Name: Philip Pham

AMATH 515 Homework Set 1

Due: Monday Jan 27th, by midnight.

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

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

Use D for the total derivative rather than abuse the ∇ notation. If $f : \mathbb{R}^n \to \mathbb{R}^m$, then the linear map $D_a f : \mathbb{R}^n \to \mathbb{R}^m$ is the total derivative of f evaluated at a. With this definition, for functions, $f : \mathbb{R}^n \to \mathbb{R}$, we have that $(D_x f)^T = \nabla f(x)$.

Proof. Clearly,

$$\lim_{h\to 0} \frac{A(x+h) - Ax - Ah}{h} = 0, \forall x \in \mathbb{R}^n$$

since A is a linear map. Thus, by definition $D_x A = A$.

The chain rule tells us that if $F(x) = G(H(x))$, then $D_x F = D_{H(x)} G \circ D_x H$. In this case, the composition operator is just matrix multiplication. We have that

$$D_x h = D_{Ax} g \circ D_x A = D_{Ax} g \circ A,$$

which implies that

$$\begin{aligned} 0 &= \lim_{h\to 0} \frac{g(A(x+h)) - g(Ax) - (D_{Ax} g \circ A)\, h}{h} \\ &= \lim_{h\to 0} \frac{g(A(x+h)) - g(Ax) - \langle A^T \left(D_{Ax} g\right)^T, h\rangle}{h} \\ &= \lim_{h\to 0} \frac{g(A(x+h)) - g(Ax) - \langle A^T \nabla g(Ax), h\rangle}{h}. \end{aligned}$$

Thus, by definition, $\nabla h(x) = A^T \nabla g(Ax)$. □

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

Proof. Define $F(x) = A^T \nabla g(Ax)$, so $F : \mathbb{R}^n \to \mathbb{R}^n$. We can apply the chain rule again:

$$D_x F = A^T D_x \nabla g(Ax) = A^T \nabla^2 g(Ax) D_x A = A^T \nabla^2 g(Ax) A.$$

Since $D_x F = \nabla^2 h(x)$, we have our desired result. □

(c) Use the formulas to compute the gradient and hessian of the logistic regression objective:

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

where a_i denote the rows of A.

Proof. We can write the terms in the summation as $g(a_i^T x)$, where $g(x) = \log(1 + \exp(x))$. We have that

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

Let

$$l(x) = \sum_{i=1}^{n} \log(1 + \exp(\langle a_i, x \rangle)) - b^T A x.$$

Applying the previous results to each term, we obtain

$$\nabla l(x) = \sum_{i=1}^{n} a_i \frac{\exp(\langle a_i, x \rangle)}{1 + \exp(\langle a_i, x \rangle)} - A^T b$$
$$\nabla^2 l(x) = \sum_{i=1}^{n} a_i a_i^T \frac{\exp(\langle a_i, x \rangle)}{(\exp(1 + \langle a_i, x \rangle))^2}.$$

□

The point here is just to show why these formulas hold. One way to proceed is to be fully explicit in all coordinates but this is time consuming and not very insightful. A better way to go is to think about efficient ways to write the product Ax that makes it easy to differentiate with respect to each coordinate.

(2) Explain why each of the following functions is convex.

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

Proof. This follows from the defintion that

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

for all x and y and $\lambda \in [0, 1]$. We have four cases: (1) if $x \in C$ and $y \in C$, then both the left-hand side and right-hand side are always 0 and the inquality holds; (2) if $x \in C$ and $y \notin C$, then the right-hand side is ∞ unless $\lambda = 1$,

in which case, both the right-hand and left-hand side are 0; (3) if $x \notin C$ and $y \in C$, this is the same as the previous case except that both the left-hand side and right-hand side are 0 when $\lambda = 0$; and (4) if $x \notin C$ and $y \notin C$, the right-hand side is always ∞. □

(b) Support function to any set: $\sigma_C(x) = \sup_{c \in C} c^T x$.

