Numerical Optimization for Large Scale Problems
Assignment Report

Giacomo Maino s338682, Mattia Molinari s337194, Alessandro Perlo s337131

February 3^{rd}, 2025

Abstract

Contents

1 Introduction

In this section we will describe the implementation details of the algorithms used to solve the optimization problems, namely modified Newton method and truncated Newton method, focusing on the differences with respect to the standard Newton method. These methods will be tested against the Rosenbrock function and three test problems from [1]. The chosen test problems are the extended Rosenbrock function (problem 25), the generalized Broyden tridiagonal function (problem 32) and the banded trigonometric function (problem 16) and results are contained in sections 3, 4 and 5 respectively.

1.1 Modified Newton Method

The modified Newton method aims to enhance robustness of the standard Newton method by ensuring positive-definiteness of the Hessian matrix. At iteration k, it is necessary to check whether the Hessian matrix H_k is positive definite: in case it is not, the matrix is modified by adding a matrix B_k in order to ensure positive definiteness. A common choice for B_k is a multiple of the identity matrix, i.e. $B_k = \tau_k I$ so that the whole spectrum of H_k is shifted by τ_k. Then we want to find the smallest τ_k such that $H_k + \tau_k I$ is positive definite, which is $-\lambda_{k,\,min} + \beta$ where $\lambda_{k,\,min}$ is the negative eigenvalue with the largest module.

To avoid to have to compute $\lambda_{k,\,min}$, we adopted the *Cholesky with Added Multiple of the Identity* algorithm outlined in [2] that consists in building a sequence of τ_k until the modified matrix is positive definite. The

sequence is built starting from $\tau_k = \min_i h_{ii} + \beta$ where $\min_i h_{ii}$ is the smallest diagonal element of H_k. Then, at each iteration:

1. positive-definitess is assessed trying to perform a Cholesky factorization of $H_k + \tau_k I$;
2. if the factorization is not successful, τ_k is increased by a factor c and the process is repeated for a limited number of times $k_{chol,\,max}$.

In all the experiments we choose $\beta = 10^{-3}$, AP: [check value for $k_{chol,\,max}$] $k_{chol,\,max} = 100$. A good value for the constant factor is $c = 2$, but as we will discuss in section AP: [add reference] for the generalized Broyden tridiagonal function a larger value $c = 5$ is beneficial. The method is endowed with a line search strategy with backtracking.

1.2 Truncated Newton method

The truncated Newton method aims to reduce the computational cost of the Newton method by adopting the following strategies:

- the newton system $H_k p_k = -\nabla f(x_k)$ is solved approximately by means of an iterative method (i.e. conjugate gradient method), with a tolerance that depends on $\|\nabla f(x_k)\|$;
- whenever a direction of negative curvature is found in the execution of the iterative method, the method is stopped and the direction is used as the search direction to prevent a non-negative curvature direction to be chosen in case of a non-positive definite H_k.

In all the experiments, we choose the tolerance for the iterative method at iteration k to be

$$\eta_k = \min\{0.5,\ \sqrt{\|\nabla f(x_k)\|}\}$$

that is a forcing term that is proven to yield a superlinear convergence rate. The method is endowed with a line search strategy with backtracking.

1.3 Finite differences

Experiments in subsequent sections will adopt both exact and finite differences gradient and Hessian to perform the optimization. When finite differences are adopted, the gradient will be estimated using centered finite differences

$$\frac{\partial f}{\partial x_k} \approx \frac{f(x + he_k) - f(x - he_k)}{2h} \tag{1}$$

while the Hessian will be estimated using forward finite differences, using the following formula

$$\frac{\partial^2 f}{\partial x_k \partial x_j} \approx \frac{f(x + he_k + he_j) - f(x + he_k) - f(x + he_j) + f(x)}{h^2} \tag{2}$$

where e_k and e_j are the k-th and the j-th canonical basis vectors respectively. Moreover, two different approaches will be adopted to choose the step size h: the first one will use a fixed step size while the second one will use a step size that depends on the current point x and that is different for each component, defined as follows

$$h_{k,\,i} = h|x_{k,\,i}|$$

where $h_{k,\,i}$ is the increment for component i at step k, h is a relative step size and $x_{k,\,i}$ is the i-th component of the point at step k. Due to the large scale nature of the problems, the finite differences method is expected to be slower than the exact method, so ad-hoc implementations that will exploit the sparsity of the Hessian matrix and the separability of the specific functions will be used.

