Distance Regression by Gauss-Newton-type Methods and Iteratively Re-weighted Least-Squares

M. Aigner, B. Jüttler
Institute of Applied Geometry, Johannes Kepler University, Linz, Austria

October 30, 2008

Abstract

We discuss the problem of fitting a curve or surface to given measurement data. In many situations, the usual least–squares approach (minimization of the sum of squared norms of residual vectors) is not suitable, as it implicitly assumes a Gaussian distribution of the measurement errors. In those cases, it is more appropriate to minimize other functions (which we will call norm-like functions) of the residual vectors. This is well understood in the case of scalar residuals, where the technique of iteratively re-weighted least-squares (IRLS), which originated in statistics [11], is known to be a Gauss-Newton-type method for minimizing a sum of norm-like functions of the residuals. We extend this result to the case of vector-valued residuals. It is shown that simply treating the norms of the vector-valued residuals as scalar ones does not work. In order to illustrate the difference we provide a geometric interpretation of the iterative minimization procedures as evolution processes.

1 Introduction

The problem of fitting a given mathematical model to a certain set of measurement data appears in the context of various applications. For instance, during the process of reconstructing geometric models from (possibly unstructured) point cloud data, one is interested in fitting a parametric curve or surface (often represented by NURBS: non–uniform rational B-splines) to a given point cloud, see [10, 18]. In statistics, regression lines or other geometric objects need to be identified in order to analyze the relations between two or more variables [11]. In biomedicine, image segmentation or the automatic recognition of biological structures is an important issue [15].

Many existing techniques rely on least-squares approximation, i.e., on minimizing the sum of the squared residuals (the deviations between the model and the measured data). This is equivalent to minimizing the ℓ_2 norm of the vector of (scalar–valued) residuals. In the case of curves and surfaces, the residuals can be chosen as the distances between the

data and the associated closest points. This leads to the special case of orthogonal distance regression [2, 4, 5, 6, 14, 18, 19, 20, 21, 22].

If the residuals depend linearly on the parameters controlling the model, then the solution can be found by solving a system of linear equations. In the case of a non-linear dependence between residuals and parameters, Gauss-Newton-type methods are widely used, as they avoid the evaluation of second derivatives. See [26] for a historical overview of Newton and Newton-like methods.

In many situations, considering the ℓ_2 norm of the residuals is not the appropriate approach. On the one hand, so–called outliers (data with large error) may decrease the quality of the approximation, since their influence grows quadratically with the distance to the curve or surface. On the other hand, if the data are very precise, then it is more appropriate to minimize the maximum deviation between the model and the data. It is clearly important to adjust the norm carefully to the problem, see [3, 12, 23, 24].

In the field of statistics [11], the usual ℓ_2 approximation is extended with the help of the concept of iteratively-reweighted least-squares (IRLS). In [16] and [25] the important connection of IRLS and Gauss-Newton for maximum likelihood estimation is analyzed. These articles are concerned with fitting regression models to given observations. As the underlying models are functions, the considered techniques are applied to scalar residuals.

In [8] the IRLS approach was used to implement the fitting of linear and non-linear models in different norms. Again, the residuals where scalar values.

A more general approach for fitting graphs of functions is proposed in [7]. The error is assumed to be both in the dependent variable and in the independent variable. This modification allows to perform orthogonal distance regression, i.e., to minimize the shortest distance from a data point to the function.

The need for dealing with vector-valued residuals has arisen in connection with curve and surface fitting. The existing literature which studies the relation between IRLS and Gauss-Newton-type techniques discusses only the case of scalar residuals. The present paper aims to extend these results to the vector–valued case.

The remainder of the paper is organized as follows. The next section presents some preliminaries and introduces the generalized fitting problem. Section 3 focuses on the Gauss-Newton method. We analyze it for the scalar and vector-valued situation in terms of an evolution process and show that simply defining scalar residuals as norms of vector-valued ones does not give good results in Section 4. Section 5 discusses the convergence behavior of the Gauss-Newton method for vector-valued residuals. Section 6 presents several examples. Finally, we conclude this paper.

2 The generalized fitting problem

Given a set of data points $\{\mathbf{P}_j\}_{j=1..N}$ in $\mathbb{R}^d$, we are interested in a surface of dimension $k < d$ which approximates them. In order to keep the notation simple, in this paper we restrict ourselves to the case of curves ($k = 1$) or surfaces (k=2). The results which are presented in this paper can easily be extended to other values of k.

We denote a parametric curve or surface by $\mathbf{c}_\mathbf{s}(t)$, with $t \in I = [a, b] \subset \mathbb{R}$ for curves or $t = (t^{(1)}, t^{(2)}) \in [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$ for surfaces. Similarly to the framework described in [1, 2], we assume that a curve/surface can be described by a vector of shape parameters $\mathbf{s} \in \mathbb{R}^n$, where n is the number of degrees of freedom.

Example 1. An ellipse in the plane can be parameterized using trigonometric functions as

$$\mathbf{c}_\mathbf{s}(t) = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} x_0 + a\sin(t) \\ y_0 + b\cos(t) \end{pmatrix}, \quad t \in I = [0, 2\pi). \tag{1}$$

The shape parameters are the coordinates of the center (x_0, y_0) along with the semi-major a and the semi-minor axes b and the rotation angle ϕ of the ellipse.

Example 2. Consider a spline surface

$$\mathbf{c}_\mathbf{s}(t) = \sum_{i=0}^{l} \sum_{j=0}^{m} \mathbf{d}_{ij} \psi_i(t^{(1)}) \psi_j(t^{(2)}) \tag{2}$$

where $\psi_i(t_1)$ are the B-splines of a certain degree k, defined over a suitable fixed knot vector, see [10]. The vector of shape parameters is obtained as the concatenation of all control points $\mathbf{d}_{ij} = (s_{3(m+1)i+3j+1}, s_{3(m+1)i+3j+2}, s_{3(m+1)i+3j+3})$, $n = 3(l+1)(m+1)$. In addition, it may also contain the knots.

