Noname manuscript No.
(will be inserted by the editor)

Lagrange optimality system for a class of nonsmooth convex optimization problems

Bangti Jin · **Tomoya Takeuchi**

Received: date / Accepted: date

Abstract In this paper, we revisit the augmented Lagrangian method for a class of nonsmooth convex optimization problems. We present the Lagrange optimality system of the augmented Lagrangian associated with the problems, and establish its connections with the standard optimality condition for the solution and the saddle point condition of the augmented Lagrangian, which provides a powerful tool for developing numerical algorithms. We employ a linear Newton method to the Lagrange optimality system, and obtain a novel algorithm for the nonsmooth convex optimization problems. Under suitable conditions, the nonsingularity of the Newton system is shown, and the local convergence and convergence rates of the algorithm are provided. The algorithm is applicable to a wide range of problems in practical applications, and illustrated on three common examples.

Keywords nonsmooth convex optimization · augmented Lagrangian method · Lagrange optimality system · Newton method

Mathematics Subject Classification (2000) 90C25 · 46N10 · 49M15

1 Introduction

In this paper we consider the augmented Lagrangian method for solving a class of nonsmooth convex optimization problems

$$\min_{x\in X} f(x)+\phi(\Lambda x), \tag{1}$$

Bangti Jin
Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
E-mail: b.jin@ucl.ac.uk

Tomoya Takeuchi
Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
E-mail: takeuchi@sat.t.u-tokyo.ac.jp

where the function $f\colon X\to\mathbb{R}$ is convex and continuously differentiable on a Banach space X, $\phi\colon H\to\mathbb{R}^+$ is a proper, lower semi-continuous and convex function on a Hilbert space H, and Λ is a bounded linear operator from X to H. We assume that the proximity operator of the convex function ϕ has a closed form expression. This problem class encompasses a wide range of optimization problems arising in practical applications, e.g., inverse problems, variational problems, image processing, signal processing and statistics to name a few [1–7].

The augmented Lagrangian method was proposed independently by Hestenes [8] and Powell [9] for solving nonlinear programming problems with equality constraints. The method was studied in relation to Fenchel duality and generalized to nonlinear programming problems with inequality constraints by Rockafellar [10,11]. Later it was further generalized to the problem (1) by Glowinski and Marroco [12] where the augmented Lagrangian is given by

$$\mathscr{L}_c(x,v,\lambda)=f(x)+\phi(v)+(\lambda,\Lambda x-v)+\frac{c}{2}\|\Lambda x-v\|^2.$$

The inner product $(\lambda,\Lambda x-v)$ dualizes the equality constraint, and the quadratic term penalizes the constraint violation for the following equality constrained problem equivalent to problem (1):

$$\min_{x\in X,v\in H} f(x)+\phi(v)\qquad\text{subject to}\qquad \Lambda x=v.$$

A solution of problem (1) can be characterized, under certain conditions on f, ϕ and Λ, as a saddle point of the augmented Lagrangian, and the strong duality theorem leads to first-order algorithms for the dual function $\theta(\lambda)=\inf_{x,v}\mathscr{L}_c(x,v,\lambda)$. In practical implementation, the combination of the dualization and the penalization alleviates the slow convergence for the ordinary Lagrangian methods and ill conditioning as $c\to\infty$ for penalty methods. Due to these advantages over the standard Lagrangian formulation and the penalty formulation, a large number of first order algorithms based on the augmented Lagrangian $\mathscr{L}_c$ have been developed for a wide variety of applications; see e.g., [1,13–15].

An alternative Lagrangian for (1) has been introduced by Fortin [16], which was obtained by employing the partial conjugate of the augmented perturbation bifunction $F_c(x,v)=f(x)+\phi(\Lambda x-v)+\frac{c}{2}\|v\|^2$ due to Rockafeller [10]:

$$\begin{aligned}L_c(x,\lambda)&=\min_{v\in H}\left((v,\lambda)+F_c(x,v)\right)=\min_{v\in H}\left((v,\lambda)+f(x)+\phi(\Lambda x-v)+\frac{c}{2}\|v\|^2\right)\\&=\min_{u\in H}\left((\Lambda x-u,\lambda)+f(x)+\phi(u)+\frac{c}{2}\|\Lambda x-u\|^2\right)\\&=f(x)+\min_{u\in H}\left(\phi(u)+(\lambda,\Lambda x-u)+\frac{c}{2}\|\Lambda x-u\|^2\right)\\&=f(x)+\phi_c(\Lambda x+\lambda/c)-\tfrac{1}{2c}\|\lambda\|^2,\end{aligned}$$

where c is a positive constant and the function $\phi_c(z)$ is the Moreau envelope (see Section 2 for the definition). It was shown that a saddle point of L_c is also a saddle point of the standard Lagrangian and conversely [16, Thm. 2.1]. A first order algorithm often referred to as the *augmented Lagrangian algorithm*, which is quite similar to the one developed in [12], was proposed for certain special cases of the function ϕ [16, Thm. 4.1]. The augmented Lagrangian method was further studied by Ito and Kunisch [17] for the following optimization problem

$$\min_{x \in C} f(x) + \phi(\Lambda x), \tag{2}$$

where C is a convex set in X. One of their major achievements is the results concerning the existence of a Lagrange multiplier for problem (2): It was shown that under appropriate conditions Lagrange multipliers of a regularized problem defined by the augmented Lagrangian L_c converge and the limit is a Lagrange multiplier of problem (2). In addition to the valuable contribution, the augmented Lagrangian algorithm by Fortin was extended to a more general class of convex functions ϕ, and the convergence of the algorithm was established. It is noted that the problem can be reformulated into problem (1), by redefining the convex function ϕ and the linear map Λ by $\phi(x,y) := \phi(x) + \chi_C(y)$ and $\Lambda x := (\Lambda x, x)$, respectively, where χ_C is the characteristic function of the convex set C. Hence, it shares an identical structure with problem (1).

The augmented Lagrangian L_c is Fréchet differentiable, cf. Section 3, which motivates the use of the Lagrange optimality system

$$D_x L_c(x,\lambda) = 0, \quad \text{and} \qquad D_\lambda L_c(x,\lambda) = 0, \tag{3}$$

to characterize the saddle point and hence the solution of problem (1). This perspective naturally leads to the application of Newton methods for solving the nonlinear system. However, the Moreau envelope involved in (3), cf. Proposition 3, is twice continuously differentiable if and only if the same is true for the convex function ϕ [18], and thus the standard (classical) Newton methods cannot be applied directly to the Lagrange optimality system. Semismooth Newton methods and quasi-Newton methods are possible alternatives for solving the Lagrange optimality system, but there are some drawbacks in their applications to the Lagrange optimality system: The inclusion appearing in the chain rule of a composite map makes it difficult to theoretically identify a generalized or limiting Jacobian of $D_{x,\lambda} L_c$ for semismooth Newton methods, while the superlinear convergence of quasi-Newton methods holds only when the system to be solved is differentiable at the solution [19]. We opt for instead linear Newton methods [20] to solve the Lagrange optimality system (3), where one replaces the generalized Jacobian of $D_{x,\lambda} L_c$ in semismooth Newton methods with a *linear Newton approximation (LNA)* of $D_{x,\lambda} L_c$. Calculus rules, which provide a systematic way of generating LNAs of a given map, reduce the construction of a LNA of the Lagrange optimality system to the computation of the (Clarke's) generalized or limiting Jacobian of the proximity operator involved in the system, cf. Section 4.

The focus of this work is twofold. First, we present the Lagrange optimality system, which was not provided in both [16] and [17], and establish its connection with the standard optimality system of problem (1) and the saddle point condition of the augmented Lagrangian. Second, on the basis of the Lagrange optimality system, we develop a Newton type algorithm for problem (1). To the best of our knowledge, this is the first work using the Lagrange optimality system systematically for developing

Newton type algorithms for nonsmooth convex optimization (1). These two aspects represent the essential contributions of this work.

