arXiv:1911.10968v3 [math.OC] 28 Aug 2021

# An Investigation on Semismooth Newton based Augmented Lagrangian Method for Image Restoration

Hongpeng Sun*

**Abstract**

Augmented Lagrangian method (also called as method of multipliers) is an important and powerful optimization method for lots of smooth or nonsmooth variational problems in modern signal processing, imaging, optimal control and so on. However, one usually needs to solve the coupled and nonlinear system together and simultaneously, which is very challenging. In this paper, we proposed several semismooth Newton methods to solve the nonlinear subproblems arising in image restoration in finite dimensional spaces, which leads to several highly efficient and competitive algorithms for imaging processing. With the analysis of the metric subregularities of the corresponding functions, we give both the global convergence and local linear convergence rate for the proposed augmented Lagrangian methods with semismooth Newton solvers.

## 1 Introduction

The augmented Lagrangian method (shorted as ALM henceforth) was originated in [12, 25]. The early developments can be found in [2, 10, 29] and the extensive studies in infinite dimensional spaces with various applications can be found in [10, 19] and so on. The comprehensive studies of ALM for convex, nonsmooth optimization and variational problems can be found in [2, 19]. In [29], the celebrated connections between the ALM and proximal point algorithms are established, where ALM is found to be equivalent to the proximal point algorithm applying to the essential dual problem. The convergence can thus be concluded in the general and powerful proximal point algorithm framework for convex optimization [28, 29].

ALM is very flexible for constrained optimization problems including both equality and inequality constraint problems [2, 19]. However, the challenging problem is solving the nonlinear and coupling systems simultaneously which usually appear while applying ALM. This is different from alternating direction method of multipliers (ADMM) type methods [10, 11], which decouple the unknown variables, deal with each subproblem separately and update them consecutively. However, for ALM, the extra effort is deserved if the nonlinear systems can be solved efficiently. This is due to the appealing asymptotic linear or superlinear convergence of ALM with increasing step sizes [24, 28, 29]. It is well known semismooth Newton methods are efficient solvers for nonsmooth and nonlinear systems. Semismooth Newton based augmented Lagrangian methods already have lots of successful applications in semidefinite programming [41], compressed sensing [23, 40], friction and contact problem [32, 33] and imaging problems [13, 20].

In this paper, we proposed several novel semismooth Newton based augmented Lagrangian methods for the total variation regularized image restoration problems with quadratic data terms.

*Institute for Mathematical Sciences, Renmin University of China, 100872 Beijing, China. Email: hpsun@amss.ac.cn.

Total variation (TV) regularization is quite fundamental for imaging problems. Besides, the regularization problem is a typical nonsmooth and convex optimization problem which is very challenging to solve and is standard for testing various algorithms [3]. To the best of our knowledge, the most related work is [13, 15, 20]. In [13], Moreau-Yosida regularized dual problem of the anisotropic ROF model in Banach space is considered and solved with the semismooth Newton based ALM method. In [20], the primal-dual active set method is employed for nonlinear systems of the ALM applying to the additionally regularized primal problems for solving the nonlinear system. In [15], semismooth Newton methods are directly applied to the strongly convex regularized ROF models with strong convexity on both the primal and dual variables without ALM method.

All the proposed algorithms are based on applying ALM to the primal model, whose strong convexity can be employed. Compared to [15], we do not need the strong convexity on the dual variables. Our contributions belong to the following parts. First, by introducing auxiliary variables, we use primal-dual semismooth Newton method [15] for nonlinear systems of ALM, which is very efficient for ALM. The proposed ALM is very efficient compared to the popular existing algorithms including the primal-dual first-order method [3]. They are especially very fast for the anisotropic total variation regularized quadratic problems. Second, we proved that the maximal monotone KKT (Kuhn-Tucker) mapping is metric subregular for the anisotropic TV regularized quadratic problem. We thus get the asymptotic linear or superlinear convergence rate of both the primal and dual sequence for the anisotropic case by the framework in [24, 28, 29]. With the calm intersection theorem [17], we also prove the metric subregularity of the maximal monotone operator associated with the dual problem for the isotropic case under mild condition. This leads to the asymptotic linear or superlinear convergence rate of the dual sequence for the isotropic case [24, 28, 29]. To the best of our knowledge, these subregularities are novel for the TV regularized quadratic problems. Third, we also give a systematical investigation of another efficient semismooth Newton method for solving the nonlinear systems that appear in ALM, which is involved with the soft thresholding operator.

The rest of this paper is organized as follows. In section 2, we give an introduction to the TV regularized quadratic problems and the corresponding ALM including the isotropic and anisotropic cases. In section 3, we present discussions on the primal-dual semismooth Newton for ALM by introducing auxiliary variables, which turns out very efficient. In section 4, we give detailed discussions on the semismooth Newton method involving the soft thresholding operators. Although all the semismooth Newton algorithms are aiming at the same nonlinear system for the isotropic or the anisotropic TV regularization. However, different formulations turn out different algorithms with efficiency. In section 5, we give the metric subregularity for the maximal monotone KKT mapping for anisotropic TV regularized quadratic problems and for the maixmal monotone operator associated with the dual problem of isotropic TV regularized quadratic problems. Together with the convergence of the semismooth Newton method, we get the corresponding asymptotic linear convergence. In section 6, we present the detailed numerical tests for all the algorithms and the comparison with efficient first-order algorithms focused on the ROF model and deblurring model. In section 7, we present some discussions.

## 2 Image restoration and Augmented Lagrangian Methods

The total variation regularized model is as follows.

$$\min_{u\in X} D(u) + \alpha\|\nabla u\|_1, \tag{P}$$

where $D(u) = \|Ku - z\|_2^2/2 + \frac{\mu}{2}\|\nabla u\|_2^2$ with $\mu \geq 0$ and $z$ being the known data for general image restoration [15, 20]. Here $K$ is a linear and bounded operator depending on the problems, e.g., $Ku = \kappa * u$ being the convolution with kernel $\kappa$ for image deblurring problems. The strong convexity of $D(u)$ is of critical importance for the proposed algorithms and analysis. Especially, the ROF model is for the case $K = I$ and $\mu = 0$. For the deblurring problems, we added the strongly regularization term $\mu\|\nabla u\|_2^2/2$ with $\mu > 0$ since $K^*K$ is usually not positive definite. Henceforth, all the variables, operators and spaces are all in finite dimensional space setting. The norm $\|\cdot\|_1$ based on the following isotropic or anisotropic norm

$$|p| := \sqrt{p_1^2 + p_2^2} \ \text{ (isotropic)} \quad \text{or} \quad |p|_1 = |p_1| + |p_2| \ \text{ (anisotropic)}, \tag{2.1}$$

where $p = (p_1, p_2)^T \in \mathbb{R}^2$. We denote $|\cdot|$ as the Euclidean norm including the absolute value for real valued scalar. Now introduce the image domain $\Omega \subset \mathbf{Z}^2$ which is the discretized grid [3]

$$\Omega = \{(i,j) \mid i,j \in \mathbb{N},\ 1 \leq i \leq M, 1 \leq j \leq N\}$$

where $M, N$ are the image dimensions. With discrete image domain $\Omega$, we define the discrete image space as $X = \{u : \Omega \to \mathbf{R}\}$ with the standard $L^2$ scalar product. Finite differences are used to discretize the operator $\nabla$ and its adjoint operator $\nabla^* = -\operatorname{div}$ with homogeneous Neumann and Dirichlet boundary conditions, respectively. We define $\nabla$ as the following operator

$$(\nabla u)_{i,j} = ((\nabla_1 u)_{i,j}, (\nabla_2 u)_{i,j})^T,$$

where forward differences are taken according to

$$(\nabla_1 u)_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j}, & \text{if } 1 \leq i < M, \\ 0, & \text{if } i = M, \end{cases} \qquad (\nabla_2 u)_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j}, & \text{if } 1 \leq j < N, \\ 0, & \text{if } j = N. \end{cases}$$

With $Y = X^2$ with the standard product scalar product, this gives a linear operator $\nabla : X \to Y$. The discrete divergence is then the negative adjoint of $\nabla$ with $\operatorname{div} = -\nabla^*$, i.e., the unique linear mapping $\operatorname{div} : Y \to X$ which satisfies for $\forall u \in X,\ \ p = (p_1, p_2)^T \in Y$

$$\langle \nabla u, p\rangle_Y = \langle u, \nabla^* p\rangle_X = -\langle u, \operatorname{div} p\rangle_X, \tag{2.2}$$

where the inner products are defined by

$$\langle u, v\rangle_X := \sum_{i=1}^{M}\sum_{j=1}^{N} u_{i,j} v_{i,j}, \quad \langle p, q\rangle_Y := \sum_{i=1}^{M}\sum_{j=1}^{N} [(p_1)_{i,j}(q_1)_{i,j} + (p_2)_{i,j}(q_2)_{i,j}], \quad q = (q_1, q_2)^T \in Y.$$

The div operator can be computed as $\operatorname{div} p = \nabla_1^- p_1 + \nabla_2^- p_2$ and we refer to [3] for the detailed backward difference operators $\nabla_1^-$ and $\nabla_2^-$.

By (2.1), the isotropic or anisotropic TV in (P) is as follows,

$$\|\nabla u\|_1 := \int_\Omega |\nabla u| dx, \quad \text{or} \quad \|\nabla u\|_1 := \int_\Omega |\nabla_1 u| dx + \int_\Omega |\nabla_2 u| dx, \tag{2.3}$$

where $dx$ in (2.3) is the area element of $\mathbb{R}^2$ and the integrals in (2.3) are understood as the finite summation over all pixels. With these preparations, by the Fenchel-Rockafellar duality theory [1, 19], the primal-dual form of (P) can be written as

$$\min_{u \in X} \max_{\lambda \in Y} \ D(u) + \langle \nabla u, \lambda\rangle_Y - \mathcal{I}_{\{\|\lambda\|_\infty \leq \alpha\}}(\lambda), \tag{2.4}$$

where $\mathcal{I}$ is the indicator function and the dual form of (P) can be written

$$\max_{\lambda\in Y}\left\{-\left(d(\lambda):=\frac{1}{2}\|\operatorname{div}\lambda+f\|_{H^{-1}}^2-\frac{1}{2}\|z\|_2^2+\mathcal{I}_{\{\|\lambda\|_\infty\le\alpha\}}(\lambda)\right)\right\}. \tag{D}$$

The notations $f$ and the positive definite, linear operator $H$ are as follows

$$f:=K^*z,\quad H:=-\mu\Delta+K^*K,\quad \|u\|_{H^{-1}}^2=\langle H^{-1}u,u\rangle_X,\quad \|u\|_H^2=\langle Hu,u\rangle_X. \tag{2.5}$$

We also need the following norms for arbitrary $u\in X$, $p=(p_1,p_2)\in Y$, $1\le t<\infty$,

$$\|u\|_t=\Big(\sum_{(i,j)\in\Omega}|u_{i,j}|^t\Big)^{1/t},\quad \|p\|_t=\Big(\sum_{(i,j)\in\Omega}\big[((p_1)_{i,j})^2+((p_2)_{i,j})^2\big]^{t/2}\Big)^{1/t}.$$

The $\|\lambda\|_\infty$ is defined as follows. For the isotropic case,

$$\|\lambda\|_\infty=\max_{(i,j)\in\Omega}\sqrt{((\lambda_1)_{i,j})^2+(\lambda_2)_{i,j})^2}.$$

For the anisotropic case, with (2.3) and applying the Fenchel-Rockafellar duality theory for $\|\nabla_1 u\|_1$ and $\|\nabla_2 u\|_1$ separately, we have

$$\mathcal{I}_{\{\|\lambda\|_\infty\le\alpha\}}(\lambda)=\mathcal{I}_{\{\|\lambda_1\|_\infty\le\alpha\}}(\lambda_1)+\mathcal{I}_{\{\|\lambda_2\|_\infty\le\alpha\}}(\lambda_2),\quad \|\lambda_k\|_\infty=\max_{(i,j)\in\Omega}|(\lambda_k)_{i,j}|,\quad k=1,2.$$

The optimality conditions on the saddle points $(\bar u,\bar\lambda)$ are as follows

$$H\bar u-f+\nabla^*\bar\lambda=0,\quad \nabla\bar u\in\partial\mathcal{I}_{\{\|\bar\lambda\|_\infty\le\alpha\}}(\bar\lambda). \tag{2.6}$$

By the Fenchel-Rockafellar duality theory, we have $\bar\lambda\in\partial(\alpha\|\nabla\bar u\|_1)$ and

$$\langle\bar\lambda,\nabla\bar u\rangle=\mathcal{I}_{\{\|\bar\lambda\|_\infty\le\alpha\}}(\bar\lambda)+\alpha\|\nabla\bar u\|_1. \tag{2.7}$$

The optimality condition for $\bar\lambda$ in (2.6) is also equivalent to

$$\mathcal{P}_\alpha(\bar\lambda+c_0\nabla\bar u)=\bar\lambda,\quad \text{for all constant } c_0>0, \tag{2.8}$$

where $\mathcal{P}_\alpha$ is the projection to the feasible set $\{\lambda:\|\lambda\|_\infty\le\alpha\}$, i.e.,

$$\mathcal{P}_\alpha(\lambda)=\frac{p}{\max(1.0,|\lambda|/\alpha)}\quad\text{or}\quad \mathcal{P}_\alpha(\lambda)=\left(\frac{\lambda_1}{\max(1,|\lambda_1|/\alpha)},\frac{\lambda_2}{\max(1,|\lambda_2|/\alpha)}\right)^T. \tag{2.9}$$

Introducing $p:=\nabla u$, the problem (P) becomes the following constrained optimization problem

$$\min_{x\in X}D(u)+\alpha\|p\|_1,\quad \text{such that}\quad \nabla u=p.$$

The augmented Lagrangian method thus follows, with nondecreasing update of $\sigma_k\to c_\infty<+\infty$,

$$(u^{k+1},p^{k+1})=\arg\min_{u,p}L(u,p;\lambda^k):=D(u)+\alpha\|p\|_1+\langle\lambda^k,\nabla u-p\rangle+\frac{\sigma_k}{2}\|\nabla u-p\|_2^2, \tag{2.10}$$

$$\lambda^{k+1}=\lambda^k+\sigma_k(\nabla u^{k+1}-p^{k+1}). \tag{2.11}$$

We will use the semismooth Newton method to solve (2.10). The optimality conditions with fixed $\sigma_k$ and $\lambda^k$ for (2.10) are

$$Hu - f + \nabla^*\lambda^k + \sigma_k\nabla^*(\nabla u - p) = 0, \tag{2.12}$$

$$\partial\alpha\|p\|_1 - \lambda^k + \sigma_k(p - \nabla u) \ni 0, \tag{2.13}$$

where $(u, p) = (u^{k+1}, p^{k+1})$ are the optimal solutions of (2.10). The equation (2.13) leads to

$$\lambda^k + \sigma_k\nabla u \in (\sigma_k I + \partial\alpha\|\cdot\|_1)p \Rightarrow p = (I + \frac{1}{\sigma_k}\partial\alpha\|\cdot\|_1)^{-1}(\frac{\lambda^k + \sigma_k\nabla u}{\sigma_k}) := S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u), \tag{2.14}$$

where $S_{\frac{\alpha}{\sigma_k}}(\cdot)$ is the soft thresholding operator for the isotropic or anisptropic $\|\cdot\|_1$ norm. With relation (2.14), the augmented Lagrangian $L(u, p; \lambda^k)$ can be reformulated as

$$\begin{aligned}\Phi_k(u; \lambda^k, \sigma_k) &:= L(u, S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u); \lambda^k)\\ &= D(u) + \alpha\|S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u)\|_1 + \frac{\sigma_k}{2}\|\frac{\lambda^k}{\sigma_k} + \nabla u - S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u)\|_2^2 - \frac{1}{2\sigma_k}\|\lambda^k\|_2^2,\end{aligned} \tag{2.15}$$

which will be more convenient than (2.10) once the globalization strategy including the line search is employed. Substituting $p$ of (2.14) into (2.12), we get

$$Hu - f + \nabla^*\lambda^k + \sigma_k\nabla^*\nabla u - \sigma_k\nabla^*(I + \frac{1}{\sigma_k}\partial\alpha\|\cdot\|_1)^{-1}(\frac{\lambda^k + \sigma_k\nabla u}{\sigma_k}) = 0. \tag{2.16}$$

Denoting $G^*(p) = \alpha\|p\|_1$, we see the Fenchel dual function of $G^*$ is $G(h) = I_{\{\|\cdot\|_\infty \le \alpha\}}(h)$. With the Moreau's identity,

$$x = (I + \tau\partial G)^{-1}(x) + \tau(I + \frac{1}{\tau}\partial G^*)^{-1}(\frac{x}{\tau}), \tag{2.17}$$

we arrive at

$$\sigma_k p = \sigma_k(I + \frac{1}{\sigma_k}\partial\alpha\|\cdot\|_1)^{-1}(\frac{\lambda^k + \sigma_k\nabla u}{\sigma_k}) = \lambda^k + \sigma_k\nabla u - (I + \sigma_k\partial G)^{-1}(\lambda^k + \sigma_k\nabla u). \tag{2.18}$$

Substituting $p$ of (2.18) into (2.12) leads to the equation of $u$

$$Hu - f + \nabla^*(I + \sigma_k\partial G)^{-1}(\lambda^k + \sigma_k\nabla u) = 0. \tag{2.19}$$

Indeed, we can solve (2.16) directly with semismooth Newton methods, which will be discussed in the subsequent sections. Now let's turn to another formulation by introducing an auxiliary variable

$$h := (I + \sigma_k\partial G)^{-1}(\lambda^k + \sigma_k\nabla u) = \mathcal{P}_\alpha(\lambda^k + \sigma_k\nabla u). \tag{2.20}$$

By the definition of the projection (2.9) and taking the isotropic case for example, (2.20) becomes

$$h = \mathcal{P}_\alpha(\lambda^k + \sigma_k\nabla u) = \frac{\lambda^k + \sigma_k\nabla u}{\max(1.0, |\lambda^k + \sigma_k\nabla u|/\alpha)}. \tag{2.21}$$

The equation (2.19) thus becomes

$$Hu - f + \nabla^* h = 0. \tag{2.22}$$

Combining (2.20), (2.21) and (2.22), we get the following equations of $(u, \lambda)$ instead of $(u, p)$,

$$\mathcal{F}(u,h) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathcal{F}(u,h) := \begin{bmatrix} Hu - f + \nabla^* h \\ -\sigma_k \nabla u - \lambda^k + \max\big(1.0, \dfrac{|\lambda^k + \sigma_k \nabla u|}{\alpha}\big) h \end{bmatrix}. \tag{2.23}$$

Once solving any of the equations (2.16) and (2.23) and obataining the solution $u$, we can update the $p$ by (2.14) or (2.18) accordingly. Then the Lagrangian multiplier $\lambda^{k+1}$ can be updated by (2.11). Actually, according to (2.18), compared to the update of $\lambda^{k+1}$ (2.11), we see the update of the multiplier can also be

$$\lambda^{k+1} = (I + \sigma_k \partial G)^{-1}(\lambda^k + \sigma_k \nabla u) = \mathcal{P}_\alpha(\lambda^k + \sigma_k \nabla u), \tag{2.24}$$

which is a nonlinear update compared to the linear update (2.11). We refer to [19] (chapter 4) for general nonlinear updates of Lagrangian multipliers with different derivations and framework of ALM. Throughout this paper, we will use the semismooth Newton methods to any of the subproblems (2.16) and (2.23). The different formulations of (2.16) and (2.23) will bring out different algorithms and it would turn out different efficiency. We begin with the semismoothness where the Newton derivative can be chosen as the Clarke generalized derivative [18, 19].