When fitting a curve/surface to a given set of points one minimizes usually certain distances from the data to the curve/surface. In this work we consider distance regression where the parameters of the points on the curve/surface are fixed a-priori with some suitable parameterization method, such as chordal, centripetal or uniform parameterization, see e.g. [10]. We shall denote the parameter values with t_j. The *residual vector* which connects a data point with an associated point on the hyper surface is given by $\mathbf{R}_j(\mathbf{s}) = \mathbf{P}_j - \mathbf{c}_\mathbf{s}(t_j)$. Its Euclidean norm gives the *scalar residual* $r_j(\mathbf{s}) = ||\mathbf{R}_j(\mathbf{s})||$.

Now, consider the *generalized fitting problem*

$$\mathbf{s}^0 = \arg\min_\mathbf{s} F(\mathbf{s}) = \arg\min_\mathbf{s} \sum_{j=1}^{M} N(\|\mathbf{R}_j(\mathbf{s})\|) = \arg\min_\mathbf{s} \sum_{j=1}^{M} N(r_j(\mathbf{s})). \tag{3}$$

It generalizes the usual fitting problems by replacing the ℓ_p norms (mostly with $p = 2$) of the vector of scalar residuals $||\mathbf{R}_j(\mathbf{s})||$ with a sum of a certain functions $N(.)$ applied to the residuals. We choose these functions from the class of norm-like functions in the following sense, cf. [17].

Definition 1. *A $\mathcal{C}^2$ function $N(x) : \mathbb{R}^+ \to \mathbb{R}^+$ is said to be* ***norm-like*** *if there exists a positive constant ε such that the derivative satisfies*

$$N'(x) = x\, w(x) \quad \textit{for} \quad x \in (0, \varepsilon] \tag{4}$$

where the associated ***weight function*** *$w(x)$ is positive. If the weight function $w(x)$ is a restriction of a function $w_0 \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ to the interval $(0, \varepsilon]$ with $w_0(x) : [0, \varepsilon] \to [c, C]$ and $c, C \in \mathbb{R}^+$, then we will call it* ***positive*** *and* ***bounded****.*

Example 3.
1. The norm–like function $N(x) = x^2$, which corresponds to the ℓ_2 norm of the vector of residuals, has the weight function $w(x) = 2$ which is positive and bounded.

2. The norm–like function $N(x) = \exp(x^2) - 1$ has the weight function $w(x) = 2 \exp x^2$. Again, the weight function is positive and bounded.

3. The norm–like functions $N(x) = x^p$ for $p \neq 2$, which correspond to the ℓ_p norm of the vector of residuals, have the weight functions $w(x) = px^{p-2}$. The weight functions are positive, but not bounded, for $1 < p < 2$, and bounded, but not positive, for $p > 2$.

For later reference we provide the gradient and the Hessian of the objective function $F(\mathbf{s})$, see (3). The gradient is the row vector of the first partial derivatives with respect to the shape parameters s_i,

$$\nabla F = \sum_{j=1}^{M} N'(\|\mathbf{R}_j\|) \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j = \sum_{j=1}^{M} w(\|\mathbf{R}_j\|) \mathbf{R}_j^\top \nabla \mathbf{R}_j \tag{5}$$

where we shall omitt the argument of the residual from now on. The Hessian is the $n \times n$–matrix

$$H_F = \nabla(\nabla F^\top) = \sum_{j=1}^{M} \frac{w_j'}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j^\top \mathbf{R}_j \mathbf{R}_j^\top \nabla \mathbf{R}_j + w_j \nabla \mathbf{R}_j^\top \nabla \mathbf{R}_j + w_j \nabla(\nabla \mathbf{R}_j^\top) \circ \mathbf{R}_j, \tag{6}$$

where we use the abbreviations $w_j = w(\|\mathbf{R}_j\|)$ and $w_j' = w'(\|\mathbf{R}_j\|)$. The second order derivative $\nabla(\nabla \mathbf{R}_j^\top)$ is a tensor, and is to be interpreted in the following way:

$$\left[\nabla(\nabla \mathbf{R}_j^\top) \circ \mathbf{R}_j\right]_{l,k} = \sum_{i=1}^{d} \left[\frac{\partial}{\partial s_l} \frac{\partial}{\partial s_k} [\mathbf{R}_j]_i \right] [\mathbf{R}_j]_i. \tag{7}$$

Here, $[\mathbf{V}]_i$ denotes the $i-$th entry of the vector $\mathbf{V}$ and $[\mathbf{M}]_{i,j}$ denotes the $j-$th entry of the $i-$th row of the matrix $\mathbf{M}$.

Alternatively, one may also express the gradient and the Hessian using the scalar residuals $r_j = \|\mathbf{R}_j\|$,

$$\nabla F = \nabla \sum_{j=1}^{M} N(r_j) = \sum_{j=1}^{M} N'(r_j) \nabla r_j = \sum_{j=1}^{M} w(r_j) r_j \nabla r_j \tag{8}$$

and

$$H_F = \nabla(\nabla F^\top) = \sum_{j=1}^{M} w_j' r_j \nabla r_j^\top \nabla r_j + w_j \nabla r_j^\top \nabla r_j + w_j \nabla(\nabla r_j^\top) r_j. \tag{9}$$

3 Iterative solution of the generalized fitting problem

Equation (3) defines a non-linear optimization problem. Using Newton's method one obtains the linearized system

$$\nabla F^\top + H_F \Delta \mathbf{s} = 0 \tag{10}$$

for the update step $\Delta \mathbf{s}$. Under certain assumptions, Newton's method exhibits quadratic convergence, see [26] and the references cited therein.

