Name: Anna Huang

AMATH 515 Homework Set 1

**Due: Monday Jan 23rd, by midnight.**

(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 $g(Ax)$, then we have two simple generalizations of the chain rule that combine linear algebra with calculus:

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

and

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

(a) Show what happens when you apply the above chain rules to the special case

$$h(x) = g(a^T x)$$

where $a$ is a vector.

**Answer to 1a:** We have $A = a^T$ and the $\nabla a^T x = \vec{a}$.

$$h(x) = g(a^T x)$$
$$\frac{dh(x)}{dx} = g'(a^T x)\frac{d(a^T x)}{dx}$$
$$\frac{dh(x)}{dx} = g'(Ax)\frac{d(Ax)}{dx}$$
$$\nabla h(x) = A^T \nabla g(Ax)$$

(b) Compute the gradient and hessian of the regularized logistic regression objective:

$$\left(\sum_{i=1}^{n} \log(1 + \exp(a_i^T x)) - b^T Ax\right) + \lambda \|x\|^2$$

where $a_i$ denote the rows of $A$.

**Answer to 1b:** Using the result from 1(a) to find the gradient of the first log term is:

$$\nabla g(a_i^T x) = \frac{\nabla \exp(a_i^T x)}{1 + \exp(a_i^T x)}$$
$$\nabla g(a_i^T x) = \frac{a_i \exp(a_i^T x)}{1 + \exp(a_i^T x)}$$

because the $\nabla a_i^T x = a_i$. Now the same thing applies to the second term $b^T z$ and using the result from 1(a):

$$\nabla h(z) = A^T \nabla g(Az)$$
$$\nabla h(b^T x) = A^T \nabla g(b^T x)$$

$$\nabla h(b^T x) = A^T b$$

and lastly, the third term:

$$\lambda||x||^2 = \lambda(x_1^2 + x_2^2 + ... + x_n^2)$$
$$\nabla\lambda||x||^2 = \nabla\lambda(x_1^2 + x_2^2 + ... + x_n^2)$$
$$\nabla\lambda||x||^2 = 2\lambda x$$

Therefore the gradient is:

$$\sum_{i=1}^{n} \frac{a_i \exp(a_i^T x)}{1 + \exp(a_i^T x)} - A^T b + 2\lambda x$$

To calculate the hessian, we use $\nabla^2 h(x) = A^T \nabla^2 g(Ax) A$. The gradient of the first term using the quotient rule becomes:

$$\nabla^2 g(a_i^T x) = \nabla(\frac{a_i \exp(a_i^T x)}{1 + \exp(a_i^T x)})$$

$$\nabla^2 g(a_i^T x) = \frac{a_i a_i^T \exp(a_i^T x)}{(1 + \exp(a_i^T x))^2}$$

The second term is $\nabla A^T b = 0$ and the third term is $\nabla 2\lambda x = 2\lambda I$ resulting in the hessain as:

$$\sum_{i=1}^{n} \frac{a_i \exp(a_i^T x) a_i^T}{(1 + \exp(a_i^T x))^2} + 2\lambda I$$

.

(c) Compute the gradient and hessian of the regularized poisson regression objective:

$$\left(\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax\right) + \lambda\|x\|^2$$

where $a_i$ denote the rows of $A$.

**Answer to 1c:** Using the result from 1(a) to find the gradient of the first log term is:

$$\nabla g(a_i^T x) = \nabla \exp(a_i^T x)$$
$$\nabla g(a_i^T x) = a_i \exp(a_i^T x)$$

because the $\nabla a_i^T x = a_i$. The second and third terms are the same as problem 1b, so combining them with the first term from 1c is given by:

$$\sum_{i=1}^{n} a_i \exp(a_i^T x) - a_i b_i + 2\lambda x$$

To calculate the hessian, we use $\nabla^2 h(x) = A^T \nabla^2 g(Ax)A$ and similarly to 1b: the second term is $\nabla A^T b = 0$ and the third term is $\nabla 2\lambda x = 2\lambda I$ resulting in:

$$\sum_{i=1}^{n} a_i a_i^T \exp(a_i^T x) + 2\lambda I$$

(d) Compute the gradient and hessian of the regularized 'concordant' regression objective

$$\|Ax - b\|_2 + \lambda\|x\|_2.$$

Give conditions on a point $x$ that ensure the gradient and Hessian exist at $x$.

