HOMEWORK PROBLEMS 05, ANLY 561, FALL 2017
DUE 10/13/17

Exercises:

1. Create Python functions

   `logistic_objective(x,y)`, `dlogistic_objective(x,y)`, `d2logisitic_objective(x,y)`

   satisfying the following specifications:

   - All functions expect two arrays/numpy arrays `x` and `y` where `x[i]` $= x_i$, `y[i]` $= y_i$, and $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R} \times \{-1, 1\}$ are training data for logistic regression.
   - `logistic_objective(x,y)` returns a *function* `f` satisfying the following specifications:
     - `f` expects a single array/numpy array `b` such that `b[0]` $= \beta_0$ and `b[1]` $= \beta_1$ for some logistic model parameters (β_0, β_1)
     - `f(b)` computes the negative log-likelihood of the parameters (β_0, β_1). That is,

     $$\texttt{f(b)} = \ell(\beta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i(\beta_0 + \beta_1 x_k)}\right).$$

   - `dlogistic_objective(x,y)` returns a *function* `df` satisfying the following specifications:
     - `df` expects a single array/numpy array `b` such that `b[0]` $= \beta_0$ and `b[1]` $= \beta_1$ for some logistic model parameters (β_0, β_1)
     - `df(b)` computes the gradient of negative log-likelihood at (β_0, β_1). That is,

     $$\texttt{df(b)} = \nabla_\beta \ell(\beta) = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0}(\beta) \\ \frac{\partial \ell}{\partial \beta_1}(\beta) \end{pmatrix}.$$

   - `d2logistic_objective(x,y)` returns a *function* `d2f` satisfying the following specifications:
     - `d2f` expects a single array/numpy array `b` such that `b[0]` $= \beta_0$ and `b[1]` $= \beta_1$ for some logistic model parameters (β_0, β_1)
     - `d2f(b)` computes the Hessian of negative log-likelihood at (β_0, β_1). That is,

     $$\texttt{d2f(b)} = \nabla_\beta^2 \ell(\beta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2}(\beta) & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1}(\beta) \\ \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1}(\beta) & \frac{\partial^2 \ell}{\partial \beta_1^2}(\beta) \end{pmatrix}$$

   Use these implementations and the data from Exercise 2 in Homework 04 to perform backtracking with both gradient descent and Newton increments. For the Newton increments, note that the computation of $\left(\nabla_\beta^2 \ell(\beta)\right)^{-1} \nabla_\beta \ell(\beta)$ is carried out by the command

   ```
   dx_newt = - numpy.linalg.solve(d2f(b), df(b)).
   ```

   For each type of increment, initialize with $\beta^{(0)} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$, run 30 steps of backtracking and provide a y-semilog plot of the point residuals $\{\|\beta^{(k+1)} - \beta^{(k)}\|\}_{i=1}^{30}$ and the function residuals $\{|\ell(\beta^{(k+1)}) - \ell(\beta^{(k)})|\}_{i=1}^{30}$. For backtracking, use $\alpha = 0.2$ and $\beta = 0.8$.

2. Consider the program

   $$\min_{(x,y)\in\mathbb{R}^2} 2x + 3y \text{ subject to } -1 \le x \le 1 \text{ and } -1 \le y \le 1.$$

   Such a program is called a **linear program** because the objective function and all the constraint functions are affine.

(a) Exhibit the KKT conditions for this program.

(b) Show that $(-1,-1)$ is the only point which satisfies the KKT conditions, and that $(1,1)$ satisfies all the KKT conditions except dual feasibility.

