Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces

Axel Munk\textsuperscript{1}, Nicolai Bissantz\textsuperscript{2}, Lutz Dümbgen\textsuperscript{3}, and Bernd Stratmann\textsuperscript{1}

\textsuperscript{1} Institute for Mathematical Stochastics
Georg-August-University Göttingen, Germany
\textsuperscript{2} Faculty of Mathematics
Ruhr-University Bochum, Germany
and
\textsuperscript{3} Institute of Mathematical Statistics and Actuarial Science
University of Bern, Switzerland

September 18, 2007

Abstract

The computation of robust regression estimates often relies on minimization of a convex functional on a convex set. In this paper we discuss a global technique for a large class of convex functionals to compute the minimizers iteratively which is closely related to majorization-minimization algorithms. Our approach is based on a quadratic approximation of the functional to be minimized. We show geometric convergence on general convex function spaces for general coercive and convex functionals $F$. This includes the iteratively reweighted least squares algorithm as a special case. The algorithm is applied to TV penalized quantile regression and is compared with a step size corrected Newton Raphson algorithm. It is found that in the first steps the iteratively reweighted least squares algorithm performs significantly better, whereas after an intermediate number of iterations both behave comparable. Finally, we indicate for bivariate regression with a unimodality constraints how this algorithm can serve as a linking device for highly efficient algorithms to nonquadratic minimization problems.

AMS 2000 subject classification: primary 62G07; secondary 62J05, 65K05, 85-08

Keywords: regression analysis, monotone regression, quantile regression, shape constraints, $L^1$ regression, nonparametric regression, total variation semi-norm, reweighted least squares, Fermat’s problem, convex approximation, quadratic approximation, pool adjacent violators algorithm

Address for correspondence: Nicolai Bissantz, Fakultät für Mathematik, Universitätstraße 150, Mathematik III, NA 3/70, D-44780 Bochum, Germany; email: nicolai.bissantz@rub.de

1 Introduction

The computation of robust parametric and nonparametric regression estimators often requires the minimization of (convex) functionals on a set $C$ which is determined by a priori information on the
model underlying the data. For example, \( C \) can be a linear finite-dimensional space (linear model) or the set of isotonic vectors \( m = (m_1, ..., m_d) \in \mathbb{R}^d \), \( m_1 \leq ... \leq m_d \), with \( d \leq n \). To this end the functional

\[
F^{(\rho)}(m) = \sum_{i=1}^{n} \rho(r_i(m))
\]

(1)

has to be minimized over \( C \subset \mathbb{R}^d \). Here \( r_i, i = 1, ..., n \) denote the (model-dependent) residuals of \( n \) data pairs \((X_i, Y_i), i = 1, ..., n \) and \( \rho \) a given function (Huber 1981). Taking \( \rho(z) = z^2/2 \) gives the ordinary least squares problem, and

\[
\rho(z) = 2|z| \cdot \begin{cases} 
p & z \geq 0 \\
1-p & z < 0
\end{cases}
\]

(2)

with \( 0 < p < 1 \) yields quantile regression (Koenker & Bassett 1978, Portnoy 1997). Other functions are logistic \( \rho(z) = \gamma z \log(\cosh(z/\gamma)) \) (Coleman et al. 1980) or Huber’s (1964) loss function

\[
\rho(z) = \begin{cases} 
z^2/2 & |z| \leq \gamma \\
\gamma|z| - \gamma^2/2 & |z| > \gamma
\end{cases}
\]

for some \( \gamma > 0 \). An important extension of (1) are functionals

\[
F(m) = F^{(\rho)}(m) + \lambda P(m), \quad \lambda \geq 0,
\]

(3)

where \( P(m) \) denotes a penalizing term such as, for instance, the discrete total variation semi-norm of \( m \in \mathbb{R}^d \),

\[
P(m) = \sum_{j=1}^{d-1} |m_j - m_{j+1}|;
\]

(4)

see Künsch (1994), Koenker, Ng & Portnoy (1994) or Mammen & van de Geer (1997). In this paper a generalization of the iteratively reweighted least squares (IRLS) algorithm - therefore named GIRLS - is considered for minimization of a functional \( F \) as in (3) over any convex subset \( \mathcal{C} \) of \( \mathbb{R}^d \). This allows us to extend the IRLS algorithm for example to situations where \( \mathcal{C} \) is defined as the space of monotone (or \( k \)-modal) vectors or to the problem of nonparametric regression estimates with total variation semi-norm penalization of its discrete derivative.

The general idea of the IRLS algorithm (and variants of it) is to approximate the functional \( F \) in a first step by smooth functionals \( F_\delta \) such that \( F_\delta \to F \) pointwise as \( \delta \searrow 0 \). The collection \( \{F_\delta\}_{\delta > 0} \) will be called a regularization of \( F \) (cf. Def. 1). In a second step, for each given base point \( f \in \mathcal{C} \) the functional \( F_\delta \) will be approximated by \( G_\delta(f, \cdot) \) (cf. Def. 2). Here \( G_\delta : \mathcal{C} \times \mathbb{R}^d \to \mathbb{R} \) is a functional which is chosen such that a quick and numerically stable iterative minimization can be performed. The resulting minimizer will serve as an approximation for the minimizer \( m_\delta^* \) of \( F_\delta \).
and hence for a minimizer $m^*$ of $F$. In particular, if it is possible to choose $G_\delta$ as a polynomial of degree two, the well known iteratively reweighted least squares algorithm may result (Lejeune & Sarda 1988, McCullagh & Nelder 1989, Dodge & Jurečková 2000).

Now, the GIRLS algorithm can be summarized schematically as follows:

**The GIRLS algorithm**

**STEP 1** Determine a regular approximation $F_\delta: \mathbb{R}^d \to \mathbb{R}$, $\delta > 0$, for $F: \mathcal{C} \to \mathbb{R}$.

**STEP 2** Determine a smooth approximation to $F_\delta$ from above $G_\delta: \mathcal{C} \times \mathbb{R}^d \to \mathbb{R}$.

**STEP 3** Initialize $m^{(0)}_\delta$ and $\delta > 0$.

**STEP 4** Iterate

$$m^{(k+1)}_\delta := \arg\min_{m \in \mathcal{C}} G_\delta(m^{(k)}_\delta, m) \quad \text{for } k = 0, 1, 2, \ldots$$

(5)

**STEP 5** Terminate the iteration for a proper $k$ and repeat for $\delta' < \delta$.

**STEP 6** Terminate $\delta$.

A brief discussion of proper selection of the regularisation parameter $\delta$ and the iteration $k$ is provided in Section 3.2.

