Problem Set 8 Solutions

This problem set is due **at 11:59pm** on **Wednesday, April 26, 2017.**

EXERCISES (NOT TO BE TURNED IN)

Note: This problem set makes heavy use of the Hessian matrix. You may want to familiarize yourself with the definition of the Hessian and how it can be used to characterize critical points of multivariate functions. The Wikipedia article for the Hessian matrix has a good overview of this topic.

Problem 8-1. Critical Points [40 points]

Gradient descent is not guaranteed to find a global, or even local optimum for nonconvex functions. In this problem, we will explore some possible failure modes of gradient descent and gain familiarity working with gradients and Hessians.

Consider functions of the form $f(x, y) = ax^2 + bx^4 + cy \ln y$, where a, b, and c are constants. Note the domain of f is restricted to $y > 0$.

(a) [10 points] Calculate the gradient $\nabla_{x,y} f(x, y)$ and the Hessian $\nabla^2_{x,y} f(x, y)$ in terms of a, b, and c.

Solution:

This is a straightforward application of the definition of the gradient and Hessian.

The gradient is

$$\begin{pmatrix} x(2a + 4bx^2) \\ c(1 + \log y) \end{pmatrix}$$

The Hessian is

$$\begin{bmatrix} 2a + 12bx^2 & 0 \\ 0 & c/y \end{bmatrix}$$

Notes for the graders:

5 points - Correct expression for gradient
5 points - Correct expression for Hessian

(b) [10 points] If $a = c = 1$ and $b = 0$, give a possible step size and starting position such that gradient descent neither converges nor diverges.

Solution: Note that we only need one coordinate to oscillate: this is easiest to analyze for the x-dimension because this function is symmetric about 0 in the x-dimension. Because the gradient of x is independent of y, it doesn't really matter where we set y, so for convenience we will set $y = y^*$ for some critical point y^*.

Now consider what happens if we set $x = 1$ with step size k. The partial derivative with respect to x is $2ax^0 = 2$, so $x^1 = 1 - 2k$. Observe that if $k = 1$, then $x^1 = -1$, and then $x^2 = 1$ and we will oscillate forever. (You can also set this up as a system of equations and solve for the k such that $x^0 = x^2$)

Thus, with step size 1 and starting position $(1, y^*)$, gradient descent will oscillate forever.

Notes for the graders:
There are many possible solutions for this problem. Check that whatever solution is given neither converges nor diverges.
Give half credit for solutions which diverge instead of oscillating.

(c) [10 points] Now let $a = -1/2, b = 1/4$, and $c = 1$. Find all the critical points of f and determine if they are local maxima, minima, or saddle points. One way to do this is by finding the eigenvalues of the Hessian you calculated in part (a) for each point. If both eigenvalues are positive, the point is a local minimum. If both eigenvalues are negative, the point is a local maximum. If the eigenvalues have different signs, the point is a saddle point.

Solution: A critical point of a multivariate function is a point at which the gradient is zero. In our case, we need to find x^*, y^* such that

$$x(2a + 4bx^2) = 0$$

$$c(1 + \log y) = 0$$

Note that these two equations are actually independent, so any combination of solutions to the individual equations is a solution to the system. In particular, $x^* = \{0, 1, -1\}$ and $y^* = 1/e$, so the critical points are $(0, 1/e)$, $(-1, 1/e)$, and $(1, 1/e)$.
The Hessian is diagonal, so the eigenvalues are exactly the diagonal entries. Evaluating the Hessian at each point reveals that $(0, 1/e)$ is a saddle point (since the eigenvalues have opposite signs) and the other two are local minima (since both eigenvalues are positive).

Notes for the graders:
3 points - set gradient components to zero
3 points - correctly identified critical points
4 points - correctly classify critical points
Note: they don't need to show work here as long the answer is correct.

(d) [10 points] Still letting $a = -1/2, b = 1/4$, and $c = 1$, find the set of starting positions such that gradient descent will never converge to a local minimum no matter what step size is used.

Solution: Since we need to find points such that gradient descent will do badly no matter what the step size is, we should look for points where the gradient is 'unhelpful' in the sense that it will never get us any closer to a local minimum.

We already know one point where the gradient is completely unhelpful: that is the saddle point, because the gradient is zero. So if we start at the saddle point, no matter what the step size is we will never move to a local minimum. However, a little thought reveals that this problem is not restricted to the saddle point.