The rest of the paper is organized as follows. In Section 2 we collect fundamental results on the Moreau envelope and the promixity operator, which provide the main tools for the analysis. In Section 3, we investigate the connection among the optimality system for the problem (1), the Lagrange optimality system and the saddle point of the augmented Lagrangian L_c. In Section 4, we develop a Newton method for problem (1), which exhibits a local Q-superlinear convergence.

1.1 Notations

We denote by X a real Banach space with the norm $|\cdot|$. The duality bracket between the dual space X^* and X is denoted by $\langle\cdot,\cdot\rangle_{X^*,X}$. For a twice continuously differentiable function f, its derivative is denoted by $Df(x)$ or $D_xf(x)$, and its Hessian by $D_x^2f(x)$. H is a Hilbert space with the inner product $(\cdot,\cdot)$, and the norm on H is denoted by $\|\cdot\|$. The set of proper, lower semicontinuous, convex functions defined on the Hilbert space H is denoted by $\Gamma_0(H)$. The effective domain of a function $\phi\in\Gamma_0(H)$ is denoted by $\mathscr{D}(\phi)=\{z\in H\mid\phi(z)\text{ is finite}\}$, and it is always assumed to be nonempty. For a function $\phi\in\Gamma_0(H)$, the convex conjugate ϕ^* is defined by $\phi^*(z^*)=\sup_{z\in H}\left((z^*,z)-\phi(z)\right)$. A *subgradient* of ϕ at $x\in H$ is $g\in H$ satisfying

$$\phi(y)\geq\phi(x)+(g,y-x),\quad\forall y\in H.$$

The *subdifferentials* of ϕ at x is the set of all subgradients of ϕ at x, and is denoted by $\partial\phi(x)$.

2 Moreau envelope and proximity operator

The central tools for analyzing the augmented Lagrangian approach are Moreau envelope and proximity operator. We recall their definitions and basic properties that are relevant to the development of the Lagrange multiplier theory. We note that for $\phi\in\Gamma_0(H)$ the strictly convex function $u\to\phi(u)+\frac{1}{2}\|u-z\|^2$ admits a unique minimizer.

Definition 1 Let $\phi\in\Gamma_0(H)$ and $c>0$. The *Moreau envelope* $\phi_c:H\to\mathbb{R}$ and the *proximity operator* $\mathrm{prox}_\phi:H\to H$ are defined respectively as

$$\begin{aligned}\phi_c(z)&=\min_{u\in H}\left(\phi(u)+\tfrac{c}{2}\|u-z\|^2\right),\\ \mathrm{prox}_\phi(z)&=\arg\min_{u\in H}\left(\phi(u)+\tfrac{1}{2}\|u-z\|^2\right),\end{aligned}$$

for $z\in H$.

By definition we have

$$\mathrm{prox}_{\frac{\phi}{c}}(z)=\arg\min_{u\in H}\left(\tfrac{\phi(x)}{c}+\tfrac{1}{2}\|u-z\|^2\right)=\arg\min_{u\in H}\left(\phi(u)+\tfrac{c}{2}\|u-z\|^2\right),$$

and

$$\phi_c(z) = \phi(\mathrm{prox}_{\frac{\phi}{c}}(z)) + \tfrac{c}{2}\|\mathrm{prox}_{\frac{\phi}{c}}(z) - z\|^2.$$

We refer interested readers to Tables 10.1 and 10.2 of [3] for closed-form expressions of a number of frequently used proximity operators.

We recall well-known properties of the Moreau envelope and proximity operator.

Proposition 1 ([21]) *Let $z \in H$ and $c > 0$. Let $\phi \in \Gamma_0(H)$.*

(a) $0 \le \phi(z) - \phi_c(z)$ *for all $z \in H$ and all $c > 0$.*

(b) $\lim_{c\to\infty} \phi_c(z) = \phi(z)$ *for all $z \in H$.*

(c) *The proximity operator* $\mathrm{prox}_{\frac{\phi}{c}}$ *is nonexpansive, that is,*

$$\|\mathrm{prox}_{\frac{\phi}{c}}(z) - \mathrm{prox}_{\frac{\phi}{c}}(w)\|^2 \le (\mathrm{prox}_{\frac{\phi}{c}}(z) - \mathrm{prox}_{\frac{\phi}{c}}(w), z - w), \quad \forall z, \forall w \in H.$$

(d) *The Moreau envelope ϕ_c is Fréchet differentiable and the gradient is given by*

$$D_z\phi_c(z) = c(z - \mathrm{prox}_{\frac{\phi}{c}}(z)), \quad \forall c > 0, \forall z \in H. \tag{4}$$

(e) *The gradient $z \to D_z\phi_c(z) \in H$ is Lipschitz continuous with a Lipschitz constant c, i.e.,*

$$\|D_z\phi_c(z) - D_z\phi_c(w)\| \le c\|z - w\|, \quad \forall z, \forall w \in H.$$

(f) *The Moreau envelope and the proximity operator of the conjugate of ϕ are related with ϕ_c and $\mathrm{prox}_{\frac{\phi}{c}}$, respectively as*

$$\phi_c(z) + (\phi^*)_{\frac{1}{c}}(cz) = \tfrac{c}{2}\|z\|^2, \quad \mathrm{prox}_{\frac{\phi}{c}}(z) + \tfrac{1}{c}\mathrm{prox}_{c\phi^*}(cz) = z.$$

All the results are standard and the proofs can be found in e.g., [21]. Here we give an alternative proof of (f) based on the duality theory.

Proof For $z \in H$, we define the function $L_z\colon H \times \mathscr{D}(\phi) \to \mathbb{R}$ by

$$L_z(u,p) := (u,p) - \phi(p) + \tfrac{1}{2c}\|u - cz\|^2.$$

Clearly, L_z is convex in u and is concave in p. We claim that L_z posses a saddle point on $H \times \mathscr{D}(\phi)$. Clearly, $\lim_{\|u\|\to\infty} L_z(u,p) = \infty$ for all $p \in \mathscr{D}(\phi)$. Thus by [4, Chap. 6, Prop. 2.3], we have

$$\inf_u \sup_p L_z(u,p) = \sup_p \inf_u L_z(u,p). \tag{5}$$

Now we compute $\inf_u \sup_p L_z(u,p)$ and $\sup_p \inf_u L_z(u,p)$ separately. First, we observe

$$\begin{aligned}\inf_u \sup_p L_z(u,p) &= \inf_u \left(\sup_p((u,p) - \phi(p)) + \tfrac{1}{2c}\|u - cz\|^2 \right)\\ &= \inf_u \left(\phi^*(u) + \tfrac{1}{2c}\|u - cz\|^2\right) = (\phi^*)_{\frac{1}{c}}(cz).\end{aligned}$$

Meanwhile, we have

$$\begin{aligned}\inf_u L_z(u,p) &= \inf_u \left(\tfrac{1}{2c}\|u - cz\|^2 + (p,u)\right) - \phi(p) = \tfrac{1}{2c}\|cp\|^2 + (p, c(z-p)) - \phi(p)\\ &= c(p,z) - \tfrac{c}{2}\|p\|^2 - \phi(p) = \tfrac{c}{2}\|z\|^2 - \left(\phi(p) + \tfrac{c}{2}\|p - z\|^2\right).\end{aligned}$$

Thus, we deduce

$$\sup_p \inf_u L_z(u,p) = \sup_p \left(\tfrac{c}{2}\|z\|^2 - \left(\phi(p) + \tfrac{c}{2}\|p-z\|^2\right)\right) = \tfrac{c}{2}\|z\|^2 - \phi_c(z).$$

Therefore, from (5) we have

$$(\phi^*)_{\frac{1}{c}}(cz) = \inf_u \sup_p L_z(u,p) = \sup_p \inf_u L_z(u,p) = \tfrac{c}{2}\|z\|^2 - \phi_c(z),$$

which shows the first relation. Differentiating both side of this equation with respect to z and using (4) result in the second relation. □

The Moreau envelope and the proximity operator provide equivalent expressions of the inclusion $\lambda \in \partial\phi(z)$.

