Weekly Assignment 1

Group 4
Yi Liu djl174

February 11, 2021

In this assignment, first we aim to implement the first and second derivatives (gradient and Hessian) of the benchmark functions from f_1 to f_5. Afterwards, we discuss and test whether these functions are actually implemented correctly. Finally, we will present the 3D plots and contour maps of these benchmark functions.

1 Benchmark functions and the derivatives

1.1 Gradient

The following are the gradients (first derivatives) of the 5 benchmark functions.

$$
\begin{aligned}
\nabla f_1(x)_i &= \frac{\partial f(x)}{\partial x_i} = 2\alpha^{\frac{i-1}{d-1}} x_i \\
\nabla f_2(x) &= \begin{pmatrix} \frac{\partial f_2(x)}{\partial x_1} \\ \frac{\partial f_2(x)}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -2 \cdot (1-x_1) - 400 \cdot (x_2 - x_1^2)x_1 \\ 200(x_2 - x_1^2) \end{pmatrix} \\
\nabla f_3(x)_i &= \frac{\partial f_3(x)}{\partial x_i} = \frac{1}{f_1(x)+\epsilon} \cdot \nabla f_1(x)_i \\
\nabla f_4(x)_i &= \frac{\partial f_4(x)}{\partial x_i} = \frac{\partial h(x_i)}{\partial x_i} + 100 \cdot \frac{\partial h(-x_i)}{\partial x_i} \\
\nabla f_5(x)_i &= \frac{\partial f_5(x)}{\partial x_i} = \frac{\partial h(x_i)^2}{\partial x_i} + 100 \cdot \frac{\partial h(-x_i)^2}{\partial x_i} = 2h(x_i)\frac{\partial h(x_i)}{\partial x_i} + 200 \cdot h(-x_i)\frac{\partial h(-x_i)}{\partial x_i}
\end{aligned}
$$

where $\frac{\partial h(x_i)}{\partial x_i} = \frac{1}{1+e^{(-qx_i)}}$ and $\frac{\partial h(-x_i)}{\partial x_i} = \frac{1}{1+e^{(qx_i)}}$

1.2 Hessian

Now consider the Hessian(second derivatives) of these benchmark functions. To show Hessian is a diagonal matrix with $(Hf(x))_{ii} = g_i^{''}(x_i)$. We take $f(x) = \sum_{i=1}^{N} g_i(x_i), x_i \in \mathbb{R}^n$. Thus we have

$$
\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i} = \frac{\partial g_i(x_i)}{\partial x_i} = g_i^{'}(x_i), \tag{1}
$$

and further

$$
(Hf(x))_{ij} = \nabla^2 f(x)_i = \frac{\partial^2 f(x)}{\partial x_i x_j} = \frac{\partial \nabla f(x)_i}{\partial x_j} = \frac{\partial g_i^i(x_i)}{\partial x_j} = \begin{cases} g_i^{''}(x_i), & i = j \\ 0, & i \neq j \end{cases} \tag{2}
$$

So we can say the Hessian is a diagonal matrix with $(Hf(x))_{ii} = g_i^{''}(x_i)$. Hence, we have the following Hessian corresponding to our five benchmark functions.

$$
\begin{aligned}
\nabla^2 f_1(x) &= (Hf_1(x))_{ii} = \frac{\partial^2 f(x)}{\partial x_i^2} = 2 \cdot \alpha^{\frac{i-1}{d-1}} \\
\nabla^2 f_2(x) &= (Hf_2(x_1, x_2)) = \begin{pmatrix} 2 - 400 \cdot x_2 + 1200 \cdot x_1^2 & -400 \cdot x_1 \\ -400 \cdot x_1 & 200 \end{pmatrix} \\
\nabla^2 f_3(x) &= (Hf_3(x))_{ii} = -\frac{\nabla f_1(x)_j}{(\epsilon + f_1(x)^2} \cdot \nabla f_1(x)_i + \frac{1}{\epsilon + f_1(x)}(Hf_1(x))_{ij} \\
\nabla^2 f_4(x) &= (Hf_4(x))_{ii} = \frac{\partial^2 h(x)}{\partial x_i^2} + 100 \cdot \frac{\partial^2 h(-x_i)}{\partial x_i^2} \\
\nabla^2 f_5(x) &= (Hf_5(x))_{ii} = 2(\frac{\partial^2 h(x)}{\partial x_i} + h(x_i) \cdot \frac{\partial^2 h(x)}{\partial x_i^2} + 100 \cdot (\frac{\partial h(-x_i)}{\partial x_i} + h(-x_i) \cdot \frac{\partial^2 h(x)}{\partial x_i^2}))
\end{aligned}
$$

where $\frac{\partial^2 h(x)}{\partial x_i^2} = \frac{q}{2+exp(qx_i)+exp(-qx_i)} = -\frac{\partial^2 h(-x_i)}{\partial x_i^2}$.