**Answer to 1d:** Using the result from 1(a) and composition equations to find the gradient and hessian of the first term
$||Ax - b||_2 = \sqrt{(Ax_1 - b_1)^2 + (Ax_2 - b_2)^2 + (Ax_3 - b_3)^2 + ...(Ax_n - b_n)^2}$. The gradient is

$$h(x) = ||Ax - b||$$
$$g(z) = ||z = b||$$
$$g(z) = \sqrt{(z_1 - b_1)^2 + (z_2 - b_2)^2 + (z_3 - b_3)^2 + .... + (z_n - b_n)^2}$$
$$\nabla g(z) = \frac{z-b}{||z-b||}$$
$$\nabla h(x) = A^T g(Ax) \text{ with } g(Ax) = g(z)$$
$$\nabla h(x) = A^T \frac{z-b}{||z-b||}$$
$$\nabla h(x) = A^T \frac{Ax-b}{||Ax-b||}$$

with $h(x) = g(Ax)$. Using the quotient rule to find $\nabla g^2(z)$, the hessian is different depending on $x_{mn}$ of the matrix. The diagonal is where it is differentiated with respect to $x_{mn}$ twice: $\frac{dg^2(z)}{dz_n dz_n}$ with $z = Ax$. The diagonal Hessian is the following:

$$\nabla g^2(z) = \frac{||z-b||^2}{||z-b||^3} - \frac{(z-b)^2}{||z-b||^3}$$
$$\nabla g^2(z) = \frac{1}{||z-b||} - \frac{(z-b)^2}{||z-b||^3}$$
$$\nabla h(x) = A^T(\frac{1}{||Ax-b||} - \frac{(Ax-b)^2}{||Ax-b||^3})A$$

The non-diagonal Hessian is the following, with $m$ representing the row term and $n$ representing the column term:

$$\nabla g^2(z) = \frac{d^2g(z)}{dx_m dx_n}$$
$$\nabla g^2(z) = -\frac{(z_m-b_m)*(z_n-b_n)}{||z-b||^3}$$

An example of the (1,2) term would be $\nabla g^2(z) = -\frac{(z_1-b_1)*(z_2-b_2)}{||z-b||^3}$.

The resulting Hessian matrix for the first term $||Ax-b||_2$ with $z = Ax$ is:

$$H1 = A^T \begin{bmatrix} \frac{1}{||z_1-b_1||} - \frac{(z_1-b_1)^2}{||z_1-b_1||^3} & -\frac{(z_1-b_1)*(z_2-b_2)}{||z-b||^3} & \cdots \\ -\frac{(z_1-b_1)*(z_2-b_2)}{||z-b||^3} & \frac{1}{||z_2-b_2||} - \frac{(z_2-b_2)^2}{||z_2-b_2||^3} & \cdots \\ \vdots & \ddots & \\ -\frac{(z_m-b_m)*(z_n-b_n)}{||z-b||^3} & & \frac{1}{||z_n-b_n||} - \frac{(z_n-b_n)^2}{||z_n-b_n||^3} \end{bmatrix} A$$

Using the result from 1(a) and composition equations to find the gradient and hessian of the second term $\lambda||x||_2 = \sqrt{(x_1)^2 + (x_2)^2 + (x_3)^2 + ...(x_n)^2}$. The gradient is

$$f(x) = \lambda||x||_2 = \sqrt{(x_1)^2 + (x_2)^2 + (x_3)^2 + ...(x_n)^2}$$
$$\nabla f(x) = \lambda\frac{x}{||x||}$$

Using the quotient rule to find $\nabla f(x)$, the hessian is different depending on $x_{mn}$ of the matrix. The diagonal is where it is differentiated with respect to $x_{mn}$ twice: $\frac{df^2(x)}{dx_n dx_n}$. The diagonal Hessian is the following:

$$\nabla^2 f(x) = \frac{1}{||x||} - \frac{x^2}{||x||^3}$$

and the non-diagonal Hessian is the following, with $m$ representing the row term and $n$ representing the column term:

$$\nabla^2 f(x) = 0 - \frac{x_n * x_m}{||x||^3}$$

An example of the (1,2) term would be $\nabla^2 f(x) = 0 - \frac{x_1 * x_2}{||x||^3}$.

