SIAM J. NUMER. ANAL.
Vol. 40, No. 3, pp. 965–994


# MINIMIZERS OF COST-FUNCTIONS INVOLVING NONSMOOTH DATA-FIDELITY TERMS. APPLICATION TO THE PROCESSING OF OUTLIERS*

MILA NIKOLOVA†

**Abstract.** We present a theoretical study of the recovery of an unknown vector $x \in \mathbb{R}^p$ (such as a signal or an image) from noisy data $y \in \mathbb{R}^q$ by minimizing with respect to $x$ a regularized cost-function $\mathcal{F}(x, y) = \Psi(x, y) + \alpha\Phi(x)$, where $\Psi$ is a data-fidelity term, $\Phi$ is a smooth regularization term, and $\alpha > 0$ is a parameter. Typically, $\Psi(x, y) = \|Ax - y\|^2$, where $A$ is a linear operator. The data-fidelity terms $\Psi$ involved in regularized cost-functions are generally smooth functions; only a few papers make an exception to this and they consider restricted situations. Nonsmooth data-fidelity terms are avoided in image processing. In spite of this, we consider both smooth and nonsmooth data-fidelity terms. Our goal is to capture essential features exhibited by the local minimizers of regularized cost-functions in relation to the smoothness of the data-fidelity term.

In order to fix the context of our study, we consider $\Psi(x, y) = \sum_i \psi(a_i^T x - y_i)$, where $a_i^T$ are the rows of $A$ and $\psi$ is $\mathcal{C}^m$ on $\mathbb{R} \setminus \{0\}$. We show that if $\psi'(0^-) < \psi'(0^+)$, then typical data $y$ give rise to local minimizers $\hat{x}$ of $\mathcal{F}(., y)$ which fit exactly a certain number of the data entries: there is a possibly large set $\hat{h}$ of indexes such that $a_i^T \hat{x} = y_i$ for every $i \in \hat{h}$. In contrast, if $\psi$ is smooth on $\mathbb{R}$, for almost every $y$, the local minimizers of $\mathcal{F}(., y)$ do not fit any entry of $y$. Thus, the possibility that a local minimizer fits some data entries is due to the nonsmoothness of the data-fidelity term. This is a strong mathematical property which is useful in practice. By way of application, we construct a cost-function allowing aberrant data (outliers) to be detected and to be selectively smoothed. Our numerical experiments advocate the use of nonsmooth data-fidelity terms in regularized cost-functions for special purposes in image and signal processing.

**Key words.** inverse problems, MAP estimation, nonsmooth analysis, perturbation analysis, proximal analysis, reconstruction, regularization, stabilization, outliers, total variation, variational methods

**AMS subject classifications.** 49N45, 62H12, 49J52, 49N60, 94A12, 94A08, 35A15, 68U10, 26B10

**PII.** S0036142901389165

## 1. Introduction.

We consider the general problem where a sought vector (e.g., an image or a signal) $\hat{x} \in \mathbb{R}^p$ is obtained from noisy data $y \in \mathbb{R}^q$ by minimizing a regularized cost-function $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the form

$$\mathcal{F}(x, y) = \Psi(x, y) + \alpha\Phi(x), \tag{1}$$

where typically $\Psi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is a data-fidelity term and $\Phi : \mathbb{R}^p \to \mathbb{R}$ is a regularization term, with $\alpha > 0$ a parameter. In many applications, the relation between $x$ and $y$ is modeled by $y_i = a_i^T x + n_i$ for $i = 1, \ldots, q$, where $a_i^T : \mathbb{R}^p \to \mathbb{R}$ are linear operators and $n_i$ accounts for perturbations. We focus on such situations and assume that $a_i^T$, $i = 1, \ldots, q$, are known and non-null. The relevant data-fidelity term assumes the form

$$\Psi(x, y) = \sum_{i=1}^{q} \psi_i(a_i^T x - y_i), \tag{2}$$

*Received by the editors May 9, 2001; accepted for publication (in revised form) December 28, 2001; published electronically August 8, 2002.
http://www.siam.org/journals/sinum/40-3/38916.html
†CNRS URA820–ENST Dpt. TSI, ENST, 46 rue Barrault, 75013 Paris, France (nikolova@tsi.enst.fr).

where $\psi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, q$, are continuous functions which decrease on $(-\infty, 0]$ and increase on $[0, +\infty)$. Usually, $\psi_i = \psi$ for all $i$. One usual choice is $\psi(t) = |t|^\rho$, for $\rho > 0$, which yields [31, 4]

$$\Psi(x, y) = \sum_{i=1}^{q} |a_i^T x - y_i|^\rho. \tag{3}$$

Let $A \in \mathbb{R}^{q \times p}$ be the matrix whose rows are $a_i^T$ for $i = 1, \ldots, q$. This matrix can be ill-posed, or singular, or invertible. Most often, $\Psi(x, y) = \|Ax - y\|^2$, that is, $\psi(t) = t^2$. Such data-fidelity terms are currently used in denoising, in deblurring, and in numerous inverse problems [37, 35, 13, 33, 1, 14, 38]. In a statistical framework, $\Psi$ accounts for both the distortion and the noise intervening between the original $x$ and the device recording the data $y$. The above quadratic form of $\Psi$ corresponds to white Gaussian noise $\{n_i\}$. Recall that many papers are dedicated to the minimization of $\Psi(., y)$ *alone* and of the form (3), i.e., $\mathcal{F} = \Psi$, mainly for $\psi(t) = t^2$ [22], in some cases for $\psi(t) = |t|$ [8], but functions $\psi(t) = |t|^\rho$ for different values for $\rho$ in the range $(0, \infty]$ also have been considered [31, 30]. Specific data-fidelity terms arise in applications such as emission and transmission computed tomography, X-ray radiography, eddy-currents evaluation, and many others [23, 20, 34, 10]. In general, for every $y$, the data-fidelity term $\Psi(., y)$ is a function which is smooth and usually convex. The introduction of nonsmooth data-fidelity terms in regularized cost-functions (1) remains very unusual. Only a few papers make an exception to this; we cite [2, 3], where $\Psi$ corresponds to $\psi(t) = |t|$ and $a_i^T x = x_i$ for all $i$. Nonsmooth data-fidelity terms $\Psi$ are avoided in image processing, for instance. In spite of this, we analyze the effects produced by both smooth and nonsmooth data-fidelity terms $\Psi$. In the latter case we suppose that $\{\psi_i\}$ *are any functions which are* $\mathcal{C}^m$*-smooth on* $\mathbb{R} \setminus \{0\}$*,* $m \geq 2$*, whereas at zero they admit finite side derivatives which satisfy* $\psi_i'(0^-) < \psi_i'(0^+)$.

The regularization term $\Phi$ usually takes the form

$$\Phi(x) = \sum_{i=1}^{r} \varphi(\|G_i^T x\|), \tag{4}$$

where $G_i^T : \mathbb{R}^p \to \mathbb{R}^s$ for $s \in \mathbb{N}^*$ are linear operators, e.g., operators yielding the differences between neighboring samples; $\|.\|$ stands for a norm on $\mathbb{R}^s$; and $\varphi : \mathbb{R} \to \mathbb{R}$ is a potential function. In a Bayesian estimation framework, $\Phi$ is the prior energy of the unknown $x$ modeled using a Markov random field [6, 17, 24]. Several customarily used potential functions $\varphi$ are [20, 29, 21, 33, 9, 7, 39, 36]

$$\begin{array}{ll}
\mathrm{L}^\nu & \varphi(t) = |t|^\nu,\ 1 \leq \nu \leq 2, \\
\text{Lorentzian} & \varphi(t) = \nu t^2/(1 + \nu t^2), \\
\text{Concave} & \varphi(t) = \nu|t|/(1 + \nu|t|), \\
\text{Gaussian} & \varphi(t) = 1 - \exp(-\nu t^2), \\
\text{Huber} & \varphi(t) = t^2 \text{ if } |t| \leq \nu, \quad \varphi(t) = \nu(\nu + 2|t - \nu|) \text{ if } |t| > \nu, \\
\text{Mean-field} & \varphi(t) = -\log\big(\exp(-\nu t^2) + 1\big),
\end{array} \tag{5}$$

where $\nu > 0$ is a parameter. Being convex and differentiable, the function $\mathrm{L}^\nu$ for $1 < \nu \leq 2$ is preferred in many applications requiring intensive computation [9, 10]. In our paper, $\Phi$ *in* (1) *is any* $\mathcal{C}^m$*-smooth function*, with $m \geq 2$.

The visual aspect of a minimizer of a cost-function is determined on the one hand by the data and on the other hand by the shape of the cost-function. Our goal is to

capture essential features expressed by the local minimizers of cost-functions of the form (1)–(2) in relation to the smoothness of the data-fidelity term $\Psi$. Note that all our results hold for local minimizers, and hence for global minimizers as well, so we systematically speak of local minimizers. There is a striking distinction in the behavior of the local minimizers relevant to smooth and nonsmooth data-fidelity terms. It concerns the possibility of fitting *exactly* a certain number of the data entries, i.e., that for $y$ given, a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ satisfies $a_i^T\hat{x} = y_i$ for some, or even for many, indexes $i$ (see section 2). Intuitively, one is unlikely to obtain such minimizers, especially when data are noisy. *Our main result states that for $\mathcal{F}$ of the form* (1)–(2)*, with $\Psi$ nonsmooth as specified, typical data $y$ give rise to local minimizers $\hat{x}$ which fit a certain number of the data entries; i.e., there is a nonempty set $\hat{h}$ of indexes such that $a_i^T\hat{x} = y_i$ for every $i \in \hat{h}$* (see sections 3 and 4). This effect is due to the nondifferentiability of $\Psi$ since it cannot occur when $\mathcal{F}$ is differentiable (see section 5). The obtained result is a strong mathematical property which can be used in different ways. Based on it, we construct a cost-function allowing aberrant data (outliers) to be detected and to be selectively smoothed from signals, or from images, or from noisy data, while preserving efficiently all the nonaberrant entries (see section 7). This is illustrated using numerical experiments.

Readers may associate cost-functions where $\Psi$ is nonsmooth (e.g., $\psi(t) = |t|$) with cost-functions where $\Psi$ is smooth and $\Phi$ is nonsmooth, e.g., $\Psi(x,y) = \|Ax - y\|^2$ and $\varphi(t) = |t|$ in (4), as in total-variation methods [33, 1, 14, 12]. Since the latter methods arouse an increasing interest in the area of image and signal restoration, we compare in section 6 nonsmooth regularization to the cost-functions considered in this paper. To this end, we use some previous results [26, 27] and illustrate the strikingly different visual effects they produce (see section 7).

**2. The problem of an exact fit for some data entries.** We shall use the symbol $\|.\|$ to denote the $\ell_2$-norm of vectors. Next, we denote by $\mathbb{N}^*$ the positive integers and $\mathbb{R}_+ = \{t \in \mathbb{R} : t \geq 0\}$. The letter $S$ will systematically denote the centered, unit sphere in $\mathbb{R}^n$, say $S := \{x \in \mathbb{R}^n : \|x\| = 1\}$, for whatever dimension $n$ is appropriate in the context. For $x \in \mathbb{R}^n$ and $\rho > 0$, we put $B(x,\rho) := \{x' \in \mathbb{R}^n : \|x' - x\| < \rho\}$. For any $i = 1, \ldots, n$ the letter $e_i$ represents the $i$th vector of the canonical basis of $\mathbb{R}^n$ (i.e., $e_i = e_i[i] = 1$ and $e_i[j] = 0$ for all $j \neq i$). The closure of a set $N$ will be denoted $\overline{N}$. For a subspace $T$; its orthogonal complement is denoted $T^\perp$. If a function $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ depends on two variables, its $k$th differential with respect to the $j$th variable is denoted $D_j^k f$. The notation $f \in \mathcal{C}^m(N)$ means that the function $f$ is $\mathcal{C}^m$-smooth on the set $N$. For a discrete, finite set $h \subset \{1, \ldots, n\}$, with $n \in \mathbb{N}^*$, the symbol $\#h$ is the cardinality of $h$ and $h^c$ is the complementary of $h$. Next we introduce a set-valued function which is constantly evoked in what follows.

DEFINITION 1. *Let $\mathcal{H}$ be the function which for every $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ yields the following set:*

$$(x,y) \quad \to \quad \mathcal{H}(x,y) = \left\{ i \in \{1, \ldots, q\} : a_i^T x = y_i \right\}. \tag{6}$$

Given $y$ and a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$, the set of all data entries which are fitted exactly by $\hat{x}$ reads $\hat{h} := \mathcal{H}(\hat{x},y)$. Furthermore, with every $h \subseteq \{1, \ldots, q\}$ we associate the following sets:

$$(h,y) \to \Theta_h(y) := \{x \in \mathbb{R}^p : \ a_i^T x = y_i \ \forall i \in h \text{ and } a_i^T x \neq y_i \ \forall i \in h^c\}, \tag{7}$$

$$h \quad \to \quad T_h := \{u \in \mathbb{R}^p : \ a_i^T u = 0 \ \ \forall\, i \in h\}, \tag{8}$$

$$h \quad \to \quad M_h := \{(x,y) \in \mathbb{R}^p \times \mathbb{R}^q : \ a_i^T x = y_i \ \forall i \in h \text{ and } a_i^T x \neq y_i \ \forall i \in h^c\}. \tag{9}$$

Note that for every $y$ and $h \neq \emptyset$, the sets $\Theta_h(y)$ and $M_h$ are composed of a finite number of connected components, whereas their closures $\overline{\Theta_h(y)}$ and $\overline{M_h}$, respectively, are affine subspaces. The family of all $\Theta_h$, when $h$ ranges over all the subsets of $\{1,\ldots,q\}$, forms a partition of $\mathbb{R}^p$. Observe that for $y \in \mathbb{R}^q$ fixed, $\{x \in \mathbb{R}^p : (x,y) \in M_h\} = \Theta_h(y)$. Notice also the equivalences

$$\mathcal{H}(x',y') = h \;\Leftrightarrow\; x' \in \Theta_h(y') \;\Leftrightarrow\; (x',y') \in M_h. \tag{10}$$

The theory in this paper is developed by analyzing how the local minimizers of every $\mathcal{F}(.,y)$ behave under small variations of the data $y$. We thus consider local minimizer functions.

Definition 2. *Let $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ and $N \subseteq \mathbb{R}^q$. The family $f(.,N) := \{f(.,y) : y \in N\}$ is said to admit a local minimizer function $\mathcal{X} : N \to \mathbb{R}^p$ if for any $y \in N$ the function $f(.,y)$ has a strict local minimum at $\mathcal{X}(y)$.*

The next lemma addresses local minimizer functions relevant to smooth cost-functions.

Lemma 1. *Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q$ be a $\mathcal{C}^m$-function with $m \geq 2$. For $y \in \mathbb{R}^q$, assume that $\hat{x} \in \mathbb{R}^p$ is such that $D_1\mathcal{F}(\hat{x},y) = 0$, and $D_1^2\mathcal{F}(\hat{x},y)$ is positive definite.*

*Then there exists a neighborhood $N \subset \mathbb{R}^q$ containing $y$ and a $\mathcal{C}^{m-1}$-function $\mathcal{X} : N \to \mathbb{R}^p$ such that for every $y' \in N$ we have $D_1\mathcal{F}(\mathcal{X}(y'),y') = 0$, and $D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is positive definite. In particular, $\hat{x} = \mathcal{X}(y)$.*

Equivalently, $\mathcal{X} : N \to \mathbb{R}^p$ is a local minimizer function relevant to $\mathcal{F}(.,N)$ such that $D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is positive definite for every $y' \in N$.

