Homework 1

Due: January 27, 2021

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

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

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

   (c) Use the formulas to compute the gradient and hessian of the logistic regression objective:
   $$\sum_{i=1}^{m} \log(1 + \exp(\langle a_i, x\rangle)) - b^T Ax$$
   where a_i denote the rows of A.

Solution.

(a) By chain rule $h' = g'(Ax)A$, then $\nabla h = (h')^T = A^T(g'(Ax))^T = A^T \nabla g(Ax)$.

(b) Applying part (a) to ∇h yields $\nabla^2 h = A^T \nabla(\nabla g(Ax)) = A^T \nabla^2 g(Ax) A$.

(c) Let $z = Ax$, then $z_i = \langle a_i, x\rangle$. Denote the objective by h and let $g(z) = \sum_{i=1}^m \log(1 + \exp(z_i))$.

$$\nabla h = A^T \nabla g(z) - A^T b, \text{ where } (\nabla g(z))_i = \frac{\exp(z_i)}{1 + \exp(z_i)}$$
$$\nabla^2 h = A^T \nabla^2 g(z) A, \text{ where } \nabla^2 g(z) = \operatorname{diag}\left(\frac{\exp(z_i)}{(1 + \exp(z_i))^2}\right)$$

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

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

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

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

Solution.

(a) Suppose $x, y \in C$, then for $\lambda \in (0, 1)$

$$\delta_c(\lambda x + (1 - \lambda)y) = 0 \leq \lambda\delta_c(x) + (1 - \lambda)\delta_c(y) = 0$$

For $x \notin C$ (same argument for $y \notin C$)

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

(b) For $\lambda \in (0,1)$

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

(c) By properties 1 and 2 of norms

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

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

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

Solution.

(a) Suppose $f = -x$ and $g = x^2$, then $h = -x^2$ is nonconvex. Suppose f is non-decreasing. Since g is convex, $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$. Since f is convex and non-decreasing

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

Thus $f(g(x))$ is convex.

(b) Suppose f is non-increasing. Since g is concave, $g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y)$. Since f is convex and non-increasing

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

Thus $f(g(x))$ is convex.

(c) Since g is affine, we can write $g = Ax + b$.

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= A(\lambda x + (1-\lambda)y) + b \\ &= \lambda(Ax + b) + (1-\lambda)(Ay + b) \\ &= \lambda g(x) + (1-\lambda)g(y) \end{aligned}$$

Then since f is convex

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

Thus $f(g(x))$ is convex.

(di) Let $g(z) = \sum_i f(z_i)$, $f(z_i) = \log(1+\exp(z_i)) - b_i z_i$, then the objective $h = g(Ax)$. Since $f''(z_i) = \frac{\exp(z_i)}{(1+\exp(z_i))^2} > 0$, f is convex. Then

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= \sum_i f(\lambda x_i + (1-\lambda)y_i) \\ &\leq \sum_i \lambda f(x_i) + (1-\lambda)f(y_i) \\ &= \lambda g(x) + (1-\lambda)g(y) \end{aligned}$$

Thus g is also convex. Since Ax is affine, $h = g(Ax)$ is convex.

(dii) Let $g(z) = \sum_i f(z_i)$, $f(z_i) = \exp(z_i) - b_i z_i$, then the objective $h = g(Ax)$. Since $f''(z_i) = \exp(z_i) > 0$, f is convex. Then we can follow the same method in (di) to show that h 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. 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) Characterize all solutions to the problem

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

(d) Does the objective function below (logistic regularized with an elastic net) have a minimizer? Is it unique? Explain.

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

Solution.