Proposition 2 *Let $c > 0$ be an arbitrary fixed constant and $\phi \in \Gamma_0(H)$. Then the following conditions are equivalent.*

(a) $\lambda \in \partial\phi(z)$.
(b) $z - \mathrm{prox}_{\frac{\phi}{c}}(z+\lambda/c) = 0$.
(c) $\phi(z) = \phi_c(z+\lambda/c) - \frac{1}{2c}\|\lambda\|^2$.

Proof Let the pair (z,λ) satisfy the condition $\lambda \in \partial\phi(z)$. This can be expressed as

$$0 \in \partial\phi(z) + c(z-(z+\lambda/c)) = \partial_x\left(\phi(x) + \frac{c}{2}\|x-(z+\lambda/c)\|^2\right)|_{x=z},$$

which is equivalent to $z = \mathrm{prox}_{\frac{\phi}{c}}(z+\lambda/c)$. This shows the equivalence between (a) and (b). Next we show that (b) implies (c). Suppose $z - \mathrm{prox}_{\frac{\phi}{c}}(z+\lambda/c) = 0$. Then by definition of ϕ_c, it follows that

$$\begin{aligned}\phi_c(z+\lambda/c) &= \phi(\mathrm{prox}_{\frac{\phi}{c}}(z+\lambda/c)) + \tfrac{c}{2}\|\mathrm{prox}_{\frac{\phi}{c}}(z+\lambda/c) - (z+\lambda/c)\|^2\\ &= \phi(z) + \tfrac{c}{2}\|z-(z+\lambda/c)\|^2 = \phi(z) + \tfrac{1}{2c}\|\lambda\|^2.\end{aligned}$$

Finally, we show that (c) implies (a). By definition of the Moreau envelope, it follows that

$$\phi_c(z+\lambda/c) \le \phi(u) + \frac{c}{2}\|u-(z+\lambda/c)\|^2, \qquad \forall u \in H,$$

which is equivalently written as

$$\phi(z) = \phi_c(z+\lambda/c) - \frac{1}{2c}\|\lambda\|^2 \le \phi(u) + \frac{c}{2}\|u-z\|^2 + (u-z,-\lambda), \qquad \forall u \in H.$$

This implies that the strictly convex function $u \to \phi(u) + \frac{c}{2}\|u-z\|^2 + (u-z,-\lambda)$ attains its minimum at z. Thus

$$0 \in \partial_u\left(\phi(u) + \frac{c}{2}\|u-z\|^2 + (u-z,-\lambda)\right)|_{u=z} = \partial\phi(u) - \lambda,$$

which proves that (c) implies (a). □

3 The optimality systems

In the classical optimization problem for a smooth cost function with equality constraints by smooth maps, it is well known that saddle points are characterized by Lagrange optimality system of the (standard) Lagrangian associated with the optimization problem. In this section, we show that the augmented Lagrangian L_c generalizes the classical result to the nonsmooth convex optimization problem (1).

Proposition 3 *Let $c > 0$, f be convex and continuously differentiable, and $\phi \in \Gamma_0(H)$. The augmented Lagrangian L_c satisfies the following properties.*

(a) *L_c is finite for all $x \in X$ and for all $\lambda \in H$.*

(b) *L_c is convex and continuously differentiable with respect to x, and is concave and continuously differentiable with respect to λ. Further, for all $(x,\lambda) \in X \times H$ and for all $c > 0$, the gradients D_xL_c and $D_\lambda L_c$ are written respectively as*

$$D_xL_c(x,\lambda) = D_xf(x) + c\Lambda^*(\Lambda x + \lambda/c - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x + \lambda/c)), \tag{6}$$

$$D_\lambda L_c(x,\lambda) = \Lambda x - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x + \lambda/c). \tag{7}$$

(c) *$D_xL_c(x,\lambda)$ can be expressed in terms of $D_\lambda L_c(x,\lambda)$ by*

$$D_xL_c(x,\lambda) = D_xf(x) + \Lambda^*\left(\lambda + cD_\lambda L_c(x,\lambda)\right). \tag{8}$$

Proof All the assertions follow directly from the differentiability and convexity of f, and Proposition 1. □

Theorem 1 *Let $c > 0$, f be convex and continuously differentiable, and $\phi \in \Gamma_0(H)$. The following conditions on a pair $(\bar{x},\bar{\lambda})$ are equivalent.*

(a) *(optimality system) A pair $(\bar{x},\bar{\lambda}) \in X \times H$ satisfies the optimality system*

$$D_xf(\bar{x}) + \Lambda^*\bar{\lambda} = 0 \quad \text{and} \quad \bar{\lambda} \in \partial\phi(\Lambda\bar{x}). \tag{9}$$

(b) *(Lagrange optimality system) A pair $(\bar{x},\bar{\lambda}) \in X \times H$ satisfies the Lagrange optimality system*

$$D_xL_c(\bar{x},\bar{\lambda}) = 0 \quad \text{and} \quad D_\lambda L_c(\bar{x},\bar{\lambda}) = 0, \tag{10}$$

where the gradients of L_c with respect to x and λ are given by (6) *and* (7)*, respectively. More precisely, $(\bar{x},\bar{\lambda})$ satisfies the nonlinear system:*

$$\begin{cases} D_xf(x) + c\Lambda^*\left(\Lambda x + \lambda/c - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x + \lambda/c)\right) = 0 \\ \Lambda x - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x + \lambda/c) = 0. \end{cases}$$

(c) *(saddle point) A pair $(\bar{x},\bar{\lambda}) \in X \times H$ is a saddle point of L_c:*

$$L_c(\bar{x},\lambda) \le L_c(\bar{x},\bar{\lambda}) \le L_c(x,\bar{\lambda}), \qquad \forall x \in X,\ \forall\lambda \in H. \tag{11}$$

Proof First we show the equivalence between (a) and (b). Suppose (a) holds, i.e., $(\bar{x},\bar{\lambda})$ satisfies the optimality system. By Propostion 2, the inclusion $\lambda \in \partial\phi(\Lambda x)$ can be written as $\Lambda x = \mathrm{prox}_{\frac{\phi}{c}}(\Lambda x + \lambda/c)$, and it holds independently of $c > 0$. Hence, it follows from (7) that

$$D_\lambda L_c(\bar{x},\bar{\lambda}) = \Lambda\bar{x} - \mathrm{prox}_{\frac{\phi}{c}}(\Lambda\bar{x} + c^{-1}\bar{\lambda}) = 0,$$

for all $c > 0$, and thereby from Proposition 3(c) and the first optimality system

$$D_x L_c(\bar{x},\bar{\lambda}) = D_x f(\bar{x}) + \Lambda^*\left(\bar{\lambda} + cD_\lambda L_c(\bar{x},\bar{\lambda})\right) = D_x f(\bar{x}) + \Lambda^*\bar{\lambda} = 0,$$

for all $c > 0$. Similarly, we can show that (b) for some $c > 0$ implies (a).