The IRLS and related algorithms are based on the idea of majorizing functionals by a sequence of quadratic approximations and subsequent minimization. These have been treated extensively in the literature, e.g. Kuhn (1972), Katz (1973), Wolke & Schwetlick (1988), O’Leary (1990), de Leeuw & Michailidis (2000), Hunter & Lange (2000), Lange, Hunter & Yang (2000), Vardi & Zhang (2000, 2001), and the references therein. However, in most cases convergence is only shown for $\mathcal{C} = \mathbb{R}^d$. This simplifies proofs notably, since the minimizers can be represented as zeros of the derivatives of the functional. For arbitrary convex $\mathcal{C}$, however, the minimizers are no longer represented solely by such equality constraints, instead inequalities occur. Notable exceptions for general convex $\mathcal{C}$ are Eckhardt (1980), where however, the convergence results are restricted to a special class of functionals, requiring e.g. $F(m) = O(\|m\|)$, or Voß & Eckhardt (1980), who show convergence on convex polyhedral sets under certain regularity conditions on the functional.

Our findings generalize these results to the case of $\mathcal{C}$ being an arbitrary convex closed set as well as to more general functionals which are only required to be coercive and convex. This appears to be close to the weakest possible set of assumptions required for a general proof of convergence. Our proof adopts various arguments from convex analysis.
It is interesting to note that in the numerical literature the IRLS algorithm is denoted as the Weiszfeld algorithm (Weiszfeld, 1936, 1937) who suggested this algorithm to solve the Fermat-Steiner-Weber problem (Weiszfeld 1936, 1937, Kuhn 1973, Katz 1974) which is known to the statistical community as the computation of the spatial median (as mentioned in Brown 1983, Brown et al. 1997, Ducharme & Milasevic 1987).

The paper is organized as follows. First, we motivate the GIRLS algorithm for the special case of $L^1$-regression in Section 2. Then we present the GIRLS algorithm in a general framework and prove various results about its convergence in Section 3. Its convergence to the minimizer $m^*_\delta$ and hence to $m^*$ as $\delta \searrow 0$ will be shown under very general assumptions (Theorem 2). Furthermore, in Theorem 5 we prove geometric, or, more precisely, at least $Q$-linear convergence of the sequence $(m_k)_k$ to $m^*_\delta$ under slightly stronger conditions (cf. Voß & Eckhardt, 1980, and Böhning & Lindsay, 1988), and guidance is provided on the choice of the number of iterates in (5) and the regularization parameter $\delta$. Finally, we show in Theorem 3 that any convex and coercive functional $F$ can be regularized by a sequence $F_\delta$ s.t. each $F_\delta$ admits a quadratic approximation $G_\delta$ from above.

We stress that an advantage of the GIRLS approach is that there is some flexibility in the choice of $F_\delta$ and $G_\delta$. This choice can be driven by various aspects, such as computational efficiency or rate of convergence (cf. Theorem 3). In this paper we emphasis the possibility to make use of efficient algorithms already available for the minimization of $G_\delta$, such as the pooling adjacent violator algorithm (PAVA) for isotonic weighted least squares approximation (see Robertson, et al. 1988 for a comprehensive treatment). This is illustrated in Section 4, where we describe the construction of $F_\delta$ and $G_\delta$ in some specific cases explicitly. In Section 5 we discuss two numerical examples. In the first example we investigate in detail numerical performance of the GIRLS algorithm for the case of TV penalized quantile regression. To this end the GIRLS is compared to a step size corrected Newton algorithm and it is found that it outperforms the latter one in the first iteration steps significantly, in particular when the initial value is far from the optimum. Moreover, both algorithms behave similar after an intermediate number of iterations. This finding coincides with other numerical experiments, e.g., where we have applied the algorithm to $L^1$-penalized Poisson regression.

In the second example we discuss an application of the GIRLS algorithm to the estimation of the variance surface of brightness data from astrophysical measurements. Here the data is on a two-dimensional grid and we impose a unimodality constraint in one direction and penalization
of a regression function’s non-smoothness. Note, that the GIRLS algorithm allows to include the PAVA for the univariate unimodal subproblem, which is in general not possible for regression with two- or higher dimensional predictor. Note also, that PAVA type algorithms are not available in general if an additional penalization term as in (3) is added. Again, the GIRLS algorithm offers a possibility to include them in each updating step.

In summary, the main advantage of the GIRLS algorithm is twofold. First, it is simple to perform and offers great flexibility for the choice of the approximating functionals \( G_\delta \). Second, it allows us to combine various restrictions and minimisation criteria (such as monotonicity constraints and roughness penalties). For such complex minimisation problems simple and quick algorithms such as PAVA or Newton type algorithms are not available in general, and more complicated and time consuming algorithms such as quadratic programming or interior point methods become necessary. Here the GIRLS algorithm represents a feasible alternative because it typically requires in each updating step the computation of minimizers (e.g. a weighted \( L^2 \) solution), which can be obtained easily. Further, our numerical experiments have shown that a rather small number of updating steps give already satisfactory results. Hence, as a practical rule of thumb, we find that the GIRLS algorithm can be used for the quick improvement of an initial value by a few iterations. Subsequent iterations can be performed by any other optimization algorithm.

2 \( L^1 \)-regression with the GIRLS algorithm

As a motivating example consider the \( L^1 \) linear regression problem for observations \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) in \( \mathbb{R}^d \times \mathbb{R} \). Assuming that \( Y_i \) equals \( X_i^\top m \) plus a random error, the goal is to compute

\[
m := \arg\min_{m \in \mathbb{R}^d} \sum_{i=1}^n |Y_i - X_i^\top m| = \arg\min_{m \in \mathbb{R}^d} F(m),
\]

an estimator of the unknown parameter vector \( m \in \mathbb{R}^d \). Iteratively reweighted least squares is based on the idea that, in a first step, the \( L^1 \) norm \( F \), being a convex functional, will be approximated (regularized) by a family of smooth convex functionals \( F_\delta, \delta > 0 \), e.g.

\[
F_\delta(m) = \sum_{i=1}^n h_\delta \left( Y_i - X_i^\top m \right),
\]

where

\[
h_\delta(z) = [z^2 + \delta]^{1/2}.
\]
The regularisation of a nonsmooth functional as in (6) by (7) is well known, of course (see e.g. Vogel & Oman (1996)). It is supposed that minimisation of \( F_\delta \) is numerically better tractable than minimisation of the original functional \( F \) in (6). Then \( m_\delta := \arg\min_{m \in \mathbb{R}^d} F_\delta(m) \) will be an approximation of \( m \) (cf. Theorem 1). In order to compute \( m_\delta \) the following recursion formula is iterated:

\[
m_\delta^{(k+1)} = \arg\min_{m \in \mathbb{R}^d} \sum_{i=1}^n \frac{(Y_i - X_i^T m)^2}{h_\delta(Y_i - X_i^T m^{(k)})}.
\]

(8)

Note, that in each updating step the computation of \( m_\delta^{(k+1)} \) means solving a simple diagonally reweighted least squares minimisation problem, which can easily be done by using standard methods such as, e.g., Householder QR decomposition. As a starting value \( m_\delta^{(0)} \) any (reasonable) choice, e.g. the least squares estimator, may serve.