The resulting Hessian matrix for the second term $\lambda||x||_2$ is:

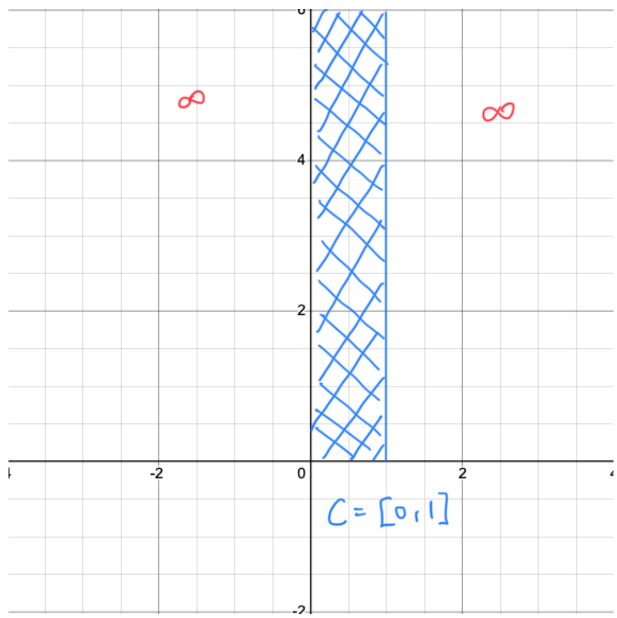
$$H2 = \begin{bmatrix} \frac{1}{||x||} - \frac{x_1^2}{||x||^3} & -\frac{x_1*x_2}{||x||^3} & \cdots \\ -\frac{x_1*x_2}{||x||^3} & \frac{1}{||x||} - \frac{x_2^2}{||x||^3} & \cdots \\ \vdots & \ddots & \\ -\frac{x_n*x_m}{||x||^3} & & \frac{1}{||x||} - \frac{x_n^2}{||x||^3} \end{bmatrix}$$

As a result, the gradient of the regularized 'concordant' regression objective is $A^T \frac{Ax-b}{||Ax-b||} + \lambda \frac{x}{||x||}$. The hessian of the regularized 'concordant' regression objective is the sum of the two matrices H1 + H2 from above. In order for the gradient and hessian of $x$ to exist, the denominator can not be 0, so $||Ax - b|| > 0$ and $||x|| > 0$ since the norms can not be negative either.

(2) Show that each of the following functions is convex.

(a) Indicator function to a convex set: $\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ \infty & \text{if} \quad x \notin C. \end{cases}$

**Answer to 2a:** Assuming that C is a convex set between C [0,1] and according to the epigraph below, all of the above means convexity. Anything outside of the epigraph, the function is taking the value to $\infty$.




(b) Support function to any set:

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

**Answer to 2b:** The support function is the supremum of $c^T x$. From the definition defined in the class, the sup of arbitrary many convex functions is convex. Since $c^T x = c_1x_1 + c_2x_2 + c_3x_3$ is linear, it is convex. As a result, the supremum of these linear functions is convex.

(c) Any norm (see Chapter 1 for the definition of a norm).

**Answer to 2c:** According to Chapter 1: a norm on a vector space holds three properties for all points $x, y \in V$ and scalars $a \in \mathbf{R}$.