2 Rosenbrock function

2.1 Exact gradient and Hessian

2.2 Finite differences gradient and Hessian

3 Extended Rosenbrock function

The extended Rosenbrock function is a generalization of the Rosenbrock function to n dimensions, defined as follows. Figure 1 shows the surface plot of the 2-dimensional extended Rosenbrock function: notice that for $n = 2$

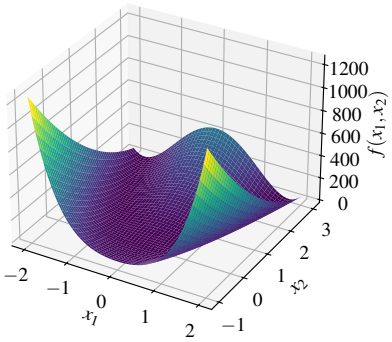

Figure 1: Surface plot of the 2-dimensional extended Rosenbrock function

Table 1: Results for Modified Newton method applied to Extended Rosenbrock with exact gradient and hessian, metrics are average metrics for succesful attempts.

preconditioning dimension	iterations		convergence rate		time		success rate	
	False	True	False	True	False	True	False	True
3	31.91	28.91	1.99	1.51	0.05	0.02	1.00	1.00
4	32.36	29.18	2.10	2.04	0.11	0.08	1.00	1.00
5	26.50	26.00	1.10	1.00	0.68	0.53	1.00	1.00

it is identical to the standard Rosenbrock function, except for the $\frac{1}{2}$ term.

$$F(x) = \frac{1}{2}\sum_{k=1}^{n} f_k^2(x), \qquad f_k(x) = \begin{cases} 10(x_k^2 - x_{k+1}), & k \mod 2 = 1 \\ x_{k-1} - 1, & k \mod 2 = 0 \end{cases} \tag{3}$$

The minimum of the function is in a very flat valley which is easy to reach, but in practice it's harder to converge to a minimum, which makes the extended Rosenbrock function a challenging optimization problem.

3.1 Exact gradient and Hessian

The gradient of the extended Rosenbrock function is given by the following expression,

$$\frac{\partial F}{\partial x_k} = \begin{cases} 200(x_k^3 - x_k x_{k+1}) + (x_k - 1), & k \mod 2 = 1 \\ -100(x_{k-1}^2 - x_k), & k \mod 2 = 0 \end{cases} \tag{4}$$

computation can be eased considering that component k depends only on f_k and f_{k+1} when k is odd, and only on f_{k-1} when k is even. The Hessian of the extended Rosenbrock function is given by the following expression.

$$\frac{\partial^2 F}{\partial x_k \partial x_j} = \begin{cases} 200(3x_k^2 - x_{k+1}) + 1, & j = k,\ k \mod 2 = 1 \\ 100, & j = k,\ k \mod 2 = 0 \\ -200x_k, & |k - j| = 1,\ k \mod 2 = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Notice that the Hessian is a sparse matrix, with only n non-zero elements on the diagonal and $n/2$ non-zero elements on the first co-diagonal.

3.2 Finite differences gradient and Hessian

When applying 1, one can notice that the terms $F(x + he_k)$ and $F(x - he_k)$ only differ by terms f_k and f_{k+1} for k odd, by terms f_{k-1} for k even. Then to make function evaluations less expensive, we can define the following

Table 2: Results for Truncated Newton method applied to Extended Rosenbrock with exact gradient and hessian, metrics are average metrics for succesful attempts.

preconditioning dimension	iterations		convergence rate		time		success rate	
	False	True	False	True	False	True	False	True
3	50.09	31.00	-2.66	11.55	0.01	0.01	1.00	1.00
4	57.27	34.73	-3.06	4.33	0.04	0.03	1.00	1.00
5	64.00	28.50	-5.58	1.62	0.45	0.23	1.00	1.00

function $F_{fd,\,k}$, which can be plugged in 1 in place of F yielding the same result.