It is instructive to indicate a proof for this simple case. The basic idea is to approximate \( h_\delta(z) \) from above for any given real number \( r \) by a quadratic function \( g_\delta(r, z) = c(r) + a(r)z^2/2 \) of \( z \) such that \( g_\delta(r, \cdot) \geq h_\delta \) and \( g_\delta(r, r) = h_\delta(r) \). This can be achieved indeed with

\[
g_\delta(r, z) = h_\delta(r) + h_\delta(r)^{-1}(z^2 - r^2)/2;
\]

(9)

see also Lemma 1 in Section 4. The intrinsic reason is that \( h_\delta \) is an even convex function whose second derivative \( h''_\delta \) is non-increasing on \([0, \infty)\). Thus \( m_\delta^{(k+1)} \) in (8) is the minimizer of

\[
G_\delta(m_\delta^{(k)}, m) := \sum_{i=1}^n g_\delta(Y_i - X_i^T m_\delta^{(k)}, Y_i - X_i^T m)
\]

over all \( m \in \mathbb{R}^d \). Note that \( F_\delta \) as well as \( G_\delta(m_\delta^{(k)}, \cdot) \) are convex functions such that \( F_\delta(m) \leq G_\delta(m_\delta^{(k)}, m) \) with equality for \( m = m_\delta^{(k)} \), and their gradients satisfy \( \nabla F_\delta(m_\delta^{(k)}) = \nabla G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) \).

Here and in the following the gradient of \( G_\delta \) is defined with respect to the second argument. Thus

\[
F_\delta(m_\delta^{(k+1)}) \leq G_\delta(m_\delta^{(k)}, m_\delta^{(k+1)}) \leq G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) = F_\delta(m_\delta^{(k)}),
\]

and the last inequality is strict if, and only if, \( m_\delta^{(k)} \) differs from the solution \( m_\delta \). Consequently, \( F_\delta(m_\delta^{(k)}) \) is either strictly decreasing in \( k \), or \( m_\delta^{(k)} = m_\delta \) for sufficiently large \( k \). This fact was established by Lejeune & Sarda (1988) for the particular problem (6). Convergence of \( m_\delta^{(k)} \) to \( m_\delta \) as \( k \to \infty \) follows from our general Theorem 2 below.
3 The GIRLS algorithm

3.1 Main theorem and convergence analysis

Now let us turn to the general setting. We always assume that our target functional $F : \mathbb{R}^d \to \mathbb{R}$ is convex and coercive, i.e. $F(x) \to \infty$ as $\|x\| \to \infty$. Moreover, let $C \subset \mathbb{R}^d$ be closed and convex. This entails that the set

$$M^* := \text{argmin}_{m \in C} F(m)$$

is a nonvoid, compact and convex subset of $C$. Now the first step is to approximate $F$ by a family of strictly convex and smooth functionals $F_\delta, \delta > 0$, converging pointwise to $F$ as $\delta \downarrow 0$. This is summarized in the following

**Definition 1.** A functional $F_\delta : \mathbb{R}^d \to \mathbb{R}$ is called regular, if $F_\delta$ is strictly convex, continuously differentiable and coercive. A regular class for $F$ (or a regularization of $F$) consists of regular functionals $F_\delta, \delta > 0$, such that $F_\delta$ converges pointwise to $F$ as $\delta \downarrow 0$.

Theorem 4 below shows that there exists always a regular class $(F_\delta)_{\delta > 0}$ for $F$. It follows from strict convexity and coercivity of $F_\delta$ that it has a unique minimizer

$$m^*_\delta := \text{argmin}_{m \in C} F_\delta(m)$$

which serves as an approximation to $M^*$. The next theorem provides an exact formulation of this fact.

**Theorem 1.** (Approximation of $M^*$).

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex and coercive functional, and let $(F_\delta)_{\delta > 0}$ be a regularization of $F$. Then, as $\delta \downarrow 0$,

$$F_\delta(m^*_\delta) \to \min_{x \in C} F(x) \quad \text{and} \quad d(m^*_\delta, M^*) := \inf_{y \in M^*} \|m^*_\delta - y\| \to 0 .$$

Before proving Theorem 1 we summarize some well known facts about convex functionals (see Rockafellar 1970) which we utilize in the subsequent proofs. A convex functional on $\mathbb{R}^d$ is automatically continuous. If a sequence of convex functionals on $\mathbb{R}^d$ converges pointwise, then the convergence is uniform on arbitrary bounded sets. Finally, if $H : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and if $C \subset \mathbb{R}^d$ is closed and convex, then $f \in C$ minimizes $H$ over $C$ if, and only if,

$$\nabla H(f)^\top (m - f) \geq 0 \quad \text{for all } m \in C . \quad (10)$$
Proof. For any set $S \subset \mathbb{R}^d$ let $\|F - F_\delta\|_S$ be the supremum norm of $F - F_\delta$ over $S$. For any fixed $\epsilon > 0$, the set $B_\epsilon := \{m \in C : d(m, M^*) \leq \epsilon\}$ is compact. Thus $\|F - F_\delta\|_{B_\epsilon}$ tends to zero as $\delta \searrow 0$. In particular, for sufficiently small $\delta > 0$,

$$\min_{m \in C : d(m, M^*) = \epsilon} F_\delta(m) > \max_{m \in M^*} F_\delta(m).$$

In order to see the last inequality, note that this holds for $F$ and use that $F_\delta \to F$ uniformly on bounded sets. But this implies that $F_\delta(m_o) > \min_{m \in B_\epsilon} F_\delta(m)$ for any $m_o \in C \setminus B_\epsilon$, i.e. $m_o^* \in B_\epsilon$.

For let $m_o$ be the metric projection of $m_o$ onto $M^*$ and write $m_o = m_* + tv$ for some unit vector $v \in \mathbb{R}^d$ and a scalar $t > \epsilon$. Then it follows from convexity of $F_\delta$ that

$$F_\delta(m_o) - F_\delta(m_*) \geq (\epsilon/t)(F_\delta(m_* + \epsilon v) - F_\delta(m_*)) > 0.$$ 

Note also that in case of $m_*^* \in B_\epsilon$,

$$|F_\delta(m_*^*) - F(m_*^*)| \leq \|F - F_\delta\|_{B_\epsilon},$$

and

$$F(m_*^*) - \min_{m \in C} F \leq \max_{m \in B_\epsilon} \left( F(m) - \min_{M^*} F \right).$$

Finally, the r.h.s. of this inequality can be made arbitrarily small for proper choice of $\delta$ and $\epsilon$ by compactness of $M^*$ and continuity of $F$. \hfill \Box

The second step is to determine $m_*^\delta$ via approximations $G_\delta(f, \cdot)$ of $F_\delta$ for various $f \in C$ as in (5). The following definition summarizes our assumptions on $G_\delta$.

**Definition 2.** Let $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be a regular functional. Another functional $G_\delta : C \times \mathbb{R}^d \to \mathbb{R}$ is called a smooth approximation of $F_\delta$ from above, if it is continuous in both arguments and satisfies the following additional properties for arbitrary $f \in C$:

(i) $G_\delta(f, \cdot)$ is strictly convex and continuously differentiable,

(ii) $G_\delta(f, m) \geq F_\delta(m)$ for all $m \in \mathbb{R}^d$ with equality for $m = f$.