(i) Absolute homogeneity: $||ax|| = |a| \cdot ||x||$

(ii) Triangle inequality: $||x + y|| <= ||x|| + ||y||$

(iii) Positivity: Equality $||x|| = 0$ holds if and only if $x = 0$

Based on the absolute homogeneity and triangle inequality properties,

$$
\begin{aligned}
||\lambda x + (1-\lambda)y|| &<= ||\lambda x|| + ||(1-\lambda)y|| \\
||\lambda x + (1-\lambda)y|| &<= \lambda||x|| + (1-\lambda)||y||
\end{aligned}
$$

Therefore, it concludes that every norm is convex simply based on the triangle inequality.

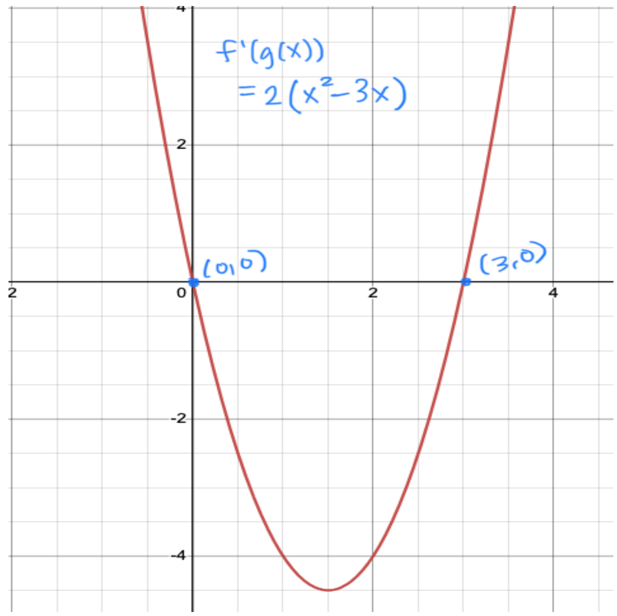
(3) Convexity and composition rules. Suppose that $f$ and $g$ are $\mathcal{C}^2$ functions from $\mathbb{R}$ to $\mathbb{R}$, with $h = f \circ g$ their composition, defined by $h(x) = f(g(x))$.

(a) If $f$ and $g$ are convex, show it is possible for $h$ to be nonconvex (give an example). Give additional conditions that ensure the composition is convex.

**Answer to 3a:** If $f$ is convex and $g$ is concave, and in order for $h(x)$ to be convex, then $h''(x) > 0$.

$$
\begin{aligned}
h(x) &= f(g(x)) \\
h''(x) &= f'(g(x))g'(x) \\
h''(x) &= f'(g(x))g''(x) + f''(g(x))g'(x)g'(x) \\
h''(x) &= f'(g(x))g''(x) + f''(g(x))(g'(x))^2
\end{aligned}
$$

If $h''(x) > 0$, then $h$ is convex. The second derivative of a convex function has to be positive, so $g''(x)$ and $f''(g(x)$ are positive. Additionally, $g'(x)^2$ is positive. So that leaves with $f'(g(x))$, which could be $<= 0$ or $>= 0$. Depending on the magnitude and sign of $f'(g(x))$, $h''(x) > 0$ or $h''(x) < 0$. An example is if we have $g(x) = x^2 - 3x$ and $f(x) = x^2$, then $f'(g(x)) = 2 * (x^2 - 3x)$. According to the graph below, when $x \in [0, 3]$, then $f'(g(x)) < 0$ , so it is non-convex.




In order to ensure that $h(x)$ is convex, then $f'(g(x)) >= 0$.

(b) If $f$ is convex and $g$ is concave, what additional hypothesis that guarantees $h$ is convex?