*Proof.* Being a local minimizer of $\mathcal{F}(.,y)$, $\hat{x}$ satisfies $D_1\mathcal{F}(\hat{x},y) = 0$. We focus on the equation $D_1\mathcal{F}(x',y') = 0$ in the vicinity of $(\hat{x},y)$ and notice that $D_1^2\mathcal{F}(\hat{x},y)$ determines an isomorphism from $\mathbb{R}^p$ to itself. From the implicit functions theorem [5], there exist $\rho_1 > 0$ and a unique $\mathcal{C}^{m-1}$-function $\mathcal{X} : B(y,\rho_1) \to \mathbb{R}^p$ such that $D_1\mathcal{F}\left(\mathcal{X}(y'),y'\right) = 0$ for all $y' \in B(y,\rho_1)$. Furthermore, since $y' \to \det D_1^2\mathcal{F}(\mathcal{X}(y'),y')$ is continuous and $\det D_1^2\mathcal{F}(\hat{x},y) > 0$, there is $\rho_2 \in (0,\rho_1]$ such that $\det D_1^2\mathcal{F}(\mathcal{X}(y'),y') > 0$ for all $y' \in B(y,\rho_2)$. □

*Remark* 1 (on the conditions required in Lemma 1). The minimizers of $\mathcal{C}^m$-functions of the form

$$\mathcal{F}(x,y) = \|Ax - y\|^2 + \alpha\Phi(x)$$

are extensively studied in [16]. It is shown there that if $\mathrm{rank}A = p$, and under some assumptions ensuring that $\mathcal{F}(.,y)$ admits local minimizers for every $y \in \mathbb{R}^q$, the data domain $\mathbb{R}^q$ contains a subset $N$ whose interior is dense in $\mathbb{R}^q$ such that for every $y \in N$, then every local minimizer $\hat{x}$ of the corresponding $\mathcal{F}(.,y)$ is strict and $D_1^2\mathcal{F}(\hat{x},y)$ is positive definite. Reciprocally, all data leading to minimizers at which the conditions of Lemma 1 fail belong to a closed negligible subset of $\mathbb{R}^q$: the chance of acquiring data placed in such subsets is null.

The central question of this paper is how the shape of a cost-function $\mathcal{F}$ favors, or inhibits, the possibility that a local minimizer $\hat{x}$ of $\mathcal{F}(.,y)$, for $y \in \mathbb{R}^q$, fits a certain number of the entries of this same $y$, i.e., that the set $\hat{h} := \mathcal{H}(\hat{x},y)$ is nonempty. It will appear that this possibility is closely related to the smoothness of $\Psi$. We recall some facts about nonsmooth functions [32].

Definition 3. *Let $E_0 \subseteq \mathbb{R}^p$ be an affine subspace and $E$ be the relevant vector space. Consider a function $f : E_0 \to \mathbb{R}$, and let $x \in E_0$ and $u \in E$. The function $f$ admits a one-sided derivative at $x$ in the direction of $u \neq 0$, denoted by $\delta g(x)(u)$, if*

*the following (possibly infinite) limit exists:*

$$\delta f(x)(u) := \lim_{t \downarrow 0} \frac{f(x+tu) - f(x)}{t}.$$

*If* $u = 0$, *put* $\delta f(x)(0) = 0$.

The downward pointing arrow above means that $t \in \mathbb{R}_+$ converges to zero by positive values. If $f$ is differentiable at $x$, then $\delta f(x)(u) = Df(x).u$. If $f : \mathbb{R} \to \mathbb{R}$, we have $\delta f(x)(1) = f'(x^+)$. The left derivative of $f$ at $x$ for $u$ is $-\delta f(x)(-u)$. In the following, $\delta_1 \mathcal{F}$ will address one-sided derivatives of $\mathcal{F}$ with respect to its first argument.

**3. Cost-functions with nonsmooth data-fidelity terms.** Here and in section 4 we focus on cost-functions which read

$$\mathcal{F}(x, y) = \Psi(x, y) + \alpha \Phi(x, y), \tag{11}$$

$$\Psi(x, y) = \sum_{i=1}^{q} \psi(a_i^T x - y_i), \tag{12}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is $\mathcal{C}^m$ on $\mathbb{R} \setminus \{0\}$, with $m \geq 2$, whereas at zero it admits finite side derivatives satisfying $\psi'(0^-) < \psi'(0^+)$. The term $\Phi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is any $\mathcal{C}^m$-function. This formulation allows us to address data-fidelity terms composed of a nonsmooth function $\Psi$ and of a smooth function $\tilde{\Psi}$, since we can write $\Phi(x, y) = \tilde{\Psi}(x, y) + \tilde{\Phi}(x)$ with $\tilde{\Phi}$ a regularization term. For example, we can have $\Phi(x, y) = \sum_i \big(\phi_i(B_i^T x - y_{q_i}) + \varphi_i(G_i^T x)\big)$, where $\phi_i : \mathbb{R}^{q_i} \to \mathbb{R}$ and $\varphi_i : \mathbb{R}^{p_i} \to \mathbb{R}$ are $\mathcal{C}^m$-functions, $y_{q_i} \in \mathbb{R}^{q_i}$ are data, and $B_i^T \in \mathbb{R}^{q_i \times p}$ and $G_i^T \in \mathbb{R}^{p_i \times p}$, with $p_i \in \mathbb{N}^*$ and $q_i \in \mathbb{N}^*$.

*Remark* 2. The results presented in sections 3 and 4 are developed for $\Psi$ of the form (12), that is, $\psi_i = \psi$ for all $i$, but we should emphasize that they remain true for $\Psi$ of the form (2) provided that all $\psi_i$, for $i = 1, \ldots, q$, have finite side derivatives at zero satisfying $\psi_i'(0^-) < \psi_i'(0^+)$. The proofs are straightforward to extend to this situation but at the expense of complicated notation which may cloud the presentation.

We start by providing a sufficient condition for a strict local minimum.

PROPOSITION 1. *For* $y \in \mathbb{R}^q$, *let* $\mathcal{F}(., y) : \mathbb{R}^p \to \mathbb{R}$ *be of the form* (11)–(12), *where* $\Phi \in \mathcal{C}^m(\mathbb{R}^p \times \mathbb{R}^q)$ *for* $m \geq 1$ *and* $\psi \in \mathcal{C}^m(\mathbb{R} \setminus \{0\})$ *satisfies* $-\infty < \psi'(0^-) < \psi'(0^+) < +\infty$. *Let* $\hat{x} \in \mathbb{R}^p$ *be such that*

1. *the restricted function* $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}}(., y) : \overline{\Theta_{\hat{h}}(y)} \to \mathbb{R}$ *reaches a strict local minimum at* $\hat{x}$,
2. $\delta_1 \mathcal{F}(\hat{x}, y)(u) > 0$ *for all* $u \in T_{\hat{h}}^{\perp} \cap S$,

*where* $\hat{h} := \mathcal{H}(\hat{x}, y)$, $\Theta_{\hat{h}}(y)$, *and* $T_{\hat{h}}$ *are determined according to* (6), (7), *and* (8), *respectively.*

*Then* $\mathcal{F}(., y)$ *reaches a strict local minimum at* $\hat{x}$.

*Proof.* The result is a tautology if $\hat{h} = \emptyset$ since then $\Theta_{\hat{h}}(y) = \mathbb{R}^p$. So consider that $\hat{h}$ is nonempty. First of all, we put $\mathcal{F}$ into a more convenient form. Define

$$\tilde{\psi}(t) := \psi(t) - \frac{t}{2}\big(\psi'(0^-) + \psi'(0^+)\big) - \psi(0). \tag{13}$$

Now we have

$$\tilde{\psi}'(0^+) = -\tilde{\psi}'(0^-) > 0 \quad \text{and} \quad \tilde{\psi}(0) = 0, \tag{14}$$

which will allow important simplifications. By means of $\tilde{\psi}$, the cost-function $\mathcal{F}$ assumes the form

$$\mathcal{F}(x,y) = \tilde{\Psi}(x,y) + \tilde{\Phi}(x,y), \tag{15}$$

$$\text{where } \tilde{\Psi}(x,y) = \sum_{i=1}^{q} \tilde{\psi}(a_i^T x - y_i)$$

$$\text{and } \tilde{\Phi}(x,y) = \sum_{i=1}^{q} \frac{\psi'(0^-) + \psi'(0^+)}{2}(a_i^T x - y_i) + q\psi(0) + \alpha\Phi(x,y).$$

Both $\tilde{\Psi}$ and $\tilde{\Phi}$ satisfy the assumptions about $\Psi$ and $\Phi$, respectively. Henceforth, we deal with the formulation of $\mathcal{F}$ given in (15). For notational convenience, we systematically write $\psi$ for $\tilde{\psi}$, $\Psi$ for $\tilde{\Psi}$, and $\Phi$ for $\tilde{\Phi}$.

Let us consider the altitude increment of $\mathcal{F}(.,y)$ at $\hat{x}$ in the direction of an arbitrary $u \in S$,

$$\mathcal{F}(\hat{x} + tu, y) - \mathcal{F}(\hat{x}, y) \quad \text{for } t \in \mathbb{R}_+.$$

In order to avoid misunderstandings, $u_0$ will denote a vector of $T_{\hat{h}}$ and $u_{\perp}$ a vector of $T_{\hat{h}}^{\perp}$. Using the fact that every $u \in S$ has a unique decomposition into

$$u = u_0 + u_{\perp} \text{ with } u_0 \in T_{\hat{h}} \cap \overline{B(0,1)} \text{ and } u_{\perp} \in T_{\hat{h}}^{\perp} \cap \overline{B(0,1)}, \tag{16}$$

we decompose the altitude increment of $\mathcal{F}(.,y)$ accordingly:

$$\mathcal{F}(\hat{x} + tu, y) - \mathcal{F}(\hat{x}, y) = \mathcal{F}(\hat{x} + tu_0 + tu_{\perp}, y) - \mathcal{F}(\hat{x} + tu_0, y) \tag{17}$$

$$+ \mathcal{F}(\hat{x} + tu_0, y) - \mathcal{F}(\hat{x}, y). \tag{18}$$

The term on the right-hand side of (17) is analyzed with the aid of assumption 2. In order to calculate the side derivative $\delta_1\mathcal{F}(\hat{x},y)$, we decompose $\mathcal{F}$ into

$$\mathcal{F}(x',y') = \Psi_{\hat{h}}(x',y') + \mathcal{F}_{\hat{h}}(x',y'), \tag{19}$$

$$\text{where } \Psi_{\hat{h}}(x',y') := \sum_{i \in \hat{h}} \psi(a_i^T x' - y_i')$$

$$\text{and } \mathcal{F}_{\hat{h}}(x',y') = \sum_{i \in \hat{h}^c} \psi(a_i^T x - y_i') + \alpha\Phi(x',y').$$

This decomposition is used recurrently in the following.

*Remark* 3. The function $\mathcal{F}_{\hat{h}}$ is $\mathcal{C}^m$ on a neighborhood of $(\hat{x},y)$ which contains $B(\hat{x},\sigma) \times B(y,\sigma)$ for

$$\sigma := \frac{1}{2(\|a\|_\infty + 1)} \min_{i \in \hat{h}^c} |a_i^T \hat{x} - y_i|, \tag{20}$$

$$\|a\|_\infty := \max_{i=1}^{q} \|a_i\|. \tag{21}$$

Indeed, for every $(x',y') \in B(\hat{x},\sigma) \times B(y,\sigma)$ we have

$$i \in \hat{h}^c \quad \Rightarrow \quad |a_i^T x' - y_i'| = \left|(a_i^T \hat{x} - y_i) + a_i^T(x' - \hat{x}) + (y_i - y_i')\right| \tag{22}$$

$$\geq \left|a_i^T \hat{x} - y_i\right| - \left|a_i^T(x' - \hat{x})\right| - |y_i - y_i'|$$

$$\geq \min_{i \in \hat{h}^c} |a_i^T \hat{x} - y_i| - \|a\|_\infty \sigma - \sigma = (\|a\|_\infty + 1)\sigma > 0,$$

since clearly $\|a\|_\infty > 0$ and $\sigma > 0$.

In contrast, $\Psi_{\hat{h}}$ is nonsmooth at $(\hat{x}, y)$. Using Definition 3 we calculate that for every $u \in \mathbb{R}^p$,

$$\delta_1 \mathcal{F}(x,y)(u) = \delta_1 \Psi_{\hat{h}}(\hat{x},y)(u) + D\mathcal{F}_{\hat{h}}(\hat{x},y).u, \tag{23}$$

$$\text{where } \ \delta_1 \Psi_{\hat{h}}(\hat{x},y)(u) = \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u|, \tag{24}$$

since $\delta\psi(a_i^T\hat{x} - y_i)(u) = \lim_{t\downarrow 0} \psi(ta_i^T u)/t = \psi'(0^+)|a_i^T u|$, for every $i \in \hat{h}$, which accounts for (14). Notice that $\delta_1\Psi_{\hat{h}}(\hat{x},y)(u) = \delta_1\Psi_{\hat{h}}(\hat{x},y)(-u) \geq 0$ for every $u \in \mathbb{R}^p$. Applying assumption 2 to both $u_\perp \in T_{\hat{h}}^\perp$ and $-u_\perp$ yields

$$|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp| < \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_\perp| \ \ \forall u_\perp \in T_{\hat{h}}^\perp. \tag{25}$$

Now consider the function

$$f : T_{\hat{h}}^\perp \cap S \to \mathbb{R},$$
$$u_\perp \to f(u_\perp) := \frac{|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp|}{\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_\perp|}.$$

Since for every $u_\perp \in T_{\hat{h}}^\perp \cap S$ there is at least one index $i \in \hat{h}$ such that $a_i^T u_\perp \neq 0$, this function is well defined and continuous. If $u_\perp \to D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp$ is not identically null on $T_{\hat{h}}^\perp$, put

$$c_0 := \sup_{u_\perp \in T_{\hat{h}}^\perp \cap S} f(u_\perp). \tag{26}$$

Since $T_{\hat{h}}^\perp \cap S$ is compact, $f$ reaches the maximum value $c_0$. By (25) we see that $0 < c_0 < 1$. If $D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp = 0$ for all $u_\perp \in T_{\hat{h}}^\perp$, we put $c_0 := 1/2$. In both cases,

$$|D\mathcal{F}_{\hat{h}}(\hat{x},y).u_\perp| \leq c_0\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_\perp| \ \ \forall u_\perp \in T_{\hat{h}}^\perp. \tag{27}$$

Using (19), the right-hand side of (17) takes the form

$$\mathcal{F}(\hat{x}+tu_0+tu_\perp, y) - \mathcal{F}(\hat{x}+tu_0, y) = \Psi_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) - \Psi_{\hat{h}}(\hat{x}+tu_0, y) \tag{28}$$
$$+ \mathcal{F}_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) - \mathcal{F}_{\hat{h}}(\hat{x}+tu_0, y). \tag{29}$$

First, we focus on the right-hand side of (28). From the definition of $\hat{h}$ and (16),

$$\Psi_{\hat{h}}(\hat{x}+tu_0, y) = 0,$$
$$\Psi_{\hat{h}}(\hat{x}+tu_0+tu_\perp, y) = \sum_{i\in\hat{h}} \psi\left(a_i^T(\hat{x}+tu_\perp+tu_0) - y_i\right) = \sum_{i\in\hat{h}} \psi(ta_i^T u_\perp).$$

