J Math Imaging Vis 27: 265–276, 2007

DOI: 10.1007/s10851-007-0650-0

On Semismooth Newton's Methods for Total Variation Minimization

MICHAEL K. NG*
Centre for Mathematical Imaging and Vision and Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
mng@math.hkbu.edu.hk

LIQUN QI†
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
maqilq@polyu.edu.hk

YU-FEI YANG‡
College of Mathematics and Econometrics, Hunan University, Changsha, China

YU-MEI HUANG
Centre for Mathematical Imaging and Vision and Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong and School of Information Science and Engineering, Lanzhou University, China

Published online: 7 March 2007

Abstract. In [2], Chambolle proposed an algorithm for minimizing the total variation of an image. In this short note, based on the theory on semismooth operators, we study semismooth Newton's methods for total variation minimization. The convergence and numerical results are also presented to show the effectiveness of the proposed algorithms.

Keywords: semismooth Newton's methods, total variation, denoising, regularization

1. Introduction

Recently, Chambolle [2] studied the following total variation denoising model [1, 3, 10]:

$$\min_u \ \mathrm{TV}(u) + \frac{1}{2\lambda}\|u - g\|_2^2, \tag{1}$$

where

$$TV(u) \equiv \int_\Omega |\nabla u(x, y)| dxdy,$$

g is the observed image in Ω, $\|\cdot\|_2$ denotes the norm in $\mathcal{L}^2(\Omega)$, and 2λ is the regularization parameter. He developed a semi-implicit gradient descent algorithm to solve (1) and also presented the convergence results for the descent algorithm.

In this note, based on the theory on semismooth operators, we further develop semismooth Newton's methods for total variation denoising algorithm for (1). The main contribution of this note is to show the convergence of the proposed method (see Section 2). Numerical results are presented in Section 3 to demonstrate the effectiveness of the proposed method.

*The research of this author is supported in part by Hong Kong Research Grants Council Grant Nos. 7035/04P and 7035/05P, and HKBU FRGs.
†The research of this author is supported in part by the Research Grant Council of Hong Kong.
‡This work was started while the author was visiting Department of Applied Mathematics, The Hong Kong Polytechnic University. The research of this author is supported in part by The Hong Kong Polytechnic University Postdoctoral Fellowship Scheme and the National Science Foundation of China (No. 60572114).

In the subsequent analysis, if $G : \mathcal{R}^n \to \mathcal{R}^m$ is differentiable at $x \in \mathcal{R}^n$ then $\nabla G(x)$ denotes the transposed Jacobian of G at x. If $G : \mathcal{R}^n \to \mathcal{R}^m$ is locally Lipschitz at $x \in \mathcal{R}^n$ then $\partial G(x)$ indicates the Clarke generalized Jacobian [4] of G at x, which is the convex hull of the following set

$$\partial_B G(x) := \left\{ \lim_{x^k \to x} \nabla G(x^k)^T \mid G \text{ is differentiable at } x^k \in \mathcal{R}^n \right\}.$$

If $G : \mathcal{R}^n \to \mathcal{R}^m$ is locally Lipschitz at $x \in \mathcal{R}^n$ and for any $h \in \mathcal{R}^n$,

$$\lim_{\substack{V \in \partial G(x+th') \\ h' \to h,\, t \downarrow 0}} \{Vh'\} \tag{2}$$

exists, then we say that G is semismooth at x. It follows from Theorem 2.3 in [9] that: if $G : \mathcal{R}^n \to \mathcal{R}^m$ is locally Lipschitz at $x \in \mathcal{R}^n$, then G is semismooth at x if and only if for any $V \in \partial G(x + h),\ h \to 0$, we have

$$Vh - G'(x; h) = o(\|h\|),$$

where $\|\cdot\|$ is the Euclidean norm. This property motivates the concept of the strong semismoothness: if G is locally Lipschitz at $x \in \mathcal{R}^n$ and for any $V \in \partial G(x + h),\ h \to 0$, we have

$$Vh - G'(x; h) = O(\|h\|^2),$$

then we call G is strongly semismooth at x.

2. The Main Results

2.1. Background Material

We first review some definitions and basic results from [2]. Then we introduce the concept of the semismooth operator, we refer reader to [8, 9] for details about its properties.

We assume that our images are matrices of size $N \times N$. Following the notations in [2], we let $\mathcal{X} = \mathcal{R}^{(N\times N)}$ and $\mathcal{Y} = \mathcal{X} \times \mathcal{X}$. In $\mathcal{X}$ and $\mathcal{Y}$, we use the Euclidean scalar product, defined in the standard way respectively by

$$\langle s, t\rangle_{\mathcal{X}} = \sum_{1\le i,j\le N} s_{i,j} t_{i,j}, \quad s, t \in \mathcal{X}$$

and

$$\langle p, q\rangle_{\mathcal{Y}} = \sum_{1\le i,j\le N} \left(p^1_{i,j} q^1_{i,j} + p^2_{i,j} q^2_{i,j}\right), \quad p = (p^1, p^2), q = (q^1, q^2) \in \mathcal{Y}.$$

The discrete gradient operator $\nabla : \mathcal{X} \to \mathcal{Y}$ is defined by

$$(\nabla u)_{i,j} = \left((\nabla u)^1_{i,j}, (\nabla u)^2_{i,j}\right)$$

with

$$(\nabla u)^1_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } i < N, \\ 0 & \text{if } i = N, \end{cases}$$

$$(\nabla u)^2_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } i < N, \\ 0 & \text{if } i = N \end{cases}$$

for $i, j = 1, \ldots, N$. The discrete total variation is defined by

$$J(u) = \sum_{1\le i,j\le N} |(\nabla u)_{i,j}|.$$

Therefore, the discretization of minimization problem (1) is given by

$$\min_{u\in\mathcal{X}} J(u) + \frac{1}{2\lambda}\|u - g\|^2_{\mathcal{X}}, \tag{3}$$

where $u, g \in \mathcal{X}$ are discretization vectors of related continuous variables, $\lambda > 0$ and $\|\cdot\|_{\mathcal{X}}$ is the Euclidean norm in $\mathcal{X}$, given by $\|u\|_{\mathcal{X}} = \langle u, u\rangle_{\mathcal{X}}$.

