Newtonian Methods with Wolfe Linesearch in Nonsmooth Optimization and Machine Learning

Miantao Chao* Boris S. Mordukhovich† Zijian Shi‡ Jin Zhang§

Abstract

This paper introduces and develops coderivative-based Newton methods with Wolfe linesearch conditions to solve various classes of problems in nonsmooth optimization and machine learning. We first propose a generalized regularized Newton method with Wolfe linesearch (GRNM-W) for unconstrained $C^{1,1}$ minimization problems (which are second-order nonsmooth) and establish global as well as local superlinear convergence of their iterates. The Newton directions in the algorithm are obtained by solving linear equations extracted from coderivatives. To deal with convex composite minimization problems (which are first-order nonsmooth and can be constrained), we combine the proposed GRNM-W with two algorithmic frameworks: the forward-backward envelope and the augmented Lagrangian method resulting in the two new algorithms called CNFB and CNAL, respectively. Finally, we present numerical results to solve Lasso and support vector machine problems appearing in, e.g., machine learning and statistics, which demonstrate the efficiency of the proposed algorithms.

Keywords: nonsmooth optimization, machine learning, variational analysis, nonsmooth Newton methods, Wolfe linesearch, Lasso problems, support vector machines

1 Introduction

The classical Newton method is an efficient second-order algorithm for unconstrained C^2-smooth minimization problems, with superlinear or quadratic local convergence when the Hessian of the objective function is positive-definite around a solution. Nevertheless, it has some serious drawbacks such as lack of global convergence, high computational cost, and restricted applicability. Accordingly, many modifications of Newton's method have been proposed, including damped Newton method, regularized Newton method, quasi-Newton methods, trust-region Newton methods, cubic regularized Newton method, etc. For such Newton-type methods, second-order smoothness of the objective function is required either in the problem formulation, or in the convergence analysis. However, in many natural models, the objective function is not second-order differentiable while we still want to utilize some generalized second-order derivatives to design *nonsmooth* Newton algorithms with the hope that they retain the fast convergence of the classical Newton method. With that in mind, we first focus our attention on the class of unconstrained $C^{1,1}$ minimization problems, where the objective functions are continuously differentiable with locally Lipschitz continuous gradients. In this way, several nonsmooth Newton methods employing different generalized differentiation constructions have been proposed and developed in the literature. The most popular by far generalized Newton method is known as the *semismooth Newton method* (SNM), which primarily addresses, along with its various modifications, to solving Lipschitzian gradient equations that arise, e.g., from stationary conditions for minimizing $C^{1,1}$ functions as well as systems that can be reduced to this framework. The main analytic tools in SNM and its versions are *generalized Jacobians* by Clarke [9]. Among an enormous amount of publications on SNM and its modifications, we refer the reader to the books [13, 22, 28] and the bibliographies

*College of Mathematics and Information Science, Guangxi University, Nanning, 530004, China. E-mail: chaomiantao@126.com. Research of this author was supported by National Science Foundation of China 12061013.

†Department of Mathematics, Wayne State University, Detroit, Michigan, USA. E-mail: aa1086@wayne.edu. Research of this author was partly supported by the US National Science Foundation under grant DMS-2204519, by the Australian Research Council under Discovery Project DP-190100555, and by Project 111 of China under grant D21024.

‡College of Mathematics and Information Science, Guangxi University, Nanning, 530004, China. E-mail: zijianshi@foxmail.com.

§Department of Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen, 518055, China. E-mail: zhangj9@sustech.edu.cn. Research of this author was supported by National Natural Science Foundation of China (12222106, 12326605), Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515020082).

therein for a variety of results, discussions, and historical comments. Due to the well-recognized limitations of SNM discussed in the aforementioned publications, the search of other nonsmooth Newtonian methods has been undertaken over the years. In particular, the semismooth* Newton method [17] and its SCD (subspace containing derivative) [18] variant have been proposed for solving set-valued inclusions. The reader can find more information on recent developments in the fresh book [43].

This paper develops a quite recent direction in variational theory and applications of nonsmooth Newton methods with algorithms constructed by employing *coderivatives* by Mordukhovich [41] instead of generalized Jacobians by Clarke. Newtonian iterations to find stationary points and local minimizers are defined in this way by using coderivative-based *second-order subdifferentials/generalized Hessians* of objective functions in the sense of [39, 41]. The latter constructions of second-order variational analysis enjoy comprehensive calculus rules and admit explicit calculations in terms of the given data for broad classes of nonsmooth functions that overwhelmingly appear in various settings of optimization and its applications to machine learning, data science, statistics, biochemical modeling, etc.; see [43] for more details and references. A variety of coderivative-based Newtonian algorithms have been developed and applied in [1, 25, 26, 27, 43, 45] with establishing their local and global convergence, convergence rates, and applications to practical modeling. To ensure *global convergence* of the generalized damped Newton method in [25, 26, 43] and the generalized regularized Newton method in [27, 43], together with their algorithmic implementations for various classes of optimization problems, the backtracking *Armijo linesearch* has been widely employed.

In this paper, we first propose and justify a generalized regularized Newton method with the *Wolfe linesearch* (GRNM-W) for $C^{1,1}$ minimization problems. Its iterative procedure is given in the form

$$x^{k+1}=x^k+\tau_k d^k \;\text{ with }\; -\nabla\varphi(x^k)\in\partial^2\varphi(x^k)(d^k)+\mu_k B_k d^k,$$

where τ_k is a stepsize determined by a linesearch that satisfies the Wolfe conditions, where $\partial^2\varphi(x^k):=(D^*\nabla\varphi)(x^k)$ is Mordukhovich's generalized Hessian (i.e., coderivative of the gradient mapping), where $\mu_k>0$ is a regularization parameter, and where $B_k\succ 0$ is a regularization matrix. Note that the nonlinear inclusion in the above formula can be implemented by solving linear equation; see Remark 1 for more details. GRNM-W has two *crucial differences* from the generalized regularized Newton method (GRNM) in [27]. On one hand, GRNM-W uses a Wolfe linesearch strategy instead of the backtracking linesearch. On the other hand, GRNM-W incorporates a more general regularization matrix than GRNM, which uses the identity matrix. We establish *global convergence* and *local superlinear convergence rates* of GRNM-W under the same assumptions as in [27]. Moreover, we present an appropriately *modified version* of GRNM-W so that it can be applied to arbitrary nonconvex functions. Global convergence with convergence rates of the modified GRNM-W are established for general nonconvex functions satisfying the *Polyak-Łojasiewicz-Kurdyka* conditions. To the best of our knowledge, the obtained results are new for nonsmooth Newton-type methods. We show that the Wolfe linesearch is *more efficient* than the backtracking Armijo linesearch employed in [27] and other nonsmooth Newton methods mentioned above, especially when evaluations of gradients are not too expensive and the starting point is far enough from a solution. One particularly impressive property of the Wolfe linesearch for Newton-type methods is that it allows us to choose *larger-than-unit stepsizes* in the initial stage of the iteration.

Starting with unconstrained problems of $C^{1,1}$ optimization, we then extend our GRNM-W method to significantly more general classes of *constrained* optimization problems, which may be even *first-order nonsmooth*. Of our primary interest here are problems of *convex composite minimization*, where one of the functions in summation is extended-real-valued and hence incorporates constraints. We combine the coderivative-based GRNM-W with the two algorithmic frameworks: *forward-backward envelope* from [49] and *augmented Lagrangian method* from [54]. These combinations lead us to the two new algorithms called CNFB (coderivative-based Newton forward-backward method) and CNAL (coderivative-based Newton augmented Lagrangian method). The main difference of CNFB from GRNM in [27] is the usage of the Wolfe linesearch instead of the Armijo one, while CNAL is different from SSNAL (semismooth Newton augmented Lagrangian method) in [31] by using coderivatives instead of generalized Jacobians and the Wolfe conditions instead of the Armijo condition. We present applications of our results to *support vector machines* and *Lasso problems* with numerical experiments that demonstrate the efficiency of our algorithms.

The remaining parts of the paper are organized as follows. In Section 2, we review preliminaries from variational analysis including generalized differentiation constructions and the semismooth and semismooth* properties. Section 3 proposes and develops a generalized regularized Newton method with the Wolfe linesearch (GRNM-W) for $C^{1,1}$ functions whose generalized Hessians are positive-semidefinite. Global con-

vergence results with local superlinear convergence rates are derived in this section. A modified version of GRNM-W is given in Section 4 in such a way that it is applicable to arbitrary nonconvex $C^{1,1}$ functions without any positive-semidefiniteness assumptions whatsoever. We then verify global convergence and convergence rates for the modified GRNM-W assuming the PLK property of the objective function. Our study of convex composite minimization problems begins in Section 5, where we propose the CNFB algorithm by combining GRNM-W with the forward-backward envelope. Section 6 develops the CNAL algorithm by embedding GRNM-W in the augmented Lagrangian method. In Section 7, we compare GRNM-W with GRNM numerically. Applications of the obtained results to support vector machines and Lasso problems together with the corresponding numerical experiments are reported in Section 8 and Section 9, respectively. Finally, Section 10 presents concluding remarks and discusses some directions of our future research.

Our notations are standard. Recall that $\mathbb{N} := \{0,1,\ldots\}$, that $\mathbb{R}$ is the field of real numbers, and that $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. The distance between $x \in \mathbb{R}^n$ and a nonempty set $\Omega \subset \mathbb{R}^n$ is defined by $\operatorname{dist}(x,\Omega) := \inf\{\|x-y\| \mid y \in \Omega\}$. The symbol $\mathbb{B}_\delta(\bar{x}) = \{x \in \mathbb{R}^n \mid \|x-\bar{x}\| < \delta\}$ stands for the open ball centered at $\bar{x}$ with radius δ. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the notation $A \succ 0$ means that A is positive-definite. The norm defined by such a matrix A is denoted by $\|x\|_A := \sqrt{\langle x, Ax\rangle}$.

2 Preliminaries from Variational Analysis

Here we overview some well-known notions and results of variational analysis broadly employed in the paper; see the books [41, 42, 43, 55] for more details and related material.

For a nonempty set $\Omega \subset \mathbb{R}^n$, the (Fréchet) *regular normal cone* to Ω at $\bar{x} \in \Omega$ is

$$\widehat{N}_\Omega(\bar{x}) := \Big\{ v \in \mathbb{R}^n \;\Big|\; \limsup_{x \xrightarrow{\Omega} \bar{x}} \frac{\langle v, x-\bar{x}\rangle}{\|x-\bar{x}\|} \le 0 \Big\}, \tag{1}$$

where the symbol $x \xrightarrow{\Omega} \bar{x}$ indicates that $x \to \bar{x}$ with $x \in \Omega$. The (Mordukhovich) *basic/limiting normal cone* to Ω at $\bar{x} \in \Omega$ is defined by

$$N_\Omega(\bar{x}) := \big\{ v \in \mathbb{R}^n \;\big|\; \exists\, x_k \xrightarrow{\Omega} \bar{x}, v_k \to v \;\text{ as }\; k \to \infty \;\text{ with }\; v_k \in \widehat{N}_\Omega(x_k) \big\}. \tag{2}$$

Note that the set $N_\Omega(\bar{x})$ is often nonconvex as, e.g., for $\Omega := \{(x,\alpha) \in \mathbb{R}^2 \mid \alpha \ge -|x|\}$ at $\bar{x} = 0$. Thus (2) cannot be generated in duality by any tangential approximation of Ω at $\bar{x}$ since duality always yields convexity. Nevertheless, the limiting normal cone (2) and the corresponding coderivative and subdifferential constructions enjoy *full calculus* based on *variational/extremal principles*.

Consider a set-valued mapping/multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ with the graph $\operatorname{gph} F := \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in F(x)\}$. Generated by the corresponding normal cone in (1) and (2), the *regular coderivative* and *limiting coderivative* of F at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ are defined, respectively, by

$$\widehat{D}^* F(\bar{x},\bar{y})(v) := \big\{ u \in \mathbb{R}^n \;\big|\; (u,-v) \in \widehat{N}_{\operatorname{gph} F}(\bar{x},\bar{y}) \big\}, \quad v \in \mathbb{R}^m, \tag{3}$$

$$D^* F(\bar{x},\bar{y})(v) := \big\{ u \in \mathbb{R}^n \;\big|\; (u,-v) \in N_{\operatorname{gph} F}(\bar{x},\bar{y}) \big\}, \quad v \in \mathbb{R}^m. \tag{4}$$

When $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is single-valued, we omit $\bar{y} = F(\bar{x})$ in the notations (3) and (4). Recall further that a multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *metrically regular* at $(\bar{x},\bar{y}) \in \operatorname{gph} F$ if there exist a positive constant $\mu > 0$ and neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that

$$\operatorname{dist}\big(x; F^{-1}(y)\big) \le \mu \operatorname{dist}\big(y; F(x)\big) \;\text{ for all }\; x,y) \in U \times V. \tag{5}$$

Given now an extended-real-valued function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ with $\operatorname{dom}\varphi := \{x \in \mathbb{R}^n \mid \varphi(x) < \infty\}$, define the *limiting subdifferential* of φ at $\bar{x} \in \operatorname{dom}\varphi$ geometrically by

$$\partial\varphi(\bar{x}) := \big\{ v \in \mathbb{R}^n \;\big|\; (v,-1) \in N_{\operatorname{epi}\varphi}\big(\bar{x}, \varphi(\bar{x})\big) \big\} \tag{6}$$

while observing that (6) admits various analytic representations that can be found in the aforementioned books. The *second-order subdifferential* (or *generalized Hessian*) of φ at $\bar{x} \in \operatorname{dom}\varphi$ for $\bar{v} \in \partial\varphi(\bar{x})$ is defined in [39] as the coderivative of the subgradient mapping $\partial^2\varphi(\bar{x},\bar{v}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ by

$$\partial^2\varphi(\bar{x},\bar{v})(u) := \big(D^*\partial\varphi\big)(\bar{x},\bar{v})(u) \;\text{ whenever }\; u \in \mathbb{R}^n. \tag{7}$$

Note that if φ is a C^2-smooth function around $\bar{x}$, then (7) reduces to the classical (symmetric) Hessian matrix $\partial^2\varphi(\bar{x})(u)=\{\nabla^2\varphi(\bar{x})(u)\}$ for all $u\in\mathbb{R}^n$. Over the years, extensive calculus rules, explicit calculations, and a variety of applications have been obtained in terms of (7), which have been summarized in the recent book [43] with the numerous references therein.

Among the striking applications of (7), we mention *complete characterizations* in its terms several notions of *variational stability* in finite and infinite dimensions. It concerns, in particular, the following major stability notion in optimization introduced in [50].