In some situations, i.e. if the exact Hessian is unknown or very costly to evaluate, one may prefer an approximation instead of the exact Hessian. This leads to Gauss-Newton-type methods. In the following we consider the approximate Hessians

$$H_F^* = \sum_{j=1}^{M} w_j \nabla \mathbf{R}_j^\top \nabla \mathbf{R}_j \tag{11}$$

and

$$H_F^\dagger = \sum_{j=1}^{M} w_j \nabla r_j^\top \nabla r_j \tag{12}$$

which are obtained by omitting the first and the last part in the expansions (6) and (9), respectively.

In the vector-valued situation and for vanishing residuals we get the following result.

Lemma 2. *Let* $\mathbf{s}^0$ *be the minimizer of* (3) *such that* $\|\mathbf{R}_j(\mathbf{s}^0)\| = 0$ *for all* j *and let* $N(x)$ *have a positive and bounded weight function. Then*

$$\lim_{\mathbf{s}\to\mathbf{s}^0} H_F = \lim_{\mathbf{s}\to\mathbf{s}^0} H_F^*. \tag{13}$$

Proof. We have to show that all but the second term of the expansion of the Hessian (6) vanish. For the last term this can be seen immediately. The first part of the Hessian is the sum of

$$T_j = w_j' \|\mathbf{R}_j\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j, \quad j = 1, \ldots, M. \tag{14}$$

We consider the spectral norm of T_j,

$$\|T_j\|_2^2 \leq \left\| w_j' \|\mathbf{R}_j\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2 \leq |w_j'|^2 \, \|\mathbf{R}_j\|^2 \, \left\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2$$

Exploiting the Cauchy-Schwarz inequality we obtain

$$\|T_j\|_2^2 \leq |w_j'|^2 \, \|\mathbf{R}_j\|^2 \, \left\| \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \right\|_2^2 \left\| \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j \right\|_2^2.$$

Finally we get

$$\|T_j\|_2^2 \leq |w_j'|^2 \, \|\mathbf{R}_j\|^2 \, \left\|\nabla \mathbf{R}_j^\top\right\|_2^2 \underbrace{\left\|\frac{\mathbf{R}_j}{\|\mathbf{R}_j\|}\right\|_2^2}_{=1} \underbrace{\left\|\frac{\mathbf{R}_j}{\|\mathbf{R}_j\|}\right\|_2^2}_{=1} \left\|\nabla \mathbf{R}_j^\top\right\|_2^2, \tag{15}$$

as the spectral norm for matrices and the ℓ_2 norm for vectors are consistent. Now, if $\mathbf{s} \to \mathbf{s}^0$, then $\|\mathbf{R}_j\| \to 0$, hence $\|T_j\|_2 \to 0$. □

Remark 1. This Lemma excludes the case of ℓ_p norms for $p > 2$ since they do not fulfill the assumption that $w(0) \neq 0$ in a zero-residual case. Nevertheless, Eq. (13) remains true, since both H_F and H_F^* converge to the null matrix. For the case $1 \leq p < 2$, however, the optimization problem is not differentiable.

The equivalent result for scalar residuals (e.g. regression analysis) is well-known, see [16]. However, if the scalar approach is applied directly to vector-valued residuals by choosing simply $r_j = \|\mathbf{R}_j\|$, this result is no longer true:

Lemma 3. *Let $\mathbf{c}_\mathbf{s}(t)$ be a B-Spline curve in $\mathbb{R}^2$ (i.e. $k = 1$ and $d = 2$) as in (2), where the shape parameters $\mathbf{s}$ are the components of the control points, and let $\mathbf{s}^0$ be the minimizer of (3) such that $\|\mathbf{R}_j(\mathbf{s}^0)\| = 0$ for all j. In addition, let $N(x)$ have a positive and bounded weight function. Then*

$$\lim_{\mathbf{s}\to\mathbf{s}^0} H_F \neq \lim_{\mathbf{s}\to\mathbf{s}^0} H_F^\dagger. \tag{16}$$

Proof. For any matrix $A = (a_{ij})$, let $\|(a_{ij})\|_p = \sqrt[p]{\sum_{i,j} |a_{ij}|^p}$. We show that there exists a positive constant D such that

$$\lim_{\mathbf{s}\to\mathbf{s}^0} \|H_F^* - H_F^\dagger\|_1 = \lim_{\mathbf{s}\to\mathbf{s}^0} \|\sum_{j=1}^{M} w_j(\nabla \mathbf{R}_j^\top \nabla \mathbf{R}_j - \nabla r_j^\top \nabla r_j)\|_1 \geq D > 0. \tag{17}$$

Due to the previous lemma, this statement is equivalent to (16).

With the identity matrix $\mathbf{E} = \mathrm{diag}(1, 1)$, we get

$$\sum_{j=1}^{M} w_j(\nabla \mathbf{R}_j^\top \nabla \mathbf{R}_j - \nabla r_j^\top \nabla r_j) = \sum_{j=1}^{M} w_j(\nabla \mathbf{R}_j^\top \mathbf{E} \nabla \mathbf{R}_j - \nabla \mathbf{R}_j^\top \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|} \nabla \mathbf{R}_j) =$$
$$= \sum_{j=1}^{M} w_j \nabla \mathbf{R}_j^\top (\mathbf{E} - \frac{\mathbf{R}_j}{\|\mathbf{R}_j\|} \frac{\mathbf{R}_j^\top}{\|\mathbf{R}_j\|}) \nabla \mathbf{R}_j = \sum_{j=1}^{M} w_j \nabla \mathbf{R}_j^\top \mathbf{S}_j \mathbf{S}_j^\top \nabla \mathbf{R}_j$$