$$F_{fd,\,k}(x) = \begin{cases} \frac{1}{2}f_k^2(x) + \frac{1}{2}f_{k+1}^2(x), & k \mod 2 = 1 \\ \frac{1}{2}f_{k-1}^2(x), & k \mod 2 = 0 \end{cases}$$

The same procedure can be applied for the Hessian, considering that:

- function evaluations to compute entry $h_{k,k}$ differ only by f_k and f_{k+1} for k odd, and only on f_{k-1} for k even;
- function evaluations to compute entry $h_{k,k+1}$ differ only by f_k and f_{k+1} for k odd.

Then to make function evaluations less expensive, we can define the functions $F_{fd,\,k,k}$ and $F_{fd,\,k,k+1}$, which can be plugged in 2 in place of F yielding the same result to compute entries $h_{k,k}$ and $h_{k,k+1}$ respectively.

$$F_{fd,\,k,k}(x) = \begin{cases} \frac{1}{2}f_k^2(x) + \frac{1}{2}f_{k+1}^2(x), & k \mod 2 = 1 \\ \frac{1}{2}f_{k-1}^2(x), & k \mod 2 = 0 \end{cases}$$

$$F_{fd,\,k,k+1}(x) = \begin{cases} \frac{1}{2}f_k^2(x) + \frac{1}{2}f_{k+1}^2(x), & k \mod 2 = 1 \\ 0, & k \mod 2 = 0 \end{cases}$$

When plugging the functions $F_{fd,\,k}$, $F_{fd,\,k,k}$ and $F_{fd,\,k,k+1}$ into 1 and 2 it's convenient to expand them so that the computation of the gradient and Hessian is not subject to numerical cancellation. After expanding the functions, the gradient and Hessian can be approximated as follows.

$$\frac{\partial F}{\partial x_k} \approx \begin{cases} 600h^2x_k - 100hx_{k+1} + \frac{1}{2}h + 350h^3 + 300hx_k^2, & k \mod 2 = 1 \\ -100x_{k-1}^2 + 100x_k, & k \mod 2 = 0 \end{cases}$$

$$\frac{\partial^2 F}{\partial x_k \partial x_j} \approx \begin{cases} 1200h_kx_k - 200x_{k+1} + 1 + 700h_k^2 + 600x_k^2, & j = k,\ k \mod 2 = 1 \\ 100, & j = k,\ k \mod 2 = 0 \\ -100h_kh_{k+1} - 200x_k, & |k-j| = 1,\ k \mod 2 = 1 \\ 0, & \text{otherwise} \end{cases}$$

4 Generalized Broyden tridiagonal function

The generalized Broyden tridiagonal function is defined as follows.

$$F(x) = \frac{1}{2}\sum_{i=1}^{n} f_k^2(x) \qquad f_k(x) = (3 - 2x_k)x_k + 1 - x_{k-1} - x_{k+1} \tag{6}$$

Figure 2 shows the surface plot of the 2-dimensional generalized Broyden tridiagonal function. Notice that the area where the minimum lies is very flat, which makes it hard to converge to the minimum.

4.1 Exact gradient and Hessian

The gradient of the generalized Broyden tridiagonal function is given by the following expression,

$$\frac{\partial F}{\partial x_k} = \begin{cases} (3 - 4x_1)f_1(x) - f_2(x), & k = 1 \\ (3 - 4x_k)f_k(x) - f_{k+1}(x) - f_{k-1}(x), & 1 < k < n \\ (3 - 4x_n)f_n(x) - f_{n-1}, & k = n \end{cases} \tag{7}$$

Table 3: Results for Modified Newton method applied to Extended Rosenbrock with absolute finite differences, metrics are average metrics for succesful attempts.