1.3 The minimizer problem

By theorem 2.1, set $\nabla f_2(x) = \begin{pmatrix} -2 \cdot (1 - x_1) - 400 \cdot (x_2 - x_1^2)x_1 \\ 200(x_2 - x_1^2) \end{pmatrix} = 0$, then we obtain a stationary point $x^* = (0, 0)$.
By Theorem 2.4 we find the strict local minimizer of the Rosenbrock function if $\nabla f(x^*) = 0$ and $\nabla^2 f_2(x)$ is positive definite. Since $\nabla^2 f_2(x) = \begin{pmatrix} 2 - 400 \cdot x_2 + 1200 \cdot x_1^2 & -400 \cdot x_1 \\ -400 \cdot x_1 & 200 \end{pmatrix}$. Therefore the determinant $\nabla^2 f_2(1, 1) = 400$ and the trace of $\nabla^2 f_2(1, 1) = 1002$ are strictly positve. So $\nabla^2 f_2(1, 1)$ must be positive definite(PD), hence the point (1,1) is a minimizer of the Rosenbrock function.

Similarly we can identify point $(0, 0)$ is the minimizer of the Ellipsoid function. Since $\nabla f_1(0, 0) = \begin{pmatrix} 2\alpha^0 \\ 2\alpha \cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\nabla^2 f_1(0, 0) = \begin{pmatrix} 2\alpha^0 & 0 \\ 0 & 2\alpha \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2000 \end{pmatrix}$, and hence $det\nabla^2 f_1(0, 0) = 4000 > 0$ and $trace\nabla^2 f_1(0, 0) = 2002 > 0$ guarantees the point $(0, 0)$ is a minimizer.

1.3.1 The Attractive-Sector Functions

Given the functions
$f_4(x) = \sum_{i=1}^{d} h(x_i) + 100 \cdot h(-x_i)$
$f_5(x) = \sum_{i=1}^{d} h(x_i)^2 + 100h(-x_i)^2$ with $h(x) = \frac{log(1+exp(q \cdot x}{q}$, where $q = 10^8$.
Here we introduce $log(1 + exp(x)) = log(1 + exp(-|x|) + max(x, 0)$ for the implementations of $h(x)$ in the attractive-sector functions.
If $x < 0$:

$$
log(1 + exp(x)) = log(1 + exp(x)) + 0 = log(1 + exp(x)) \tag{3}
$$

If $x \geq 0$:

$$
\begin{aligned}
log(1 + exp(x)) = log(1 + exp(x)) - x + x &= log(1 + exp(x)) - log(exp(x)) + x \\
&= log\left(\frac{1 + exp(x)}{exp(x)}\right) + max(x, 0) \\
&= log(1 + exp(-x)) + max(x, 0) \\
&= log(1 + exp(-|x|)) + max(x, 0).
\end{aligned}
$$

$log(1 + exp(-|x|) + max(x, 0)$ is more like a linear function, we don't have to worry about overflow if x is too large, so the error is very small. With LHS we have $log(1 + exp(x))$, which can cause overflow if x is large, since the corresponding exponential increases rapidly.

2 Testing

Suppose $g : \mathbb{R} \to \mathbb{R}$ is a univariative function and let $\epsilon \in \mathbb{R}$. According to Taylor's theorem, we have

$$g(x + \epsilon) = g(x) + \nabla g(x + t\epsilon)^T \epsilon \tag{4}$$

for some $t \in (0, 1)$. Moreover, if g is twice continuously differentiable, we have

$$\nabla g(x + \epsilon) = \nabla g(x) + \int_0^1 \nabla^2 g(x + t\epsilon)\, dt \tag{5}$$

and that

$$g(x + \epsilon) = g(x) + \nabla g(x)^T \epsilon + \frac{1}{2}\epsilon^2 \nabla^2 g(x + t\epsilon)\epsilon \tag{6}$$

then we have

$$\nabla g(x)^T = \frac{g(x + \epsilon) - g(x)}{\epsilon} - \frac{1}{2}\epsilon^2 \int_0^1 \nabla^2 g(x + t\epsilon)\, dt \tag{7}$$

If the ϵ is finite and sufficiently small then $\frac{1}{2}\epsilon^2 \int_0^1 \nabla^2 g(x + t\epsilon)\, dt$ will be converge to 0. Hence the first derivative $g^{'}(x)$

$$\frac{dg}{dx} = g^{'}(x) \stackrel{def}{=} \lim_{\epsilon \to 0} \frac{g(x + \epsilon) - g(x)}{\epsilon} \tag{8}$$

is a good approximation of the gradient. Similarly, the second derivative is obtained by substituing g by $g^{'}$ in this same formula, that is