It is shown in [2] that the solution u of problem (3) is given by

$$u = g - \pi_{\lambda\mathcal{K}}(g), \tag{4}$$

where $\pi_{\lambda\mathcal{K}}$ is the orthogonal projection of g onto the convex set $\lambda\mathcal{K}$,

$$\mathcal{K} := \{\text{div} p \ :\ p \in \mathcal{Y},\ |p_{i,j}| \le 1\ \forall i,\ j = 1, \ldots, N\}$$

and the discrete divergence operator $\text{div} : \mathcal{Y} \to \mathcal{X}$ is defined by $\text{div} = -\nabla^*$, that is, for any $p \in \mathcal{Y}$ and $u \in \mathcal{X}$, $\langle -\text{div}\, p, u\rangle_{\mathcal{X}} = \langle p, \nabla u\rangle_{\mathcal{Y}}$. It is clear that div

is given by

$$(\operatorname{div} p)_{ij} = \begin{cases} p^1_{i,j} - p^1_{i-1,j} & \text{if } 1 < i < N, \\ p^1_{i,j} & \text{if } i = 1, \\ -p^1_{i-1,j} & \text{if } i = N, \end{cases} + \begin{cases} p^2_{i,j} - p^2_{i,j-1} & \text{if } 1 < j < N, \\ p^2_{i,j} & \text{if } j = 1, \\ -p^2_{i,j-1} & \text{if } j = N, \end{cases}$$

for any $p \in \mathcal{Y}$. Hence, computing the solution of (3) hinges on computing the nonlinear projection $\pi_{\lambda\mathcal{K}}$. The latter amounts to solving the constrained minimization problem

$$\begin{aligned} &\min \|\lambda \operatorname{div} p - g\|^2_{\mathcal{X}} \quad \text{subject to } p \in \mathcal{Y}, \quad \text{and} \\ &|p_{i,j}|^2 - 1 \le 0 \ \ \forall i, \quad j = 1, \dots, N. \end{aligned} \tag{5}$$

By the optimal condition of the above problem, there exists a Lagrange multiplier $\eta \in \mathcal{X}$ such that the following KKT system holds:

$$\begin{aligned} &\text{For each } \ i, j = 1, \dots, N \\ &\quad -(\nabla(\lambda \operatorname{div} p - g))_{i,j} + \eta_{i,j} p_{i,j} = 0, \\ &|p_{i,j}|^2 - 1 \le 0, \quad \eta_{i,j} \ge 0, \quad \text{and} \quad \eta_{i,j}(|p_{i,j}|^2 - 1) = 0. \end{aligned} \tag{6}$$

Let us set $M := N^2$. For the sake of convenience, for a vector $p = (p^1, p^2) \in \mathcal{Y}$, we array first p^1 then p^2 row-wise in a vector $v \in \mathcal{R}^{2M}$. Similarly, for a vector $\eta \in \mathcal{X}$, we array it row-wise in a vector $\xi \in \mathcal{R}^M$. In this way, we can equivalently rewrite (6) as the following system:

$$\begin{aligned} &\text{For each } \ k = 1, \dots, M \\ &-(f_k(v), f_{M+k}(v)) + \xi_k(v_k, v_{M+k}) = 0, \\ &1 - |(v_k, v_{M+k})|^2 \ge 0, \quad \xi_k \ge 0, \quad \text{and} \\ &\xi_k(1 - |(v_k, v_{M+k})|^2) = 0, \end{aligned} \tag{7}$$

where $f : \mathcal{R}^{2M} \to \mathcal{R}^{2M}$ is defined by

$$\begin{aligned} f_k(v) &:= (\nabla(\lambda \operatorname{div} p - g))^1_{i,j} \ \text{ and} \\ f_{M+k}(v) &:= (\nabla(\lambda \operatorname{div} p - g))^2_{i,j} \end{aligned}$$

for $1 \le k \le M$ with $k = N(i-1) + j$ and $1 \le i, j \le N$.

It is well known that the Fisher-Burmeister function $\phi : \mathcal{R}^2 \to \mathcal{R}$, defined by

$$\phi(a, b) = \sqrt{a^2 + b^2} - a - b,$$

possesses the following characterization:

$$\phi(a, b) = 0 \iff a \ge 0, \ b \ge 0 \text{ and } ab = 0.$$

By using the above property, it is easy to deduce that solving the system (7) is equivalent to solving the system of nonsmooth equations

$$\Phi(z) = 0,$$

where $z := (v, \xi) \in \mathcal{R}^{2M} \times \mathcal{R}^M$ and $\Phi : \mathcal{R}^{3M} \to \mathcal{R}^{3M}$ is defined by

$$\begin{aligned} \Phi_k(z) &:= -f_k(v) + \xi_k v_k, \\ \Phi_{M+k}(z) &:= -f_{M+k}(v) + \xi_k v_{M+k}, \\ \Phi_{2M+k}(z) &:= \phi\big(1 - v_k^2 - v_{M+k}^2, \xi_k\big) \end{aligned} \tag{8}$$

for $1 \le k \le M$.

On the other hand, it is easy to deduce from (7) that

$$\xi_k = |(f_k(v), f_{M+k}(v))| \quad \forall k = 1, \dots, M.$$

Therefore, (7) can be rewritten equivalently as the nonsmooth overdetermined system

$$H(v) = 0,$$

where $H : \mathcal{R}^{2M} \to \mathcal{R}^{3M}$ is defined by

$$\begin{aligned} H_k(v) &:= -f_k(v) + |(f_k(v), f_{M+k}(v))| v_k, \\ H_{M+k}(v) &:= -f_{M+k}(v) + |(f_k(v), f_{M+k}(v))| v_{M+k}, \\ H_{2M+k}(v) &:= \phi\big(1 - v_k^2 - v_{M+k}^2, \big|(f_k(v), f_{M+k}(v))\big|^2\big) \end{aligned} \tag{9}$$

for $1 \le k \le M$. Note that in the definition of H_{2M+k}, $|(f_k(v), f_{M+k}(v))|^2$, instead of $|(f_k(v), f_{M+k}(v))|$, is used to replace ξ_k in Φ_{2M+k}.

Here we give two remarks of the above formulation. The detailed explanation will be given in the next subsection.

Remark 1. If a function $\varphi : \mathcal{R}^2 \to \mathcal{R}$ possesses the following property:

$$\varphi(a, b) = 0 \iff a \ge 0, \ b \ge 0 \text{ and } ab = 0,$$

then φ is called a NCP function. In fact, any NCP function can reformulate the complementarity condition in (5) as a system of nonsmooth equations.

The Fisher-Burmeister function ϕ is a NCP function and is strongly semismooth, this property can guarantee the functions Φ and H to be strongly semismooth so that the generalized Newton methods for solving nonsmooth systems $\Phi(z) = 0$ and $H(v) = 0$ possess fast locally (quadratically) convergent rate.