Definition 1 (tilt stability). For $\varphi:\mathbb{R}^n\to\overline{\mathbb{R}}$, a point $\bar{x}\in\operatorname{dom}\varphi$ is called a *tilt-stable local minimizer* of φ if there exists a number $\gamma>0$ such that the mapping

$$M_\gamma: v\mapsto \operatorname{argmin}\big\{\varphi(x)-\langle v,x\rangle \;\big|\; x\in\mathbb{B}_\gamma(\bar{x})\big\}$$

is single-valued and Lipschitz continuous on some neighborhood of $0\in\mathbb{R}^n$ with $M_\gamma(0)=\{\bar{x}\}$. A Lipschitz constant of M_γ around 0 is called a *modulus of tilt stability* of φ at $\bar{x}$.

The next notion of semismoothness was introduced in [35] for real-valued functions and then extended to vector-valued functions in [29, 53] with applications to generalized Newton's methods for directionally differentiable Lipschitz continuous functions.

Definition 2 (semismoothness). A mapping $f:\mathbb{R}^n\to\mathbb{R}^m$ is said to be *semismooth* at $\bar{x}$ if it is locally Lipschitz continuous around $\bar{x}$ and the limit

$$\lim_{\substack{A\in\operatorname{co}\overline{\nabla}f(\bar{x}+tu')\\ u'\to u,t\downarrow 0}} Au' \tag{8}$$

exists for all $u\in\mathbb{R}^n$, where 'co' stands for the convex hull of the set in question, and where

$$\overline{\nabla}f(x):=\big\{A\in\mathbb{R}^{m\times n}\;\big|\;\exists\, x_k\overset{\Omega_f}{\to}x \text{ with } \nabla f(x_k)\to A\big\}$$

with Ω_f signifying the set of points at which f is differentiable.

Note that the set $\overline{\nabla}f(x)$ above is also written as $\partial_B f(x)$ and is called the *B-subdifferential* of f at x. The convex hull $\operatorname{co}\overline{\nabla}f(x)$ is also written as $\partial_C f(x)$ and is called the (Clarke) *generalized Jacobian* of f at x. Observe the relationship (with A^T standing for the matrix transposition)

$$\operatorname{co}D^*f(\bar{x})(v)=\big\{A^T\;\big|\;A\in\partial_C f(\bar{x})\big\},\quad v\in\mathbb{R}^m,$$

between the coderivative (4) and the generalized Jacobian of f at $\bar{x}$, valid for any $f\colon\mathbb{R}^n\to\mathbb{R}^m$ locally Lipschitzian around $\bar{x}$. The following characterization of semismoothness is taken from [53, Theorem 2.3].

Proposition 1 (characterization of semismoothness). *Let $f\colon\mathbb{R}^n\to\mathbb{R}^m$ be locally Lipschitzian around $\bar{x}$. Then f is semismooth at $\bar{x}$ if and only if it is directionally differentiable at $\bar{x}$ in every direction and for any $x\to\bar{x}$ and $A\in\partial_C f(\bar{x})$ we have the condition*

$$f(x)-f(\bar{x})-A(x-\bar{x})=o(\|x-\bar{x}\|). \tag{9}$$

To proceed further, recall that the *directional normal cone* to a set $\Omega\subset\mathbb{R}^s$ at $\bar{z}\in\Omega$ in the direction $d\in\mathbb{R}^s$ is introduced in [19] by

$$N_\Omega(\bar{z};d):=\big\{v\in\mathbb{R}^s\;\big|\;\exists\, t_k\downarrow 0,\ d_k\to d,\ v_k\to v \text{ with } v_k\in\widehat{N}_\Omega(\bar{z}+t_kd_k)\big\}.$$

The *directional coderivative* of $F:\mathbb{R}^n\rightrightarrows\mathbb{R}^m$ at $(\bar{x},\bar{y})\in\operatorname{gph}F$ in the direction $(u,v)\in\mathbb{R}^n\times\mathbb{R}^m$ is

$$D^*F\big((\bar{x},\bar{y});(u,v)\big)(q):=\big\{p\in\mathbb{R}^n\;\big|\;(p,-q)\in N_{\operatorname{gph}F}\big((\bar{x},\bar{y});(u,v)\big)\big)\big\},\quad q\in\mathbb{R}^m, \tag{10}$$

as defined in [16]. Using (10), the semismoothness was extended in [17] to set-valued mappings $F\colon\mathbb{R}^n\rightrightarrows\mathbb{R}^m$ as follows: F is said to be *semismooth** at $(\bar{x},\bar{y})\in\operatorname{gph}F$ if for all $(u,v)\in\mathbb{R}^n\times\mathbb{R}^m$ we have the condition

$$\langle p,u\rangle=\langle q,v\rangle \;\text{ whenever }\; (q,p)\in\operatorname{gph}D^*F\big((\bar{x},\bar{y});(u,v)\big).$$

We refer the reader to [17] and [43, Section 9.1.2] for various properties of semismooth* mappings. In particular, if $F=f\colon\mathbb{R}^n\to\mathbb{R}^m$ is single-valued and locally Lipschitzian around $\bar{x}$, then its semismooth* property at $\bar{x}$ is equivalent to condition (9) without assuming the directional differentiability of f at $\bar{x}$.

In the algorithms developed in this paper, we achieve several convergence rates defined, e.g., in [13]. The following technical lemma from [27, Lemma 4] is useful below.

Lemma 2 (linear convergence of sequences). *Let $\{\alpha_k\},\{\beta_k\},\{\gamma_k\}$ be sequences of positive real numbers. Assume that there exist constants $c_1,c_2,c_3>0$ satisfying the estimates for large $k\in\mathbb{N}$:*

$$\alpha_k-\alpha_{k+1}\le c_1\beta_k^2,\quad \beta_k\le c_2\gamma_k,\quad c_3\gamma_k^2\le\alpha_k.$$

Then $\{\alpha_k\}$ converges to zero Q-linearly while $\{\beta_k\}$ and $\{\gamma_k\}$ converge to zero R-linearly.

3 Regularized Newtonian Method with Wolfe Linesearch

Consider the following unconstrained optimization problem:

$$\text{minimize}\ \ \varphi(x)\ \ \text{subject to}\ \ x\in\mathbb{R}^n, \tag{11}$$

where φ is of *class* $C^{1,1}$, i.e., it is continuously differentiable with the locally Lipschitzian gradient. We propose and justify the *generalized regularized Newton method* with the *Wolfe linesearch* (GRNM-W) for solving the second-order nonsmooth problem (11). Here is the algorithm:

Algorithm 1 Generalized regularized Newton method with the Wolfe linesearch (GRNM-W)

Input: $x^0\in\mathbb{R}^n$, $c>0$, $0<\sigma_1<\sigma_2<1$, $\rho\in(0,1]$.

1: **for** $k=0,1,\ldots$ **do**

2: If $\nabla\varphi(x^k)=0$, stop; otherwise set $\mu_k=c\|\nabla\varphi(x^k)\|^\rho$ and go to next step.

3: Choose $B_k\succ 0$. Find $d^k\in\mathbb{R}^n$ such that $-\nabla\varphi(x^k)\in\partial^2\varphi(x^k)(d^k)+\mu_kB_kd^k$.

4: Set $\tau_k=1$ and check the Wolfe conditions:

$$\varphi(x^k+\tau_kd^k)\le\varphi(x^k)+\sigma_1\tau_k\langle\nabla\varphi(x^k),d^k\rangle,$$

$$\langle\nabla\varphi(x^k+\tau_kd^k),d^k\rangle\ge\sigma_2\langle\nabla\varphi(x^k),d^k\rangle.$$

If these conditions do not hold, adjust τ_k (using any specific implementation of the Wolfe linesearch) until it satisfies the Wolfe conditions. We assume that in the implementation there exists an upper bound $\tau_{\max}$ on the maximum stepsize allowed.

5: Set $x^{k+1}=x^k+\tau_kd^k$.

6: **end for**

The proposed algorithm is a counterpart of the globally convergent coderivative-based GRNM from [27] with replacing the Armijo linesearch by the Wolfe one. The reader can consult [11, 52] and the references therein for some other versions of globally convergent regularized Newton methods in the case of convex C^2-smooth objective functions using the Armijo linesearch for globalization.

Remark 1 (implementation of Algorithm 1). Let us address the issue of implementing the inclusion

$$-\nabla\varphi(x^k)\in\partial^2\varphi(x^k)(d^k)+\mu_kB_kd^k \tag{12}$$

in line 3 of Algorithm 1. The generalized Hessian $\partial^2\varphi(x)(d):=D^*\nabla\varphi(x)(d)$ is nonlinear in d in general, but we can always extract a linear mapping from it as follows. Let $\mathscr{S}^*\nabla\varphi(x)$ be the SC limiting coderivative defined in [18]. By [18, Lemma 3.7, Lemma 3.11], we can always pick a linear subspace $L_x\in\mathscr{S}^*\nabla\varphi(x)$ represented by a matrix $A_x\in\mathbb{R}^{n\times n}$ such that $A_xd\subset D^*\nabla\varphi(x)(d)$ for all $d\in\mathbb{R}^n$. Therefore, it is possible to solve the nonlinear inclusion (12) by solving the linear equation in d^k:

$$-\nabla\varphi(x^k)=(A_k+\mu_kB_k)d^k, \tag{13}$$

where A_k is chosen from the SC limiting coderivative $\mathscr{S}^*\nabla\varphi(x^k)$, which is always nonempty and is contained in the limiting coderivative $D^*\nabla\varphi(x^k)=:\partial^2\varphi(x^k)$. Note that once the coderivative is calculated, it is usually easy to extract such a linear mapping.

To proceed with the justification of Algorithm 1, we first present the following lemma.

Lemma 3 (existence of Newton-Wolfe directions). *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be of class $C^{1,1}$ around a point $x \in \mathbb{R}^n$ such that $\nabla\varphi(x) \neq 0$ and $\partial^2\varphi(x)$ is positive-semidefinite, i.e.,*

$$\langle z, u\rangle \geq 0 \text{ for all } u \in \mathbb{R}^n \text{ and all } z \in \partial^2\varphi(x)(u). \tag{14}$$

Then for any positive-definite symmetric matrix $B \in \mathbb{R}^{n\times n}$, there exists $d \neq 0$ such that

$$-\nabla\varphi(x) \in \partial^2\varphi(x)(d) + Bd.$$

Proof. This can be distilled from the similar proof of [27, Theorem 3(i)], and hence is omitted. □

Now we get the well-posedness of the proposed CRNM-W.

Theorem 4 (well-posedness of Algorithm 1). *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be of class $C^{1,1}$ and be bounded from below, and let $x^0 \in \mathbb{R}^n$ be a starting point such that for all $x \in \mathbb{R}^n$ with $\varphi(x) \leq \varphi(x^0)$ the generalized Hessian $\partial^2\varphi(x)$ is positive-semidefinite.*[1] *If $\nabla\varphi(x^k) \neq 0$, then there exists $d^k \neq 0$ such that the condition in line 3 of Algorithm 1 holds and the Wolfe linesearch is well defined. Moreover, Algorithm 1 either stops after finitely many iterations at a stationary point, or generates a sequence $\{x^k\}$ such that $\{\varphi(x^k)\}$ is decreasing.*

Proof. The existence of d^k follows from Lemma 3. Since $\partial^2\varphi(x^k)$ is positive-semidefinite (by the sufficient decrease condition and the assumption on x^0), and since $-\nabla\varphi(x^k) - \mu_k B_k d^k \in \partial^2\varphi(x^k)(d^k)$, we get $\langle -\nabla\varphi(x^k) - \mu_k B_k d^k, d^k\rangle \geq 0$ giving us the estimate

$$\langle -\nabla\varphi(x^k), d^k\rangle \geq \mu_k \|d^k\|^2_{B_k}. \tag{15}$$

Using $\mu^k > 0$ and $d^k \neq 0$ ensures that $\langle -\nabla\varphi(x^k), d^k\rangle > 0$. This tells us that d^k is a descent direction. By [47, Lemma 3.1], a stepsize satisfying the Wolfe conditions always exists when the objective function is C^1-smooth and bounded from below. By design, Algorithm 1 either stops at a stationary point after finitely many iterations, or generates a sequence of iterates $\{x^k\}$ such that $\varphi(x^{k+1}) < \varphi(x^k)$ for all $k \in \mathbb{N}$. □

The next theorem establishes stationarity of accumulation points of iterates in GRNM-W.

Theorem 5 (stationarity of accumulation points in Algorithm 1). *Under the assumptions of Theorem 4, suppose that the eigenvalues of the regularization matrices B_k in Algorithm 1 are contained in $[m^2_{\text{lower}}, m^2_{\text{upper}}]$ with $m_{\text{upper}} \geq m_{\text{lower}} > 0$, which implies that $m_{\text{lower}}\|x\| \leq \|x\|_{B_k} \leq m_{\text{upper}}\|x\|$ for all $x \in \mathbb{R}^n$. Then every accumulation point of $\{x^k\}$ is stationary for problem (11).*

Proof. It follows from Theorem 4 that Algorithm 1 either stops after finitely many iterations (in which case it must land at a stationary point), or generates a sequence of iterates $\{x^k\}$ such that $\varphi(x^{k+1}) < \varphi(x^k)$ for all $k \in \mathbb{N}$. This tells us that $\{x^k\} \subset \Omega := \{x \in \mathbb{R}^n \mid \varphi(x) \leq \varphi(x^0)\}$. Suppose that the algorithm does not stop after finitely many steps. Then we have $\nabla\varphi(x^k) \neq 0$ for all $k \in \mathbb{N}$. Recall that $\mu_k := c\|\nabla\varphi(x^k)\|^\rho$ with $c > 0$ and $\rho \in (0, 1]$. Define the modified directions

$$\tilde{d}^k := \|\nabla\varphi(x^k)\|^{\rho-1} d^k, \quad k \in \mathbb{N}, \tag{16}$$

which is possible due to condition $\|\nabla\varphi(x^k)\| > 0$. Let us show that the sequence $\{\tilde{d}^k\}$ is bounded. Indeed, the algorithm design (line 3 of Algorithm 1) provides

$$-\nabla\varphi(x^k) - \mu_k B_k d^k \in \partial^2\varphi(x^k)(d^k) \text{ for all } k \in \mathbb{B},$$

which ensures by the positive-semidefiniteness of $\partial^2\varphi(x^k)$ that

$$\langle -\nabla\varphi(x^k), d^k\rangle \geq \mu_k \|d^k\|^2_{B_k} \geq \mu_k m^2_{\text{lower}}\|d^k\|^2. \tag{17}$$

By the Cauchy-Schwarz inequality with the choice of the regularization parameter $\mu_k = c\|\nabla\varphi(x^k)\|^\rho$ and the regularization matrices B_k, we have

$$c\|\nabla\varphi(x^k)\|^\rho m^2_{\text{lower}}\|d^k\|^2 \leq c\|\nabla\varphi(x^k)\|^\rho \|d^k\|^2_{B_k} \leq \|\nabla\varphi(x^k)\|\|d^k\|,$$

[1] This condition always holds when φ is convex; see [7].

which brings us to the estimate

$$\|\nabla\varphi(x^k)\|^{\rho-1}\|d^k\| \le \frac{1}{m_{\text{lower}}^2 c}, \quad k\in\mathbb{N}.$$

This justifies that the sequence of $\|\tilde{d}^k\| = \|\nabla\varphi(x^k)\|^{\rho-1}\|d^k\|$ is bounded.

