Name: Hongda Li
Class: AMATH 515 Winter 2021
HW1: Theoretical Portion

Notations and Rules I used:

1. $f^E(x)$ where $x \in \mathbb{R}^n$ is the vectorized function, defined as:

$$f^E(x) := \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}$$

2. Taking composite scalar function ($f(x) : \mathbb{R}^n \mapsto \mathbb{R}$) with a Multi-Variable function ($G : \mathbb{R}^m \mapsto \mathbb{R}^n$) wrt to a certain variable:

$$\partial_{xi}(f(G(x))) = \left(\frac{\partial G(x)}{\partial x_i}\right)^T \nabla f(G(x))$$

3. Disambiguate the way we use ∇. $\nabla[f]$ means taking the gradient of f, and $\nabla[f(g(x))]$ means take the gradient on the composition, which is $\nabla f \circ g(x)$. However $\nabla f(g(x))$ means putting $g(x)$ as input for the gradient of f.

4. Take derivative of composition of multiple-input and output functions, which is used for briefly on deriving the Hessian of the Logistic objective. Let F be $\mathbb{R}^m \mapsto \mathbb{R}^n$ and G be $\mathbb{R}^K \mapsto \mathbb{R}^m$.
Assuming that we are taking the derivative on the i th input wrt the whole multi-variable function, which should give us a gradient.

$$\frac{\partial F(G(x))}{\partial x_i} = \nabla F(G(x)) \left(\frac{\partial G(x)}{\partial x_i}\right)$$

Where we take the Jacobian of F, put in $F(G(x))$ and them multiply it by the vector obtained by taking G wrt to x_i.

Problem 1

Let $g : \mathbb{R}^m \to \mathbb{R}$ is a twice differentiable function, $A \in \mathbb{R}^{m \times n}$ any matrix, and h is the composition $g(Ax)$,.

(a)

Show $\nabla h(x) = A^T \nabla g(Ax)$ Consider taking derivative with respect to one of the input variable on $h(x)$:

$$\partial_{xi} h(x) = \partial_{xi} g(Ax)$$

Using rule 1, we have:

$$\left(\frac{\partial Ax}{\partial_{xi}}\right)^T \nabla g(Ax)$$

Take notice that, the first part of the expression is taking derivative on a linear function, therefore the derivative is just the i th column of the matrix, where we denoted it as: $A_{\text{col}(i)}$, hence:

$$(A_{\text{col}(i)})^T \nabla g(Ax)$$

Take notice that, the gradient is all the differential with respect to all the variable stacked together into a vector, therefore we have:

$$\nabla[g(Ax)] = \begin{bmatrix} \partial_{x1}(g(Ax)) \\ \partial_{x2}(g(Ax)) \\ \vdots \\ \partial_{x_n}(g(Ax)) \end{bmatrix} = \begin{bmatrix} (A_{\text{col}(1)})^T \nabla g(Ax) \\ (A_{\text{col}(2)})^T \nabla g(Ax) \\ \vdots \\ (A_{\text{col(n)}})^T \nabla g(Ax) \end{bmatrix} = A^T \nabla g(Ax)$$

(b)

Show that $\nabla^2 h(x) = A^T \nabla^2 g(Ax) A$.
Note that there are multiple outputs and multiple input, so for each output, we are taking it wrt a particular input variable, resulting in the Jacobian of the gradient.
Consider taking the j th output of the function wrt to the i th variable x_i:

$$\partial_{xi} \left[(A^T \nabla g(Ax))_j\right]$$

With matrix multiplication, the j th output is the j th row multiplied multiplied by the vector:

$$\partial_{xi} \left[(A^T)_{\text{row(j)}} \nabla g(Ax)\right] = \partial_{xi} \left[(A)^T_{\text{col(j)}} \nabla g(Ax)\right]$$

Notice that, the differential operator can move pass the vector dot product because the vector consists of only constant.