If $x = -1, 1,$ or 0 then the gradient with respect to x is zero. This is not a problem for $x = -1, 1$ because we can still converge to a local minimum, but if $x = 0$ at the starting position then x will always equals zero. Because neither of the local minima have $x = 0$, we can't possibly ever converge to a minimum if we start at any point with $x = 0$.

Notes for the graders:
3 points - Claims saddle point is the only such starting position.
10 points - Noting that start point where $x = 0$ suffices. This answer should receive full credit even if they don't explicitly point out the saddle point.

Problem 8-2. Logistic Regression and Gradient Descent [60 points]

In this problem, we will consider using gradient descent to minimize the following function:

$$l(\omega) = -y \log \sigma(\omega^T x) - (1-y)\log(1-\sigma(\omega^T x)) \quad \omega, x \in \mathbb{R}^n, y \in \mathbb{R} \tag{1}$$

where $\sigma(t) = 1/(1+e^{-t})$. Note that $\sigma(t) \in (0,1) \quad \forall t$.

For this problem, the following property of the sigmoid function may prove useful.

$$\frac{\partial}{\partial t}\sigma(t) = \sigma(t)(1-\sigma(t))$$

You can easily verify that this property holds using the definition of $\sigma(t)$.

You don't have to know this, but if you're curious, l is the loss function for logistic regression. Minimizing l corresponds to finding the best coefficient ω for logistic regression on a single point (x, y)

(a) [5 points] Derive the gradient descent step to minimize $l(\omega)$. Your step should be of the form $\omega^{t+1} = \omega^t - \eta g(\omega^t)$ for some function g.

Solution: By the definition of gradient descent, $g = \nabla_\omega l$. If we let x_i represent the ith element of x and ω_i represent the ith element of ω, then we can take partial derivatives with respect to each ω_i to find the gradient. This is most easily done by using equation and never expanding the sigmoid function at all.

$$\begin{aligned}
(\nabla_\omega l)_i &= -\frac{x_i y \sigma(\omega^T x)(1-\sigma(\omega^T x))}{\sigma(\omega^T x)} - \frac{-(1-y)x_i\sigma(\omega^T x)(1-\sigma(\omega^T x))}{1-\sigma(\omega^T x)} \\
(\nabla_\omega l)_i &= (\sigma(\omega^T x) - y)x_i \\
\nabla_\omega l &= (\sigma(\omega^T x) - y)x
\end{aligned}$$

So our gradient descent step is $\omega^{t+1} = \omega^t - \eta\left(\sigma((\omega^t)^T x) - y\right)x$

Notes for the graders:
Full credit should be awarded for a correct answer even without work.
3 points if they attempted to find the gradient but made a math error.

(b) [5 points] Evaluate the Hessian of $l(w)$ to show that

$$\nabla^2 l(\omega) = xx^T\sigma(\omega^T x)(1-\sigma(\omega^T x))$$

Solution: The easiest way to evaluate the Hessian here is to use the definition of the Hessian:

$$\nabla^2 l(\omega) = \left[\frac{\partial^2 l(\omega)}{\partial\omega_i\partial\omega_j}\right]_{i,j}$$

We already know the partial derivatives $\partial l(w)/\partial \omega_i$ from the gradient we calculated in part (a), so all that remains is to take the partial derivatives of each entry of $\nabla l(w)$. Again, we can do this without expanding σ.

$$(\nabla^2 l(\omega))_{ij} = x_i x_j \sigma(\omega^T x)(1 - \sigma(\omega^T x))$$
$$\nabla^2 l(\omega) = x x^T \sigma(\omega^T x)(1 - \sigma(\omega^T x))$$

(c) [15 points] Now show that gradient descent will always converge to the correct minimum by proving that $l(\omega)$ is convex, but **not** $\alpha-$convex for any $\alpha > 0$.
Hint: A function $f(x)$ is convex if $\nabla^2 f(x)$ is positive semi-definite for all x, that is, if

$$z^T \nabla^2 f(x) z \geq 0 \quad \forall x, z$$

Solution: We can show that $\nabla^2 l(\omega)$ is positive semidefinite for any ω. For any vector z, we have that

$$z^T(\nabla^2 l(\omega))z = \sigma(\omega^T x)(1 - \sigma(\omega^T x)) z^T x x^T z = (z^T x)^2 \sigma(\omega^T x)(1 - \sigma(\omega^T x)) \geq 0$$