Let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ such that $x^{k_j}\to\bar{x}$ as $j\to\infty$, i.e., $\bar{x}$ is an accumulation point of $\{x^k\}$. Since $\varphi(\bar{x})$ is clearly an accumulation point of the decreasing sequence $\{\varphi(x^k)\}$, we have $\varphi(x^k)\to\varphi(\bar{x})$ as $k\to\infty$. It follows from the Wolfe conditions

$$\varphi(x^{k+1})-\varphi(x^k)\le\sigma_1\tau_k\langle\nabla\varphi(x^k),d^k\rangle<0$$

that the passage to the limit leads us to

$$\lim_{k\to\infty}\tau_k\langle\nabla\varphi(x^k),d^k\rangle=0. \tag{18}$$

The boundedness of $\{\tilde{d}^k\}$ provides a convergent subsequence, which can be taken as $\{\tilde{d}^{k_j}\}$ without loss of generality. We get that $\{d^{k_j}\}$ is convergent with limit $\bar{d}:=\lim_{j\to\infty}d^{k_j}=\lim_{j\to\infty}\|\nabla\varphi(x^{k_j})\|^{1-\rho}\tilde{d}^{k_j}$ by the continuity of the function $\|\cdot\|^{1-\rho}$ on $\mathbb{R}$, which is equivalent to $\rho\le 1$. Next we claim that

$$\langle\nabla\varphi(\bar{x}),\bar{d}\rangle=0. \tag{19}$$

If $\limsup_{j\to\infty}\tau_{k_j}>0$, then (19) follows from (18). Hence we only need to consider the case where $\limsup_{j\to\infty}\tau_{k_j}=0$, which is the same as $\lim_{j\to\infty}\tau_{k_j}=0$ (since $0\le\liminf_{j\to\infty}\tau_{k_j}\le\limsup_{j\to\infty}\tau_{k_j}=0$). Note that

$$\langle\nabla\varphi(x^{k_j}+\tau_{k_j}d^{k_j}),d^{k_j}\rangle\ge\sigma_2\langle\nabla\varphi(x^{k_j}),d^{k_j}\rangle \text{ for all } j\in\mathbb{N}.$$

Letting $j\to\infty$ yields $\langle\nabla\varphi(\bar{x}),\bar{d}\rangle\ge 0$ by $\sigma_2<1$. On the other hand, taking limits in $\langle\nabla\varphi(x^{k_j}),d^{k_j}\rangle<0$ ensures that $\langle\nabla\varphi(\bar{x}),\bar{d}\rangle\le 0$ and justifies (19) in this case. Combining further (19) with (17), we have

$$m_{\text{lower}}^2 c\|\nabla\varphi(\bar{x})\|^\rho\|\bar{d}\|^2=m_{\text{lower}}^2\lim_{j\to\infty}\mu_{k_j}\|d^{k_j}\|^2=0. \tag{20}$$

If $\|\bar{d}\|\ne 0$, then $\|\nabla\varphi(\bar{x})\|^\rho=0$ by (20). Since $\rho>0$, we get $\|\nabla\varphi(\bar{x})\|=0$. If $\|\bar{d}\|=0$, we can still show that $\|\nabla\varphi(\bar{x})\|=0$. Indeed, recall that φ is of class $C^{1,1}$ around $\bar{x}$ and $-\nabla\varphi(x^k)-\mu_k B_k d^k\in\partial^2\varphi(x^k)(d^k)$. Then it follows from [41, Theorem 1.44] that there is $l>0$ with

$$\left\|\nabla\varphi(x^{k_j})+\mu_{k_j}B_{k_j}d^{k_j}\right\|\le l\left\|d^{k_j}\right\|$$

for all large j. Letting $j\to\infty$ verifies $\|\nabla\varphi(\bar{x})\|\le l\|\bar{d}\|=0$ and thus completes the proof. □

Theorem 6 (convergence and convergence rates of GRNM-W). *In the setting of Theorem 5, let $\bar{x}$ be an accumulation point of $\{x^k\}$ such that $\nabla\varphi$ is metrically regular around $\bar{x}$. Then $\bar{x}$ is a tilt-stable local minimizer of φ, and Algorithm 1 converges to $\bar{x}$ with the convergence rates as follows:*

(i) *The sequence $\{\varphi(x^k)\}$ converges Q-linearly to $\varphi(\bar{x})$.*

(ii) *The sequences $\{x^k\}$ and $\{\nabla\varphi(x^k)\}$ converge R-linearly to $\bar{x}$ and 0, respectively.*

(iii) *The convergence rates of $\{x^k\}$, $\{\varphi(x^k)\}$, and $\{\nabla\varphi(x^k)\}$ are Q-superlinear if $\nabla\varphi$ is semismooth* at $\bar{x}$ and one of the following two groups of conditions holds:*

(a) *$\nabla\varphi$ is directionally differentiable at $\bar{x}$ and $\sigma_1\in(0,\frac{1}{2})$.*

(b) *$\sigma_1\in(0,\frac{1}{2l\kappa})$ and $\sigma_2\in(1-\frac{\kappa}{l},1)$, where $\kappa>0$ and $l>0$ are moduli of the metric regularity and Lipschitz continuity of $\nabla\varphi$ around $\bar{x}$, respectively.*

Proof. We split the proof into seven claims. Note that the proof under the wolfe conditions is rather different and more involved that the proof of [27, Theorem 4] in the Armijo case. In particular, we have to establish that the unit stepsize satisfies the Wolfe conditions, which are more stringent than the Armijo one.

Claim 1: *$\bar{x}$ is a tilt-stable local minimizer of φ.* By Theorem 5, $\bar{x}$ is a stationary point of φ being such that $\varphi(\bar{x})\le\varphi(x^0)$. The positive-semidefiniteness of $\partial^2\varphi(\bar{x})$ and the metric regularity of $\nabla\varphi$ around $\bar{x}$ with

modulus κ imply by [12, Theorem 4.13] that $\bar{x}$ is a tilt-stable local minimizer of φ with the same modulus.

Claim 2: *Let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ with $x^{k_j}\to\bar{x}$ as $j\to\infty$. Then the subsequence of stepsizes τ_{k_j} in Algorithm 1 is bounded from below by a positive number γ and satisfies*

$$\varphi(x^{k_j})-\varphi(x^{k_j+1})\geq\frac{\sigma_1\gamma}{\kappa}\|d^{k_j}\|^2 \text{ for all large } j\in\mathbb{N}.$$

To verify this, suppose on the contrary that $\{\tau_{k_j}\}$ is not bounded from below by a positive number, and so there exists a subsequence of $\{\tau_{k_j}\}$ that converges to 0. We can assume without loss of generality that $\tau_{k_j}\to 0$ as $j\to\infty$. By the second-order characterization of tilt-stable minimizers from [44, Theorem 3.5] and [8, Proposition 4.6], there exists $\delta>0$ such that

$$\langle z,w\rangle\geq\frac{1}{\kappa}\|w\|^2 \text{ for all } z\in\partial^2\varphi(x)(w),\ x\in\mathbb{B}_\delta(\bar{x}), \text{ and } w\in\mathbb{R}^n. \tag{21}$$

Since $-\nabla\varphi(x^{k_j})-\mu_{k_j}B_{k_j}d^{k_j}\in\partial^2\varphi(x^{k_j})(d^{k_j})$, it follows that

$$\langle-\nabla\varphi(x^{k_j}),d^{k_j}\rangle\geq\left(\mu_{k_j}m^2_{\text{lower}}+\frac{1}{\kappa}\right)\|d^{k_j}\|^2\geq\frac{1}{\kappa}\|d^{k_j}\|^2 \text{ for large } j\in\mathbb{N}. \tag{22}$$

Combining the Cauchy-Schwarz inequality, the Lipschitz continuity of $\nabla\varphi$ around $\bar{x}$ with Lipschitz constant l, and the Wolfe conditions yields

$$l\tau_{k_j}\|d^{k_j}\|^2\geq\langle\nabla\varphi(x_{k_j}+\tau_{k_j}d^{k_j})-\nabla\varphi(x^{k_j}),d^{k_j}\rangle\geq-(1-\sigma_2)\langle\nabla\varphi(x^{k_j}),d^{k_j}\rangle.$$

This tells us together with the estimates in (22) that

$$\tau_{k_j}\geq\frac{(1-\sigma_2)\langle-\nabla\varphi(x^{k_j}),d^{k_j}\rangle}{l\|d^{k_j}\|^2}\geq\frac{1-\sigma_2}{l\kappa}>0,$$

which shows in turn that the subsequence $\{\tau_{k_j}\}$ is bounded from below by $\gamma:=\frac{1-\sigma_2}{l\kappa}$. Using finally the Wolfe conditions and (22) justifies the claimed assertion by

$$\varphi(x^{k_j})-\varphi(x^{k_j+1})\geq\sigma_1\tau_{k_j}\langle-\nabla\varphi(x^{k_j}),d^{k_j}\rangle\geq\frac{\sigma_1\gamma}{\kappa}\|d^{k_j}\|^2 \text{ for all large } j.$$

Claim 3: *The sequence $\{x^k\}$ is convergent.* We show this by applying the convergence criterion based on Ostrowski's condition [13, Proposition 8.3.10]. Let us first check that $\bar{x}$ is an isolated accumulation point of $\{x^k\}$. Indeed, if $\tilde{x}\in\mathbb{B}_\delta(\bar{x})$ is an accumulation point of $\{x^k\}$, we get by Theorem 5 that $\tilde{x}$ is a stationary point of φ. It follows from (21) and the second-order characterization of strong convexity for $C^{1,1}$ functions in [7, Theorem 5.2(i)] that φ is strongly convex with modulus κ^{-1} on $\mathbb{B}_\delta(\bar{x})$, which ensures that $\tilde{x}=\bar{x}$. To verify further Ostrowski's condition, let $\{x^{k_j}\}$ be a subsequence of $\{x^k\}$ that converges to $\bar{x}$. We need to show that $\lim_{j\to\infty}\|x^{k_j+1}-x^{k_j}\|=0$. To see this, deduce from Claim 2 that

$$\left\|x^{k_j+1}-x^{k_j}\right\|^2=\tau^2_{k_j}\|d^{k_j}\|^2\leq\tau_{\max}\|d^{k_j}\|^2\leq\frac{\tau_{\max}\kappa}{\sigma_1\gamma}\left(\varphi(x^{k_j})-\varphi(x^{k_j+1})\right)\to 0$$

as $j\to\infty$. Applying finally [13, Proposition 8.3.10] yields the convergence of $\{x^k\}$ to $\bar{x}$ as $k\to\infty$.

Claim 4: *$\{\varphi(x^k)\}$ converges at least Q-linearly, while $\{x^k\}$ and $\{\nabla\varphi(x^k)\}$ converge at least R-linearly.* We use the strong convexity of φ with modulus κ^{-1} on $\mathbb{B}_\delta(\bar{x})$ to get the estimates

$$\varphi(x)\geq\varphi(u)+\langle\nabla\varphi(u),x-u\rangle+\frac{1}{2\kappa}\|x-u\|^2, \tag{23}$$

$$\langle\nabla\varphi(x)-\nabla\varphi(u),x-u\rangle\geq\frac{1}{\kappa}\|x-u\|^2 \tag{24}$$

for all $x,u\in\mathbb{B}_\delta(\bar{x})$. Since $x^k\to\bar{x}$, this shows that $x^k\in\mathbb{B}_\delta(\bar{x})$ for all k sufficiently large. Letting $x=x^k,u=\bar{x}$ and using the Cauchy-Schwarz inequality lead us to the conditions

$$\varphi(x^k)\geq\varphi(\bar{x})+\frac{1}{2\kappa}\|x^k-\bar{x}\|^2,\quad\|\nabla\varphi(x^k)\|\geq\frac{1}{\kappa}\|x^k-\bar{x}\|. \tag{25}$$

Due to [22, Lemma A.11], the Lipschitz continuity of $\nabla\varphi$ around $\bar{x}$ with modulus $l>0$ implies that

$$\varphi(x^k)-\varphi(\bar{x})=\left|\varphi(x^k)-\varphi(\bar{x})-\langle\nabla\varphi(\bar{x}),x^k-\bar{x}\rangle\right|\le\frac{l}{2}\|x^k-\bar{x}\|^2. \tag{26}$$

By $-\nabla\varphi(x^k)-\mu_k B_k d^k\in\partial^2\varphi(x^k)(d^k)$, it follows from [41, Theorem 1.44] that

$$\left\|\nabla\varphi(x^k)+\mu_k B_k d^k\right\|\le l\|d^k\|. \tag{27}$$

Since $x^k\to\bar{x}$ and $\nabla\varphi(\bar{x})=0$, we have $\mu_k=c\|\nabla\varphi(x^k)\|^\rho\to 0$ as $k\to\infty$, and so $\mu_k\le l$ when k is large. Thus $\|\nabla\varphi(x^k)\|\le\|\nabla\varphi(x^k)+\mu_k d^k\|+\mu_k\|d^k\|\le 2l\|d^k\|$. Together with Claim 2, this yields

$$\varphi(x^k)-\varphi(x^{k+1})\ge\frac{\sigma_1\gamma}{\kappa}\|d^k\|^2\ge\frac{\sigma_1\gamma}{4\kappa l^2}\|\nabla\varphi(x^k)\|^2. \tag{28}$$

Combining the estimates in (25)–(28) and applying Lemma 2 with the choices of $\alpha_k=\varphi(x^k)-\varphi(\bar{x})$, $\beta_k=\|\nabla\varphi(x^k)\|$, $\gamma_k=\|x^k-\bar{x}\|$, $c_1=\frac{\sigma\gamma}{4\kappa l^2}$, $c_2=\frac{1}{\kappa}$, and $c_3=\frac{l}{2}$ justifies this claim.