(c) Explain why the point $(0,0)$ is an interior point of this program, and carry out the log-barrier method to numerically produce a solution to this program using $(0,0)$ to initialize. Use $M=10$, 10 centering steps, 5 iterations in the outer loop, and 3 iterations in each inner loop. Whenever backtracking is called, use Newton increments and the standard parameters $\alpha=0.2$ and $\beta=0.8$. Display the answer you compute. You may find it helpful to note that

$$\nabla\phi(x,y)=\begin{pmatrix}-\frac{a}{ax+by+c}\\ -\frac{b}{ax+by+c}\end{pmatrix}\text{ and }\nabla^2\phi(x,y)=\begin{pmatrix}\frac{a^2}{(ax+by+c)^2} & \frac{ab}{(ax+bx+c)^2}\\ \frac{ab}{(ax+bx+c)^2} & \frac{b^2}{(ax+by+c)^2}\end{pmatrix}$$

for $\phi(x,y)=-\log(-(ax+bx+c))$ on the set $\{(x,y)\in\mathbb{R}^2 : ax+by+c<0\}$ and where $a,b,c\in\mathbb{R}$.

3. For a function $f:\mathbb{R}^n\to\mathbb{R}$ defined by $f(\mathbf{x})=f(x_1,x_2,\dots,x_n)$ where $\mathbf{x}=(x_1,\ x_2,\dots,\ x_n)$, if all partial derivatives of f, $\frac{\partial f}{\partial x_i}$, exist and are continuous, we write $f\in C^1(\mathbb{R}^n)$ and define the **gradient** of f as $\nabla f:\mathbb{R}^n\to\mathbb{R}^n$ given by

$$\nabla f(\mathbf{x})=\begin{pmatrix}\frac{\partial f}{\partial x_1}(\mathbf{x})\\ \frac{\partial f}{\partial x_2}(\mathbf{x})\\ \vdots\\ \frac{\partial f}{\partial x_n}(\mathbf{x})\end{pmatrix}.$$

If in addition we have $g\in C^1(\mathbb{R})$, then $g\circ f:\mathbb{R}^n\to\mathbb{R}$ defined by $(g\circ f)(\mathbf{x})=g(f(\mathbf{x}))$ is also in $C^1(\mathbb{R})$. Prove the chain rule:

$$\nabla(g\circ f)(\mathbf{x})=g'(f(\mathbf{x}))\nabla f(\mathbf{x}).$$

4. We represent a function $f:\mathbb{R}^n\to\mathbb{R}^m$ in terms **coordinate** or **component** functions $f_i:\mathbb{R}^n\to\mathbb{R}$ so that

$$f(x_1,\ x_2,\dots,\ x_n)=\begin{pmatrix}f_1(x_1,\ x_2,\dots,\ x_n)\\ f_2(x_1,\ x_2,\dots,\ x_n)\\ \vdots\\ f_m(x_1,\ x_2,\dots,\ x_n)\end{pmatrix}.$$

If all the first-order partial derivatives of the component functions exist and are continuous, we say that $f\in C^1(\mathbb{R}^n;\mathbb{R}^m)$ and we define the **Jacobian** of f by

$$Df(\mathbf{x})=\begin{pmatrix}\frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x})\\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x})\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x})\end{pmatrix}$$

Note that $Df:\mathbb{R}^n\to M_{m,n}$ is a matrix-valued function.

(a) Show that the ith row of $Df(x)$ is simply $\nabla f_i(\mathbf{x})^T$ for all $i=1,\dots,m$ so that

$$Df(\mathbf{x})=\begin{pmatrix}\nabla f_1(\mathbf{x})^T\\ \nabla f_2(\mathbf{x})^T\\ \vdots\\ \nabla f_n(\mathbf{x})^T\end{pmatrix}.$$

(b) If $g\in C^1(\mathbb{R}^m)$, then $g\circ f:\mathbb{R}^n\to\mathbb{R}$ is in $C^1(\mathbb{R}^n)$. Prove the **chain rule**:

$$\nabla(g\circ f)(\mathbf{x})=Df(\mathbf{x})^T\nabla g(f(\mathbf{x})).$$

(c) If $g\in C^1(\mathbb{R}^k,\mathbb{R}^m)$ and $f\in C^1(\mathbb{R}^n,\mathbb{R}^k)$, then $g\circ f\in C^1(\mathbb{R}^n,\mathbb{R}^m)$. Prove the **chain rule**:

$$D(g\circ f)(\mathbf{x})=Dg(f(\mathbf{x}))Df(\mathbf{x}).$$