**Definition 1** (Newton differentiable and Newton Derivative [19])**.** $F : D \subset X \to Z$ is called Newton differentiable at $x$ if there exist an open neighborhood $N(x) \subset D$ and mapping $G : N(x) \to \mathcal{L}(X, Z)$ such that (Here the spaces $X$ and $Z$ are Banach spaces.)

$$\lim_{|h| \to 0} \frac{|F(x+h) - F(x) - G(x+h)h|_Z}{|h|_X} = 0. \tag{2.25}$$

The family $\{G(s) : s \in N(x)\}$ is called an Newton derivative of $F$ at $x$.

If $F : \mathbb{R}^n \to \mathbb{R}^m$ and the set of mapping $G$ is the Clarke generalized derivative $\partial F$, we call $F$ is semismooth [17].

**Definition 2** (Semismoothness [22, 23, 35])**.** Let $F : O \subseteq X \to Y$ be a locally Lipschitz continuous function on the open set $O$. $F$ is said to be semismooth at $x \in O$ if $F$ is directionally differentiable at $x$ and for any $V \in \partial F(x + \Delta x)$ with $\Delta x \to 0$,

$$F(x + \Delta x) - F(x) - V\Delta x = o(\|\Delta x\|).$$

The Newton derivatives of vector-valued functions can be computed component-wisely [35] (Proposition 2.10) or [5] (Theorem 9.4). Together with the definition of semismoothness, we have the following lemma.

**Lemma 1.** *Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ and $F = (F_1(x), F_2(x), \cdots, F_l(x))^T$ with $F_i : \mathbb{R}^n \to \mathbb{R}^{l_i}$ being semismooth. Here $l_i \in \mathbb{Z}^+$ and $\sum_{i=1}^l l_i = m$. Assuming the Newton derivative $D_N F_i(x) \in \partial F_i(x)$, $i = 1, 2, \cdots, l$, then the Newton derivative of $F$ can be chosen as*

$$D_N F(x) = \big[D_N F_1(x), D_N F_2(x), \cdots, D_N F_l(x)\big]^T. \tag{2.26}$$

Now we turn to the semismooth Newton method for solving (2.16) and (2.23). The semismooth Newton method for the general nonlinear equation $F(x) = 0$ can be written as

$$\mathcal{V}(x^l)^{-1} \delta x^{l+1} = -F(x^l), \tag{2.27}$$

where $\mathcal{V}(x^l)$ is a semismooth Newton derivative of $F$ at $x^l$, for example $\mathcal{V}(x^l) \in \partial F(x^l)$. Additionally, $\mathcal{V}(x)$ needs to satisfy the *regularity condition* [5, 35]. Henceforth, we say $\mathcal{V}(x)$ satisfies the *regularity condition* if $\mathcal{V}(x)^{-1}$ exist and are uniformly bounded in a small neighborhood of the solution $x^*$ of $F(x^*) = 0$. When the globalization strategy including line search is necessary, one can get the Newton update $x^{l+1}$ with the Newton direction $\delta x^{l+1}$ in (2.27). Once the globalization strategy is not needed, the semismooth Newton iteration can also be written as follows with solving $x^{l+1}$ directly

$$\mathcal{V}(x^l)x^{l+1} = \mathcal{V}(x^l)x^l - F(x^l). \tag{2.28}$$

# 3 Augmented Lagrangian Method with Primal-Dual Semismooth Newton Method

## 3.1 The Isotropic Total Variation

Now we turn to the semismoothness of the nonlinear system (2.23). The only nonlinear or nonsmooth part comes from the function $\max(1.0, |\lambda^k + \sigma_k \nabla u|/\alpha)$.

**Lemma 2.** *The function* $\mathcal{G}(u) := \max(1.0, \dfrac{|\lambda^k + \sigma_k \nabla u|}{\alpha})$ *is semismooth on* $u$ *and its Clarke's generalized gradient for* $u$ *is as follows,*

$$\left\{ \chi^s_{\lambda^k,u} \frac{\sigma_k}{\alpha} \frac{\langle \lambda^k + \sigma_k \nabla u, \nabla \cdot \rangle}{|\lambda^k + \sigma_k \nabla u|} \mid s \in [0,1] \right\} = \partial_u(\max(1.0, \frac{|\lambda^k + \sigma_k \nabla u|}{\alpha})), \tag{3.1}$$

*where* $\chi^s_{\lambda^k,u}$ *is the generalized Clarke derivatives of* $\max(\cdot, 1.0)$,

$$\chi^s_{\lambda^k,u} = \begin{cases} 1, & |\lambda^k + \sigma_k \nabla u|/\alpha > 1.0, \\ s, & |\lambda^k + \sigma_k \nabla u|/\alpha = 1.0, \ s \in [0,1], \\ 0, & |\lambda^k + \sigma_k \nabla u|/\alpha < 1.0. \end{cases} \tag{3.2}$$

*Furthermore,* $\mathcal{F}(u,h)$ *in* (2.23) *is semismooth on* $(u,h)$. *Henceforth, we choose* $s = 1$ *for the Newton derivative of* $\mathcal{G}(u)$ *in* (3.1) *and denote* $\chi_{\lambda^k,u} := \chi^1_{\lambda^k,u}$.

*Proof.* We claim that $\mathcal{G}(u)$ is actually a $PC^\infty$ function of $u$. Introduce $\mathcal{G}_1(u) = 1.0$ and $\mathcal{G}_2(u) = |\lambda^k + \sigma_k \nabla u|/\alpha$ which are *selection functions* of $\mathcal{G}(u)$. $\mathcal{G}(u)$ is called as the *continuous selection* of the functions $\mathcal{G}_1(u)$ and $\mathcal{G}_2(u)$ [31] (Chapter 4) (or Definition 4.5.1 of [9]). We see $\mathcal{G}_1(u)$ is a smooth function and $\mathcal{G}_2(u)$ is smooth in any open set outside the closed set $D_0 := \{u \mid |\lambda^k + \sigma_k \nabla u| = 0\}$. Thus for any $u \in D_\alpha := \{u \mid |\lambda^k + \sigma_k \nabla u| = \alpha\}$, there exists a small open neighborhood of $u$ such that $\mathcal{G}_1(u)$ and $\mathcal{G}_2(u)$ are smooth functions. We thus conclude that $\mathcal{G}(u)$ is a $PC^\infty$ function of $u$. $\mathcal{G}(u)$ is also $PC^1$ and hence is semismooth on $u$ [35] (Proposition 2.26), since $\nabla_u \mathcal{G}_1(u) = 0$ and $\nabla_u \mathcal{G}_2(u) = \sigma_k \langle \lambda^k + \sigma_k \nabla u, \nabla \cdot \rangle / (\alpha |\lambda^k + \sigma_k \nabla u|)$ out side $D_0$. For any $u \in D_\alpha$, by [31] (Proposition 4.3.1), we see

$$\partial_u \mathcal{G}(u) = \mathrm{co}\{\nabla_u \mathcal{G}_1(u), \nabla_u \mathcal{G}_2(u)\},$$

where "co" denotes the convex hull of the corresponding set [4]. We thus obtain the equation (3.1). Since each component of $\mathcal{F}(u,h)$ is an affine function on $h$ and is semismoooth on $(u,h)$, the semismooth of $\mathcal{F}(u,h)$ on $(u,h)$ then follows [35] (Proposition 2.10). □

By Lemma 1 and 2, denoting $x^l = (u^l, h^l)$, $x = (u, h)$ and $\mathcal{F}(u, h) = (\mathcal{F}_1(u, h), \mathcal{F}_2(u, h))^T$, the Newton derivative of $F(u, h)$ can be chosen as

$$D_N\mathcal{F} = (D_N\mathcal{F}_1(u, h), D_N\mathcal{F}_2(u, h))^T. \tag{3.3}$$

Thus the Newton derivative of the nonlinear equation (2.23) can be chosen as

$$\mathcal{V}^I(x^l) = \begin{bmatrix} H & \nabla^* \\ -\sigma_k\nabla + \chi_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle\lambda^k + \sigma_k\nabla u^l, \nabla\cdot\,\rangle}{|\lambda^k + \sigma_k\nabla u^l|}h^l & \max\big(1.0, \dfrac{|\lambda^k + \sigma_k\nabla u^l|}{\alpha}\big) \end{bmatrix}. \tag{3.4}$$

Let's introduce

$$D_l = U_{\sigma_k}(\lambda^k, u^l) := \max\big(1.0, \frac{|\lambda^k + \sigma_k\nabla u^l|}{\alpha}\big), \quad B_l = \left[\chi_{\lambda^k,u^l}\frac{\sigma_k}{\alpha}\frac{\langle\lambda^k + \sigma_k\nabla u^l, \nabla\cdot\rangle}{|\lambda^k + \sigma_k\nabla u^l|}h^l\right],$$
$$C_l = -\sigma_k\nabla + B_l.$$

It can be readily verified that

$$\mathcal{V}^I(x^l)x^l - \mathcal{F}(x^l) = \begin{bmatrix} f \\ \lambda^k + \chi_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle\lambda^k + \sigma_k\nabla u^l, \nabla u^l\rangle}{|\lambda^k + \sigma_k\nabla u^l|}h^l \end{bmatrix} = \begin{bmatrix} f \\ \lambda^k + B_l u^l \end{bmatrix} := \begin{bmatrix} f \\ b_2^l \end{bmatrix}.$$

Next, we turn to solve the Newton update (2.28)

$$\begin{bmatrix} H & \nabla^* \\ C_l & D_l \end{bmatrix}\begin{bmatrix} u^{l+1} \\ h^{l+1} \end{bmatrix} = \begin{bmatrix} f \\ b_2^l \end{bmatrix}. \tag{3.5}$$

For solving the linear system (3.5), it is convenient to solve $u^{l+1}$ first, i.e., solving the equation of the Schur complement $\mathcal{V}^I(x^l)/D_l$. Substituting

$$h^{l+1} = \left(b_2^l + \sigma_k\nabla u^{l+1} - \chi_{\lambda^k,u^l}\frac{\sigma_k}{\alpha}\frac{\langle\lambda^k + \sigma_k\nabla u^l, \nabla u^{l+1}\rangle}{|\lambda^k + \sigma_k\nabla u^l|}h^l\right)\Big/\left(\max\big(1.0, \frac{|\lambda^k + \sigma_k\nabla u^l|}{\alpha}\big)\right) \tag{3.6}$$

into the first equation on $u^{k+1}$, we have

$$(H - \nabla^* D_l^{-1}C_l)u^{l+1} = f + \operatorname{div}\frac{b_2^l}{U_{\sigma_k}(\lambda^k, u^l)}, \tag{3.7}$$

which is also the following equation in detail

$$Hu^{l+1} - \operatorname{div}\frac{\sigma_k\nabla u^{l+1} - \chi_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle\lambda^k + \sigma_k\nabla u^l, \nabla u^{l+1}\rangle}{|\lambda^k + \sigma_k\nabla u^l|}h^l}{U_{\sigma_k}(\lambda^k, u^l)} = f + \operatorname{div}\frac{b_2^l}{U_{\sigma_k}(\lambda^k, u^l)}. \tag{3.8}$$

After calculating $u^{l+1}$ in (3.8), we get $h^{l+1}$ by (3.6).

We can also first calculate $h^{l+1}$ following the calculation of $u^{l+1}$, i.e., solving the equation of the Schur complement $\mathcal{V}^I(x^l)/H$. Solving dual variables first can also be found in [14], where the primal-dual semismooth Newton is employed for total generalized variation. Substituting

$$u^{l+1} = H^{-1}(\operatorname{div} h^{l+1} + f), \tag{3.9}$$

into (3.6), we obtain the linear equation of $h^{l+1}$

$$(D_l - C_l H^{-1}\nabla^*)h^{l+1} = b_2^k - C_l H^{-1} f, \tag{3.10}$$

which is the folloing equation in detail,

$$\begin{aligned}&\max\big(1.0, \frac{|\lambda^k + \sigma_k \nabla u^l|}{\alpha}\big)h^{l+1} - \sigma_k \nabla H^{-1} \operatorname{div} h^{l+1} + \chi_{\lambda^k, u^l} \frac{\sigma_k}{\alpha} \frac{\langle \lambda^k + \sigma_k \nabla u^l, \nabla H^{-1} \operatorname{div} h^{l+1}\rangle}{|\lambda^k + \sigma_k \nabla u^l|} h^l \\ &= b_2^l - C_l H^{-1} f \end{aligned} \tag{3.11}$$

We then recover $u^{l+1}$ by (3.9) after calculating $h^{l+1}$.

We have the following lemma for the regularity conditions of the Newton derivative.

**Lemma 3.** *If the feasibility of $h^l$ is satisfied, i.e., $|h^l| \le \alpha$ by (2.20), we have the positive definiteness of the Schur complement $\mathcal{V}^I(x^l)/D_l = (H - \nabla^* D_l^{-1} C_l)$,*

$$\langle -\nabla^* D_l^{-1} C_l u, u\rangle_X \ge 0 \Rightarrow \langle (\mathcal{V}^I(x^l)/D_l)u, u\rangle_X \ge \|u\|_H^2 = \langle Hu, u\rangle_X. \tag{3.12}$$

*Thus $\mathcal{V}^I(x^l)/D_l$ satisfies the regularity condition. Furthermore $\mathcal{V}^I(x^l)$ can be chosen as a Newton derivative of $\mathcal{F}(u,h)$. The linear operator $\mathcal{V}^I(x^l)$ and the Schur complement $\mathcal{V}^I(x^l)/H$ satisfy the regularity condition for any fixed $\sigma_k$ and $\lambda^k$, i.e., $\mathcal{V}^I(x^l)$ and $\mathcal{V}^I(x^l)/H$ are nonsingular and the corresponding inverse are uniformly bounded for any fixed $\sigma_k$ and $\lambda^k$.*

*Proof.* We first prove the regular condition of $\mathcal{V}^I(x^l)/D_l$, whose proof is essentially similar to the proof of Lemma 3.3 in [15]. Denote

$$\Omega^+ := \{(x_1, x_2) \in \Omega : |\lambda^k + \sigma_k \nabla u| \ge \alpha\}, \quad \Omega^- := \{(x_1, x_2) \in \Omega : |\lambda^k + \sigma_k \nabla u| < \alpha\}, \quad \Omega = \Omega^+ \cup \Omega^-.$$

Since for any $u \in X$, we have

$$\begin{aligned}&\langle -\nabla^* D_l^{-1} C_l u, u\rangle_X = -\langle D_l^{-1} C_l u, \nabla u\rangle_Y = \langle -D_l^{-1}(-\sigma_k \nabla + B_l)u, \nabla u\rangle_Y \\ &= \langle \sigma_k D_l^{-1} \nabla u, \nabla u\rangle_2 - \langle D_l^{-1} B_l u, \nabla u\rangle_Y \\ &= \int_{\Omega^-} \sigma_k D_l^{-1} |\nabla u|^2 dx + \int_{\Omega^+} \sigma_k D_l^{-1} |\nabla u|^2 dx - \int_{\Omega^+} \langle D_l^{-1} \frac{\sigma_k}{\alpha} \frac{\langle \lambda^k + \sigma_k \nabla u^l, \nabla u\rangle}{|\lambda^k + \sigma_k \nabla u^l|} h^k, \nabla u\rangle dx. \end{aligned} \tag{3.13}$$

With the assumption $|h^l| \le \alpha$ and direct calculations, we see

$$|\langle D_l^{-1} \frac{\sigma_k}{\alpha} \frac{\langle \lambda^k + \sigma_k \nabla u^l, \nabla u\rangle}{|\lambda^k + \sigma_k \nabla u^l|} h^l, \nabla u\rangle| \le D_l^{-1} \frac{\sigma_k}{\alpha} \frac{|\lambda^k + \sigma_k \nabla u^l||\nabla u|}{|\lambda^k + \sigma_k \nabla u^l|} |h^l||\nabla u| \le \sigma_k D_l^{-1} |\nabla u|^2. \tag{3.14}$$

Combining (3.13) and (3.14), we arrive at

$$\langle -\nabla^* D_l^{-1} C_l u, u\rangle_X \ge \int_{\Omega^-} \sigma_k D_l^{-1} |\nabla u|^2 dx \ge 0,$$

which leads to (3.12). We thus conclude that $\mathcal{V}^I(x^l)/D_l \ge H$ and $(\mathcal{V}^I(x^l)/D_l)^{-1}$ is bounded by the positive definiteness of $H$. For the regularity condition of $\mathcal{V}^I(x^l)$, it is known that ([39] formula 0.8.1 which is similar to the Banachiewicz inversion formula)

$$\mathcal{V}^I(x^l)^{-1} = \begin{bmatrix} (\mathcal{V}^I(x^l)/D_l)^{-1} & -(\mathcal{V}^I(x^l)/D_l)^{-1} \nabla^* D_l^{-1} \\ -D_l^{-1} C_l (\mathcal{V}^I(x^l)/D_l)^{-1} & D_l^{-1} + D_l^{-1} C_l (\mathcal{V}^I(x^l)/D_l)^{-1} \nabla^* D_l^{-1}. \end{bmatrix}$$

By the boundedness of $(\mathcal{V}^I(x^l)/D_l)^{-1}$, $C_l$ and $D_l^{-1}$, we get the boundedness of $\mathcal{V}^I(x^l)^{-1}$. The boundedness of $D_l^{-1}$ is because $D_l \geq I$ and the boundedness of $C_l$ comes from the boudnedness of $\nabla$ and $B_l$.

