Michael Leibert
Math 504
Homework 6

2. (a) Let $f(x)$ be a function from $\mathbb{R}^n$ to $\mathbb{R}$. Suppose we would like to **maximize** $f(x)$. Show that if $Hf(x)$ is negative definite then the Newton's method direction at x, $-[Hf(x)]^{-1}\nabla f(x)$, is an ascent direction. What does this imply for Newton's method with backtracking? (We did this in class, except that we considered the minimization case and $Hf(x)$ as positive definite; here I want you to go through the argument yourself for this slightly altered case.)

Assume $Hf(x)$ is negative definite, $f(x)$ is concave. Then the eigenvalues, λ_i $i = 1, 2, ..., n$, of $Hf(x)$ are negative.

The inverse of $Hf(x)$, $[Hf(x)]^{-1}$, has eigenvalues $\dfrac{1}{\lambda_i}$ $i = 1, 2, ..., n$. Because the eigenvalues of $Hf(x)$ are negative, all of the eigenvalues of $\left[Hf(x)\right]^{-1}$ are negative thus $[Hf(x)]^{-1}$ is negative definite.

We check if the direction,

$$d^{(i)} = -\left[Hf(x)\right]^{-1}\nabla f\left(x^{(i-1)}\right),$$

is an ascent direction by determining,

$$\begin{aligned} 0 >& \nabla f\left(x^{(i-1)}\right)\cdot d^{(i)} \\ 0 >& \nabla f\left(x^{(i-1)}\right)\cdot -\left[Hf(x)\right]^{-1}\nabla f\left(x^{(i-1)}\right). \end{aligned}$$

We know this will always be true, that the direction is less than zero, because this is the definition of negative definiteness. This is true since $z^t - \left[Hf(x)\right]^{-1} z < 0$, for $z \neq 0$, it is not just true for the gradient, but for any point in $\mathbb{R}^n$.

Damped Newton's method here will be an ascent algorithm $x^{(i)} = x^{(i-1)} + d^{(i)}$. Our step size while loop terminates if $d^{(i)}$ is an ascent direction, but because we will always be ascending towards the maximum so it is unnecessary to backtrack.

(b) Consider the negative of the log-likelihood from previous homeworks:

$$-\log L(\alpha) = -\sum_{i=1}^{N}\left((1-y_i)(-\alpha_0-\alpha_1 x_i) - \log(1+\exp(-\alpha_0-\alpha_1 x_i))\right) \qquad (1)$$

Show that the negative log-likelihood is convex (here we're flipping the viewpoint (a), in which we considered maximization, while now we consider minimization). You can show convexity by checking that the Hessian is positive definite at various α or, more rigorously, show that the Hessian is positive definite for all α. You can also come up with your own approach.

Let $\eta_i = x^T\alpha = \alpha_0 + \alpha_1 x_i$.

If $y_i \sim \textit{Bernoulli}$, we can write this in exponential family form $\exp\left\{\dfrac{y_i\theta_i - b(\theta_i)}{a\left(\phi\right)} + c(y_i, \phi_{)}\right\}$.

$$
\begin{aligned}
f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} &= \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i) \\
&= \exp\left\{ y_i \underbrace{\log\left(\frac{\pi_i}{1-\pi_i}\right)}_{\theta_i} + \underbrace{\log(1-\pi_i)}_{b(\theta_i)} \right\} \\
&= \exp\left\{ y_i\theta_i - b(\theta_i) \right\} \\
\text{where } \theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) &\text{ and } \pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}
\end{aligned}
$$

So the negative log-likelihood is,

$$
\begin{aligned}
-\ell(\boldsymbol{\alpha}) &= -\sum_i \log\left(\exp\left\{ y_i\theta_i - b(\theta_i) \right\} \right) \\
&= -\sum_i y_i\theta_i - b(\theta_i).
\end{aligned}
$$