where $\mathbf{S}_j$ is a unit vector which is orthogonal to $\mathbf{R}_j$. Now we note that $\nabla \mathbf{R}_j^\top \mathbf{S}_j \mathbf{S}_j^\top \nabla \mathbf{R}_j = (\mathbf{S}_j \nabla \mathbf{R}_j)^\top (\mathbf{S}_j^\top \nabla \mathbf{R}_j)$ and that the trace of $(\mathbf{S}_j^\top \nabla \mathbf{R}_j)^\top (\mathbf{S}_j^\top \nabla \mathbf{R}_j)$ is $\|\mathbf{S}_j^\top \nabla \mathbf{R}_j\|_2^2$. From this

and using the Cauchy-Schwarz inequality we conclude

$$\begin{aligned}
&\|\sum_{j=1}^{M} w_j \nabla \mathbf{R}_j^\top \mathbf{S}_j \mathbf{S}_j^\top \nabla \mathbf{R}_j\|_1 \geq \sum_{j=1}^{M} w_j \mathrm{trace}(\nabla \mathbf{R}_j^\top \mathbf{S}_j \mathbf{S}_j^\top \nabla \mathbf{R}_j) \geq \sum_{j=1}^{M} w_j \|\mathbf{S}_j^\top \nabla \mathbf{R}_j\|_2^2 \\
&= \sum_{j=1}^{M} w_j \|\mathbf{S}_j^\top \begin{pmatrix} \psi_0(t_j) & 0 & \dots & \psi_m(t_j) & 0 \\ 0 & \psi_0(t_j) & \dots & 0 & \psi_m(t_j) \end{pmatrix} \|_2^2 \\
&= \sum_{j=1}^{M} w_j \sum_{i=0}^{m} (\psi_i(t_j))^2 \underbrace{(\frac{1}{m+1} \sum_{i=0}^{m} 1^2)}_{=1} \geq \sum_{j=1}^{M} w_j \frac{1}{m+1} \underbrace{(\sum_{i=0}^{m} \psi_i(t_j) \cdot 1)^2}_{=1} \geq \frac{Mc}{m+1} > 0
\end{aligned}$$

as the B-splines $\psi_i(t_j)$ sum to one, $w_j > c$ (see Definition 1) and $\|\mathbf{S}_j\| = 1$. □

Consequently, at least for the class of B-Spline curves (which includes polynomial ones), the direct application of the scalar approach is not useful. Hence, we use the approximate Hessian H_F^*, which is based on vector-valued residuals, in order to define a Gauss-Newton-type method. With the help of (5) and (11), we are led to formulate the following definition.

Definition 4. *In each step of the vector-valued-residual-based Gauss-Newton (VGN) method for the generalized fitting problem* (3) *we solve the system of linear equations*

$$\sum_{j=1}^{M} w(\|\mathbf{R}_j(\mathbf{s}^c)\|) \nabla \mathbf{R}_j^\top(\mathbf{s}^c) \nabla \mathbf{R}_j(\mathbf{s}^c) \Delta \mathbf{s} + \sum_{j=1}^{M} w(\|\mathbf{R}_j(\mathbf{s}^c)\|) \nabla \mathbf{R}_j^\top(\mathbf{s}^c) \mathbf{R}_j(\mathbf{s}^c) = 0 \tag{18}$$

for the update step $\Delta \mathbf{s}$ *and compute the update* $\mathbf{s}^+ = \mathbf{s}^c + h\Delta \mathbf{s}$*, where* h *is the step size* $h \leq 1$*.*

It is well known that the Gauss-Newton approach for the generalized fitting problem for scalar residuals is equivalent to a weighted version of the usual Gauss-Newton approach, see [15]. It can be directly seen that this is also true for its vector-valued version VGN. Indeed, (18) is the minimum condition of

$$\sum_{i=1}^{M} w(\|\mathbf{R}_j(\mathbf{s}^c)\|) \|\mathbf{R}_j(\mathbf{s}^c) + \nabla \mathbf{R}_j(\mathbf{s}^c) \Delta \mathbf{s}\|^2 \to \min_{\Delta \mathbf{s}} \tag{19}$$

Again the generalized fitting problem can be solved via a least-squares approach where the summands are multiplied with individual weights. During an iteration step, these weights are kept constant, and then updated with the new residuals. In statistics this approach is called iteratively re-weighted least squares(IRLS).

4 Fitting as an evolution

A geometric interpretation of iterative methods for orthogonal distance regression was introduced in [2]. The intermediate results of the iterative approximation are seen as instances of a continuous family of curves or surfaces generated by an evolution process. We adapt this framework to the case of arbitrary norm–like functions and general distance regression. We assume that the shape parameters $\mathbf{s} = \mathbf{s}(\tau)$ of the curve or surface depend on a time-like variable τ. This produces a time–dependent family of evolving curves or surfaces. Each point $\mathbf{c_s}(t^*)$ with the velocity

$$\mathbf{v}(t^*) = \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{c_s}(t)\,\Big|_{t=t^*} = \nabla_{\mathbf{s}}\mathbf{c_s}(t)\,\Big|_{t=t^*}\,\dot{\mathbf{s}} \tag{20}$$

where the dot indicates the derivative with respect to τ. We restrict ourselves to the case of distance regression, where the parameter values $t^* = t_j$, that are associated with the given points, are kept fixed. The velocities of these points are then given by

$$\mathbf{v}_j = \mathbf{v}(t_j) = \nabla_{\mathbf{s}}\mathbf{c_s}(t)\,\Big|_{t=t_j}\,\dot{\mathbf{s}} = -\nabla\mathbf{R}_j\dot{\mathbf{s}}, \tag{21}$$

due to the definition of the residual vectors.