Similarly, for the existence and boundedness of $(\mathcal{V}^I(x^l)/I)^{-1}$, by the Duncan inversion formula (see 0.8.1 and 0.8.2 of [39]) or Woodbury formula, we have

$$(\mathcal{V}^I(x^l)/H)^{-1} = (D_l - C_l H^{-1}\nabla^*)^{-1} = D_l^{-1} + D_l^{-1} C_l (\mathcal{V}^I(x^l)/D_l)^{-1} \nabla^* D_l^{-1}.$$

We thus get the boundedness of $(\mathcal{V}^I(x^l)/H)^{-1}$. □

## 3.2 The Anisotropic Total Variation

For the anisotropic $l_1$ norm in (2.1), with (2.9), the projection (2.21) becomes

$$\mathcal{P}_\alpha^A(\lambda^k + \sigma_k \nabla u) = \left( \frac{\lambda_1^k + \sigma_k \nabla_1 u}{\max(1.0, |\lambda_1^k + \sigma_k \nabla_1 u|/\alpha)}, \frac{\lambda_2^k + \sigma_k \nabla_2 u}{\max(1.0, |\lambda_2^k + \sigma_k \nabla_2 u|/\alpha)} \right)^T. \tag{3.15}$$

With similar analysis as the isotropic case, the equation (2.23) becomes

$$\mathcal{F}^A(u,h) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathcal{F}^A(u,h) := \begin{bmatrix} Hu - f + \nabla^* h \\ -\sigma_k \nabla_1 u - \lambda_1^k + \max\left(1.0, \dfrac{|\lambda_1^k + \sigma_k \nabla_1 u|}{\alpha}\right) h_1 \\ -\sigma_k \nabla_2 u - \lambda_2^k + \max\left(1.0, \dfrac{|\lambda_2^k + \sigma_k \nabla_2 u|}{\alpha}\right) h_2 \end{bmatrix}. \tag{3.16}$$

Similar to Lemma 2, we have the following lemma, whose proof is completely similar to Lemma 2 and we omit here.

**Lemma 4.** *The functions* $\max(1.0, |\lambda_1^k + \sigma_k \nabla_1 u|/\alpha)$ *and* $\max(1.0, |\lambda_2^k + \sigma_k \nabla_2 u|/\alpha)$ *are semismooth functions of* $u$ *and their Clarke generalized gradients are as follows,*

$$\left\{ \chi_{\lambda^k,u}^{i,s} \frac{\sigma_k}{\alpha} \frac{\langle \lambda_i^k + \sigma_k \nabla_i u, \nabla_i \cdot \rangle}{|\lambda_i^k + \sigma_k \nabla_i u|} \mid s \in [0,1] \right\} = \partial_u \left( \max\left(1.0, \frac{|\lambda_i^k + \sigma_k \nabla_i u|}{\alpha}\right) \right), \quad i = 1,2 \tag{3.17}$$

*where* $\chi_{\lambda^k,u}^{1,s}$ *and* $\chi_{\lambda^k,u}^{2,s}$ *are the generalized derivatives of* $\max(\cdot, 1.0)$,

$$\chi_{\lambda^k,u}^{1,s} = \begin{cases} 1, & |\lambda_1^k + \sigma_k \nabla_1 u|/\alpha > 1.0, \\ s, & |\lambda_1^k + \sigma_k \nabla_1 u|/\alpha = 1.0, \\ 0, & |\lambda_1^k + \sigma_k \nabla_1 u|/\alpha < 1.0, \end{cases} \quad \chi_{\lambda^k,u}^{2,s} = \begin{cases} 1, & |\lambda_2^k + \sigma_k \nabla_2 u|/\alpha \geq 1.0, \\ s, & |\lambda_2^k + \sigma_k \nabla_2 u|/\alpha = 1.0, \\ 0, & |\lambda_2^k + \sigma_k \nabla_2 u|/\alpha < 1.0. \end{cases} \; s \in [0,1], \tag{3.18}$$

*Furthermore* $F^A(u,h)$ *is semismooth on* $(u,h)$. *Henceforth, denote* $\chi_{\lambda^k,u}^1 := \chi_{\lambda^k,u}^{1,1}$ *and* $\chi_{\lambda^k,u}^2 := \chi_{\lambda^k,u}^{2,1}$ *for* $s = 1$ *cases in* (3.18).

Let's introduce

$$D_l^A = \begin{bmatrix} U_{\sigma_k}^1 & 0 \\ 0 & U_{\sigma_k}^2 \end{bmatrix}, \quad B_l^A = \begin{bmatrix} \chi_{\lambda^k,u^k}^1 \dfrac{\sigma_k}{\alpha} \dfrac{\langle \lambda_1^k + \sigma_k \nabla_1 u^l, \nabla_1 \cdot \rangle}{|\lambda_1^k + \sigma_k \nabla_1 u^l|} h_1^l \\ \chi_{\lambda^k,u^l}^2 \dfrac{\sigma_k}{\alpha} \dfrac{\langle \lambda_2^k + \sigma_k \nabla_2 u^l, \nabla_2 \cdot \rangle}{|\lambda_2^k + \sigma_k \nabla_1 u^l|} h_2^l \end{bmatrix}, \quad C_l^A = -\sigma\nabla + B_l^A,$$

where

$$U_{\sigma_k}^1(\lambda^k, u^l) = \max\left(1.0, \frac{|\lambda_1^k + \sigma_k \nabla_1 u^l|}{\alpha}\right), \quad U_{\sigma_k}^2(\lambda^k, u^l) = \max\left(1.0, \frac{|\lambda_2^k + \sigma_k \nabla_2 u^l|}{\alpha}\right).$$

By Lemma 1 and 4, since $F^A(u,h)$ is an affine function of $h$, together with Lemma 4, it can be readily verified that, we can choose the Newton derivative of the nonlinear equation (3.16) as

$$\mathcal{V}^A = \begin{bmatrix} H & \nabla^* \\ C_l^A & D_l^A \end{bmatrix}. \tag{3.19}$$

The right-hand side becomes

$$\mathcal{V}^A(x^l)x^l - \mathcal{F}^A(x^l) = \begin{bmatrix} f \\ B_l f \end{bmatrix} = \begin{bmatrix} f \\ \lambda_1^k + \chi^1_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle \lambda_1^k + \sigma_k\nabla_1 u^l, \nabla_1 u^l\rangle}{\|\lambda_1^k + \sigma_k \nabla_1 u^l\|}h_1^l \\ \lambda_2^k + \chi^2_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle \lambda_2^k + \sigma_k\nabla_2 u^l, \nabla_2 u^l\rangle}{\|\lambda_2^k + \sigma_k \nabla_2 u^l\|}h_1^l \end{bmatrix} := \begin{bmatrix} f \\ b_1^l \\ b_2^l \end{bmatrix}.$$

For solving $u^{k+1}$ first, the Newton update becomes

$$(H - \nabla^* {D_l^A}^{-1} C_l^A)u^{l+1} = f + \operatorname{div} {D_l^A}^{-1} b^l, \tag{3.20}$$

where $b^l = (b_1^l, b_2^l)^T$. Then $h^{l+1}$ can be recovered by

$$h^{l+1} = \begin{bmatrix} \left(b_1^l + \sigma_k\nabla_1 u^{l+1} - \chi^1_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle \lambda_1^k + \sigma_k\nabla_1 u^l, \nabla_1 u^{l+1}\rangle}{|\lambda_1^k + \sigma_k \nabla_1 u^l|}h_1^l\right) \Big/ U^1_{\sigma_k}(\lambda^k, u^l) \\ \left(b_2^l + \sigma_k\nabla_2 u^{l+1} - \chi^2_{\lambda^k,u^l}\dfrac{\sigma_k}{\alpha}\dfrac{\langle \lambda_2^k + \sigma_k\nabla_2 u^l, \nabla_2 u^{l+1}\rangle}{|\lambda_2^k + \sigma_k \nabla_2 u^l|}h_2^l\right) \Big/ U^2_{\sigma_k}(\lambda^k, u^l) \end{bmatrix}. \tag{3.21}$$

For calculating $h^{+1}$ first with solving the Schur complement $\mathcal{V}^A(x^l)/H$, we have

$$(D_l^A - C_l H^{-1}\nabla^*)h^{l+1} = b^l - C_l H^{-1} f. \tag{3.22}$$

Then $u^{l+1}$ can be recovered through

$$u^{l+1} = H^{-1}(f - \nabla^* h^{l+1}).$$

For the positive definiteness and regularity condition of the Schur complement $\mathcal{V}^A(x^l)/D_l^A$ , we have the following lemma, whose proof is completely similar to the proof of Lemma 3 and we omit here.

**Lemma 5.** *If the feasibility of $h^l$ is satisfied, i.e., $|h_i^l| \le \alpha$ by (2.20), $i = 1, 2$ we have the positive definiteness of $(H - \nabla^* {D_l^A}^{-1} C_l^A)$,i.e.,*

$$\langle -\nabla^* {D_l^A}^{-1} C_l^A u, u\rangle_X \ge 0 \Rightarrow \langle (H - \nabla^* {D_l^A}^{-1} C_l^A)u, u\rangle_X \ge \|u\|_H^2. \tag{3.23}$$

*We thus conclude $\mathcal{V}^A(x^l)$ can be chosen as a Newton derivative of $\mathcal{F}^A(u,h)$. $\mathcal{V}^A(x^l)/D_l^A$ satisfies the regularity condition. The linear operator $\mathcal{V}^A(x^l)$ and the Schur complement $\mathcal{V}^A(x^l)/H$ satisfy the regularity condition for any fixed $\sigma_k$ and $\lambda^k$.*

We conclude the section 3 by the following algorithm 1 and 2 which are subproblems for the $k$th iteration of ALM applying to (P).

*Remark* 1. The projection to the feasible set $\{h : \|h\|_\infty \le \alpha\}$ is very important for the positive definiteness of (3.7) or (3.20). It can bring out more efficiency for solving the linear systems (3.7) or (3.20) numerically as in our numerical tests (see also [15]).

**Algorithm 1** Primal-dual semismooth Newton method with solving the primal variable first (SSNPDP) for (2.19) with auxiliary (2.23) or (3.16)

Given corrupted image $f$, multiplier $\lambda^k$, step size $\sigma_k$ of ALM, $u^0$, auxiliary variable $h^0$ in the feasible set $\{h : \|h\|_\infty \leq \alpha\}$.
Iterate the following steps for $l = 0, 1, \cdots$ , unless some stopping criterion associated with the nonlinear system (2.23) is satisfied.
**Step 1:** Solve the linear system (3.8) for $u^{l+1}$ with some stopping criterion for the isotropic case (or (3.20) for the anisotropic case) with iterative method (BiCGSTAB):
**Step 2:** Update $h^{l+1}$ by (3.6) (or (3.21) for the anisotropic case)
**Step 3:** Project $h^{l+1}$ to the feasible set $\{h : \|h\|_\infty \leq \alpha\}$, i.e., $h^{l+1} = \mathcal{P}_\alpha(h^{l+1})$. Set $(u^{l+1}, h^{l+1})$ as the initial value for the next Newton iteration and go to **Step 1**.
Output $u^{k+1}$, $h^{k+1}$

**Algorithm 2** Primal-dual semismooth Newton method with solving the dual variable first (SSNPDD) for (2.19) with auxiliary (2.23) or (3.16)

Given corrupted image $f$, multiplier $\lambda^k$, step size $\sigma_k$ of ALM, $u^0$, auxiliar variable $h^0$ in the feasible set $\{h : \|h\|_\infty \leq \alpha\}$.
Iterate the following steps for $l = 0, 1, \cdots$ , unless some stopping criterion associated with the nonlinear system (2.23) is satisfied.
**Step 1:** Unless some stopping criterion is satisfied, solve the linear system (3.11) for $h^{l+1}$ for the isotropic case (or (3.22) for the anisotropic case) with iterative method (BiCGSTAB):
**Step 2:** Update $u^{l+1} = H^{-1}(f + \operatorname{div} h^{l+1})$
**Step 3:** Project $h^{l+1}$ to the feasible set $\{h : \|h\|_\infty \leq \alpha\}$, i.e., $h^{l+1} = \mathcal{P}_\alpha(h^{l+1})$. Set $(u^{l+1}, h^{l+1})$ as the initial value for the next Newton iteration and go to **Step 1**.
Output $u^{k+1}$, $h^{k+1}$

# 4 ALM with Semismooth Newton Involving Soft Thresholding Operators

## 4.1 ALM with Semismooth Newton: the Isotropic Case

For the isotropic TV, we can rewrite the equation (2.16) as follows

$$F^I(u) = 0, \quad F^I(u) := Hu - f + \nabla^*\lambda^k + \sigma_k\nabla^*\nabla u - \sigma_k\nabla^* S^I_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u). \tag{4.1}$$

We now turn to the semismooth Newton method to solve (4.1). The main problem comes from the soft thresholding operator $S^I_{\frac{\alpha}{\sigma_k}}(\cdot)$. Still, let's introduce the active sets

$$\chi^+_{u,\hat{\lambda}^k} = \begin{cases} 1, & |\hat{\lambda}^k + \nabla u| \geq \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k + \nabla u| < \alpha/\sigma_k. \end{cases}$$

With the Moreau's indentity (2.18), with direct calculations, we have

$$S^I_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u) = \hat{\lambda}^k + \nabla u - \mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u). \tag{4.2}$$

By [4] (Corollary 2 in section 2.3.3), we have

$$\partial_u(S^I_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)) = \nabla - \partial_u(\mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)). \tag{4.3}$$

Denote $P(u) = \mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)$. Let's introduce the active sets

$$\chi^+_{u,\hat{\lambda}^k} = \begin{cases} 1, & |\hat{\lambda}^k + \nabla u| \geq \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k + \nabla u| < \alpha/\sigma_k, \end{cases} \qquad \chi^-_{u,\hat{\lambda}^k} = \begin{cases} 1, & |\hat{\lambda}^k + \nabla u| < \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k + \nabla u| \geq \alpha/\sigma_k. \end{cases} \tag{4.4}$$

Actually, we have the following lemma.

**Lemma 6.** *$P(u) = \mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)$ is a semismooth function of $u$. Furthermore, we have*

$$\left\{ A_u^{P,s}(\nabla\cdot\ ) \mid s \in [0,1] \right\} \subset (\partial_u \mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)). \tag{4.5}$$

*It means that for any $v$ and $s \in [0,1]$, we have*

$$A_u^{P,s}(\nabla\cdot\ ) = \begin{cases} A_u^-(\nabla\cdot\ ) := \nabla\cdot\ , & |\hat{\lambda}^k + \nabla u| < \alpha/\sigma_k, \\ \dfrac{\alpha}{\sigma_k}\left( \dfrac{\nabla\cdot}{|\hat{\lambda}^k + \nabla u|} - s\dfrac{\langle \hat{\lambda}^k + \nabla u, \cdot\ \rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3} \right), & |\hat{\lambda}^k + \nabla u| = \alpha/\sigma_k, \\ A_u^+(\nabla\cdot\ ) = \dfrac{\alpha}{\sigma_k}\left( \dfrac{\nabla\cdot}{|\hat{\lambda}^k + \nabla u|} - \dfrac{\langle \hat{\lambda}^k + \nabla u, \nabla\cdot\ \rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3} \right), & |\hat{\lambda}^k + \nabla u| > \alpha/\sigma_k. \end{cases} \tag{4.6}$$

*Throughout this paper, we choose $s = 1$ for the Newton derivatives in* (4.5) *and* (4.6)*, i.e.,*

$$A^P_{u,I} := \chi^+_{u,\hat{\lambda}^k} A_u^+ + \chi^-_{u,\hat{\lambda}^k} A_u^-. \tag{4.7}$$

*Proof.* The semismoothness of $P(u)$ is as follows. It is known that $L(y) = \mathcal{P}_{\frac{\alpha}{\sigma_k}}(y)$ is $PC^\infty$ function [35] (Example 5.16) (or see [9] Theorem 4.5.2 for more general projections of $PC^1$ function). Since $y(u) = \nabla u + \hat{\lambda}^k$ is differentiable and affine on $u$, $y(u)$ is also semismooth on $u$. We thus get the semismoothness of $P(u) = L(y(u))$ on $u$ [35] (Proposition 2.9).