		iterations		convergence rate		time		success rate	
	preconditioning	False	True	False	True	False	True	False	True
dimension	h								
3	1e-02	149.91	150.82	1.00	1.00	0.07	0.07	1.00	1.00
	1e-04	33.18	32.09	1.01	1.00	0.02	0.02	1.00	1.00
	1e-06	32.09	30.18	1.96	2.09	0.03	0.02	1.00	1.00
	1e-08	32.00	30.18	2.04	2.23	0.02	0.02	1.00	1.00
	1e-10	32.00	30.18	2.00	2.23	0.02	0.02	1.00	1.00
	1e-12	32.00	30.18	2.00	2.23	0.02	0.02	1.00	1.00
4	1e-02	160.09	160.55	1.00	1.00	0.34	0.36	1.00	1.00
	1e-04	34.18	32.55	1.00	1.00	0.12	0.09	1.00	1.00
	1e-06	32.55	30.27	1.89	1.96	0.12	0.08	1.00	1.00
	1e-08	32.55	30.27	1.95	1.98	0.12	0.08	1.00	1.00
	1e-10	32.55	30.27	1.95	1.98	0.11	0.08	1.00	1.00
	1e-12	32.55	30.27	1.95	1.98	0.11	0.08	1.00	1.00
5	1e-02	169.36	169.82	1.00	1.00	3.26	3.36	1.00	1.00
	1e-04	34.82	33.73	1.00	1.00	1.08	0.79	1.00	1.00
	1e-06	34.00	31.73	2.13	2.00	1.17	0.75	1.00	1.00
	1e-08	33.27	31.73	2.60	2.10	1.17	0.73	1.00	1.00
	1e-10	33.45	31.73	1.93	2.10	1.13	0.73	1.00	1.00
	1e-12	33.45	31.73	1.90	2.10	1.15	0.73	1.00	1.00

Table 4: Results for Modified Newton method applied to Extended Rosenbrock with relative finite differences, metrics are average metrics for succesful attempts.

		iterations		convergence rate		time		success rate	
	preconditioning	False	True	False	True	False	True	False	True
dimension	h								
3	1e-02	139.82	141.45	1.00	1.00	0.06	0.06	1.00	1.00
	1e-04	33.73	32.27	1.00	1.00	0.02	0.02	1.00	1.00
	1e-06	32.09	30.18	1.89	2.06	0.02	0.02	1.00	1.00
	1e-08	32.00	30.18	2.04	2.23	0.02	0.02	1.00	1.00
	1e-10	32.00	30.18	2.00	2.23	0.02	0.02	1.00	1.00
	1e-12	32.00	30.18	2.00	2.23	0.02	0.02	1.00	1.00
4	1e-02	149.55	150.55	1.00	1.00	0.32	0.32	1.00	1.00
	1e-04	34.45	32.45	1.00	1.00	0.12	0.09	1.00	1.00
	1e-06	32.55	30.27	1.81	1.96	0.12	0.09	1.00	1.00
	1e-08	32.55	30.27	1.95	1.98	0.12	0.08	1.00	1.00
	1e-10	32.55	30.27	1.95	1.98	0.12	0.08	1.00	1.00
	1e-12	32.55	30.27	1.95	1.98	0.11	0.08	1.00	1.00
5	1e-02	159.18	160.27	1.00	1.00	3.05	3.17	1.00	1.00
	1e-04	34.64	34.45	1.00	1.00	1.09	0.80	1.00	1.00
	1e-06	33.45	31.73	2.56	2.00	1.16	0.74	1.00	1.00
	1e-08	33.09	31.73	1.84	2.10	1.13	0.73	1.00	1.00
	1e-10	33.45	31.73	1.90	2.10	1.12	0.74	1.00	1.00
	1e-12	33.45	31.73	1.90	2.10	1.14	0.77	1.00	1.00

Table 5: Results for Truncated Newton method applied to Extended Rosenbrock with absolute finite differences, metrics are average metrics for succesful attempts.