Applying Definition 3 to $\psi'(0^+)$ shows that there is $\eta_0 \in (0, \sigma]$ such that

$$\frac{\psi(t)}{t} \geq \psi'(0^+) - \frac{1-c_0}{2}\psi'(0^+) \ \ \forall t \in (0, \|a\|_\infty \eta_0)\,,$$

since $(1-c_0)/2 \in (0,1)$. On the other hand, $|a_i^T u| \leq \|a_i\|\|u\| \leq \|a\|_\infty$ for all $u \in \overline{B(0,1)}$ and for all $i \in \{1,\ldots,q\}$. Then

$$t \in (0,\eta_0) \;\Rightarrow\; \psi(ta_i^T u_{\perp\!\!\!\perp}) \geq \frac{c_0+1}{2}\psi'(0^+)\, t\, |a_i^T u_{\perp\!\!\!\perp}| \;\; \forall u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp} \cap \overline{B(0,1)}.$$

Hence, taking $t \in (0,\eta_0)$ ensures that for all $u \in S$, decomposed into $u = u_0 + u_{\perp\!\!\!\perp}$ as in (16), we have

$$\Psi_{\hat{h}}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) \geq \frac{c_0+1}{2}\, t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}|. \tag{30}$$

Second, we consider (29). Define the constants

$$c_1 := \min_{u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp} \cap S} \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}|, \tag{31}$$

$$c_2 := c_1 \psi'(0^+) \frac{1-c_0}{4}, \tag{32}$$

and notice that $c_1 > 0$ and $c_2 > 0$, and that (31) implies

$$\sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| \geq c_1 \|u_{\perp\!\!\!\perp}\| \quad \forall u_{\perp\!\!\!\perp} \in T_{\hat{h}}^{\perp}. \tag{33}$$

Since $\mathcal{F}_{\hat{h}}(.,y) \in \mathcal{C}^1\left(B(\hat{x},\sigma)\right)$ (see Remark 3), the mean-value theorem [5] shows that for every $u \in S$ and for every $t \in [0,\sigma)$ there exists $\theta \in (0,1)$ such that

$$\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) - \mathcal{F}_{\hat{h}}(\hat{x} + tu_0, y) = tD_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}, \tag{34}$$

where $u = u_0 + u_{\perp\!\!\!\perp}$ is decomposed as in (16). Moreover, there is $\eta_1 \in (0,\eta_0)$ such that for every $t \in (0,\eta_1)$,

$$\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp} - D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| \leq c_2\|u_{\perp\!\!\!\perp}\| \quad \forall u \in S, \;\; \forall \theta \in (0,1),$$

and hence

$$\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}\right| \leq \left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| + c_2\|u_{\perp\!\!\!\perp}\| \;\; \forall u \in S, \;\; \forall \theta \in (0,1). \tag{35}$$

Starting with (28)–(29), we derive

$$\begin{aligned}
&\mathcal{F}(\hat{x} + tu_0 + tu_{\perp\!\!\!\perp}, y) - \mathcal{F}(\hat{x} + tu_0, y) \qquad (36)\\
&\geq \frac{c_0+1}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - t\left|D_1\mathcal{F}_{\hat{h}}(\hat{x} + tu_0 + \theta t u_{\perp\!\!\!\perp}, y).u_{\perp\!\!\!\perp}\right| \quad \text{[by (30) and (34)]}\\
&\geq \frac{c_0+1}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - t\left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp\!\!\!\perp}\right| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (35)]}\\
&\geq \frac{1-c_0}{2} t\, \psi'(0^+) \sum_{i\in\hat{h}} |a_i^T u_{\perp\!\!\!\perp}| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (27)]}\\
&\geq \frac{1-c_0}{2} \psi'(0^+) tc_1\|u_{\perp\!\!\!\perp}\| - tc_2\|u_{\perp\!\!\!\perp}\| \quad \text{[by (33)]}\\
&= \frac{1-c_0}{4} \psi'(0^+) tc_1\|u_{\perp\!\!\!\perp}\|. \quad \text{[by (32)]} \qquad (37)
\end{aligned}$$

Consequently,

$$(38)\ t \in (0,\eta_1) \quad \Rightarrow \quad \mathcal{F}(\hat{x}+tu_0+tu_\perp,y)-\mathcal{F}(\hat{x}+tu_0,y)>0 \quad \forall u\in S \text{ with } u_\perp \neq 0.$$

From assumption 1, there exists $\eta_2 \in (0,\eta_1]$ such that

$$(39)\quad t\in(0,\eta_2) \quad \Rightarrow \quad \mathcal{F}(\hat{x}+tu_0,y)-\mathcal{F}(\hat{x},y)>0 \quad \forall u_0 \in T_{\hat{h}} \cap \overline{B(0,1)} \setminus \{0\}.$$

If $u_0=0$, then (38) holds since $\|u_\perp\|=1$, whereas if $u_\perp=0$, then (39) is true since $\|u_0\|=1$. Introducing (38) and (39) into (17)–(18) shows that if $t\in(0,\eta_2)$, then $\mathcal{F}(\hat{x}+tu,y)-\mathcal{F}(\hat{x},y)>0$ for every $u\in S$. ∎

*Remark* 4. The conditions required in Proposition 1 are pretty weak. Indeed, if an arbitrary function $\mathcal{F}(.,y):\mathbb{R}^p\to\mathbb{R}$ has a strict minimum at $\hat{x}$, then assumption 1 is trivially true and necessarily $\delta_1\mathcal{F}(\hat{x},y)(u)\geq 0$ for all $u\in T_{\hat{h}}^{\perp}\cap S$ [32]. In comparison, assumption 2 requires only that the latter inequality be strict.

Observe that the above sufficient condition for strict minimum concerns the behavior of $\mathcal{F}(.,y)$ on two orthogonal subspaces *separately*. This occurs because of the nonsmoothness of $\psi$.

**4. Minimizers that fit exactly some data entries.** The theorem below states the main contribution of this work.

THEOREM 1. *Consider $\mathcal{F}$ as given in* (11)–(12)*, where $\Phi\in\mathcal{C}^m(\mathbb{R}^p\times\mathbb{R}^q)$ for $m\geq 2$, and $\psi\in\mathcal{C}^m(\mathbb{R}\setminus\{0\})$ has finite side derivatives at zero such that $\psi'(0^-)<\psi'(0^+)$. Given $y\in\mathbb{R}^q$ and $\hat{x}\in\mathbb{R}^p$, let $\hat{h}:=\mathcal{H}(\hat{x},y)$, $\Theta_{\hat{h}}(y)$, and $T_{\hat{h}}$ be obtained by* (6)*,* (7)*, and* (8)*, respectively. Suppose the following:*

1. *The set $\{a_i : i\in\hat{h}\}$ is linearly independent;*
2. *for every $u\in T_{\hat{h}}\cap S$ we have $D_1(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}})(\hat{x},y).u=0$ and $D_1^2(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}})(\hat{x},y)(u,u)>0$;*
3. *for every $u\in T_{\hat{h}}^{\perp}\cap S$ we have $\delta_1\mathcal{F}(\hat{x},y)(u)>0$.*

*Then there is a neighborhood $N\subset\mathbb{R}^q$ containing $y$ and a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}:N\to\mathbb{R}^p$ relevant to $\mathcal{F}(.,N)$ (see Definition 2) yielding, in particular, $\hat{x}=\mathcal{X}(y)$, whereas for every $y'\in N$,*

$$(40)\qquad \begin{aligned} a_i^T\mathcal{X}(y') &= y'_i \quad \text{if} \quad i\in\hat{h},\\ a_i^T\mathcal{X}(y') &\neq y'_i \quad \text{if} \quad i\in\hat{h}^c. \end{aligned}$$

*The latter means that $\mathcal{H}(\mathcal{X}(y'),y')=\hat{h}$ is constant on $N$.*

*Proof.* If $\hat{h}=\emptyset$, then $\Theta_{\hat{h}}(y')=\mathbb{R}^p$ for all $y'$. Applying Lemma 1 shows the existence of $\tilde{N}\subset\mathbb{R}^q$ and of a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}$ relevant to $\mathcal{F}(.,\tilde{N})$. By the continuity of $\mathcal{X}$, there is $N\subset\tilde{N}$ where (40) holds, in which case (40) is reduced to $a_i^T\mathcal{X}(y')\neq y'_i$ for all $i\in\{1,\dots,q\}$.

In the following we consider that $\hat{h}$ is nonempty. As in the proof of Proposition 1, we use the formulation of $\mathcal{F}$ given in (13)–(15) and write $\psi$ for $\tilde{\psi}$ and $\Phi$ for $\tilde{\Phi}$. This proof is based on two lemmas given next.

LEMMA 2. *Let assumptions* 1 *and* 2 *of Theorem* 1 *be satisfied. Then there exist $\nu>0$ and a $\mathcal{C}^{m-1}$-function $\mathcal{X}:B(y,\nu)\to\mathbb{R}^p$ so that for every $y'\in B(y,\nu)$ the point $\hat{x}':=\mathcal{X}(y')$ belongs to $\Theta_{\hat{h}}(y')$ and satisfies*

$$(41)\qquad D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\hat{x}',y').u=0 \ \text{ and } \ D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\hat{x}',y')(u,u)>0 \quad \forall u\in T_{\hat{h}}\setminus\{0\}.$$

*In particular,* $\hat{x} = \mathcal{X}(y)$.

*Proof of Lemma* 2. We start by commenting on the restricted functions in (41).

*Remark* 5. For $\sigma$ as in (20), the inequality reached in (22) shows that for all $(x', y') \in B(\hat{x}, \sigma) \times B(y, \sigma)$ we have $\mathcal{H}(x', y') \subseteq \hat{h}$. On the other hand, if $x' \in \overline{\Theta_{\hat{h}}(y')}$, then $\mathcal{H}(x', y') \supseteq \hat{h}$. If we put

$$B_{\hat{h}}\left((\hat{x}, y), \sigma\right) := \left(B(\hat{x}, \sigma) \times B(y, \sigma)\right) \cap \overline{M_{\hat{h}}}, \tag{42}$$

where $M_{\hat{h}}$ is given in (9), we have

$$(x', y') \in B_{\hat{h}}\left((\hat{x}, y), \sigma\right) \quad \Rightarrow \quad \mathcal{H}(x', y') = \hat{h},$$

and $B_{\hat{h}}\left((\hat{x}, y), \sigma\right) \subset M_{\hat{h}}$. By (7) and (10), for every $(x', y') \in M_{\hat{h}}$ we find $\Psi_{\hat{h}}(x', y') = 0$ and hence $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y') = \mathcal{F}_{\hat{h}}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y')$. Since $\mathcal{F}_{\hat{h}} \in \mathcal{C}^m\left(B(\hat{x}, \sigma) \times B(y, \sigma)\right)$ (see Remark 3), we get

$$\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}} \in \mathcal{C}^m\left(B_{\hat{h}}\left((\hat{x}, y), \sigma\right)\right) \text{ and } \mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x', y') = \mathcal{F}_{\hat{h}}(x', y') \ \forall\ (x', y') \in B_{\hat{h}}\left((\hat{x}, y), \sigma\right).$$

We now pursue the proof of the lemma. Let the indexes contained in $\hat{h}$ read $\hat{h} = \{j_1, \ldots, j_{\#\hat{h}}\}$. Let $I_{\hat{h}}$ be the $\#\hat{h} \times q$ matrix with entries $I_{\hat{h}}[i, j_i] = 1$ for $i = 1, \ldots, \#\hat{h}$, the remaining entries being null. Thus $y_{\hat{h}} := I_{\hat{h}} y \in \mathbb{R}^{\#\hat{h}}$ is composed of only those entries of $y$ whose indexes are in $\hat{h}$. Similarly, put $A_{\hat{h}} := I_{\hat{h}} A$; then $A_{\hat{h}} \in \mathbb{R}^{\#\hat{h} \times p}$ and $A_{\hat{h}} \hat{x} = y_{\hat{h}}$. With this notation, $\overline{M_{\hat{h}}} = \left\{(x', y') \in \mathbb{R}^p \times \mathbb{R}^q : A_{\hat{h}} x' - I_{\hat{h}} y' = 0\right\}$. By assumption 1, $\mathrm{rank} A_{\hat{h}} = \#\hat{h}$. Then for every $y'$ we have the following dimensions: $\dim \overline{\Theta_{\hat{h}}(y')} = \dim T_{\hat{h}} = p - \#\hat{h}$ while $\dim \overline{M_{\hat{h}}} = p - \#\hat{h} + q$. Recalling that $A_{\hat{h}} A_{\hat{h}}^T$ is invertible, put

$$P_{\hat{h}} := A_{\hat{h}}^T \left(A_{\hat{h}} A_{\hat{h}}^T\right)^{-1} I_{\hat{h}}. \tag{43}$$

Let $C_{\hat{h}} : T_{\hat{h}} \to \mathbb{R}^{p - \#\hat{h}}$ be an isomorphism. The affine mapping

$$\begin{aligned} \Gamma : \quad & \overline{M_{\hat{h}}} \to \mathbb{R}^{p - \#\hat{h}}, \\ & (x', y') \to \Gamma(x', y') = C_{\hat{h}}\left(x' - \hat{x} - P_{\hat{h}}(y' - y)\right) \end{aligned} \tag{44}$$

is well defined for every $y' \in \mathbb{R}^q$ since on the one hand $\hat{x} + P_{\hat{h}}\left(y' - y\right)$ is the orthogonal projection[1] of $\hat{x}$ onto $\overline{\Theta_{\hat{h}}(y')}$, whereas on the other hand $x' \in \overline{\Theta_{\hat{h}}(y')}$ by (10). Consider also the conjugate mapping

$$\begin{aligned} \Gamma^\dagger : \quad & \mathbb{R}^{p - \#\hat{h}} \times \mathbb{R}^q \to \overline{\Theta_{\hat{h}}(y')}, \\ & (z, y') \to \Gamma^\dagger(z, y') = C_{\hat{h}}^{-1} z + \hat{x} + P_{\hat{h}}(y' - y), \end{aligned} \tag{45}$$

[1] The orthogonal projection of $\hat{x}$ onto $\overline{\Theta_{\hat{h}}(y')}$, denoted by $\hat{x}_{y'}$, is unique and is determined by solving the problem

$$\text{minimize } \ \|\hat{x}_{y'} - \hat{x}\| \quad \text{subject to } \ \hat{x}_{y'} \in \overline{\Theta_{\hat{h}}(y')}.$$

The latter constraint also reads $A_{\hat{h}} \hat{x}_{y'} = y'_{\hat{h}}$ if we denote $y'_{\hat{h}} = I_{\hat{h}} y'$. It is easily calculated that the solution to this problem reads

$$\hat{x}_{y'} = \hat{x} - A_{\hat{h}}^T \left(A_{\hat{h}} A_{\hat{h}}^T\right)^{-1} \left(A_{\hat{h}} \hat{x} - y'_{\hat{h}}\right).$$

Recalling that $A_{\hat{h}} \hat{x} = I_{\hat{h}} y$ from the definition of $\hat{h}$, we obtain that $\hat{x}_{y'} = \hat{x} + P_{\hat{h}}\left(y' - y\right)$.

which is also well defined. Let

$$\nu_0 := \frac{\sigma}{2}\,\min\left\{1,\,\left(\sup_{z\in S}\|C_{\hat{h}}^{-1}z\| + \sup_{y'\in S}\|P_{\hat{h}}y'\|\right)^{-1}\right\}. \tag{46}$$