When $|\hat{\lambda}^k + \nabla u| < \alpha/\sigma_k$,

$$P(u) = P_1(u) := \mathcal{P}_\alpha(\hat{\lambda}^k + \nabla u) = (\hat{\lambda}^k + \nabla u).$$

Since $P_1(u)$ is an affine and differentiable function on $u$, we have

$$P_1(u) = \hat{\lambda}^k + \nabla u, \quad \nabla_u P_1(u) = \nabla.$$

The Gâteaux derivative of $\mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)$ can be directly calculated as in (4.6). We thus have the first case of (4.6).

While $|\hat{\lambda}^k + \nabla u| > \frac{\alpha}{\sigma_k}$, $P(u)$ is a smooth function on $u$. Similarly, we have

$$P(u) = P_2(u) := \mathcal{P}_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u) = \frac{\alpha}{\sigma_k}\frac{(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|}, \quad |\hat{\lambda}^k + \nabla u| > \alpha/\sigma_k, \tag{4.8}$$

$$\nabla_u P_2(u) = \frac{\alpha}{\sigma_k}\left[ \frac{\nabla\cdot}{|\hat{\lambda}^k + \nabla u|} - \frac{\langle \hat{\lambda}^k + \nabla u, \nabla\cdot\ \rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3} \right], \quad |\hat{\lambda}^k + \nabla u| > \alpha/\sigma_k. \tag{4.9}$$

Actually, it can be readily verified by the directional derivative as follows.

$$\partial_u \frac{(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|}(v) = \lim_{t\to 0} \frac{1}{t}\left(\frac{\hat{\lambda}^k + \nabla u + t\nabla v}{|\hat{\lambda}^k + \nabla u + t\nabla v|} - \frac{\hat{\lambda}^k + \nabla u}{|\hat{\lambda}^k + \nabla u|}\right) \tag{4.10}$$

$$= \lim_{t\to 0} \frac{1}{t}\left(\frac{\hat{\lambda}^k + \nabla u + t\nabla v}{\sqrt{|\hat{\lambda}^k + \nabla u|^2 + 2t\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle + t^2|\nabla v|^2}} - \frac{\hat{\lambda}^k + \nabla u}{|\hat{\lambda}^k + \nabla u|}\right)$$

$$= \lim_{t\to 0} \frac{1}{t}\left(\frac{\hat{\lambda}^k + \nabla u + t\nabla v}{|\hat{\lambda}^k + \nabla u|}\left(1 - \frac{1}{2}\frac{2t\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle + t^2|\nabla v|^2}{\|\hat{\lambda}^k + \nabla u\|^2} + \mathcal{O}(t^2)\right) - \frac{\hat{\lambda}^k + \nabla u}{|\hat{\lambda}^k + \nabla u|}\right)$$

$$= \lim_{t\to 0} \frac{1}{t}\left(\frac{t\nabla v}{|\hat{\lambda}^k + \nabla u|} - \frac{t\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3} + \mathcal{O}(t^2)\right) \tag{4.11}$$

$$= \frac{\nabla v}{|\hat{\lambda}^k + \nabla u|} - \frac{\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}. \tag{4.12}$$

For any $u$ such that $|\hat{\lambda}^k + \nabla u| = \alpha/\sigma_k$, by [31] (Proposition 4.3.1), we have

$$\mathrm{co}\{(\nabla_u P_1(u)), (\nabla_u P_2(u))\} = A_u^P(\nabla\cdot) \subset \partial_u P(u),$$

which leads to (4.5) and (4.6). By [4] (Corollary 2.6.6), we have

$$\nabla^* A_u^P(\nabla v) \subset (\nabla^* \partial_u P(u))(v) = \partial_u(\nabla^* P(u))(v).$$

□

Similarly, we have the following lemma.

**Lemma 7.** *$F^I(u)$ is a semismooth function of $u$. We have*

$$A_u^S(\nabla v) \subset (\partial_u S^I_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)(v), \tag{4.13}$$

$$A_u^S(\nabla v) = \begin{cases} 0, & |\hat{\lambda}^k + \nabla u| < \frac{\alpha}{\sigma_k} \\ \left(\nabla v - \frac{\alpha}{\sigma_k}\left(\frac{\nabla v}{|\hat{\lambda}^k + \nabla u|} - s\frac{\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}\right)\right), \ s \in [0,1], & |\hat{\lambda}^k + \nabla u| = \frac{\alpha}{\sigma_k} \\ A_u^+(\nabla v) := \left(\nabla v - \frac{\alpha}{\sigma_k}\left(\frac{\nabla v}{|\hat{\lambda}^k + \nabla u|} - \frac{\langle \hat{\lambda}^k + \nabla u, \nabla v\rangle(\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}\right)\right). & |\hat{\lambda}^k + \nabla u| > \frac{\alpha}{\sigma_k} \end{cases} \tag{4.14}$$

*Throughout this paper, we choose the following generalized gradient for computations*

$$A_u^I(\nabla v) := \chi^+_{u,\hat{\lambda}^k} A_u^+(\nabla v) \subset A_u^S(\nabla v),$$

*where we always choose $s = 1$ in (4.14). The Newton derivative of (4.1) can be chosen as*

$$H + \sigma_k \nabla^*\nabla - \sigma_k \nabla^* A_u^I \nabla, \tag{4.15}$$

*which is positive definite with lower bound on $u$ and thus satisfies the regularity condition,*

$$\langle (H + \sigma_k \nabla^*\nabla - \sigma_k \nabla^* A_u^I \nabla)u, u\rangle_X \geq \|u\|_H^2. \tag{4.16}$$

*Proof.* By Lemma 6 and (4.2), we see $S^I_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u)$ is semismooth and the semismoothness of $F^I(u)$ then follows. Furthermore, by (4.1), we obtain (4.13). Since

$$H + \sigma_k \nabla^* \nabla - \sigma_k \nabla^* A^I_u \nabla \subset \partial F^I(u) = H + \sigma_k \nabla^* \nabla - \sigma_k \nabla^* \partial_u (S^I_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u)),$$

we can choose $I + \sigma_k \nabla^* \nabla - \sigma_k \nabla^* A^I_u \nabla$ as a Newton derivative for $F^I(u)$.

For the positive definiteness of the Newton derivative (4.15), we have

$$\begin{aligned}
&\langle (H + \sigma_k \nabla^* \nabla - \nabla^* A^I_u \nabla) u, u \rangle_X = \|u\|^2_H + \sigma_k \langle \nabla u, \nabla u \rangle_Y - \sigma_k \langle A^I_u \nabla u, \nabla u \rangle_Y \\
=&\|u\|^2_H + \sigma_k \langle (I - \chi^+_{u,\hat{\lambda}}) \nabla u, \nabla u \rangle_Y + \frac{\alpha}{\sigma_k} \langle \chi^+_{u,\hat{\lambda}} \frac{\nabla u}{|\hat{\lambda}^k + \nabla u|} - \chi^+_{u,\hat{\lambda}} \frac{\langle \hat{\lambda}^k + \nabla u, \nabla u \rangle (\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}, \nabla u \rangle_Y \\
=&\|u\|^2_H + \sigma_k \langle (I - \chi^+_{u,\hat{\lambda}}) \nabla u, \nabla u \rangle_Y + \frac{\alpha}{\sigma_k} \langle \chi^+_{u,\hat{\lambda}} \frac{\nabla u}{|\hat{\lambda}^k + \nabla u|}, \nabla u \rangle_Y \\
&- \alpha \langle \chi^+_{u,\hat{\lambda}} \frac{\langle \hat{\lambda}^k + \nabla u, \nabla u \rangle (\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}, \nabla u \rangle_Y \geq \|u\|^2_H,
\end{aligned}$$

since $\chi^+_{u,\hat{\lambda}} \leq I$ and by the comparison the integral functions

$$\begin{aligned}
&\langle \chi^+_{u,\hat{\lambda}} \frac{\langle \hat{\lambda}^k + \nabla u, \nabla u \rangle (\hat{\lambda}^k + \nabla u)}{|\hat{\lambda}^k + \nabla u|^3}, \nabla u \rangle = \chi^+_{u,\hat{\lambda}} \frac{\langle \hat{\lambda}^k + \nabla u, \nabla u \rangle^2}{|\hat{\lambda}^k + \nabla u|^3} \\
&\leq \chi^+_{u,\hat{\lambda}} \frac{|\hat{\lambda}^k + \nabla u|^2 |\nabla u|^2}{|\hat{\lambda}^k + \nabla u|^3} = \chi^+_{u,\hat{\lambda}} \frac{|\nabla u|^2}{|\hat{\lambda}^k + \nabla u|}.
\end{aligned}$$

□

The semismooth Newton method for solving (2.16) follows

$$(H + \sigma_k \nabla^* \nabla - \sigma_k \nabla^* A^I_{u^l} \nabla) \delta u^{l+1} = -F^I(u^l). \tag{4.17}$$

## 4.2 ALM with Semismooth Newton: the Anisotropic Case

For the anisotropic $l_1$ norm (2.1), we have

$$p = (p_1, p_2)^T = S_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k + \nabla u) = (S_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k_1 + \nabla_1 u), S_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k_2 + \nabla_2 u))^T, \tag{4.18}$$

$$p_i = S_{\frac{\alpha}{\sigma_k}}(\hat{\lambda}^k_i + \nabla_i u) = \begin{cases} \hat{\lambda}^k_i + \nabla_i u - \alpha/\sigma_k, & \hat{\lambda}^k_i + \nabla_i u > \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k_i + \nabla_i u| \leq \alpha/\sigma_k, \\ \hat{\lambda}^k_i + \nabla_i u + \alpha/\sigma_k, & \hat{\lambda}^k_i + \nabla_i u < -\alpha/\sigma_k. \end{cases} \quad i = 1, 2, \tag{4.19}$$

With (4.19), the equation (2.16) becomes

$$F^A(u) = 0, \quad F^A(u) := Hu - f + \nabla^* \lambda^k + \sigma_k \nabla^* \nabla u - \sigma_k \nabla^* (I + \frac{\alpha}{\sigma_k} \partial \|\cdot\|_1)^{-1} (\lambda^k/\sigma_k + \nabla u). \tag{4.20}$$

The Newton derivative is of critical importance for the semismooth Newton method to solve the equation (4.20). Let's first introduce

$$\chi^{+,s}_{1,\hat{\lambda}^k} = \begin{cases} 1, & |\hat{\lambda}^k_1 + \nabla_1 u| > \alpha/\sigma_k, \\ s, & |\hat{\lambda}^k_1 + \nabla_1 u| = \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k_1 + \nabla_1 u| < \alpha/\sigma_k, \end{cases} \qquad \chi^{+,s}_{2,\hat{\lambda}^k} = \begin{cases} 1, & |\hat{\lambda}^k_2 + \nabla_2 u| > \alpha/\sigma_k, \\ s, & |\hat{\lambda}^k_2 + \nabla_2 u| = \alpha/\sigma_k, \\ 0, & |\hat{\lambda}^k_2 + \nabla_2 u| < \alpha/\sigma_k. \end{cases} \quad s \in [0, 1].$$

Introduce $\chi^{+}_{i,\hat{\lambda}^k} := \chi^{+,s}_{1,\hat{\lambda}^k}$ for $s = 1$ with $i = 1, 2$. For the Newton derivative of equation (4.20), we have the following lemma.

**Lemma 8.** *The function $F^A(u)$ is semismooth on $u$. We have*

$$\left\{\chi^{i,s}_{\lambda^k,u}\nabla_i \mid s \in [0,1]\right\} = \partial_u(S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k_i}{\sigma_k} + \nabla_i u)), \quad i = 1, 2. \tag{4.21}$$

*The Newton derivative of the equation* (4.20) *can be choose as*

$$H + \sigma_k\nabla^*\nabla - \sigma_k\nabla^* A^A_u(\nabla), \tag{4.22}$$

*where $A^A_u(\nabla)$ can be chosen as the Newton derivative of $H(u)$*

$$A^A_u(\nabla) = \begin{bmatrix} \chi^{+}_{1,\hat{\lambda}^k} & 0 \\ 0 & \chi^{+}_{2,\hat{\lambda}^k} \end{bmatrix} \begin{bmatrix} \nabla_1 \\ \nabla_2 \end{bmatrix}. \tag{4.23}$$

*Furthermore, the Newton derivative is positive definite and satisfies the regularity condition,*

$$\langle (H + \sigma_k\nabla^*\nabla - \sigma_k\nabla^* A^A_u(\nabla))u, u\rangle_X \geq \|u\|^2_H. \tag{4.24}$$

*Proof.* It can be seen that $H_i(u) = S_{\frac{\alpha}{\sigma_k}}(\lambda^k_i/\sigma_k + \nabla_i u)$ is piecewise differentiable, $i = 1, 2$. Take $H_1(u)$ for example. Denote $H^1_1(u) = \lambda^k_1/\sigma_k + \nabla_1 u - \alpha/\sigma_k$ and $H^2_1(u) = 0$ being *selection functions* of $H_1(u)$. For $|\lambda^k_1/\sigma_k + \nabla_1 u - \alpha/\sigma_k| > \alpha/\sigma_k$,

$$H_1(u) = H^1_1(u), \quad \nabla_u H_1(u) = \nabla_u H^1_1(u) = \nabla_1,$$

and while $|\lambda^k_1/\sigma_k + \nabla_1 u - \alpha/\sigma_k| < \alpha/\sigma_k$,

$$H_1(u) = H^2_1(u), \quad \nabla_u H_1(u) = \nabla_u H^2_1(u) = 0.$$

Finally, while $|\lambda^k_1/\sigma_k + \nabla_1 u - \alpha/\sigma_k| = \alpha/\sigma_k$,

$$\partial_u H_1(u) = \mathrm{co}\{\nabla_u H^1_1(u), \nabla_u H^2_1(u)\} = \mathrm{co}\{0, \nabla_1\},$$

which leads to (4.21). By Lemma 1, we found that $A_k(u)$ in (4.23) can be chosen as a Newton derivative of $H(u)$. Still, since $\nabla^*$ is a linear, bounded and thus is a differentiable mapping, the Newton derivative of $F^A(u)$ can be choose as (4.22) [19] (Lemma 8.15). For the positive definiteness, with $\chi^{+}_{1,\hat{\lambda}^k} \leq I$ and $\chi^{+}_{2,\hat{\lambda}^k} \leq I$, we see

$$\begin{aligned}\langle (H + \sigma_k\nabla^*(\nabla - A^A_u(u)))u, u\rangle_X &= \|u\|^2_H + \sigma_k\left[\langle \nabla_1 u, (I - \chi^{+}_{1,\hat{\lambda}^k})\nabla_1 u\rangle_X + \langle \nabla_2 u, (I - \chi^{+}_{2,\hat{\lambda}^k})\nabla_2 u\rangle_X\right] \\ &\geq \|u\|^2_H.\end{aligned}$$

□

Then the semismooth Newton method for solving equation (4.20) follows

$$(H + \sigma_k\nabla^*\nabla - \sigma_k\nabla^* A^A_{u^l}(\nabla))(\delta u^{l+1}) = -F^A(u^l). \tag{4.25}$$

It can be checked that the linear operator in (4.15) is also self-adjoint. We thus can use the efficient conjugate gradient (CG) to solve (4.17) and (4.25). We conclude this section by the following Algorithm 3, i.e., the semismooth Newton method for the primal problem involving with the soft thresholding operator (SSNPT) (2.16), which is subproblem for the $k$th iteration of ALM applying to (P). The soft thresholding operators are frequently seen in compressed sensing and we refer [23] for the celebrated framework of semismooth Newton based ALM for compressed sensing, which gives us a lot of inspiration.

**Algorithm 3** Semismooth Newton method with solving primal problem involving the soft thresholding operator (SSNPT) for (2.16)

Given corrupted image $f$, multiplier $\lambda^k$, initial value for Newton $\delta u^0 = 0$, step size $\sigma_k$ of ALM, choose $\mu \in (0, 1/2)$, $\theta \in (0, 1)$, $\eta^0 = 1$.
Iterate the following steps for $l = 0, 1, \cdots$ , unless some stopping criterion associated with the nonlinear system (2.16) is satisfied.
**Step 1**: Unless some stopping criterion is satisfied, solve the linear system (4.17) for $\delta u^{k+1}$ of the isotropic case with iterative method (CG) (or solve (4.25) of the anisotropic case with CG, i.e., Conjugate Gradient method)
**Step 2**: Do backtracking Armijo line search as follows:
Compute $\xi^l = (H+\sigma_k\nabla^*\nabla-\sigma_k\nabla^* A^X_{u^l}(u^l))(u^l) \in \partial_u\Phi(u;\lambda^k,\sigma_k)|_{u=u^l}$. $A^X_{u^l} = A^I_{u^l}$ or $A^X_{u^l} = A^A_{u^l}$ and $\Phi(u;\lambda^k,\sigma_k)$ is in (2.15) depending on the isotropic or the anisotropic case.
**while**

$$\Phi(u^l + \eta^j\delta u^{l+1};\lambda^k,\sigma_k) > \Phi(u^l;\lambda^k,\sigma_k) + \mu\eta^j\langle\xi^l,\delta u^{l+1}\rangle,$$

**do**

$$\eta^{j+1} = \eta^j\theta,$$

**end while**
**Step 3**: Update $u^{l+1} = u^l + \eta^j\delta u^{l+1}$
Output $u^{k+1}$

# 5 Convergence of the Augmented Lagrangian Methods and the Corresponding Semismooth Newton Methods

For the proposed semismooth Newton based ALM methods, we need the convergence both for the inner semismooth Newton iterations and the outer ALM iterations. Let's begin with the inner Newton iterations followed by the convergence of ALM.

