4.1. Numerical Aspects

As evident in figure (4.1) as well as figure (4.2), the optimization is stopped after 214 iterations. The reason for this is neither that a sufficiently large gradient is no longer available nor that the set R is empty. So all conditions of the main loop stated in step 3 of algorithm 4 are still fulfilled. Rather, a stable state has occurred which does not allow any further improvement of the results. Reaching a stable state during an optimization is one of several reasons for an optimization to stop.

Facing a stable state involves a very special case of index shifts between the sets R and P, which must be handled separately by the algorithm. Starting from a gradient calculated in step i), the new index j in R is searched for which has the maximum gradient. This new index is then moved from the set R to the set P and the constrained least squares problem is calculated as specified in step e). Since the vector s^P now has exactly one negative entry, the inner loop must be entered in the next step. It turns out that exactly that index j of s is negative which was added by the shift from j to P in the last main loop. Since the corresponding entry x_j is zero and s has only one negative entry, the scaling factor α also has a value of zero. This leads to the fact that the shift specified in step ii. has no effect and the solution vector x remains the same. In step iii. the previously determined index j will then be shifted back from the set P to the set R. In the next iteration of the main loop, the index with the largest gradient is found again as the previously moved index j. This will lead to what has just been described, resulting in an infinite loop. A case such as this therefore only occurs if the following two conditions occur simultaneously:

1. The solution vector s^P of the constrained least squares problem has one negative entry.
2. This negative entry is in vector s exactly at the index that was transferred from set R to set P in the current loop pass in step b).

Remark 4.6. *If the stopping reason of the optimization algorithm is the reaching of a stable status, the number of non-negative entries of the solution vector x must be equal for the last two iterations. This means that in figure (4.1) the graph has to end with a horizontal line.*

Another reason that can lead to an unplanned termination of the optimization algorithm is related to solving process of the restricted least squares problem in step e) and iv). In these steps the solution of the constrained least squares problem is calculated which requires the use of an inverse matrix. If the inverse of $((A^P)^\top A^P)^{-1}$ does not exist, an alternative solution has to be found. The most practicable approach is to test whether an inverse exists using the determinant of $(A^P)^\top A^P$. If this is not the case, remove the just added index j from the set P and replace it with a j' which has the second largest gradient. Not only the existence of the inverse matrix is important but also its numerical stability. Therefore it should be ensured that with a small change of the matrix A, the inverse matrix A^{-1} do not change significantly in order to get stable results. To determine whether the inversion of a matrix A is numerically stable, a new concept must be introduced which characterizes the numerical stability.

Definition 4.2. *Let V and W be normed vector spaces and $f : V \to W$ a linear operator. Then the operator norm is given by:*

$$\|f\| := \sup_{x \in V \setminus \{0\}} \frac{\|f(x)\|_W}{\|x\|_V} = \sup_{\|x\|_V = 1} \|f(x)\|_V \tag{4.3}$$

Remark 4.7. *Since every real valued matrix $A \in \mathbb{R}^{m \times n}$ corresponds to a linear map from $\mathbb{R}^n$ to R^m, each pair of norms induces an operator norm. In the special case of choosing the Euclidean norm for both vector spaces, (4.3) simplifies to:*

$$\|A\|_2 := \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_{2=1}} \|Ax\|_2 \tag{4.4}$$

which is a naturally induced matrix norm called spectral norm.

The natural matrix norm thus vividly corresponds to the greatest possible stretching factor, which results from the application of the linear mapping (i.e. matrix) to a unit vector.