Remark 2. Another important property of the Fisher-Burmeister function ϕ is that ϕ^2 is continuously differentiable on the whole space $\mathcal{R}$, this property can guarantee $\frac{1}{2}\|\Phi\|^2$ and $\frac{1}{2}\|H\|^2$ to be continuously differentiable so that the damped modified Gauss-Newton methods for solving nonsmooth systems $\Phi(z) = 0$ and $H(v) = 0$ are globally convergent. However, not every one of NCP functions possesses this property. For example, the function $\varphi_{min}(a, b) := \min\{a, b\}$ is a NCP function and is strongly semismooth, but it is easy to prove that $\varphi^2_{min}(a, b)$ is not continuously differentiable whenever $a = b \neq 0$.

2.2. Semismooth Newton-Type Methods

In this subsection, we first prove that Φ and H are strongly semismooth with smooth least squares formulation, and then give an approach to estimate their generalized Jacobians. Finally we present two semismooth Newton-type methods for the solution of total variation minimization problem (1) and give their convergence properties. Our methods are based on the modified Gauss-Newton methods for solving nonsmooth systems $\Phi(z) = 0$ and $H(v) = 0$.

Proposition 1. *Φ and H are strongly semismooth on $\mathcal{R}^{3M}$ and $\mathcal{R}^{2M}$ respectively.*

Proof: Since ϕ is strongly semismooth, the result follows from the fact that composite functions of strongly semismooth functions are also strongly semismooth. □

Next we give an approach to compute the expression of elements in $\partial\Phi(z)^T$ and $\partial H(v)^T$. By using the rules on the evaluation of the generalized Jacobian, we have

$$\partial\Phi(z)^T \subseteq \partial\Phi_1(z)^T \times \cdots \times \partial\Phi_{3M}(z)^T.$$

Therefore, we are only required to compute $\partial\Phi_l(z)^T$ and $\partial H_l(v)^T$ for each l. It is obvious that $\partial\Phi_l(z)^T = \nabla\Phi_l(z)$ whenever $\Phi_l(z)$ is differentiable.

Let us consider how to compute $\partial\Phi_l(z)^T$. For each $k = 1, \ldots, M$, $\Phi_k(z)$ and $\Phi_{M+k}(z)$ are obviously differentiable and

$$\nabla\Phi_k(z) = -\nabla_z f_k(v) + \xi_k e_k + v_k e_{2M+k},$$
$$\nabla\Phi_{M+k}(z) = -\nabla_z f_{M+k}(v) + \xi_k e_{M+k} + v_{M+k} e_{2M+k},$$

where e_l indicates the l-th column of the $3M \times 3M$ identity matrix, and

$$\nabla_z f_l(v) := \left(\nabla f_l(v)^T, 0_M^T\right)^T$$

where 0_M is the zero vector in $\mathcal{R}^M$. If $k \in \{1, \ldots, M\}$ is such that $|(v_k, v_{M+k})| \neq 1$ or $\xi_k \neq 0$, then $\Phi_{2M+k}(z)$ is also differentiable and

$$\nabla\Phi_{2M+k}(z) = -2(\rho_{2M+k} - 1)(v_k e_k + v_{M+k} e_{M+k}) + (\gamma_{2M+k} - 1)e_{2M+k}$$

with

$$\rho_{2M+k} = \frac{1 - v_k^2 - v_{M+k}^2}{\sqrt{\left(1 - v_k^2 - v_{M+k}^2\right)^2 + \xi_k^2}} \quad \text{and}$$
$$\gamma_{2M+k} = \frac{\xi_k}{\sqrt{\left(1 - v_k^2 - v_{M+k}^2\right)^2 + \xi_k^2}}.$$

If $k \in \{1, \ldots, M\}$ is such that $|(v_k, v_{M+k})| = 1$ and $\xi_k = 0$, by using the theorem on the generalized gradient of a composite function, we deduce

$$\partial\Phi_{2M+k}(z)^T = \{-2(\rho_{2M+k} - 1)(v_k e_k + v_{M+k} e_{M+k}) + (\gamma_{2M+k} - 1)e_{2M+k} : |(\rho_{2M+k}, \gamma_{2M+k})| \leq 1\}.$$

Now we compute $\partial H_l(v)^T$. If $k \in \{1, \ldots, M\}$ is such that $|(f_k(v), f_{M+k}(v))| \neq 0$, then $H_k(v)$ and $H_{M+k}(v)$ are obviously differentiable and

$$\nabla H_k(v) = -\nabla f_k(v) + (\rho_k \nabla f_k(v) + \gamma_k \nabla f_{M+k}(v))v_k + |(f_k(v), f_{M+k}(v))|\hat{e}_k,$$
$$\nabla H_{M+k}(v) = -\nabla f_{M+k}(v) + (\rho_{M+k}\nabla f_k(v) + \gamma_{M+k}\nabla f_{M+k}(v))v_{M+k} + |(f_k(v), f_{M+k}(v))|\hat{e}_{M+k}$$

with

$$\rho_k = \rho_{M+k} = \frac{f_k(v)}{|(f_k(v), f_{M+k}(v))|} \quad \text{and}$$
$$\gamma_k = \gamma_{M+k} = \frac{f_{M+k}(v)}{|(f_k(v), f_{M+k}(v))|},$$

where $\hat{e}_l$ indicates the l-th column of the $2M \times 2M$ identity matrix. If $k \in \{1, \ldots, M\}$ is such that $|(f_k(v), f_{M+k}(v))| = 0$, we deduce

$$\begin{aligned}\partial H_k(v)^T = \{&-\nabla f_k(v) + (\rho_k \nabla f_k(v) + \gamma_k \nabla f_{M+k}(v))v_k \\ &+ |(f_k(v), f_{M+k}(v))|\hat{e}_k : |(\rho_k, \gamma_k)| \le 1\},\\ \partial H_{M+k}(v)^T = \{&-\nabla f_{M+k}(v) + (\rho_{M+k}\nabla f_k(v) \\ &+ \gamma_{M+k}\nabla f_{M+k}(v))v_{M+k} \\ &+ |(f_k(v), f_{M+k}(v))|\hat{e}_{M+k} : \\ &|(\rho_{M+k}, \gamma_{M+k})| \le 1\}.\end{aligned}$$

