Problem Sheet 2

Problems in Part A will be discussed in class. Problems in Part B come with solutions and should be tried at home.

Part A

(2.1) Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be a sequence of vectors in $\mathbb{R}^n$ and assume that this sequence converges linearly to some $\boldsymbol{x}^* \in \mathbb{R}^n$ with respect to a norm $\|\cdot\|$,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq r \cdot \|\boldsymbol{x}_k - \boldsymbol{x}^*\|, \quad k \geq 0,$$

for some constant $r \in (0, 1)$.

(a) Show that for $M = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq r^k \cdot M.$$

(b) Let $\varepsilon > 0$ be given. Show that if N is an integer such that

$$N > \frac{1}{1-r}\left(\ln(M) + \ln\left(\frac{1}{\varepsilon}\right)\right),$$

then $\|\boldsymbol{x}_N - \boldsymbol{x}^*\| \leq \varepsilon$. In words, N iterations are sufficient to reach accuracy ε.

(c) Now assume that the sequence converges quadratically,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq C\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

for some constant $C > 0$. Show that

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq C^k \cdot M^{2^k}.$$

How many steps will guarantee a solution up to an error ε?

(2.2) Consider the function

$$f(x) = \sqrt{x^2 + 1}.$$

Determine a value $\overline{x} \in \mathbb{R}$ such that

- for $x_0 < \overline{x}$, Newton's method converges to a minimizer,
- for $x_0 > \overline{x}$, Newton's method does not converge.

What happens if $x_0 = \overline{x}$ is chosen as starting point?

(2.3) Describe a steepest descent method with respect to the norm

$$\|\boldsymbol{x}\|_\infty = \max_{1\leq i\leq n} |x_i|.$$

What are the descent directions? How can the best descent direction be found?

Part B

(2.4) Iterative algorithms will never find the exact solution, so it is important to have suitable criteria for stopping an algorithm. Given a tolerance $\varepsilon > 0$, consider the following candidates for stopping criteria:

(a) $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \varepsilon$;

(b) $\|\nabla f(\boldsymbol{x}_k)\| < \varepsilon$;

(c) $k = 100$.

Apply Newton's method for minimizing the function

$$f(x) = (x-1)^6$$

(in Python, MATLAB, or by hand) with each of the given stopping criteria. Which of these is the most efficient? In general, describe the benefits or disadvantages of each of these stopping criteria.

(2.5) When implementing descent algorithms one often uses *backtracking* to select a good step length. Given a descent direction $\boldsymbol{p}_k$, one first tries the step length 1 and then successively scales it down until the *sufficient decrease* condition (Lecture 4) is satisfied.

Fixing a decrease parameter $c \in (0, 1/2)$ and a scaling parameter $s \in (0, 1)$, backtracking works as follows.

- $\alpha = 1$;
- while $f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k) \geq f(\boldsymbol{x}_k) + \alpha \cdot c \cdot \langle \boldsymbol{p}_k, \nabla f(\boldsymbol{x}_k) \rangle$: $\alpha = \alpha \cdot s$;
- $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha \boldsymbol{p}_k$

Consider the function

$$f(\boldsymbol{x}) = e^{x_1 + 3x_2 - 0.1} + e^{x_1 - 3x_2 - 0.1} + e^{-x_1 - 0.1}$$

on $\mathbb{R}^2$, with level sets given in the contour plot in Figure 1. Using the starting point $\boldsymbol{x}_0 = (-1, 0.7)^\top$, plot the trajectory of

(a) Gradient descent with step length 0.1;

(b) Gradient descent with backtracking, using $c = 0.1$ and $s = 0.5$;

(c) Newton's method.

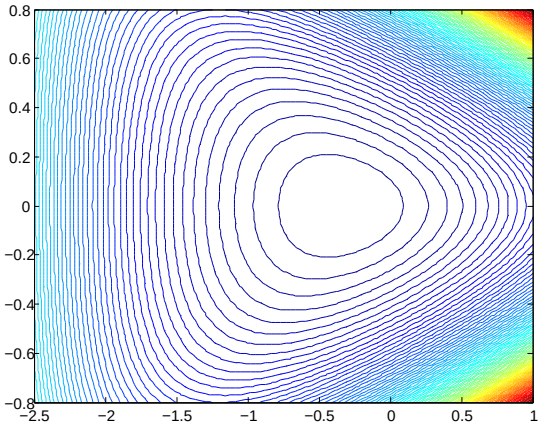

Figure 1: Contour plot

(2.6) (Logistic Regression.) Let Y be a random variable that satisfies

$$\mathbb{P}\{Y = 1\} = p, \quad \mathbb{P}\{Y = 0\} = 1 - p,$$

where $\mathbb{P}\{\}$ denotes the probability of an event. This random variable models an event with two possible outcomes. The *logistic model* for p has the form

$$p = \frac{e^{\langle \boldsymbol{a}, \boldsymbol{u} \rangle + b}}{1 + e^{\langle \boldsymbol{a}, \boldsymbol{u} \rangle + b}}, \tag{1}$$

$\boldsymbol{u} \in \mathbb{R}^n$ is a vector of *explanatory variables*, and $\boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}$ are the *model parameters* that explain how p depends on the variables. For example, the variable Y could represent a choice in an election and the vector $\boldsymbol{u}$ encodes demographic information, or the variable Y could indicate the presence of a disease and $\boldsymbol{u}$ encodes data about a patient.

Given vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and corresponding observed outcomes $y_1, \ldots, y_m$, we would like to estimate the parameters $\boldsymbol{a}$ and b. Assuming the observed outcomes $y_1 = \cdots = y_k = 1$ and $y_{k+1} = \cdots y_m = 0$, the *log likelihood* function is defined as

$$\sum_{i=1}^{k} \ln(p_i) + \sum_{i=k+1}^{m} \ln(1 - p_i), \tag{2}$$

where p_i is the probability (1) computed using $\boldsymbol{u}_i$, and the goal is to find parameters $\boldsymbol{a}, b$ that maximize this function.

(a) Show that the negative of the log likelihood function is given by

$$f(\boldsymbol{a}, b) = -\sum_{i=1}^{k} (\langle \boldsymbol{a}, \boldsymbol{u}_i \rangle + b) + \sum_{i=1}^{m} \ln\left(1 + e^{\langle \boldsymbol{a}, \boldsymbol{u}_i \rangle + b}\right)$$

Show that the sum of convex functions and the composition of a convex function with a linear one are convex, and deduce that the log likelihood function is convex.

(b) Consider the following table, in which the first row represents the hours of prepration and the second row whether an exam was passed (1) or not (0). Find the

Hours	0.5	1	1.5	2	2.5	3	3	4.5	4	4.5	4.75	5
Pass	0	0	0	1	0	1	0	1	1	1	1	1

(negative) log likelihood function $f(a, b)$. Solve the corresponding minimization problem

$$\text{minimize} \quad f(a, b)$$

in two dimensions using either gradiend descent or Newton's method, and plot the resulting probability function (1), giving the relationship of hours of study to probability of success.