	2.11	5.53	2.30
8	-1.40	2.34	0.08
9	-2.36	4.47	0.20
10	5.26	0.31	0.63

Note that the z-coordinate represents altitude and is never negative. The village wants to build a new hospital for its tourists that minimizes the average squared distance to the hotels; we want to find the coordinates of such a location (which are hopefully not inside of a hill). That is, we want to find
$$(x^*, y^*, z^*) = \underset{x,y \in \mathbb{R}; z \geq 0}{\text{argmin}} \left(\sum_{i=1}^{10} (x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2 \right).$$
   (a) Solve for (x^*, y^*, z^*) analytically.

(b) Derive the Gauss-Newton update rule for estimating (x^*, y^*, z^*).

(c) Derive the Newton-Raphson update rule for estimating (x^*, y^*, z^*). How would things change if instead we wanted to find

$$\underset{x,y\in\mathbb{R};z\geq 0}{\operatorname{argmin}} \left(\sum_{i=1}^{10} \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} \right)?$$

4. Derive the Newton-Raphson update rule for estimating (p_a, p_b, p_o) in the Newton-Raphson Example 2 on blood types from Class 2, where p_a is the true frequency of allele a in the population, p_b is the true frequency of allele b, and p_o is the true frequency of allele o. You can assume that $\mathbb{P}(\text{allele pair } xy) = p_x p_y$, where $x, y \in \{a, b, o\}$. Remember that $p_a + p_b + p_o = 1$.

5. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ function and set

$$Q(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \nabla^2 g(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Show that

$$g(\boldsymbol{y}) = Q(\boldsymbol{x}, \boldsymbol{y}) + \int_0^1 \int_0^t \big\langle \big(\nabla^2 g(\boldsymbol{x} + s(\boldsymbol{y} - \boldsymbol{x})) - \nabla^2 g(\boldsymbol{x})\big)(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle \, \mathrm{d}s \, \mathrm{d}t.$$

6. A function $g : E \subseteq \mathbb{R} \to \mathbb{R}$ is *convex* if it satisfies

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda) g(y)$$

for all $x, y \in E$ and $\lambda \in [0, 1]$.

(a) Suppose $g \in \mathcal{C}^1$. Show that g is convex if and only if

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle$$

for all $x, y \in E$.

(b) Suppose $g \in \mathcal{C}^1$. Show that g is convex if and only if

$$\langle \nabla g(y) - \nabla g(x), y - x \rangle \geq 0$$

for all $x, y \in E$.

7. A function $g : E \subseteq \mathbb{R} \to \mathbb{R}$ is *α-strongly convex* if

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle + \frac{\alpha}{2} ||y - x||^2$$

for all $x, y \in E$. Show that g is α-strongly convex if and only if

$$f(x) = g(x) - \frac{\alpha}{2} ||x||^2$$

is convex.

8. Consider the functions g and m_k defined in Class 2, where the gradient descent update rule was derived for the L-Lipschitz $\mathcal{C}^2$ function g. Let $x^* = \operatorname{argmin} g(x)$.

(a) Show that if g is convex, then m_k is L-strongly convex.

(b) Show that

$$g(x_{k+1}) \leq m_k(x^*) - \frac{L}{2} ||x_{k+1} - x^*||^2.$$

(c) Show that

$$\underset{1\leq j\leq M}{\operatorname{argmin}}\, g(x_j) - g(x^*) \leq \frac{L}{2M} ||x_0 - x^*||^2.$$

(d) (Tougher) Combine these with results proven in class to further show that

$$\min_{1\leq j\leq M} ||\nabla g(x_j)|| \leq \frac{2L}{M} ||x_0 - x^*||.$$