The functional $G_\delta$ is called a quadratic approximation of $F_\delta$ from above if, in addition, $G_\delta(f, \cdot)$ is always a polynomial of order two, i.e.

$$G_\delta(f, m) = F_\delta(f) + \nabla F_\delta(f)^\top (m - f) + 2^{-1} (m - f)^\top B(f)(m - f)$$ (11)

for some symmetric, positive definite matrix $B(f) \in \mathbb{R}^{d \times d}$.

Note that from (i) and (ii) it follows that $\nabla F_\delta(f) = \nabla G_\delta(f, f)$. The next theorem is the main result of this paper.
Theorem 2. (Convergence of the GIRLS algorithm).

Let \( C \subset \mathbb{R}^d \) be a closed convex set and \( F_\delta : \mathbb{R}^d \to \mathbb{R} \) be a regular functional which can be smoothly approximated from above by \( G_\delta \). Then the GIRLS algorithm, defined by (5) with an arbitrary starting point \( m_\delta^{(0)} \in C \), yields a sequence \( (m_\delta^{(k)}) \) converging to \( m_\delta^* \).

Proof. At first we prove that \( F_\delta(m_\delta^{(k)}) \) is decreasing in \( k \). It follows from property (ii) in Definition 2 that the gradients \( \nabla G_\delta(m_\delta^{(k)}, m) \) (with respect to the second argument) and \( \nabla F_\delta(m) \) coincide for \( m = m_\delta^{(k)} \). Thus it follows from (10) that \( m_\delta^{(k+1)} = m_\delta^{(k)} \) if, and only if, \( m_\delta^{(k)} = m_\delta^* \). Otherwise,

\[
F_\delta(m_\delta^{(k+1)}) \leq G_\delta(m_\delta^{(k)}, m_\delta^{(k+1)}) < G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) = F_\delta(m_\delta^{(k)}).
\]

By monotonicity of \( (F_\delta(m_\delta^{(k)}))_k \), all points \( m_\delta^{(k)} \) lie in the set \( \{ m \in C : F_\delta(m) \leq F_\delta(m_\delta^{(0)}) \} \), which is compact by continuity and coercivity of \( F_\delta \). Hence it is sufficient to show that any limit point \( m_o \) equals \( m_\delta^* \). Now, take an arbitrary convergent subsequence \( (m_\delta^{(k_\ell)})_\ell \) with limit \( m_o \). For any \( v \in C \),

\[
F_\delta(m_\delta^{(k_\ell+1)}) \leq G_\delta(m_\delta^{(k_\ell)}, m_\delta^{(k_\ell+1)}) \leq G_\delta(m_\delta^{(k_\ell)}, v) \to G_\delta(m_o, v) \text{ as } \ell \to \infty,
\]

by continuity of \( G_\delta \). But

\[
\lim_{\ell \to \infty} F_\delta(m_\delta^{(k_\ell+1)}) \geq \lim_{\ell \to \infty} F_\delta(m_\delta^{(k_\ell+1)}) = F_\delta(m_o) = G_\delta(m_o, m_o).
\]

Thus \( G_\delta(m_o, m_o) \leq G_\delta(m_o, v) \) for all \( v \in C \), i.e. \( m_o \) is the unique minimizer of \( G_\delta(m_o, \cdot) \). As argued above, this entails that \( m_o = m_\delta^* \).

The next theorem states that convex and coercive functionals \( F \) can always be regularized and approximated quadratically from above. Hence GIRLS is, in principle, always applicable.

Theorem 3. (Regularization and approximation of \( F \)).

Let \( F : \mathbb{R}^d \to \mathbb{R} \) be a convex and coercive functional. Then there exists a regularization \( (F_\delta)_{\delta > 0} \) of \( F \) such that each \( F_\delta \) admits a quadratic approximation \( G_\delta \) from above.

In order to prove Theorem 3 we require the following result.
**Theorem 4.** Let $F$ be a nonnegative, coercive, convex functional on $\mathbb{R}^d$. Then there are strictly convex and infinitely often differentiable functionals $F_\delta \geq F$, $\delta > 0$, such that $F_\delta \rightarrow F$ pointwise as $\delta \searrow 0$.

**Proof.** Let $K(x) := 1\{\|x\| < 1\} C \exp \left(- (1 - \|x\|^2)^{-1}\right)$, where $C$ is chosen such that $K$ integrates to one. This is a well-known example of an infinitely differentiable, nonnegative, even kernel function with compact support $\{x : \|x\| \leq 1\}$. For $\delta > 0$ we define $K_\delta(x) := \frac{1}{\delta} K\left(\frac{1}{\delta} x\right)$ and $F_\delta(x) := \int F(y) K_\delta(x - y) \, dy = \int F\left(x + \delta z\right) K(z) \, dz$.

Elementary calculus shows that $F_\delta$ is convex and infinitely often differentiable with limit $F$ pointwise. Moreover, since $K$ is even, $F_\delta(x) = \int F\left(x + \delta z\right) + F\left(x - \delta z\right) \, dz \geq \int F(x) K(z) \, dz = F(x)$, by convexity of $F$. Finally, if $F_\delta$ fails to be strictly convex, we may add to $F_\delta$ the strictly convex function $x \mapsto \delta \|x\|^2$.

We mention that the construction of $F_\delta$ given here is mainly for theoretical purposes, and may in practice be difficult to evaluate numerically due to the high dimensionality of the integral.

**Proof of Theorem 3.** Let $(F_\delta)_{\delta > 0}$ be a regularization of $F$ such that $D^2 F_\delta$ is positive definite everywhere; cf. Theorem 4 and its proof. It may happen that $\limsup_{\|m\| \to \infty} \frac{F_\delta(m)}{\|m\|^2} = \infty$, rendering quadratic approximation of $F_\delta$ from above impossible. Thus we modify the functions $F_\delta$ as follows: Let

$$c_\delta := \max_{\|m\| \leq \delta^{-1}} \lambda_{\text{max}}(D^2 F_\delta(m))$$

with $\lambda_{\text{max}}(A)$ denoting the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d \times d}$. Starting from the representation

$$F_\delta(m) = F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top D^2 F_\delta(tm) m (1 - t) \, dt,$$

we define

$$\tilde{F}_\delta(m) := F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top \min(D^2 F_\delta(tm), c_\delta I) m (1 - t) \, dt.$$