Clearly, $0 < \nu_0 < \sigma$. It is worth noticing that

$$\Gamma^\dagger(z,y') \in \overline{\Theta_{\hat{h}}(y')} \cap B(\hat{x},\sigma) \subset \Theta_{\hat{h}}(y') \quad \forall (z,y') \in B(0,\nu_0)\times B(y,\nu_0), \tag{47}$$

since on the one hand (45) shows that $\Gamma^\dagger(z,y') \in \overline{\Theta_{\hat{h}}(y')}$ whereas, on the other hand,

$$\|\Gamma^\dagger(z,y') - \hat{x}\| \le \|C_{\hat{h}}^{-1}\|\,\|z\| + \|P_{\hat{h}}\|\,\|y'-y\| \le (\|C_{\hat{h}}^{-1}\| + \|P_{\hat{h}}\|)\,\nu_0 < \sigma.$$

Now we introduce the function

$$\begin{aligned} &\mathcal{G} : \mathbb{R}^{p-\#\hat{h}} \times \mathbb{R}^q \to \mathbb{R}, \\ &\qquad (z,y') \to \mathcal{G}(z,y') := \mathcal{F}_{\hat{h}}\left(\Gamma^\dagger(z,y'),\ y'\right). \end{aligned} \tag{48}$$

Since for every $y' \in \mathbb{R}^q$ we have

$$z = \Gamma(x',y') \quad \Leftrightarrow \quad x' = \Gamma^\dagger(z,y'),$$

then

$$\mathcal{G}\left(\Gamma(x',y'),y'\right) = \mathcal{F}_{\hat{h}}(x',y') = \mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(x',y') \quad \forall (x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right),$$

where the last equality comes from Remark 5. Now for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right)$, the derivatives of $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}$, mentioned in (41), can be calculated in terms of $\mathcal{G}$ and $\Gamma$ as follows:

$$D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(x',y').u_0 = D_1\mathcal{G}\left(\Gamma(x',y'),y'\right).C_{\hat{h}}u_0 \qquad \forall u_0 \in T_{\hat{h}}, \tag{49}$$

$$D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y)}}\right)(x',y')(u_0,u_0) = D_1^2\mathcal{G}\left(\Gamma(x',y'),y'\right).\left(C_{\hat{h}}u_0, C_{\hat{h}}u_0\right) \qquad \forall u_0 \in T_{\hat{h}}. \tag{50}$$

Since $C_{\hat{h}}$ is an isomorphism, $D_1\Gamma(x',y').u_0 = C_{\hat{h}}.u_0 \neq 0$ for every $u_0 \in T_{\hat{h}} \setminus \{0\}$, whereas $C_{\hat{h}}.T_{\hat{h}} = \mathbb{R}^{p-\#\hat{h}}$. Then assumption 2, combined with the fact that $\Gamma(\hat{x},y) = 0$ by construction, yields

$$\begin{aligned} D_1\mathcal{G}(0,y) &= 0, \\ D_1^2\mathcal{G}(0,y)(u,u) &> 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}. \end{aligned}$$

By Lemma 1, there exist $\nu \in (0,\nu_0]$ and a unique $\mathcal{C}^{m-1}$-function $\mathcal{Z} : B(y,\nu) \to B(0,\nu_0)$ such that

$$D_1\mathcal{G}\left(\mathcal{Z}(y'),\ y'\right) = 0 \ \text{ and } \ D_1^2\mathcal{G}\left(\mathcal{Z}(y'),\ y'\right) \text{ is positive definite } \quad \forall y' \in B(y,\nu), \tag{51}$$

with, in particular, $\mathcal{Z}(y) = 0$. Next we express the derivatives in (51) in terms of $\mathcal{F}_{\hat{h}}$ and $\Gamma^\dagger$. From (47) and Remark 5 it follows that $\mathcal{F}_{\hat{h}}$ is $\mathcal{C}^m$ at every $\left(\Gamma^\dagger(z,y'),y'\right)$ relevant to $(z,y') \in B(0,\nu_0)\times B(y,\nu)$, in which case (48) gives rise to

$$D_1\mathcal{G}(z,y').u = D_1\mathcal{F}_{\hat{h}}(\Gamma^\dagger(z,y'),y').C_{\hat{h}}^{-1}u, \tag{52}$$

$$D_1^2\mathcal{G}(z,y')(u,u) = D_1^2\mathcal{F}_{\hat{h}}(\Gamma^\dagger(z,y'),y')\left(C_{\hat{h}}^{-1}u, C_{\hat{h}}^{-1}u\right). \tag{53}$$

Put

$$\mathcal{X}(y') := \Gamma^{\dagger}\left(\mathcal{Z}(y'), y'\right) \quad \forall y' \in B(y,\nu), \tag{54}$$

and notice that $\mathcal{X}(y') \in \Theta_{\hat{h}}(y')$. Then (51) implies that for every $y' \in B(y,\nu)$,

$$D_1\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y').C_{\hat{h}}^{-1}u = 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}},$$
$$D_1^2\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y')\left(C_{\hat{h}}^{-1}u, C_{\hat{h}}^{-1}u\right) > 0 \quad \forall u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}.$$

Since $C_{\hat{h}}^{-1}u \neq 0$ for all $u \in \mathbb{R}^{p-\#\hat{h}} \setminus \{0\}$ and $C_{\hat{h}}^{-1}.\mathbb{R}^{p-\#\hat{h}} = T_{\hat{h}}$, it follows that for every $y' \in B(y,\nu)$,

$$D_1\mathcal{F}_{\hat{h}}\left(\mathcal{X}(y'),\ y'\right).u_0 = 0 \ \text{ and } \ D_1^2\mathcal{F}_{\hat{h}}(\mathcal{X}(y'), y').(u_0,u_0) > 0 \ \ \forall u_0 \in T_{\hat{h}} \setminus \{0\}.$$

Again applying Remark 5 allows us to write that if $y' \in B(y,\nu)$, then

$$D_1\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)\left(\mathcal{X}(y'),\ y'\right).u_0 = 0 \ \text{ and } \ D_1^2\left(\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}\right)(\mathcal{X}(y'), y')(u_0,u_0) > 0$$
$$\forall u_0 \in T_{\hat{h}} \setminus \{0\}.$$

The proof of Lemma 2 is complete. □

The next lemma addresses assumption 3 of the theorem.

LEMMA 3. *Given $\hat{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, let $\hat{h} = \mathcal{H}(\hat{x}, y) \neq \emptyset$. Let assumption 3 of Theorem 1 hold.*

*Then there exists $\mu > 0$ such that*

$$y' \in B(\hat{x},\mu) \ \text{ and } \ x' \in \Theta_{\hat{h}}(y') \cap B(\hat{x},\mu) \ \Rightarrow \ \delta_1\mathcal{F}(x',y')(u_{\perp}) > 0 \ \ \forall u_{\perp} \in T_{\hat{h}}^{\perp} \cap S. \tag{55}$$

*Proof of Lemma* 3. We decompose $\mathcal{F}$ according to (19). Let $\sigma$ and $B_{\hat{h}}\left((\hat{x},y),\sigma\right)$ be defined according to (20) and (42), respectively. Remark 5 applies to $B_{\hat{h}}\left((\hat{x},y),\sigma\right)$. Similarly to (23)–(24), for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\sigma\right)$ we have

$$\delta_1\mathcal{F}(x',y')(u) = \psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u| + D_1\mathcal{F}_{\hat{h}}(x',y').u \quad \forall u \in \mathbb{R}^p. \tag{56}$$

By the continuity of $D_1\mathcal{F}_{\hat{h}}$, there is $\mu \in (0,\sigma]$ such that for every $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\mu\right)$,

$$\left|D_1\mathcal{F}_{\hat{h}}(x',y').u_{\perp} - D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp}\right| \leq \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| \quad \forall u_{\perp} \in T_{\hat{h}}^{\perp}, \tag{57}$$

where $c_0 \in (0,1)$ and $c_1 > 0$ are the constants given in (26) and (31), respectively. We derive the following inequality chain which holds for all $(x',y') \in B_{\hat{h}}\left((\hat{x},y),\mu\right)$ and for all $u_{\perp} \in T_{\hat{h}}^{\perp}$:

$$\begin{aligned}
&\left|D_1\mathcal{F}_{\hat{h}}(x',y').u_{\perp}\right| \\
&\leq \left|D_1\mathcal{F}_{\hat{h}}(\hat{x},y).u_{\perp}\right| + \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| && \text{[by (57)]} \\
&\leq c_0\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| + \frac{1-c_0}{2}\psi'(0^+)c_1\|u_{\perp}\| && \text{[by (27)]} \qquad (58)\\
&\leq c_0\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| + \frac{1-c_0}{2}\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}| && \text{[by (33)]} \\
&= \frac{c_0+1}{2}\,\psi'(0^+)\sum_{i\in\hat{h}}|a_i^T u_{\perp}|. && (59)
\end{aligned}$$

On the other hand, (56) shows that for every $(x', y') \in B_{\hat{h}}((\hat{x}, y), \mu)$ and for all $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$, we have

$$\begin{aligned}\delta_1 \mathcal{F}(x', y')(u_{\perp}) &\geq \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u_{\perp}| - |D_1 \mathcal{F}_{\hat{h}}(x', y').u_{\perp}| \\ &\geq \left(1 - \frac{c_0 + 1}{2}\right) \psi'(0^+) \sum_{i \in \hat{h}} |a_i^T u_{\perp}| > 0. \qquad \text{[by (59)]}\end{aligned}$$

The last inequality is strict since for every $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$, there is at least one index $i \in \hat{h}$ for which $a_i^T u_{\perp} \neq 0$. □

We now complete the proof of Theorem 1. Consider $\nu > 0$ and $\mu > 0$ the radii found in Lemmas 2 and 3 and $\mathcal{X}$ the function exhibited in Lemma 2. By the continuity of $\mathcal{X}$, there exists $\xi \in (0, \min\{\mu, \nu\}]$ such that $\mathcal{X}(y') \in B(\hat{x}, \mu)$ for every $y' \in B(y, \xi)$. For any $y' \in B(y, \xi)$, consider the point $\hat{x}' := \mathcal{X}(y')$. From Lemma 2, $\hat{x}' \in \Theta_{\hat{h}}(y')$ and $\hat{x}'$ is a strict local minimizer of $\mathcal{F}|_{\overline{\Theta_{\hat{h}}(y')}}(., y')$. From Lemma 3, $\delta_1 \mathcal{F}(\hat{x}', y')(u_{\perp}) > 0$ for all $u_{\perp} \in T_{\hat{h}}^{\perp} \cap S$. All the conditions of Proposition 1 being satisfied, $\mathcal{F}(., y')$ reaches a strict local minimum at $\hat{x}'$. It follows that $\mathcal{X} : B(y, \xi) \to \mathbb{R}^p$ is the sought-after $\mathcal{C}^{m-1}$ minimizer function. □

We now focus on the assumptions involved in this theorem. Assumption 2 is nothing else but the very classical sufficient condition for a strict local minimum of a smooth function over an affine subspace. Assumption 3 was used in Proposition 1 and was discussed therein.

*Remark* 6 (on assumption 1). The subset $\{a_i : i \in \hat{h}\}$ in assumption 1 is determined by (6). With the notation introduced in the beginning of Lemma 2, $y_{\hat{h}} := I_{\hat{h}} y \in \mathbb{R}^{\#\hat{h}}$ belongs to the range of $A_{\hat{h}}$, denoted by $\mathcal{R}(A_{\hat{h}})$. Since $\dim \mathcal{R}(A_{\hat{h}}) = \text{rank} A_{\hat{h}}$, it follows that if $\text{rank} A_{\hat{h}} < \#\hat{h}$, then all $y'_{\hat{h}}$ belonging to $\mathcal{R}(A_{\hat{h}})$ belong to a subspace of dimension strictly smaller than $\#\hat{h}$. Thus, assumption 1 fails to hold only if $y$ is included in a subspace of dimension smaller than $q$. But the chance that noisy data $y$ belong to such a subspace is null. Hence, assumption 1 is satisfied for almost all $y \in \mathbb{R}^q$.

It is worth emphasizing that the independence of the whole set $\{a_i : i \in \{1, \ldots, q\}\}$ is not required. Thus, Theorem 1 addresses any matrix $A$ whether it be ill conditioned, or singular, or invertible.

Theorem 1 entails some important consequences which are discussed next.

*Remark* 7 (stability of minimizers). The fact that there is a $\mathcal{C}^{m-1}$ local minimizer function shows that, in spite of the nonsmoothness of $\mathcal{F}$, for any $y$, all the strict local minimizers of $\mathcal{F}(., y)$ which satisfy the conditions of the theorem are *stable under weak perturbations of data* $y$. This result extends Lemma 1 to nonsmooth functions of the form (11)–(12). Moreover, if for every $y \in \mathbb{R}^q$ the function $\mathcal{F}(., y)$ is strictly convex, then the unique minimizer function $\mathcal{X} : \mathbb{R}^q \to \mathbb{R}^p$, relevant to $\mathcal{F}(., \mathbb{R}^q)$, is $\mathcal{C}^{m-1}$ on $\mathbb{R}^q$.

*Remark* 8 (stability of $\hat{h}$). The result formulated in (40) means that *the set-valued function* $y' \to \mathcal{H}(\mathcal{X}(y'), y')$ *is constant on* $N$, *i.e., that* $\mathcal{H}$ *is constant under small perturbations of* $y$. Equivalently, *all residuals* $(a_i^T \mathcal{X}(y') - y'_i)$ *for* $i \in \hat{h}$ *are null on* $N$.

*Remark* 9 (data domain). Theorem 1 reveals that the data domain $\mathbb{R}^q$ contains *volumes of positive measure* composed of data that lead to local minimizers which

fit exactly the data entries belonging to the same set (e.g., for $A$ invertible, $\alpha = 0$ yields $\hat{h} = \{1, \ldots, q\}$ and the data volume relevant to this $\hat{h}$ is $\mathbb{R}^q$). For a meaningful choice of $\psi$, $\Phi$, and $\alpha$, there are volumes corresponding to various $\hat{h}$, and they are large enough so that noisy data come across them. That is why in practice, nonsmooth data-fidelity terms yield minimizers fitting exactly a certain number of the data entries. The resultant numerical effect is observed in section 7.

Next we present a simple example which illustrates Theorem 1.

*Example* 1 (nonsmooth data-fidelity term). Consider the function

$$\mathcal{F}(x, y) = \sum_{i=1}^{q} |x_i - y_i| + \alpha \sum_{i=1}^{q} \frac{x_i^2}{2},$$

where $\alpha > 0$. For every $y \in \mathbb{R}^q$, the function $\mathcal{F}(., y)$ is strictly convex, so it has a unique minimizer and the latter is strict. Moreover,

$$\min_x \mathcal{F}(x, y) = \sum_{i=1}^{q} \min_{x_i} f(x_i, y_i),$$
$$\text{where} \quad f(x_i, y_i) = |x_i - y_i| + \frac{\alpha x_i^2}{2} \quad \text{for} \quad i = 1, \ldots, q.$$

For $y \in \mathbb{R}^q$, let $\hat{x}$ be the minimizer of $\mathcal{F}(., y)$. Now $\hat{h} = \{i : \hat{x}_i = y_i\}$. For every $i$, the fact that $f(., y_i)$ has a minimum at $\hat{x}_i$ means that $\delta_1 f(\hat{x}_i, y_i)(u) \geq 0$ for every $u \in \mathbb{R}$. Then for every $u \in \mathbb{R}$ we have

$$\text{if} \ \ (i \in \hat{h}^c \ \Leftrightarrow \ \hat{x}_i \neq y_i), \ \text{then} \ \delta_1 f(x_i, y_i)(u) = Df(x_i, y_i).u = (\operatorname{sign}(x_i - y_i) + \alpha x_i) .u \geq 0;$$
$$\text{if} \ \ (i \in \hat{h} \ \Leftrightarrow \ \hat{x}_i = y_i), \ \text{then} \ \delta_1 f(\hat{x}_i, y_i)(u) = |u| + (\alpha y_i) .u \geq 0.$$

From Proposition 1, the entries of the minimizer function $\mathcal{X}$ are

$$\text{if} \ \ |y_i| > \frac{1}{\alpha}, \quad \text{then} \quad \mathcal{X}_i(y) = \frac{1}{\alpha} \operatorname{sign}(y_i);$$
$$\text{if} \ \ |y_i| \leq \frac{1}{\alpha}, \quad \text{then} \quad \mathcal{X}_i(y) = y_i.$$

Theorem 1 applies, provided that $|y_i| \neq 1/\alpha$ for every $i \in \hat{h}$, which corresponds to assumption 3. In such a case, we can take for the neighborhood exhibited in Theorem 1

$$N = B(y, \xi) \ \text{ with } \ \xi = \min_{i=1}^{q} \left| \ |y_i| - \frac{1}{\alpha} \right|.$$

We see that $y' \to \mathcal{H}(\mathcal{X}(y'), y')$ reads

$$\mathcal{H}(\mathcal{X}(y'), y') = \left\{ i \in \{1, \ldots, q\} : |y'_i| \leq \frac{1}{\alpha} \right\}$$

and is constant on $N$. The above expression shows also that the cardinality of $\hat{h}$ increases when $\alpha$ decreases.