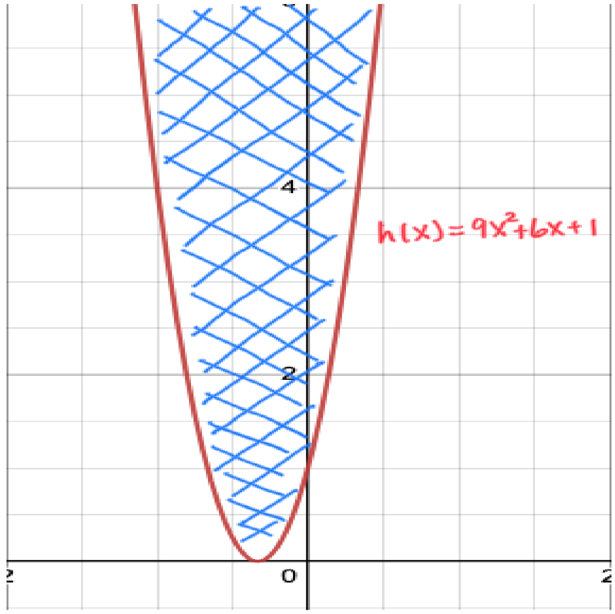
**Answer to 3b:** If $f$ is convex, and $g$ is concave then $g''(x) <= 0$ and $f''(x) >= 0$. Since $h''(x) = f'(g(x))g''(x) + f''(g(x))(g'(x))^2$, then $f''(g(x)) > 0$, $g''(x) <= 0$ and $(g'(x))^2 > 0$. In order for $h$ to be convex, then $f'(g(x)) <= 0$ in order to have $h''(x) > 0$.

(c) Show that if $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ affine, then $h$ is convex.

**Answer to 3c:** An example of $f(x) = x^2$ is convex and $g(x) = ax+b$ is affine. Therefore:

$$
\begin{aligned}
h(x) &= f(g(x)) \\
h(x) &= f(g(ax+b)) \\
h(x) &= (ax+b)^2 \\
h(x) &= a^2x^2 + 2axb + b^2
\end{aligned}
$$

$h(x)$ will always be convex because the epigraph epi $h$ is set above the graph of the function $h(x)$ since h(x) is a positive parabola every time at any point. Given an example where $a = 3$ and $b = 1$, see the epigraph below: $h(x) = 9x^2 + 6x + 1$.




(d) Show that the following functions are convex:

(i) Logistic regression objective: $\sum_{i=1}^{n} \log(1 + \exp(a_i^T x)) - b^T Ax$

**Answer to 3di:** According to Theorem 2.16, the function is convex if is twice differentiable and the Hessian is positive definite. Based on problem 1b, the logistic regression objective is twice differentiable. Additionally, the Hessian is:

$$\sum_{i=1}^{n} \frac{a_i a_i^T \exp(a_i^T x)}{(1 + \exp(a_i^T x))^2}$$

From the Hessian, both the numerator and denominator are positive because of the exp term on the numerator and squared term on the denominator. Additionally, linear matrix operations on the convex function will still be a convex set. As a result, the hessian is everywhere positive definite, and the logistic regression objective is convex.

(ii) Poisson regression objective: $\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax$.

**Answer to 3dii:** According to Theorem 2.16, the function is convex if is twice differentiable and the Hessian is positive semi-definite. Based on problem 1c, the poisson regression objective is twice differentiable. Additionally, the Hessian is:

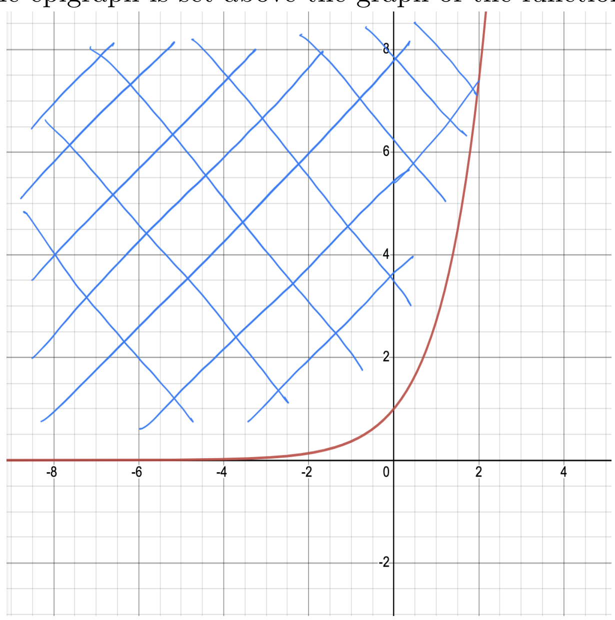
$$\sum_{i=1}^{n} a_i a_i^T \exp(a_i^T x)$$

The Hessian of the poisson regression is positive semi-definite because of the exp term. Additionally, linear matrix operations on the convex function will still be a convex set. As a result, the poisson regression objective is convex.