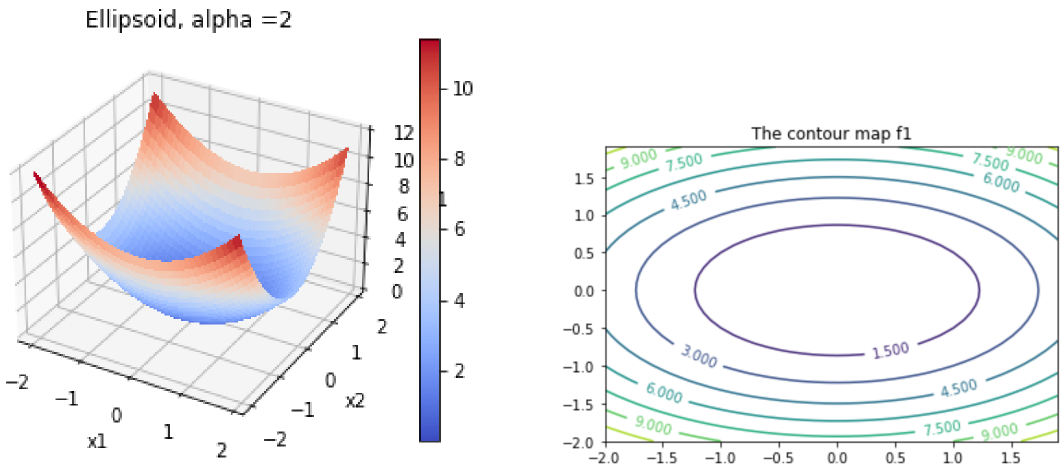
$$\frac{d^2 g}{dx^2} = g^{''}(x) \stackrel{def}{=} \lim_{\epsilon \to 0} \frac{g^{'}(x + \epsilon) - g^{'}(x)}{\epsilon} \tag{9}$$

a good approximation for the gradient as well.
Equivalently, we have

$$\lim_{\epsilon \to 0} \frac{g(x + \epsilon) - g(x)}{\epsilon} = \lim_{\epsilon \to 0} \nabla g(x)^T \epsilon + \lim_{\epsilon \to 0} \frac{1}{2}\epsilon \nabla^2 g(x + t\epsilon)\epsilon \tag{10}$$

Where the LHS is our estimation, and the RHS we have the target and the estimation error respectively. So we tend to choose the small value of epsilon that close to the minimal point to reduce the error. Also, we need to consider the floating-point arithmetic, the smallest value of epsilon is 10^{-16}.

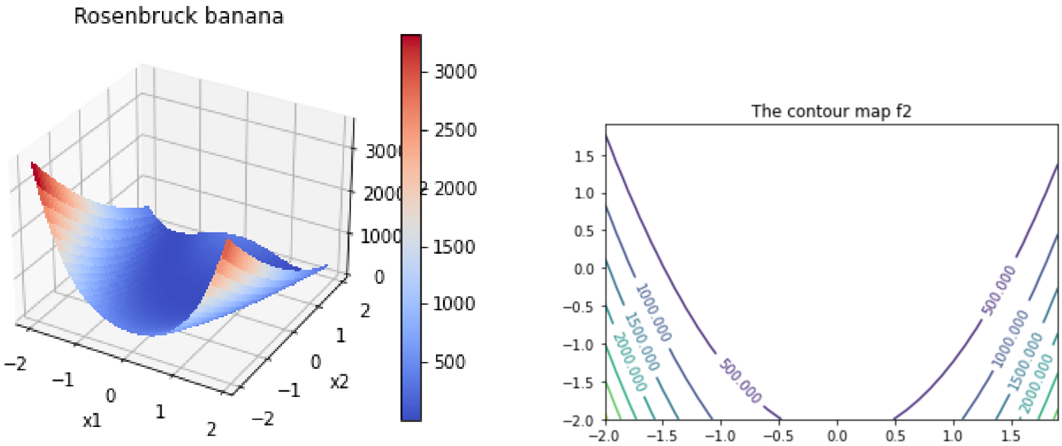
2.1 Plots



(a) figure 1 (b) figure 2

Figure 1: 3D plot and Contour map of the Ellipsoid function

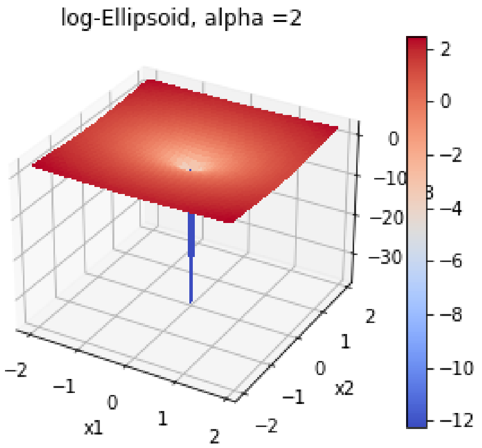
The above plot is overall nicely behaved, convex and has a global minimum. For the Ellipsoid function in $\mathbb{R}^2$, We can see the lower function values are depicted in blue with global minimum at point $(0, 0)$.