Next we show the equivalence between (b) and (c). Suppose that $(\bar{x},\bar{\lambda})$ is a saddle point of L_c. Then from [4, Chap. 6, Prop. 1.2] we have

$$\min_{x\in X}\sup_{\lambda\in H} L_c(x,\lambda) = L_c(\bar{x},\bar{\lambda}) = \max_{\lambda\in H}\inf_{x\in X} L_c(x,\lambda).$$

The second equality implies that $\bar{x}$ solves the convex optimization problem $\min_{x\in X} L_c(x,\bar{\lambda})$, and hence it follows that $D_x L_c(x,\bar{\lambda}) = 0$ at $x = \bar{x}$. The similar argument proves that $D_\lambda L_c(\bar{x},\bar{\lambda}) = 0$. Conversely, if $(\bar{x},\bar{\lambda})$ satisfies the Lagrange optimality system, then from the convexity $L_c(x,\lambda)$ with respect to x, we have

$$L_c(x,\bar{\lambda}) - L_c(\bar{x},\bar{\lambda}) \geq \langle D_x L_c(\bar{x},\bar{\lambda}), x - \bar{x}\rangle_{X^*,X} = 0 \quad \forall x \in X.$$

Similarly, by the concavity of $L_c(x,\cdot)$, we deduce $L_c(\bar{x},\lambda) \leq L_c(\bar{x},\bar{\lambda})$. □

Corollary 1 *If one of the conditions in Theorem 1 holds, then $\bar{x}$ is a solution of problem* (1).

Proof Assume that there exists a pair $(\bar{x},\bar{\lambda})$ satisfying the optimality system (9). The system implies that $0 \in D_x f(\bar{x}) + \Lambda^*\partial\phi(\Lambda\bar{x})$. By [4, Chap. 1, Prop. 5.7]) we have

$$\Lambda^*\partial\phi(\Lambda x) \subset \partial(\phi\circ\Lambda)(x), \quad \forall x \in X.$$

Therefore it follows that

$$0 \in D_x f(\bar{x}) + \Lambda^*\partial\phi(\Lambda\bar{x}) \subset D_x f(\bar{x}) + \partial(\phi\circ\Lambda)(\bar{x}) = \partial(f + \phi\circ\Lambda)(\bar{x}),$$

which shows that $\bar{x}$ is a solution of the minimization problem (1). □

Remark 1 We refer to [7, Chap. 4] for a sufficient condition for the existence of a pair satisfying the optimality system (9).

Corollary 2 *The Lagrange optimality system can also be written as*

$$D_x f(\bar{x}) + \Lambda^*\bar{\lambda} = 0 \quad \textit{and} \quad \Lambda\bar{x} - \mathrm{prox}_{\frac{\phi}{c}}(\Lambda\bar{x} + \bar{\lambda}/c) = 0. \tag{12}$$

Proof It follows directly from Proposition 3, (7) and (8). □

The Lagrange optimality system (10) is closely related to the optimality system derived in [17,22] which is given by using *the generalized Moreau-Yosida approximation* $\psi_c(z,\lambda)$ defined by

$$\psi_c(z,\lambda) = \phi_c(z+\lambda/c) - \tfrac{1}{2c}\|\lambda\|^2.$$

Let us assume that a pair $(\bar{x},\bar{\lambda}) \in X \times Z$ satisfies the optimality system (9). It is shown in [17, Thm. 4.5] that the pair satisfies the following optimality condition for every $c > 0$.

$$\bar{x} = \min_x L_c(x,\bar{\lambda}) \quad \text{and} \quad \bar{\lambda} = (D_x\psi_c)(\Lambda\bar{x},\bar{\lambda}).$$

The first relation implies the inequality $L_c(\bar{x},\bar{\lambda}) \leq L_c(x,\bar{\lambda})$ for all $x \in X$, which is the second inequality of (11). Meanwhile, by the definition of $\psi_c(x,\lambda)$ and Proposition 1(d), we have

$$\begin{aligned}(D_x\psi_c)(\Lambda x,\lambda) &= \phi_c'(\Lambda x+\lambda/c)\\ &= c(\Lambda x+\lambda/c - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x+\lambda/c))\\ &= \lambda + c(\Lambda x - \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x+\lambda/c)).\end{aligned}$$

In view of the expression (7), the second relation implies $D_\lambda L_c(\bar{x},\bar{\lambda}) = 0$, which is the second equation of the Lagrange optimality system (10). Alternatively, the following optimality condition in the form of equation is given in [22]:

$$D_x f(\bar{x}) + \Lambda^*\bar{\lambda} = 0 \quad \text{and} \quad \bar{\lambda} = (D_x\psi_c)(\Lambda\bar{x},\bar{\lambda}).$$

Similarly, one can show that this optimality system is equivalent to (12).

4 Linear Newton method for the Lagrange optimality system

In this section, we present a linear Newton method for the nonsmooth optimization problem (1) on the basis of the Lagrange optimality system. We also illustrate the method on three elementary examples. To keep the presentation simple, we restrict our discussions to finite-dimensional spaces.

4.1 Linear Newton method

We begin with the concept of linear Newton approximation, which provides a building block for designing Newton type algorithms for problem (1). For a comprehensive treatment and for further references on the subject one may refer to [20].

Definition 2 Let $\Phi\colon \mathbb{R}^m \to \mathbb{R}^n$ be locally Lipschitz continuous. We say that the map Φ admits a *linear Newton approximation (LNA)* at $\bar{\xi} \in \mathbb{R}^m$ if there exists a set-valued map $T\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^{n\times m}$ such that:

(a) The set of matrices $T(\xi)$ is nonempty and compact for each $\xi \in \mathbb{R}^m$;
(b) T is upper semicontinuous at $\bar{\xi}$;

(c) The following limit holds:

$$\lim_{\substack{\bar{\xi}\neq\xi\to\bar{\xi}\\ V\in T(\xi)}} \frac{\|\Phi(\xi)+V(\bar{\xi}-\xi)-\Phi(\bar{\xi})\|}{\|\xi-\bar{\xi}\|}=0.$$

We also say that T is a *linear Newton approximation scheme* of Φ.

A linear Newton iteration for solving the nonlinear equation $\Phi(\xi)=0$ is defined by

$$\xi^{k+1}=\xi^k-V_k^{-1}\Phi(\xi^k), \text{ with } V_k\in T(\xi^k). \tag{13}$$

The local convergence of the iterate is ensured if the matrix V_k is singular for all k.

Theorem 2 ([20, Thm. 7.5.15]) *Let $\Phi\colon \mathbb{R}^n\to\mathbb{R}^n$ be locally Lipschitz continuous and admit a LNA T at $\xi^*\in\mathbb{R}^n$ such that $\Phi(\xi^*)=0$. If every matrix $V\in T(\xi^*)$ is nonsingular, then the iterate* (13) *converges Q-superlinearly to the solution ξ^* provided that ξ^0 is sufficiently close to ξ^*.*

In addition to the Newton iteration (13) we can also define inexact version of linear Newton methods, the Levenberg-Marquardt (LM) method and the inexact version of LM method, and establish their local convergence as well as characterize their convergence rate, see. e.g., [20]. The linear Newton method for the Lagrange optimality system, which we shall develop later in the section, can be extended for these methods along similar lines, but we restrict ourselves to the basic Newton method (13).

To provide a class of Lipschitz maps that admit a LNA, we shall make use of the notion of generalized Jacobian and semismoothness. Let $\Phi\colon\mathbb{R}^m\to\mathbb{R}^n$ be a locally Lipschitz continuous map. Rademacher's Theorem [23, Sect. 3.1.2] states that a locally continuous map is differentiable almost everywhere. Denote by N_Φ a set of measure zero such that Φ is differentiable on $\mathbb{R}^m\setminus N_\Phi$. The *limiting Jacobian* of Φ at ξ is the set

$$\partial_B\Phi(\xi):=\left\{G\in\mathbb{R}^{n\times m}\mid \exists\{\xi^k\}\subset\mathbb{R}^m\setminus N_\Phi \text{ with } \xi^k\to\xi, D_x\Phi(\xi^k)\to G\right\}.$$

The *(Clarke's) generalized Jacobian* $\partial\Phi(\xi)$ of Φ at $\xi\in\mathbb{R}^m$ is the convex hull of the limiting Jacobian:

$$\partial\Phi(\xi)=\mathrm{conv}(\partial_B\Phi(\xi)).$$

We denote by $\partial_B\Phi$ the set valued map $\xi\to\partial_B\Phi(\xi)$ for $\xi\in\mathbb{R}^m$. The set valued map $\partial\Phi$ for the generalized Jacobian is defined analogously.