In order to get a continuous version of VGN, we replace in (19) $\mathbf{s}^c$ by $\mathbf{s}(\tau)$ and $\Delta\mathbf{s}$ by $\dot{\mathbf{s}}$ and obtain

$$\sum_{i=1}^{M} w_j \left(\mathbf{R}_j + \nabla\mathbf{R}_j\dot{\mathbf{s}}\right)^2 = \sum_{i=1}^{M} w_j \left(\,\mathbf{R}_j - \mathbf{v}_j\right)^2 \to \min_{\dot{\mathbf{s}}}. \tag{22}$$

which means that the expected velocity vectors should be equal to the residuals. The geometric interpretation is as follows: *The evolution tries to make the velocity vector* $\mathbf{v}_j$ *of each point* $\mathbf{c_s}(t_j)$ *equal to the corresponding residual vector* $\mathbf{R}_j$,

$$\mathbf{v}_j \approx \mathbf{R}_j \quad (j = 1, \ldots, M), \tag{23}$$

see Fig. 1, left, and this system is solved in the least–squares sense with weights w_j.

Remark 2. Similarly, one can also built up a Gauss-Newton system using $H_F^{\dagger}$.

The geometric interpretation is now as follows: *The evolution tries to make the projection of the velocity vector* $\mathbf{v}_j$ *onto the direction of the residual vector of each point* $\mathbf{c_s}(t_j)$ *equal to the corresponding scalar residual* $\|\mathbf{R}_j\|$,

$$\frac{\mathbf{R}_j^{\top}}{\|\mathbf{R}_j\|}\mathbf{v}_j \approx \|\mathbf{R}_j\| \quad (j = 1, \ldots, M), \tag{24}$$

see Fig. 1, right, and this system is again solved in the least–squares sense with weights w_j.

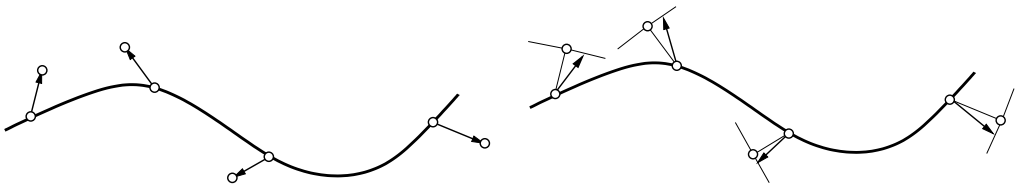


Figure 1: Evolution defined via vector-valued (left) and scalar residuals(right).

These different geometric interpretations (23), (24) have important consequences. In the first place, the approximate Hessian $H_F^\dagger$ that corresponds to the scalar approach does not converge to the true Hessian in the limit. In addition, it may easily encounter singular situations during the approximation process, e.g., if all residual vectors happen to be parallel. Consequently, the uniqueness of the solution is not guaranteed.

Remark 3. In order to guarantee a stable evolution one should introduce a stepsize control that adjusts the value of h in each iteration step. The geometrical interpretation as an evolution offers a simple strategy to do so. In each point $\mathbf{c_s}(t_j)$ one can compute the velocity and choose h such that $\max_j h\mathbf{v}_j \leq \epsilon D$, where D is related to the size of the data (e.g. the diameter of some bounding box) and $\epsilon \leq 1$ some user defined constant. This approach ensures that the points $\mathbf{c_s}(t_j)$ do not move more than a certain percentage (defined by epsilon) of the size of the data.

5 Convergence results

The convergence results for Gauss-Newton methods for scalar residuals in the ℓ_2 case, i.e., $N(x) = x^2$, are well known. Under certain conditions the standard Gauss-Newton method converges quadratically in the zero-residual case, cf. [13]. For the sake of completeness we derive the corresponding statements in the vector-valued situation for a general $N(x)$.

Theorem 5. *Let* $\mathbf{s}^0$ *be a minimizer of*

$$\sum_{j=1}^{M} N(\|\mathbf{R}_j(\mathbf{s})\|) \to \min_{\mathbf{s}}$$

with a norm-like function $N(x)$ *and* $w(x)$ *be a positive and bounded weight function. The residuals* $\mathbf{R}_j$ *shall be Lipschitz continuously differentiable and* $\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j$ *be regular in* $\mathcal{B}_\delta(\mathbf{s}^0) = \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}^0\| < \delta\}$ *with* $\delta > 0$*. Then there exists for the Gauss-Newton method* (18) *some constant* $K \in \mathbb{R}$ *such that*

$$\|\mathbf{s}^+ - \mathbf{s}^0\| \leq K(\|\mathbf{s}^c - \mathbf{s}^0\|^2 + \sum\nolimits_{j=1}^{M} \|\mathbf{R}_j(\mathbf{s}^0)\| \, \|\mathbf{s}^c - \mathbf{s}^0\|).$$

Proof. In order to simplify the notation we shall write $\mathbf{R}_j^c = \mathbf{R}_j(\mathbf{s}^c)$ and where $\mathbf{R}_j^0 = \mathbf{R}_j(\mathbf{s}^0)$, $w_j^c = w_j(\mathbf{R}_j(\mathbf{s}^c))$ and $w_j^0 = w_j(\mathbf{R}_j(\mathbf{s}^0))$. With $\Delta\mathbf{s}$ from Eq. (18), the error after one update

step is given by

$$\begin{aligned}\mathbf{s}^+ - \mathbf{s}^0 &= \mathbf{s}^c - \mathbf{s}^0 + \Delta\mathbf{s} = (\mathbf{s}^c - \mathbf{s}^0) - [\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \nabla\mathbf{R}_j^c]^{-1} \sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \mathbf{R}_j^c \\ &= [\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \nabla\mathbf{R}_j^c]^{-1} \sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c [\nabla\mathbf{R}_j^c(\mathbf{s}^c - \mathbf{s}^0) - \mathbf{R}_j^c].\end{aligned}$$