## 5.1 Convergence of the semismooth Newton method

**Theorem 1** (Superlinear Convergence [19])**.** *Suppose $x^*$ is a solution to $F(x) = 0$ and that $F$ is Newton differentiable at $x^*$ with Newton derivative $G$. If $G$ is nonsingular for all $x \in N(x^*)$ and $\{\|G(x)^{-1}\| : x \in N(x^*)\}$ is bounded, then the Newton iteration*

$$x^{l+1} = x^l - G(x^l)^{-1}F(x^l),$$

*converges superlinearly to $x^*$ provided that $|x^0 - x^*|$ is sufficiently small.*

For the semismooth Newton method, once the Newton derivative exists and is nonsingular at the solution point, we can employ semismooth Newton method [19] possibly with some globalization strategy [7]. We now turn to the semismoothness for our cases, where each of the Newton derivative satisfies the regularity condition and is nonsingular (uniform regular) with a lower bound; see lemmas 3, 5, 8, 7. For the convergence of the semismooth Newton methods including Algorithm 1 and 2, we refer to [15, 34] for the analysis of semismooth Newton under perturbations. For the convergence of the semismooth Newton method in Algorithm 3 with line search, we refer to [23] (Theorem 3.6) and [41] (Theorems 3.4 and 3.5).

Here, we follow the standard stopping criterion for the inexact augmented Lagrangian method

[28, 29] and [23, 40].

$$\Phi_k(u^{k+1}, h^{k+1}) - \inf \Phi_k \leq \epsilon_k^2/2\sigma_k, \quad \sum_{k=0}^{\infty} \epsilon_k < \infty, \tag{A}$$

$$\Phi_k(u^{k+1}, h^{k+1}) - \inf \Phi_k \leq \delta_k^2/2\sigma_k \|\lambda^{k+1} - \lambda^k\|^2, \quad \sum_{k=0}^{\infty} \delta_k < +\infty, \tag{B1}$$

$$\mathrm{dist}(0, \partial\Phi_k(u^{k+1}, h^{k+1})) \leq \delta_k'/\sigma_k \|\lambda^{k+1} - \lambda^k\|, \quad 0 \leq \delta_k' \to 0. \tag{B2}$$

We conclude this section with the following algorithmic framework of ALM with Algorithm 4. Henceforth, we denote ALM-PDP or ALM-PDD as the ALM with the Algorithm 1 (SSNPDP) or Algorithm 2 (SSNPDD). We also denote ALM-PT as the ALM with the Algorithm 3 (SSNPT).

**Algorithm 4** General framework of semismooth Newton based ALM for (P)

**Input**: corrupted image $f$, regularization parameter $\alpha$ for TV, initial Lagrangian multiplier $\lambda^0$, initial step size $\sigma_0$ and the largest step size $\sigma_\infty$ of ALM, step size update parameter $c > 1$

**Iteration**: Iterate the following steps for $k = 0, 1, \cdots$, unless some stopping criterion associated with the problem (P) is satisfied.

**Step 1**: With $\sigma_k$, $\lambda^k$, solve the minimization problem (2.10) with $u^k$ (possibly $h^k$ for Algorithm 1 or Algorithm 2) by

Algorithm 1 or Algorithm 2 for the nonlinear system (2.23) or (3.16) for $(u^{k+1}, h^{k+1})$; or Algorithm 3 for the nonlinear system (2.16) for $u^{k+1}$.

**Step 2**: Update Lagrangian multiplier $\lambda^{k+1}$:

For Algorithm 3 using (2.11), i.e.,

$$\lambda^{k+1} = \lambda^k + \sigma_k(\nabla u^{k+1} - p^{k+1}), \quad \text{with} \;\; p^{k+1} = S_{\frac{\alpha}{\sigma_k}}(\frac{\lambda^k}{\sigma_k} + \nabla u^{k+1}).$$

For Algorithm 1 or 2 using (2.24), i.e.,

$$\lambda^{k+1} = \mathcal{P}_\alpha(\lambda^k + \sigma_k \nabla u^{k+1}).$$

**Step 3**: Update step size $\sigma_{k+1} = c\sigma_k \leq \sigma_\infty$.

## 5.2 Convergence of the Augmented Lagrangian Method

It is well-known that the augmented Lagrangian method can be seen as applying proximal point algorithm to the dual problem [28, 29]. The convergence and the corresponding rate of augmented Lagrangian method are closely related to the convergence of the proximal point algorithm. Especially, the local linear convergence of the multipliers, primal or dual variables is mainly determined by the metric subregularities of the corresponding maximal monotone operators [21, 24, 28, 29]. We now turn to analysis the metric subregularity of the corresponding maximal monotone operators which is usually efficient for the asymptotic (or local) linear convergence of ALM.

Now we introduce some basic definitions and properties of multivalued mapping from convex analysis [8, 23]. Let $F : X \Longrightarrow Y$ be a multivalued mapping. The graph of $F$ is defined as the set

$$\mathrm{gph}F := \{(x, y) \in X \times Y | y \in F(x)\}.$$

The inverse of $F$, i.e., $F^{-1} : Y \Longrightarrow X$ is defined as the multivalued mapping whose graph is $\{(y, x)|(x, y) \in \text{gph}F\}$. The distance $x$ from the set $C \subset X$ is defined by

$$\text{dist}(x, C) := \inf\{\|x - x'\| \mid x' \in C\}.$$

Let's introduce the metrical subregularity for $F$ and the calmness [8, 23].

**Definition 3** (Metric Subregularity [8])**.** A mapping $F : X \Longrightarrow Y$ is called metrically subregular at $\bar{x}$ for $\bar{y}$ if $(\bar{x}, \bar{y}) \in \text{gph}F$ and there exists modulus $\kappa \geq 0$ along with a neighborhoods $U$ of $\bar{x}$ and $V$ of $\bar{y}$ such that

$$\text{dist}(x, F^{-1}(\bar{y})) \leq \kappa \text{dist}(\bar{y}, F(x) \cap V) \quad \text{for all} \quad x \in U. \tag{5.1}$$

**Definition 4** (Calmness [8])**.** A mapping $S : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is called calm at $\bar{y}$ for $\bar{x}$ if $(\bar{y}, \bar{x}) \in \text{gph } S$, and there is a constant $\kappa \geq 0$ along with neighborhoods $U$ of $\bar{x}$ and $V$ of $\bar{y}$ such that

$$S(y) \cap U \subset S(\bar{y}) + \kappa|y - \bar{y}|\mathbb{B}, \quad \forall y \in V. \tag{5.2}$$

In (5.2), $\mathbb{B}$ denotes the closed unit ball in $\mathbb{R}^n$.

Let's now vectorize the variables disscussed in detail. Suppose

$$\begin{aligned}
&\nabla_1 \in \mathbb{R}^{m\times n} : \mathbb{R}^n \to \mathbb{R}^m, \quad \nabla_2 \in \mathbb{R}^{m\times n} : \mathbb{R}^n \to \mathbb{R}^m, \quad u = (u^1, \cdots, u^n)^T \in \mathbb{R}^n, \\
&p = (p_1, \cdots, p_m)^T \in \mathbb{R}^{2m}, \quad p_i = (p_i^1, p_i^2)^T \in \mathbb{R}^2, \\
&\lambda = (\lambda_1, \cdots, \lambda_m)^T \in \mathbb{R}^{2m}, \quad \lambda_i = (\lambda_i^1, \lambda_i^2)^T \in \mathbb{R}^2.
\end{aligned}$$

For the anisotropic case, we see

$$\|p\|_1 = \sum_{i=1}^m \|p_i\|_1 = \sum_{i=1}^m |p_i^1| + \sum_{i=1}^m |p_i^2| = \sum_{i=1}^m (|p_i^1| + |p_i^2|),$$

which is a polyhedral function. For the istropic case, we notice

$$\|p\|_1 = \sum_{i=1}^m |p_i| = \sum_{i=1}^m \sqrt{{p_i^1}^2 + {p_i^2}^2}$$

which is not a polyhedral function. Fortunately, it is a group Lasso norm [37, 40]. Now, let's turn to the anisotropic case first. Introduce the Lagrangian function

$$l(u, p, \lambda) = \frac{1}{2}\langle Qu, u\rangle_2 - \langle u, b\rangle + \frac{1}{2}\|z\|_2^2 + \alpha\|p\|_1 + \langle \nabla u - p, \lambda\rangle, \tag{5.3}$$

where $Q$ is a linear, positive semidefinite operator and $b$ is known. For (P), we have $Q = H = -\mu\Delta + K^*K$ and $b = K^*z$. It is well-known that $l$ is a convex-concave function on $(u, p, \lambda)$. Define the maximal monotone operator $T_l$ by

$$T_l(u, p, \lambda) = \{(u', p', \lambda')|(u', p', -\lambda') \in \partial l(u, p, \lambda)\}, \tag{5.4}$$

z and the corresponding inverse is given by

$$T_l^{-1}(u', p', \lambda') = \{(u, p, \lambda)|(u', p', -\lambda') \in \partial l(u, p, \lambda)\}. \tag{5.5}$$

**Theorem 2.** *For the anisotropic ROF model, assuming the KKT system has at least one solution, then $T_l$ is metrically subregular at $(\bar{u}, \bar{p}, \bar{\lambda})^T$ for the origin.*

*Proof.* Let's consider the general case including the case in (P). Actually, we have

$$T_l(u,p,\lambda) = (Qu-b, -\lambda+\alpha\partial\|p\|_1, p-\nabla u)^T := \mathcal{A}(x)+\mathcal{B}(x),$$

where

$$\mathcal{A}\begin{pmatrix}u\\p\\\lambda\end{pmatrix} := \begin{pmatrix}0 & 0 & 0\\ 0 & \alpha\partial\|\cdot\|_1 & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}u\\p\\\lambda\end{pmatrix}, \quad \mathcal{B}\begin{pmatrix}u\\p\\\lambda\end{pmatrix} := \begin{pmatrix}Q & 0 & \nabla^*\\ 0 & 0 & -I\\ -\nabla & I & 0\end{pmatrix}\begin{pmatrix}u\\p\\\lambda\end{pmatrix} + \begin{pmatrix}-b\\0\\0\end{pmatrix}. \tag{5.6}$$

It can be seen that the monotone operator $\mathcal{A}$ is polyhedral since the anisotropic $\|\cdot\|_1$ (2.1) is a polyhedral convex function and the operator $\mathcal{B}$ is a maximal monotone and affine operator. Thus $T_l$ is a polyhedral mapping [27]. By the corollary in [27], we see $T_l$ is metrically subregular at $(\bar{u},\bar{p},\bar{\lambda})^T$ for the origin. □

Let's now turn to the metric subregularity of $\partial d$ for the dual problem (D), supposing $(\partial d)^{-1}(0) \neq \emptyset$ and there exists $\bar{\lambda}$ such that $0 \in (\partial d)(\bar{\lambda})$,

$$(\partial d)(\lambda) = \operatorname{div}^* H^{-1}(\operatorname{div}\lambda + f) + \partial g(\lambda), \quad g(\lambda) := \mathcal{I}_{\{\|\lambda\|_\infty \le \alpha\}}(\lambda). \tag{5.7}$$

For the anisotropic case, actually, the constraint set is a polyhedral convex set in $\mathbb{R}^{2m}$, since

$$g(\lambda) = 0 \Leftrightarrow \left\{\lambda = (\lambda_1,\lambda_2,\cdots,\lambda_m)^T \mid \lambda_i \in \mathbb{R}^2,\ |\lambda_i^k| \le \alpha,\ i=1,2,\cdots,m,\ k=1,2\right\}.$$

Together with $\operatorname{div}^* H^{-1}(\operatorname{div}\lambda + f)$ being an affine and monotone mapping, $\partial g$ is a polyhedral mapping by [27]. This leads to that $\partial g$ is metrically subregular at $\bar{\lambda}$ for the origin with similar argument as in Theorem 2.

Now we turn to the isotropic case. The metric subregularity of $\partial d$ is more subtle, since the constraint set

$$g(\lambda) = 0 \Leftrightarrow \left\{\lambda = (\lambda_1,\cdots,\lambda_m)^T \mid \lambda_i \in \mathbb{R}^2,\ |\lambda_i| = \sqrt{(\lambda_i^1)^2+(\lambda_i^2)^2} \le \alpha,\ i=1,\cdots,m\right\}$$

is not a polyhedral set. Now introduce

$$g_i(\lambda_i) = \mathcal{I}_{\{|\lambda_i|\le\alpha\}}(\lambda_i), \quad i=1,2,\cdots,m.$$

Henceforth, let's denote $\mathbb{B}_a(\bar{\lambda}_i)$ or $\mathbb{B}_a^k(\bar{\lambda}_i)$ as the Euclidean closed ball centered at $\bar{\lambda}_i \in \mathbb{R}^2$ with radius $a$, i.e.,

$$\mathbb{B}_\alpha^i(0) := \left\{\lambda_i := (\lambda_i^1,\lambda_i^2)^T \in \mathbb{R}^2 \mid |\lambda_i| = \sqrt{(\lambda_i^1)^2+(\lambda_i^2)^2} \le \alpha\right\},\ \ i=1,\cdots,m,\ \alpha>0. \tag{5.8}$$

Furthermore, denote $\mathbb{B}_a(\lambda) = \Pi_{i=1}^m \mathbb{B}_a(\bar{\lambda}_i)$ with $\bar{\lambda} = (\bar{\lambda}_1,\cdots,\bar{\lambda}_m)^T$. We can thus write

$$\partial g = \Pi_{i=1}^m \partial g_i = \Pi_{i=1}^m \partial\mathcal{I}_{\mathbb{B}_\alpha^i(0)}(\lambda_i).$$

It is known that each $\partial\mathcal{I}_{\mathbb{B}_\alpha^i(0)}(\lambda_i)$ is metrically subregular at $(\bar{\lambda}_i,\bar{v}_i) \in \operatorname{gph}\partial\mathcal{I}_{\mathbb{B}_\alpha^i(0)}$ [37] (see Lemma 6 therein). For the metric subregularity of $\partial g$, we have the following lemma.

**Lemma 9.** *For any $(\bar{\lambda},\bar{v})^T \in \operatorname{gph}\,\partial g$, $\partial g$ is metrically subregular at $\bar{\lambda}$ for $\bar{v}$.*

*Proof.* For any $(\bar{\lambda}, \bar{v})^T \in \text{gph}\ \partial g$, and $V$ of a neignborhoods of $\bar{\lambda}$, since

$$\begin{aligned}
&\text{dist}^2(\lambda, (\partial g)^{-1}(\bar{v})) = \sum_{i=1}^{m} \text{dist}^2(\lambda_i, (\partial g_i)^{-1}(\bar{v}_i)) \\
&\leq \sum_{i=1}^{m} \kappa_i^2 \text{dist}^2(\bar{v}_i, (\partial g_i)(\bar{\lambda}_i)) \leq \sum_{i=1}^{m} \max(\kappa_i^2, i = 1, \cdots, m) \text{dist}^2(\bar{v}_i, (\partial g_i)(\bar{\lambda}_i)) \\
&= \max(\kappa_i^2, i = 1, \cdots, m) \text{dist}^2(\bar{v}, (\partial g)(\bar{\lambda})).
\end{aligned}$$

Thus with choice $\kappa = \sqrt{\max_{i=1}^{m}(\kappa_i^2, i = 1, \cdots, m)}$, we found that $\partial g$ is metrically subregular at $\bar{\lambda}$ for $\bar{v}$ with modulus $\kappa$. □

By [15] (Theorem 2.1) (or formula (3.4) of [20]), the solution $\bar{u}$ of the primal problem (P) and the solution $\bar{\lambda}$ of the dual problem (D) have the following relations

$$H\bar{u} - \text{div}\,\bar{\lambda} = f, \tag{5.9}$$

$$-\alpha \nabla \bar{u} + |\nabla \bar{u}| \bar{\lambda} = 0, \quad |\bar{\lambda}| = \alpha, \tag{5.10}$$

$$\nabla \bar{u} = 0, \quad |\bar{\lambda}| < \alpha, \tag{5.11}$$

which can be derived from the optimality conditions (2.6). Now we turn to a more general model compared to (D). Suppose $A \in \mathbb{R}^{m \times 2m} : \mathbb{R}^{2m} \to \mathbb{R}^m$,

$$f(\lambda) = f_1(\lambda) + \Pi_{i=1}^{m} \mathcal{I}_{\mathbb{B}_\alpha^i(0)}(\lambda_i), \quad f_1(\lambda) := \frac{\|A\lambda - b\|^2}{2} + \langle q, \lambda \rangle + a_0, \quad b \in \mathbb{R}^m, \tag{5.12}$$

where $\mathcal{I}_{\mathbb{B}_\alpha^i(0)}(x)$ is the indicator function for the $l_2$ ball in (5.8) and $a_0$ is a constant. We claim that the model (5.12) also covers the dual problem (D) by setting $A = H^{-\frac{1}{2}}\,\text{div}$, $b = -H^{-\frac{1}{2}} f$ and $a_0 = \frac{1}{2}\|z\|_2^2$. For the existence of the square root $H^{1/2}$ of the positive definite linear operator $H$, we refer to [26] (Theorem VI.9). Supposing $g(\lambda) := \Pi_{i=1}^{m} \mathcal{I}_{\mathbb{B}_\alpha^i(0)}(\lambda_i)$, let's introduce

$$\begin{aligned}
&\mathcal{X} := \{\lambda \mid A\lambda = \bar{y}, \quad -\bar{g} \in \partial g(\lambda)\}, \\
&\Gamma_1(p^1) = \{\lambda \mid A\lambda - \bar{y} = p^1\}, \quad \Gamma_2(p^2) = \{\lambda \mid p^2 \in \bar{g} + \partial g(\lambda)\}, \\
&\hat{\Gamma}(p^1) = \Gamma_1(p^1) \cap \Gamma_2(0) = \{\lambda \mid p^1 = A\lambda - \bar{y},\ 0 \in \bar{g} + \partial g(\lambda)\},
\end{aligned}$$

where $\mathcal{X}$ is actually the solution set of (5.12), since

$$\bar{g} := A^T \nabla h(\bar{y}) + q = (\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m)^T,\ h(y) = \|y - b\|^2/2, \quad \bar{g}_i \in \mathbb{R}^2.$$

We also need another two set valued mapping,

$$\Gamma(p^1, p^2) := \{\lambda \mid p^1 = A\lambda - \bar{y}, \quad p^2 \in \bar{g} + \partial g(\lambda)\}, \tag{5.13}$$

$$S(p) := \{\lambda \mid p \in \nabla f_1(\lambda) + \partial g(\lambda)\} \Rightarrow \mathcal{X} = S(0). \tag{5.14}$$

Actually the metric subregularity of $\partial f$ at $(\bar{\lambda}, 0)$ is equivalent to the calmness $S$ at $(0, \bar{\lambda})$ [8]. Now we turn to the calmness of $S$. By [37] (Theorem 5), the calmness of $S$ at $(0, \bar{\lambda})$ is equivalent to the calmness of $\hat{\Gamma}$ at $(0, \bar{\lambda})$ for any $\bar{\lambda} \in S(0)$ if $\partial g$ is metrically subregular for $f$ in (5.12). We would use the following calm intersection theorem to prove the calmness of $\Gamma$.

