Numerical Optimization - Hand-In 1

Isabela Blucher

February 9, 2018

Exercise 2.1

Firstly, we calculate the gradient $\nabla f(x)$ of the Rosenbrock function $f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$.

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= 200(x_2 - x_1^2)(-2x_1) + 2(1 - x_1)(-1) \\ &= -400(x_2x_1 - x_1^3) - 2(1 - x_1) = 400x_1^3 - 400x_1x_2 + 2x_1 - 2\end{aligned}$$

$$\frac{\partial f}{\partial x_2} = 200(x_2 - x_1^2)(1) + 2(0) = 200x_2 - 200x_1^2$$

Now that we have our gradient, the Hessian $\nabla^2 f(x)$ can be calculated.

$$\frac{\partial^2 f}{\partial x_1^2} = 1200x_1^2 - 400x_2 + 2 \qquad \frac{\partial^2 f}{\partial x_1 x_2} = \frac{\partial^2 f}{\partial x_2 x_1} = -400x_1 \qquad \frac{\partial^2 f}{\partial x_2^2} = 200$$

Since the Rosenbrock function is continuous and differentiable it is possible to do first and second derivative tests to verify its stationary points. In the first derivative test, we solve the equations to find the stationary point (1, 1). In the second derivative test, using the (1, 1) in the Hessian matrix and calculating its determinant we find $D(1,1) = det(H(1,1)) = f_{xx}(1,1)f_{yy}(1,1) - (f_{xy}(1,1))^2 = 160400 - 160000 = 400$. Since the determinant is positive and so is $f_{xx}(1,1)$, (1, 1) is a the only local minimum. To verify that the Hessian at (1,1) is positive definite we calculate its eigenvalues, which are $501 + \sqrt{250601}$ and $501 - \sqrt{250601}$, and since they are positive, the matrix is positive definite.

Exercise 2.2

For the function $f(x_1, x_2) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$ we take the partial derivatives and equal them to zero, to find the stationary points.

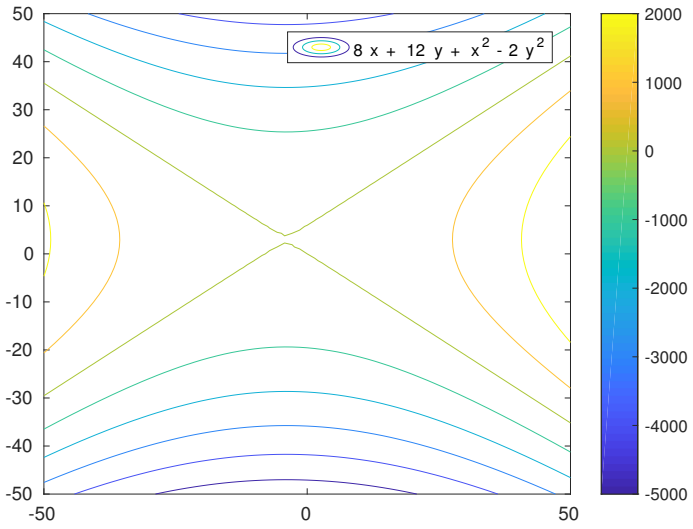
$$\frac{\partial f}{\partial x_2} = 12 - 4x_2 \qquad \frac{\partial f}{\partial x_2} = 12 - 4x_2$$

With that, we find that the only stationary point possible is (-4, 3). To verify if it is a maximum, minimum or saddle point, we compute the determinant of the Hessian matrix.

$$\frac{\partial^2 f}{\partial x_1^2} = 2 \qquad \frac{\partial^2 f}{\partial x_1 x_2} = 0 \qquad \frac{\partial^2 f}{\partial x_2^2} = -4$$

The determinant is -8. Because the determinant of the Hessian at (-4, 3) is negative, we can say that there is a saddle point in these coordinates.

Below a contour plot of the given function is shown, with both axes scaled from -50 to 50.