		iterations		convergence rate		time		success rate	
	preconditioning	False	True	False	True	False	True	False	True
dimension	h								
3	1e-02	182.00	179.27	1.00	1.00	0.02	0.04	1.00	1.00
	1e-04	54.36	37.91	1.00	1.00	0.01	0.01	1.00	1.00
	1e-06	52.82	36.36	-1.48	2.39	0.01	0.01	1.00	1.00
	1e-08	52.91	35.91	-2.12	2.86	0.01	0.01	1.00	1.00
	1e-10	53.36	35.45	-4.10	2.21	0.01	0.01	1.00	1.00
	1e-12	55.64	36.45	-2.49	3.01	0.01	0.01	1.00	1.00
4	1e-02	203.91	192.18	1.00	1.00	0.16	0.17	1.00	1.00
	1e-04	61.64	44.45	1.00	1.00	0.06	0.05	1.00	1.00
	1e-06	59.18	42.55	-1.35	4.36	0.06	0.05	1.00	1.00
	1e-08	61.64	42.09	-1.78	2.34	0.06	0.05	1.00	1.00
	1e-10	62.27	41.91	-2.46	3.14	0.06	0.05	1.00	1.00
	1e-12	61.73	43.00	-2.89	2.65	0.06	0.05	1.00	1.00
5	1e-02	224.55	206.82	1.00	1.00	1.84	1.83	1.00	1.00
	1e-04	74.64	42.91	1.00	1.01	0.73	0.41	1.00	1.00
	1e-06	79.09	49.27	-2.17	2.06	0.81	0.50	1.00	1.00
	1e-08	75.64	49.91	-2.83	7.31	0.75	0.47	1.00	1.00
	1e-10	78.00	51.64	-1.21	3.21	0.75	0.50	1.00	1.00
	1e-12	78.00	53.27	-2.11	2.50	0.77	0.52	1.00	1.00

Table 6: Results for Truncated Newton method applied to Extended Rosenbrock with relative finite differences, metrics are average metrics for succesful attempts.

		iterations		convergence rate		time		success rate	
	preconditioning	False	True	False	True	False	True	False	True
dimension	h								
3	1e-02	169.91	169.91	1.00	1.00	0.02	0.03	1.00	1.00
	1e-04	54.45	37.27	1.00	1.02	0.01	0.01	1.00	1.00
	1e-06	51.45	36.27	-2.69	2.35	0.01	0.01	1.00	1.00
	1e-08	53.64	36.27	-3.43	1.91	0.01	0.01	1.00	1.00
	1e-10	54.64	35.27	-2.94	2.01	0.01	0.01	1.00	1.00
	1e-12	53.82	35.91	-4.18	2.56	0.01	0.01	1.00	1.00
4	1e-02	191.09	182.64	1.00	1.00	0.15	0.17	1.00	1.00
	1e-04	64.27	42.55	1.00	1.01	0.06	0.05	1.00	1.00
	1e-06	62.18	43.00	-0.96	2.62	0.06	0.05	1.00	1.00
	1e-08	60.73	40.91	-2.95	2.87	0.06	0.05	1.00	1.00
	1e-10	60.82	42.73	-1.22	3.98	0.06	0.05	1.00	1.00
	1e-12	63.09	42.91	-2.01	2.51	0.06	0.05	1.00	1.00
5	1e-02	213.55	198.64	1.00	1.00	1.68	1.74	1.00	1.00
	1e-04	75.91	40.82	1.00	1.00	0.74	0.42	1.00	1.00
	1e-06	77.91	49.00	-2.96	6.91	0.77	0.49	1.00	1.00
	1e-08	77.64	50.18	-1.96	2.28	0.77	0.48	1.00	1.00
	1e-10	75.00	53.27	-4.07	3.38	0.75	0.52	1.00	1.00
	1e-12	75.09	52.45	-2.09	3.48	0.75	0.52	1.00	1.00

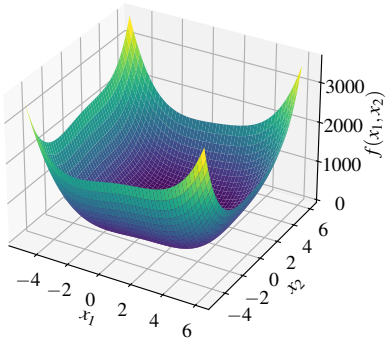

Figure 2: Surface plot of the 2-dimensional generalized Broyden tridiagonal function

computation can be eased considering that component k depends only on f_k, f_{k+1} and f_{k-1}. The Hessian of the generalized Broyden tridiagonal function is given by the following expression.