The last factor [...] can be rewritten as

$$\begin{aligned}\nabla\mathbf{R}_j^c(\mathbf{s}^c - \mathbf{s}^0) - \mathbf{R}_j^c &= \nabla\mathbf{R}_j^c(\mathbf{s}^c - \mathbf{s}^0) - \mathbf{R}_j^0 + \mathbf{R}_j^0 - \mathbf{R}_j^c \\ &= -\mathbf{R}_j^0 + \left(\nabla\mathbf{R}_j^c(\mathbf{s}^c - \mathbf{s}^0) + \mathbf{R}_j^0 - \mathbf{R}_j^c\right) = -\mathbf{R}_j^0 - \varphi'(0) + \varphi(1) - \varphi(0)\end{aligned}$$

with $\varphi(t) = \mathbf{R}_j(\mathbf{s}^c + t(\mathbf{s}^0 - \mathbf{s}^c))$. As

$$-\varphi'(0) + \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) - \varphi'(0)dt$$

and the Lipschitz continuity holds it follows with

$$\|\varphi'(t) - \varphi'(0)\| = \|(\mathbf{R}_j(\mathbf{s}^c + t(\mathbf{s}^0 - \mathbf{s}^c)) - \nabla\mathbf{R}_j(\mathbf{s}^c))(\mathbf{s}^0 - \mathbf{s}^c)\| \leq \alpha t \|\mathbf{s}^0 - \mathbf{s}^c\|^2$$

that

$$\|\nabla\mathbf{R}_j^c(\mathbf{s}^c - \mathbf{s}^0) + \mathbf{R}_j^0 - \mathbf{R}_j^c\| \leq \frac{\alpha}{2}\|\mathbf{s}^c - \mathbf{s}^0\|^2$$

Hence we obtain

$$\|\mathbf{s}^+ - \mathbf{s}^0\| \leq \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \nabla\mathbf{R}_j^c]^{-1}\| \cdot (\|\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \mathbf{R}_j^0\| + \sum_{j=1}^{M} \|\nabla\mathbf{R}_j^{c\top} w_j^c\| \frac{\alpha}{2}\|\mathbf{s}^c - \mathbf{s}^0\|^2).$$

Finally,

$$\|\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \mathbf{R}_j^0\| = \|\sum_{j=1}^{M} (w_j^0 \nabla\mathbf{R}_j^0 - w_j^c \nabla\mathbf{R}_j^c)^\top \mathbf{R}_j^0\| \leq \beta \sum_{j=1}^{M} \|\mathbf{R}_j^0\| \, \|\mathbf{s}^c - \mathbf{s}^0\|$$

since $\sum_{j=1}^{M} w_j^0 (\nabla\mathbf{R}_j^0)^\top \mathbf{R}_j^0 = 0$ and since the Lipschitz continuity of $\nabla\mathbf{R}_j$ implies the Lipschitz continuity of $w_j \nabla\mathbf{R}_j$ with some constant β because w_j is bounded. Summing up,

$$\begin{aligned}\|\mathbf{s}^+ - \mathbf{s}^0\| &\leq \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^{c\top} w_j^c \nabla\mathbf{R}_j^c]^{-1}\| \, [\, \beta \sum_{j=1}^{M} \|\mathbf{R}_j^0\| \, \|\mathbf{s}^c - \mathbf{s}^0\| + \frac{\alpha}{2} \sum_{j=1}^{M} \|\nabla\mathbf{R}_j^{c\top} w_j^c\| \, \|\mathbf{s}^c - \mathbf{s}^0\|^2 \,] \leq \\ &\leq K \, [\, \sum_{j=1}^{M} \|\mathbf{R}_j^0\| \, \|\mathbf{s}^c - \mathbf{s}^0\| + \|\mathbf{s}^c - \mathbf{s}^0\|^2 \,]\end{aligned}$$

with the constant

$$K = \max_{\mathbf{s} \in \mathcal{B}_\delta(\mathbf{s}^0)} \|[\sum_{j=1}^{M} \nabla\mathbf{R}_j^\top w_j \nabla\mathbf{R}_j]^{-1}\| \, [\, \beta + \frac{\alpha}{2} \sum_{j=1}^{M} \|\nabla\mathbf{R}_j^\top w_j\| \,].$$

This concludes the proof. □

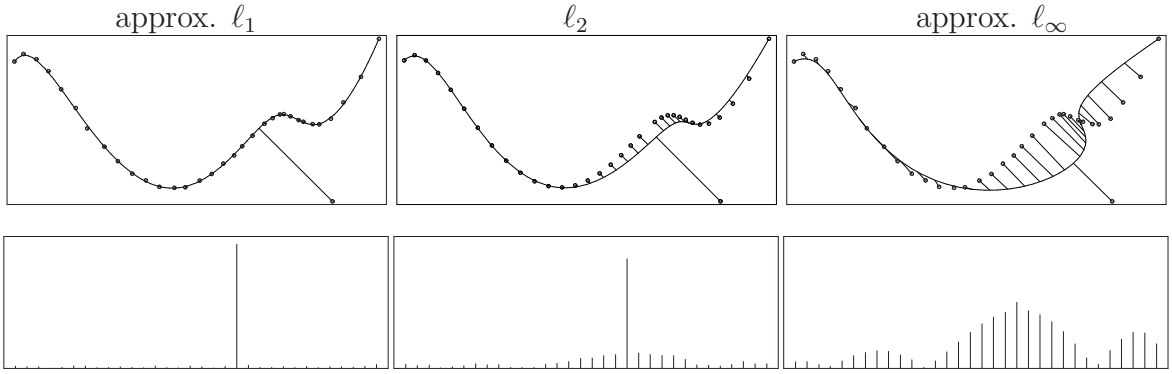


Figure 2: Approximation of point data with different $N(x)$. Top: points and approximating curves. Bottom: error distribution.