**Proposition 1** (Calm intersection theorem [17, 20, 37])**.** *Let $T_1 : \mathbb{R}^{q_1} \rightrightarrows \mathbb{R}^n$, $T_2 : \mathbb{R}^{q_2} \rightrightarrows \mathbb{R}^n$ be two set-valued maps. Define set-valued maps*

$$T(p^1, p^2) := T_1(p^1) \cap T_2(p^2), \tag{5.15}$$

$$\hat{T}(p^1) := T_1(p^1) \cap T_2(0). \tag{5.16}$$

*Let $\tilde{x} \in T(0,0)$. Suppose both the set-valued maps $T_1$ and $T_2$ are calm at $(0,\tilde{x})$ and $T_1^{-1}$ is pseudo-Lipschitiz at $(0,\tilde{x})$. Then $T$ is calm at $(0,0,\tilde{x})$ if and only if $\hat{T}$ is calm at $(0,\tilde{x})$.*

We need the following assumption first, which is actually a mild condition on the solution set $\mathcal{X}$ of the dual problem (5.12) by (5.10) and (5.11).

**Assumption 1.** *With the normal cone $\mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i) = \partial\mathcal{I}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i)$, let's assume that $\bar{\lambda} \in \mathcal{X}$ and*

i. *Either $\bar{\lambda}_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0)$ and there exists $\bar{g}_i \neq 0$ such that $-\bar{g}_i \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i)$,*

ii. *Either $\bar{\lambda}_i \in \mathrm{int}\mathbb{B}^i_\alpha(0)$.*

**Theorem 3.** *For the dual problem (5.12), supposing it has at least one solution $\bar{\lambda}$ satisfying the Assumption 1, then $\partial f$ is metrically subregular at $\bar{\lambda}$ for the origin.*

*Proof.* We mainly need to prove the calmness of $\hat{\Gamma}(p^1)$ at $(0,\bar{\lambda})$. By metric subregularity of $\partial g$ by Lemma 9 and the fact that $\Gamma_1^{-1}$ is pseudo-Lipschitiz and bounded metrically subregular at $(0,\bar{\lambda})$ ([37], Lemma 3), with the Calm intersection theorem as in Proposition 1, we get calmness of $\Gamma$ at $(0,0,\bar{\lambda})$. We thus get the calmness of $S$ at $(0,\bar{\lambda})$ and the metric subregular of $\partial f$ at $\bar{\lambda}$ for the origin.

Now let's focus the the calmness of $\hat{\Gamma}(p^1)$ at $(0,\bar{\lambda})$. Without loss of generality, suppose

$$\begin{aligned}
&\bar{\lambda}_i \in \mathrm{int}\mathbb{B}^i_\alpha(0), \quad i = 1, \cdots, L;\\
&\bar{\lambda}_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0), \quad -\bar{g}_i \neq 0 \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i), \quad i = L+1, \cdots, m, \quad 1 < L < m.
\end{aligned}$$

The existence of $\bar{g}_i \neq 0$ is guaranteed by the Assumption 1. For $i = 1, \cdots, L$, $\bar{\lambda}_i \in \mathrm{int}\mathbb{B}^i_\alpha(0)$, we have $\bar{g}_i \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i) = \{0\}$. We thus conclude $\bar{g}_i = 0$ and

$$\Gamma_2^i(0) = \{\lambda_i | 0 \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\lambda_i)\} = \mathbb{B}^i_\alpha(0), \quad i = 1, \cdots, L. \tag{5.17}$$

Now, let's turn to $\Gamma_2^i(0)$ for $i = L+1, \cdots, m$. While $\bar{\lambda}_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0)$, by the Assumption 1 on the solution $\bar{\lambda}$, there exists $\bar{g}_i \neq 0$ such that $-\bar{g}_i \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i)$. Acutally, we know that the normal cone at the boundary point of a closed ball is

$$\mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\lambda_i) = \{s\lambda_i | s \geq 0\}, \quad \forall \lambda_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0).$$

Still with the definition of $\Gamma_2^i(0)$, we have

$$\Gamma_2^i(0) = \{\lambda_i | -\bar{g}_i \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\lambda_i)\}.$$

While $\lambda_i \in \mathrm{int}\mathbb{B}^i_\alpha(0)$, by the definition of $\Gamma_2(0)$, together with $\mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\lambda_i) = \{0\}$, we see $\bar{g}_i = 0$, which is contracted with the assumption 1. While $\lambda_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0)$, together with the definition of $\Gamma_2$, we see $\lambda_i = -\bar{g}_i \frac{\alpha}{|\bar{g}_i|}$ with some $\frac{\alpha}{|\bar{g}_i|} > 0$. Since $\bar{g}_i$ is fixed as in definition of $\Gamma_2^i$ for $i = L+1, \cdots, m$ and $-\bar{g}_i \in \mathcal{N}_{\mathbb{B}^i_\alpha(0)}(\bar{\lambda}_i)$ and $\bar{\lambda}_i \in \mathrm{bd}\mathbb{B}^i_\alpha(0)$, we have $\lambda_i = \bar{\lambda}_i = -\bar{g}_i \frac{\alpha}{|\bar{g}_i|}$ and the only choice is

$$\Gamma_2^i(0) = \{\bar{\lambda}_i\}, \quad i = L+1, \cdots, m. \tag{5.18}$$

Choose $\epsilon > 0$ small enough such that $\mathbb{B}^i_{4\epsilon}(\bar{\lambda}_i) \subset \mathbb{B}_\alpha(0)$, $i = 1, \cdots, L$. We thus conclude that

$$\Gamma_2(0) \cap \mathbb{B}_\epsilon(\bar{\lambda}) = (\mathbb{B}^1_\epsilon(\bar{\lambda}_1), \cdots, \mathbb{B}^L_\epsilon(\bar{\lambda}_L), \bar{\lambda}_{L+1}, \cdots, \bar{\lambda}_m)^T. \tag{5.19}$$

Suppose $p = (p_1, p_2, \cdots, p_m)^T$ and $\lambda \in \Gamma_1(p) \cap \Gamma_2(0) \cap \mathbb{B}_\epsilon(\bar{\lambda})$, $p_i \in \mathbb{R}^2$, $i = 1, 2, \cdots, m$. Introduce the following constraint on $\lambda$

$$\mathcal{R} := \{\lambda \mid \lambda_i \in \mathbb{R}^2,\ i = 1, \cdots, L,\ \lambda_i = \bar{\lambda}_i,\ i = L+1, \cdots, m\},$$

which is a convex and closed polyhedral set. It can be seen as follows. Denote $\bar{L} = m - L$ and $0_{2L\times 2m} \in \mathbb{R}^{2L\times 2m}$, $0_{2\bar{L}\times 2L} \in \mathbb{R}^{2\bar{L}\times 2L}$ as the zero matrix with elements are all zero. Denote $I_{2\bar{L}\times 2\bar{L}} \in \mathbb{R}^{2\bar{L}\times 2\bar{L}}$ as the identity matrix. Introduce

$$E_+ = [0_{2L\times 2m}; 0_{2\bar{L}\times 2L}\ I_{2\bar{L}\times 2\bar{L}}] \in \mathbb{R}^{2m\times 2m}, \quad E_- = [0_{2L\times 2m}; 0_{2\bar{L}\times 2L}\ -I_{2\bar{L}\times 2\bar{L}}] \in \mathbb{R}^{2m\times 2m},$$

$$E = [E_+; E_-] \in \mathbb{R}^{4m\times 2m}, \quad \bar{\lambda}_0 = [0, \cdots, 0, \bar{\lambda}_{L+1}, \cdots, \bar{\lambda}_m] \in \mathbb{R}^{2m}.$$

The set $\mathcal{R} = \{\lambda \mid E\lambda \le E\bar{\lambda}_0\}$ is thus a polyhedral set. Actually, the following set

$$M(p) := \{\lambda \mid A\lambda - \bar{y} = p, \quad \lambda \in \mathcal{R}\} = \{\lambda \mid A\lambda - \bar{y} = p,\ E\lambda \le E\bar{\lambda}_0\}, \tag{5.20}$$

is also a polyhedral set.

For any $\lambda \in \Gamma_1(p) \cap \Gamma_2(0) \cap \mathbb{B}_\epsilon(\bar{\lambda}) = \hat{\Gamma}(p) \cap \mathbb{B}_\epsilon(\bar{\lambda})$, denote $\tilde{\lambda}$ as its projection on $M(0)$. Since $\bar{\lambda} \in M(0)$, we thus have

$$\|\lambda - \tilde{\lambda}\| \le \|\lambda - \bar{\lambda}\| \le \epsilon \Rightarrow \tilde{\lambda} \in \mathbb{B}_\epsilon(\lambda) \Rightarrow \tilde{\lambda} \in \mathbb{B}_\alpha(0).$$

Together with $\tilde{\lambda} \in M(0)$ and $\tilde{\lambda} \in \mathcal{R}$, we see $\tilde{\lambda} \in \Gamma_2(0)$ by (5.17) and (5.18). We thus conclude that $\tilde{\lambda} \in \hat{\Gamma}(0) = \Gamma_1(0) \cap \Gamma_2(0)$. By the celebrated results of Hoffman error bound [16] based on the polyhedral set in (5.20), for any $\lambda \in \hat{\Gamma}(p) \cap \mathbb{B}_\epsilon(\bar{\lambda})$, there exists a constant $\kappa$ such that

$$\operatorname{dist}(\lambda, \hat{\Gamma}(0)) \le \|\lambda - \tilde{\lambda}\| = \operatorname{dist}(\lambda, M(0)) \le \kappa\|p\|, \quad \forall \lambda \in \hat{\Gamma}(p) \cap \mathbb{B}_\epsilon(\bar{\lambda}), \tag{5.21}$$

since $E\lambda \le E\bar{\lambda}_0$ by $\lambda \in \hat{\Gamma}(p) = \Gamma_1(p) \cap \Gamma_2(0)$. We thus get the calmness of $\hat{\Gamma}(p)$ at $(0, \bar{\lambda})$. While $L = 0$, i.e., $\bar{\lambda}_i \in \operatorname{bd}\mathbb{B}^i_\alpha(0)$, $i = 1, \cdots, m$, one can readily check that $\hat{\Gamma}(p) = \Gamma_1(p) \cap \Gamma_2(0) = \emptyset$ whenever $p_i \neq 0$ with $p = (p_1, \cdots, p_m)^T$. The calmness follows by definition and the proof is thus finished. □

*Remark* 2. Similar result is also given in [38] (see Example 4.1(ii) of [38]), where the delicate analysis based on LMI-representable (Linear Matrix Inequalities representable) functions are employed. It is proved that while the solution $\bar{\lambda} \in \mathcal{X}$ satisfies $0 \in \operatorname{ri} \partial f(\bar{\lambda})$, then $f$ has the Kurdyka-Łojasiewicz exponent 1/2 where "ri" denotes the relative interior.

*Remark* 3. For the the Assumption 1 for the model (D), with the optimality conditions (5.7) and (5.9) and the notation $(\nabla\bar{u})_k$ for the vectorized $\nabla\bar{u}$, we see $\bar{u} = H^{-1}(\operatorname{div}\bar{\lambda} + f)$ and

$$0 \in -\nabla\bar{u} + \partial g(\bar{\lambda}) \to (\nabla\bar{u})_k \in \mathcal{N}_{\mathbb{B}^k_\alpha(0)}(\bar{\lambda}_k).$$

Assumption 1 means there exists a solution $(\bar{u}, \bar{\lambda})$ for the saddle-point problem that either $(\nabla\bar{u})_k = 0$ if $\bar{\lambda}_k \in \operatorname{int}\mathbb{B}^k_\alpha(0)$ either $(\nabla\bar{u})_k \neq 0$ if $\bar{\lambda}_k \in \operatorname{bd}\mathbb{B}^k_\alpha(0)$ (i.e. $|\bar{\lambda}_k| = \alpha$). It is not strict and coincides with the optimality conditions in (5.10) and (5.11).

Henceforth, we denote $\mathcal{X}^A$ and $\mathcal{X}^I$ as the solution sets for the dual problem (D) for the anisotropic TV case and the isotropic TV case respectively. With the stopping criterion (A), we have the following global and local convergence.

**Theorem 4.** *For the anisotropic TV regularized problem* (P)*, denote the iteration sequence* $(u^k, p^k, \lambda^k)$ *generated by ALM-PDP, ALM-PDD, ALM-PT or ALM-PP with stopping criteria* (A)*. The sequence* $(u^k, p^k, \lambda^k)$ *is bounded and convergences to* $(u^*, p^*, \lambda^*)$ *globally. If* $T_d := \partial d$ *satisfies Assumption 1 and is thus metrically subregular for the origin with modulus* $\kappa_d$ *through Assumption 1, with the additional stopping criteria* (B1)*, the sequence* $\{\lambda^k\}$ *converges to* $\lambda^* \in \mathcal{X}^A$*. For sufficiently large* $k$*, we have the following local linear convergence*

$$\mathrm{dist}(\lambda^{k+1}, \mathcal{X}^A) \leq \theta_k \mathrm{dist}(\lambda^k, \mathcal{X}^A), \tag{5.22}$$

*where*

$$\theta_k = [\kappa_d(\kappa_d^2 + \sigma_k^2)^{-1/2} + \delta_k](1 - \delta_k)^{-1}, \text{ as } k \to \infty, \ \theta_k \to \theta_\infty = \kappa_d(\kappa_d^2 + \sigma_\infty^2)^{-1/2} < 1.$$

*If in addition that* $T_l$ *is metrically subregular at* $(u^*, p^*, \lambda^*)$ *for the origin with modulus* $\kappa_l$ *and the stopping criteria* (B2) *is employed, then for sufficiently large* $k$*, we have*

$$\|(u^{k+1}, p^{k+1}) - (u^k, p^k)\| \leq \theta'_k \|\lambda^{k+1} - \lambda^k\|, \tag{5.23}$$

*where* $\theta'_k = \kappa_l(1 + \delta'_k)/\sigma_k$ *with* $\lim_{k\to\infty} \theta'_k = \kappa_l/\sigma_\infty$.

*Proof.* Since $X$ is finite dimensional reflexive space and the primal function (P) is l.s.c. proper convex functional and strongly convex, hence coercive. Thus the existence of the solution can be guaranteed [19] (Theorem 4.25). Furthermore, since $\mathrm{dom}\, D = X$ and $\mathrm{dom} \|\cdot\|_1 = Y$, by Fenchel-Rockafellar theory [19] (Chapter 4.3) (or Theorem 5.7 of [5]), the solution to the dual problem (D) is not empty and

$$\inf_{u\in X} D(u) + \alpha\|\nabla u\|_1 = \sup_{\lambda\in Y} -d(\lambda).$$

By [29] (Theorem 4) (or Theorem 1 of [28] where the augmented Lagrangian method essentially comes from proximal point method applying to the dual problem $\partial d$), with criterion (A), we get the boundedness of $\{\lambda^k\}$. The uniqueness of $(u^*, p^*)$ follows from the strongly convexity of $F(u)$ and the $p^* = \nabla u^*$ which is one of the KKT conditions. The boundedness of $(u^k, p^k)$ and convergence of $(u^k, p^k, \lambda^k)$ then follows by [29] (Theorem 4).

