

# Chapter 6

## Nonlinear least squares

In this short chapter we consider the very popular special case of unconstrained minimization where the objective function can be written as the least squares norm of a vector of residuals. We have seen such instances in Examples 2.5 and 4.5.

Let us recall the notation used for nonlinear systems of equations: we would have liked to solve

$$\mathbf{r}(\mathbf{x}) = \mathbf{0}.$$

But now in general we cannot, because

$$\mathbf{r} : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

and we may have  $m > n$  (i.e. more equations than unknowns). So, we consider the general problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^m r_i^2(\mathbf{x}) = \frac{1}{2} \|\mathbf{r}\|^2. \quad (6.1)$$

This is an unconstrained optimization problem of a special form.

In Section 5.3 we saw (in  $f_2$  of (5.36)) a special case of (6.1) where  $\mathbf{r}(\mathbf{x})$  is linear. Now we consider the nonlinear case.

Recall the notation for the Jacobian matrix

$$J = \frac{\partial \mathbf{r}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \frac{\partial r_1}{\partial x_2} & \cdots & \frac{\partial r_1}{\partial x_n} \\ \frac{\partial r_2}{\partial x_1} & \frac{\partial r_2}{\partial x_2} & \cdots & \frac{\partial r_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial r_m}{\partial x_1} & \frac{\partial r_m}{\partial x_2} & \cdots & \frac{\partial r_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}. \quad (6.2)$$

Then,

$$\nabla f(\mathbf{x}) = \sum_{i=1}^m r_i(\mathbf{x}) \nabla r_i(\mathbf{x}) = J^T \mathbf{r}, \quad (6.3a)$$

$$\nabla^2 f(\mathbf{x}) = J^T J + \sum_{i=1}^m r_i(\mathbf{x}) \nabla^2 r_i(\mathbf{x}). \quad (6.3b)$$

## 6.1 Gauss-Newton method and modifications

The material in this section and more is covered in Chapter 10 of [23].

Let us assume that there are more rows than columns in  $J$ ,  $m \geq n$ , and that  $J$  has full column rank. Then the matrix  $J^T J$  appearing in (6.3b) is nonsingular.

Newton's direction,  $\mathbf{p}_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k$ , has two potential disadvantages:

1. The 2nd order information term  $\sum_{i=1}^m r_i(\mathbf{x}) \nabla^2 r_i(\mathbf{x})$  may be much more difficult to obtain and evaluate than  $J$ .
2. The Hessian matrix  $\nabla^2 f_k$  may not always be positive definite, even though  $\nabla^2 f(\mathbf{x}^*)$  is.

The *Gauss-Newton method* (GN) addresses both difficulties by setting

$$B_k = J_k^T J_k \quad (6.4a)$$

and calculating

$$\mathbf{p}_k = -B_k^{-1} \nabla f_k = -[J_k^T J_k]^{-1} J_k^T \mathbf{r}_k. \quad (6.4b)$$

In fact, the Gauss-Newton direction  $\mathbf{p}_k$  is the solution using the normal equations of the *linear least squares* problem

$$\min_{\mathbf{p}} \|J_k \mathbf{p} + \mathbf{r}_k\|. \quad (6.5)$$

So, the Gauss-Newton method corresponds to linearizing  $\mathbf{r}(\mathbf{x})$  inside the norm in the expression for the objective function  $f$ .<sup>1</sup>

---

<sup>1</sup>This simple and obvious observation accounts to a fact that seems to have caused an inordinate amount of confusion among mathematicians and practitioners alike, because the full Newton direction also arises from linearization – of  $f$  itself!

Note that there are better ways from the point of view of numerical stability to solve linear least squares problems than forming and solving the normal equations (6.4b). For instance, use QR decomposition or SVD. In particular, using SVD it is possible to solve the linear problem (6.5) in a meaningful way also when  $J$  does not have a full column rank (so  $B_k$  is singular).

The GN method is extremely popular in practice. It converges under standard assumptions. The speed of convergence depends, though, on how small the neglected term of the Hessian is. Thus, we have the peculiar situation that for data fitting problems with little noise in the data (corresponding to  $\|\mathbf{r}\|$  being small) the method converges faster than for data fitting problems with a lot of noise in the data. Even more interesting is the case of parameter or model identification, where  $\mathbf{r}$  represents an attempt at modeling whatever has generated the data. The better the model then, the smaller  $\mathbf{r}$  and so the closer Gauss-Newton is to Newton. Some people have consequently argued that if the GN method does not converge quickly then the model is not good and should be discarded!

There are, of course, special variants of the CG algorithm to solve large scale linear least squares problems of the form (6.5).

## Levenberg-Marquardt method

If  $J_k$  is (column) rank-deficient, or nearly so, then  $J_k^T J_k$  is positive semi-definite and singular. Then the Levenberg-Marquardt method sets

$$B_k = J_k^T J_k + \mu I \quad (6.6)$$

for some appropriate  $\mu > 0$ , making  $B_k$  positive definite and the resulting search direction well-defined. We have seen this idea before, of course. A trust region approach may be used to select  $\mu$ .

Historically, the Levenberg-Marquardt method was a precursor for trust region methods.