Claim 5: *If $\nabla\varphi$ is semismooth* at $\bar{x}$, then $\|x^k+d^k-\bar{x}\|=o(\|x^k-\bar{x}\|)$ as $k\to\infty$.* To verify this, deduce from the subadditivity property of coderivatives in [26, Lemma 5.6] that

$$\partial^2\varphi(x^k)(d^k)\subset\partial^2\varphi(x^k)(x^k+d^k-\bar{x})+\partial^2\varphi(x^k)(-x^k+\bar{x}).$$

On the other hand, by $-\nabla\varphi(x^k)-\mu_k B_k d^k\in\partial^2\varphi(x^k)(d^k)$, there is $v^k\in\partial^2\varphi(x^k)(-x^k+\bar{x})$ with

$$-\nabla\varphi(x^k)-\mu_k B_k d^k-v^k\in\partial^2\varphi(x^k)(x^k+d^k-\bar{x}).$$

It follows from (21) and the Cauchy-Schwarz inequality that

$$\|x^k+d^k-\bar{x}\|\le\kappa\left\|\nabla\varphi(x^k)+v^k+\mu_k B_k d^k\right\|\le\kappa\left(\|\nabla\varphi(x^k)-\nabla\varphi(\bar{x})+v^k\|+\mu_k\|B_k d^k\|\right). \tag{29}$$

Combining $-\nabla\varphi(x^k)-\mu_k B_k d^k\in\partial^2\varphi(x^k)(d^k)$ and (21) gives us

$$\langle-\nabla\varphi(x^k),d^k\rangle\ge(\kappa^{-1}+\mu_k m_{\text{lower}}^2)\|d^k\|^2\ge\kappa^{-1}\|d^k\|^2. \tag{30}$$

Using the Cauchy-Schwarz inequality again together with the Lipschitz continuity of $\nabla\varphi$, we have

$$\|d^k\|\le\kappa\|\nabla\varphi(x^k)\|=\kappa\|\nabla\varphi(x^k)-\nabla\varphi(\bar{x})\|\le\kappa l\|x^k-\bar{x}\|. \tag{31}$$

Moreover, the Lipschitz continuity of $\nabla\varphi$ on $\mathbb{B}_\delta(\bar{x})$ guarantees that

$$\mu_k=c\|\nabla\varphi(x^k)\|^\rho=c\|\nabla\varphi(x^k)-\nabla\varphi(\bar{x})\|^\rho\le c\,l^\rho\|x^k-\bar{x}\|^\rho. \tag{32}$$

Employing now the semismooth* property of of $\nabla\varphi$ at $\bar{x}$ and the inclusion $v^k\in\partial^2\varphi(x^k)(-x^k+\bar{x})$ allows us to deduce from [25, Lemma 5.2] that

$$\|\nabla\varphi(x^k)+v^k\|=\|\nabla\varphi(x^k)-\nabla\varphi(\bar{x})+v^k\|=o(\|x^k-\bar{x}\|). \tag{33}$$

Combining finally the estimates in (29) and (31)–(33) together with $\rho\in(0,1]$ tells us that

$$\begin{aligned}\|x^k+d^k-\bar{x}\|\le&\kappa\left(\|\nabla\varphi(x^k)-\nabla\varphi(\bar{x})+v^k\|+\mu_k\|d^k\|\right)\\ \le&\kappa\,o(\|x^k-\bar{x}\|)+\kappa c\,l^\rho\|x^k-\bar{x}\|^\rho\,\kappa l\|x^k-\bar{x}\|\\ =&o(\|x^k-\bar{x}\|)+O(\|x^k-\bar{x}\|^{1+\rho})\\ =&o(\|x^k-\bar{x}\|)\ \text{ as }\ k\to\infty,\end{aligned} \tag{34}$$

which justifies the claimed convergence rate.

Claim 6: *We have $\tau_k=1$ for all k sufficiently large if $\nabla\varphi$ is semismooth* at $\bar{x}$, and if either condition* (a) *or condition* (b) *of this theorem holds.* Observe that in case (a), the directional differentiability and

semismoothness* of $\nabla\varphi$ at $\bar{x}$ ensure by [17, Corollary 3.8] that the gradient mapping $\nabla\varphi$ is semismooth at $\bar{x}$. From (30), we conclude by [13, Proposition 8.3.18] that the Newton direction satisfies the sufficient decrease condition. Let us check that it also satisfies the Wolfe curvature condition therein. To furnish this, denote $\partial_C^2\varphi := \operatorname{co}\overline{\nabla}(\nabla\varphi)$ and deduce from the semismoothness of $\nabla\varphi$ at $\bar{x}$ that

$$\nabla\varphi(x^k+d^k)=\nabla\varphi(\bar{x})+H_k(x^k+d^k-\bar{x})+o(\|x^k+d^k-\bar{x}\|),$$

where $H^k\in\partial_C^2\varphi(x^k+d^k)$. By (34), we have $\|x^k+d^k-\bar{x}\|=o(\|x^k-\bar{x}\|)$. It follows from [9, Proposition 2.6.2(d)] that the matrix sequence $\{H_k\}$ is bounded, and thus $\nabla\varphi(x^k+d^k)=o(\|x^k-\bar{x}\|)$. Employing now [13, Lemma 7.5.7] yields $\lim\limits_{k\to\infty}\frac{\|x^k-\bar{x}\|}{\|d^k\|}=1$, and hence

$$\nabla\varphi(x^k+d^k)=o(\|d^k\|). \tag{35}$$

It follows therefore that for large $k\in\mathbb{N}$, we get the relationships

$$\begin{aligned}\langle\nabla\varphi(x^k+d^k),d^k\rangle-\sigma_2\langle\nabla\varphi(x^k),d^k\rangle&=-\sigma_2\langle\nabla\varphi(x^k),d^k\rangle+o(\|d^k\|^2)\\&\geq\frac{\sigma_2}{\kappa}\|d^k\|^2+o(\|d^k\|^2)>0,\end{aligned}$$

where the equality holds by (35) and the inequality is valid due to (30).

In case (b), we deduce from estimate (23) that the Newton direction satisfies the sufficient decrease condition by [27, Lemma 1]. To check that the Wolfe curvature condition also holds, observe by $\sigma_2>1-\frac{1}{l\kappa}$ and $\lim\limits_{k\to\infty}\frac{\|x^k-\bar{x}\|}{\|d^k\|}=1$ (due to [13, Lemma 7.5.7]) that

$$\begin{aligned}&\langle\nabla\varphi(x^k+d^k),-d^k\rangle-\sigma_2\langle\nabla\varphi(x^k),-d^k\rangle\\&\leq\langle\nabla\varphi(x^k),-d^k\rangle-\frac{1}{\kappa}\|d^k\|^2-\sigma_2\langle\nabla\varphi(x^k),-d^k\rangle\\&\leq\|d^k\|^2\Big(-\frac{1}{\kappa}+(1-\sigma_2)l\frac{\|x^k-\bar{x}\|}{\|d^k\|}\Big)<0,\end{aligned}$$

where the first inequality follows from (24). Thus the claim is justified.

Claim 7: *The assertions about superlinear convergence hold.* To verify this, we have by Claim 6 that $\tau_k=1$ for large k, and hence

$$\|x^{k+1}-\bar{x}\|=\|x^k+\tau_kd^k-\bar{x}\|=\|x^k+d^k-\bar{x}\|=o(\|x^k-\bar{x}\|)\ \text{ as }\ k\to\infty.$$

Then the Q-superlinear convergence of $\{\varphi(x^k)\}$ follows from estimates (25) and (26), while the Q-superlinear convergence of $\{\nabla\varphi(x^k)\}$ follows from estimate (25) and the Lipschitz continuity of $\nabla\varphi$. Combining Claims 1–7, we thus validate all the assertions of the theorem. □

4 Modified GRNM-W under PLK Conditions

A characteristic feature of GRNM-W (the same as for its "Armijo" predecessor in [27]) is the *positive-semidefiniteness* assumption on the generalized Hessian required for its well-posedness (i.e., the existence of iterates satisfying the algorithmic procedures); see Theorem 4. In this section, we propose a *modification* of GRNM-W, labeled as *GRNM-WM*, which is well posed for general nonconvex functions of class $C^{1,1}$ without *any requirements* on the generalized Hessian. Then we establish global convergence with explicit convergence rates for GRNM-WM under the fulfillment of the corresponding Polyak-Łojasiewicz-Kurdyka (PLK) conditions; see Definition 3 below and the discussions around it.

Here is the proposed algorithm for *general* $C^{1,1}$ *functions*.

Algorithm 2 Modified regularized Newton method with Wolfe linesearch (GRNM-WM)

Input: $x^0 \in \mathbb{R}^n$, $c > 0$, $0 < \sigma_1 < \sigma_2 < 1$, $M > m > 0$.

1: **for** $k = 0, 1, \ldots$ **do**

2: If $\nabla\varphi(x^k) = 0$, stop; otherwise go to next step.

3: Choose $\mu_k \geq 0$ in such a way that there exists $d^k \in \mathbb{R}^n$ with

$$-\nabla\varphi(x^k) \in \partial^2\varphi(x^k)(d^k) + \mu_k d^k,$$

$$m\|d^k\|^2 \leq \langle -\nabla\varphi(x^k), d^k\rangle, \quad \|\nabla\varphi(x^k)\| \leq M\|d^k\|.$$

4: Set $\tau_k = 1$ and check the Wolfe conditions:

$$\varphi(x^k + \tau_k d^k) \leq \varphi(x^k) + \sigma_1\tau_k\langle\nabla\varphi(x^k), d^k\rangle,$$

$$\langle\nabla\varphi(x^k + \tau_k d^k), d^k\rangle \geq \sigma_2\langle\nabla\varphi(x^k), d^k\rangle.$$

If the conditions do not hold, adjust τ_k until it satisfies the Wolfe conditions. We assume that in the implementation there exists an upper bound $\tau_{\max}$ on the allowed maximum stepsize.

5: Set $x^{k+1} = x^k + \tau_k d^k$.

6: **end for**

We first show that Algorithm 2 is *well posed* in the general nonconvex setting.

Theorem 7 (well-posedness of Algorithm 2). *Suppose that the set $\Omega = \{x \in \mathbb{R}^n \mid \varphi(x) \leq \varphi(x^0)\}$ is bounded. Then for any $k \in \mathbb{N}$, there exists μ_k such that the inclusion in line 3 of* Algorithm 2 *is solvable and the required conditions therein are satisfied.*

Proof. It follows from [22, Proposition 1.51] that $\emptyset \neq \partial_B(\nabla\varphi)(x^k)d \subset \partial^2\varphi(x^k)(d)$. Pick any matrix $H_k \in \partial_B(\nabla\varphi)(x^k)$, which is always symmetric. Then the linear system $(H_k + \mu_k)d = -\nabla\varphi(x^k)$ is solvable if $\mu_k > \max(0, -\lambda_{\min}(H_k))$, where $\lambda_{\min}(H_k)$ is the smallest eigenvalue of H_k that may be negative since φ is nonconvex. Moreover, the solution is nonzero whenever $\nabla\varphi(x^k)$ is nonzero. Choose $\mu_k = \max(0, -\lambda_{\min}(H_k)) + m$ with $m > 0$ and observe that the symmetric matrix $H_k + \mu_k I$ is positive-definite and that $\langle -\nabla\varphi(x^k), d^k\rangle = \langle (H_k + \mu_k I)d^k, d^k\rangle \geq m\|d^k\|^2$. Furthermore, since $\nabla\varphi$ is locally Lipschitzian on the compact set Ω, it is Lipschitz continuous on Ω with some constant $l > 0$. By [41, Theorem 1.44], we have that $\|w\| \leq l\|d\|$ for any d and any $w \in \partial^2\varphi(x^k)(d)$. Combining the latter with the definition of μ_k brings us to

$$\|\nabla\varphi(x^k)\| = \|(H_k + \mu_k I)d^k\| \leq \|H_k d^k\| + \|\mu_k d^k\| \leq l\|d^k\| + l\|d^k\| + m\|d^k\| = M\|d^k\|$$

with $M := 2l + m > m$ and thus completes the proof of the theorem. □

Remark 2 (discussions on GRNM-WM).

(i) It follows from the proof of Theorem 7 that the regularization parameter μ_k can be chosen as $\mu_k := l + m$, where $m > 0$ is arbitrary and l is the Lipschitz constant of $\nabla\varphi$ on Ω. Observe that, even without the prior knowledge of l, we can adaptively choose μ_k in the following way: pick any number $m > 0$ and set $\mu_k := m + \mu r^j$, where $\mu > 0$, $r > 1$ and $j \geq 0$ is the first nonnegative integer such that the linear system is solvable and the conditions on d^k hold. This procedure terminates in *finitely many steps*.

(ii) In Algorithm 2, we can also use a more *general regularization matrix* $B_k \succ 0$ similarly to Algorithm 1. An appropriate modification of the proof of Theorem 7 shows that such an algorithm is still well posed. Although our convergence analysis below can be easily extended to this more general case, for simplicity we focus on the case of the identity regularization matrix.

Next we formulate and discuss some efficient conditions, which provide global convergence and convergence rates for numerical algorithms of optimization; in particular, those developed in this paper. For functions of class $C^{1,1}$, Polyak [51] introduced the condition

$$\|\nabla f(x)\| \geq (1/2M)|f(x) - f(\bar{x})|^{1/2}, \quad c > 0,$$

and used it to prove a linear convergence of the gradient descent method In Hilbert spaces. Independently, Łojasiewicz [34] introduced the inequality

$$\|\nabla f(x)\| \geq b\,|f(x)-f(\bar{x})|^q, \quad b := 1/M(1-q), \quad q \in [0,1), \tag{36}$$

for analytic functions in the finite-dimensional framework of semialgebraic geometry with no applications to optimization. The gradient inequality (36) is referred to (especially in the literature on machine learning and computer science) as the *Polyak-Łojasiewicz* (PL) condition; see, e.g., [24]. The subsequent algebraic-geometric extension of (36) was developed by Kurdyka [30] for the class of definable differentiable functions on $\mathbb{R}^n$. A nonsmooth extension of that condition was first proposed in [5] in the form of Definition 3(i) via Clarke's subdiffential under the name of "Kurdyka-Łojasiewicz (KL) inequality" without any reference to the pioneering work by Polyak. We suggest using the name *Polyak-Łojasiewicz-Kurdyka* (PKL) conditions for the properties of this type formulated below.

Definition 3 (PLK conditions). Let $f:\mathbb{R}^n \to \overline{\mathbb{R}}$ be an extended-real-valued lower semicontinuous (l.s.c.) function, and let $\bar{x} \in \operatorname{dom} f$. We say that:

(i) The *basic PLK property* holds for f at $\bar{x}$ if there exist a number $\eta \in (0,\infty]$), a neighborhood U of $\bar{x}$, and a function $\psi:[0,\eta) \to \mathbb{R}_+$ such that ψ is concave and C^1-smooth on $(0,\eta)$ with $\psi(0)=0$ and $\psi'(s)>0$ for all $s \in (0,\eta)$, and that we have

$$\psi'\big(f(x)-f(\bar{x})\big)\operatorname{dist}\big(0,\partial f(x)\big) \geq 1 \ \text{ for all } \ x \in U \cap \big\{x \in \mathbb{R}^n \;\big|\; f(\bar{x}) < f(x) < f(\bar{x})+\eta\big\}.$$

(ii) If ψ can be chosen in (i) as $\psi(s)=cs^{1-\theta}$ with $\theta \in [0,1)$ for some $c>0$, then f satisfies the *exponent PLK property* at $\bar{x}$ with the exponent θ.