Exercise 2.3

We have that a is a given n-vector and A is a nxn symmetric matrix. Firstly, we'll compute the gradient and Hessian of $f_1(x) = a^T x = \sum_{i=1}^{n} a_i x_i$.

$$\nabla f_1(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} \\ \vdots \\ \frac{\partial f_1}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

$$\nabla^2 f_1(x) = \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_1^2} & \frac{\partial^2 f_1}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f_1}{\partial x_n x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_1}{\partial x_n x_1} & \frac{\partial^2 f_1}{\partial x_n x_2} & \cdots & \frac{\partial^2 f_1}{\partial x_n^2} \end{bmatrix} = 0$$

Now we do the same for $f_2(x) = x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$.

$$\nabla f_2(x) = \begin{bmatrix} \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_2}{\partial x_2} \\ \vdots \\ \frac{\partial f_2}{\partial x_n} \end{bmatrix} = \left[\sum_j A_{sj} x_j + \sum_i A_{is} x_i\right]_s = 1..n = 2Ax$$

That last step is valid due to A being symmetric.

$$\nabla^2 f_2(x) = \begin{bmatrix} \frac{\partial^2 f_2}{\partial x_1^2} & \frac{\partial^2 f_2}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f_2}{\partial x_n x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_2}{\partial x_n x_1} & \frac{\partial^2 f_2}{\partial x_n x_2} & \cdots & \frac{\partial^2 f_2}{\partial x_n^2} \end{bmatrix} = \left[\frac{\partial^2 \sum_i \sum_j A_{ij} x_i x_j}{\partial x_s \partial x_t}\right]_{s=1..n, t=1..n} = \left[A_{st} + A_{ts}\right]_{s=1..n, t=1..n} = 2A$$

Exercise 2.4

The second-order Taylor expansion is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2$$

Applying it to $f_1(x) = \frac{1}{cos(x)}$ around a non-zero point, we get

$$f(x) = \frac{1}{cos(x_0)} + \frac{1}{x^2}sin(\frac{1}{x})(x - x_0) + (\frac{-1}{2x^4})(cos(\frac{1}{x}) + 2xsin(\frac{1}{x}))(x - x_0)^2$$

The third-order Taylor expansion ins

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f^{(3)}(x_0)}{3!}(x - x_0)^3$$

Applying it to $f_2(x) = cos(x)$ around any point x, we get

$$f(x) = cos(x_0) + -sin(x)(x - x_0) + \frac{-cos(x)}{2}(x - x_0)^2 + \frac{sin(x)}{6}(x - x_0)^3$$

For the specific case where $x_0 = 1$ we have

$$f(x) = cos(1) + -sin(1)(x - 1) + \frac{-cos(1)}{2}(x - 1)^2 + \frac{sin(1)}{6}(x - 1)^3$$

Programming Assignment

Rosenbrock function

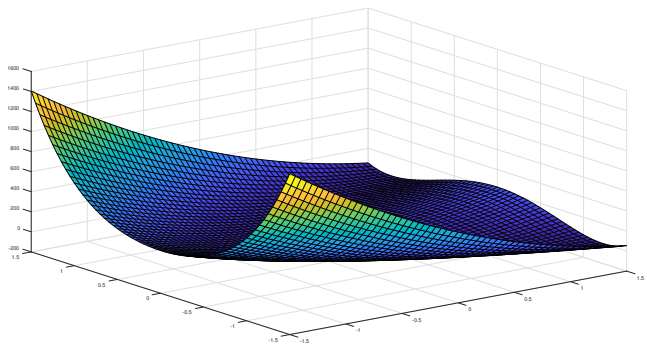
Implementing the Rosenbrock function in MATLAB, and verifying its gradient and Hessian at the local minimizer (1,1) we get, as expected, a zero vector and a positive-definite matrix, respectively.