(a) Let $f : (0,1) \to \mathbb{R}$ be defined as $f(x) = x^2$.

$$(\lambda x + (1-\lambda)y)^2 = \lambda^2 x^2 + (1-\lambda)^2 y^2 + 2\lambda(1-\lambda)xy$$

We want to show that $(\lambda x + (1-\lambda)y)^2 < \lambda x^2 + (1-\lambda)y^2$, i.e.

$$\begin{aligned} \lambda x^2 + (1-\lambda)y^2 &> \lambda^2 x^2 + (1-\lambda)^2 y^2 + 2\lambda(1-\lambda)xy \\ \lambda(1-\lambda)x^2 + \lambda(1-\lambda)y^2 - 2\lambda(1-\lambda)xy &> 0 \\ \lambda(1-\lambda)(x-y)^2 &> 0 \end{aligned}$$

which holds if $x \neq y$. Thus f is strictly convex. f has no minimizer because we can always find $x_1 < x_2$ such that $f(x_1) < f(x_2)$.

(b) Let f be strictly convex and g be convex, then

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

Thus $f + g$ is strictly convex.

(c) We have the condition for optimal solution

$$\begin{aligned} \nabla f(x) &= 0 \\ A^T(Ax - b) &= 0 \end{aligned}$$

Then if A is invertible, we have $x = (A^T A)^{-1} A^T b$. This minimizer is unique because $\nabla^2 f(x) = A^T A > 0$.

(d) Yes, the minimizer is unique because this function is strictly convex. To prove strict convexity, we show that the first term is strictly convex and second term is convex, so their sum is strictly convex by (c). We notice that $f''(z_i) > 0$ holds for $f(z_i) = \log(1 + \exp(z_i))$, thus the first term is strictly convex. The second term is convex because norms are convex. Thus the objective is strictly convex.

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

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

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

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

Solution.

(a) We have $\nabla f = A^TAx - A^Tb$, then

$$||(A^TAx - A^Tb) - (A^TAy - A^Tb)|| = ||A^TA(x-y)|| \leq ||A^TA||||x-y||$$

This proves ∇f is β-Lipschitz continuous. Thus a global bound is $\beta \geq ||A^TA||$.

(b) Let $g(z) = \sum_i \log(1+\exp(z_i))$, then $f(x) = g(Ax) + \frac{\lambda}{2}\|x\|^2$. We have $\nabla f(x) = A^T\nabla g(Ax) + \lambda x$. Also $(\nabla g(z))_i = \frac{\exp(z_i)}{1+\exp(z_i)}$. By the mean value theorem, there exists a point c in (x_i, y_i) with

$$\frac{g'(x_i) - g'(y_i)}{x_i - y_i} = g''(c)$$

then $||g'(x_i) - g'(y_i)|| = ||g''(c)||||x_i - y_i||$. We know $g''(x_i) = \frac{\exp(x_i)}{(1+\exp(x_i))^2} \leq g''(0) = \frac{1}{4}$.

$$||\nabla g(x) - \nabla g(y)||^2 = \sum_i ||g'(x_i) - g'(y_i)||^2$$
$$||\nabla g(x) - \nabla g(y)||^2 \leq \frac{1}{16}\sum_i ||x_i - y_i||^2$$
$$||\nabla g(x) - \nabla g(y)|| \leq \frac{1}{4}||x-y||$$

Thus $\beta = \frac{1}{4}$ for ∇g, then for ∇f

$$\begin{aligned}
||\nabla f(x) - \nabla f(y)|| &= ||(A^T\nabla g(Ax) + \lambda x) - (A^T\nabla g(Ay) + \lambda y)|| \\
&= ||A^T(\nabla g(Ax) - \nabla g(Ay)) + \lambda(x-y)|| \\
&\leq \frac{1}{4}\left\|A^T\right\|\|A\|\|x-y\| + \lambda\|x-y\| \\
&\leq (\frac{1}{4}\left\|A^T\right\|\|A\| + \lambda)\|x-y\|
\end{aligned}$$

Thus a global bound is $\beta \geq \frac{1}{4} \left\|A^T\right\| \|A\| + \lambda$.

(c) No. If we look at $\nabla^2 f$, the diagonal is $\exp(\langle a_i, x\rangle)$, which is unbounded. Then the eigenvalues of $\nabla^2 f$ is unbounded, thus the objective is not β-smooth.

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.

   (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?

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

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

   Solution.

   (a-c) See Figure 1 and Figure 2.