Proof.

$$\begin{aligned}
\sigma_C(\lambda x + (1-\lambda)y) &= \sup_{c \in C} c^T \left(x + (1-\lambda)y\right) \\
&= \sup_{c \in C} \left(\lambda c^T x + (1-\lambda) c^T y\right) \\
&\leq \sup_{c \in C} \lambda c^T x + \sup_{c \in C} (1-\lambda) c^T y \\
&= \lambda \sup_{c \in C} c^T x + (1-\lambda) \sup_{c \in C} c^T y \\
&= \lambda \sigma_C(x) + (1-\lambda)\sigma_C(y),
\end{aligned}$$

so by definition σ_C is convex. □

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

Proof. Let $\|\cdot\|$ be a norm, meaning that it satisfies *absolute homogeneity* and the *triangle inequality*. We have that

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda \|x\| + (1-\lambda)\|y\|$$

for all $\lambda \in [0, 1]$, where we first apply the triangle inequality and then absolute homogeneity. By defintion, $\|\cdot\|$ is convex. □

(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). What additional condition ensures the convexity of the composition?

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

(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.

(d) Show that the following functions are convex:

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

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

(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. Explain why your function is strictly convex.

(b) Show that a sum of a strictly convex function and a convex function is strictly convex.

(c) What conditions (if any) are necessary to ensure that the following problems have a unique minimizer?

(i) Least squares: $\min_x \frac{1}{2}\|Ax-b\|^2$

(ii) Elastic net logistic:

$$\min_x \sum_{i=1}^{n} \log(1+\exp(\langle a_i, x\rangle)) + \lambda(\alpha\|x\|_1 + (1-\alpha)\|x\|^2), \quad \lambda > 0, \alpha \in (0,1)$$

(5) Lipschitz constants and β-smoothness. Remember that f is β smooth when its gradient is β-Lipschitz continuous.

(a) Find a global bound for β of the least-squares objective $\frac{1}{2}\|Ax-b\|^2$.

Proof. If $l(x) = \frac{1}{2}\|Ax-b\|^2$, then $\nabla l(x) = A^T(Ax-b)$.

We have that $\nabla l(y) - \nabla l(x) = A^TAy - A^TAx = A^TA(y-x)$. Since A^TA is positive semidefinite (it's the the Gramian matrix), we can write $x - y = c_1v_1 + \cdots + c_nv_n$, where the v_i are orthonormal eigenvectors. Moreover,

$$\|x-y\| = \sqrt{c_1^2 + \cdots + c_n^2}.$$

Let v_i have the corresponding eigenvalues λ_i and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

$$\begin{aligned}\|\nabla l(y) - \nabla l(x)\| &= \|A^TA(y-x)\| \\ &= \sqrt{\lambda_1^2c_1^2 + \cdots + \lambda_n^2c_n^2} \\ &\leq \lambda_1\sqrt{c_1^2 + \cdots + c_n^2} = \lambda_1\|y-x\|.\end{aligned}$$

Thus, we have that $\beta = \lambda_1 = \lambda_{\max}\left(A^TA\right)$, that is, the maximum eigenvalue of A^TA. □

(b) Find a global bound for β of the regularized logistic objective

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

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

$$\nabla l(x) = \sum_{i=1}^{n} a_i \frac{\exp(\langle a_i, x \rangle)}{1 + \exp(\langle a_i, x \rangle)} + \lambda x$$

$$\nabla^2 l(x) = \sum_{i=1}^{n} a_i a_i^T \frac{\exp(\langle a_i, x \rangle)}{(1 + \exp(\langle a_i, x \rangle))^2} + \lambda I \leq \frac{1}{4} \sum_{i=1}^{n} a_i a_i^T + \lambda I = \frac{1}{4} A^T A + \lambda I.$$

from the derivation in Problem 1. Thus, $\beta = \lambda_{\max}\left(\frac{1}{4}A^T A + \lambda I\right)$. □

(c) Do the gradients for Poisson regression admit a global Lipschitz constant?