Here $\min(A, c_\delta I) \in \mathbb{R}^{d \times d}$ is obtained from the spectral representation of $A$ by replacing each eigenvalue $\lambda_i(A)$ with $\min(\lambda_i(A), c_\delta)$. Note that $\tilde{F}_\delta$ is twice continuously differentiable with $\tilde{F}_\delta(0) = F_\delta(0)$, $\nabla \tilde{F}_\delta(0) = \nabla F_\delta(0)$ and $D^2 \tilde{F}_\delta = \min(D^2 F_\delta, c_\delta I)$. The hessian matrix is positive.
definite with largest eigenvalue never exceeding \( c_\delta \). In addition, \( \tilde{F}_\delta = F_\delta \) on \( \{ m : \| m \| \leq \delta^{-1} \} \). Thus for sufficiently small \( \delta > 0 \), \( \tilde{F}_\delta \) is regular, and a quadratic approximation of \( \tilde{F}_\delta \) from above is given by

\[
G_\delta(f, m) := \tilde{F}_\delta(f) + \nabla \tilde{F}_\delta(f)^T (m - f) + c_\delta \| m - f \|^2 / 2.
\]

\[
\square
\]

**Remark 1.** In Definition 1 we assume that \( F_\delta \) is strictly convex. This property is only required for notational convenience, because it guarantees uniqueness of the minimizer \( m_\delta^* \). A careful inspection of the proof of Theorem 1 shows, however, that convergence continues to hold if strict convexity is replaced with convexity. Only the assertion \( d(m_\delta^*, M^*) \to 0 \) has to be replaced by

\[
\sup_{x \in M_\delta^*} \inf_{y \in M^*} \| x - y \| \to 0,
\]

where \( M_\delta^* := \arg\min_{m \in C} F_\delta(m) \). An analogous modification holds for Theorem 2.

We close the section with the following result, which shows under additional regularity conditions on \( F_\delta \) and \( C \) geometric, or, more precisely, at least \( Q \)-linear convergence of the GIRLS algorithm (cf. Böhning & Lindsay, 1988, Theorem 4.1, for a related result).

**Theorem 5.** (Geometric convergence of the GIRLS algorithm).

Let \( F_\delta : \mathbb{R}^d \to \mathbb{R} \) be coercive and twice continuously differentiable with positive definite hessian matrix \( D^2 F(m_\delta^*) =: A \). Further let \( G_\delta : \mathbb{R}^d \times \mathbb{R}^d \) be a quadratic approximation of \( F_\delta \) from above with hessian matrix \( B(m_\delta^*) =: B \) as in (11). Then the GIRLS algorithm yields a sequence \( (m_\delta^{(k)})_{k=0}^\infty \) converging to \( m_\delta^* = \arg\min_C F_\delta \) such that

\[
\limsup_{k \to \infty} \frac{\| m_\delta^{(k+1)} - m_\delta^* \|_A}{\| m_\delta^{(k)} - m_\delta^* \|_A} \leq 1 - \lambda_{\min} (B^{-1} A) \in [0, 1).
\]

Here \( \| v \|_A := (v^T A v)^{1/2} \), and \( \lambda_{\min}(B^{-1} A) \in (0, 1] \) denotes the smallest eigenvalue of \( B^{-1} A \).

**Proof.** According to Theorem 2, \( \lim_{k \to \infty} m_\delta^{(k)} = m_\delta^* \). Since \( C = \mathbb{R}^d \), \( \nabla F_\delta(m_\delta^*) = 0 \) and

\[
\begin{align*}
m_\delta^{(k+1)} &= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \nabla F_\delta(m_\delta^{(k)}) \\
&= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \int_0^1 D^2 F_\delta \left( (1-t)m_\delta^* + tm_\delta^{(k)} \right) (m_\delta^{(k)} - m_\delta^*) \, dt \\
&= m_\delta^{(k)} - B^{-1} A (m_\delta^{(k)} - m_\delta^*) + o(\| m_\delta^{(k)} - m_\delta^* \|).
\end{align*}
\]

Thus

\[
\frac{\| m_\delta^{(k+1)} - m_\delta^* \|_A}{\| m_\delta^{(k)} - m_\delta^* \|_A} = \frac{\| (I - B^{-1} A)(m_\delta^{(k)} - m_\delta^*) \|_A}{\| m_\delta^{(k)} - m_\delta^* \|_A} + o(1),
\]

11
and for any vector \( v \in \mathbb{R}^d \),
\[
\frac{\| (I - B^{-1} A) v \|_A^2}{\| v \|_A^2} = \frac{v^\top (I - AB^{-1}) A (I - B^{-1} A) v}{v^\top A v} = \frac{w^\top A^{-1/2} (I - AB^{-1}) A (I - B^{-1} A) A^{-1/2} w}{\| w \|_2^2}
\]
(\text{with } w := A^{1/2} v)
\[
= \frac{w^\top C^2 w}{\| w \|_2^2}
\]
(\text{with } C := I - A^{1/2} B^{-1} A^{1/2})
\[
\leq \lambda_{\max}(C^2)
\]
\]
It follows from property (ii) of \( G_\delta \) in Definition 2 that \( B - A \) is nonnegative definite, which implies that \( \lambda_i(B^{-1} A) = \lambda_i(A^{1/2} B^{-1} A^{1/2}) \in (0, 1] \). This entails that \( C \) is nonnegative definite with \( \lambda_{\max}(C^2) = \lambda_{\max}(C)^2 = (1 - \lambda_{\min}(B^{-1} A))^2 \).

\[ \square \]

3.2 Proper choice of \( \delta \) and the number of iterations

In practical applications the points \( m_\delta^* \) are never calculated exactly. Instead after finitely many, say \( I(\delta) \), iterations of (5) the iteration is terminated and the regularization parameter \( \delta \) is decreased, e.g. replaced with \( \delta/2 \). An obvious question is how to choose the iteration numbers \( I(\delta) \). We found empirically in most cases that for a fixed parameter \( \delta > 0 \), the values \( F(m_\delta^{(k)}) \) are decreasing for \( k \leq k(\delta) \) and increasing in \( k \geq k(\delta) \) for some fixed \( k(\delta) \in \mathbb{N} \). Hence we may take
\[
I(\delta) := \min \left\{ k \in \mathbb{N}_0 : F(m_\delta^{(k+1)})/F(m_\delta^{(k)}) \geq 1 - \epsilon \right\} \cup \{ k_{\max} \}
\]
(12)
for a small constant \( \epsilon > 0 \) and a large maximal number \( k_{\max} \). In the examples discussed subsequently, we found that for \( \epsilon = 10^{-5} \) and \( k_{\max} = 100 \), the number \( I(\delta) \) was never larger than 30, which seems to compensate for the fact that the sequence \( m_\delta^{(k)} \) converges only geometrically. This is similar to numerical findings of an implementation of an algorithm by Lejeune & Sarda (1988, Section 5) for the median and various parametric regression models.

Having determined \( I(\delta) \) and \( m_\delta^{(I(\delta))} \) for one particular \( \delta > 0 \), we define \( m^{(0)}_{\delta/2} := m_\delta^{(I(\delta))} \) and repeat the same procedure with \( \delta/2 \) in place of \( \delta \), provided that \( I(\delta) > 0 \). We proceed as long as \( F \) is decreased, otherwise we terminate the algorithm.