Here we make use of the fact that since $z^T x$ and $x^T z$ are both ways of expresing the same dot product, $z^T x x^T z = (z^T x)^2$. Since the square of any number is nonnegative and $\sigma(\omega^T x) \in (0, 1)$, the last inequality is true. Thus, $l(\omega)$ is convex.
We can also use the Hessian to show that l is not strongly convex.
$l(\omega)$ is α-convex if for any z and ω, $z^T(\nabla^2 l(\omega))z \geq \alpha \|z\|^2$. By the same argument as above, but now using the fact that $(z^T x)^2 \leq \|x\|^2 \|z\|^2$ because of the properties of the dot product, we have that

$$z^T(\nabla^2 l(\omega))z = (z^T x)^2 \sigma(\omega^T x)(1 - \sigma(\omega^T x)) \leq \|x\|^2 \sigma(\omega^T x)(1 - \sigma(\omega^T x)) \|z\|^2$$

Unfortunately, $\sigma(\omega^T x)(1 - \sigma(\omega^T x))$ could be arbitrarily close to zero if $\omega^T x \to \infty$ or $\omega^T x \to -\infty$, which means that for any $\alpha > 0$ we can pick a point ω such that $z^T \nabla^2 l(w) z < \alpha \|z\|^2$. Thus, $l(\omega)$ is not strongly convex.

Notes for the graders:
6 points - Used nonnegativity of $z^T x x^T z$ and σ to argue for convexity
6 points - Upper bounded $(z^T x)^2$ and argued that $\sigma(t)(1 - \sigma(t)) \to 0$ as $t \to \infty, -\infty$
3 points - Correctly argued that because of the above, we can't set any $\alpha > 0$, so $l(w)$ is not strongly convex.
It is also OK if they use a different way to prove [strong] convexity as long as their argument is valid.

(d) [10 points] Because $l(\omega)$ is not strongly convex, we can't place a useful bound on how long gradient descent takes to converge. We can slightly adjust the form of l to make it strongly convex by adding a regularization term. In this case, we will use ridge regularization with a positive constant λ and define our new function to be

$$l'(\omega) = -y \log \sigma(\omega^T x) - (1-y) \log(1 - \sigma(\omega^T x)) + \lambda \|\omega\|^2 \tag{2}$$

Now $l'(\omega)$ is λ-strongly convex. Prove that $l'(\omega)$ is β-smooth for $\beta = 0.25\|x\|^2 + 2\lambda$

Solution: To make our life easier when trying to find $\nabla^2 l'(w)$, we can note that $l'(w) = l(w) + \lambda\|w\|^2$. This means that

$$\nabla^2 l'(w) = \nabla^2 l(w) + \nabla^2(\lambda \omega^T \omega) = \nabla^2 l(w) + 2\lambda I$$

$$\nabla^2 l'(w) = x x^T \sigma(\omega^T x)(1 - \sigma(\omega^T x)) + 2\lambda I$$

$$z^T(\nabla^2 l'(w))z = (z^T x)^2 \sigma(\omega^T x)(1 - \sigma(\omega^T x)) + 2\lambda\|z\|^2 \leq (0.25\|x\|^2 + 2\lambda)\|z\|^2$$

Where the last inequality holds because $\sigma(\omega^T x)(1 - \sigma(\omega^T x)) \leq 0.25$ and $(z^T x)^2 \leq \|x\|^2\|z\|^2$.

Notes for the graders:
4 points - correctly evaluated $\nabla^2 l'(w)$
6 points - correctly upper bounded $z^T \nabla^2 l'(w) z$ by bounding $\sigma(\omega^T x)(1-\sigma(\omega^T x))$ and $z^T x$.

(e) [10 points] We can now place a useful bound on the convergence rate of l'. Prove that if we set $\omega^0 = 0$ and perform gradient descent with step size $\eta = O(1/(\|x\|^2 + \lambda))$, then the objective value will converge to within ϵ of the true minimum within t steps, where

$$t = O\left(\frac{\|x\|^2}{\lambda} \log \frac{\|\omega^*\|^2(\|x\|^2 + \lambda)}{\epsilon}\right)$$

and ω^* is the global optimum which minimizes $l'(\omega)$.
Hint: Try to bound $f(\omega^0) - f(\omega^*)$ in terms of $\|\omega^0 - \omega^*\|$.