Proof. They do not. The poisson objective is $l(x) = \sum_{i=1}^{n} \exp(\langle a_i, x \rangle) - b^T A x$, the gradient and hessian is

$$\nabla l(x) = \sum_{i=1}^{n} a_i \exp(\langle a_i, x \rangle) - A^T b$$

$$\nabla^2 l(x) = \sum_{i=1}^{n} a_i a_i^T \exp(\langle a_i, x \rangle).$$

There is no global upper bound for $\nabla^2 l(x)$ since the $\exp(\langle a_i, x \rangle)$ factor is unbounded as function of x, so ∇l is not Lipschitz-continuous. □

(6) Behavior of steepest descent for logistic vs. poisson regression.

(a) Given the sample (logistic) data set and starter code, implement gradient descent for ℓ_2-regularized logistic regression. Plot (a) the objective value and (b) the norm of the gradient (as a measure of optimality) on two separate figures. For the figure in (b), make sure the y-axis is on a logarithmic scale.

See Figure 1. I used the result in problem 5(b) to choose a constant step size of $\frac{1}{\beta}$.

(b) Implement Newton's method for the same problem. Does the method converge? If necessary, use the line search routine provided to scale your updated directly to ensure descent. Add the plots for Newton's method (a) and (b) to your Figures 1 and 2. What do you notice?

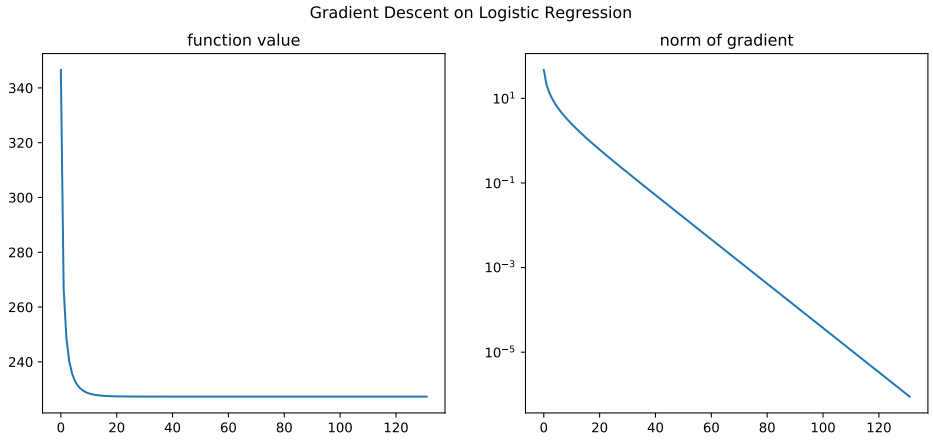

Figure 1: Gradient descent used to minimize the logisitic regression objective on the randomly generated sample data.

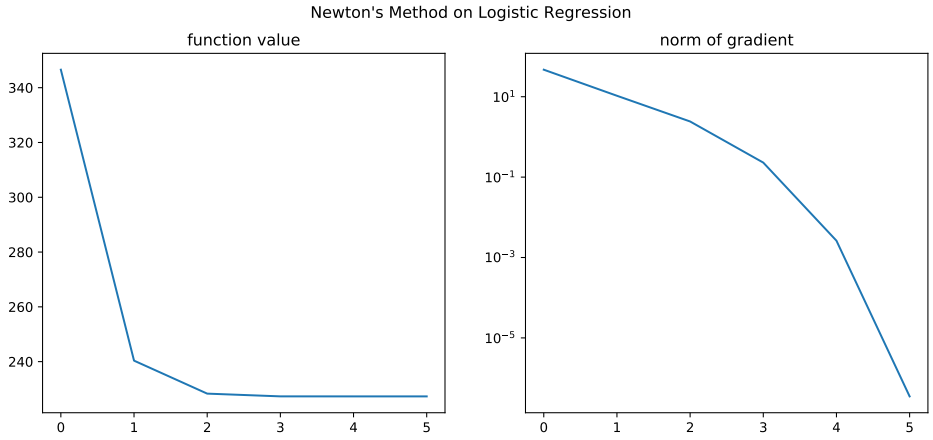

Figure 2: Newton's method used to minimize the logisitic regression objective on the randomly generated sample data.