We generate the score equations using the chain rule. Although $\boldsymbol{\alpha}$ does not appear in these equations, it is implicitly there through π_i, since $\pi_i = g^{-1}(\eta_i)$

$$
\frac{\partial \ell(\boldsymbol{\alpha})}{\partial \alpha_j} = \sum_i \frac{(y_i - \pi_i)\; x_{ij}}{Var(Y_i)} \frac{\partial \pi_i}{\partial \eta_i}
$$

Because logistic regression uses the canonical link, $\eta_i = \theta_i$, $\dfrac{\partial \pi_i}{\partial \eta_i} = Var(Y_i)$ and the score becomes

$$
-\sum_i (y_i - \pi_i)\, x_{ij}.
$$

We now take the second derivatives of the log-likelihood function.

$$
\frac{\partial \ell_i}{\partial \alpha_j \partial \alpha_k} = b''(\theta_i) x_{ij} x_{ik}
$$

These are the the entries for the hessian of the logistic log-likelihood, $H\ell(\boldsymbol{\alpha})$, they are positive. So the $z^t\left[H\ell(\boldsymbol{\alpha})\right]^{-1} z < 0$, for $z \neq 0$, and it is negative definite.

3. Here we will revisit the linear regression problem from hw 1. Attached is the datafile `economic_data.txt` .

   (a) Form the model matrix M and compute the condition number of M, $M^T M$. Explain why R gave an error when you tried to compute the normal equations.

```
dat<-read.table("economic_data.txt",header=T)
options(scipen = 999)

y<-dat[,ncol(dat)]
B<-as.matrix(dat[,2:(ncol(dat)-1)])
B<-cbind(1,B);M<-B
solve(t(M) %*% M ) %*% t(M) %*% y
```

```
## Error in solve.default(t(M) %*% M): system is computationally singular:  reciprocal
condition number = 3.50595e-20
kappa(M)
## [1] 5617818514
kappa(t(M) %*% M )
## [1] 22722032332043558912
```

The condition number is $2.2722032332043558912 \cdot 10^{19}$. *R* gave us an error when we tried to compute the normal equations because we do not have the significant digits to compute invertibility.

(b) Force R to apply Gaussian elimination by adjusting the **tol** flag in the **solve** function. Compute the solution of the normal equations and compare to the result you get using **lm**. Explain the result..

```
    options(scipen = 999)
    dat<-dat[,-1]
summary( lm(B˜.,data=dat) )
##
## Call:
## lm(formula = B ˜ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -409.7 -158.0  -27.5  101.5  455.9
##
## Coefficients:
##                   Estimate     Std. Error t value Pr(>|t|)
## (Intercept) -3475440.82413   887996.11981  -3.914 0.003544 **
## A1                14.78948       84.71941   0.175 0.865281
## A2                -0.03575        0.03341  -1.070 0.312495
## A3                -2.02020        0.48727  -4.146 0.002499 **
## A4                -1.03277        0.21379  -4.831 0.000933 ***
## A5                -0.04912        0.22553  -0.218 0.832448
## A6              1825.54365      454.23394   4.019 0.003023 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 304 on 9 degrees of freedom
## Multiple R-squared:  0.9955,Adjusted R-squared:  0.9925
## F-statistic: 332.3 on 6 and 9 DF,  p-value: 0.0000000004853
solve(t(M) %*% M , tol = 1e-20) %*% t(M) %*% y
##                  [,1]
##    -3475440.84060007
## A1       14.78948545
## A2       -0.03574762
## A3       -2.02019513
## A4       -1.03276578
## A5       -0.04911940
## A6     1825.54366146
```

The coefficient estimates are identical. While the default tolerance flag for the solve function in *R* did not allow the function to run, if we adjust the tolerance we can get the function to run. Although we adjusted the tolerance and lost precision using the solve function, we do not lose it at a decimal place that is relevant to our analysis.