(iii) f is a *PLK function (of exponent $\theta \in [0,1)$)* if f enjoys the basic PLK property (of exponent θ, respectively) at every point $\bar{x} \in \operatorname{dom} f$.

A large class of descent algorithms for which the PLK conditions are instrumental for deriving impressive convergence properties is described in [2] via the following generic properties of iterative sequences:

(H1) There exists $a>0$ such that for all $k \geq 0$ we have

$$\varphi(x^{k+1}) \leq \varphi(x^k) - a\|x^{k+1}-x^k\|^2. \tag{H1}$$

(H2) There exists $b>0$ such that for all $k \geq 0$ we have

$$\|\nabla\varphi(x^{k+1})\| \leq b\|x^{k+1}-x^k\|. \tag{H2}$$

The next lemma presents convergence results under the PLK properties for abstract descent algorithms satisfying conditions (H1) and (H2).

Lemma 8 (convergence of abstract descent algorithms under PLK conditions). *Let $\varphi:\mathbb{R}^n \to \mathbb{R}$ be of class $C^{1,1}$, and let $\{x^k\}$ satisfy conditions (H1) and (H2). The following hold:*

(i) *If φ is a PLK function, then we have $\sum_{k=0}^{\infty}\|x^{k+1}-x^k\|<\infty$. In particular, $\{x^k\}$ converges to a stationary point $\bar{x}$ as $k \to \infty$.*

(ii) *If φ is a PLK function of exponent $\theta \in (0,1)$, then $\{x^k\}$ converges to $\bar{x}$ with the rates:*

(a) *When $\theta \in (0,\frac{1}{2})$, for any $\varepsilon>0$ and any large $k \in \mathbb{N}$ we have*

$$\|x^{k+1}-\bar{x}\| \leq \varepsilon\|x^k-\bar{x}\|^{\frac{1}{2\theta}}.$$

If furthermore $\bar{x}$ is a local minimizer of φ, then there is no function φ of class $C^{1,1}$ satisfying the PLK property at $\bar{x}$ with such an exponent.

(b) *When $\theta=\frac{1}{2}$, there exist $\gamma>0$ and $q \in (0,1)$ such that for all large $k \in \mathbb{N}$ we have*

$$\|x^k-\bar{x}\| \leq \sum_{j=k}^{\infty}\|x^{j+1}-x^j\| \leq \gamma q^k.$$

(c) *When $\theta \in (\frac{1}{2},1)$, there exists $\gamma>0$ such that*

$$\|x^k-\bar{x}\| \leq \sum_{j=k}^{\infty}\|x^{j+1}-x^j\| \leq \gamma k^{\frac{1-\theta}{1-2\theta}}.$$

Proof. Assertions (i) and (ii) in cases (b) and (c), as well as the convergence rates in case (a) of (ii) for arbitrary stationary points of φ, are taken from [48, Theorems 3.1 and 3.2] with a small rewording. The inconsistency between the class $C^{1,1}$ and the PLK property of φ with exponent $\theta \in (0, \frac{1}{2})$ at local minimizers of φ has been recently observed in [4, Theorem 4]. □

Remark 3 (modified generic conditions). For some important descent algorithms, condition (H2) does not hold, while its replacement

(H2') There exists $b > 0$ such that for all $k \geq 0$ we have

$$\|\nabla\varphi(x^k)\| \leq b\|x^{k+1} - x^k\|$$

does; see more discussions in [4] on such algorithms for general l.s.c. functions. It follows from [4, Theorem 2] that for any abstract algorithm satisfying (H1) and (H2'), the PLK property of exponent $\theta \in (0, \frac{1}{2})$ at a stationary point $\bar{x}$ of φ yields the *finite termination* of the algorithm.

The next lemma shows that the iterative sequence $\{x^k\}$ generated by GRNM-WM (Algorithm 2) for minimizing of $C^{1,1}$ functions satisfies conditions (H1) and (H2).

Lemma 9 (embedding GRNM-WM into the generic scheme). *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a function of class $C^{1,1}$, and let $x^0 \in \mathbb{R}^n$ be a starting point. Suppose that the set $\Omega = \{x \in \mathbb{R}^n \mid \varphi(x) \leq \varphi(x^0)\}$ is bounded. Then the sequence $\{x^k\}$ generated by Algorithm 2 satisfies conditions (H1) and (H2).*

Proof. To verify condition (H1), we deduce from the algorithmic design that

$$\begin{aligned}
\varphi(x^{k+1}) - \varphi(x^k) &\leq \sigma_1\tau_k\langle\nabla\varphi(x^k), d^k\rangle \\
&= -\sigma_1\tau_k\langle-\nabla\varphi(x^k), d^k\rangle \\
&\leq -\sigma_1\tau_k m\|d^k\|^2 \\
&= -\frac{\sigma_1 m}{\tau_k}\|\tau_k d^k\|^2 \\
&\leq -\frac{\sigma_1 m}{\tau_{\max}}\|\tau_k d^k\|^2 \\
&= -\frac{\sigma_1 m}{\tau_{\max}}\|x^{k+1} - x^k\|^2,
\end{aligned}$$

where the first inequality is the sufficient decrease property, the second one follows from the condition $\langle-\nabla\varphi(x^k), d^k\rangle \geq m\|d^k\|^2$ guaranteed in Algorithm 2, and the third inequality holds since τ_k is upper bounded by $\tau_{\max}$, the maximum stepsize allowed.

Next we show that condition (H2) holds. Note that φ is Lipschitz continuous on Ω with some constant l since Ω is compact and φ is locally Lipschitz continuous by the imposed assumptions. It follows from the curvature condition as in the proof of Claim 6 of Theorem 6 that

$$(\sigma_2 - 1)\langle\nabla\varphi(x^k), d^k\rangle \leq \langle\nabla\varphi(x^{k+1}) - \nabla\varphi(x^k), d^k\rangle \leq l\tau_k\|d^k\|^2.$$

From this and $m\|d^k\|^2 \leq \langle-\nabla\varphi(x^k), d^k\rangle$, we deduce that $\tau_k \geq \frac{(1-\sigma_2)m}{l}$. Therefore,

$$\begin{aligned}
\|\nabla\varphi(x^{k+1})\| &\leq \|\nabla\varphi(x^{k+1}) - \nabla\varphi(x^k)\| + \|\nabla\varphi(x^k)\| \\
&\leq l\|x^{k+1} - x^k\| + M\|d^k\| \\
&= l\|x^{k+1} - x^k\| + \frac{M}{\tau_k}\|\tau_k d^k\| \\
&\leq l\|x^{k+1} - x^k\| + \frac{Ml}{(1-\sigma_2)m}\|x^{k+1} - x^k\| \\
&= \left(l + \frac{Ml}{(1-\sigma_2)m}\right)\|x^{k+1} - x^k\|,
\end{aligned}$$

where the second inequality follows from the condition $\|\nabla\varphi(x^k)\| \leq M\|d^k\|$ in Algorithm 2. □

The following theorem summarizes the main results for GRNM-WM obtained above.

Theorem 10 (performance of CRNM-WM). *Let $\varphi:\mathbb{R}^n\to\mathbb{R}$ be of class $C^{1,1}$, and let $\{x^k\}$ be the sequence generated by Algorithm 2. Suppose that the set $\Omega=\{x\in\mathbb{R}^n\mid \varphi(x)\le\varphi(x^0)\}$ is bounded. Then Algorithm 2 is well defined and exhibits the convergence properties of Lemma 8 under the fulfillment of the corresponding PLK conditions therein.*

Proof. This follows directly from Theorem 7 and Lemmas 8, 9. □

5 Coderivative-Based Newton Forward-Backward Method

This section is devoted to a class of convex composite minimization problems, which are *first-order nonsmooth* and can incorporate *constraints*. Therefore, the coderivative-based Newton methods CRNM-W and CRNM-WM proposed and developed in previous sections cannot be applied directly. Following the approach in [27], implemented there for the case of Armijo's linesearch, we now employ the *forward-backward envelope* (FBE) machinery from [49] to the novel Wolfe linesearch in coderivative-based Newtonian algorithms. For brevity, our main attention is paid here to applying CRNM-W to the FBE setting. We label the new algorithm as the *coderivative-based forward-backward Newton method* (abbr. CNFB).

Consider the class of *convex composite minimization problem*

$$\text{minimize}\ \ \varphi(x)=f(x)+g(x)\ \ \text{over all}\ \ x\in\mathbb{R}^n, \tag{37}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a C^2-smooth convex function and $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ is proper, l.s.c., and convex. The word "composition" signifies here that the functions f and g have completely different natures. Being extended-real-valued, the function g allows us to implicitly incorporate constraints in the seemingly unconstrained framework of optimization. Problems of type (37) arise in many areas of research and practical modeling like machine learning, data science, signal processing, and statistics, where the nonsmooth term plays a role of regularizators. The main idea of CNFB is applying CRNM-W to the FBE associated with φ in (37), which happens to be of class $C^{1,1}$ and allows us to eventually solve the original problem.

To begin with, recall the relevant definitions and facts needed in what follows. The first constructions are classical; see, e.g., in [55]. Given a proper l.s.c. function $\varphi:\mathbb{R}^n\to\overline{\mathbb{R}}$, the *Moreau envelope* of φ with a parameter $\gamma>0$ and the associated *proximal mapping* of φ are defined, respectively, by

$$e_\gamma\varphi(x):=\inf_{y\in\mathbb{R}^n}\Big\{\varphi(y)+\frac{1}{2\gamma}\|y-x\|^2\Big\},\quad x\in\mathbb{R}^n, \tag{38}$$

$$\operatorname{Prox}_{\gamma\varphi}(x):=\operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{\varphi(y)+\frac{1}{2\gamma}\|y-x\|^2\Big\},\quad x\in\mathbb{R}^n. \tag{39}$$

Let $\varphi:\mathbb{R}^n\to\overline{\mathbb{R}}=f+g$, where $f:\mathbb{R}^n\to\mathbb{R}$ is C^1-smooth and where $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ is proper l.s.c. The *forward-backward envelope* (abbr. FBE) of φ with parameter $\gamma>0$ is introduced in [49] by

$$\varphi_\gamma(x):=\inf_{y\in\mathbb{R}^n}\Big\{f(x)+\langle\nabla f(x),y-x\rangle+g(y)+\frac{1}{2\gamma}\|y-x\|^2\Big\},\quad x\in\mathbb{R}^n. \tag{40}$$

The following properties of FBEs are taken from [49, 56].

Proposition 11 (properties of FBEs). *Consider the class of functions $\varphi=f+g$, where $f:\mathbb{R}^n\to\mathbb{R}$ is convex, C^2-smooth and being such that ∇f is Lipschitz continuous with modulus $l>0$, and where $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ is proper l.s.c., and convex. Then we have the assertions:*

(i) *The FBE φ_γ of φ in (40) is C^1-smooth function with the gradient*

$$\nabla\varphi_\gamma(x)=\gamma^{-1}\big(I-\gamma\nabla^2 f(x)\big)\big(x-\operatorname{Prox}_{\gamma g}\big(x-\gamma\nabla f(x)\big)\big).$$

Moreover, the set of optimal solutions to problem (37) coincides with the set of stationary points of the FBE φ_γ for all parameter values $\gamma\in(0,l^{-1})$.

(ii) *If $f(x)=\frac{1}{2}\langle Ax,x\rangle+\langle b,x\rangle+\alpha$ with $A\in\mathbb{R}^{n\times n}$ being a symmetric and positive-semidefinite matrix, $b\in\mathbb{R}^n$, and $\alpha\in\mathbb{R}$, then for all $\gamma\in(0,\frac{1}{l})$, the FBE φ_γ of φ is convex and its gradient $\nabla\varphi_\gamma$ is Lipschitz continuous with modulus $L:=2(1-\gamma\lambda_{\min(A)})/\gamma$. If A is positive-definite, then φ_γ is strongly convex with modulus $K:=\min\{(1-\gamma\lambda_{\min(A)})\lambda_{\min(A)},(1-\gamma\lambda_{\max(A)})\lambda_{\max(A)}\}$.*

Consider further the unconstrained optimization problem

$$\text{minimize} \;\; \varphi_\gamma(x) \;\; \text{over all} \;\; x\in\mathbb{R}^n \tag{41}$$

whose stationary points are optimal solutions to (37) when f is C^2-smooth with the Lipschitz continuous gradient. Since the gradient $\nabla\varphi_\gamma$ may not be locally Lipschitz continuous, (41) is not a $C^{1,1}$ optimization problem in general. From now on, we assume that f is a *quadratic function* as in Proposition 11(ii), and thus $\nabla\varphi_\gamma$ is Lipschitz continuous.

In what follows, we focus on the convex composite minimization problem

$$\text{minimize} \;\; \varphi(x)=\frac{1}{2}\langle Ax,x\rangle+\langle b,x\rangle+g(x) \;\; \text{over all} \;\; x\in\mathbb{R}^n, \tag{42}$$

where $A\in\mathbb{R}^{n\times n}$ is a positive-semidefinite symmetric matrix, where $b\in\mathbb{R}^n$, and where $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper l.s.c. convex function. To implement GRNM-W for problem (42), we need to use the generalized Hessian of φ_γ, which is calculated in [27].

Proposition 12 (calculating of generalized Hessians of FBEs). *Let $\varphi=f+g$ be as in (42), and let $\gamma>0$ be such that $R:=I-\gamma A\succ 0$. Then the generalized Hessian of φ_γ is calculated by*

$$\begin{aligned}&\bar z\in\partial^2\varphi_\gamma(\bar x)(w)\\ \Longleftrightarrow\;& R^{-1}\bar z-Aw\in\partial^2 g\Big(\operatorname{Prox}_{\gamma g}(\bar u),\frac{1}{\gamma}(\bar u-\operatorname{Prox}_{\gamma g}(\bar u))\Big)(w-\gamma R^{-1}\bar z)\end{aligned}$$

for $\bar x\in\mathbb{R}^n, w\in\mathbb{R}^n, \bar u:=\bar x-\gamma(A\bar x+b)$. This can be equivalently expressed as

$$\partial^2\varphi_\gamma(\bar x)(w)=\gamma^{-1}R\big(w+D^*(-\operatorname{Prox}_{\gamma g})(\bar u)(Rw)\big). \tag{43}$$

We employ (43) in the construction of CNFB to solve (42).

Algorithm 3 Coderivative-based Newton forward-backward method (CNFB)

Input: $x^0\in\mathbb{R}^n$, $\gamma>0$ such that $R:=I-\gamma A\succ 0$, $0<\sigma_1<\sigma_2<1$, $c>0$, $\rho\in(0,1]$.