$$\frac{\partial^2 F}{\partial x_k \partial x_j} = \begin{cases} (3 - 4x_1)^2 - 4f_1(x) + 1, & k = j = 1 \\ (3 - 4x_k)^2 - 4f_k(x) + 2, & 1 < k = j < n \\ (3 - 4x_n)^2 - 4f_n(x) + 1, & k = j = n \\ 4x_k + 4x_{k+1} - 6, & |k - j| = 1 \\ 1, & |k - j| = 2 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Notice that the Hessian is a banded matrix, with only n non-zero elements on the diagonal, $n - 1$ non-zero elements on the first co-diagonal and $n - 2$ non-zero elements on the second codiagonal.

4.2 Finite differences gradient and Hessian

When applying 1, one can notice that the terms $F(x + he_k)$ and $F(x - he_k)$ only differ by terms f_k, f_{k+1} and f_{k-1}. Then to make function evaluations less expensive, we can define the following function $F_{fd,\,k}$, which can be plugged in 1 in place of F yielding the same result.

$$F_{fd,\,k}(x) = \frac{1}{2} f_k^2(x) + \frac{1}{2} f_{k+1}^2(x) + \frac{1}{2} f_{k-1}^2(x)$$

The same procedure can be applied for the Hessian, considering that:

- function evaluations to compute entry $h_{k,k}$ differ only by f_k, f_{k+1} and f_{k-1};
- function evaluations to compute entry $h_{k,k+1}$ differ only by f_k and f_{k+1};
- function evaluations to compute entry $h_{k,k+2}$ differ only by f_{k-1}.

Then to make function evaluations less expensive, we can define the functions $F_{fd,\,k,k}$, $F_{fd,\,k,k+1}$, $F_{fd,\,k,k+2}$, which can be plugged in 2 in place of F yielding the same result to compute entries $h_{k,k}$, $h_{k,k+1}$ and $h_{k,k+2}$ respectively.

$$\begin{aligned} F_{fd,\,k,k}(x) &= \frac{1}{2} f_k^2(x) + \frac{1}{2} f_{k-1}^2(x) + \frac{1}{2} f_{k+1}^2(x) \\ F_{fd,\,k,k+1}(x) &= \frac{1}{2} f_k^2(x) + \frac{1}{2} f_{k+1}^2(x) \\ F_{fd,\,k,k+2}(x) &= \frac{1}{2} f_{k-1}^2(x) \end{aligned}$$

When plugging the functions $F_{fd,\,k}$, $F_{fd,\,k,k}$, $F_{fd,\,k,k+1}$ and $F_{fd,\,k,k+2}$ into 1 and 2 it's convenient to expand them so that the computation of the gradient and Hessian is not subject to numerical cancellation as previously done for the extended Rosenbrock function in subsection 3.2.

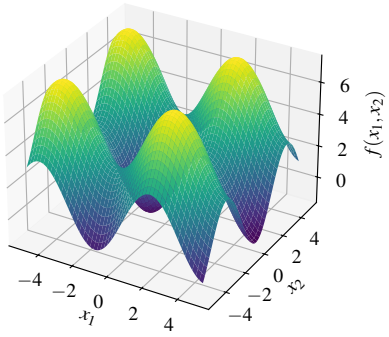

Figure 3: Surface plot of the 2-dimensional banded trigonometric function

5 Banded trigonometric function

$$f(x) = \sum_{i=1}^{n} i[(1 - \cos x_i) + \sin x_{i-1} - \sin x_{i+1}]$$

5.1 Exact gradient and Hessian

5.2 Finite differences gradient and Hessian

6 Conclusions

References

[1] Ladislav Luksan and Jan Vlček. *Test Problems for Unconstrained Optimization*. Nov. 2003.

[2] Jorge Nocedal and Stephen J. Wright. "Numerical optimization". English (US). In: *Springer Series in Operations Research and Financial Engineering*. Springer Series in Operations Research and Financial Engineering. Springer Nature, 2006, pp. 1–664.