The local convergence rate (5.22) with metrical subregularity of $T_g$ and the stopping criteria (A) (B1) can be obtained from [29] (Theorem 5) (or Theorem 2 of [28]). Now we turn to the local convergence rate of $(u^k, p^k)$. By the metrical subregularity of $T_l$, for sufficiently large $k$, we have

$$\|(u^{k+1}, p^{k+1}) - (u^*, p^*))\| + \mathrm{dist}(\lambda^k, \mathcal{X}^A) \leq \kappa_l \mathrm{dist}(0, T_l(u^{k+1}, p^{k+1}, \lambda^{k+1})).$$

Together with the stopping criteria (B2) and [29] (Theorem 5 and Corollary with formula (4.21)), we arrive at

$$\begin{aligned}\|(u^{k+1}, p^{k+1}) - (u^*, p^*))\| &\leq \kappa_l \sqrt{\delta_k'^2 \sigma_k^{-2} \|\lambda^{k+1} - \lambda^k\|^2 + \sigma_k^{-2} \|\lambda^{k+1} - \lambda^k\|^2} \\ &= \kappa_l \sqrt{\delta_k'^2 + 1} \sigma_k^{-1} \|\lambda^{k+1} - \lambda^k\| \leq \theta'_k \|\lambda^{k+1} - \lambda^k\|,\end{aligned}$$

which leads to (5.23). □

For the isotropic case, we can get similar results with metric subregularity of the dual problem.

**Theorem 5.** *For the isotropic TV regularized problem* (P)*, denote the iteration sequence* $(u^k, p^k, \lambda^k)$ *generated by ALM-PDP, ALM-PDD, or ALM-PT with stopping criteria* (A)*. Then the sequence* $(u^k, p^k, \lambda^k)$ *is bounded and convergences to* $(u^*, p^*, \lambda^*)$ *globally. If* $T_d := \partial d$ *satisfies Assumption 1 and is thus metrically subregular for the origin with modulus* $\kappa_d$ *and with the additional stopping criteria* (B1)*, then the sequence* $\{\lambda^k\}$ *converges to* $\lambda^* \in \mathcal{X}^I$ *and for arbitrary sufficiently large* $k$*,*

$$\mathrm{dist}(\lambda^{k+1}, \mathcal{X}^I) \leq \theta_k \mathrm{dist}(\lambda^k, \mathcal{X}^I), \tag{5.24}$$

*where*

$$\theta_k = [\kappa_d(\kappa_d^2 + \sigma_k^2)^{-1/2} + \delta_k](1 - \delta_k)^{-1}, \text{ as } k \to \infty, \ \theta_k \to \theta_\infty = \kappa_d(\kappa_d^2 + \sigma_\infty^2)^{-1/2} < 1.$$

# 6 Numerical Experiments

In the numerics, we will first focus on the ROF denoising model for testing all the proposed algorithms, i.e., $K = I$, $\mu = 0$ with $H = I$. We employ the standard finite difference discretization of the discrete gradient $\nabla$ and divergence operator div [3], which satisfies (2.2) and are convenient for operator actions based implementation. The initial value of $u^0$ is chosen as the noisy or degraded image $z$ and the dual variable is initialized zero value for all compared algorithms. We employ the scaled sum of the residuals of $u$ and $\lambda$ as our stopping criterion for all compared algorithms,

$$\mathrm{Err}(u^{k+1}, \lambda^{k+1}) := [\mathrm{res}(u)(u^{k+1}, \lambda^{k+1}) + \mathrm{res}(\lambda)(u^{k+1}, \lambda^{k+1})]/\|f\|_F, \tag{6.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm henceforth. The residual of $u$ is

$$\mathrm{res}(u)(u^{k+1}, \lambda^{k+1}) := \|u^{k+1} - f + \mathrm{div}\, \lambda^{k+1}\|_F, \tag{6.2}$$

and the residual of $\lambda$ is defined by

$$\mathrm{res}(\lambda)(u^{k+1}, \lambda^{k+1}) := \|\lambda^{k+1} - \mathcal{P}_\alpha(\lambda^{k+1} + c_0 \nabla u^{k+1})\|_F. \tag{6.3}$$

The residual $\mathrm{res}(u)$ is the primal error of the optimality conditions in (2.6) with Frobenius norm and $\mathrm{res}(\lambda)$ is used the measure the dual error in (2.6). Thus $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ in (6.1) is employed to measure the primal and dual errors of the original primal model (P).

Besides the stopping criterion (6.1), the following residual originated in [20] is also useful

$$\mathrm{Res1}(u^{k+1}, \lambda^{k+1}) := \|\mathcal{I}_{\{\|\lambda\|_\infty \leq \alpha\}}(\lambda^{k+1}) + \alpha|\nabla u^{k+1}| - \langle \lambda^{k+1}, \nabla u^{k+1}\rangle\|_F. \tag{6.4}$$

We will also consider the following residual for measuring the optimality conditions (5.10) and (5.11) originated from [15]

$$\mathrm{Res2}(u^{k+1}, \lambda^{k+1}) = \|\alpha \nabla u^{k+1} - |\nabla u^{k+1}|\lambda^{k+1}\|_F. \tag{6.5}$$

The following primal-dual gap for the ROF model as in [3] is also a useful residual

$$\mathfrak{G}(u^{k+1}, \lambda^{k+1}) = \frac{\|u^{k+1} - f\|_2^2}{2} + \alpha\|\nabla u^{k+1}\|_1 + \frac{\|\mathrm{div}\, \lambda^{k+1} + f\|_2^2}{2} - \frac{\|f\|_2^2}{2} + \mathcal{I}_{\{\|\lambda\|_\infty \leq \alpha\}}(\lambda^{k+1}).$$

We use the following normalized primal-dual gap [3]

$$\mathrm{gap}(u^{k+1}, \lambda^{k+1}) := \mathfrak{G}(u^{k+1}, \lambda^{k+1})/NM, \quad \text{with} \quad NM = N * M, \quad u^k \in \mathbb{R}^{N \times M}. \tag{6.6}$$

Let's now turn to the stopping criterion for linear iterative solvers including CG (conjugate gradient) and BiCGSTAB ( biconjugate gradient stabilized method) for each linear system for the Newton update in Algorithm 1, 2, 3. For Algorithm 3, we use CG due to the symmetric linear system (4.25), and (4.17). For any other linear system, e.g., (3.8), (3.11), we use BiCGSTAB (see Figure 9.1 of [36]), which is very efficient for nonsymmetric linear system. The following stopping criterion is employed for solving linear systems to get Newton updates with BiCGSTAB or CG [15],

$$\text{tol}_{k+1} := .1 \min \left\{ \left( \frac{\text{res}_k}{\text{res}_0} \right)^{1.5}, \ \frac{\text{res}_k}{\text{res}_0} \right\}, \tag{6.7}$$

which can help catch the superlinear convergence of semismooth Newton we employ. The $\text{res}_k$ in (6.7) denotes the relative residual of the Newton linear system after the $k$-th BiCGSTAB or CG iteration while $\text{res}_0$ denotes the original residual before BiCGSTAB or CG iterations.

Now, we turn to the most important stopping criterion (A), (B1), (B2) of each ALM iteration for determining how many Newton iterations are needed when solving the corresponding nonlinear systems (2.23) or (2.16). The criterion (B1) is not practical. New stopping criterion of ALM for cone programming can be found in [6]. We found the following empirical stopping criterion for each ALM iteration is efficient numerically. For the primal-dual semismooth Newton with auxiliary variable, we empirical employ

$$\text{res-alm}^l_{SSNPD} = \| - \sigma_k \nabla u^l - \lambda^k + \max\big(1.0, \frac{|\lambda^k + \sigma_k \nabla u^l|}{\alpha}\big) h^l \|_F, \tag{6.8}$$

where $h^l$ is computed by Algorithm 1 or 2 before the projection to the feasible set. For the semismooth Newton method involving the soft thresholding operator (SSNPT), we employ

$$\text{res-alm}^l_{SSNPT} := \|u^l - f + \nabla^* \lambda^k + \sigma_k \nabla^* \nabla u^l - \sigma_k \nabla^* (I + \frac{1}{\sigma_k} \partial \alpha \| \cdot \|_1)^{-1} (\frac{\lambda^k + \sigma_k \nabla u^l}{\sigma_k}) \|_F. \tag{6.9}$$

Here we use $\text{res-alm}^l_{SSNPD}$ without $\text{res}(u)$ compared to (B2) since we found $\text{res}(u)$ is usually much smaller compared to (6.8) in numerics. We employ the following stopping criterion

$$\text{res-alm}^l_{SSNPD} \quad \text{or} \quad \text{res-alm}^l_{SSNPT} \le \delta_k / \sigma_k, \tag{6.10}$$

where $\delta_k$ is a small parameter which can be chosen as fixed constants $10^{-2}$, $10^{-4}$ and so on in our numerical tests. We emphasis that divided by $\sigma_k$ is of critical importance for the convergence of ALM, which is also required by the stopping criterion (A), (B1), (B2).

For numerical comparisons, we mainly choose the accelerated primal-dual algorithm ALG2 [3] with asymptotic convergence rate $\mathcal{O}(1/k^2)$, which is very efficient, robust and standard algorithms for imaging problems. We follow the same parameters for ALG2 as in [3] and the corresponding software for the ROF model. Here we use the stopping criterion (6.1) instead of the normalized primal-dual gap (6.6) as in [3]. The test images Lena and Cameraman are with size $256 \times 256$ and images Monarch and Sails are with size $768 \times 512$. The original, noisy and denoised images can be seen in Figure 1 and 2. We acknowledge that the residuals $\text{Res1}(u^{k+1}, \lambda^{k+1})$ and $\text{Gap}(u^{k+1}, \lambda^{k+1})$ can be unbounded without projection of $\lambda^{k+1}$ to the feasible set as in ALM-PT. We leave them in the tables for references.

From Table 1 and Table 3, it can be seen that the proposed ALM-PDP, ALM-PDD and ALM-PT are very efficient and competitive for the anisotropic ROF model and are very robust for different sizes of images. From Table 2 and Table 4, it can be seen that the proposed ALM-PDP are still very efficient and competitive for the isotropic cases. The performance of ALG2 seems to hold nearly the same efficiency for the isotropic and anisotropic cases. Although the residual

| Anisotropic TV: $\alpha = 0.1$ | | | | Lena: $256 \times 256$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n(t)$ | res($u$) | res($\lambda$) | Res1 | Res2 | Gap | PSNR | Err |
| ALM-PDP | 4(5.42s) | 2.46e-5 | 6.96e-3 | 4.98e-5 | 3.53e-5 | 5.89e-8 | 19.25 | 1e-4 |
| ALM-PDD | 4(23.98s) | 1.68e-13 | 6.96e-3 | 4.98e-5 | 3.53e-5 | 5.89e-8 | 19.25 | 1e-4 |
| ALM-PT | 7(13.80s) | 2.26e-4 | 7.57e-3 | 5.49e-5 | 3.89e-5 | 5.88e-8 | 19.25 | 1e-4 |
| ALG2 | 1032(6.71s) | 1.30e-2 | 6.92e-4 | 7.86e-6 | 5.56e-6 | 1.00e-8 | 19.25 | 1e-4 |
| ALM-PDP | 6(11.45s) | 3.66e-5 | 9.12e-6 | 6.91e-8 | 4.93e-8 | 2.58e-11 | 19.25 | 1e-6 |
| ALM-PDD | 7(56.49s) | 3.66e-5 | 7.78e-5 | 1.55e-6 | 1.11e-6 | 7.85e-11 | 19.25 | 1e-6 |
| ALM-PT | 9(18.29s) | 8.66e-6 | 3.73e-5 | 2.87e-7 | 2.04e-7 | 1.15e-10 | 19.25 | 1e-6 |
| ALG2 | 14582(90.28s) | 1.37e-4 | 2.26e-7 | 2.46e-9 | 1.83e-9 | 4.93e-13 | 19.25 | 1e-6 |
| ALM-PDP | 9(18.83s) | 1.06e-6 | 3.74e-10 | 3.03e-12 | 2.16e-12 | 4.02e-16 | 19.25 | 1e-8 |
| ALM-PDD | 7(69.76s) | 2.80e-7 | 1.07e-7 | 8.62e-10 | 6.22e-10 | 2.02e-13 | 19.25 | 1e-8 |
| ALM-PT | 11(28.26s) | 6.41e-7 | 3.12e-9 | 5.45e-12 | 3.94e-12 | 1.29e-15 | 19.25 | 1e-8 |
| ALG2 | 365019(2069.36s) | 1.37e-6 | 2.88e-13 | 3.37e-15 | 2.38e-15 | <eps | 19.25 | 1e-8 |

**Table 1:** For $n(t)$ of the first column of each algorithm, $n$ presents the outer ALM iteration number or iteration number for ALG2 for the scaled residual $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ less than the stopping value, $t$ denoting the CPU time. The notation "<eps" denotes the corresponding quality less than the machine precision, i.e., the "eps" in Matlab.

| Isotropic TV: $\alpha = 0.1$ | | | | Cameraman: $256 \times 256$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n(t)$ | res($u$) | res($\lambda$) | Res1 | Res2 | Gap | PSNR | Err |
| ALM-PDP | 3(4.75s) | 2.32e-3 | 1.14e-2 | 5.61e-5 | 9.36e-5 | 9.70e-8 | 19.18 | 1e-4 |
| ALM-PDD | 5(57.29s) | 8.61e-3 | 2.61e-3 | 1.37e-5 | 2.19e-5 | 2.01e-8 | 19.18 | 1e-4 |
| ALM-PT | 8(174.07s) | 5.23e-5 | 4.30e-3 | 2.31e-5 | 3.64e-5 | 3.34e-8 | 19.18 | 1e-4 |
| ALG2 | 479(2.65s) | 1.59e-2 | 6.03e-6 | 7.38e-6 | 7.01e-4 | 8.69e-9 | 19.18 | 1e-4 |
| ALM-PDP | 7(20.35s) | 6.75e-5 | 8.08e-5 | 4.46e-7 | 6.96e-7 | 4.01e-10 | 19.18 | 1e-6 |
| ALM-PDD | 7(191.50s) | 3.98e-5 | 4.94e-5 | 2.73e-7 | 4.24e-7 | 2.28e-10 | 19.18 | 1e-6 |
| ALM-PT | 11(787.01s) | 1.89e-5 | 6.97e-5 | 3.91e-7 | 6.02e-7 | 3.29e-10 | 19.18 | 1e-6 |
| ALG2 | 10808(55.16s) | 1.60e-4 | 1.60e-9 | 2.46e-9 | 2.63e-7 | 6.56e-13 | 19.18 | 1e-6 |
| ALM-PDP | 10(243.57s) | 1.07e-6 | 4.22e-7 | 2.43e-9 | 3.82e-9 | 1.04e-12 | 19.18 | 1e-8 |
| ALM-PDD | — | — | — | — | — | — | — | 1e-8 |
| ALM-PT | — | — | — | — | — | — | — | 1e-8 |
| ALG2 | 262811(1348.55s) | 1.60e-6 | 5.01e-13 | 7.47e-13 | 8.11e-11 | <eps | 19.18 | 1e-8 |

**Table 2:** For $n(t)$ of the first column of each algorithm, $n$ presents the outer ALM iteration number or iteration number for ALG2 for the scaled residual $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ less than the stopping value, $t$ denoting the CPU time. "—" denotes the iteration time more than 2000 seconds. The notation "<eps" is the same as in Table 1.

| | Anisotropic TV: $\alpha = 0.1$ | | | | Sails: $768 \times 512$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $n(t)$ | res($u$) | res($\lambda$) | Res1 | Res2 | Gap | PSNR | Err |
| ALM-PDP | 6(25.50s) | 1.56e-2 | 1.15e-3 | 7.31e-5 | 2.28e-5 | 1.61e-5 | 18.90 | 1e-4 |
| ALM-PDD | 4(159.38) | 7.60e-3 | 1.97e-2 | 2.73e-4 | 2.28e-4 | 4.76e-8 | 18.90 | 1e-4 |
| ALM-PT | 7(101.59s) | 2.72e-4 | 1.53e-2 | 1.15e-4 | 8.16e-5 | 4.55e-8 | 18.90 | 1e-4 |
| ALG2 | 933(28.23s) | 3.35e-2 | 1.83e-3 | 2.10e-5 | 1.49e-5 | 1.08e-8 | 18.90 | 1e-4 |
| ALM-PDP | 7(52.52s) | 5.68e-5 | 2.25e-6 | 2.03e-8 | 1.45e-8 | 9.42e-13 | 18.90 | 1e-6 |
| ALM-PDD | 6(377.60s) | 2.18e-5 | 1.13e-5 | 9.03e-8 | 6.43e-8 | 1.25e-11 | 18.90 | 1e-6 |
| ALM-PT | 8(337.23s) | 2.20e-5 | 5.58e-5 | 4.42e-7 | 3.17e-7 | 6.11e-11 | 18.90 | 1e-6 |
| ALG2 | 14842(476.20s) | 3.53e-4 | 3.40e-7 | 3.94e-9 | 2.99e-9 | 3.09e-13 | 18.90 | 1e-6 |
| ALM-PDP | 7(54.51s) | 5.65e-7 | 2.25e-6 | 2.03e-8 | 1.45e-8 | 9.38e-13 | 18.90 | 1e-8 |
| ALM-PDD | 14(122.23s) | 2.68e-6 | 3.69e-12 | 3.44e-14 | 2.47e-14 | <eps | 18.90 | 1e-8 |
| ALM-PT | 10(363.27s) | 3.30e-6 | 1.71e-8 | 1.61e-10 | 1.14e-10 | 1.44e-14 | 18.90 | 1e-8 |
| ALG2 | 371060(11339.88s) | 3.54e-6 | 6.98e-13 | 8.20e-15 | 5.80e-15 | <eps | 18.90 | 1e-8 |

**Table 3:** For $n(t)$ of the first column of each algorithm, $n$ presents the outer ALM iteration number or iteration number for ALG2 for the scaled residual $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ less than the stopping value, $t$ denoting the CPU time. The notation "<eps" is the same as in Table 1.