$$(A_{\text{col(j)}})^T \frac{\partial \nabla g(Ax)}{\partial_{xi}} = (A_{\text{col(j)}})^T \left(\nabla^2 g(Ax) \left(\frac{\partial Ax}{\partial_{xi}}\right)\right)$$

We use the rule 4 to take the derivative and get the Hessian $\nabla^2 g$, and the last term in the expression is constant, therefore we get out one of the column of the matrix for that:

$$(A_{\text{col(j)}})^T \left(\nabla^2 g(Ax) \left(A_{\text{col(i)}}\right)\right)$$

And the above expression is the element for in position i, j, therefore, the whole matrix can be compactly written as:

$$\nabla^2 h(x) = A^T \nabla^2 g(Ax) A$$

(c)

Compute the gradient and Hessian for the scalar function:

$$\sum_{i=1}^{m} \log(1 + \exp(a_i^T x)) - b^T A x$$

Where a_i^T denotes the i th row of the matrix A. The above formulation can be written as a vector notation (We keep the notation, using $g(x)$ to represent the function we are taking the gradient and Hessian of.):

$$g(Ax) = J_m^T \log(1 + \exp^E(Ax)) - b^T A x$$

Where the vector J_n is a vector full of ones with length n. The function can be implicitly written wrt to x, and we have:

$$g(x) = J_m^T \log(1 + \exp^E(x)) - b^T x$$

Then the gradient is just:

$$\nabla g(x) = \frac{\exp^E(x)}{1 + \exp^E(x)} - b$$

Therefore, we can use the formula we had and get the **Gradient:**

$$\nabla[g(Ax)] = \nabla[Ax]^T \nabla g(Ax) = A^T \frac{\exp^E(Ax)}{1 + \exp^E(Ax)} - A^T b$$

And now we can write the gradient in a big vector and then find the Jacobian of the Gradient.

$$\nabla[g(Ax)] = A^T \begin{bmatrix} \frac{\exp(a_1^T x)}{1+\exp(a_1^T x)} - b_1 \\ \frac{\exp(a_2^T x)}{1+\exp(a_2^T x)} - b_2 \\ \vdots \\ \frac{\exp(a_m^T x)}{1+\exp(a_m^T x)} - b_m \end{bmatrix}$$

Notice that, each of the row is only associated wit one element, therefore the Jacobian of the Gradient (Hessian) is going to be a diagonal matrix:

$$\nabla^2 g(x) = \begin{bmatrix} \frac{\exp(x_1)}{(1+\exp(x_1))^2} & & & \\ & \frac{\exp(x_2)}{(1+\exp(x_2))^2} & & \\ & & \ddots & \\ & & & \frac{\exp(x_n)}{(1+\exp(x_n))^2} \end{bmatrix}$$

And here is how I the derivative is computed:

$$\partial_x \left(\frac{\exp(x)}{1 + \exp(x)} \right) = \left(\frac{\exp(x)}{1 + \exp(x)} - \frac{(\exp(x))^2}{(1 + \exp(x))^2} \right) = \frac{\exp(x)}{(1 + \exp(x))^2}$$

And using the formula we had earlier, we know that:

$$\nabla^2[g(Ax)] = A^T \begin{bmatrix} \frac{\exp(a_1^T x)}{(1+\exp(a_1^T x))^2} & & & \\ & \frac{\exp(a_2^T x)}{(1+\exp(a_2^T x))^2} & & \\ & & \ddots & \\ & & & \frac{\exp(a_n^T x)}{(1+\exp(a_n^T x))^2} \end{bmatrix} A$$

Problem 2

(a)

Prove that the convex indicator function is a convex function. The convex indicator function is:

$$\delta_c(x) := \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases} \tag{2.a.1}$$

Where C is a convex set.

Proof Strategy: We prove that the epigraph of the function is another convex set, therefore by definition of convexity, the function is convex.