(4) A function $f$ is *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y), \quad \lambda \in (0,1).$$

(a) Give an example of a strictly convex function that does not have a minimizer.
**Answer to 4a:** An example of a strictly convex function that does not have a minimizer is $f(x) = e^x$. Since $f''(x) = e^x$ is strictly positive, it is sufficient to conclude that it is strictly positive. According to the epigraph below, you can see that the epigraph is set above the graph of the function $f(x) = e^x$.




(b) Show that a sum of a strictly convex function and a convex function is strictly convex.
**Answer to 4b:** A sum of a strictly convex function and a convex function is strictly convex based on the algebraic argument below with f(x) as strictly convex and g(x) as convex. This is done by adding the two inequalities with $h(x) = f(x) + g(x)$. Assuming $\lambda \in (0,1)$.

$$\begin{aligned}
h(\lambda x + (1-\lambda)y) &= (f+g)(\lambda x + (1-\lambda)y) \\
(f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\
&<= \lambda f(x) + (1-\lambda)f(y) + g(\lambda x + (1-\lambda)y) \\
&<= \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\
&= \lambda h(x) + (1-\lambda)h(y)
\end{aligned}$$

So therefore $h(\lambda x + (1-\lambda)y) <= \lambda h(x) + (1-\lambda)h(y)$. If either $f$ or $g$ is strictly convex, then one inequality is strict, so the whole inequality becomes strict, and the sum $f + g$ is strictly convex.

(c) Characterize all solutions to the problem

$$\min_x \frac{1}{2}\|Ax - b\|^2$$

**Answer to 4c:** In order to characterize all solutions to the problem, we set the gradient = 0. Using the composition rule from 1(a), the gradient $A^T(Ax - b) = 0$. All the x's that satisfy this linear system will solve the optimization problem. The only possible solutions to the problems are unique (1 solution when $A^T A$ is invertible) or infinitely many solutions when $A^T A$ has a nontrival nullspace.

(5) A function $f$ is $\beta$-smooth when its gradient is $\beta$-Lipschitz continuous.

(a) Find a global bound for $\beta$ of the least-squares objective $\frac{1}{2}\|Ax - b\|^2$.
**Answer to 5a:** The global bound for $\beta$ of the least-square objective $\frac{1}{2}\|Ax-b\|^2$ is $\lambda_{max}(A^T A)$ as discussed in class.

(b) Find a global bound for $\beta$ of the regularized logistic objective

$$\sum_{i=1}^{n} \log(1 + \exp(\langle a_i, x\rangle)) + \frac{\lambda}{2}\|x\|^2.$$

**Answer to 5b:** The global bound for $\beta$ of the regularized logistic objective is $\frac{1}{4}\lambda_m ax(A^T A)$ as shown by the inequality below:

$$\textstyle\sum_{i=1}^{n} \frac{a_i \exp(a_i^T x) a_i^T}{(1+\exp(a_i^T x))^2} <= |\frac{1}{4}\sum_{i=1} a_i^T a_i|$$

$$\textstyle\sum_{i=1}^{n} \frac{a_i \exp(a_i^T x) a_i^T}{(1+\exp(a_i^T x))^2} <= |\frac{1}{4}\sum_{i=1} A^T A|$$

$$\textstyle\sum_{i=1}^{n} \frac{a_i \exp(a_i^T x) a_i^T}{(1+\exp(a_i^T x))^2} <= \frac{1}{4}\lambda_{max}(A^T A)$$

(c) Do the gradients for Poisson regression admit a global Lipschitz constant?
**Answer to 5c:** The gradients of Poisson regression does not admit a global Lipschitz constant because Poisson regression has no single $B$ when trying to determine if the gradient is $B$-Lipschitz continuous. Since it is not $B$-smooth, there is no quadratic upper bound.

(6) Please complete the coding homework (starting with the notebook uploaded to Canvas).