A possible choice for a LNA scheme of a locally Lipschitz map is the limiting or generalized Jacobian of the map. This attempt, in the absence of additional assumption on Φ, is doomed because both of them do not necessarily satisfy the approximation property of condition (c) in Definition 2. This drawback can be ameliorated by employing the notion of semismoothness, which narrows down the class of Lipschitz maps so that each of $\partial\Phi$ and $\partial_B\Phi$ provides a LNA scheme of the map.

Definition 3 Let $\varPhi\colon \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitz map. We say that $\varPhi$ is *semismooth* at $\bar{\xi} \in \mathbb{R}^m$ if $\varPhi$ is directionally differentiable near $\bar{\xi}$ and the following limit holds:

$$\lim_{\bar{\xi}\neq\xi\to\bar{\xi}\to 0} \frac{\|\varPhi'(\xi;\xi-\bar{\xi})-\varPhi'(\bar{\xi};\xi-\bar{\xi})\|}{\|\xi-\bar{\xi}\|}=0,$$

where $\varPhi'(\xi;h)$ denotes the directional derivative of $\varPhi$ at $\xi \in \mathbb{R}^m$ along the direction $h \in \mathbb{R}^m$.

Proposition 4 *Assume that a locally Lipschitz map $\varPhi\colon \mathbb{R}^m \to \mathbb{R}^n$ is semismooth at $\xi \in \mathbb{R}^m$, then each of $\partial\varPhi$ and $\partial_B\varPhi$ defines a LNA scheme of $\varPhi$ at ξ.*

Proof It follows from [20, Prop. 7.1.4] that the set valued map $\partial\varPhi$ satisfies the condition (a) and (b) of Definition 2, while, from [20, Thm. 7.4.3], the map also satisfies the condition (c). We refer the proof for the limiting Jacobian to [20, Prop. 7.5.16].□

4.2 Linear Newton method for the Lagrange optimality system

We are ready to present a Newton algorithm for problem (1). Let the map $\varPhi_c\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n+m}$ be defined by

$$\varPhi_c(x,\lambda)=\begin{bmatrix} D_x L_c(x,\lambda) \\ D_\lambda L_c(x,\lambda)\end{bmatrix}.$$

Proposition 3 shows that the map $\varPhi_c$ is the difference of a smooth and nonsmooth part

$$\varPhi_c(x,\lambda)=\varPhi_s(x,\lambda)-\varPhi_{ns}(x,\lambda),$$

where

$$\varPhi_s(x,\lambda):=\begin{bmatrix} D_x f(x)+c\Lambda^*\Lambda x+\Lambda^*\lambda \\ \Lambda x\end{bmatrix} \quad \text{and} \quad \varPhi_{ns}(x,\lambda)=\begin{bmatrix} c\Lambda^*\operatorname{prox}_{\frac{\phi}{c}}(\Lambda x+\lambda/c) \\ \operatorname{prox}_{\frac{\phi}{c}}(\Lambda x+\lambda/c)\end{bmatrix}.$$

The Jacobian of $\varPhi_s(x,\lambda)$ is

$$D_{x,\lambda}\varPhi_s(x,\lambda)=\left[\begin{array}{c|c} D_x^2 f(x)+c\Lambda^*\Lambda & \Lambda^* \\ \hline \Lambda & 0\end{array}\right],$$

and the (matrix valued) map $D_{x,\lambda}\varPhi_s$ defines a LNA scheme of the smooth map $\varPhi_s$ at every point (x,λ). By the sum rule (see, e.g., [20, Thm. 7.5.18]), a LNA scheme of $\varPhi_c$ is provide by $T=D_{x,\lambda}\varPhi_s-T_{ns}$ where T_{ns} is a LNA scheme of $\varPhi_{ns}$. The next result shows that the task of determining T_{ns} is reduced to the one of computing a LNA scheme of the proximity operator.

Lemma 1 *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ and $c>0$. Let T_p be a LNA scheme of the proximity operator $\operatorname{prox}_{\frac{\phi}{c}}$. Then the set-valued map*

$$T_{ns}(x,\lambda):=\left\{\begin{bmatrix} c\Lambda^* \\ I\end{bmatrix} G\begin{bmatrix}\Lambda & c^{-1}I\end{bmatrix} \mid G\in T_p(\Lambda x+\lambda/c)\right\}\subset \mathbb{R}^{n+m,n+m}$$

is a LNA of the map $\varPhi_{ns}$.

Proof Since T_p is upper semi-continuous and the set $T_p(z)$ is compact by definition, so is the set-valued map $(x,\lambda) \to T_{ns}(x,\lambda)$, which implies that the T_{ns} satisfies the conditions (a) and (b) in Def. 2. One can verify that the set valued map T_{ns} satisfies the condition (c) in the definition by employing the sum rule ([20, Thm. 7.5.18]) and the chain rule ([20, Thm. 7.5.17]). □

We now turn our attention to define a possible LNA scheme of a proximity operator. By Proposition 1, the proximity operator is nonexpansive, and therefore it is Lipschitz continuous. Hence the limiting Jacobian $\partial_B(\operatorname{prox}_{\phi/c})(z)$ is well-defined for all $z \in \mathbb{R}^m$, and so also is the generalized Jacobian $\partial(\operatorname{prox}_{\phi/c})(z)$. The next result, due to [24, Thm. 3.2], gives the basic properties of the generalized Jacobian of the proximity operator.

Proposition 5 *For any $\phi \in \Gamma_0(\mathbb{R}^m)$, every $G \in \partial(\operatorname{prox}_{\frac{\phi}{c}})(z)$ is a symmetric positive semidefinite matrix with $\|G\| \leq 1$.*

Now we can specify a LNA scheme of the map $D_{x,\lambda}L_c$ at (x,λ).

Proposition 6 *Let $\phi \in \Gamma_0(\mathbb{R}^m)$ and $c > 0$. Assume that the proximity operator* $\operatorname{prox}_{\frac{\phi}{c}}$ *is semismooth. Then the set-valued map $T : \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^{(n+m)\times(n+m)}$ defined by*

$$T(x,\lambda) := \left\{ \left[\begin{array}{c:c} D_x^2 f(x) + c\Lambda^*(I-G)\Lambda & ((I-G)\Lambda)^* \\ \hdashline (I-G)\Lambda & -c^{-1}G \end{array} \right] \mid G \in \partial(\operatorname{prox}_{\frac{\phi}{c}})(z) \right\}, \tag{14}$$

with $z = \Lambda x + \lambda/c$, is a LNA scheme of the map Φ_c at $(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m$.

Proof The symmetry of the generalized Jacobian of a proximity operator allows to write $\Lambda^* G = (G\Lambda)^*$ for $G \in \partial(\operatorname{prox}_{\frac{\phi}{c}})(\Lambda x + \lambda/c)$, which yields

$$\left[\begin{array}{c:c} D_x^2 f(x) + c\Lambda^*(I-G)\Lambda & ((I-G)\Lambda)^* \\ \hdashline (I-G)\Lambda & -c^{-1}G \end{array} \right] = \left[\begin{array}{c:c} D_x^2 f(x) + c\Lambda^*\Lambda & \Lambda^* \\ \hdashline \Lambda & 0 \end{array} \right] - \begin{bmatrix} c\Lambda^* \\ I \end{bmatrix} G \begin{bmatrix} \Lambda & c^{-1}I \end{bmatrix}.$$

From Proposition 4 and the assumption that $\operatorname{prox}_{\frac{\phi}{c}}$ is semismooth, it follows that the generalized Jacobian $\partial(\operatorname{prox}_{\frac{\phi}{c}})(z)$ is a LNA scheme of the proximity operator $\operatorname{prox}_{\frac{\phi}{c}}(z)$, which together with Lemma 1 shows that $T_{ns}(x,\lambda)$ with $T_p(\Lambda x + \lambda/c) = \partial(\operatorname{prox}_{\frac{\phi}{c}})(\Lambda x + \lambda/c)$ defines a LNA scheme of Φ_{ns} at (x,λ). Thus $T = D_{x,\lambda}\Phi_s - T_{ns}$ defines a LNA scheme of Φ_c at (x,λ). □

Remark 2 One can replace the generalized Jacobian $\partial(\operatorname{prox}_{\frac{\phi}{c}})(z)$ in (14) with the limiting Jacobian $\partial_B(\operatorname{prox}_{\frac{\phi}{c}})(z)$.