By definition of epigraph, we have that:

$$\text{epi}(\delta_C) = \left\{ \begin{bmatrix} \alpha \\ x \end{bmatrix} : \alpha \geq \delta_C(x) \right\} \tag{2.a.2}$$

Firstly, the second element in the vector will has to be in the convex set C for it to be in the epigraph of the function because:

$$x \notin C \implies \forall \alpha : \alpha < \infty \implies \begin{bmatrix} \alpha \\ x \end{bmatrix} \notin \text{epi}(\delta_C)$$

Therefore, it has to be the case that $x \in C$.

Now that we have the epigraph of the function well-defined, we choose 2 arbitrary elements from it and prove that their convex combination is in the epigraph too.

let the 2 arbitrary elements be:

$$\forall \begin{bmatrix} \alpha \\ x \end{bmatrix}, \begin{bmatrix} \beta \\ y \end{bmatrix} \in \text{epi}(\delta_C) : \alpha, \beta \geq 0$$

This is by the definition of the epigraph of the function. Therefore:

$$\alpha, \beta \geq 0 \implies \alpha + \lambda(\beta - \alpha) \geq 0 \quad \forall \lambda \in [0, 1]$$

At the same time:

$$x, y \in C \implies x + \lambda(y - x) \in C$$

By the definition of the convexity of the graph too. Therefore, we have the statement that:

$$\begin{bmatrix} \alpha \\ x \end{bmatrix} + \lambda \begin{bmatrix} \beta - \alpha \\ y - x \end{bmatrix} \in \text{epi}(\delta_C)$$

Therefore, the epigraph of the function is a convex set, and using the definition for the convexity of function, we know that the convex indicator function is a convex function as well.

(b)

Show that the support function of any set is a convex function. The support function is defined as:

$$\sigma_C(x) = \sup_{c \in C} \{c^T x\}$$

Using the definition of the convex function, we need to show that:

$$\sigma_C(x + \lambda(y - x)) \leq \sigma_C(x) + \lambda \sigma_C(y - x) \tag{2.b.1}$$

Notice that by definition of the support function:

$$\sigma_C(x + \lambda(y - x)) = \sup_{C} \{c^T x + \lambda c^T (y - x)\}$$

And notice that, by the properties of the supremum, we know that:

$$\sup_{c \in C} \{c^T x + \lambda c^T (y - x)\} \leq \sup_{c \in C} \{c^T x\} + \lambda \sup_{c \in C} \{c^T (y - x)\}$$

And substituting with the definition of the function, we will reveal that the convexity definition and it's the same as 2.b.1.

(c)

There are 2 important properties of norm that will show that norm balls are convex functions:

(1) $\|x + y\| \leq \|x\| + \|y\|$

(2) $\|\alpha x\| \leq |\alpha| \|x\|$

Consider the expression:

$$\|x + \lambda(y - x)\|$$

Using the properties above we know that:

$$\|x + \lambda(y - x)\| \leq \|x\| + |\lambda| \|y - x\|$$

And therefore, we have show that the function is convex, because this is the definition of a convex function.

Problem 3

(a)

Let $f(x) := -x$, which is linear, and linear functions are convex, now let $g(x) := 0$ which is convex, because it satisfies the **Gradient inequality**. However, their composite is not a convex function because:

$$f(g(x)) = -x^2$$

This is not convex because the line with both end points lies on $(1, \pm 1)$ is below the function, and therefore, invalidates the convex definition.

(b)

Claim: "If $f(x)$ is convex and **non-increasing**, and $g(x)$ is concave, then $f(g(x))$is a convex function."

Proof strategies: We are going to use the first differential characterization of the convex function to prove it.