According to this result, the Gauss-Newton-type method (Definition 4) performs with quadratic convergence rate in the zero-residual case. However, the assumptions restrict the choice of the weight function $w(x)$, and therefore of the norm–like function $N(x)$. More precisely, the weight function $w(x)$ must be positive and has to be bounded for $x \in [0, \epsilon]$.

The weights of the ℓ_p norms for $p \neq 2$ do not fulfill these conditions. In the case $p < 2$, even a single vanishing residual causes the Hessian to diverge.

Consider the case $p > 2$. In a zero–residual case, all residuals vanish simultaneously which causes the Hessian to be singular. In the general situation, some of the weights may vanish in the limit. This has the same effect as if the points with the associated weights would be excluded from the data set. Since we assume that $M \gg n$ this will not influence the regularity compared to the ℓ_2 case too much, since sufficiently many points can be expected to remain. Consequently, one may expect that the method works well for most data sets.

6 Examples

We present three numerical examples.

Example 4. Fig. 2 shows a point cloud which was approximated with a polynomial curve of degree 5 that interpolates the first and the last data point. The three pictures in the top row show the sensitivity of approximations obtained using different norm–like functions $N(x)$ with respect to outliers. The first result was obtained with $N(x) = 1 - \exp(-(\eta x)^2)$ which shall serve as an approximation of the ℓ_1 norm, where the choice of η will be clearified in the next remark. In the statistics literature, the weight derived from this function is called the Welsch weight [9]. The second result is the ℓ_2 fit, where $N(x) = x^2$. For the third result, the maximal distance from the curve to the data points was to be minimized. This corresponds to the ℓ_∞ norm. We replaced this non-differentiable norm by $N(x) = \exp((\eta x)^2) - 1$. The plots in the bottom row visualize the lengths of the error vectors. From left to right the maximal error decreases, while the distribution of the error gets

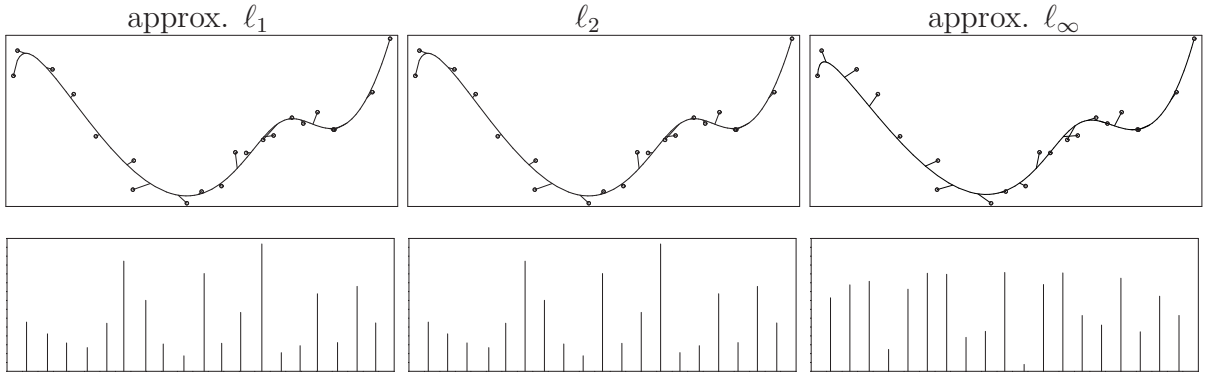


Figure 3: Approximation of point data with different $N(x)$. Top: fitted curves. Bottom: distances from data to curve.

more uniform.

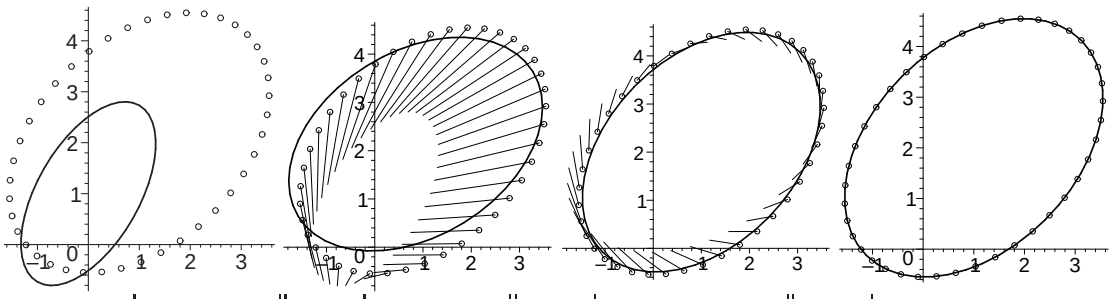
Remark 4. The results obtained by minimizing the objective functions with $N(x) = 1 - \exp(-x^2)$ and $N(x) = \exp(x^2) - 1$ are not invariant under scaling. The parameter η is introduced in order to avoid this dependency. A possible strategy for choosing the parameter η is the following. First we compute the ℓ_2 fit of the data points and identify the point which has the largest distance to the curve or surface. Then η is chosen such that this data point has some prescribed weight w_0. This does not require much additional computational effort, since the ℓ_2 approximation should be computed anyway, in order to obtain a suitable initial value for the iteration process.

Example 5. Fig. 3 shows a point set that was obtained from the data of the previous example by adding some additional noise. Also the single outlier was eliminated. From left to right one can see again an approximate ℓ_1 fit, an ℓ_2 approximation and an approximate ℓ_∞ fit. Again, the magnitudes of the corresponding errors in the bottom line show that the maximal error decreases, while the distribution becomes more uniform. However, the difference between the approximate ℓ_1 fit and the ℓ_2 fit are not that significant, since the data points were perturbed with a uniform random error.