If $k \in \{1, \ldots, M\}$ is such that $|(v_k, v_{M+k})| \neq 1$ or $|(f_k(v), f_{M+k}(v))| \neq 0$, then $H_{2M+k}(v)$ is also differentiable and

$$\begin{aligned}\nabla H_{2M+k}(v) = &-2(\rho_{2M+k} - 1)(v_k\hat{e}_k + v_{M+k}\hat{e}_{M+k}) \\ &+ 2(\gamma_{2M+k} - 1)(f_k(v)\nabla f_k(v) \\ &+ f_{M+k}(v)\nabla f_{M+k}(v))\end{aligned}$$

with

$$\rho_{2M+k} = \frac{1 - v_k^2 - v_{M+k}^2}{\sqrt{\left(1 - v_k^2 - v_{M+k}^2\right)^2 + |(f_k(v), f_{M+k}(v))|^4}}$$

and

$$\gamma_{2M+k} = \frac{|(f_k(v), f_{M+k}(v))|^2}{\sqrt{\left(1 - v_k^2 - v_{M+k}^2\right)^2 + |(f_k(v), f_{M+k}(v))|^4}}.$$

If $k \in \{1, \ldots, M\}$ is such that $|(v_k, v_{M+k})| = 1$ and $|(f_k(v), f_{M+k}(v))| = 0$, we have

$$\begin{aligned}\partial H_{2M+k}(v)^T = \{&-2(\rho_{2M+k} - 1)(v_k\hat{e}_k + v_{M+k}\hat{e}_{M+k}) \\ &+ 2(\gamma_{2M+k} - 1)(f_k(v)\nabla f_k(v) \\ &+ f_{M+k}(v)\nabla f_{M+k}(v)) \\ &: |(\rho_{2M+k}, \gamma_{2M+k})| \le 1\}.\end{aligned}$$

Based on the above analysis, the semismooth Newton method developed in [8] can be applied to the system of semismooth equations $\Phi(z) = 0$: Given an iterate z^k, one can compute a direction d^k by solving the generalized Newton equation

$$V_k^T d + \Phi(z^k) = 0, \tag{10}$$

where $V_k \in \partial_B \Phi(z^k)^T$; then let $z^{k+1} = z^k + d^k$. In order to globalize this method, a line search technique is used to achieve a sufficient decrease of the natural merit function

$$\Psi(z) := \frac{1}{2}\|\Phi(z)\|^2.$$

Notice that the generalized Newton Eq. (10) is required to be solvable. To overcome this drawback, the modified Gauss-Newton method used in solving systems of smooth equations can be generalized to $\Phi(z) = 0$: Given an iterate z^k, one can compute a direction d^k by solving the modified Gauss-Newton equation

$$V_k\Phi(z^k) + \left(V_k V_k^T + \nu_k I\right)d = 0, \tag{11}$$

where $\nu_k > 0$ is a fixed constant.

The next proposition plays a crucial role in convergence analysis of algorithms presented subsequently.

Proposition 2. *The following statements hold. (i) Ψ is continuously differentiable on $\mathcal{R}^{3M}$ and $\nabla\Psi(z) = \partial\Phi(z)^T\Phi(z)$. (ii) Let $\theta : \mathcal{R}^{2M} \to \mathcal{R}$ be defined by $\theta(v) = \frac{1}{2}\|H(v)\|^2$. Then θ is continuously differentiable on $\mathcal{R}^{2M}$ and $\nabla\theta(v) = \partial H(v)^T H(v)$.*

Proof: (i) By using the calculus of generalized gradients, we have

$$\partial\Psi(z) = \partial\Phi(z)^T\Phi(z) = \sum_{l=1}^{3M} \partial\Phi_l(z)^T\Phi_l(z).$$

It follows from the above analysis that: if l is such that Φ_l is not differentiable at $z \in \mathcal{R}^{3M}$, then it holds that $\Phi_l(z) = 0$ and hence $\partial\Phi_l(z)^T\Phi_l(z) = 0$. Therefore, $\partial\Psi(z)$ is single valued everywhere. The assertion then follows from the corollary to Theorem 2.2.4 in [4].

Figure 1. The original image.

(ii) Similarly, we have

$$\partial\theta(v) = \partial H(v)^T H(v) = \sum_{l=1}^{3M} \partial H_l(v)^T H_l(v).$$

If l is such that H_l is not differentiable at $v \in \mathcal{R}^{2M}$, then $H_l(v) = 0$ and hence we obtain $\partial H_l(v)^T H_l(v) = 0$. Therefore, $\partial\theta(v)$ is single valued everywhere, which implies that the assertion holds. □

Now we explain why we use $|(f_k(v), f_{M+k}(v))|^2$ in the definition of H_{2M+k} in (9). Assume that we use $|(f_k(v), f_{M+k}(v))|$ as usual, i.e., we define

$$\hat{H}_{2M+k}(v) := \phi\big(1 - v_k^2 - v_{M+k}^2, |(f_k(v), f_{M+k}(v))|\big) \quad \forall k = 1, \ldots, M.$$

If $k \in \{1, \ldots, M\}$ is such that $|(f_k(v), f_{M+k}(v))| = 0$, then $\hat{H}_{2M+k}$ must be not differentiable at $v \in \mathcal{R}^{2M}$. However, when $|(f_k(v), f_{M+k}(v))| = 0$ but $|(v_k, v_{M+k})| > 1$, $\hat{H}_{2M+k}(v) = 2|(v_k, v_{M+k})|^2 - 2 > 0$ such that $\partial\hat{H}_{2M+k}(v)^T \hat{H}_{2M+k}(v)$ is not single valued, which implies that one cannot guarantee that the least squares of the function defined above is continuously differentiable everywhere.

In the following we summarize the damped modified Gauss-Newton method for solving $\Phi(z) = 0$.

Algorithm 1. (Damped Modified Gauss-Newton Method I)

Step 0. Choose $\sigma, \rho \in (0, 1)$, $\nu_0 > 0$, and a starting point $z^0 \in R^{3M}$. Set $k := 0$.

Step 1. Choose $V_k \in \partial_B \Phi(z^k)^T$ and solve the modified Gauss-Newton Eq. (11). Let d^k be the solution of (11). If $d^k = 0$, the algorithm terminates.