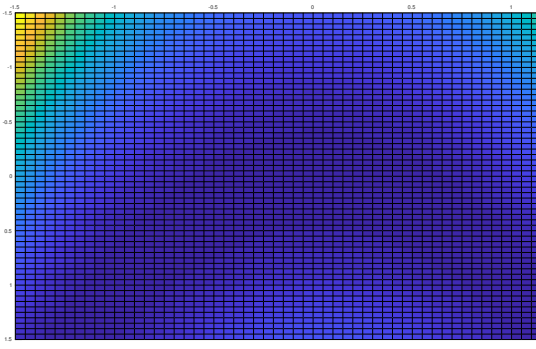
```
rosegrad(1,1) = [0 0]
rosehess(1,1) =  [802  -400
                 -400   200]
```

Computing the gradient and Hessian of the Rosenbrock function on the points (0,0) and (-1, -1) we get the following vectors and matrices

```
rosegrad(0,0) = [-2 0]
rosegrad(-1,-1) = [-804 -400]
rosehess(0,0) =  [2     0
                  0   200]
rosehess(-1, -1) = [1602     400
                    400      200]
```

After computing the gradient and Hessian, the Rosenbrock function was also plotted. I am posting two images, so that in one of them the "banana shape" is seen from above. The points (x, y) generated to plot the function are from -1.5 to 1.5 in steps of size 0.05.

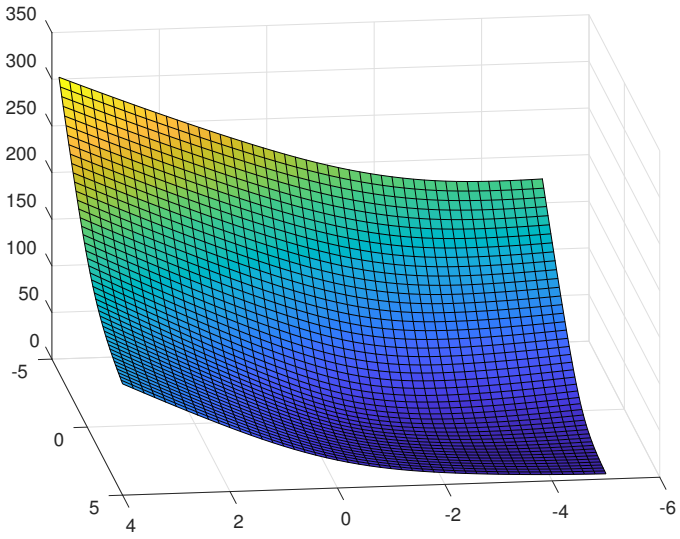




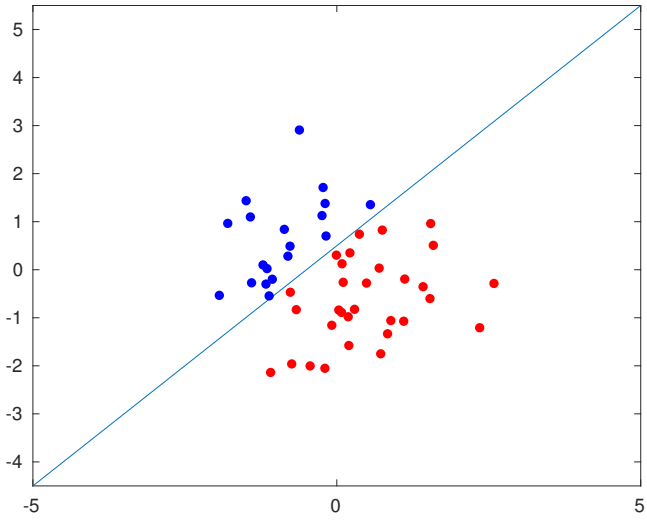
Logistic Regression

After the implementation of the error function $L(\theta) = \sum_{i=1}^{l} log(1 + exp(-y^{(i)} \cdot f_\theta(x^{(i)})))$ and generation of the random dataset with 50 2D points from a standard normal distribution, θ was set to $\theta^* = (0.5, 1, -1)$ and the labels were chosen as $y = sign(f_\theta(x))$.