We now illustrate Remark 9. For $h \subset \{1, \ldots, q\}$, put

$$V_h := \left\{ y \in \mathbb{R}^q : \ |y_i| \leq \frac{1}{\alpha} \ \forall i \in h \ \text{ and } \ |y_i| > \frac{1}{\alpha} \ \forall i \in h^c \right\}.$$

Obviously, every $y' \in V_h$ gives rise to a minimizer $\hat{x}'$ of $\mathcal{F}(.,y')$ satisfying $\mathcal{H}(\hat{x}',y') = h$. That is, the function $y' \to \mathcal{H}(\mathcal{X}(y'),y')$ is constant on $V_h$. Note that $V_\emptyset = \{y \in \mathbb{R}^q : |y_i| > 1/\alpha \text{ for all } i\}$ and that $V_\emptyset = \emptyset$ if $\alpha = 0$. Moreover, *for every* $h \subset \{1,\dots,q\}$, *the set* $V_h$ *has a positive volume in* $\mathbb{R}^q$, whereas the family of all $V_h$, when $h$ ranges over the family of all the subsets of $\{1,\dots,q\}$ (including the empty set), is a *partition* of $\mathbb{R}^q$.

**5. Smooth data-fidelity terms.** In this section we focus on smooth cost-functions with the goal of checking whether we can get minimizers which fit exactly a certain number of data entries. We start with an illuminating example.

*Example* 2 (smooth cost-function). For $A \in \mathbb{R}^{q\times p}$ and $G \in \mathbb{R}^{r\times p}$ with $r \in \mathbb{N}^*$, consider the cost-function $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$,

$$\mathcal{F}(x,y) = \|Ax - y\|^2 + \alpha\|Gx\|^2. \tag{60}$$

Recall that since the publication of [37], cost-functions of this form are among the most widely used tools in signal and image estimations [25, 22, 35, 13]. Under the classical assumption $\ker A^TA \cap \ker G^TG = \emptyset$, it is seen that for every $y \in \mathbb{R}^q$, $\mathcal{F}(.,y)$ is strictly convex and its unique minimizer $\hat{x}$ is determined by solving the equation

$$D_1\mathcal{F}(\hat{x},y) = 0 \quad \text{where} \quad D_1\mathcal{F}(\hat{x},y) = 2(A\hat{x} - y)^TA + 2\alpha\hat{x}^TG^TG.$$

The relevant minimizer function $\mathcal{X} : \mathbb{R}^q \to \mathbb{R}^p$ reads

$$\mathcal{X}(y) = (A^TA + \alpha\, G^TG)^{-1}A^T.\, y. \tag{61}$$

We now determine the set of *all* data points $y \in \mathbb{R}^q$ for which $\hat{x} := \mathcal{X}(y)$ fits exactly the $i$th data entry $y_i$. To this end, we have to solve with respect to $y$ the equation

$$a_i^T\,\mathcal{X}(y) = y_i. \tag{62}$$

Using (61), this is equivalent to solving the equation

$$\begin{aligned} &p_i(\alpha).y = 0, \\ \text{where} \quad &p_i(\alpha) \quad = a_i^T(A^TA + \alpha\, G^TG)^{-1}A^T - e_i^T. \end{aligned} \tag{63}$$

We can have $p_i(\alpha) = 0$ only if $\alpha$ belongs to the discrete set of several values which satisfy a data-independent system of $q$ polynomials of degree $p$. However, $\alpha$ will almost never belong to such a set so, in general, $p_i(\alpha) \neq 0$. Then (63) implies $y \in \{p_i(\alpha)\}^\perp$. More generally, we have the implication

$$\exists i \in \{1,\dots,q\} \ \text{ such that } \ \mathcal{X}_i(y) = y_i \quad \Rightarrow \quad y \in \bigcup_{j=1}^{q} \{p_j(\alpha)\}^\perp.$$

Since every $\{p_i(\alpha)\}^\perp$ is a subspace of $\mathbb{R}^q$ of dimension $q-1$, the union on the right-hand side above is a *closed, negligible subset of* $\mathbb{R}^q$. The chance that noisy data come across this union is null. Hence, the chance that noisy data $y$ yield a minimizer $\mathcal{X}(y)$ which fits even one data entry, i.e., that there is at least one index $i$ such that (62) holds, is null.

The theorem stated below generalizes this example.

THEOREM 2. *Consider a* $\mathcal{C}^m$*-function* $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$, *with* $m \geq 2$, *of the form* (1)–(2), *and let* $h \subset \{1\dots,q\}$ *be nonempty. Assume the following:*

1. *For all* $i = 1, \dots, q$, *the functions* $\psi_i : \mathbb{R} \to \mathbb{R}$ *satisfy* $\psi_i''(t) > 0$ *for all* $t \in \mathbb{R}$;
2. $A$ *is invertible* (*recall that for every* $i = 1, \dots, q$, *the* $i$*th row of* $A$ *is* $a_i^T$);
3. *there is an open domain* $N_0 \subset \mathbb{R}^q$ *so that* $\mathcal{F}(., N_0)$ *admits a* $\mathcal{C}^{m-1}$ *local minimizer function* $\mathcal{X} : N_0 \to \mathbb{R}^p$, *such that* $D_1^2\mathcal{F}(\mathcal{X}(y), y)$ *is positive definite, for all* $y \in N_0$;
4. *for every* $x \in \mathcal{X}(N_0) \subset \mathbb{R}^p$ *and for every* $i \in h$ *we have* $D^2\Phi(x).[A^{-1}]_i \neq 0$, *where* $[A^{-1}]_i$ *denotes the* $i$*th column of* $A^{-1}$, *for* $i = 1, \dots, q$.

*For a given set of constants* $\{\theta_i, i \in h\}$, *and for any* $N \subset N_0$ *a closed subset of* $\mathbb{R}^q$, *put*

$$\Upsilon_h := \left\{ y \in N : a_i^T \mathcal{X}(y) = y_i + \theta_i \ \forall i \in h \right\}. \tag{64}$$

*Then* $\Upsilon_h$ *is a closed subset of* $\mathbb{R}^q$ *whose interior is empty.*

*Proof.* For every $h$ nonempty we have

$$\Upsilon_h = \bigcap_{i \in h} \Upsilon_{\{i\}}.$$

It is hence sufficient to consider $\Upsilon_{\{i\}}$ for some $i \in h$. For simplicity, in the following we write $\Upsilon_i$ for $\Upsilon_{\{i\}}$. Since $\mathcal{X}$ is continuous on $N$, every $\Upsilon_i$ is closed in $N$ and hence in $\mathbb{R}^q$. Our reasoning below is developed *ad absurdum*. So suppose that $\Upsilon_i$ contains an open, connected subset of $\mathbb{R}^q$, say $\tilde{N} \subset \Upsilon_i \subset N$. We can hence write

$$a_i^T \mathcal{X}(y) = y_i + \theta_i \quad \forall y \in \tilde{N}. \tag{65}$$

Differentiating both sides of this identity with respect to $y$ yields

$$a_i^T D\mathcal{X}(y) = e_i^T \quad \forall y \in \tilde{N}. \tag{66}$$

We next determine the form of $D\mathcal{X}$. Since for every $y \in \tilde{N}$ the point $\mathcal{X}(y)$ is a local minimizer of $\mathcal{F}(., y)$, it satisfies $D_1\mathcal{F}(\mathcal{X}(y), y) = 0$. Differentiating both sides of the latter identity leads to

$$D_1^2\mathcal{F}\left(\mathcal{X}(y), y\right) D\mathcal{X}(y) + D_{1,2}\mathcal{F}\left(\mathcal{X}(y), y\right) = 0 \quad \forall y \in \tilde{N}. \tag{67}$$

The Hessian of $x \to \mathcal{F}(x, y)$, denoted $H(x, y) := D_1^2\mathcal{F}(x, y)$, reads

$$\begin{aligned} H(x, y) &= D_1^2\Psi(x, y) + \alpha D^2\Phi(x) \\ &= A^T \operatorname{Diag}\left(\ddot{\psi}(x, y)\right) A + \alpha D^2\Phi(x), \end{aligned} \tag{68}$$

where for every $x$ and $y$, $\ddot{\psi}(x, y) \in \mathbb{R}^q$ is the vector whose entries read

$$[\ddot{\psi}(x, y)]_i = \psi_i''(a_i^T x - y_i) \ \text{ for } \ i = 1, \dots, q.$$

By assumption 3, $H\left(\mathcal{X}(y), y\right)$ is an invertible matrix for every $y \in \tilde{N}$. Furthermore,

$$D_{1,2}\mathcal{F}(x, y) = -A^T \operatorname{Diag}\left(\ddot{\psi}(x, y)\right).$$

Inserting the last expression and (68) into (67) shows that

$$D\mathcal{X}(y) = \left(H(\mathcal{X}(y), y)\right)^{-1} A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y), y)\right) \quad \forall y \in \tilde{N}. \tag{69}$$

Now introducing (69) into (66) yields

$$a_i^T \left(H(\mathcal{X}(y),y)\right)^{-1} A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right) = e_i^T \quad \forall y \in \tilde{N}. \tag{70}$$

By assumption 1, $\operatorname{Diag}\big(\ddot{\psi}(\mathcal{X}(y),y)\big)$ is invertible for every $y \in \tilde{N}$. Its inverse is a diagonal matrix whose diagonal terms are $\left(\psi_i''(a_i^T\mathcal{X}(y)-y_i)\right)^{-1}$ for $i=1\ldots,q$. Noticing that

$$e_i^T \left(\operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right)\right)^{-1} = \frac{e_i^T}{\psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)},$$

we find that (70) equivalently reads

$$\psi_i''(a_i^T\mathcal{X}(y)-y_i)\,.a_i^T \left(H(\mathcal{X}(y),y)\right)^{-1} = e_i^T A^{-T} \quad \forall y \in \tilde{N},$$

where $A^{-T} := \left(A^T\right)^{-1}$. Then, taking into account (68),

$$\psi_i''(a_i^T\mathcal{X}(y)-y_i)\,.a_i^T = e_i^T A^{-T}\left(A^T \operatorname{Diag}\left(\ddot{\psi}(\mathcal{X}(y),y)\right) A + \alpha D^2\Phi(\mathcal{X}(y))\right) \quad \forall y \in \tilde{N}.$$

By the invertibility of $A$ (assumption 2), and noticing that $e_i^T A = a_i^T$, the latter expression is simplified to

$$\psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)\,.a_i^T = \psi_i''\left(a_i^T\mathcal{X}(y)-y_i\right)\,.a_i^T + \alpha e_i^T A^{-T} D^2\Phi(\mathcal{X}(y)) \quad \forall y \in \tilde{N},$$

and finally to

$$D^2\Phi(\mathcal{X}(y)).A^{-1}e_i = 0 \quad \forall y \in \tilde{N}.$$

However, the obtained identity contradicts assumption 4. Hence the conclusion. □

Let us comment on the assumptions taken in this theorem. Recall first that assumption 3 was discussed in Lemma 1 and Remark 1. In the typical case when $\Psi$ is a data-fidelity measure, every $\psi_i$ is a strictly convex function satisfying $\psi_i(0)=0$ and $\psi_i(t)=\psi_i(-t)$.

*Remark* 10 (on assumption 2). This proposition also addresses the case when

$$\mathcal{F}(x,y) = \|Ax-y\|^2 + \alpha\Phi(x) \;\text{ with }\; \operatorname{rank}A = p \le q.$$

Indeed, for $p<q$, $\mathcal{F}$ can equivalently be expressed in terms of an invertible $p\times p$ matrix $\tilde{A}$ with $\tilde{A}^T\tilde{A} = A^TA$ in place of $A$.

*Remark* 11 (on assumption 4). By the invertibility of $A$ (assumption 2), we see that $[A^{-1}]_i = A^{-1}e_i \neq 0$ for every $i=1,\ldots,q$. It would be a "pathological" situation to have some of the columns of $A^{-1}$ in $\ker D^2\Phi(x)$ for some $x$. For instance, focus on the classical case given in (4) with $G_i^T : \mathbb{R}^p \to \mathbb{R}$. Let $G$ denote the $r\times p$ matrix whose rows are $G_i^T$ for $i=1,\ldots,r$. Then $D^2\Phi(x) = G^T\operatorname{Diag}\left(\ddot{\varphi}(Gx)\right)G$, where $\ddot{\varphi}(Gx)\in\mathbb{R}^r$ is the vector with entries $[\ddot{\varphi}(Gx)]_i = \varphi''(G_i^Tx)$ for $i=1,\ldots,r$. Focus on the case when $\varphi''(t)>0$ for all $t\in\mathbb{R}$ (e.g., $\varphi$ is strictly convex) and $G$ yields first-order differences between neighboring samples. Then $\operatorname{Ker}D^2\Phi(x)$ is composed of the constant vectors $\kappa[1,\ldots,1]^T$, $\kappa\in\mathbb{R}$. Then assumption 4 is satisfied provided that $A^{-1}$ does not involve constant columns.

*Remark* 12 (meaning of the theorem). If for some $y\in\mathbb{R}^q$ a minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ satisfies an affine equation of the form $a_i^T\hat{x} = y_i + \theta_i$, then Theorem 2 asserts that

$y$ belongs to a closed subset of $\mathbb{R}^q$ whose interior is empty. There is no chance that noisy data $y$ yield local minimizers of a smooth cost-function $\mathcal{F}(.,y)$ satisfying such an equation.

The next proposition states the same conclusions but under different assumptions.