Solution:
First, we can use the smoothness guarantee to bound $l'(\omega^0) - l'(\omega^*)$. In particular, using a lemma from page 6 of the L15 notes, we have that

$$l'(\omega^0) - l'(\omega^*) \leq (\omega^0 - \omega^*)^T \nabla l'(\omega^*) + \frac{\beta}{2}\|\omega^0 - \omega^*\|^2$$

Note that $\nabla l'(\omega^*) = 0$ because ω^* is a minimum of f. We can also plug in for ω^0 and β to find that

$$l'(\omega^0) - l'(\omega^*) \leq \frac{0.25\|x\|^2 + \lambda}{2}\|\omega^*\|^2$$

Alternatively, we can use the Cauchy-Schwartz inequality to show that

$$l'(\omega^0) - l'(\omega^*) \leq \|\nabla l'(\omega^0)\|^2 \|\omega^0 - \omega^*\|^*$$

$$\|\nabla l'(\omega^0)\|^2 = \|x(\sigma(0) - y) + 2\lambda 0\|^2 \leq \|x\|^2$$

Now we can use the guarantee on convergence from the lecture notes. Plugging in the appropriate expressions for β and α for the condition number and using the bound we proved above on $l'(\omega^0) - l'(\omega^*)$, we find that

$$t \leq O\left(\frac{0.25\|x\|^2 + \lambda}{\lambda} \log \frac{0.25\|\omega^*\|^2(\|x\|^2 + \lambda)}{2\epsilon}\right)$$

We can simplify this expression because we only want to bound t up to a constant multiple.

$$t \leq O\left(\frac{\|x\|^2}{\lambda} \log \frac{\|\omega^*\|^2(\|x\|^2 + \lambda)}{\epsilon}\right)$$

Notes for the graders:

5 points - correctly bounded $l'(\omega^0) - l'(\omega^*)$
5 points - plugged correct values into the bound from lecture to obtain t

(f) [5 points] Unfortunately, while we can prove the convergence of the adjusted objective function, we are now minimizing a different function! Luckily, we can show that the value of the optimum of the regularized function is close to the value of the optimum of the original.

Let ω^* be the optimal minimizer for l, and ω'^* be the optimal minimizer for l'. Show that $l(\omega'^*) \leq l(\omega^*) + \lambda\|\omega^*\|^2$.

Solution: Since ω'^* is the optimizer for l', we know that $l'(\omega'^*) \leq l'(\omega^*)$. We can use the definition of $l'(\omega) = l(\omega) + \lambda\|\omega\|^2$

$$l(\omega'^*) + \lambda\|\omega'^*\|^2 \leq l(\omega^*) + \lambda\|\omega^*\|^2$$
$$l(\omega'^*) \leq l(\omega^*) + \lambda\|\omega^*\|^2$$

(g) [10 points] Using the inequalities found in part (f) and the runtime obtained in part (e), determine how to set λ such that if we use gradient descent to obtain a point $\widetilde{\omega}'^*$ which is within ϵ of optimum for l', then $\widetilde{\omega}'^*$ is also within 2ϵ of the optimum for l. Also determine how many iterations are necessary to converge to this approximation. Your answers should be in terms of ϵ, $\|\omega^*\|$, and $\|x\|$.

Solution: Note that here we are trying to set λ such that an approximate solution for l' is also an approximate solution for l.
The total error of $\widetilde{\omega}'^*$, is

$$l(\omega^*) - l(\widetilde{\omega}'^*) = (l(\omega^*) - l(\omega'^*)) + (l(\omega'^*) - l(\widetilde{\omega}'^*))$$
$$l(\omega^*) - l(\widetilde{\omega}'^*) \leq \epsilon + \lambda\|\omega^*\|^2$$

We need to set λ such that $\epsilon + \lambda\|\omega^*\|^2 \leq 2\epsilon$. Thus, $\lambda \leq \epsilon/\|\omega^*\|^2$.
Plugging this back into the runtime from part (e), we find that we need

$$t = O\left(\frac{\|x\|^2\|\omega^*\|^2}{\epsilon}\log\frac{\|\omega^*\|^2(\|x\|^2 + \frac{\epsilon}{\|\omega^*\|^2})}{\epsilon}\right) = O\left(\frac{\|x\|^2\|\omega^*\|^2}{\epsilon}\log\frac{\|x\|^2\|\omega^*\|^2}{\epsilon}\right)$$

iterations to find such a $\widetilde{\omega}'^*$

Notes for the graders:
5 points - Decomposed error into additive components $(l(\omega^*) - l(\omega'^*)) + (l(\omega'^*) - l(\widetilde{\omega}'^*))$
2 points - Obtained correct expression for λ
3 points - Correct bound on t