1: **for** $k=0,1,\ldots$ **do**

2: If $\nabla\varphi_\gamma(x^k)=0$, stop; otherwise set $u^k:=x^k-\gamma(Ax^k+b)$, $v^k:=\operatorname{Prox}_{\gamma g}(u^k)$, $\mu_k=c\|\nabla\varphi_\gamma(x^k)\|^\rho=c\gamma^{-1}\|R(x^k-v^k)\|^\rho$ and go to next step.

3: Choose $B_k\succ 0$. Find $d^k\in\mathbb{R}^n$ such that $-\nabla\varphi_\gamma(x^k)\in\partial^2\varphi_\gamma(x^k)(d^k)+\mu_k B_k d^k$, i.e.,

$$\gamma^{-1}R(v^k-x^k)\in\gamma^{-1}R\big(d^k+D^*(-\operatorname{Prox}_{\gamma g})(u^k)(Rd^k)\big)+\mu_k B_k d^k.$$

4: Set $\tau_k=1$ and check the Wolfe conditions:

$$\varphi_\gamma(x^k+\tau_k d^k)\le\varphi_\gamma(x^k)+\sigma_1\tau_k\langle\nabla\varphi_\gamma(x^k),d^k\rangle,$$

$$\langle\nabla\varphi_\gamma(x^k+\tau_k d^k),d^k\rangle\ge\sigma_2\langle\nabla\varphi_\gamma(x^k),d^k\rangle.$$

If the latter conditions do not hold, adjust τ_k (using any specific implementation of the Wolfe linesearch) until it satisfies the Wolfe conditions. We assume that in the implementation there exists an upper bound $\tau_{\max}$ on the maximum stepsize allowed.

5: Set $x^{k+1}=x^k+\tau_k d^k$.

6: **end for**

To proceed with establishing the convergence properties of Algorithm 3, we need to recall the two results obtained in [27]. The first one is taken from [27, Proposition 5].

Proposition 13 (metric regularity of FBEs). *Let $\varphi=f+g$ be as in (42), and let $\gamma>0$ be such that $R:=I-\gamma A\succ 0$. For any $\bar x\in\mathbb{R}^n$ with $0\in\partial\varphi(\bar x)$, the following assertions hold:*

(i) *$\partial\varphi$ is metrically regular around $(\bar x,0)$ if and only if $\nabla\varphi_\gamma$ is metrically regular around $\bar x$.*

(ii) *$\bar x$ is a tilt-stable local minimizer of φ if and only if $\bar x$ is a tilt-stable local minimizer of φ_γ.*

(iii) $\|\partial^2\varphi_\gamma(\bar x)^{-1}\|\le\|\partial^2\varphi(\bar x,0)^{-1}\|+\gamma\|R^{-1}\|$.

In the next proposition taken from [27, Propositions 6], we use the notion of *twice epi-differentiability* of extended-real-valued functions, which is studied and applied to optimization in [55] and more recent papers; see, e.g., [36, 37, 38], where the reader can find more details and references.

Proposition 14 (semismoothness* and directional differentiability for FBEs). *In the setting of Proposition 13, the following assertions hold:*

(i) $\nabla\varphi_\gamma$ *is semismooth* at* $\bar{x}$ *if* ∂g *is semismooth* at* $(\bar{x},\bar{v})$*, where* $\bar{v}:=-A\bar{x}-b$*;*

(ii) $\nabla\varphi_\gamma$ *is directionally differentiable at* $\bar{x}$ *if* g *is twice epi-differentiable at* $\bar{x}$ *for* $\bar{v}$*.*

Now we are ready to establish comprehensive convergence results for CNFB.

Theorem 15 (performance of CNFB). *Consider the convex composite minimization problem (42), where the symmetric matrix* A *is positive-semidefinite. Then we have the assertions:*

(i) *CNFB (Algorithm 3) either stops after finitely many iterations at a minimizer of* φ*, or generates a sequence* $\{x^k\}$ *whose accumulation points are optimal solutions to problem (42).*

(ii) *If the subgradient mapping* $\partial\varphi$ *is metrically regular around* $(\bar{x},0)$ *with modulus* $\kappa>0$ *(which is satisfied, in particular, when* A *is positive-definite), where* $\bar{x}$ *is an accumulation point of* $\{x^k\}$*, then the sequence* $\{x^k\}$ *converges with local R-linear rate to* $\bar{x}$*, and* $\bar{x}$ *is a tilt-stable local minimizer of* φ*.*

(iii) *The local convergence rate of* $\{x^k\}$ *is Q-superlinear if the subgradient mapping* ∂g *is semismooth* at* $(\bar{x},\bar{v})$*, where* $\bar{v}:=-A\bar{x}-b$*, and if either one of the following conditions holds:*

(a) g *is twice epi-differentiable at* $\bar{x}$ *for* $\bar{v}$*.*

(b) *The linesearch constants satisfy the conditions* $\sigma_1\in(0,\frac{1}{2LK})$ *and* $\sigma_2\in(1-\frac{K}{L},1)$*, where* $L:=2(1-\gamma\lambda_{\min(A)})/\gamma$ *and* $K:=\kappa+\gamma\|B^{-1}\|$*.*

Proof. By Proposition 11(i), minimizing φ reduces to minimizing the FBE function φ_γ of class $C^{1,1}$ when the parameter $\gamma>0$ is sufficiently small. We now verify each claim of the theorem.

(i) Proposition 11(ii) tells us that the FBE φ_γ is convex and its gradient $\nabla\varphi_\gamma$ is Lipschitz continuous with modulus $L:=2(1-\gamma\lambda_{\min(A)})/\gamma$. By [7, Theorem 3.2], the generalized Hessian $\partial^2\varphi_\gamma(x)$ is positive-semidefinite for all $x\in\mathbb{R}^n$. Then GRNM-W is well defined, and the claimed assertion (i) follows from Theorems 4 and 5.

(ii) By [12, Proposition 4.5], the tilt stability of φ at $\bar{x}$ with modulus κ follows from the metric regularity of $\partial\varphi$ and the convexity of φ. By Proposition 13 (i), the gradient mapping $\nabla\varphi_\gamma$ is metrically regular around $\bar{x}$. Then the R-linear convergence of $\{x^k\}$ follows from Theorem 6(ii). We also need to show that if A is positive-definite, then $\partial\varphi$ is metrically regular around $(\bar{x},0)$. Indeed, the positive-definiteness of A ensures by Proposition 11(ii) that φ_γ is strongly convex with the Lipschitz continuous gradient. Using [7, Theorem 5.1] tells us that $\partial^2\varphi_\gamma(x)$ is positive-definite for all $x\in\mathbb{R}^n$. This implies by [12, Proposition 4.5 and Theorem 4.6] that $\nabla\varphi_\gamma$ is metrically regular around $\bar{x}$. Therefore, we deduce from Proposition 13(i) that $\partial\varphi$ is metrically regular around $(\bar{x},0)$.

(iii) Proposition 14(i) verifies that $\nabla\varphi_\gamma$ is semismooth* at $\bar{x}$. In case (a), $\nabla\varphi_\gamma$ is directionally differentiable at $\bar{x}$ by Proposition 14(ii). Then Theorem 6(iii,a) yields assertion (iii) in this case.

To complete the proof, it remains to consider case (b) in (iii). It follows from Proposition 11(ii) that the number L therein is a Lipschitz constant of $\nabla\varphi_\gamma$ around $\bar{x}$. Then Proposition 13(i) and the Mordukhovich criterion from [55, Theorem 9.40] (see [40]) ensure that the FBE φ_γ is metrically regular around $\bar{x}$ with the modulus K defined above, and thus the claimed assertion (iii) holds by Theorem 6(iii,b). □

6 Coderivative-Based Newton Augmented Lagrangian Method

In this section, we develop another algorithm for a class of convex composite minimization problems, first-order nonsmooth and constrained, by embedding GRNM-W into the augmented Lagrangian method. Inspired by the semismooth Newton augmented Lagrangian method (SSNAL) [31], an augmented Lagrangian method employing the semismooth Newton method as inner problem solver, we propose here the new *coderivative-based Newton augmented Lagrangian method* (abbr. CNAL).

Consider the following *linear-convex composite minimization problem* written in the unconstrained extended-real-valued format as

$$(\mathbf{P})\quad \text{minimize}\ \ f(x):=h(\mathscr{A}x)-\langle c,x\rangle+p(x)\ \ \text{over all}\ \ x\in\mathbb{R}^n, \tag{44}$$

where $\mathscr{A}:\mathbb{R}^n\to\mathbb{R}^m$ is a linear mapping, $h:\mathbb{R}^m\to\mathbb{R}$ is a l.s.c. convex function, $p:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper l.s.c. convex function, and $c\in\mathbb{R}^n$. The *dual problem* of (44) is given by

$$\begin{aligned}(\mathbf{D})\quad &\text{minimize}\quad h^*(y)+p^*(z)\\ &\text{subject to}\quad \mathscr{A}^*y+z=c,\end{aligned}\tag{45}$$

where h^* and p^* are the Fenchel conjugates of h and p, respectively, and $\mathscr{A}^*=\mathscr{A}^T$ is the adjoint/transpose mapping of $\mathscr{A}$; see, e.g., [55, Example 11.41].

Given $\sigma>0$, the *augmented Lagrangian* associated with (45) is

$$\mathscr{L}_\sigma(y,z;x):=h^*(y)+p^*(z)-\langle x,\mathscr{A}^*y+z-c\rangle+\frac{\sigma}{2}\|\mathscr{A}^*y+z-c\|^2$$

whenever $(y,z,x)\in\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^n$. From now on, the following additional assumption is imposed:

Assumption 1. *The function h in (44) is strongly convex and C^2-smooth.*

Now we propose the coderivative-based Newton augmented Lagrangian method (CNAL) to solve the dual problem (45)(and thus the primal problem (44) by strong duality) designed as follows:

Algorithm 4 Coderivative-based Newton augmented Lagrangian method (CNAL)

Input: $\sigma_0>0$, $(y^0,z^0,x^0)\in\mathbb{R}^m\times\operatorname{dom}p^*\times\mathbb{R}^n$.

1: **for** $k=0,1,\ldots$ **do**

2: Compute

$$(y^{k+1},z^{k+1})\approx\operatorname{argmin}\left\{\Psi_k(y,z):=\mathscr{L}_{\sigma_k}(y,z;x^k)\right\}\tag{46}$$

via the coderivative-based Newton method GRNM-W. The stopping criterion is: $\Psi_k(y^{k+1},z^{k+1})-\inf\Psi_k\le\varepsilon_k^2/2\sigma_k$ with $\sum_{k=0}^\infty\varepsilon_k<\infty$.

3: Compute $x^{k+1}=x^k-\sigma_k(\mathscr{A}^*y^{k+1}+z^{k+1}-c)$ and update $\sigma_{k+1}\uparrow\sigma_\infty\le\infty$.

4: **end for**

The next theorem establishes convergence properties of iterates in Algorithm 4.

Theorem 16 (convergence of CNAL). *Suppose that the primal problem (44) admits an optimal solution and that all the assumptions imposed above are satisfied. Then for any infinite sequence of iterates $\{(y^k,z^k,x^k)\}$ generated by Algorithm 4, we have that $\{x^k\}$ converges to an optimal solution of problem (44) while $\{(y^k,z^k)\}$ converges to an optimal solution of the dual problem (45).*

Proof. Observe first that Assumption 1 ensures by [55, Proposition 12.60] that h^* is strongly convex and C^2-smooth. It follows from the classical Fenchel duality theorem that strong duality holds in the setting of (44) and (45). Employing finally [54, Theorem 4] verifies the claimed convergence results. □

Next we consider the subproblem in (46) formulated as

$$\text{minimize}\ \ \Psi(y,z):=\mathscr{L}_\sigma(y,z;\tilde{x})\ \ \text{over}\ \ (y,z)\in\mathbb{R}^m\times\mathbb{R}^n$$

for which the optimal solution is given by

$$\bar{y}=\operatorname{argmin}\psi(y),\quad \bar{z}=\operatorname{Prox}_{p^*/\sigma}(\tilde{x}/\sigma-\mathscr{A}^*\bar{y}+c)$$

via the proximal mapping (39), where the function ψ is defined by

$$\psi(y):=\inf_z\Psi(y,z)=h^*(y)+e_{\frac{1}{\sigma}}p^*\big(\tilde{x}/\sigma-(\mathscr{A}^*y-c)\big)-\frac{1}{2\sigma}\|\tilde{x}\|^2\tag{47}$$

via the Moreau envelope (38). To solve subproblem (46), we apply GRNM-W to minimizing the function ψ. The gradient of ψ is computed by

$$\begin{aligned}\nabla\psi(y)&=\nabla h^*(y)-\mathscr{A}\nabla e_{\frac{1}{\sigma}}p^*\big(\tilde{x}/\sigma-(\mathscr{A}^*y-c)\big)\\ &=\nabla h^*(y)-\mathscr{A}\big(\sigma(u'-\operatorname{Prox}_{\frac{1}{\sigma}p^*}(u'))\big)\\ &=\nabla h^*(y)-\mathscr{A}\big(\sigma(\sigma^{-1}\operatorname{Prox}_{\sigma p}(\sigma u'))\big)\\ &=\nabla h^*(y)-\mathscr{A}\operatorname{Prox}_{\sigma p}\big(\tilde{x}-\sigma(\mathscr{A}^*y-c)\big),\end{aligned}$$

where $u' := \tilde{x}/\sigma - (\mathscr{A}^* y - c)$. Note that $\nabla\psi$ is locally Lipschitz continuous while being nonsmooth due to the presence of the proximal mapping.

To implement Algorithm 4, we need to constructively evaluate the generalized Hessian of ψ (i.e., the limiting coderivative of $\nabla\psi$) in line 2. The precise calculation of the generalized Hessian $\partial^2\psi$ in this case is challenging, but we can give an upper estimate that is sufficient for computational purposes.