PROPOSITION 2. *Consider a $\mathcal{C}^m$-function $\mathcal{F}:\mathbb{R}^p\times\mathbb{R}^q\to\mathbb{R}$, with $m\geq 2$, of the form (1)–(2) and let $h\subset\{1\ldots,q\}$ be nonempty. Assume the following:*

1. *There is a domain $N_0\subset\mathbb{R}^q$ so that $\mathcal{F}(.,N_0)$ admits a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{X}:N_0\to\mathbb{R}^p$ such that $D_1^2\mathcal{F}(\mathcal{X}(y),y)$ is positive definite for all $y\in N_0$;*
2. *for every $y\in N_0$ and for every $i\in h$ there exists $j\in\{1,\ldots,q\}$ such that the function $\mathcal{K}_{i,j}$,*

$$\mathcal{K}_{i,j}(y'):=\psi_i''\left(a_j^T\mathcal{X}(y')-e_j^Ty'\right).a_i^T\left(H(\mathcal{X}(y'),y')\right)^{-1}.a_j,$$

*where $H$ was given in (68), is nonconstant on any neighborhood of $y$.*

*For $\{\theta_i\in\mathbb{R}:i\in h\}$ given, and for every $N\subset N_0$ a closed subset of $\mathbb{R}^q$, put*

$$\Upsilon_h:=\left\{y\in N:a_i^T\mathcal{X}(y)=y_i+\theta_i\ \forall i\in h\right\}. \tag{71}$$

*Then $\Upsilon_h$ is a closed subset of $\mathbb{R}^q$ whose interior is empty.*

*Proof.* As in the proof of Theorem 2, we focus on $\Upsilon_i$ for $i\in h$ and develop our reasoning by contradiction. So suppose that $\Upsilon_i$ contains an open ball $\tilde{N}$. Then (65) and (66) are true. In particular, comparing (66) for $y'\neq y$ with the same equality for $y$ yields

$$a_i^TD\mathcal{X}(y')=a_i^TD\mathcal{X}(y)\quad\forall y'\in\tilde{N}. \tag{72}$$

Notice that $A^T\operatorname{Diag}\big(\ddot{\psi}(x,y')\big)$ is a matrix whose $j$th column reads $\psi''(a_j^Tx-y_j').a_j$. Introducing (69) into (72) shows that the latter is equivalent to the system

$$\mathcal{K}_{i,j}(y')=\mathcal{K}_{i,j}(y)\ \ \forall j\in\{1,\ldots,q\},\ \ \forall y'\in\tilde{N}.$$

The obtained result contradicts assumption 2. ∎

*Remark* 13 (on assumption 2). Although a general proof of the validity of this assumption appears to be more intricate than important, we conjecture that it is usually satisfied. The intuitive arguments are the following. Let us focus on the classical case when $\Phi$ is as in (4). The entries of $H(x',y')$ read

$$[H(x',y')]_{m,n}=\sum_{j=1}^q\eta_{j,m}^2\psi''(a_jx'-y_j')+\sum_{j=1}^r\kappa_{j,n}^2\varphi''(G_jx')\ \text{ for }\ (m,n)\in\{1,\ldots,p\}^2, \tag{73}$$

where $\eta_{j,m}$, $j=1,\ldots,q$, and $\kappa_{j,n}$, $j=1,\ldots,r$, are constants that are calculated from $G$ and $A$. From Cramer's rule for matrix inversion, for every $j$, the term $a_i^T\left(H(x',y')\right)^{-1}a_j$ is the fraction of two polynomials. The entries of the numerator read $\beta_{s,m,n}([H(x',y')]_{m,n})^s$ for all $(m,n)\in\{1,\ldots,p\}^2$ with $\beta_{s,m,n}\in\mathbb{R}$ for $s=0,\ldots,p-1$. In the denominator we have $\gamma_{s,m,n}([H(x',y')]_{m,n})^s$ for all $(m,n)\in\{1,\ldots,p\}^2$ with $\gamma_{s,m,n}\in\mathbb{R}$ for $s=0,\ldots,p$. For $\mathcal{X}$ a minimizer function and $j$ and $i$ given, $\mathcal{K}_{i,j}$ has the form

$$\mathcal{K}_{i,j}(y')=\psi''\left(a_j^T\mathcal{X}(y')-y_j'\right).\frac{\sum_{s=1}^{p-1}\sum_{(m,n)}\beta_{s,m,n}([H(\mathcal{X}(y'),y')]_{m,n})^s}{\sum_{s=1}^{p}\sum_{(m,n)}\gamma_{s,m,n}([H(\mathcal{X}(y'),y')]_{m,n})^s}. \tag{74}$$

Assumption 2 requires that for $i\in h$, there is at least one index $j\in\{1,\ldots,q\}$ for which the relevant function $\mathcal{K}_{i,j}$ does not remain constant on any neighborhood of $y$.

**6. Nonsmooth regularization versus nonsmooth data-fidelity.** In this section we compare cost-functions involving nonsmooth data-fidelity terms to cost-functions involving nonsmooth regularization terms. The visual effects produced by these classes of cost-functions can be seen in section 7.

Cost-functions with *nonsmooth regularization* typically have the form (1), where $\Psi$ is a $\mathcal{C}^m$-function, $m \geq 2$, whereas $\Phi$ is as in (4) with $\varphi$ nonsmooth at zero. Most often, $\Psi(x, y) = \|Ax - y\|^2$. Nonsmooth functions $\varphi$ are, for instance, the $L^1$- and concave functions in (5). Since the publication of [33, 18], such cost-functions are customarily used in signal and image restoration [18, 1, 14, 11, 12, 38]. Visually, the obtained minimizers exhibit a *staircasing effect* since they typically involve many constant regions—see, for instance, Figures 6 and 10 in section 8. This effect is discussed by many authors [18, 15, 14, 12]. In particular, the ability of the $L^1$-function to recover noncorrelated "nearly black" images in the simplest case when $G_i = e_i$ for all $i$ was interpreted in [15] using mini-max decision theory. Total-variation methods, corresponding to $\varphi(t) = |t|$ also, were observed to yield "blocky images" [14, 12]. The concave function was shown to transform ramp-shaped data into a step-shaped minimizer [19].

A theoretical explanation of staircasing was given in [26, 27, 28]. It was shown there that regularization of the form (4) with $\varphi$ nonsmooth at zero yields local minimizers $\hat{x}$ which satisfy $G_i^T \hat{x} = 0$ *exactly* for many indexes $i$. For instance, if $G_i^T$, $i = 1, \ldots, r$, yield first-order differences between neighboring samples (if $x$ is a signal of $\mathbb{R}^p$, $G_i^T x = x_i - x_{i+1}$ for $i = 1, \ldots, p-1$), the relevant minimizers $\hat{x}$ are constant over many zones. If $G_i^T$, $i = 1, \ldots, r$, yield second-order differences, then $\hat{x}$ involves many zones over which it is affine, etc. More generally, the sets of indexes $i$ for which $G_i^T \hat{x} = 0$ determine zones which can be said to be *strongly homogeneous* [27]. Staircasing is due to a special form of stability property which is explained next. Let a data point $y$ give rise to a local minimizer $\hat{x}$ which satisfies $G_i^T \hat{x} = 0$ for all $i \in \hat{h}$, where $\hat{h} \neq \emptyset$. It is shown in [26, 27, 28] that $y$ is in fact contained in a neighborhood $N \in \mathbb{R}^q$ whose elements $y' \in N$ (noisy data) give rise to local minimizers $\hat{x}'$ of $\mathcal{F}(., y')$, placed near $\hat{x}$, which satisfy $G_i^T \hat{x}' = 0$ for all $i \in \hat{h}$. Since every such $N$ is a volume of positive measure, noisy data come across these volumes and yield minimizers satisfying $G_i^T \hat{x}' = 0$ for many indexes $i$. Notice that this behavior is due to the nonsmoothness of $\varphi$ at zero since it cannot occur with differentiable cost-functions [27, 28].

The behavior of the minimizers of cost-functions with *nonsmooth data-fidelity*, as considered in Theorem 1, is opposite. If $y$ leads to a minimizer $\hat{x}$ which fits exactly a set $\hat{h}$ of entries of $y$, Theorem 1 shows that $y$ is contained in a neighborhood $N$ such that the relevant minimizer function $\mathcal{X}$ follows closely every small variation of all data entries $y'_i$ for $i \in \hat{h}$ when $y'$ ranges over $N$. Thus $a_i^T \mathcal{X}(y')$ is never constant in the vicinity of $y$ for $i \in \hat{h}$.

**7. Nonsmooth data-fidelity to detect and smooth outliers.** Our objective now is to process data in order to detect, and possibly to smooth, outliers and impulsive noise. To this end, take $a_i = e_i$ for every $i \in \{1, \ldots, q\}$ in (2). Focus on

$$\mathcal{F}(x, y) = \sum_{i=1}^{q} \psi(x_i - y_i) + \alpha \sum_{i=1}^{r} \varphi(G_i^T x), \tag{75}$$

where $G_i^T : \mathbb{R}^p \to \mathbb{R}$ for $i = 1, \ldots, r$ yield differences between neighboring samples (e.g., $G_i^T x = x_i - x_{i+1}$ if $x$ is a signal); $\psi$ and $\varphi$ are even and strictly increasing on $[0, \infty)$, with $\psi'(0^+) > 0$ and $\varphi$ smooth on $\mathbb{R}$. Suppose that $\hat{x}$ is a strict minimizer

of $\mathcal{F}(.,y)$ and put $\hat{h}=\mathcal{H}(\hat{x},y)$. Based on the results in section 4, we naturally come to the following method for the detection of outliers. Since every $y_i$ corresponding to $i\in\hat{h}$ is kept intact in the minimizer $\hat{x}$, that is, $\hat{x}_i=y_i$, every such $y_i$ can be considered as a *faithful data entry*. In contrast, every $y_i$ with $i\in\hat{h}^c$ corresponds to $\hat{x}_i\neq y_i$ which can indicate that this $y_i$ is aberrant. In other words, *given $y\in\mathbb{R}^q$, we posit that $\hat{h}^c$, the complementary of $\hat{h}=\mathcal{H}(\mathcal{X}(y),y)$, provides an estimate of the locations of the outliers in $y$*. The possibility of keeping intact all faithful data entries is both spectacular and valuable from a practical point of view, e.g., to preprocess data.

*Remark* 14 (stability of the detection of outliers). If a minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ for $y\in\mathbb{R}^q$ gives rise to $\hat{h}=\mathcal{H}(\hat{x},y)$, then Theorem 1 ensures that all data $y'$ placed near $y$ yield minimizers $\hat{x}'$ which recover exactly the same set of outlier positions $\hat{h}^c$. Hence, the suggested method for detection of outliers is stable under small data variations.

We also can envisage *smoothing* outliers since the value of every $\hat{x}_i$ for $i\in\hat{h}^c$ is obtained from the values of neighboring data samples through the terms $\alpha\varphi(G_j^T\hat{x})$ for all $j$ neighbor of $i$. Small values of $\alpha$ make the weight of $\Psi$ more important, so the relevant minimizers $\hat{x}$ fit larger sets of data entries, i.e., $\hat{h}$ is larger. At the same time, all samples $\hat{x}_i$ for $i\in\hat{h}^c$ incur an only-weak smoothing and may remain close to $y_i$. In contrast, large values of $\alpha$ improve smoothing since they increase the weight of $\Phi$. To resume, small values of $\alpha$ are better adapted for the detection of outliers while large values of $\alpha$ are better suited for smoothing of outliers. We are hence faced with a compromise between efficiency of detection and quality of smoothing. The next example, as well as the experiments presented below, corroborate this conjecture.

*Example* 3. Consider the following cost-function:

$$\mathcal{F}(x,y)=\sum_{i=1}^{q}|x_i-y_i|+\alpha\sum_{i=1}^{p-1}(x_i-x_{i+1})^2.$$

Let $\hat{x}$ be a minimizer of $\mathcal{F}(.,y)$ for which $\hat{h}:=\mathcal{H}(\hat{x},y)$ is nonempty. Focus on $i\in\hat{h}^c$. Since $\hat{x}_i\neq y_i$, then

$$0=\frac{\partial\mathcal{F}(\hat{x},y)}{\partial\hat{x}_i}=\operatorname{sign}(\hat{x}_i-y_i)+2\alpha\left((\hat{x}_i-\hat{x}_{i+1})-(\hat{x}_{i-1}-\hat{x}_i)\right),$$

which yields

$$\hat{x}_i=\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{\operatorname{sign}(\hat{x}_i-y_i)}{4\alpha}. \tag{76}$$

Hence, $\hat{x}_i$ takes the form (76) only if we have

$$\text{either}\quad y_i>\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}+\frac{1}{4\alpha}\quad\text{or}\quad y_i<\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{1}{4\alpha}.$$

We remark that (76) does not involve $y_i$ but only the sign of $(\hat{x}_i-y_i)$. Thus, if $y_i$ is an outlier, the value of $\hat{x}_i$ relies only on faithful data entries $y_j$ for $j\in\hat{h}$ by means of $\hat{x}_{i-1}$ and $\hat{x}_{i+1}$. Moreover, the smoothing incurred by $\hat{x}_i$ is stronger for large values of $\alpha$, since then $\hat{x}_i$ is closer to the mean of $\hat{x}_{i-1}$ and $\hat{x}_{i+1}$. Otherwise, if $i\in\hat{h}$, we have $\delta_1\mathcal{F}(\hat{x},y)(e_i)\geq 0$, which yields

$$\hat{x}_i=y_i\quad\Leftrightarrow\quad\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}-\frac{1}{4\alpha}\leq y_i\leq\frac{\hat{x}_{i-1}+\hat{x}_{i+1}}{2}+\frac{1}{4\alpha}.$$

This inequality is easier to satisfy if $\alpha$ is small, in which case numerous data samples are fitted exactly, whereas only a few samples are detected as outliers.

Concrete results depend on the shape of $\psi$, $\varphi$, $\{G_i^T\}$, and $\alpha$. We leave this crucial question for future work. In order to recover and smooth outliers, we take the following cost-function:

$$\mathcal{F}(x,y) = \sum_{i=1}^{q} |x_i - y_i| + \alpha \sum_{i=1}^{p} \sum_{j \in \mathcal{N}(i)} |x_i - x_j|^\nu \quad \text{for} \quad \nu \in (1,2], \tag{77}$$

where for every $i = 1, \ldots, p$ the set $\mathcal{N}(i)$ contains the indexes of all samples $j$ which are neighbors to $i$. In all the restorations presented below, $\mathcal{N}(i)$ is composed of the eight nearest neighbors. Since the publication of [9], we can expect that $\nu > 1$ but close to 1 allow edges to be better preserved when outliers are smoothed. Based on this, all the experiments with (77) in the following correspond to $\nu = 1.1$.