By the hypothesis that $f(x)$ is convex and the range of $g(x)$ falls into the domain of function $f(x)$, we know that:

$$f(g(y)) - f(g(x)) \geq f'(g(x))(g(y) - g(x)) \quad \textbf{(3.b.1)}$$

Now with the assumption that $g(x)$ is concave, which means that $-g(x)$ is convex, which means that swapping the inequality sign for the convexity for $g(x)$ will give us the correct statement:

$$g(y) - y(x) \leq g'(x)(y - x)$$

By the assumption that $f(x)$ is a non-increasing function, we know that $f(x) \leq 0$ for all x in the domain, which means multipliying $f'(g(x))$ on the above equation will swap the inequality giving us:

$$f'(g(x))(g(y) - g(x)) \geq f'(g(x))g'(x)(y - x) \tag{3.b.2}$$

Chaining it with expression (3.b.1) will give us the desire inequality:

$$f(g(y)) - f(g(x)) \geq f'(g(x))g'(x)(y - x)$$

Therefore, the composite of these 2 functions is a convex function.
Corollary: If the function $f(x)$ is convex and non-decreasing, and function $g(x)$ is convex, then the composite, $f(g(x))$ is convex. If $g(x)$ is convex, then:

$$g(y) - y(x) \geq g'(x)(y - x)$$

And at the same time, because $f(x)$ is non decreasing, we have:

$$f'(g(x))(g(y) - g(x)) \geq f'(g(x))g'(x)(y - x)$$

And we arrive at the sane place as 3.b.2, the composite function is convex.

(c)

Claim: "$f : \mathbb{R}^m \mapsto \mathbb{R}$ is convex, $g : \mathbb{R}^n \mapsto \mathbb{R}^m$" is affine linear, then $f(g(x))$ is convex. $g(x)$ being affine linear means that:

$$g(x + h) = g(x) + \nabla g(x)h \tag{3.c.1}$$

Using the first statement of differential characterization on convex functions on f, we have:

$$f(\alpha + \beta) \geq f(\alpha) + \nabla f(\alpha)^T \beta \tag{3.c.2}$$

Using 3.c.2 and 3.c.1we can say that:

$$f(g(x+\lambda(y-x))) = f(g(x)+\lambda\nabla g(x)(y-x)) \geq \nabla f(g(x))^T(\lambda\nabla g(x)(y-x)) \tag{3.c.3}$$

Take note that, the above equation if true for all $\lambda \in [0,1]$ and $\nabla g(x)$ denotes the Jacobian of the function $g(x)$. Substituting $\lambda = 1$ into 3.c.3, we have:

$$f(g(y)) \geq f(g(x)) + \nabla f(g(x))^T \nabla g(x)(y-x)$$

And this is the definition of a convex function. $f(g(x))$ is a convex function.

(d)(i)

Claim(3.d.i.1): The function: $\sum_{i=1}^{n} \log(1+\exp(a_i^T x)) - b^T Ax$ is a convex function.
Proof strategies The function is the sum of composite function of an affine linear function and a convex function, therefore the whole function is convex.
Claim (3.d.i.2) The second derivative of the logistic objective function $\log(1+\exp(x))$ is positive definite. This is true because the second derivative is:

$$\frac{\exp(x)}{(1+\exp(x))^2} \geq 0 \quad \forall x$$

And therefore, for each of the term in the summation, we have: $\log(1+\exp(a_i^T x))$ is convex (by 3(c)). Using the fact that sum of convex functions are convex (see d.ii.3), the function is convex.

(d)(ii)

Claim: The function $\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax$ is a convex function.
Proof strategies: The function is a sum over a lot of functions, and all of them are convex.
Claim (d.ii.1): $b^T Ax$ is convex, this is true because it's a linear function.
Claim (d.ii.2): $\exp(a_i^T x)$ is a convex function.
This is true because, $a_i^T x$ is a linear function, hence it's convex, and $\exp(x)$ is a convex and non-decreasing function (by second derivative of e^x) is positive, and the composite of a non-decreasing convex function with a convex function is convex (by part corollary from part (b)).
Claim (d.ii.3): Sum of convex function is still convex.

$$f_i(x+\lambda(y-x)) \leq f_i(x) + \lambda f_i(y-x) \quad \forall 1 \leq i \leq n$$
$$\sum_{i=1}^{n} f_i(x+\lambda(y-x)) \leq \sum_{i=1}^{n} f_i(x) + \lambda f_i(y-x)$$

This is just by the properties of the inequalities.
By claims (d.ii.1)(d.ii.2)(d.ii.3), we proved that the first claim, that the Poisson objective function is a convex function.