4 Regularization and quadratic approximation for different types of regression problems

In the subsequent data examples the target functional \( F(m) \) is always of type (1) or (3), i.e.
\[
F(m) = \sum_{i=1}^n \rho(r_i(m)) + \lambda P(m)
\]
(13)
with \( \lambda \geq 0 \), where each residual \( r_i(m) \) is an affine linear functional of \( m \in \mathbb{R}^d \). Here each summand of \( F \) is regularized and approximated separately. We will start with an auxiliary result justifying the quadratic approximation (9).

**Lemma 1.** Let \( h : \mathbb{R} \to \mathbb{R} \) be even and twice differentiable such that \( h'' \) is non-negative and non-increasing on \([0, \infty)\). For \( r, z \in \mathbb{R} \) define

\[
g(r, z) := h(r) + \frac{(h'(r)/r)(z^2 - r^2)}{2},
\]

where \( h'(0)/0 := h''(0) \). Then \( g(r, z) \geq h(z) \) with equality if \( z = \pm r \).

**Proof.** One verifies easily that \( g(r, z) \) is even in both arguments with \( g(r, r) = h(r) \). Thus it suffices to show that \( g(r, z) \geq h(z) \) for any \( r, z \geq 0 \). Now,

\[
g(r, z) - h(z) = g(r, z) - h(r) - (h(z) - h(r)) = \frac{(h'(r)/r)(z^2 - r^2)}{2} - \frac{1}{r} \int_r^z (h'(t) - h'(r)) \, dt
\]

\[
= \frac{(h'(r)/r)(z - r)^2}{2} - \frac{1}{r} \int_r^z (h'(t) - h'(r)) \, dt
\]

\[
= \int_r^z \left( h(r, 0) - \tilde{h}(r, t) \right) (t - r) \, dt
\]

\[
= \int_{\max(r,z)}^{\min(r,z)} \left( h(r, 0) - \tilde{h}(r, t) \right) |t - r| \, dt, \tag{14}
\]

where \( \tilde{h}(r, t) := (h'(t) - h'(r))/(t - r) \) for \( t \neq r \), and \( \tilde{h}(r, r) := h''(r) \). One can deduce easily from \( h'' \) being non-increasing on \([0, \infty)\) that \( \tilde{h}(r, \cdot) \) has the same property. Thus the integrand of (14) is non-negative. \( \square \)

Let us first describe how to approximate \( \rho \) itself in three special cases. After this we will discuss several penalisations \( P \) in (13). Finally we comment on isotonic regression, an example with \( C \neq \mathbb{R}^d \).

**Quantile regression.** Let \( \rho(z) \) be given by (2). This may be rewritten as

\[
\rho(z) = |z| + (2p - 1)z.
\]

Hence we utilize the functions \( h_\delta \) and \( g_\delta \) from (7) and (9), which yields the regularization

\[
z \mapsto h_\delta(z) + (2p - 1)z
\]

and by means of Lemma 1 the quadratic approximation

\[
z \mapsto g_\delta(r, z) + (2p - 1)z = c_\delta(r) + h_\delta(r)^{-1}z^2/2 + (2p - 1)z
\]
of $z \mapsto \rho(z)$, where $c_\delta(r)$ is an irrelevant constant.

**$L^q$-regression.** Let $\rho(z) := |z|^q$ for some $q \in [1, \infty)$. If $1 \leq q < 2$, one may generalize definitions (7) and (9) immediately as follows:

$$
\begin{align*}
  h_\delta(z) &:= (z^2 + \delta)^{q/2}, \\
g_\delta(r, z) &:= h_\delta(r) + q(r^2 + \delta)^{1-q}(z^2 - r^2)/2 \\
&= c_\delta(r) + q(r^2 + \delta)^{1-q}z^2/\delta.
\end{align*}
$$

Again it follows from Lemma 1 that $g_\delta(r, z) \geq h_\delta(z)$ with equality for $z = \pm r$.

In case of $q > 2$, the second derivative of $z \mapsto |z|^q$ is increasing in $|z|$ and unbounded, hence Lemma 1 cannot be applied directly. To circumvent this problem, one could redefine

$$
h_\delta(z) := \begin{cases} 
  |z|^q & \text{if } |z| \leq \delta^{-1} \\
a_\delta + b_\delta|z| + q(q - 1)\delta^{-2-q}z^2/2 & \text{otherwise}
\end{cases}
$$

with constants $a_\delta, b_\delta$ such that $h_\delta$ is twice continuously differentiable, and then use the quadratic approximation

$$
g_\delta(r, z) := h_\delta(r) + h_\delta'(r)(z - r) + q(q - 1)\delta^{-2-q}(z - r)^2/2. \tag{15}
$$

**Logistic regression.** For data sets with a covariable $X$ and a dichotomous response $Y \in \{0, 1\}$, maximum likelihood estimation of $M(X) := \log [P(Y = 1 \mid X)/P(Y = 0 \mid X)]$ involves “residuals” $z = (1/2 - Y)M(X)$ and

$$
  \rho(z) := h(z) + z \quad \text{with} \quad h(z) := \log[e^z + e^{-z}].
$$

Note that $h$ satisfies the conditions of Lemma 1 with $h'(r) = \tanh(r)$ and $h''(r) = 1 - \tanh(r)^2$. Thus regularization is superfluous, while quadratic approximation is straightforward. In this case, the well known IRLS algorithm results (McCullagh & Nelder 1989).

**Roughness penalties.** Let us start with two particular examples for $P(m)$. For given real numbers $x_1 < x_2 < \cdots < x_d$ let $M$ be a function on $[x_1, x_d]$ and $m := (M(x_j))_{j=1}^d$. Then let

$$
  \text{TV}^{(0)}(m) := \sum_{j=1}^{d-1} |m_j - m_{j+1}|,
$$

$$
  \text{TV}^{(1)}(m) := \sum_{j=2}^{d-1} |\Delta_j m| \quad \text{with} \quad \Delta_j m := \frac{m_{j+1} - m_j}{x_{j+1} - x_j} - \frac{m_j - m_{j-1}}{x_j - x_{j-1}}.
$$

If $M$ is continuous and piecewise linear with knots in $\{x_1, \ldots, x_d\}$, then $\text{TV}^{(0)}(m)$ and $\text{TV}^{(1)}(m)$ are the total variation of $M$ and its first derivative, respectively. One could also think about
smoother functions $M$ and approximate the total variation of its second or higher order derivative by suitable divided differences of $m$.