The minimizer $\hat{x}$ of $\mathcal{F}(.,y)$ for $y \in \mathbb{R}^q$ is calculated by continuation. Using that the Huber function (5),

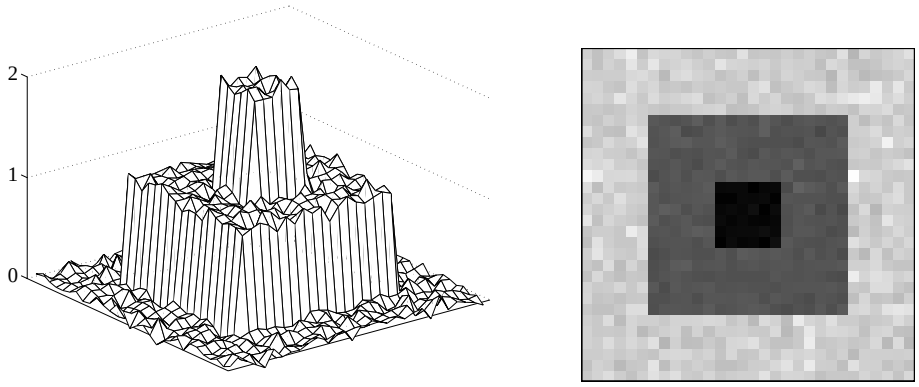
$$\psi_\nu(t) = \begin{cases} t^2 & \text{if} \quad |t| \leq \nu, \\ \nu(\nu + 2|t - \nu|) & \text{if} \quad |t| > \nu, \end{cases} \quad \text{where} \quad \nu > 0,$$

satisfies $\psi_\nu(t) \to |t|$ when $\nu \downarrow 0$, we construct a family of functions $\mathcal{F}_\nu(.,y)$ indexed by $\nu > 0$:

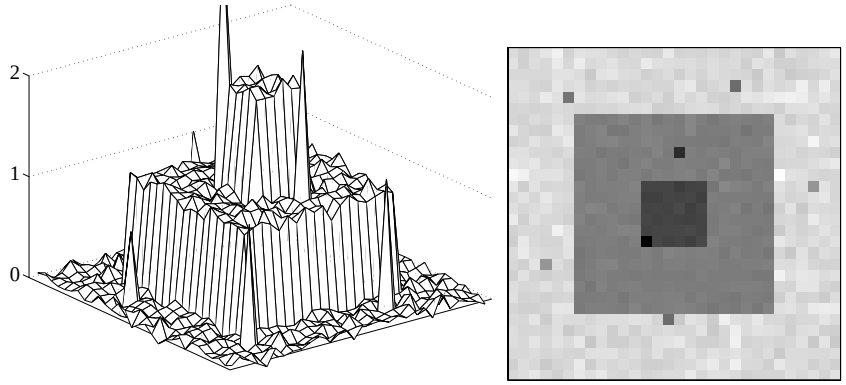
$$\mathcal{F}_\nu(x,y) := \sum_{i=1}^{q} \psi_\nu(a^T x - y_i) + \Phi(x).$$

Being strictly convex and differentiable, every $\mathcal{F}_\nu(.,y)$ has a unique minimizer, denoted by $\hat{x}_\nu$, which is calculated by a gradient descent. Since by construction having $\nu > \nu'$ entails $\mathcal{F}_\nu(x,y) \geq \mathcal{F}_{\nu'}(x,y)$ for all $x \in \mathbb{R}^p$, we see that $\mathcal{F}_\nu(\hat{x}_\nu, y)$ decreases monotonically when $\nu$ decreases to 0. It is easy to check that, moreover, as $\nu \downarrow 0$, we have $\mathcal{F}_\nu(\hat{x}_\nu, y) \to \mathcal{F}(\hat{x}, y)$, and hence $\hat{x}_\nu \to \hat{x}$, since every $\mathcal{F}_\nu(.,y)$ has a unique minimizer and the latter is strict. Total-variation methods are similar from a numerical point of view since they involve $\varphi(t) = |t|$. Many authors used smooth approximations [33, 38], e.g., $\varphi_\nu = \sqrt{t^2 + \nu}$. However, approximation using the Huber function has the numerical advantage of involving only quadratic and affine segments. At the same time, the fact that $\psi'_\nu$ is discontinuous at $\pm\nu$ is of no practical importance since the chance of obtaining a minimizer $\hat{x}_\nu$ involving a difference whose modulus is exactly $\nu$ is null [27].

**First experiment.** The original image $x$ in Figure 1(a) can be assumed to be a noisy version of an ideal piecewise constant image. Data $y$ in Figure 1(b) are obtained by adding aberrant impulsions to $x$ whose locations are seen in Figure 4, left. Recall that our goal is to detect, and possibly smooth, the outliers in $y$, while preserving all the remaining entries of $y$.




(a) Original $x$.

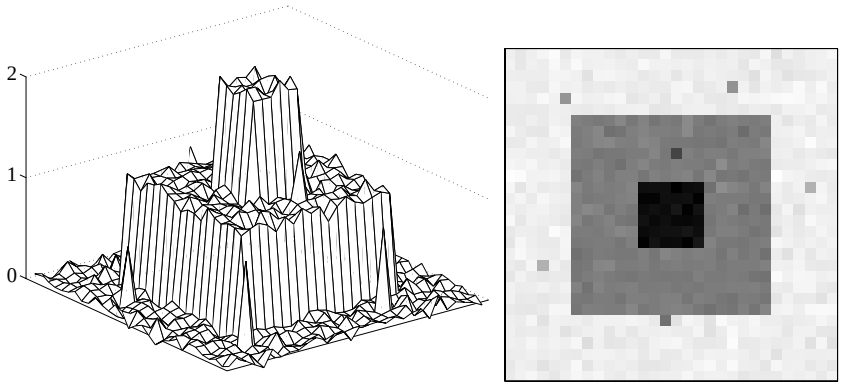



(b) Data $y = x +$ outliers.

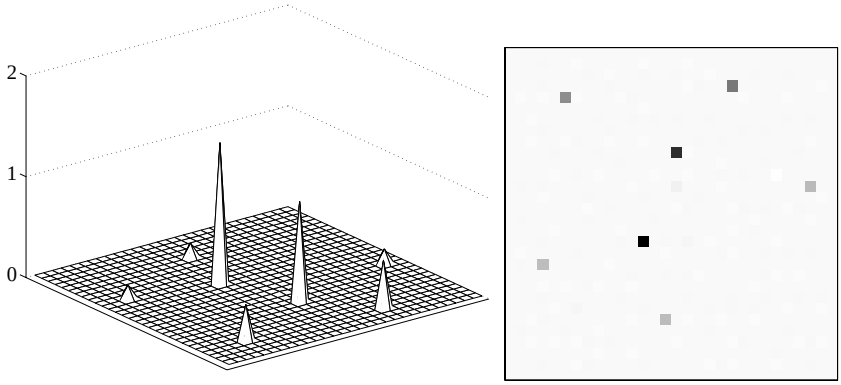
FIG. 1. *Original $x$ and data $y$ degraded by outliers.*

The image in Figure 2(a) is the minimizer $\hat{x}$ of the cost-function $\mathcal{F}(., y)$ proposed in (77), with $\nu = 1.1$ and $\alpha = 0.14$. The outliers are clearly visible although their amplitudes are considerably reduced. The image of the residuals $y - \hat{x}$, shown in Figure 2(b), is null everywhere except at the positions of the outliers in $y$. Reciprocally, the pixels corresponding to nonzero residuals (i.e., the elements of $\hat{h}^c$) provide a faithful estimate of the locations of the outliers in $y$, as seen in Figure 4, middle. Next, in Figure 3(a) we show a minimizer $\hat{x}$ of the same $\mathcal{F}(., y)$ obtained for $\alpha = 0.25$. This minimizer does not contain visible outliers and is very close to the original image $x$. The image of the residuals $y - \hat{x}$ in Figure 3(b) is null only on restricted areas but has a very small magnitude everywhere beyond the positions of the outliers. However, applying the above detection rule now leads to numerous false detections, as seen in Figure 4, right. These experiments confirm our conjecture about the role of $\alpha$.

The issue of the minimization of a smooth cost-function, namely, $\mathcal{F}$ in (75) with $\psi(t) = \varphi(t) = t^2$ and $\alpha = 0.2$, is shown in Figure 5(a). As expected, edges are blurred, whereas outliers are clearly seen. The residuals in Figure 5(b) are large everywhere, which shows that $\hat{x}$ does not fit any data entry. The minimizer in Figure 6(a) is obtained using nonsmooth regularization, where $\mathcal{F}$ is of the form (75) with $\psi(t) = t^2$, $\varphi(t) = |t|$, and $\alpha = 0.2$. In accordance with our discussion in section 6, $\hat{x}$ exhibits




(a) Restoration $\hat{x}$ for $\alpha = 0.14$.



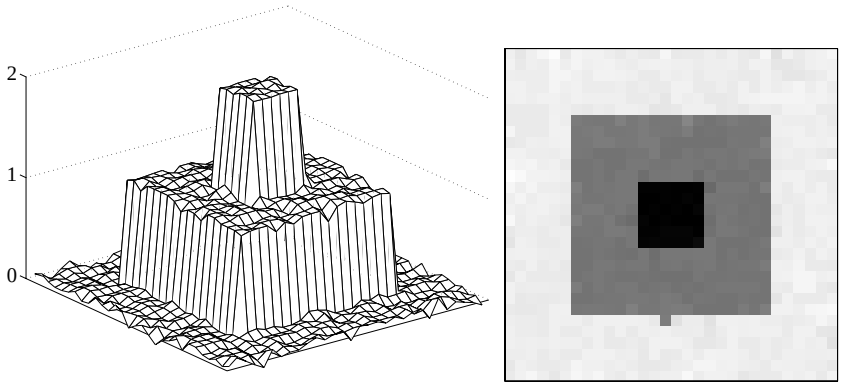

(b) Residuals $y - \hat{x}$.

FIG. 2. *Restoration using the proposed cost-function $\mathcal{F}$ with nonsmooth data-fidelity in* (77) *for $\nu = 1.1$ and $\alpha = 0.14$. The residuals provide a faithful estimator for the locations of outliers.*

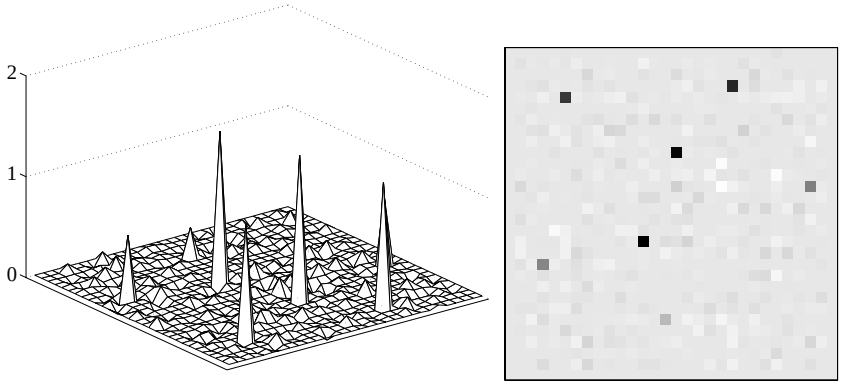
staircasing since it is constant on very large regions.

**Second experiment.** The original, clean image $x$ is shown in Figure 7(a). The data $y$, shown in Figure 7(b), are obtained by adding to $x$ 770 impulsions with random locations and random amplitudes in the interval $(0, 1.2)$.

In Figure 8(a) we show a zoom of the histograms of $x$ (up) and of $y$ (down). Figure 8(b) shows the result from applying to $y$ two iterations of median filtering. The obtained image contains only a few outliers with weak amplitude but the entire image is degraded and, in particular, the edges are blurred. The $\ell_1$-norm of the error $\|\hat{x} - x\|_1 = \sum_i |\hat{x}_i - x_i|$ is 523. The next two restorations in Figure 9 are obtained by minimizing the cost-function $\mathcal{F}$ with nonsmooth data-fidelity proposed in (77), where $\nu = 1.1$. The minimizer in Figure 9(a) corresponds to $\alpha = 0.2$ and it fits exactly the data everywhere except for several hundred pixels, where it detects outliers. This detection gives rise to 50 erroneous nondetections and to 15 false alarms, the remaining detections being correct. Figure 9(b) is obtained for $\alpha = 0.55$. The minimizer $\hat{x}$ does not contain outliers any longer but it fits exactly only a restricted number of the data entries. Nevertheless, it remains very close to all data entries which are not outliers, since the $\ell_1$-norm of the error is 126. This minimizer provides a very clean restoration,




(a) Restoration $\hat{x}$ for $\alpha = 0.25$.




(b) Residuals $y - \hat{x}$.

FIG. 3. *Restoration using the proposed cost-function $\mathcal{F}$ in* (77) *for $\nu = 1.1$ and $\alpha = 0.25$. The outliers are well smoothed in $\hat{x}$, whereas the residuals remain small everywhere beyond the outlier locations.*

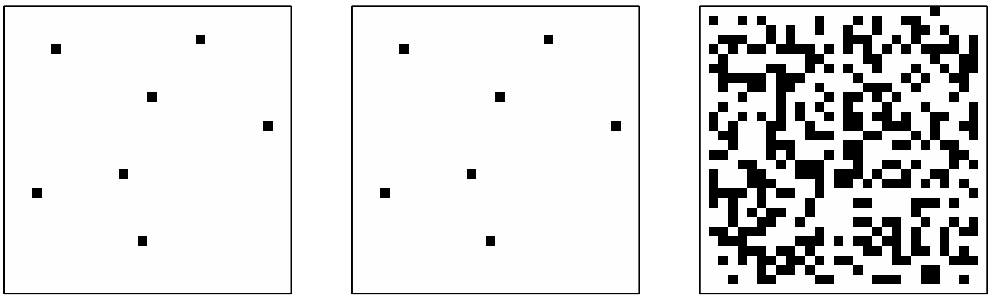
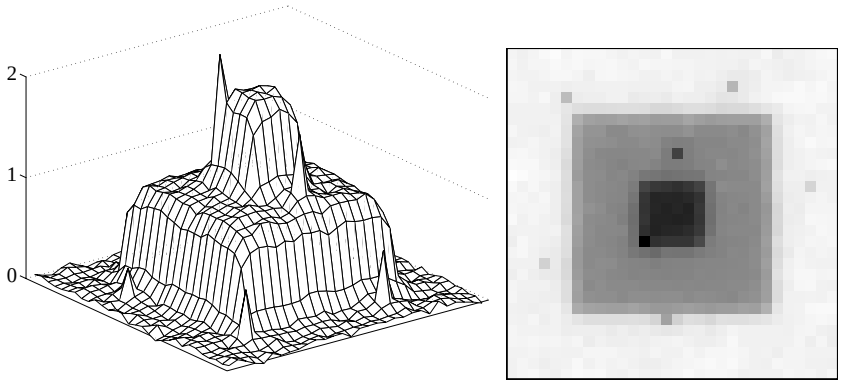
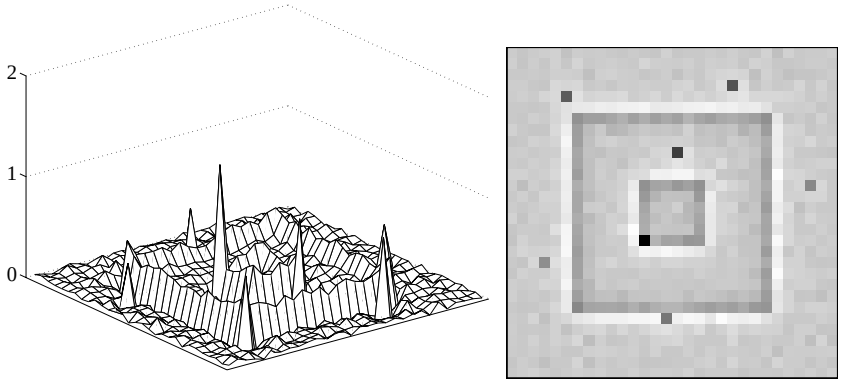


FIG. 4. *Left: The locations of the outliers in $y$. Middle: The locations of the pixels $i$ of $\hat{x}$ at which $\hat{x}_i \neq y_i$, where $\hat{x}$ is the minimizer obtained for $\alpha = 0.14$ given in Figure* 2. *Right: The same locations for $\hat{x}$ the minimizer relevant to $\alpha = 0.25$, shown in Figure* 3.

where both edges and smoothly varying areas are nicely preserved. The restoration in Figure 10(a) results from a smooth cost-function $\mathcal{F}$, as in (75) with $\psi(t) = \varphi(t) = t^2$




(a) Restoration from $yo\ \hat{x}$ for $\alpha = 0.2$.