Example 6. In order to demonstrate the quadratic convergence rate in the zero-residual case, we fitted an ellipse-shaped point cloud, as one can see in Fig. 4. The model curve and the shape parameters were chosen according to Example 1. The errors in the first 7 steps are reported in the table. After 3 steps the new error is roughly the square of the previous one, which indicates the quadratic convergence.

7 Conclusion

The generalized fitting problem (3) can be solved iteratively by using two different Gauss-Newton-type methods. These methods are in fact equivalent to the technique of iteratively



step	error	step	error	step	error	step	error
0	11.583450	2	0.4580835	4	0.5437303e-6	6	0.9761834e-30
1	3.9724490	3	0.0041590	5	0.5572782e-14	7	0.1796254e-61

Figure 4: Top: Zero-residual approximation with an ellipse; initial position and after 1, 2 and 4 steps and associated error vectors. Bottom: approximation errors.

re-weighted least-squares, where the weights $w(x)$ and the corresponding norm-like function $N(x)$ are related by $N'(x) = x\,w(x)$. The direct application of the results which are available in the scalar case, which is well understood in statistics does not give useful results. The main difference between both approaches were illustrated by interpreting both techniques as evolution processes.

In addition, based on classical results about Gauss-Newton methods, we analyzed the convergence properties. By choosing other norm-like functions than $N(x) = x^2$, which corresponds to the ℓ_2 norm, one may efficiently deal with data containing outliers, and one may find curves and surfaces that minimize the maximum distance error.

Acknowledgment The authors were supported by the Austrian Science Fund (FWF) through the research network S92 "Industrial Geometry", subproject 2.

References

[1] M. Aigner and B. Jüttler. Approximation flows in shape manifolds. In P. Chenin, T. Lyche, and L.L. Schumaker, editors, *Curve and Surface Design: Avignon 2006*, pages 1–10. Nashboro Press, 2007.

[2] M. Aigner, Z. Šír, and B. Jüttler. Evolution-based least-squares fitting using Pythagorean hodograph spline curves. *Comp. Aided Geom. Design*, 24:310–322, 2007.

[3] I. Al-Subaihi and G.A. Watson. The use of the l_1 and l_∞ norms in fitting parametric curves and surfaces to data. *Appl. Numer. Anal. Comput. Math.*, 1:363–376, 2004.

[4] M. Alhanaty and M. Bercovier. Curve and surface fitting and design by optimal control methods. *Computer–Aided Design*, 33:167–182, 2001.

[5] A. Atieg and G.A. Watson. A class of methods for fitting a curve or surface to data by minimizing the sum of squares of orthogonal distances. *J. Comput. Appl. Math.*, 158:277–296, 2003.

[6] A. Blake and M. Isard. *Active contours*. Springer, New York, 1998.

[7] P. T. Boggs, R. H. Byrd, and R. B. Schnabel. A stable and efficient algorithm for nonlinear orthogonal distance regression. *J. Sci. Stat. Comput.*, 8(6):1052–1078, 1987.

[8] K. P. Bube and R. T. Langan. Hybrid ℓ_1/ℓ_2 minimization with application to tomography. *Geophysics*, 62:1183–1195, 1997.

[9] P.W. Holland and R. Welsch. Robust regression using iteratively re-weighted least-squares. *Communications in Statistics: Theory and Methods*, 9(A6):813–827, 1977.

[10] J. Hoschek and D. Lasser. *Fundamentals of computer aided geometric design*. A K Peters, Wellesley, MA, 1993.

[11] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[12] B. Jüttler. Computational methods for parametric discrete ℓ_1 and ℓ_∞ curve fitting. *Int. J. Shape Modelling*, 4:21–34, 1998.

[13] C. T. Kelley. *Iterative Methods for Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[14] Y. Liu and W. Wang. A revisit to least squares orthogonal distance fitting of parametric curves and surfaces. In F. Chen and B. Jüttler, editors, *Advances in Geometric Modeling and Processing, GMP 2008*, pages 384–397. 2008.

[15] V. Mahadevan, H. Narasimha-Iyer, B. Roysam, and H. L. Tanenbaum. Robust model-based vasculature detection in noisy biomedical images. *IEEE Transactions on Information Technology in Biomedicine*, 8(3):360–376, 2004.

[16] P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman & Hall, London, 1998.

[17] M. R. Osborne. *Finite Algorithms in Optimization and Data Analysis*. John Wiley & Sons, New York, 1985.

[18] H. Pottmann, S. Leopoldseder, M. Hofer, T. Steiner, and W. Wang. Industrial geometry: recent advances and applications in CAD. *Comp.-Aided Design*, 37:751–766, 2005.

[19] D. Rogers and N. Fog. Constrained B-spline curve and surface fitting. *Computer–Aided Design*, 21:641–648, 1989.

[20] B. Sarkar and C.-H. Menq. Parameter optimization in approximating curves and surfaces to measurement data. *Comp. Aided Geom. Design 8*, 8:267–280, 1991.

[21] T. Speer, M. Kuppe, and J. Hoschek. Global reparametrization for curve approximation. *Comp. Aided Geom. Design*, 15:869–877, 1998.

[22] W. Wang, H. Pottmann, and Y. Liu. Fitting B-spline curves to point clouds by curvature-based squared distance minimization. *ACM Trans. Graph.*, 25(2):214–238, 2006.

[23] G. A. Watson. Approximation in normed linear spaces. *J. Comput. Appl. Math.*, 121:1–36, 2000.

[24] G. A. Watson. On the Gauss-Newton method for ℓ_1 orthogonal distance regression. *IMA J. Num. Anal.*, (22):345–357, 2001.

[25] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974.

[26] T. Yamamoto. Historical developments in convergence analysis for Newton's and Newton-like methods. *J. Comput. Appl. Math.*, 124(1-2):1–23, 2000.