Theorem 17 (upper estimate of generalized Hessians). *The generalized Hessian of ψ from (47) admits the following upper estimate, where $u = \tilde{x} - \sigma(\mathscr{A}^* y - c)$:*

$$\partial^2\psi(y)(w) \subset \nabla^2 h^*(y)(w) - \sigma\mathscr{A}\big(D^* \operatorname{Prox}_{\sigma p}\big)(u)(-\mathscr{A}^* w), \quad w \in \mathbb{R}^m.$$

Proof. Note that the standing Assumption 1 implies that h^* is C^2-smooth. It follows from the coderivative sum rule in [42, Theorem 3.9] that

$$\big(D^*\nabla\psi\big)(y)(w) = \nabla^2 h^*(y)(w) + \big(D^*(-\mathscr{A}S)\big)(y)(w),$$

where $S(y) := \operatorname{Prox}_{\sigma p}(\tilde{x} - \sigma(\mathscr{A}^* y - c))$. By [42, Theorem 3.11, (i)], we have

$$\big(D^*(-\mathscr{A}S)\big)(y)(w) \subset D^*S(y)(-\mathscr{A}^* w) \subset -\sigma\mathscr{A}\big(D^* \operatorname{Prox}_{\sigma p}\big)(u)(-\mathscr{A}^* w)$$

since $\mathscr{A}$ and $\operatorname{Prox}_{\sigma p}$ are Lipschitz continuous. □

Remark 4. For ψ from (47), we write $\hat{\partial}^2\psi(y)(w) := \nabla^2 h^*(y)(w) - \sigma\mathscr{A}\big(D^* \operatorname{Prox}_{\sigma p}\big)(u)(-\mathscr{A}^* w)$. Theorem 17 tells us that $\partial^2\psi(y)(w) \subset \hat{\partial}^2\psi(y)(w)$ for all $w \in \mathbb{R}^m$. Suppose that $D^* \operatorname{Prox}_{\sigma p}$ is positive-semidefinite (in the sense that $\langle w, d\rangle \geq 0$ for $d \in \mathbb{R}^n$ and $w \in D^* \operatorname{Prox}_{\sigma p}(u)(d)$) and that $\operatorname{Prox}_{\sigma p}$ is semismooth*. These assumptions hold in many practical models such as , e.g., the Lasso problem considered in Section 9. Then we can prove, by using similar arguments as in Section 3, that GRNM-W (Algorithm 1) with $\hat{\partial}^2\psi(y)$ (instead of $\partial^2\psi(y)$) also converges superlinearly for the minimization of (47).

7 Numerical Comparison with GRNM

In this section, we compare our algorithm GRNM-W with GRNM [27], which employs Armijo line search.

Consider the following test minimization problem:

$$\min_{x\in\mathbb{R}^n} \varphi_\gamma(x) := \frac{1}{2}\|Ax-b\|^2 - \frac{\gamma}{2}\|A^*(Ax-b)\|^2 + e_{\gamma g}(x - \gamma A^*(Ax-b)), \tag{48}$$

where $\gamma > 0$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $g(x) := \lambda\|x\|_1$ with $\lambda > 0$, and $e_{\gamma g}$ is the Moreau envelope (38) of g. This problem is a reformulation of the Lasso problem (57) via the forward-backward envelope (40). We omit the implementation details since our purpose is only to compare the line searches in GRNM-W and GRNM.

In our experiments. the matrix A is generated randomly with iid standard Gaussian entries, and the vector b is generated randomly with iid standard uniform components. We set $\lambda = 10^{-3}$, $\gamma = \frac{1}{2}\|A^*A\|^{-1}$ and choose an initial point x^0 with large components: each of the components of x^0 is 10^3. In this way, x^0 is far from the sparse minimizer. This allows us to test the performance of GRNM-W and GRNN when the initial point is far from the optimal solution. The (absolute) KKT residual $\eta_k = \|x^k - \operatorname{Prox}_{\lambda\|\cdot\|_1}(x^k - A^*(Ax^k - b))\|$ is used to measure the accuracy of an approximate minimizer and also serves as the stopping criterion. The results are displayed in Table 1, where 'iter' stands for the number of iterations and 'CPU time' refers to the time needed to obtain an approximate solution of prescribed accuracy 10^{-6}. In Table 2, we also show the

results for different initial points.

Table 1: Test problems

	Size		CPU time		iter	
	m	n	GRNM	GRNM-W	GRNM	GRNM-W
$m > n$	20	10	0.17s	0.01s	3850	65
	50	10	0.16s	0.01s	3385	58
	100	10	0.15s	0.01s	3238	55
	200	100	9.63s	0.72s	11181	76
	500	100	12.39s	0.96s	10597	67
	1000	100	13.76s	1.25s	10239	64
$m = n$	10	10	0.16s	0.02s	3939	95
	50	50	2.58s	0.48s	11510	725
	100	100	8.05s	2.46s	15641	775
	500	500	248.54s	32.12s	34973	1113
	1000	1000	1259.01s	119.31s	48904	1252
$m < n$	10	20	0.59s	0.06s	6487	127
	10	50	2.24s	0.47s	9259	302
	10	100	7.26s	3.02s	12711	565
	100	200	73.90s	18.37s	24490	1152

Table 2: Test problems with different initial points

	Size		x^0	CPU time		iter	
	m	n		GRNM	GRNM-W	GRNM	GRNM-W
$m > n$	1000	500	0	0.10s	0.15s	6	6
			1	0.28s	0.35s	35	13
			10	2.21s	0.78s	273	19
			100	19.31s	1.72s	2632	27
$m = n$	1000	1000	0	1.54s	2.77s	30	30
			1	3.97s	5.94s	77	66
			10	13.12s	9.87s	512	123
			100	125.39s	32.86s	4909	308
$m < n$	500	1000	0	60.81s	207.82s	1402	1402
			1	52.19s	176.94s	1221	1196
			10	72.58s	185.06s	1758	1284
			100	207.29s	181.70s	6846	1383

It is clear from Table 1 and Table 2 that the Wolfe linesearch used in GRNM-W is more efficient than the Armijo linesearch used in GRNM when the initial point is far from the solution, since the former can greatly reduce the number of iterations. In practice, we cannot make sure that the initial point is close to an optimal solution, so the Wolfe line search is particularly attractive in such situations.

8 Applications of CNFB to Support Vector Machines

In this section, we apply the proposed CNFB to optimization problems arising in support vector machines, which can be formulated as a quadratic programming problem with some special structure. The constructive implementations of CNFB to solve such problems with conducting numerical experiments require *explicit calculations* of the *generalized Hessians* for the functions in question.

8.1 Support Vector Machines

A *support vector machine* (SVM) is a machine learning model for binary classification. Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1,-1\}$, the aim of SVMs is to select an appropriate class of *classifiers* for the training data and to optimize its characteristics under the imposed requirements; see, e.g., [23, Chapter 6] with the references therein.

For the class of linear classifiers, the optimization problem in SVMs is formulated as follows:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i \ \text{ over } \ \mathbf{w}, b, \xi \\ &\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge 1 - \xi_i, \quad \xi_i \ge 0 \ \text{ for } \ i = 1,\ldots,n. \end{aligned} \tag{49}$$

The dual problem of (49) is defined by

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\lambda^T D\lambda - \lambda^T\mathbf{e} \ \text{ over } \ \lambda \in \mathbb{R}^n \\ &\text{subject to} \quad \lambda^T\mathbf{y} = 0, \quad \mathbf{0} \le \lambda \le C\mathbf{e}, \end{aligned} \tag{50}$$

where $\mathbf{e} \in \mathbb{R}^n$ is the all-one vector, and where $D \in \mathbb{R}^{n\times n}$ is a positive-semidefinite symmetric matrix with $D_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$. We can see that (50) is a convex quadratic programming problem with a single linear constraint and bound constraints on the variables (i.e., the feasible region is the intersection of a hyperplane and a box). Note that for nonlinear classifiers in SVMs, the dual optimization problem can also be written in such a form by using a kernel function.

Having this in mind, we consider below the following general class of *convex quadratic programs* with a *single linear constraint* and *bound constraints* on the variables (abbr. SLBQP):

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}x^T Qx + c^T x \ \text{ over } \ x \in \mathbb{R}^n \\ &\text{subject to} \quad a^T x = b, \quad l \le x \le u, \end{aligned} \tag{51}$$

where $Q \in \mathbb{R}^{n\times n}$ is a positive-semidefinite symmetric matrix, $c \in \mathbb{R}^n$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}$, $l \in (\mathbb{R}\cup\{-\infty\})^n$, and $u \in (\mathbb{R}\cup\{\infty\})^n$. This class includes support vector machine and bound-constrained quadratic programming problems as special cases.

Observe that the quadratic program (51) is a special case of the convex composite minimization problems of type (42) with $\varphi(x) := f(x) + g(x)$, $f(x) := \frac{1}{2}x^T Qx + c^T x$, $g(x) := \delta_\Gamma(x)$, where δ_Γ is the indicator function of the feasible set $\Gamma := \{x \in \mathbb{R}^n \mid a^T x = b,\ l \le x \le u\}$. To apply the proposed CNFB to solving problem (51), we need to explicitly calculate the generalized Hessian of FBE, which reduces to calculating the coderivative of the (minus) proximal mapping (see Step 3 of Algorithm 3) and eventually to calculating the coderivative of the (minus) projection operator Proj_Γ. This goal is achieved in the next subsection.

8.2 Coderivatives of Polyhedral Projectors

Let Γ be a nonempty convex polyhedral set given by

$$\begin{aligned} \Gamma &:= \big\{x \in \mathbb{R}^n \;\big|\; \langle a_i, x\rangle = b_i \ \text{ for } \ i \in R, \quad \langle c_i, x\rangle \le d_i \ \text{ for } \ i \in S\big\} \\ &= \big\{x \in \mathbb{R}^n \;\big|\; A^T x = b,\ C^T x \le d\big\}, \end{aligned} \tag{52}$$

where $b \in \mathbb{R}^{|R|\times 1}$, $d \in \mathbb{R}^{|S|\times 1}$, and where the matrices $A^T \in \mathbb{R}^{|R|\times n}$ and $C^T \in \mathbb{R}^{|S|\times n}$ are formed by the row vectors a_i^T as $i \in R$ and c_i^T as $i \in S$, respectively.

The following result is taken from [21, Corollary 4.3].

Lemma 18 (coderivative calculations for normals to polyhedra). *Given a convex polyhedron Γ in (52), define the family of index sets*

$$\mathscr{I}_\Gamma := \big\{S' \subset S \;\big|\; \exists x \in \Gamma \ \text{ such that } \ \langle c_i, x\rangle = d_i,\ i \in S' \ \text{ and } \ \langle c_i, x\rangle < d_i,\ i \in S\backslash S'\big\}$$

and pick any $\bar{x}^* \in N_\Gamma(\bar{x})$ with $\bar{x} \in \Gamma$. Then we have $z \in D^* N_\Gamma(\bar{x};\bar{x}^*)(w)$ if and only if

$$\begin{cases} -w \in B_{J,K} := \big\{x \;\big|\; \langle a_i, x\rangle = 0,\ i \in R;\ \langle c_i, x\rangle = 0,\ i \in J;\ \langle c_i, x\rangle \le 0,\ i \in K\big\}, \\ z \in A_{J,K} := B_{J,K}^* = \text{span}\big\{a_i,\ i \in R\big\} + \text{span}\big\{c_i,\ i \in J\big\} + \text{cone}\big\{c_i,\ i \in K\backslash J\big\}, \end{cases} \tag{53}$$

where $K \in \mathscr{I}_\Gamma$, $J \subset K \subset I(\bar{x}) := \big\{i \in S \;\big|\; \langle c_i, \bar{x}\rangle = d_i\big\}$, $\bar{x}^* \in \text{span}\big\{a_i,\ i \in R\big\} + \text{cone}\big\{c_i,\ i \in J\big\}$.

The next lemma follows from the fact that for any convex set Γ, we have $\mathrm{Proj}_\Gamma = (I+N_\Gamma)^{-1}$.

Lemma 19 (coderivatives of projections and normal cone mappings). *Let $u \in \mathbb{R}^n$. Then the inclusion $w \in D^*(-\mathrm{Proj}_\Gamma)(u)(d)$ is equivalent to*

$$w+d \in D^*N_\Gamma\big(\mathrm{Proj}_\Gamma(u), u-\mathrm{Proj}_\Gamma(u)\big)(-w).$$

Combining the two lemmas above, we arrive at the explicit formula to calculate coderivatives of (minus) projection operators.

Theorem 20 (calculating coderivatives of polyhedral projections). *We have $w \in D^*(-\mathrm{Proj}_\Gamma)(u)(d)$ if and only if $w \in B_{J,K}$ and $w+d \in A_{J,K}$, where all the data are taken from Lemma 18.*

It is computationally convenient to extract a *linear mapping* from the above coderivative descriptions. It can be done, e.g., by selecting the data as follows:

$$J = K = I\big(\mathrm{Proj}_\Gamma(u)\big) = \big\{i \in S \;\big|\; \langle c_i, \mathrm{Proj}_\Gamma(u)\rangle = d_i\big\}.$$

Then we get from (53) the description

$$\begin{cases} w \in B_{J,J} = \big\{x \;\big|\; \langle a_i, x\rangle = 0,\ i \in R;\ \langle c_i, x\rangle = 0,\ i \in J\big\}, \\ w+d \in A_{J,J} = \mathrm{span}\big\{a_i,\ i \in R\big\} + \mathrm{span}\big\{c_i,\ i \in J\big\} = B^\perp_{J,J}, \end{cases}$$

which is equivalent to the simple inclusions

$$d \in -w + B^\perp_{J,J}, \quad -w \in B_{J,J}.$$

Therefore, we arrive at the projection expression

$$-w = \mathrm{Proj}_{B_{J,J}}(d) \Longleftrightarrow w = -\mathrm{Proj}_{B_{J,J}}(d),$$

which finally brings us to the exact explicit formula

$$w = -(I-B^\dagger B)d, \;\text{ where }\; B := \begin{pmatrix} c_i^T, & i \in J = I\big(\mathrm{Proj}_\Gamma(u)\big) \\ a_i^T, & i \in R \end{pmatrix}$$

with $B^\dagger$ standing for the Moore-Penrose inverse of the matrix B.

Now we examine computing the coderivative of the projector to a *specific convex polyhedron*, the intersection of a hyperplane and a box, given by

$$\begin{aligned} \Gamma :=& \big\{x \in \mathbb{R}^n \;\big|\; \langle a, x\rangle = b,\ l \le x \le L\big\} \\ =& \big\{x \in \mathbb{R}^n \;\big|\; a^T x = b,\ Cx \le d\big\}, \end{aligned} \tag{54}$$

where $a \in \mathbb{R}^n$ and $l, L \in \overline{\mathbb{R}}^n$ represent lower and upper bounds (which can be infinite), and where

$$C := \begin{pmatrix} I_n \\ -I_n \end{pmatrix}, \quad d = \begin{pmatrix} L \\ -l \end{pmatrix}.$$

Without loss of generality, assume that $l_i < L_i$ for all $1 \le i \le n$. According to the above, we take $B := \begin{pmatrix} C_J \\ a^T \end{pmatrix}$, where $C_J \in \mathbb{R}^{|J|\times n}$ is the matrix formed by the rows corresponding to the index set $J \subset [2n]$ defined by

$$J := I\big(\mathrm{Proj}_\Gamma(u)\big) = \big\{i \;\big|\; \mathrm{Proj}_\Gamma(u)_i = L_i\big\} \cup \big\{n+i \;\big|\; \mathrm{Proj}_\Gamma(u)_i = l_i\big\}.$$

Proposition 21 (coderivative calculations for specific polyhedra). *Let Γ be given in (54), and let the matrix P (depending on $u \in \mathbb{R}^n$) be defined by*

$$P := I - B^\dagger B = \Sigma - \Sigma a(a^T\Sigma a)^\dagger a^T\Sigma = \begin{cases} \Sigma & \text{if} \quad \Sigma a = 0, \\ \Sigma - \|\Sigma a\|^{-2}\Sigma a a^T\Sigma & \text{if} \quad \Sigma a \ne 0, \end{cases}$$

where $\Sigma := I - \Theta$, and where $\Theta := \mathrm{diag}(\theta)$ with

$$\theta_i := \begin{cases} 0 \;\text{ if }\; l_i < \mathrm{Proj}_\Gamma(u)_i < L_i, \\ 1 \;\text{ otherwise}. \end{cases}$$

Then we have the inclusion $-Pd \subset D^*(-\mathrm{Proj}_\Gamma)(u)(d)$.