Step 2. Let $\lambda_k = \rho^{i_k}$ where i_k is the smallest nonnegative integer i such that

$$\Psi(z^k + \rho^i d^k) - \Psi(z^k) \le \sigma\rho^i \nabla\Psi(z^k)^T d^k.$$

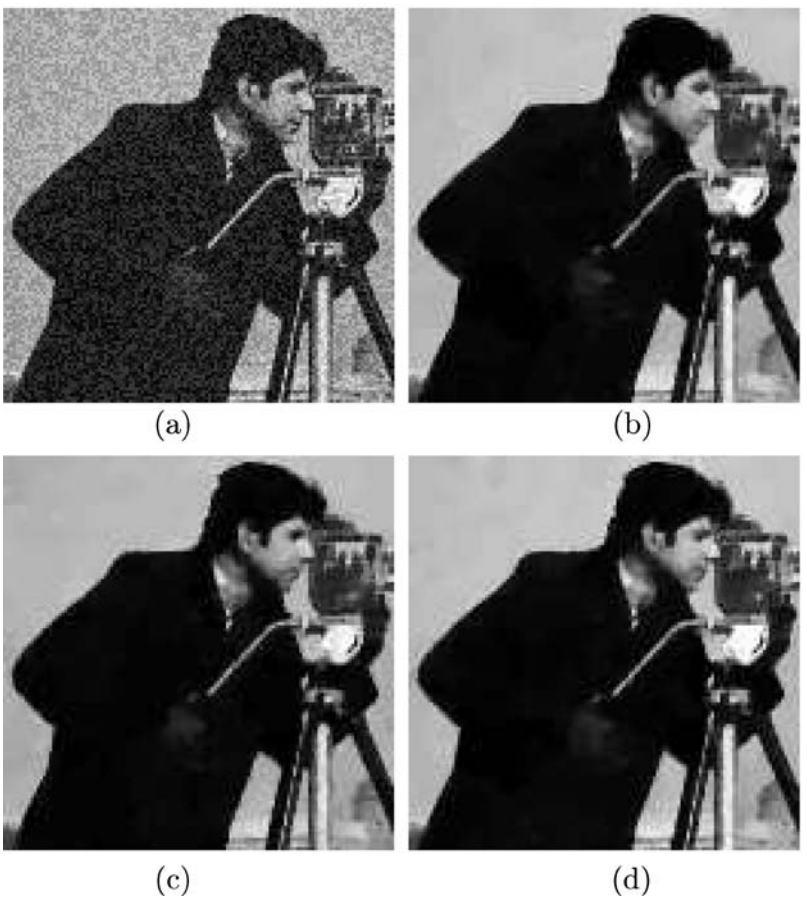


Figure 2. (a) The noisy image; (b) the denoised image by the Chambolle algorithm, (c) Algorithm 1 and (d) Algorithm 2 for a uniform noise with $\sigma = 70$ and $\epsilon = 0.001$. Here the chosen regularization parameter λ is 18.

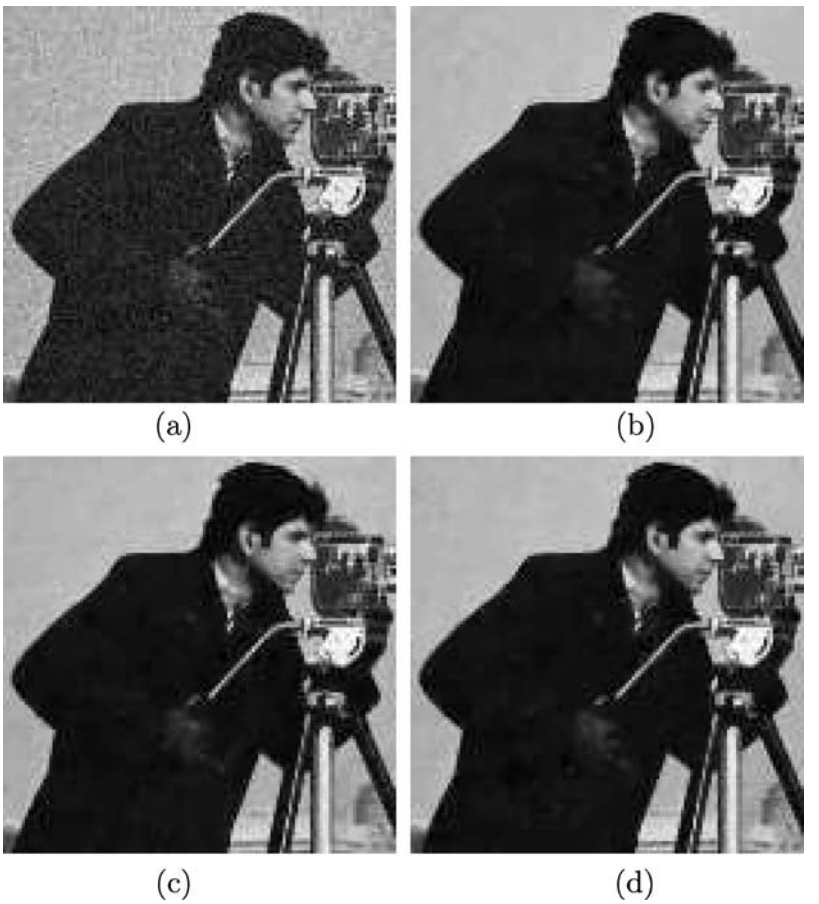


Figure 3. (a) The noisy image; (b) the denoised image by the Chambolle algorithm, (c) Algorithm 1 and (d) Algorithm 2 for a Gaussian white noise with $\sigma = 12$ and $\epsilon = 0.001$. Here the chosen regularization parameter λ is 8.

Step 3. Choose $\nu_{k+1} > 0$. Let $z^{k+1} := z^k + \lambda_k d^k$ and $k := k + 1$. Go to Step 1.

It is obvious that: if $d^k = 0$, then z^k must be a stationary point of Ψ. Therefore, we assume that Algorithm 1 does not terminate in finitely many steps. Based on Propositions 1 and 2, we have the following convergence results whose proof can be referred to [6].

Theorem 1. *Assume that z^* is an accumulation point of $\{z^k\}$ generated by Algorithm 1. Then the following assertions hold.*

(i) *z^* is a stationary point of Ψ if both $\{\nu_k\}$ and $\{d^k\}$ are bounded. Moreover, z^* is a solution of $\Phi(z) = 0$ if there exists a nonsingular element in $\partial\Phi(z^*)$.*
(ii) *z^* is a stationary point of Ψ if $\bar{\nu} > \nu_k > \underline{\nu}$ for some $\bar{\nu} > \underline{\nu} > 0$.*
(iii) *Let $\nu_k = \min\{\Psi(z^k), \|\nabla\Psi(z^k)\|\}$. Then z^* is a stationary point of Ψ; furthermore, z^* is a solution of $\Phi(z) = 0$ and $\{z^k\}$ converges to z^* Q-quadratically if $\sigma \in (0, \frac{1}{2})$ and any element in $\partial_B\Phi(z^*)$ is nonsingular.*

As was shown in the previous section, solving (7) is also equivalent to solving the nonsmooth overdetermined system $H(v) = 0$. Similar to that in solving $\Phi(z) = 0$, we can also state a damped modified Gauss-Newton method for solving $H(v) = 0$ as follows.