See Figure 2. The method converges much more quickly: gradient descent takes 131 steps, and Newton's method merely takes 5 steps.

(c) Using the sample (Poisson) data and starter code provided, implement gradient descent and Newton's method for ℓ_2-regularized Poisson regression. You may

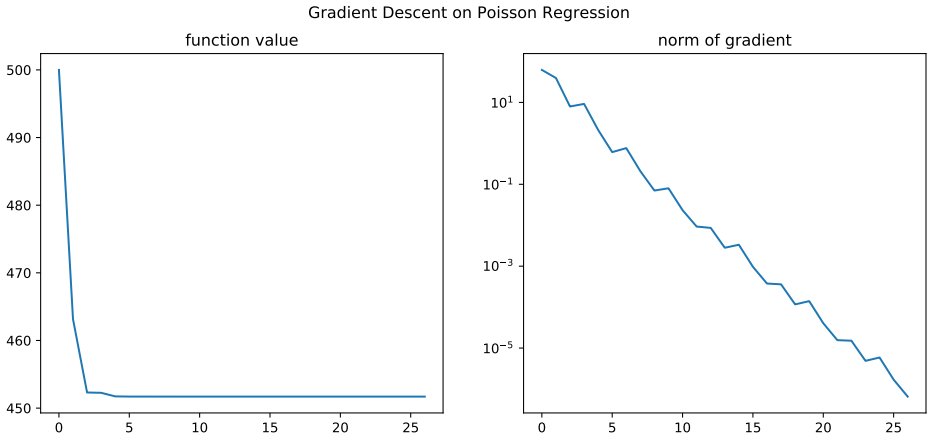

Figure 3: Gradient descent used to minimize the Poisson regression objective on the randomly generated sample data.

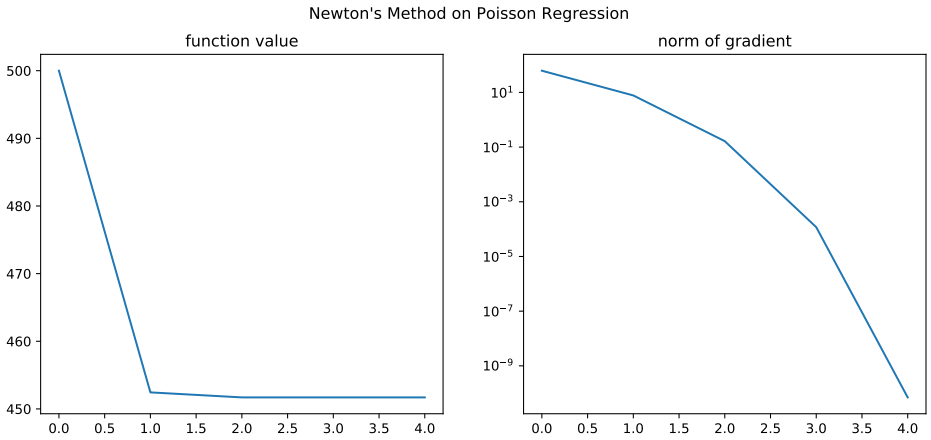

Figure 4: Newton's method used to minimize the Poisson regression objective on the randomly generated sample data.

need to use the line search routine for both algorithms. Make the same plots as you did for the logistic regression examples.

See Figures 3 and 4.

(d) What do you notice qualitatively about steepest descent vs. Newton?

Steepest descent takes many more iterations to converge. In particular, the linear rate of convergence is evident in the plots of the gradient norms for steepest descent. Likewise, in the plots of the gradient norms for Newton's method, we see quadratic rate of convergence as expected.