Problem 4

A function is strictly convex if:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \quad \forall \lambda \in (0,1) \tag{def1}$$

(a)

The function e^x is strictly convex because it's second derivative is strictly positive and it's a C2 smooth function. This function doesn't have a minimizer because it decreases monotonicall as $x \to -\infty$.

(b)

Claim: (4.b.1) The sum of strictly convex functions and convex function is strictly convex.
Proof Strategies Direct Proof.
Let $f(x)$ and $g(x)$ be convex and strictly convex, then:

$$f(x + \lambda(y-x)) \leq f(x) + \lambda f(y-x) \tag{4.b.2}$$

$$g(x + \lambda(y-x)) < g(x) + \lambda g(y-x) \tag{4.b.3}$$

Using the property of inequality, summing up 4.b.2 and 4.b.3 will get rid of the equality case, giving us:

$$f(x + \lambda(y-x)) + g(x + \lambda(y-x)) < f(x) + \lambda f(y-x) + g(x) + \lambda g(y-x)$$

And that is the definition of the sum of the 2 function being convex, proving the claim.

(c)

Task (4.c.1) Characterize all solution to the problem:

$$\min_x \frac{1}{2}\|Ax-b\|^2 \tag{4.c.1}$$

Strategies: We will need to show that the function is convex, and then look for the local minimizer/minimizers. and use the Corollary of Characterization of Convexity, where it stated that, a local minimizer is the global minimizer.
Claim (4.c.2): The function is convex and its gradient can be zero (by Corollary of Characterization of Convexity), the function has minimizers.
This claim is true because Norm balls are convex and non-decreasing, and affine linear function is convex, in this case, the function is the norm of a linear affine function, and there composite is convex.
Computing Gradient Simplify:

$$\begin{aligned} \|Ax-b\|^2 &= (Ax-b)^T(Ax-b) \\ &= \|Ax\|^2 - 2(Ax)^T b + \|b\|^2 \end{aligned} \tag{4.c.3}$$

Take notice that, the objective function can be view as a composite function (Ignoring the constant at the front which doesn't change the minimization problem.):

$$g(Ax) = \|Ax\|^2 - 2(Ax)^T b + \|b\|^2 \tag{4.c.4}$$

Which it implies that:

$$g(x) = \|x\|^2 - 2x^T b + \|b\|^2 \tag{4.c.5}$$

And in this case, it's not hard to figure out the gradient as:

$$\begin{aligned} \nabla[g(x)] &= \nabla g(x) = 2x - 2b \\ \nabla[g(Ax)] &= A^T(2Ax - 2b) \\ \nabla[g(Ax)] &= 2A^T Ax - 2b \end{aligned} \tag{4.c.6}$$

Looking for Minimizer: One of the necessary condition for convex optimality is Gradient equals to zero, which means that:

$$A^TAx - A^Tb = \mathbf{0} \tag{4.c.7}$$
$$A^TAx = A^Tb$$

Therefore, the minimizer is $x^* = A^\dagger b$. By corollary of Different Characterization of convexity, these set of minimizers are also the global minimizer because the function is convex by 4.c.2. However there are 2 cases:

1. The matrix A is full rank. In this case the matrix A^TA is invertible and it will have infinitely many solution.
2. The matrix A is not full-ranked, in this case there will be infinitely many solution, hence, infinitely many minimizers.