Algorithm 2. (Damped Modified Gauss-Newton Method II)

Step 0. Choose $\sigma, \rho \in (0, 1)$, $\nu_0 > 0$, and a starting point $v^0 \in R^{2M}$. Set $k := 0$.

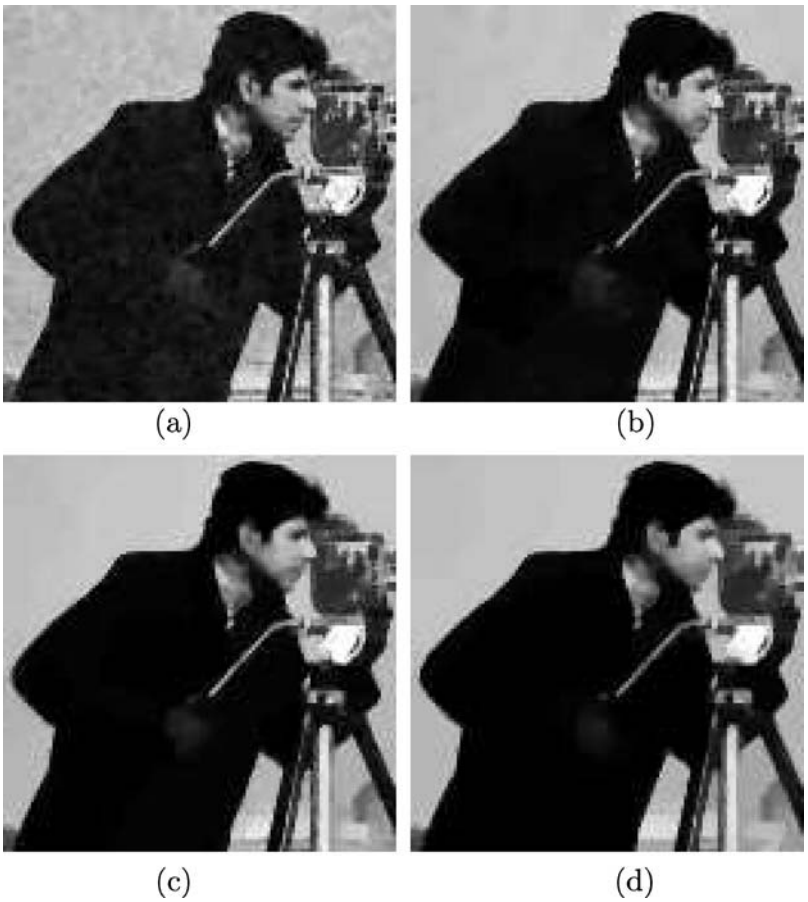


Figure 4. The denoised images when (a) $\lambda = 10$, (b) $\lambda = 18$, (c) $\lambda = 25$ and (d) $\lambda = 30$ for the case of uniform noise in Fig. 2.

Step 1. Choose $U_k \in \partial_B H(v^k)^T$ and solve the modified Gauss-Newton equation

$$U_k H(v^k) + \left(U_k U_k^T + \nu_k \hat{I}\right)s = 0. \tag{12}$$

Let s^k be the solution of (12). If $s^k = 0$, the algorithm terminates.

Step 2. Let $\lambda_k = \rho^{i_k}$ where i_k is the smallest nonnegative integer i such that

$$\theta(v^k + \rho^i s^k) - \theta(v^k) \leq \sigma \rho^i \nabla\theta(v^k)^T s^k.$$

Step 3. Choose $\nu_{k+1} > 0$. Let $v^{k+1} := v^k + \lambda_k s^k$ and $k := k + 1$. Go to Step 1.

The main difference between Algorithms 1 and 2 lies in that: at each iteration, Algorithm 2 only requires to solve a system of linear equations of dimension $2M$ instead of dimension $3M$ in Algorithm 1 so that the computational amount in Algorithm 2 is smaller than that in Algorithm 1. Similarly, we have also the following convergence results.

Theorem 2. *Assume that v^* is an accumulation point of $\{v^k\}$ generated by Algorithm 2. Then the following assertions hold.*

(i) *v^* is a stationary point of θ if both $\{\nu_k\}$ and $\{s^k\}$ are bounded.*
(ii) *v^* is a stationary point of θ if $\bar{\nu} > \nu_k > \underline{\nu}$ for some $\bar{\nu} > \underline{\nu} > 0$.*
(iii) *Let $\nu_k = \min\{\theta(v^k), \|\nabla\theta(v^k)\|\}$. Then v^* is a stationary point of θ; furthermore, $\{v^k\}$ converges to v^* Q-quadratically if v^* is a solution of $H(v) = 0$, $\sigma \in (0, \frac{1}{2})$ and every element in $\partial_B H(v^*)$ has full column rank.*

Similar to the proof of Theorem 3.1 in [2], we deduce the following results.

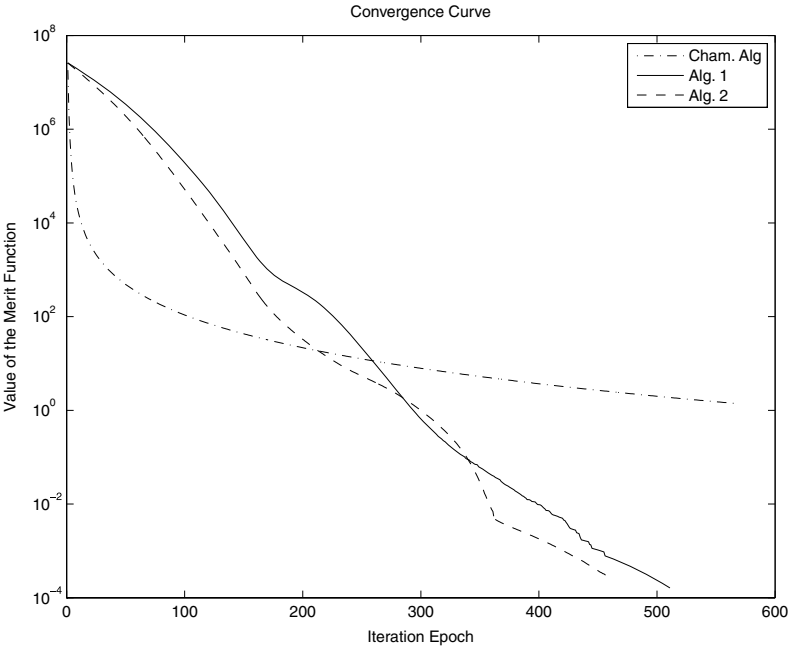


Figure 5. The convergence curves of different algorithms for the case of uniform noise in Fig. 2.