| | Isotropic TV: $\alpha = 0.1$ | | | | Monarch: $768 \times 512$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $n(t)$ | res($u$) | res($\lambda$) | Res1 | Res2 | Gap | PSNR | Err |
| ALM-PDP | 3(25.12) | 1.12e-2 | 1.43e-2 | 7.31e-5 | 1.20e-4 | 4.29e-8 | 29.89 | 1e-4 |
| ALM-PDD | 5(370.53s) | 9.89e-3 | 5.82e-3 | 3.01e-5 | 4.89e-5 | 1.54e-8 | 29.89 | 1e-4 |
| ALM-PT | 8(2009.60s) | 8.34e-5 | 9.73e-3 | 5.13e-5 | 8.19e-5 | 2.66e-8 | 29.89 | 1e-4 |
| ALG2 | 525(15.77s) | 2.86e-2 | 7.69e-6 | 1.02e-5 | 1.07e-3 | 4.76e-9 | 29.89 | 1e-4 |
| ALM-PDP | 7(145.27s) | 7.98e-5 | 1.80e-4 | 9.79e-7 | 1.54e-6 | 3.15e-10 | 29.89 | 1e-6 |
| ALM-PDD | 7(1379.33s) | 1.47e-7 | 1.09e-4 | 5.96e-7 | 9.33e-7 | 1.78e-10 | 29.89 | 1e-6 |
| ALM-PT | — | — | — | — | — | — | — | 1e-6 |
| ALG2 | 12049(387.21s) | 2.87e-4 | 2.61e-9 | 3.79e-9 | 4.27e-7 | 3.63e-13 | 29.89 | 1e-6 |
| ALM-PDP | 10(1549.91s) | 1.79e-6 | 1.06e-6 | 5.90e-9 | 9.05e-9 | 8.94e-13 | 29.89 | 1e-8 |
| ALM-PDD | — | — | — | — | — | — | — | 1e-8 |
| ALM-PT | — | — | — | — | — | — | — | 1e-8 |
| ALG2 | 292442(8590.28s) | 2.87e-6 | 1.56e-12 | 2.11e-12 | 2.33e-10 | <eps | 29.89 | 1e-8 |

**Table 4:** For $n(t)$ of the first column of each algorithm, $n$ presents the outer ALM iteration number or iteration number for ALG2 for the scaled residual $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ less than the stopping value, $t$ denoting the CPU time. For ALM based algorithms, "—" denotes the cases when the iteration time more than 3000 seconds. We keep ALG2 for comparison.

| | Isotropic TV deblurring: $\alpha = 0.1$ | | | | Pepper: $512 \times 512$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Err= $10^{-5}$, $\mu = 10^{-6}$ | | | | Err= $10^{-5}$, $\mu = 10^{-9}$ | | | |
| | $n(t)$ | res($u$) | res($\lambda$) | PSNR | $n(t)$ | res($u$) | res($\lambda$) | PSNR |
| ALM-PDP | 3(126.08s) | 1.06e-4 | 1.08e-3 | 28.48 | 3(117.37s) | 4.36e-4 | 1.07e-3 | 28.48 |
| ALG2 | 4775(178.18s) | 2.02e-3 | 5.48e-4 | 28.36 | 4762(170.65s) | 2.02e-3 | 5.72e-4 | 28.36 |
| | Err= $10^{-6}$, $\mu = 10^{-6}$ | | | | Err= $10^{-6}$, $\mu = 10^{-9}$ | | | |
| | $n(t)$ | res($u$) | res($\lambda$) | PSNR | $n(t)$ | res($u$) | res($\lambda$) | PSNR |
| ALM-PDP | 4(196.48s) | 3.07e-5 | 9.69e-5 | 28.48 | 4(186.79s) | 7.08e-5 | 1.70e-4 | 28.48 |
| ALG2 | 11017(397.71s) | 2.02e-4 | 5.58e-5 | 28.47 | 10911(395.16s) | 2.02e-4 | 5.54e-5 | 28.47 |

**Table 5:** For $n(t)$ of the first column of each algorithm, $n$ presents the outer ALM iteration number or iteration number for ALG2 for the scaled residual $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ less than the stopping value, $t$ denoting the CPU time. "—" denotes the iteration time more than 2000 seconds. The notation "<eps" is the same as in Table .

| | $k=1$ | $k=2$ | $k=3$ | $k=4$ | $k=5$ | $k=6$ |
|---|---|---|---|---|---|---|
| res($u$) | 3.50e-6 | 5.30e-7 | 1.53e-5 | 6.14e-8 | 3.29e-8 | 1.94e-7 |
| res($\lambda$) | 9.70 | 1.09 | 9.99e-2 | 6.96e-3 | 3.07e-4 | 9.13e-6 |
| Gap | 1.61e-4 | 1.41e-5 | 1.14e-6 | 5.89e-8 | 1.58e-9 | 2.58e-11 |
| $N_{SSN}$ | 10 | 9 | 8 | 11 | 9 | 7 |
| $N_{ABCG}$ | 20 | 29 | 51 | 56 | 102 | 116 |

**Table 6:** Image Lena denoised by anisotropic TV with algorithm ALM-PDP. $N_{SSN}$ denotes the number of Newton iterations for each ALM iteration. $N_{ABCG}$ denotes average number of BiCGSTAB iterations for each Newton iteration. Here $\sigma_0 = 4$, $\sigma_{k+1} = 4\sigma_k$.

res($\lambda$) for smaller $\mathrm{Err}(u^{k+1}, \lambda^{k+1})$ are much smaller for ALG2 than the proposed semismooth Newton based ALM methods, however, the residual res($u$) of ALG2 is more higher than the proposed ALM methods. It seems ALG2 is more efficient for decreasing the dual error res($\lambda$).

Our proposed algorithms are much more efficient for the anisotropic case compared to the isotropic case for ROF denoising model. Besides Theorem 2 and 3, we present another comparison between the isotropic TV and anisotropic TV with algorithm ALM-PDP. For Table 6, we give enough Newton iterations compared to previous tables with the stopping the criterion $\text{res-alm}^l_{SSNPD} \leq 10^{-3}/\sigma_k$ for each ALM as in (6.10) and $\mathrm{tol}_{k+1} \leq 10^{-5}$ in (6.7) for each Newton iteration. The ALM iterations stop while $\mathrm{Err}(u^{k+1}, \lambda^{k+1}) \leq 10^{-7}$ both for the anisotropic TV.

Finally, let's turn to the image deblurring problem. For the image deblurring with $Ku = \kappa * u$ being the convolution with kernel $\kappa$ and additional regularization term $\frac{\mu}{2}\|\nabla u\|_2^2$, we mainly focus on the isotropic case as in Table 5. The original, degraded and reconstructed Pepper images with size $512 \times 512$ can be found in Figure 1. Since the primal model (P) is only strongly convex on the primal variable $u$ instead being strongly convex on both $u$ and $\lambda$, we thus employ ALG2 for comparison. The strongly convex parameter for $D(u)$ is a subtle issue since $K^*K$ is not invertible with the kernel. We use the strongly convex parameter of $\frac{\mu}{2}\|\nabla u\|_2^2$ instead. It can be seen that both ALM-PDP and ALG2 are very efficient for small $\mu$ case. It is interesting that both ALM-PDP and ALG2 seem more efficient for the case $\mu = 10^{-9}$ than the case $\mu = 10^{-6}$.

**(a)** Original image: lena

**(b)** Noisy image: 10% Gaussian

**(c)** ALM-PDP($10^{-6}$)

**(d)** Original image: Cameraman

**(e)** Noisy image: 10% Gaussian

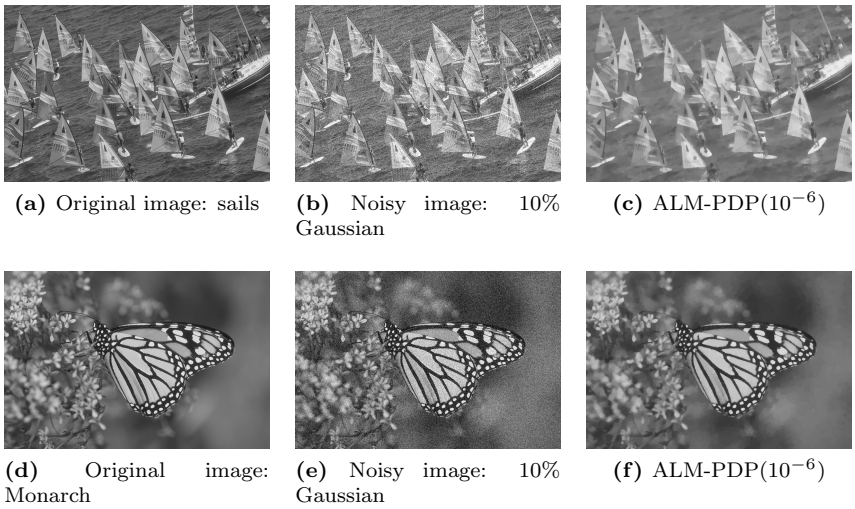
**(f)** ALM-PDP($10^{-6}$)

**(g)** Original image: Pepper

**(h)** Degraded image:

**(i)** ALM-PDP($10^{-5}$)

**Figure 1:** Images (a) and (d) show the original $256 \times 256$ Lena and Cameraman images. Image (g) shows the original $512 \times 512$ Pepper image. (b) and (e) are noisy versions corrupted by 10% Gaussian noise. Image (h) is the degraded Pepper image ($\approx$ 40 pixels motion blur, Gaussian noise, standard deviation 0.01). (c) and (f) show the denoised images with ALM-PDP with $\mathrm{Err}(u^{k+1}, \lambda^{k+1}) < 10^{-6}$. Image Lena is denoised by anisotropic TV while image Cameraman is denoised by isotropic TV. Image (i) is the deblurred Pepper image with isotropic TV by ALM-PDP with $\mathrm{Err}(u^{k+1}, \lambda^{k+1}) < 10^{-5}$.



**(a)** Original image: sails **(b)** Noisy image: 10% Gaussian **(c)** ALM-PDP($10^{-6}$)

**(d)** Original image: Monarch **(e)** Noisy image: 10% Gaussian **(f)** ALM-PDP($10^{-6}$)

**Figure 2:** Images (a) and (d) show the original $768 \times 512$ Sails and Monarch images. (b) and (e) are noisy versions corrupted by 10% Gaussian noise. (c) and (f) show the denoised images with ALM-PDP with $\mathrm{Err}(u^{k+1}, \lambda^{k+1}) < 10^{-6}$. Image Sails is denoised by anisotropic TV while image Monarch is denoised by isotropic TV.

# 7 Discussion and Conclusions

In this paper, we proposed several semismooth Newton based ALM algorithms. The proposed algorithms are very efficient and competitive especially for anisotropic TV. Global convergence and the corresponding asymptotic convergence rates are also discussed with metrical subregularites. Numerical tests show that compared with first-order algorithm, more computation efforts are deserved with ALM. Currently, no preconditioners are employed by BiCGSTAB or CG for the linear systems solving Newton updated. Actually, preconditioners for BiCGSTAB or CG are desperately needed, especially for the isotropic cases. Additionally, the asymptotic convergence rate of the KKT residuals is also an interesting topic for future study [6].

**Acknowledgements** The author acknowledges the support of NSF of China under grant No. 11701563. The work was originated during the author's visit to Prof. Defeng Sun of the Hong Kong Polytechnic University in October 2018. The author is very grateful to Prof. Defeng Sun for introducing the framework on semismooth Newton based ALM developed by him and his collaborators and for his suggestions on the self-adjointness of the corresponding operators for Newton updates where CG can be employed. The author is also very grateful to Prof. Kim-Chuan Toh, Dr. Chao Ding, Dr. Xudong Li and Dr. Xinyuan Zhao for the discussion on the semismooth Newton based ALM. The author is also very grateful to Prof. Michael Hintermüller for the discussion on the primal-dual semismooth Newton method during the author's visit to Weierstrass Institute for Applied Analysis and Stochastics (WIAS) supported by Alexander von Humboldt Foundation during 2017.

# References

[1] H. H. Bauschke, P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2011.

[2] D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, Paris, 1982.

[3] A. Chambolle, T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging and Vis., 40(1), pp. 120–145, 2011.

[4] F. H. Clarke, Optimization and Nonsmooth Analysis, Vol. 5, Classics in Applied Mathematics, SIAM, Philadelphia, 1990.

[5] C. Clason, Nonsmooth Analysis and Optimization, Lecture notes, `https://arxiv.org/abs/1708.04180`, 2018.

[6] Y. Cui, D. Sun, K. Toh, *On the R-superlinear convergence of the KKT residues generated by the augmented Lagrangian method for convex composite conic programming*, Math. Program., Ser. A, 178:38–415, 2019, https://doi.org/10.1007/s10107-018-1300-6.

[7] T. De Luca, F. Facchinei, C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Program. 75, pp. 407–439, 1996.

[8] A. L. Dontchev, R. T. Rockafellar, Functions and Solution Mappings: A View from Variational Analysis, Second Edition, Springer Science+Business Media, New York 2014.

[9] F. Facchinei, J. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I, Springer-Verlag New York, Inc, 2003.

[10] M. Fortin, R. Glowinski (eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, North-Holland, Amsterdam, 1983.

[11] R. Glowinski, S. Osher, W. Yin (eds.), Splitting Methods in Communication, Imaging, Science, and Engineering, Springer, 2016.

[12] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4, pp. 303–320, 1968

[13] M. Hintermüller, K. Kunisch, *Total bounded variation regularization as a bilaterally constrained optimization problem*, SIAM J. Appl. Math, 64(4), pp. 1311–1333.

[14] M. Hintermüller, K. Papafitsoros, C. N. Rautenberg, H. Sun, *Dualization and automatic distributed parameter selection of total generalized variation via bilevel optimization*, preprint, to appear, 2019.

[15] M. Hintermüller, G. Stadler, *A infeasible primal-dual algorithm for total bounded variaton-based inf-convolution-type image restoration*, SIAM J. Sci. Comput., 28(1), pp. 1–23, 2006.

[16] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Research Nat. Bur. Standards, 49 (1952), pp. 263–265.

[17] D. Klatte, B. Kummer, *Constrained minima and Lipschitzian penalties in metric spaces*, SIAM J. Optim., 13(2), pp. 619–633, 2002.

[18] D. Klatte, B. Kummer, Nonsmooth Equations in Optimization. Regularity, Calculus, Methods and Applications, Series Nonconvex Optimization and Its Applications, Vol. 60, Springer, Boston, MA, 2002.

[19] K. Ito, K. Kunisch, Lagrange Multiplier Approach to Variational Problems and Applications, Advances in design and control 15, Philadelphia, SIAM, 2008.

[20] K. Ito, K. Kunisch, *An active set strategy based on the augmented Lagrangian formulation for image restoration*, RAIRO, Math. Mod. and Num. Analysis, 33(1), pp. 1–21, 1999.

[21] D. Leventhal, *Metric subregularity and the proximal point method*, J. Math. Anal. Appl., 360(2009), pp. 681-688, 2009.

[22] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15(6), pp. 959–972, 1977.

[23] X. Li, D. Sun, C. Toh, *A highly efficient semismooth Newton augmented Lagrangian method for solving lasso problems*, SIAM J. Optim., 28(1), pp. 433–458, 2018.

[24] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22(2), pp. 277–293, 1984.

[25] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, pp. 283–298, 1968.

[26] M. Reed, B. Simon, Methods of Modern Mathematical Physics: Functional Analysis, Methods of Modern Math. Phys. 1, Academic Press, New York, 1980.

[27] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, in Mathematical Programming at Oberwolfach, Math. Program. Stud., Springer, Berlin, Heidelberg, pp. 206–214, 1981.

[28] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14(5), pp. 877–898, 1976.

[29] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1(2), pp. 97-116, 1976.

[30] L. I. Rudin, S. Osher, E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D., 60(1-4), pp. 259–268, 1992.

[31] S. Scholtes, Introduction to Piecewise Differentiable Equations, Springer Briefs in Optimization, Springer, New York, 2012.

[32] G. Stadler, *Semismooth Newton and augmented Lagrangian methods for a simplified friction problem*, SIAM J. Optim., 15(1), pp. 39–62, 2004.

[33] G. Stadler, Infinite-Dimensional Semi-Smooth Newton and Augmented Lagrangian Methods for Friction and Contact Problems in Elasticity, PhD thesis, University of Graz, 2004.

[34] D. Sun and J. Han, *Newton and quasi-Newton methods for a class of nonsmooth equations and related problems*, SIAM J. Optim., 7, pp. 463–480, 1997.

[35] M. Ulbrich, Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces, MOS-SIAM Series on Optimization, 2011.

[36] H. A. Van der Vorst, Iterative Krylov Methods for Large Linear Systems, Cambridge University Press, Cambridge, 2003.

[37] J. Ye, X. Yuan, S. Zeng, J. Zhang, *Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems*, Set-Valued Var. Anal., https://doi.org/10.1007/s11228-021-00591-3, 2021.

[38] P. Yu, G. Li, TK. Pong, *Deducing Kurdyka-Lojasiewicz exponent via inf-projection*, Found. Comput. Math., https://doi.org/10.1007/s10208-021-09528-6, 2021.

[39] F. Zhang (eds.), The Schur Complement and Its Applications, Numerical Methods and Algorithms 4, Springer US, 2005.

[40] Y. Zhang, N. Zhang, D. Sun, K. Toh, *An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems*, Mathematical Programming A, `https://doi.org/10.1007/s10107-018-1329-6`, 2018.

[41] X. Zhao, D. Sun, K. Toh, *A Newton-CG augmented Lagrangian method for semidefinite programming*, SIAM J. Optim., 20(4), pp. 1737–1765, 2010.