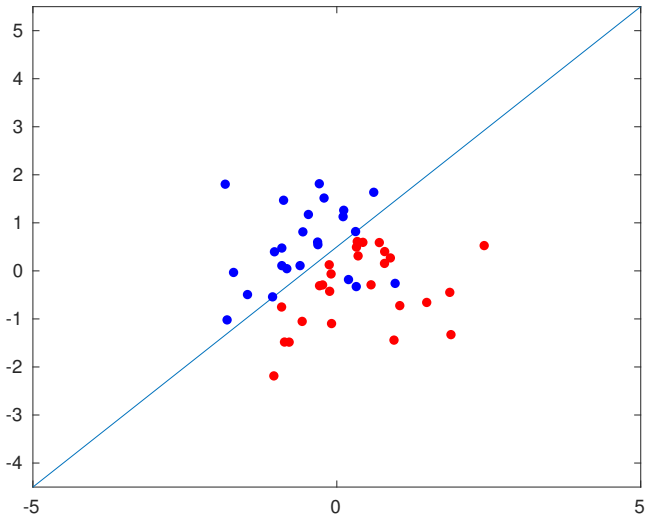
The following graph is a plot of $L(\theta)$ where $\theta_0 = 0.5$, using a surface with a meshgrid in the interval -5 to 5 with steps of size 0.2. It was plotted using the previously generated dataset.



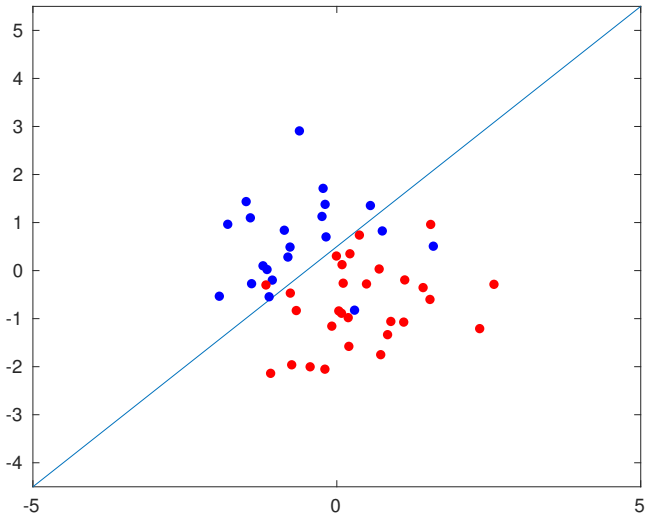
On $L(\theta)$ the model is seen to be optimal. With the help of the scatter plot below, one can see that it perfectly divides different labels. The red dots represent the data points that are labeled as +1 and the blue represents the ones that are labeled as -1. The $f_\theta(x) = 0$ line, perfectly separates the differently labeled data.



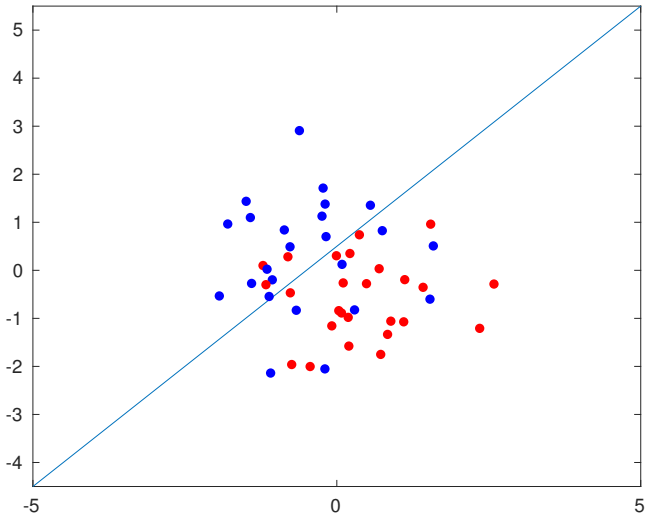
If the generation of the dataset label values is flipped with some given probability, the model will not be optimal anymore. As we can see in the graphs below, for probabilities of 5%, 10% and 20%, the data is increasingly mixed.



a) Labels flipped with 5% probability



b) Labels flipped with 10% probability



c) Labels flipped with 20% probability