Remark 3 The class of semismooth maps is broad enough to include a variety of proximity operators frequently encountered in practice, see, e.g., [24, Sect. 5].

The proposed algorithm is given in Algorithm 1.

Algorithm 1 Linear Newton algorithm for the Lagrange optimality system.

```
1: Chose $(x^0, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^m$.
2: If $\Phi_c(x^k, \lambda^k) = 0$, stop.
3: Let $z^k = \Lambda x^k + \lambda^k / c$, and compute an element $G_k$ of the generalized Jacobian $\partial(\operatorname{prox}_{\frac{\phi}{c}})(z^k)$.
4: Compute a direction $(d_x^k, d_\lambda^k)$ by
   $$\left[\begin{array}{c:c} D_x^2 f(x^k) + c\Lambda^*(I - G_k)\Lambda & ((I - G_k)\Lambda)^* \\ \hdashline (I - G_k)\Lambda & -c^{-1} G_k \end{array}\right] \begin{bmatrix} d_x^k \\ d_\lambda^k \end{bmatrix} = - \begin{bmatrix} D_x L_c(x^k, \lambda^k) \\ D_\lambda L_c(x^k, \lambda^k) \end{bmatrix}. \qquad (15)$$
5: Set $x^{k+1} = x^k + d_x^k$ and $\lambda^{k+1} = \lambda^k + d_\lambda^k$.
6: Go back to Step 2.
```

Remark 4 Proposition 4 allows to replace the generalized Jacobian $\partial(\operatorname{prox}_{\frac{\phi}{c}})(z)$ with the limiting Jacobian $\partial_B(\operatorname{prox}_{\frac{\phi}{c}})(z)$.

Remark 5 A simple calculation using Theorem 3 shows that the update at Steps 4 and 5 can be replaced with

$$\left[\begin{array}{c:c} D_x^2 f(x^k) & \Lambda^* \\ \hdashline (I - G_k)\Lambda & -c^{-1} G_k \end{array}\right] \begin{bmatrix} x^{k+1} \\ \lambda^{k+1} \end{bmatrix} = \begin{bmatrix} D_x^2 f(x^k) x^k - D_x f(x^k) \\ \operatorname{prox}_{\frac{\phi}{c}}(z^k) - G_k z^k \end{bmatrix}. \tag{16}$$

The local convergence of Algorithm 1 follows from Theorem 2, if every element of $T(x, \lambda)$ defined by (14) is nonsingular. The next result gives one sufficient condition for the nonsingularity.

Proposition 7 *Assume that Λ is surjective, $D^2 f(x)$ is strictly positive definite, and the norm is bound from below uniformly in x, that is, there exists a $\delta > 0$ such that*

$$(D_x^2 f(x) d, d) > \delta \|d\|^2 \quad \forall d \in \mathbb{R}^n.$$

Then every element of $T(x, \lambda)$ is nonsingular for all (x, λ).

Proof A saddle point matrix of the form

$$\begin{bmatrix} A & B^* \\ B & -C \end{bmatrix},$$

where A is symmetric positive definite and C is symmetric positive semidefinite, is nonsingular if and only if $\ker(C) \cap \ker(B^*) = 0$, see, e.g., [25, Thm. 3.1]. Note that $D_x^2 f(x)$ is symmetric positive definite by assumption, and G and $I - G$ are symmetric positive semidefinite, cf. Proposition 5. Hence the matrix $D_x^2 f(x) + c\Lambda^*(I - G)\Lambda$ is symmetric positive definite. Now let $d \in \ker(G) \cap \ker(((I - G)\Lambda)^*)$. We then have

$$Gd = 0 \quad \text{and} \quad ((I - G)\Lambda)^* d = 0.$$

Appealing again to the identity $G^* = G$ from Proposition 5, it immediately follows that $\Lambda^* d = 0$. Then the surjectivity of Λ implies $d \in \ker(\Lambda^*) = \operatorname{Im}(\Lambda)^\perp = 0$. □

The local convergence of Algorithm 1 follows from Theorem 2, Propositions 6 and Proposition 7.

Theorem 3 *Let f be smooth, $\phi \in \Gamma_0(\mathbb{R}^m)$, and $c > 0$. Let us assume there exits a unique solution $(\bar{x}, \bar{\lambda})$ of the Lagrange optimality system* (10)*. We also assume that the assumptions on f and Λ in Proposition 7 are satisfied, and that the proximity operator is semismooth on $\mathbb{R}^m$. Then the Newton system* (15) *is solvable, and the sequence (x^k, λ^k) generated by Algorithm 1 converges to the solution $(\bar{x}, \bar{\lambda})$ superlinearly in a neighborhood of $(\bar{x}, \bar{\lambda})$.*

4.3 Examples

We illustrate Algorithm 1 on three examples: bilateral constraints, ℓ^1 penalty and total variation penalty. We begin with a useful result for computing the generalized (limiting) Jacobian for (block) separable functions [24, Prop. 3.3]. Let $(m_1, \ldots, m_N)$ be an N partition of m, i.e., $\sum_{i=1}^N m_i = m$, and $z \in \mathbb{R}^m$ be decomposed into N blocks of variables with $z_i \in \mathbb{R}^{m_i}$. The function $\phi \in \Gamma_0(\mathbb{R}^m)$ is said to be *(block) separable* if $\phi(z) = \sum_{i=1}^N \phi_i(z_i)$ for N functions $\phi_i \in \Gamma_0(\mathbb{R}^m)$.

Proposition 8 *If $\phi \in \Gamma_0(\mathbb{R}^m)$ is (block) separable then every element of the generalized Jacobian $\partial(\mathrm{prox}_{\frac{\phi}{c}})(x)$ is also a (block) diagonal matrix.*

Example 1 Let us consider the following optimization problem with bilateral inequality constraints

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad a \le \Lambda x \le b,$$

where f is a smooth function, $a, b \in \mathbb{R}^m$ and $\Lambda \in \mathbb{R}^{m\times n}$.