(d)

Task: Characterize the minimizers of the following equation(if there is any.):

$$\min_x \left(\sum_{i=1}^{n} \left(\log(1 + \exp(a_i^T x)) \right) + \lambda \|x\|_1 + (1-\gamma)\|x\|^2 \right) \quad \lambda > 0, \gamma \in (0,1) \tag{4.d.0}$$

Claim: 4.d.1 "If function $g(x)$ is convex then $g(x) + \gamma\|x\|^2/2$ is an alpha strongly convex function.", This statement is left as an exercise in lecture 4 (It's going to be proved here).

Let $g(x)$ be convex, now let's define $f(x)$ to be:

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

And consider:

$$f(x) + \nabla f(x)^T(y-x) + \frac{\alpha||y-x||^2}{2} \tag{4.d.2}$$
$$g(x) + \frac{\alpha\|x\|}{2} + (\nabla g(x)^T + \alpha x)(y-x) + \frac{\alpha}{2}||y-x||^2$$
$$\left[g(x) + \nabla g(x)^T(y-x)\right] + \left[\frac{\alpha||x||^2}{2} + \alpha x^T(y-x) + \frac{\alpha}{2}||y-x||^2\right]$$

Let's focus what has been written inside the bracket and we have:

$$\frac{\alpha||x||^2}{2} + \alpha x^T(y-x) + \frac{\alpha}{2}||y-x||^2 \tag{4.d.3}$$
$$\frac{\alpha||x||^2}{2} + \alpha x^T(y-x) + \frac{\alpha}{2}\left(\|y\|^2 + \|x\|^2 - 2x^Ty\right)$$
$$\alpha\|x\|^2 - \alpha||x||^2 + \alpha x^Ty + \frac{\alpha}{2}\left(\|y\|^2 - 2x^Ty\right)$$
$$\frac{\alpha}{2}||y||^2$$

Therefore, 4.d.2 can be simplified into:

$$f(x) + \nabla f(x)^T(y-x) + \frac{\alpha||y-x||^2}{2} = \left[g(x) + \nabla g(x)^T(y-x)\right] + \left[\frac{\alpha||y||^2}{2}\right] \tag{4.d.4}$$

However at the same time, we know that the function $g(x)$ is convex, using the definition of its convexity we have:

$$g(y) \geq \nabla g(x) + g(x)^T(y-x) \tag{4.d.5}$$
$$g(y) + \frac{\alpha||y||^2}{2} \underset{(1)}{\geq} \left[g(x) + \nabla g(x)^T(y-x)\right] + \left[\frac{\alpha||y||^2}{2}\right]$$
$$g(y) + \frac{\alpha||y||^2}{2} \geq f(x) + \nabla f(x)^T(y-x) + \left[\frac{\alpha||y-x||^2}{2}\right]$$

(1): by substituting 4.d.4

Setting the α to be $\alpha \leq (1-\gamma)$, which is the regularization coefficient, we have:

$$f(x) \geq f(x) + \nabla f(x)^T (y-x) + \left[\frac{\alpha ||y-x||^2}{2}\right]$$

Where:

$$f(x) = g(y) + \frac{(1-\gamma)||y||^2}{2}$$

Therefore the function $f(x)$ is strongly convex.

Claim: 4.d.2 "A strongly convex function has a unique minimizer". This claim is left as a statement in lecture 4.

From problem 3(d) we know that the logistic regression is a convex function.

From problem 2(c) we know that norm with $p \geq 1$ are all convex function.

Therefore the following function $g(x)$ is going to be convex:

$$g(x) := \sum_{i=1}^{n} \left(\log(1 + \exp(a_i^T x))\right) + \lambda \|x\|_1$$

Using claim **4.d.1** and **4.d.2** and $g(x)$ to be the logistic objective we know that the logistic regression with 2-Norm Regularization is a Alpha Strongly Convex function.

Note: There might be a typo in the HW where the $-b^T Ax$ term is missing for problem 4(d), but also notice that with the additional term $-b^T Ax$ inserted into the definition for $g(x)$, the convexity remains unchanged and the same argument follows.

Problem 5

For the whole problem 5, I assume that we are using the 2-norm induced by vector for matrices.