   (d) We notice that Newton's method converges with much fewer iterations than steepest descent and the norm of gradient drops faster when it gets small.

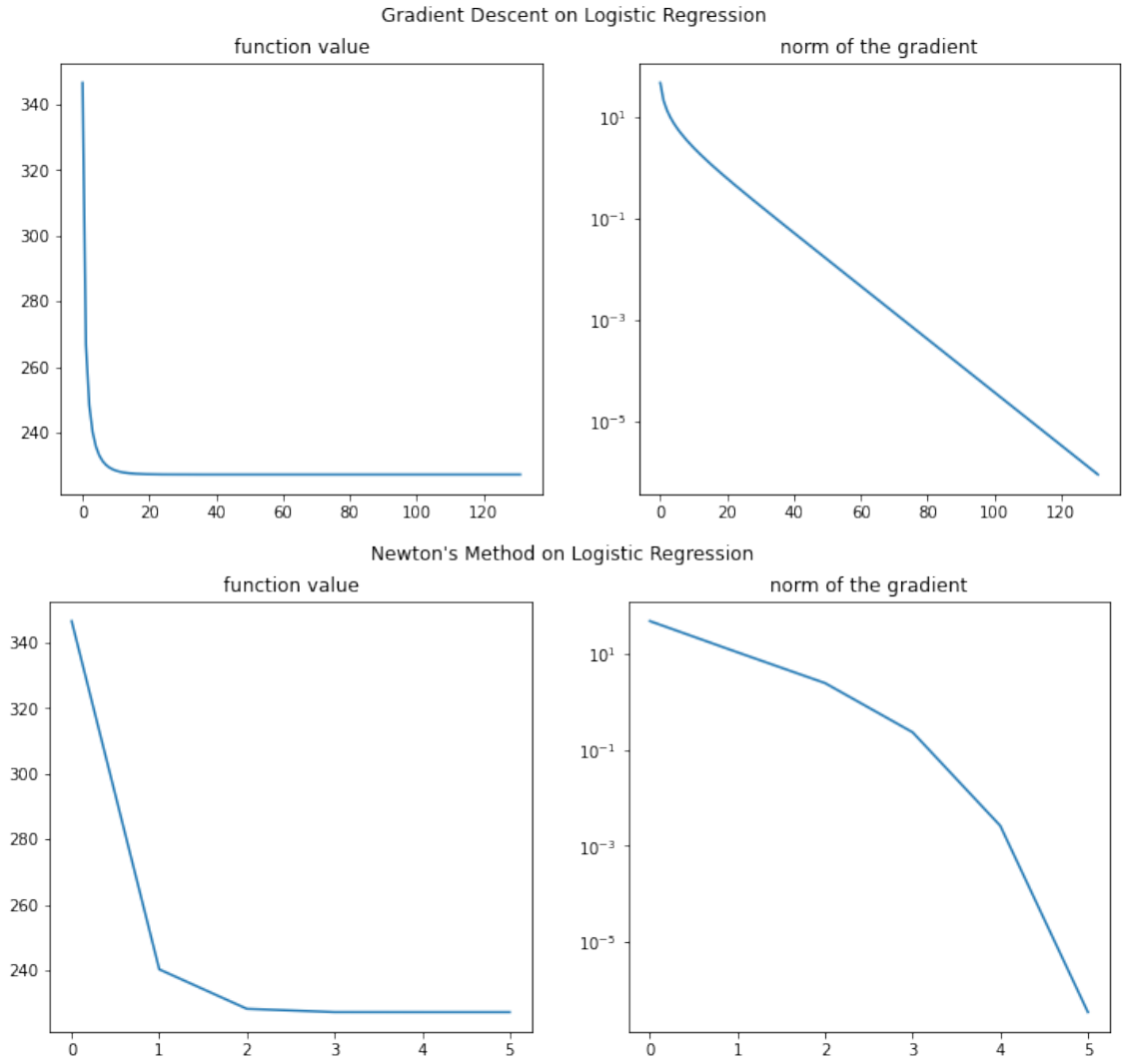

Figure 1: Gradient descent and Newton's method for logistic regression.