The problem can be reformulated into (1) with $\phi(z) = I_S(z)$, where $I_S(z)$ is the characteristic function of the set $S = \{z \in \mathbb{R}^m \mid a_j \le z_j \le b_j,\ j = 1, \ldots, m\}$. Clearly, the proximity operator $\mathrm{prox}_{\frac{\phi}{c}} : \mathbb{R}^m \to \mathbb{R}^m$ is given by

$$\mathrm{prox}_{\frac{\phi}{c}}(z) = [\max(a_1, \min(b_1, z_1)), \ldots, \max(a_m, \min(b_m, z_m))]^{\mathrm{T}}.$$

Since the proximity operator is separable, a limiting Jacobian $G \in \partial_B\, \mathrm{prox}_{\frac{\phi}{c}}(z)$ is diagonal matrix by Proposition 8:

$$G_{j,j} = \begin{cases} 1 & \text{if } a_j < z_j < b_j, \\ \{0, 1\} & \text{if } z_j \in \{a_j, b_j\}, \\ 0 & \text{otherwise.} \end{cases}$$

Now let (x, λ) be the current iterate, and $z = \Lambda x + \lambda/c$. We denote by $\mathbf{o}$ the index set $\{j \mid G_{j,j} = 0\} \subset \{1, 2, \ldots, m\}$, and by $\mathbf{i}$ its complement. Then $\mathbf{i} \cap \mathbf{o} = \emptyset$ and $\mathbf{i} \cup \mathbf{o} = \{1, 2, \ldots, m\}$. We shall denote by $x_{\mathbf{o}}$ the subvector of x, consisting of entries of x whose indices are listed in $\mathbf{o}$. The submatrix of Λ denoted by $\Lambda_{\mathbf{o}}$ is defined analogously. For example, if $\mathbf{o} = \{o_1, o_2, \ldots, o_\ell\}$ where ℓ is the number of elements of the set $\mathbf{o}$, then $x_{\mathbf{o}}$ is $\ell \times 1$ column vector, and $A_{\mathbf{o}}$ is $\ell \times n$ matrix given respectively by

$$x_{\mathbf{o}} = \begin{bmatrix} x_{o_1} \\ x_{o_2} \\ \vdots \\ x_{o_m} \end{bmatrix} \quad \text{and} \quad A_{\mathbf{o}} = \begin{bmatrix} A_{o_1,1} & A_{o_1,2} & \cdots & A_{o_1,n} \\ A_{o_2,1} & A_{o_2,2} & \cdots & A_{o_2,n} \\ \vdots & \vdots & & \vdots \\ A_{o_\ell,1} & A_{o_\ell,2} & \cdots & A_{o_\ell,n} \end{bmatrix}.$$

With the new updates denoted by x^+ and λ^+, the Newton update (16) yields

$$\begin{cases} \lambda_{\mathbf{i}}^+ = c(z - \operatorname{prox}_{\frac{\phi}{c}}(z))_{\mathbf{i}}, \\ \begin{bmatrix} D_x^2 f(x) & \Lambda_{\mathbf{o}}^* \\ \Lambda_{\mathbf{o}} & \end{bmatrix} \begin{bmatrix} x^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} D_x^2 f(x)x - D_x f(x) - \Lambda_{\mathbf{i}}^* \lambda_{\mathbf{i}}^+ \\ \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix}. \end{cases}$$

In this example, we have $z_{\mathbf{i}} = \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{i}}$, and the Newton update is further simplified as

$$\begin{bmatrix} D_x^2 f(x) & \Lambda_{\mathbf{o}}^* \\ \Lambda_{\mathbf{o}} & \end{bmatrix} \begin{bmatrix} x^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} D_x^2 f(x)x - D_x f(x) \\ \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix} \quad \text{and} \quad \lambda_{\mathbf{i}}^+ = 0.$$

In particular if f is a quadratic function $f(x) = \frac{1}{2}(x, Ax) - (b, x)$, the algorithm reduces to the primal-dual active set algorithm developed in [7,26]:

$$\begin{bmatrix} A & \Lambda_{\mathbf{o}}^* \\ \Lambda_{\mathbf{o}} & \end{bmatrix} \begin{bmatrix} x^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} b \\ \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix} \quad \text{and} \quad \lambda_{\mathbf{i}}^+ = 0.$$

Example 2 Consider the following ℓ^1 type optimization problem

$$\min_{x \in \mathbb{R}^b} f(x) + \alpha |\Lambda x|_{\ell^1},$$

where f is smooth function, $\Lambda \in \mathbb{R}^{m \times n}$, $|z|_{\ell^1}$ is the ℓ^1 norm, and $\alpha > 0$ is a regularization parameter.

Let $\phi(z) = \alpha |z|_{\ell^1}$. Its proximity operator $\operatorname{prox}_{\frac{\phi}{c}}$ is the well known soft-thresholding operator

$$\begin{aligned} \operatorname{prox}_{\frac{\phi}{c}}(z) &= [\operatorname{prox}_{\frac{\alpha}{c}|\cdot|}(z_1), \dots, \operatorname{prox}_{\frac{\alpha}{c}|\cdot|}(z_m)]^{\mathrm{T}}, \\ \operatorname{prox}_{\frac{\alpha}{c}|\cdot|}(s) &= \max(s - \tfrac{\alpha}{c}, \min(s + \tfrac{\alpha}{c}, 0)), \quad s \in \mathbb{R}. \end{aligned}$$

A limiting Jacobian $G \in \partial_B(\operatorname{prox}_{\frac{\phi}{c}})(z)$ is diagonal matrix given by

$$G_{j,j} = \begin{cases} 1 & \text{if } |z_j| > \frac{\alpha}{c}, \\ \{0,1\} & \text{if } |z_j| = \frac{\alpha}{c}, \\ 0 & \text{othewise.} \end{cases}$$

We denote by $\mathbf{o}$ the index set $\{j \mid G_{j,j}=0\} \subset \{1,2,\ldots,m\}$, and by $\mathbf{i}$ its complement, and $z=\Lambda x+\lambda/c$. We note that the relation $c(z-\mathrm{prox}_{\frac{\phi}{c}}(z))_{\mathbf{i}} = c\,\mathrm{sign}(z_{\mathbf{i}})$ holds. An argument similar to Example 1 yields the following Newton update

$$\begin{cases} \lambda_{\mathbf{i}}^{+} = c\,\mathrm{sign}(z_{\mathbf{i}}), \\ \begin{bmatrix} D_x^2 f(x) & \Lambda_{\mathbf{o}}^* \\ \Lambda_{\mathbf{o}} & \end{bmatrix} \begin{bmatrix} x^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} D_x^2 f(x)x - D_x f(x) - \Lambda_{\mathbf{i}}^* \lambda_{\mathbf{i}}^+ \\ \mathrm{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix}. \end{cases}$$

For the quadratic function $f=\frac{1}{2}(x,Ax)-(b,x)$, we obtain a primal-dual active set algorithm for ℓ^1 norm regularization

$$\begin{cases} \lambda_{\mathbf{i}}^{+} = c\,\mathrm{sign}(z_{\mathbf{i}}), \\ \begin{bmatrix} A & \Lambda_{\mathbf{o}}^* \\ \Lambda_{\mathbf{o}} & \end{bmatrix} \begin{bmatrix} x^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} b - \Lambda_{\mathbf{i}}^* \lambda_{\mathbf{i}}^+ \\ \mathrm{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix}. \end{cases}$$

Example 3 We consider the total variation denoising problem of recovering an image from a given noisy image $F \in \mathbb{R}^{n,n}$

$$\min_{u\in\mathbb{R}^{n^2}} \tfrac{1}{2}\|u-f\|_{\mathbb{R}^{n^2}}^2 + \alpha\phi(\Lambda u),$$

where $f \in \mathbb{R}^{n^2}$ stacks the columns of the two dimensional image F, and α is the regularization parameter. $\Lambda_0 u = [(D_1 u)^{\mathrm{T}}, (D_2 u)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2n^2}$, where $D_1 = I\otimes D$ and $D_2 = D\otimes I$ are finite difference matrices for y- and x-direction, respectively; D is the one-dimensional finite difference matrix with periodic boundary condition; and I is the identity matrix. The matrix Λ is defined as $\Lambda = P_\sigma \Lambda_0$ where P_σ is the permutation matrix associated with the permutation σ defined by

$$\sigma(i) = \begin{cases} 2i-1, & 1\le i\le n^2, \\ 2(i-n^2), & n^2+1 \le i \le 2n^2. \end{cases}$$

The regularization term ϕ is defined by

$$\phi(z) = \sum_{i=1,3,\ldots,2n^2-1} \sqrt{z_i^2+z_{i+1}^2} \quad z\in\mathbb{R}^{2n^2}.$$