(a)

Task: Find a global bound for β of the least-square objective $\|Ax - b\|^2$
The gradient of the lest-square objective is:

$$A^T(Ax - b) \tag{5.a.1}$$

This gradient of the objective is Lipschitz Continuous because:

$$\begin{aligned} &\|A^TAy - b - \left(A^Tx - b\right)\| \\ &\|A^TA(y - x)\| \\ &\leq \|A^TA\|\|y - x\| \end{aligned} \tag{5.a.2}$$

Under the use of 2-norm, we know that $\|A\|_2$ is the largest singular value of the matrix, and in the case of Symmetric Matrix A^TA, the largest eigenvalue is the largest singular value of the matrix.

Therefore the upper bound for β is:

$$\beta = \max_i(|\lambda_i|) = \|A^TA\| \tag{5.a.3}$$

Where λ_i are the eigenvalues of the matrix A^TA also the singular value of the matrix too.

(b)

Task: Find the β for the regularized logistic objective:

$$\sum_{i=0}^{m}\left(\log(1 + \exp(a_i^Tx))\right) - b^TAx + \frac{\lambda}{2}\|x\|^2$$

Where a_k^T are rows of a $m \times n$ matrix.
Theorem 2: f is C^2 smooth if and only if $\beta I \succeq \nabla^2 f(x)$ and β will be the bound for β smoothness.
Let D denotes the diagonal matrix we derived for the Hessian in problem 1, let $g(Ax)$ denotes the objective without regularization.

$$D = \text{diag}\left(\frac{\exp^E(x)}{(1 + \exp^E(x))^2}\right)$$

Then the Hessian of the objective will be:

$$\nabla^2[g(Ax) + \frac{\lambda}{2}\|x\|^2] = A^TDA + \lambda I \tag{5.b.1}$$

And at this point, it's better to figure out the 2-norm of A^TDA so we have an upperbound for the maximal absolute eigenvalue (Matrix is symmetric too). Here we assume that the matrix A has SVD decomposition $U\Sigma V^T$.

$$\begin{aligned} &\|A^TDA\| = \|V\Sigma U^TDU\Sigma V^T\| \\ &\|A^TDA\| \leq \|\Sigma^2\|\|D\| \\ &\|\Sigma^2\|\|D\| \underset{(1)}{=} \max_i|\sigma_i|^2\|D\| \quad \text{Where } \sigma_i \text{ are singular values of } A \end{aligned} \tag{5.b.2}$$

(1): Induced 2 norm is the maximal absolute Eigenvalue of $\sqrt{\Sigma^4}$, therefore it's the maximal singular value squared; which is also the same as the maximum absolute value of the eigenvalues of A^TA
Here, I should point out that:

$$\sup_{x\in\mathbb{R}}\left(\frac{\exp(x)}{(1 + \exp(x))^2}\right) = \frac{1}{4}$$

By this point we have the inequality:

$$\|A^T DA\| = \frac{\max_i |\sigma_i|^2}{4} \tag{5.b.3}$$

Therefore, continue with 5.b.1 we have inequality:

$$\begin{aligned} \|A^T DA + \lambda x\| &\leq \|A^T DA\| + \lambda \\ &\leq \frac{\max_i |\sigma_i|^2}{4} + \lambda \\ \beta = \frac{\max_i |\sigma_i|^2}{4} + \lambda &= \frac{\max_i(\lambda_i)}{4} + \lambda \end{aligned} \tag{5.b.4}$$

Where λ_i are the eigenvalues of matrix $A^T A$ and this is one of the possible value for β.