4. This problem will provide some examples intended to help you understand the condition number of a matrix.

   (a) Let Q be an $n \times n$ orthonormal matrix.. Show that $\|Qx\| = \|x\|$. (Hint write down what $\|Qx\|^2$ is as dot product). Show then that $\kappa(Q) = 1$. (Orthonormal matrices achieve the best possible condition number.)

   $\|Qx\| = \|x\|$

$$\begin{aligned} \|Q\mathbf{x}\| &= \|\mathbf{x}\| \\ \|Q\mathbf{x}\|^2 &= \|\mathbf{x}\|^2 \\ Q\mathbf{x} \cdot Q\mathbf{x} &= \mathbf{x} \cdot \mathbf{x} \\ (Q\mathbf{x})^T Q\mathbf{x} &= \\ \mathbf{x}^T \underbrace{Q^T Q}_{1} \mathbf{x} &= \\ \mathbf{x}^T \mathbf{x} &= \\ \mathbf{x} \cdot \mathbf{x} &= \mathbf{x} \cdot \mathbf{x} \\ \|\mathbf{x}\| &= \|\mathbf{x}\| \end{aligned}$$

   We note $A\mathbf{w} = \lambda\mathbf{w}$, where $\mathbf{w}$ is an eigenvector and λ is an eigenvalue.

$$\begin{aligned} \|Q\mathbf{v}\|^2 &= \|\lambda\mathbf{v}\|^2 \\ Q\mathbf{v} \cdot Q\mathbf{v} &= (\lambda\mathbf{v})^T \lambda\mathbf{v} \\ (Q\mathbf{v})^T Q\mathbf{v} &= (\lambda\mathbf{v})^T \lambda\mathbf{v} \\ \mathbf{v}^T Q^T Q\mathbf{v} &= \lambda^2 \left(\mathbf{v}^T\mathbf{v}\right) \\ \mathbf{v}^T\mathbf{v} &= \lambda^2 \left(\mathbf{v}^T\mathbf{v}\right) \\ \pm 1 &= \lambda \end{aligned}$$

   Since $\kappa(Q) = \dfrac{\lambda_{max}}{\lambda_{min}}$, the fraction for $\kappa(Q)$ will either be $\dfrac{-1}{-1}$ or $\dfrac{1}{1}$, so $\kappa(Q) = 1$.

   (b) Consider the following matrix:

$$M = \begin{pmatrix} a & a\cos(\theta) \\ 0 & a\sin(\theta) \end{pmatrix} \qquad (2)$$

   where $a \in \mathbb{R}$ and $\theta \in [0, \pi/2]$. Notice that as $\theta \to 0$, the two columns of M become closer to being linearly dependent but that a simply multiplies both columns. The determinant and condition number both provide information on the columns forming the matrix. How does a affect the determinant and the condition number? How does θ affect the determinant and the condition number? Explain why the condition number is a better measure of linear independence than the determinant.

   As $|a|$ increases, the value of the determinant also increases, as long as $theta \neq 0$. The condition number stays constant no matter the value of a.

   Because the determinant is $\sin(\theta) \cdot a^2$, the determinant is just the sin function being scaled by some constant. It should always be increasing with $\theta \in \left[0, \frac{\pi}{2}\right]$ and the scalar is always positive. However, as $\theta \to \dfrac{\pi}{2}$ the condition number decreased regardless of a.

   While the determinant can tell us if the matrix is linearly independent or not; it really does not measure how "strong" the independence is. For example consider $\theta = \frac{\pi}{2}$ and $a = 1, 2, 3, 4,$ or 5. For eah different a we get a different determinant (1,4,9,16,25). We know this matrix is linearly independent for the values of $a \neq 0$, but there are infinitely many different determinants based on what the value of a is. However, regardless of a the condition number is 1. So we know this is "strongly" linearly independent. The condition number is giving me some sort of quality measure of the independence which is important to determine the computational singularity.