Theorem 3. *The following statements hold.*

(i) *Let z^* be an accumulation point of $\{z^k\}$ and $\{p^k\} \subset \mathcal{Y}$ be the sequence corresponding to the v-component of $\{z^k\}$. Assume $\Phi(z^*) = 0$. Then $\lambda\, div\ p^k$ converges to $\pi_{\lambda\mathcal{K}}(g)$ as $k \to \infty$.*

(ii) *Let v^* be an accumulation point of $\{v^k\}$ and $\{p^k\} \subset \mathcal{Y}$ be the sequence corresponding to $\{v^k\}$. Assume $H(v^*) = 0$. Then $\lambda\, div\ p^k$ converges to $\pi_{\lambda\mathcal{K}}(g)$ as $k \to \infty$.*

3. Numerical Results

In this section, we illustrate the effectiveness of the proposed algorithms for total variation minimization. In the computer simulation, a 128×128 image was taken to be the original image (see Fig. 1). A Gaussian white noise with MATLAB command "$\sigma *$randn(128,128)" or a uniform noise with MATLAB command "$\sigma *$rand(128,128)" is added to the original image.

Figures 2 and 3 show the noisy images and their denoised images using the Chambolle algorithm, our Algorithms 1 and 2 respectively. The criterion for stopping the iteration of all three algorithms consists in checking that the maximum variation between $p_{i,j}^{(n)}$ and $p_{i,j}^{(n+1)}$ is less than ϵ. According to the figures, the denoised images look alike. We also see from the figures that the total variation minimization is quite effective in image denoising. For the selection of regularization parameter λ, it is chosen (by trial and error) to minimize the relative error between the denoised image and the original image. We show in Fig. 4 for the case of uniform noise that the chosen regularization parameter also gives a denoised image with good visual quality.

Figures 5 and 6 show the convergence of the Chambolle algorithm, our Algorithms 1 and 2 corresponding to the denoised images in Figs. 2 and 3 respectively. All the algorithms use the zero vector as an initial guess. According to Figs. 5 and 6, it is clear that the proposed algorithm converges much faster. For instance, within first 500 iterations, the Chambolle algorithm can reduce the value of the merit function from about 10^8 to about 10^0, but Algorithms 1 and 2 can reduce the value from about 10^8 to about 10^{-4}. We further check that the Chambolle algorithm needs about 18000 iterations in order to reduce the value to 10^{-4}. In Fig. 7, we also test the convergence performance of Algorithm 2 for different parameters λ as we showed in Fig. 4. When λ is large, the total variation term is dominant. It implies the denosing problem is more ill-conditioned and therefore the number of iterations required for convergence would be more. We remark that Algorithm 1 has the

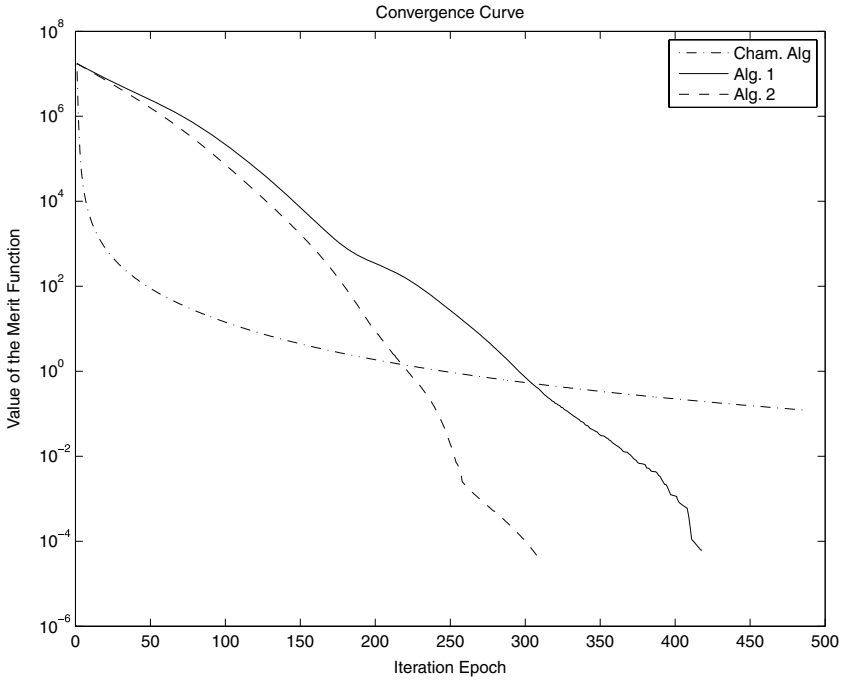


Figure 6. The convergence curves of different algorithms for the case of Gaussian noise in Fig. 3.

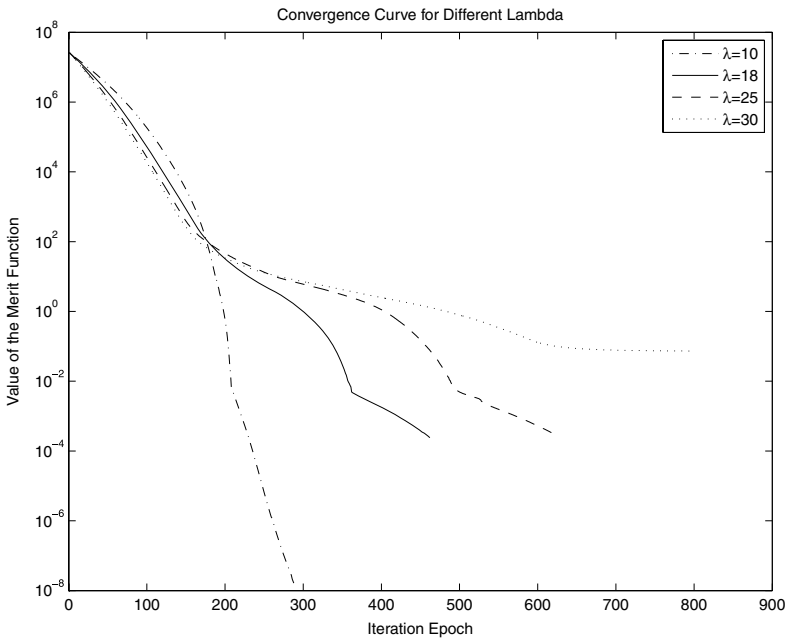


Figure 7. The convergence curves of different λ in Algorithm 2 for the case of uniform noise in Fig. 2.

similar convergence performance for different values of λ.