(c)

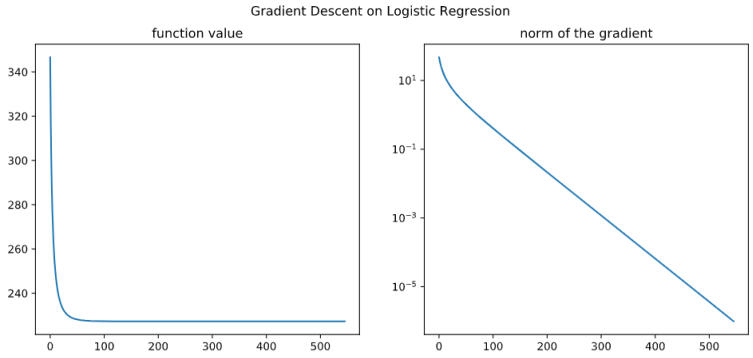
The poisson objective function:

$$\sum_{i=1}^{n} \left(\exp(a_i^T x)\right) - b^T Ax$$

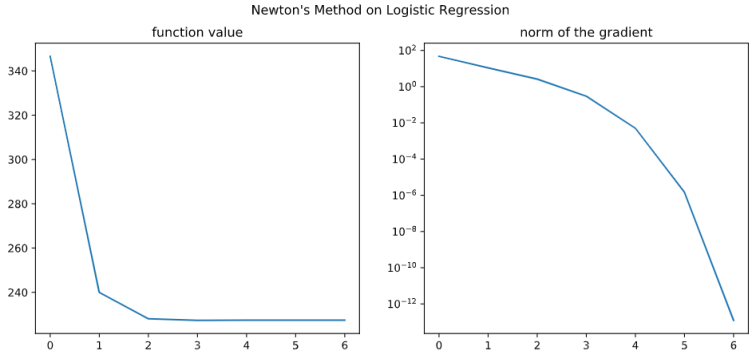
Which is not Liptschitz continuous because exponential function has exponential function as its derivative and exponential function cannot be bounded globally with a constant β. In the case of Poisson objective, it's manifested as unbounded Eigenvalues for the Hessian of the exponential function.

Problem 6

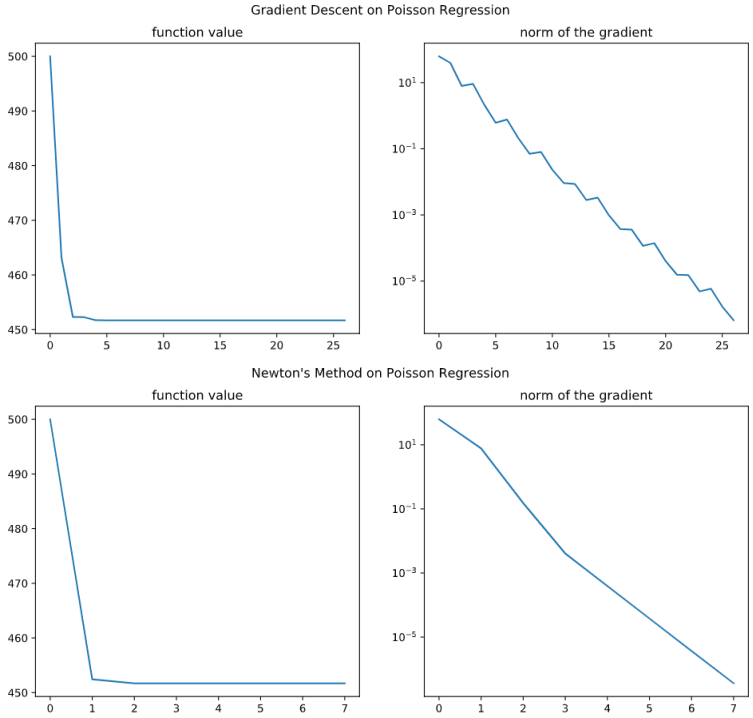
(a)



(b)



(c)



(d)

The steepest descend method is used for functions that are non beta smooth. In our HW coding part, the Poisson Objective function is not beta smooth. Qualitatively, The newton's method and the Gradient Descend with Backtracking Line search has the same Linear Convergence rate for the poisson Objective. However for the Logistic Objective, the Newton's method's convergence rate is quadratic.