Generally, let $P(m)$ be a sum of several functionals of the form

$$m \mapsto |v^\top m|$$

with a given vector $v \in \mathbb{R}^d \setminus \{0\}$. For instance, $\text{TV}^{(0)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i = j + 1, \\ 0 & \text{else} \end{cases}$$

for $1 \leq j < d$, while $\text{TV}^{(1)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} (x_j - x_{j-1})^{-1} & \text{if } i = j - 1, \\ - (x_j - x_{j-1})^{-1} - (x_{j+1} - x_j)^{-1} & \text{if } i = j, \\ (x_{j+1} - x_j)^{-1} & \text{if } i = j + 1, \\ 0 & \text{else} \end{cases}$$

for $1 < j < d$. Now an obvious strategy is to regularize $m \mapsto |v^\top m|$ by $m \mapsto h_\delta(v^\top m)$ and approximate this quadratically by

$$m \mapsto g_\delta(v^\top f, v^\top m) = c_\delta(v^\top f) + h_\delta(v^\top f)^{-1} \left( v^\top m \right)^2 / 2.$$

Often it is desirable to work with quadratic approximations $G(f, \cdot)$ whose Hessian matrix $B(f)$ is diagonal. For that purpose one can modify the quadratic term $Q(m) := (v^\top m)^2$ as follows:

$$Q(m) = \left( v^\top f \right)^2 + 2 f^\top vv^\top (m - f) + \left( v^\top (m - f) \right)^2 \leq \left( v^\top f \right)^2 + 2 f^\top vv^\top (m - f) + \|v\|^2 \sum_{i:v_i \neq 0} (m_i - f_i)^2 = c(v, f) - 2w(v, f)^\top m + \|v\|^2 \sum_{i:v_i \neq 0} m_i^2$$

for some irrelevant constant $c(v, f)$ and $w(v, f)_i := 1_{v_i \neq 0} \|v\|^2 f_i - v^\top f v_i$.

**Isotonic regression.** In some applications one seeks to minimize a functional such as (13) over all vectors in $\mathcal{C} := \{ m \in \mathbb{R}^d : m_1 \leq \cdots \leq m_d \}$. In the simplest case, $d = n$ and $\rho(r_i(m)) = (Y_i - m_i)^q$ for some $q \in [1, \infty]$, where $q = \infty$ corresponds to supremum norm of $Y - m$. For this special case it is well-known (see Barlow & Ubhaya 1971) that an explicit solution exists only for $q = 1, 2, \infty$. In general, via regularization and suitable quadratic approximation from above, each updating step of the GIRLS algorithm involves minimisation of

$$G_\delta(f, m) = C(f) + \sum_{i=1}^d w_i(f) (m_i - b_i(f))^2$$
over all vectors $m \in C$ with certain weights $w_i(f) > 0$, an irrelevant constant $C(f)$ and certain numbers $b_i(f)$. This minimisation problem can be solved explicitly by means of the PAVA (Robertson, Wright & Dykstra, 1988).

5 Numerical examples

In this section we discuss the numerical performance of the GIRLS algorithm in practical applications. Our first example is about quantile regression and shows that GIRLS outperforms a Newton type algorithm, at least in the first iterations. In the second example we consider a unimodal regression problem. For ordinary isotonic regression in one-dimensional settings, PAVA is known to be highly efficient (cf. e.g. Robertson et al. 1988) in the sense of producing an exact solution in $O(d)$ steps. Replacing the isotonicity constraint on $m$ with a penalty term $TV^{(0)}(m)$ leads to minimization problems which can be solved efficiently with the closely related “taut string algorithm” (cf. Davies and Kovac 2001). However, analogous problems in multidimensional regression settings, or using $TV^{(k)}(m)$ with $k \geq 1$, are computationally more involved. We mention e.g. Hinterberger et al. (2003) for a reformulation as a bilateral contact problem and a rather involved solution by a two level iterative method requiring a semi-smooth Newton method and a primal dual active set algorithm. GIRLS offers a different strategy: PAVA can be applied after approximating the target functional suitably by a quadratic program; hence we can also understand GIRLS as a linking device to apply efficient special purpose algorithms for quadratic programs such as PAVA to corresponding non-quadratic and/or constrained problems.

Example 1. (estimation of smooth quantile functions)

We applied the GIRLS algorithm and a competitor to a dataset containing the income ($X$) and the expenditure for food ($Y$) in the year 1973 for 7125 households in Great Britain (Family Expenditure Survey 1968–1983). This dataset has also been analyzed by Härdle and Marron (1991). In order to enhance the visual quality we reduced these data to a random subset of size $n = 2000$. Moreover, since the empirical distributions of both variables are strongly skewed to the right, we plotted only the 1936 pairs $(X_i, Y_i)$ in the range $[0, 10^5] \times [0, 3 \cdot 10^4]$. Figure 1 shows the data together with estimated $p$–quantile curves $M_p$ for $p = 0.1, 0.25, 0.5, 0.75, 0.9$. Here we used the functional

$$F(m) = \sum_{i=1}^{n} \rho(Y_i - m_{C(i)}) + \lambda TV^{(1)}(m)$$

on $\mathbb{R}^d$ with $\rho$ given by (2), where $x_1 < x_2 < \cdots < x_d$ are the $d = 1945$ different $X$–values in the
Figure 1: Quantile curves for food expenditure data.

Figure 2: Mean performance of GIRLS and a Newton-Raphson-algorithm. The dashed and dotted curves show the mean value of the function \( F(m) \) in dependence of the iteration number \( m \) for the Newton-Raphson algorithm and GIRLS, respectively, and the solid curve the stepwise rate of improvement \( \frac{F_{\text{GIRLS}}(m+1) - F_{\text{GIRLS}}(m)}{F_{\text{NEWTON}}(m+1) - F_{\text{NEWTON}}(m)} \) of \( F(m) \) by the two methods. Here, a value larger than 1 indicates that in the \( m \)-th iteration \( F(m) \) decreased more for GIRLS than for the Newton-Raphson-algorithm.
sample, \( m_j = M_p(x_j), \) \( X_i = x_{C(i)} \) and \( F \) was regularized as discussed in the last section. The tuning parameter \( \lambda \) was chosen to be \( 2 \cdot 10^6 \) by visual inspection.

We now turn to the numerical performance of the GIRLS algorithm and compare it to a Newton-Raphson algorithm (robustified with a standard step-size correction as in Dümbgen et al. 2006). In the first step we compare the performance of the two algorithms in the particular setting of our data example; the second step will consist of a small simulation study with artificial data. All computations where performed with Matlab on a 1.86Ghz Pentium-processor with 1GB Ram. As starting value of the iterations we used the polynomial regression \( p \)-quantile of order 1, and the tuning parameter \( \delta \) was selected in a data-driven way from this starting value as the smaller of the median of its absolute residuals, and the median of its first order differences, in both cases divided by 1000. In the first step, we compared the computational efficiency of the GIRLS and the Newton-Raphson algorithm applied to quantile regression with the family expenditure data. To this end we recorded the computing times and number of iterations required for determination of the 25%-quantile curve by GIRLS and the Newton algorithm, where we stopped the iterations as soon as the relative improvement of the function \( F(m) \) between two subsequent iterations fell below a threshold parameter \( \varepsilon = 10^{-12} \). Whereas GIRLS required 5.0s CPU time and 59 iterations to find the solution, the Newton-Raphson algorithm turned out to be significantly more expensive with 46.4s CPU time and 425 iterations. We also performed the same computations for a wide range of threshold parameters \( \varepsilon = 10^{-6} \ldots 10^{-24} \), without significant change in the relative computational expense of the two methods.