The proximity operator $\mathrm{prox}_{\frac{\phi}{c}}$ for $\phi(z)$ is block-separable, and its i-th and $(i+1)$-th components for $i=1,3,\ldots,2n^2-1$ are given by

$$([\mathrm{prox}_{\frac{\phi}{c}}(z)]_i, [\mathrm{prox}_{\frac{\phi}{c}}(z)]_{i+1}) = \begin{cases} (z_i - \frac{z_i}{c'r_i}, z_{i+1} - \frac{z_{i+1}}{c'r_i}) & \text{if } c'r_i \ge 1 \\ (0,0) & \text{if } c'r_i < 1 \end{cases}$$

where $c' = \frac{c}{\alpha}$ and $r_i = (z_i^2 + z_{i+1}^2)^{1/2}$. A limiting Jacobian $G \in \partial_B(\mathrm{prox}_{\frac{\phi}{c}})(z)$ for $z\in\mathbb{R}^{2n^2}$ is block-diagonal: The i-th block 2 by 2 matrix, in other words the $(i,i),(i,i+$

$1),(i+1,i),(i+1,i+1)$ components of $G \in \mathbb{R}^{2n^2 \times 2n^2}$ are given by

$$\begin{bmatrix} G_{i,i} & G_{i,i+1} \\ G_{i+1,i} & G_{i+1,i+1} \end{bmatrix} = \begin{cases} I - \dfrac{1}{c' r_i^3} \begin{bmatrix} z_{i+1}^2 & -z_i z_{i+1} \\ -z_i z_{i+1} & z_i^2 \end{bmatrix} & \text{if } c' r_i > 1 \\ 0 & \text{if } c' r_i < 1, \end{cases}$$

and any of these two matrices if $c' r_i = 1$ (see e.g., [24, Sect. 5.2.3]). Hence, with $\lambda \in \mathbb{R}^{2n^2}$ being the Lagrange multiplier and $z = \Lambda u + \lambda / c \in \mathbb{R}^{2n^2}$, the Newton update is give by

$$u^+ + \Lambda^* \lambda^+ = f, \quad (I - G)\Lambda u^+ - c^{-1} G \lambda^+ = \operatorname{prox}_{\frac{\phi}{c}}(z) - Gz. \tag{17}$$

Let $\mathbf{o}$ be the index set $\{2i-1, 2i \mid i\text{-th block matrix of } G \text{ is } 0\} \subset \{1,2,\ldots,2n^2\}$, and $\mathbf{i}$ be its complement. We denote by $G_{\mathbf{i}}$ the block diagonal matrix composed of i-th block for $i \in \mathbf{i}$. Then the system (17) is written as

$$\begin{aligned} & u^+ + \Lambda_{\mathbf{o}}^* \lambda_{\mathbf{o}}^+ + \Lambda_{\mathbf{i}}^* \lambda_{\mathbf{i}}^+ = f, \\ & \Lambda_{\mathbf{o}} u^+ = \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}}, \quad (I_{\mathbf{i}} - G_{\mathbf{i}})\Lambda_{\mathbf{i}} u^+ - c^{-1} G_{\mathbf{i}} \lambda_{\mathbf{i}}^+ = \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{i}} - (Gz)_{\mathbf{i}}. \end{aligned}$$

Hence we arrive at a prima-dual active set algorithm for the total variation regularization:

$$\begin{cases} \left[\begin{array}{c:c} A + c\Lambda_{\mathbf{i}}^* G_{\mathbf{i}}^{-1}(I_{\mathbf{i}} - G_{\mathbf{i}})\Lambda_{\mathbf{i}} & \Lambda_{\mathbf{o}}^* \\ \hdashline \Lambda_{\mathbf{o}} & 0 \end{array}\right] \begin{bmatrix} u^+ \\ \lambda_{\mathbf{o}}^+ \end{bmatrix} = \begin{bmatrix} f + \Lambda_{\mathbf{i}}^* c G_{\mathbf{i}}^{-1}(\operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{i}} - (Gz)_{\mathbf{i}}) \\ \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{o}} \end{bmatrix}, \\ \lambda_{\mathbf{i}}^+ = c G_{\mathbf{i}}^{-1}((I_{\mathbf{i}} - G_{\mathbf{i}})\Lambda_{\mathbf{i}} u^+ - \operatorname{prox}_{\frac{\phi}{c}}(z)_{\mathbf{i}} + (Gz)_{\mathbf{i}}). \end{cases}$$

5 Conclusion

In this paper, we have developed the classical Lagrange multiplier approach to a class of nonsmooth convex optimization problems arising in various application domains. We presented the Lagrange optimality system, and established the equivalence among the Lagrange optimality system, the standard optimality condition and the saddle point condition of the augmented Lagrangian. The Lagrange optimality system was used to derive a novel Newton-type algorithm. We proved the nonsingularity of the Newton system and established the local convergence of the algorithm.

In order to make the proposed Newton-type algorithm applicable to real word applications, a further study is needed on several important issues including: to construct a merit function for the globalization of the algorithm; to develop efficient solvers for the (possibly) large linear system (Newton update); to provide a stopping criterion, and to report the numerical performance of the algorithm. These issues will be investigated in future work.

Acknowledgements This research was supported by the Aihara Innovative Mathematical Modelling Project, the Japan Society for the Promotion of Science (JSPS) through the "Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program)," initiated by the Council for Science and Technology Policy (CSTP).

References

1. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learning **3**, 1–122 (2011)
2. Chan, T.F., Shen, J.: Image Processing and Analysis. SIAM, Philadelphia, PA (2005)
3. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)
4. Ekeland, I., Témam, R.: Convex Analysis and Variational Problems. SIAM, Philadelphia, PA (1999)
5. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer-Verlag, Berlin (2008)
6. Ito, K., Jin, B.: Inverse Problems: Tikhonov Theory and Algorithms. World Scientific, Singapore (2014)
7. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. SIAM, Philadelphia, PA (2008)
8. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
9. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Optimization (Sympos., Univ. Keele, Keele, 1968), pp. 283–298. Academic Press, London (1969)
10. Rockafellar, R.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Programming **5**, 354–373 (1973)
11. Rockafellar, R.: The multiplier method of Hestenes and Powell applied to convex programming. J. Optim. Theory Appl. **12**, 555–562 (1973)
12. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. ESAIM: Math. Model. Numer. Anal. **9**, 41–76 (1975)
13. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia, PA (1989)
14. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**, 123–231 (2013)
15. Wu, C., Tai, X.C.: Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. SIAM J. Imaging Sci. **3**, 300–339 (2010)
16. Fortin, M.: Minimization of some non-differentiable functionals by the augmented Lagrangian method of Hestenes and Powell. Appl. Math. Optim. **2**, 236–250 (1975)
17. Ito, K., Kunisch, K.: Augmented lagrangian methods for nonsmooth, convex optimization in Hilbert spaces. Nonlin. Anal. Ser. A Theory Methods **41**, 591–616 (2000)
18. Lemaréchal, C., Sagastizábal, C.: Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. SIAM J. Optim. **7**, 367–385 (1997)
19. Ip, C.M., Kyparisis, J.: Local convergence of quasi-Newton methods for B-differentiable equations. Math. Programming **56**, 71–89 (1992)
20. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems. Vol. II. Springer-Verlag, New York (2003)
21. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
22. Ito, K., Kunisch, K.: An active set strategy based on the augmented Lagrangian formulation for image restoration. ESAIM: Math. Model. Numer. Anal. **33**, 1–21 (1999)
23. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. CRC Press, Boca Raton, FL (1992)
24. Patrinos, P., Stella, L., Bemporad, A.: Forward-backward truncated Newton methods for convex composite optimization. preprint, arXiv:1402.6655v2 (2014)
25. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
26. Bergounioux, M., Ito, K., Kunisch, K.: Primal-dual strategy for constrained optimal control problems. SIAM J. Control Optim. **37**, 1176–1194 (1999)