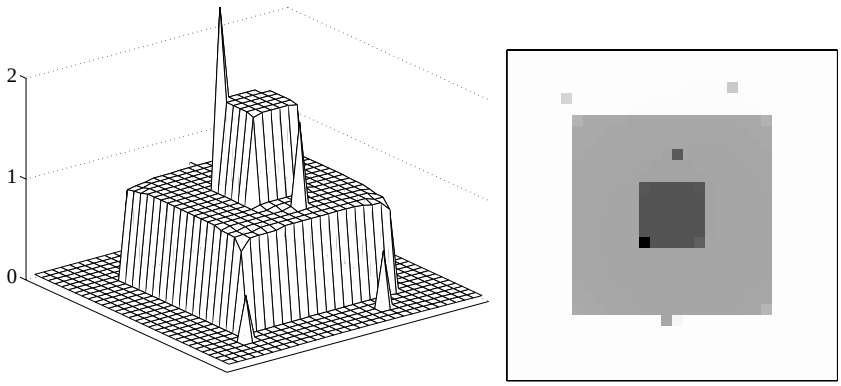



(b) Residuals $y - \hat{x}$.

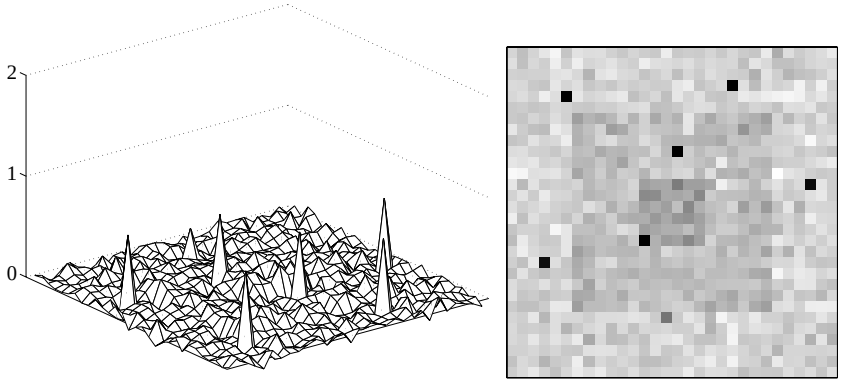
FIG. 5. *Restoration using a smooth cost-function, namely,* $\mathcal{F}$ *in* (75) *with* $\psi(t) = \varphi(t) = t^2$ *and* $\alpha = 0.2$.

and $\alpha = 0.2$. This image fits no data entry while edges are smooth. Figure 10(b) illustrates the staircasing effect induced by nonsmooth regularization. This minimizer corresponds to $\mathcal{F}$, of the form (75) with $\psi(t) = t^2$ and $\varphi(t) = |t|$, for $\alpha = 0.4$ and it still contains several outliers.

**8. Conclusion.** We showed that taking nonsmooth data-fidelity terms in a regularized cost-function yields minimizers which fit exactly a certain number of the data entries. In contrast, this cannot occur for a smooth cost-function. These are strong properties which can be used in different ways. We proposed a cost-function with a nonsmooth data-fidelity term in order to process outliers. The obtained results advocate the use of nonsmooth data-fidelity terms in image processing.

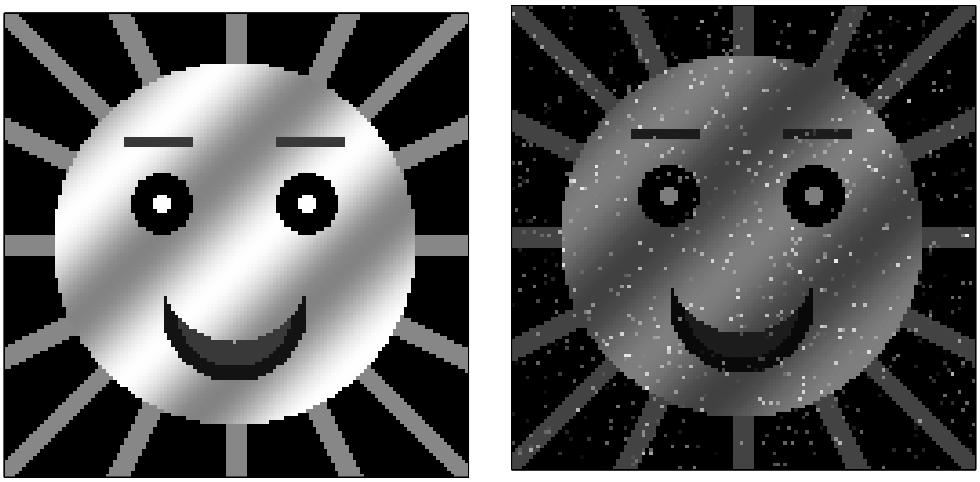



(a) Restoration $\hat{x}$ for $\alpha = 0.2$.



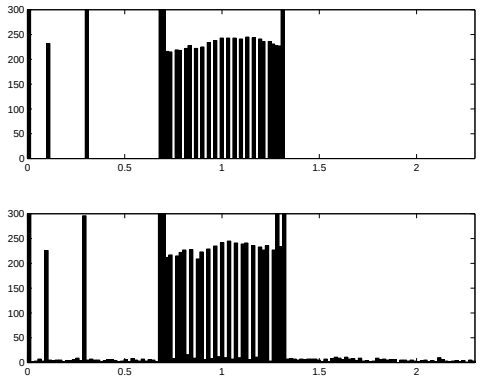

(b) Residuals $y - \hat{x}$.

FIG. 6. *Restoration involving nonsmooth regularization:* $\mathcal{F}$ *is as in* (75) *with* $\psi(t) = t^2$ *and* $\varphi(t) = |t|$ *for* $\alpha = 0.2$. *The minimizer* $\hat{x}$ *is constant over large regions.*

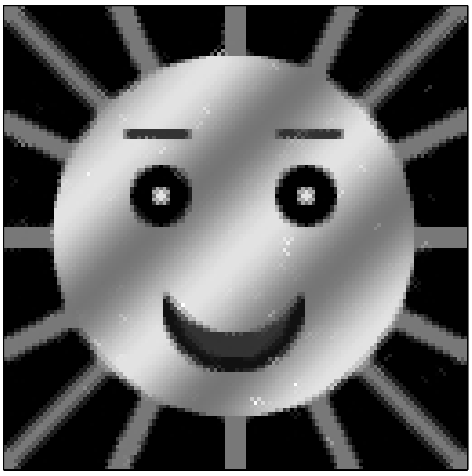


(a) Original image $x$. (b) Data $y = x + 770$ outliers.

FIG. 7. *Original image $x$ and data $y$ obtained by adding to $x$ 770 outliers with random location and random amplitude.*

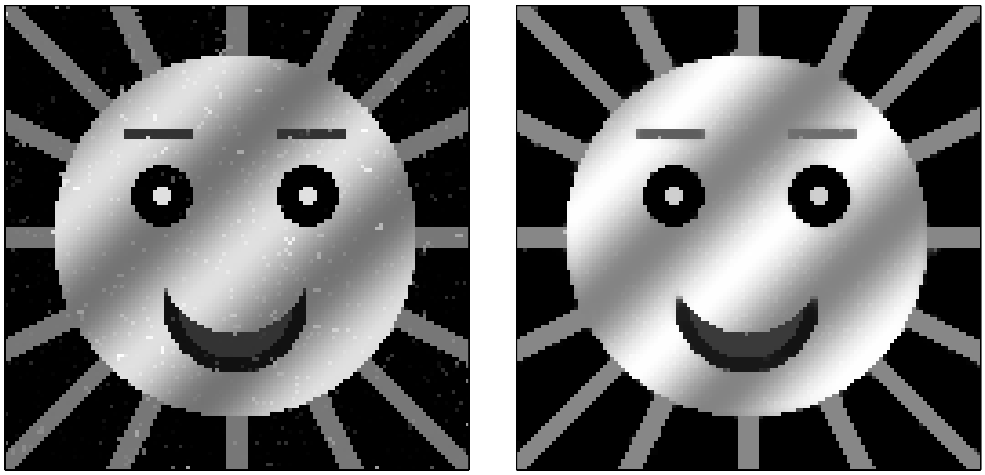



(a) Histograms: $x$ (up), $y$ (down).



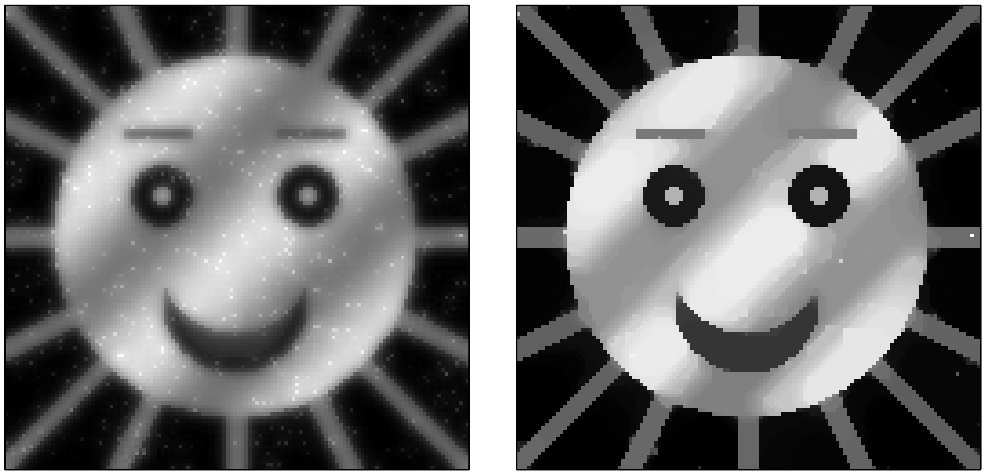
(b) Restoration by median filtering.

FIG. 8. (a) *Zoom of the histograms of the original $x$ (up) and of the data $y$ (down).* (b) *Restoration using two iterations of median filtering.*



(a) Minimizer obtained for $\alpha = 0.2$. (b) Minimizer calculated for $\alpha = 0.55$.

FIG. 9. *Minimizers obtained using the proposed cost-function $\mathcal{F}$ in* (77) *involving a nonsmooth data-fidelity term.* (a) *For $\alpha = 0.2$ there are* 720 *correct and* 65 *erroneous detections of outliers. Outliers are only weakly smoothed.* (b) *For $\alpha = 0.55$, outliers are well smoothed and the error is weak.*



(a) Smooth cost-function. (b) Nonsmooth regularization.

FIG. 10. *Minimizers obtained by minimizing $\mathcal{F}$ of the form* (75). (a) *For $\psi(t) = t^2 = \varphi(t)$ and $\alpha = 0.2$. Outliers are clearly seen, whereas edges are degraded.* (b) *For $\psi(t) = t^2$, $\varphi(t) = |t|$, and $\alpha = 0.4$. Only several outliers remain visible. Staircasing is clearly present.*

## REFERENCES


[1] R. Acar and C. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, IEEE Trans. Image Process., 10 (1994), pp. 1217–1229.

[2] S. Alliney, *Digital filters as absolute norm regularizers*, IEEE Trans. Medical Imaging, 12 (1993), pp. 173–181.

[3] S. Alliney, *A property of the minimum vectors of a regularizing functional defined by means of absolute norm*, IEEE Trans. Signal Process., 45 (1997), pp. 913–917.

[4] S. Alliney and S. A. Ruzinsky, *An algorithm for the minimization of mixed $l_1$ and $l_2$ norms with application to Bayesian estimation*, IEEE Trans. Signal Process., 42 (1994), pp. 618–627.

[5] A. Avez, *Calcul différentiel*, Masson, Paris, 1991.

[6] J. E. Besag, *On the statistical analysis of dirty pictures (with discussion)*, J. Roy. Statist. Soc. Ser. B, 48 (1986), pp. 259–302.

[7] M. Black and A. Rangarajan, *On the unification of line processes, outlier rejection, and robust statistics with applications to early vision*, Internat. J. Computer Vision, 19 (1996), pp. 57–91.

[8] B. Bloomfield and W. L. Steiger, *Least Absolute Deviations: Theory, Applications and Algorithms*, Birkhäuser, Boston, 1983.

[9] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Trans. Image Process., 2 (1993), pp. 296–310.

[10] C. Bouman and K. Sauer, *A unified approach to statistical tomography using coordinate descent optimization*, IEEE Trans. Image Process., 5 (1996), pp. 480–492.

[11] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math., 76 (1997), pp. 167–188.

[12] T. F. Chan and C. K. Wong, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.

[13] G. Demoment, *Image reconstruction and restoration: Overview of common estimation structure and problems*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 2024–2036.

[14] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1199.

[15] D. Donoho, I. Johnstone, J. Hoch, and A. Stern, *Maximum entropy and the nearly black object*, J. Roy. Statist. Soc. Ser. B, 54 (1992), pp. 41–81.

[16] S. Durand and M. Nikolova, *Stability of image restoration by minimizing regularized objective functions*, in Proceedings of the IEEE Int. Conf. on Computer Vision/Workshop on Variational and Level-Set Methods, Vancouver, Canada, 2001, pp. 73–80.

[17] D. Geman, *Random fields and inverse problems in imaging*, in École d'Été de Probabilités de Saint-Flour XVIII 1988, Lecture Notes in Math. 1427, Springer-Verlag, Berlin, 1990, pp. 117–193.

[18] D. Geman and G. Reynolds, *Constrained restoration and recovery of discontinuities*, IEEE Trans. Pattern Anal. Machine Intelligence, 14 (1992), pp. 367–383.

[19] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization*, IEEE Trans. Image Process., 4 (1995), pp. 932–946.

[20] S. Geman and D. McClure, *Statistical methods for tomographic image reconstruction*, in Proceedings of the 46th Session of the International Statistical Institute, Vol. 4 (Tokyo, 1987), Bull. Inst. Internat. Statist., 52 (1987), pp. 5–21.

[21] P. J. Green, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Trans. Medical Imaging, 9 (1990), pp. 84–93.

[22] T. Kailath, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, 20 (1974), pp. 146–181.

[23] A. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, NY, 1987.

[24] S. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, New York, 1995.

[25] K. S. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.

[26] M. Nikolova, *Estimées localement fortement homogènes*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 665–670.

[27] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658.

[28] M. Nikolova, *Weakly Constrained Minimization. Application to the Estimation of Images and Signals Involving Constant Regions*, Tech. report, CMLA—ENS de Cachan, France, 2001. Available online at http://www.cmla.ens-cachan.fr/Cmla/index.html

[29] P. Perona and J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intelligence, 12 (1990), pp. 629–639.
[30] T. T. Pham and R. J. P. De Figueiredo, *Maximum likelihood estimation of a class of non-Gaussian densities with application to $l_p$ deconvolution*, IEEE Trans. Signal Process., 37 (1989), pp. 73–82.
[31] J. R. Rice and J. S. White, *Norms for smoothing and estimation*, SIAM Rev., 6 (1964), pp. 243–256.
[32] R. T. Rockafellar and J. B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[33] L. Rudin, S. Osher, and C. Fatemi, *Nonlinear total variation based noise removal algorithm*, Phys. D, 60 (1992), pp. 259–268.
[34] K. Sauer and C. Bouman, *A local update strategy for iterative reconstruction from projections*, IEEE Trans. Signal Process., 41 (1993), pp. 534–548.
[35] A. Tarantola, *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier Science Publishers, Amsterdam, 1987.
[36] S. Teboul, L. Blanc-Féraud, G. Aubert, and M. Barlaud, *Variational approach for edge-preserving regularization using coupled PDE's*, IEEE Trans. Image Process., 7 (1998), pp. 387–397.
[37] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*, Winston, Washington, DC, 1977.
[38] C. R. Vogel and M. E. Oman, *Fast, robust total variation-based reconstruction of noisy, blurred images*, IEEE Trans. Image Process., 7 (1998), pp. 813–824.
[39] J. Weickert, *Anisotropic Diffusion in Image Processing*, B. G. Teubner, Stuttgart, 1998.