In the second step of our analysis we performed 100 simulations of the quantile regression problem with (artificial) regression data from the model

\[
Y_i = \sin \left( X_i \cdot \frac{\pi}{2} \right) + \varepsilon_i, \quad i = 1, \ldots, 1000.
\]

Here, \( X_i \sim U[0, 1] \) are uniformly distributed, independent design variables, and \( \varepsilon_i \sim N(0, 0.01) \) i.i.d. noise terms. We used both GIRLS and the Newton-Raphson method to estimate the 25%-quantile curve of the data by determining the minimizer of the function \( F(m) \). From visual inspection of a pilot simulation, we chose the tuning parameter \( \lambda = 2 \cdot 10^4 \) for the subsequent simulations.

Figure 2 compares the mean performance of the GIRLS algorithm with the performance of the Newton-Raphson-method with step-size correction. The two curves in the top show the value of the function \( F(m) \) in dependence of the iteration number \( m \), and the bottom curve represents the
mean improvement (in the simulations) of the solution, measured by the decrease of the target function $F(m)$. From the figure, we conclude that the first steps of GIRLS are significantly more efficient than for the Newton-Raphson method, whereas the latter catches up after approximately the 8th iteration.

In summary, in the computations with the penalized quantile regression problem, GIRLS outperformed the Newton-Raphson method for the family expenditure data. Moreover, in our simulation we found GIRLS to improve the solution much faster than the Newton-Raphson method in the first $\approx 8$ iterations. Only if the initial value of the Newton algorithm has been chosen very close to the true minimizer $M^*$, we found the performance of the Newton-algorithm is superior. Hence, for practical purposes it seems be advisable to combine both algorithms such that GIRLS will be used at least as an initial algorithm which efficiently provides a good initial value for a subsequently performed Newton type algorithm. Moreover, if both GIRLS and the Newton-Raphson-method are available, it may be useful to compute both a GIRLS and a Newton step, with subsequent selection of the better one.

**Example 2.** (GIRLS as a linking device to apply efficient quadratic algorithms to constrained non-quadratic problems)

In our second example we illustrate the flexibility of the GIRLS algorithm, where we combine an unimodality constraint together with TV penalization. Ordinary isotonic regression involves the
solution of a weighted least squares problem and efficient algorithms such as in particular PAVA are available. If we instead use TV penalization, the problem isn’t described as a quadratic programming problem anymore, and hence PAVA would be not applicable. By replacing the problem to be solved by a sequence of quadratic problems, GIRLS makes it possible again to apply PAVA for isotonic regression with TV penalization.

A simple application of this idea is the following. In order to reconstruct the spatial luminosity distribution of the Milky Way, the surface brightness of the Milky Way at the sky was measured on an equidistant grid of angles running in horizontal direction from \(-89.25^\circ\) to \(89.25^\circ\) and in vertical direction from \(-29.25^\circ\) to \(29.25^\circ\) in steps of \(1.5^\circ\), respectively. The data set is part of the DIRBE experiment on board the COBE satellite (cf. Spergel et al. 1996 for a comprehensive description of the data, proper calibration and dust correction). Bissantz & Munk (2001) fitted a parametric model of the spatial luminosity distribution of the Milky Way to this data and a map of the resulting squared residuals is displayed there, which suggests that the variability of the data increases in the outer regions of the Milky Way. Moreover, due to this pronounced heteroscedasticity of the data, the reliability of any conclusions about physical properties (such as scale lengths) of the Milky Way from the parameters of a (parametric) model would be improved substantially if the variance surface of the data is known and used in the fitting process.

In the following we will consider estimation of the variance surface from the squared residuals by a penalized least squares fit with an unimodality constraint. The assumption of unimodality is suggested by physical reasoning. Such an estimate of the variance surface can then be used in a further investigation of the surface brightness data.

In more detail, the data considered here, denoted as \(Y_{ij}, i = 1, \ldots, 120, j = 1, \ldots, 40\), are the squared residuals of the the parametric fit in Bissantz & Munk (2001) to the observed surface brightness data.

First, we consider a least squares fit \(m = (m_{ij})_{i,j}\) to the data \(Y = (Y_{ij})_{i,j}\). Namely, we want to minimize the sum \(F(m)\) of

\[
\|Y - m\|^2 = \sum_{ij} (Y_{ij} - m_{ij})^2 ,
\]

and the total variation penalty \(\lambda \text{TV}(m)\), where \(\lambda > 0\) and

\[
\text{TV}(m) := \sum_{i=1}^{119} \sum_{j=1}^{40} |m_{i+1,j} - m_{ij}| + \sum_{i=1}^{120} \sum_{j=1}^{39} |m_{i,j+1} - m_{ij}| .
\]

For the regularization \(F_\delta\) and quadratic approximation \(G_\delta(f,\cdot)\), the least squares term is kept unchanged, while each summand \(|m_{(a)} - m_{(b)}|\) of \(\text{TV}(m)\) is treated as described in section 4:
First we regularize it by $h_\delta(m(a) - m(b))$. Then as a first quadratic approximation we use

$$g_\delta(f(a) - f(b), m(a) - m(b)) = C_\delta(f(a), f(b)) + \frac{(m(a) - m(b))^2}{2h_\delta(f(a) - f(b))}.$$ 

Alternatively, one might replace the enumerator $(m(a) - m(b))^2$ with

$$2\left( m(a) - \frac{f(a) + f(b)}{2} \right)^2 + 2\left( m(b) - \frac{f(a) + f(b)}{2} \right)^2,$$

which is never less than $(m(a) - m(b))^2$ with equality for $m = f$. The advantages of this latter quadratic approximation are computational simplicity and feasibility of isotonic least squares algorithms when incorporating additional monotonicity constraints. Moreover, we impose in addition unimodality in vertical direction with minimum mode at the 0-line. Hence, in each step for each vertical half line the isotonic regression has to be calculated by means of some standard algorithm calculating the isotonic weighted least squares fits like PAVA (Robertson, et al., 1988). This is feasible with our second quadratic approximation. Figure 3 shows the resulting surface. Note that smoothness along the vertical direction of the estimated variance surface is already guaranteed by the unimodality constraint, and does not rely on a proper selection of the smoothing parameter $\lambda$. This relaxes significantly the requirements for a data-driven estimation of the variance surface, and, subsequently, of the spatial luminosity distribution of the Milky Way.

ACKNOWLEDGEMENTS

The authors are thankful to three anonymous referees for their constructive criticism that helped to improve this note. Moreover, they would like to thank G. Jongbloed and O. Scherzer for helpful comments. The first author is indebted to Y. Vardi for interesting discussions on a previous version of this paper. Y. Vardi passed away unexpectedly in Jan. 13th 2005 and we would like to dedicate this work to his memory. Parts of this paper were written while L. Dümbgen was visiting Göttingen University as lecturer within the PhD Program ’Applied Statistics and Empirical Methods’, and while N. Bissantz was visiting the University of Bern. Financial support of the Graduiertenkolleg 1023 ”Identification in Mathematical Models”, SFB 475 ”Komplexitätsreduktion in multivariaten Datenstrukturen” and the DAAD is gratefully acknowledged.
References