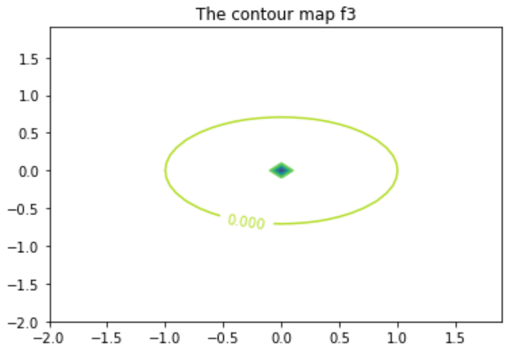
(a) figure 1 (b) figure 2

Figure 2: 3D plot and Contour map of the Rosenbrock Banana Function

The Rosenbrock function is not as nice as the Ellipsoid functions because of it's death vally, and it can be challenge for optimisation. For the Rosenbrock function in $\mathbb{R}^2$, we can see the lower function values are depicted in blue with global minimum at (1, 1) which lies in a valley.



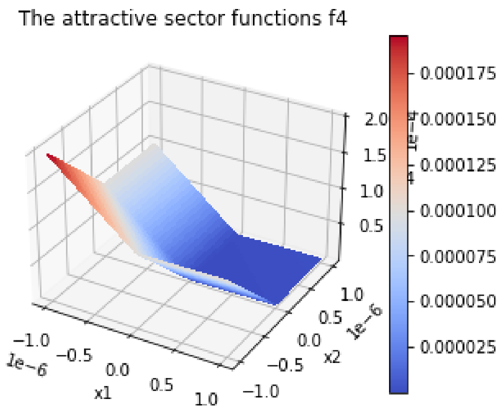
(a) figure 1



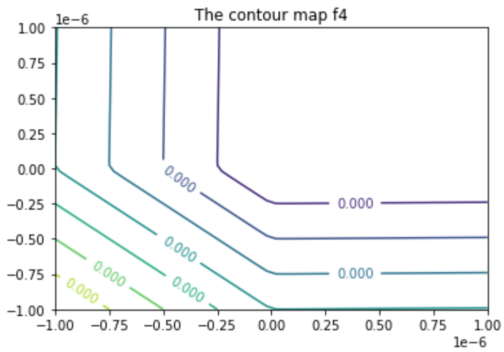
(b) figure 2

Figure 3: 3D plot and Contour map of the Log-Ellipsoid Function

The following Attractive-Sector functions are twisted, and they are non-convex and asymmetric around the minimum. For attractive sector funtions in $\mathbb{R}^2$, we can see the lower function values are depicted in blue with global minimum around (0,0).

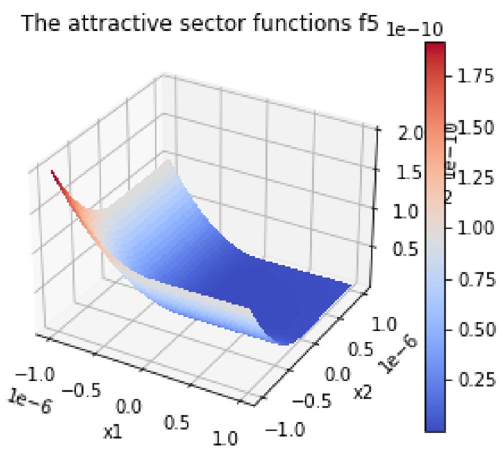


(a) figure 1

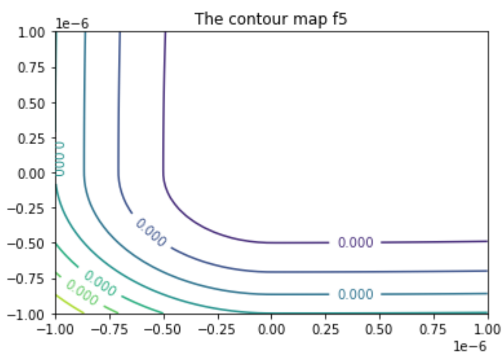


(b) figure 2

Figure 4: 3D plot and Contour map of the Attractive-Sector Functions f4



(a) figure 1



(b) figure 2

Figure 5: 3D plot and Contour map of the Attractive-Sector Functions f5