Remark 4.8. *The naturally induced matrix norm satisfies the three norm axioms:*

- $\|A\| = 0$ *iff* $A = 0$ *(being definite)*
- $\|\alpha A\| = |\alpha| \|A\|$ *(being absolutely homogeneous)*
- $\|A + B\| \leq \|A\| + \|B\|$ *(being sub-additive)*

Remark 4.9 ([**stewart1998matrix**]). *The 2-norm has the following properties:*

1. $\|A\|_2 = \sigma_{max}(A)$ *largest singular value of A*
2. $\|A\|_2^2 = \lambda_{max}(A^TA)$ *largest eigenvalue value of A^TA*
3. $\|A\|_2 = \|A^T\|_2$

Definition 4.3. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix and $\|.\|$ a matrix norm. Then the condition number of A is defined as:*

$$\kappa(A) := \|A^+\|\|A\| \tag{4.5}$$

Remark 4.10. *For the special case that matrix A is regular, the pseudo inverse can be replaced with the inverse matrix and the condition number becomes:*

$$\kappa(A) = \|A^{-1}\|\|A\| \tag{4.6}$$

Remark 4.11. *Let A be a nonsingular matrix then a lower bound for the condition number is given by:*

$$1 \leq \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A)$$

Remark 4.12. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix and $\|.\|_2$ the induced matrix norm. Then (4.5) simplifies to:*

$$\kappa_2(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$$

As already mentioned above, it is necessary to estimate to what extent changes in the input variables have an effect on the output variables. In the case of the calculation of an inverse matrix the following approach can be followed. Starting from a matrix A which should be inverted another matrix E is defined. This matrix E represents small changes of matrix A and should therefore be seen as perturbation of matrix A. In order to determine how numerically stable the inversion of a matrix A is, the following term is considered.

$$\|A^{-1} - (A+E)^{-1}\| \tag{4.7}$$

The task now is to find a constant c, so that for all sufficiently small matrices E it applies that:

$$\|A^{-1} - (A+E)^{-1}\| \leq c\|E\| \tag{4.8}$$

Remark 4.13. *It can be shown that:*

$$(A+E)^{-1} = (I + A^{-1}E)^{-1}A^{-1}$$

Corollary 4.1 ([**AnaII**]). *Suppose that B is a bounded linear operator on a Banach space X with $\|B\| < 1$. Then*

$$S = \sum_{k=0}^{\infty} B^k = (I - B)^{-1} \tag{4.9}$$

It therefore follows by applying remark 4.13 and corolarry 4.1 that:

$$\begin{aligned}(A+E)^{-1} &= (I + A^{-1}E)^{-1}A^{-1} \\ &= (I - (A^{-1}E) + (A^{-1}E)^2 - (A^{-1}E)^3 + H.O.T.)A^{-1} \\ &= A^{-1} + A^{-1}EA^{-1} + H.O.T.\end{aligned}$$

For the estimation of the stability the following inequality can be obtained by using the previous results as well as (4.7):

$$\begin{aligned}\|A^{-1} - (A+E)^{-1}\| &= \|A^{-1} - (A^{-1} - A^{-1}EA^{-1} + H.O.T.)\| \\ &= \|A^{-1}EA^{-1} - H.O.T.\| \\ &\leq \|A^{-1}\|^2\|E\| + H.O.T.\end{aligned}$$

The relative error caused by applying a perturbation to matrix A has therefore an upper bound which is given by:

$$\frac{\|A^{-1} - (A+E)^{-1}\|}{\|A^{-1}\|} \leq \|A\|\|A^{-1}\|\frac{\|E\|}{\|A\|} + H.O.T. \tag{4.10}$$

Remark 4.14. *If the naturally induced spectral norm is used as the matrix norm, (4.10) simplifies to:*

$$\begin{aligned}\frac{\|A^{-1} - (A+E)^{-1}\|}{\|A^{-1}\|} &\leq \kappa_2(A)\frac{\|E\|}{\|A\|} + H.O.T. \\ &= \frac{\sigma_{max}}{\sigma_{min}}\frac{\|E\|}{\|A\|} + H.O.T.\end{aligned}$$

To reduce the influence of numerical instabilities, in general a system of equations $Ax = b$ is not solved by using the inverse A^{-1} for $x = A^{-1}b$ but by other methods. One of the most frequently used methods is that of QR decomposition.

4.2. QR - decomposition