Next we test the convergence performance of different algorithms when a good initial guess is used. Here we set the initial guess to be the sum of the solution of Chambolle algorithm in Fig. 2 and a noise where it is given by $0.05 \times \sin(\text{rand}(128,128))$. We see from Fig. 8 that the convergence of our Algorithms 1 and 2

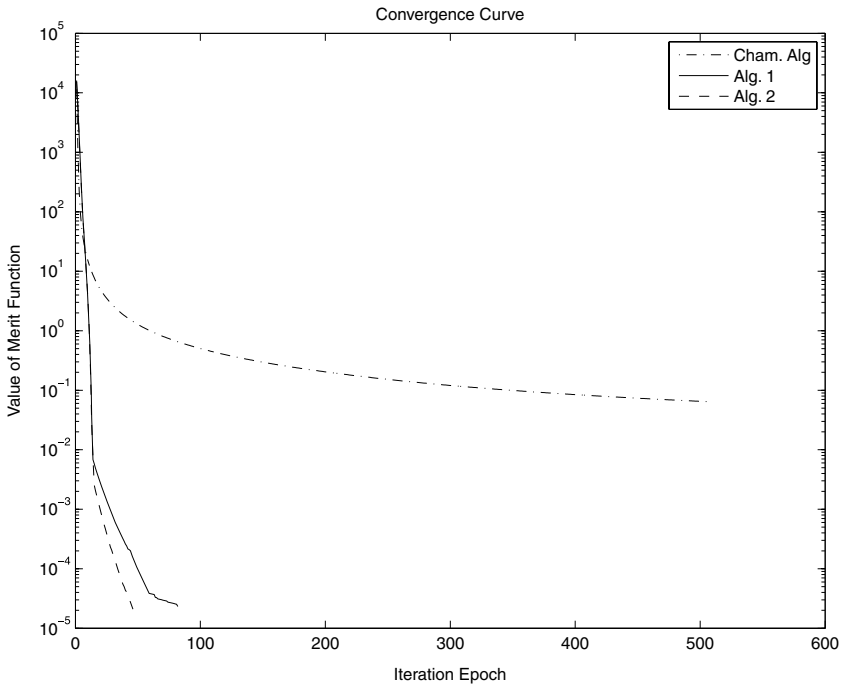


Figure 8. The convergence curves of different algorithms for the case of uniform noise in Fig. 2 when a better initial guess is used.

(Newton's methods) is much faster, while the convergence performance of Chambolle algorithm is not improved significantly.

Finally, we consider the computational cost of our method. The sizes of the linear systems required to be solved in (10) and (11) are 32768-by-32768 and 49152-by-49152 respectively. However, both coefficient matrices are sparse matrices. In our experiments, we check that the percentage of the number of nonzero entries is about 0.13%. By using sparse solvers in MATLAB under Intel Pentium IV 3.2GHz, the computational time for solving (10) (or (11)) is about 2.9 (or 3.2) seconds. We also remark that the computational time of the Chambolle algorithm per iteration is about 0.05 seconds. The cost per iteration of the Chambolle algorithm is less expensive than that of our method. In order to reduce the computational time of each iteration of our method, we consider and study preconditioning techniques [7] for the linear systems in (10) and (11). With the use of effective preconditioning techniques, the Newton method will be more competitive with the Chambolle algorithm.

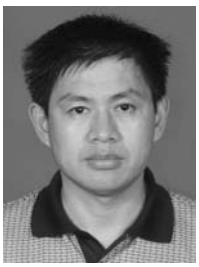
Finally, we summarize this short note that semismooth Newton's methods for total variation minimization are studied and their convergence results are also presented. Numerical results show that the proposed methods are quite competitive.

References

1. R. Acar and C.R. Vogel, "Analysis of bounded variation penalty methods for ill-posed problems," *Inverse Problems*, Vol. 10, pp. 1217–1229, 1994.
2. A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, Vol. 20, pp. 89–97, 2004.
3. T.F. Chan, G.H. Golub, and P. Mulet, "A nonlinear primal-dual method for total variation-based image restoration," *SIAM J. Sci. Comput.*, Vol. 20, pp. 1964–1977, 1999.
4. F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley: New York, 1983.
5. G.H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, 1996.
6. H. Jiang and D. Ralph, "Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares," in *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, M. Fukushima and L. Qi (eds.), Kluwer Acad. Publ., Dordrecht, 1998, pp. 181–209.
7. G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, 1996.
8. L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations," *Math. Oper. Res.*, Vol. 18, pp. 227–244, 1993.
9. L. Qi and J. Sun, "A Nonsmooth version of Newton's method," *Mathematical Programming*, Vol. 58, pp. 353–367, 1993.
10. L. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithm," *Physica D*, Vol. 60, pp. 259–268, 1992.

Michael Ng is a Professor in the Department of Mathematics at the Hong Kong Baptist University. As an applied mathematician, Michael's main research areas include Bioinformatics, Data Mining, Operations Research and Scientific Computing. Michael has published and edited 5 books, published more than 140 journal papers. He is the principal editor of the Journal of Computational and Applied Mathematics, and the associate editor of SIAM Journal on Scientific Computing.

Liqun Qi received his B.S. in Computational Mathematics at Tsinghua University in 1968, his M.S, and Ph.D. degree in Computer Sciences at University of Wisconsin-Madison in 1981 and 1984, respectively. Professor Qi has taught in Tsinghua University, China, University of Wisconsin-Madison, USA, University of New South Wales, Australia, and The Hong Kong Polytechnic University. He is now Chair Professor of Applied Mathematics at The Hong Kong Polytechnic University. Professor Qi has published more than 140 research papers in international journals. He established the superlinear and quadratic convergence theory of the generalized Newton method, and played a principal role in the development of reformulation methods in optimization. Professor Qi's research work has been cited by the researchers around the world. According to the authoritative citation database ISIHighlyCited.com, he is one of the world's most highly cited 300 mathematicians during the period from 1981 to 1999.



Yu-Fei Yang received the B.Sc., M.S. and Ph.D. degrees in mathematics from Hunan University, P. R. China, in 1987, 1994 and 1999, respectively. From 1999 to 2001, he stayed at the University of New South Wales, Australia as visiting fellow. From 2002 to 2005, he held research associate and postdoctoral fellowship positions at the Hong Kong Polytechnic University. He is currently professor in the College of Mathematics and Econometrics, at Hunan University, P. R. China. His research interests includes optimization theory and methods, and partial differential equations with applications to image analysis.

Yu-Mei Huang received her M.Sc. in Computer science from Lanzhou University in 2000. She is now pursuing her doctoral studies in computational mathematics in Hong Kong Baptist University. Her research interests are in image processing and numerical linear algebra.