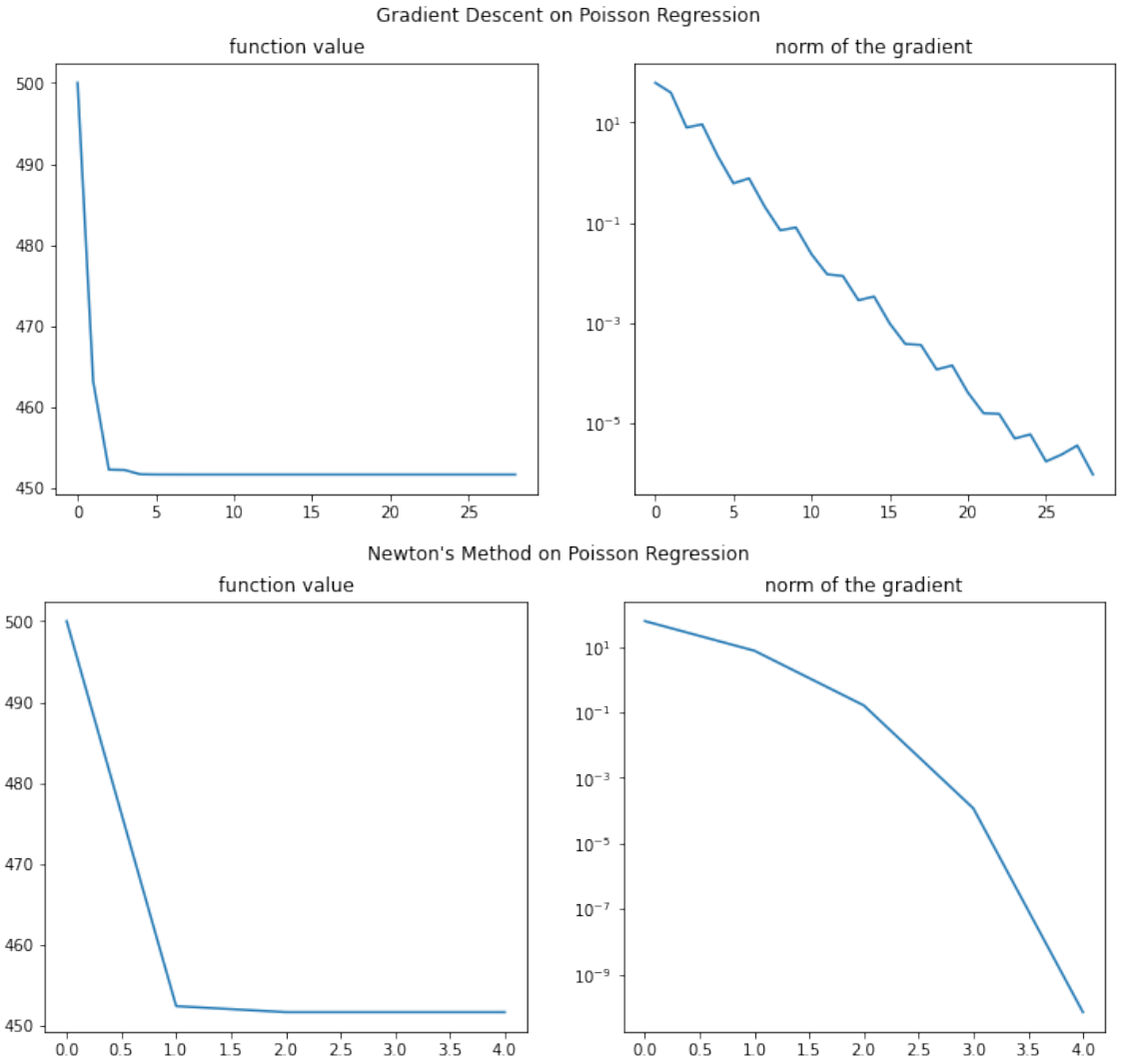

Figure 2: Gradient descent and Newton's method for Poisson regression.