Proof. This follows from Theorem 20 by calculating the dagger matrix $B^{\dagger}$. □

By choosing the regularization matrices as $B_k = R := I - \gamma Q \succ 0$, we get the following Newton system in our algorithm CNFB:

$$\big((1+\mu_k)I - PR\big)d = \mathrm{Proj}_{\Gamma}(u^k) - x^k, \ \text{ where } \ u^k := x^k - \gamma(Qx^k + c).$$

Keeping in mind the structure of P, the Newton system in CNFB can be solved as follows:

• If $\Sigma a = 0$, then we have

$$\begin{cases} (1+\mu_k)d_\delta = (v^k - x^k)_\delta, \\ (1+\mu_k)I_{\bar\delta} - R_{\bar\delta\times\bar\delta}d_{\bar\delta} = (v^k - x^k)_{\bar\delta} + R_{\bar\delta\times\delta}d_\delta, \end{cases} \tag{55}$$

where $v^k := \mathrm{Proj}_{\Gamma}(u^k)$, $\delta := \{i \mid v_i^k = l_i \ \text{ or } \ v_i^k = L_i\}$, and $\bar\delta = [n] - \delta = \{i \mid l_i < v_i^k < L_i\}$.

• If $\Sigma a \neq 0$, then we have

$$\begin{cases} (1+\mu_k)d_\delta = (v^k - x^k)_\delta, \\ (1+\mu_k)I_{\bar\delta} - P_{\bar\delta\times\bar\delta}R_{\bar\delta\times\bar\delta})d_{\bar\delta} = (v^k - x^k)_{\bar\delta} + P_{\bar\delta\times\bar\delta}R_{\bar\delta\times\delta}d_\delta. \end{cases} \tag{56}$$

8.3 Numerical Results of CNFB for SLBQP

Here we compare our algorithm CNFB for SLBQP with the following algorithms:

(1) Gurobi, a commercial software for QPs (which implements a highly optimized IPM).

(2) MATLAB's 'quadprog' solver for QPs.

(3) QPPAL [33], a two-phase proximal augmented Lagrangian method for QPs.

(4) P2GP [10], a two-phase gradient method specialized for the class of SLBQPs.

The positive-semidefinite symmetric matrix $Q \in \mathbb{R}^{n\times n}$ in our experiments is generated as $Q = C^T C$, where $C \in \mathbb{R}^{r\times n}$ ($r \le n$) is a random matrix with i.i.d. Gaussian entries. Thus the rank of Q is expected to be r. The results are reported in Table 3. It can be observed that in the high rank case, P2GP is the most efficient algorithm followed by CNFB while Gurobi, MATLAB, QPPAL are less efficient. In the low-rank case, Gurobi has the best performance followed by MATLAB and CNFB, while P2GP and QPPAL are less efficient. Therefore, our proposed CNFB is the only algorithm that is competitive regardless of the rank of the matrix Q in the objective function.

9 Applications of CNAL to Lasso Problems

This section presents some implementations of the proposed CNAL to the class of Lasso problems. We confine our attention to the basic Lasso problem introduced in [57] as the *ℓ_1-regularized least squares regression model* formulated as

$$\text{minimize } \ \varphi(x) := \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \ \text{ over } \ x \in \mathbb{R}^n, \tag{57}$$

where $A \in \mathbb{R}^{m\times n}$ is the data matrix (with m being the number of samples and n being the number of features), $b \in \mathbb{R}^m$, $\lambda > 0$, and $\|\cdot\|_1$, $\|\cdot\|_2$ are the ℓ_1-norm and ℓ_2-norm, respectively. We see that the Lasso problem (57) is a special case of (44) with $h(Ax) = \frac{1}{2}\|Ax\|^2$, $c = A^*b$, $p(x) = \lambda\|x\|_1$, $h^*(y) = \frac{1}{2}\|y\|^2$, and $p^*(z) = \delta_{B_\infty(\lambda)}$, where $B_\infty(\lambda) := \{z \in \mathbb{R}^n \mid \|z\|_\infty \le \lambda\}$ is the closed ball with radius λ in the ℓ_∞-norm.

The proximal mapping of σp is the *soft-thresholding operator*

$$[\mathrm{Prox}_{\sigma p}(u)]_i = \mathrm{sign}(u_i)(|u_i| - \sigma\lambda)_+ \ \text{ for } \ i = 1,\ldots,n,$$

where $(|u_i| - \sigma)_+ := \max\{0, |u_i| - \sigma\}$.

In order to apply CNAL to the Lasso problem (57), we need to compute the coderivative of $\mathrm{Prox}_{\sigma p}$. Since $\mathrm{Prox}_{\sigma p}$ is separable and piecewise linear, its coderivative is easily calculated by

$$\big(D^*\mathrm{Prox}_{\sigma p})\big)(u)(d) = \big\{w \in \mathbb{R}^n \ \big|\ w_i \in G_i(u_i, d_i)\big\},$$

Table 3: Solving random SLBQPs

rank	size	measure	CNFB	P2GP	MATLAB	Gurobi	QPPAL
$r=n$	1000	time	0.31s	0.24s	0.79s	0.72s	5.30s
		residual	7.38e-13	4.16e-09	3.17e-08	2.94e-06	3.79e-06
	2000	time	0.91s	0.28s	4.70s	3.16s	16.06s
		residual	2.25e-12	9.86e-09	6.12e-07	2.39e-05	7.11e-07
	5000	time	7.67s	1.55s	47.92s	39.98s	208.36s
		residual	7.86e-12	3.68e-10	1.01e-06	1.46e-04	1.91e-04
$r=0.9n$	1000	time	0.26s	0.21s	0.78s	0.67s	5.64s
		residual	8.39e-13	5.43e-09	7.40e-10	8.54e-05	1.27e-06
	2000	time	0.99s	0.35s	4.67s	3.18s	16.13s
		residual	1.93e-12	9.08e-09	4.43e-06	9.28e-05	1.20e-06
	5000	time	9.20s	1.44s	50.11s	41.90s	424.25s
		residual	7.90e-12	3.62e-08	1.30e-05	8.97e-05	1.73e-05
$r=0.5n$	1000	time	0.84s	0.55s	0.73s	0.43s	5.10s
		residual	6.85e-13	2.09e-09	4.17e-06	3.34e-05	1.60e-05
	2000	time	4.56s	2.07s	4.18s	1.53s	24.89s
		residual	1.92e-12	8.58e-09	1.64e-07	7.23e-06	3.32e-05
	5000	time	36.42s	12.80s	39.57s	13.40s	1010.14s
		residual	6.34e-12	2.58e-08	1.45e-05	1.42e-04	1.97e-05
$r=0.1n$	1000	time	2.75s	17.10s	0.71s	0.16s	6.76s
		residual	1.69e-09	1.19e+00	3.10e-09	2.11e-11	2.91e-04
	2000	time	8.76s	27.08s	5.20s	0.47s	70.24s
		residual	4.22e-08	2.77e+00	2.70e-06	2.43e-08	1.58e-05
	5000	time	52.85s	120.59s	51.60s	3.07s	1710.12s
		residual	1.18e-08	1.86e-03	1.22e-05	2.13e-05	3.63e-05

where the set-valued mapping $G:\mathbb{R}\times\mathbb{R}\rightrightarrows\mathbb{R}$ is defined by

$$G_i(u_i,d_i):=\begin{cases} d_i & \text{if} \quad |u_i|>\sigma\lambda,\\ \{0,d_i\} & \text{if} \quad u_i=\sigma\lambda, d_i>0,\\ [0,d_i] & \text{if} \quad u_i=\sigma\lambda, d_i\le 0,\\ 0 & \text{if} \quad |u_i|<\sigma\lambda,\\ [0,d_i] & \text{if} \quad u_i=-\sigma\lambda, d_i\ge 0,\\ \{0,d_i\} & \text{if} \quad u_i=-\sigma\lambda, d_i<0. \end{cases}$$

We can directly extract a linear mapping from $(D^*\mathrm{Prox}_{\sigma p})(u)$ as follows: for all $d\in\mathbb{R}^n$, take $Pd\subset D^*(\mathrm{Prox}_{\sigma p})(u)(d)$, where $P\in\mathbb{R}^{n\times n}$ is a diagonal matrix with the entries

$$P_{ii}:=\begin{cases} 1 & \text{if} \quad |u_i|>\sigma\lambda,\\ 0 & \text{if} \quad |u_i|\le\sigma\lambda. \end{cases}$$

Note that for the Lasso problem under consideration, the proposed CNAL is closely related to SSNAL designed to solve Lasso models in [32], because the Newton systems in both algorithms are identical in this case with the major difference in the linesearch strategy. CNAL uses the Wolfe linesearch for the coderivative Newton subproblem solver GRNM-W while SSNAL employs the backtracking Armijo linesearch for the semismooth Newton subproblem solver as stated in [32]. In the numerical implementations, our codes are adapted from SuiteLasso, with the major change being a *different Wolfe linesearch* that appears to be more efficient, at least for random instances.

9.1 Numerical Results of CNAL for Lasso

Here we present the results of numerical experiments to compare our algorithm CNAL with the following well-known first-order and second-order algorithms to solve Lasso problems:

(1) The *Semismooth Newton Augmented Lagrangian Method* (SSNAL)[2] [31], one of the most efficient methods for Lasso.

(2) The *Matrix-Free Interior Point Method* (mfIPM)[3] [14], an interior point method that is highly optimized for Lasso.

(3) The (Nesterov) *Accelerated Proximal Gradient Method* (APG) [46] and the closely related in this case *Fast Iterative Shrinkage-Thresholing Algorithm* (FISTA)[4] [3], a simple and efficient first-order method.

(4) The *Alternating Direction Method of Multipliers* (ADMM)[5] [15, 20], a classical and popular primal-dual splitting method.

We tested these algorithms on the standard data set LIBSVM [6]. The parameter λ in (57) is chosen to be $\lambda = \lambda_c \|A^T b\|_\infty$ where $\lambda_c = 10^{-3}$. The relative KKT residual

$$\eta = \frac{\|\tilde{x} - \mathrm{Prox}_{\lambda\|\cdot\|_1}(\tilde{x} - A^T(A\tilde{x} - b))\|}{1 + \|\tilde{x}\| + \|A^T(A\tilde{x} - b)\|}$$

is used to measure the accuracy of an approximate minimizer $\tilde{x}$ of (57). The results are reported in Table 4. In the table, 'NA' means 'not applicable' (it appears because mfIPM is not applicable by design in the $m > n$ case). We can observe that CNAL is as efficient as SSNAL and both algorithms are highly accurate and much faster than the others in comparison.

Table 4: Solving Lasso problems in LIBSVM

Problem	Measure	CNAL	SSNAL	mfIPM	APG	ADMM
covtype	time	0.01s	0.01s	NA	70.49s	11.95s
581012;54	residual	2.17e-07	2.18e-07	NA	2.75e-05	4.26e-04
YearPredictionMSD	time	0.01s	0.01s	NA	326.84s	55.99s
463715;90	residual	2.48e-07	3.74e-07	NA	5.66e-04	6.18e-04
E2006.test	time	0.14s	0.14s	2.04s	0.10s	21.85s
3308;72812	residual	8.25e-07	1.60e-07	4.62e-09	2.97e-07	4.68e-07
E2006.train	time	0.40s	0.38s	5.84s	0.36s	187.12s
16087;150348	residual	8.51e-07	1.65e-07	5.65e-09	2.94e-06	8.54e-08
news20_tfidf_test	time	0.34s	0.33s	14.71s	21.02s	60.41s
7532;49909	residual	5.02e-07	6.77e-07	1.69e-05	3.01e-06	8.03e-07
housing_expanded7	time	1.91s	1.93s	249.77s	114.20s	104.52s
506;77520	residual	8.83e-07	8.83e-07	2.97e-01	7.57e-04	1.78e-04
pyrim_expanded5	time	1.14s	1.15s	875.85	63.90s	156.27s
74;169911	residual	9.37e-07	9.37e-07	4.46e-07	2.57e-03	3.50e-04

10 Conclusions

This paper proposes the globally convergent coderivative-based generalized regularized Newton method with the Wolfe linesearch (GRNM-W) to solve $C^{1,1}$ optimization problems. The Newton directions are found by solving linear equations extracted from coderivatives. The local convergence rate of GRNM-W is at least linear and becomes superlinear under the semismooth* property of the gradient mapping. We also presented a modified version of GRNM-W that is applicable to arbitrary nonconvex functions. Under the imposed PLK properties of the objective function, we proved convergence and convergence rate results for the modified GRNM-W. We further combined the coderivative-based Newton method GRNM-W with

[2] https://github.com/MatOpt/SuiteLasso

[3] http://www.maths.ed.ac.uk/ERGO/mfipmcs/

[4] We use the implementation in SLEP: http://yelabs.net/software/SLEP/

[5] We use the implementation at https://web.stanford.edu/~boyd/papers/admm/lasso/lasso.html

the forward-backward envelope and the augmented Lagrangian method while proposing the two algorithms CNFB and CNAL for solving convex composite minimization problems that are first-order nonsmooth and constrained. Numerical experiments on support vector machines (formulated as special quadratic programming problems) and the Lasso problem indicate that both CNFB and CNAL are efficient, which confirms the effectiveness of the main novel algorithmic scheme GRNM-W. Our future research includes further applications of the proposed nonsmooth Newton algorithms to a large variety of optimization problems arising in machine learning, data science, statistics, and other fields.

Acknowledgments Acknowledgments are not compulsory. Grant or contribution numbers may be acknowledged. Please refer to the journal-level guidance for any specific requirements.

Data availability The data that supports the findings of this study are available from the corresponding author upon request.

Code availability The code that supports the findings of this study is available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

[1] F. J. Aragón-Artacho, B. S. Mordukhovich and P. Peréz-Aros, Coderivative-based semi-Newton method in nonsmooth difference programming, *Math. Program.*, https://doi.org/10.1007/s10107-024-02142-8, 2024.

[2] H. Attouch, J. Bolté and B. F. Svaiter, Convergence of descent methods for semialgebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods, *Math. Program.*, 137:91–129, 2013.

[3] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.*, 2:183–202, 2009.

[4] G. C. Bento, B. S. Mordukhovich, T. S. Mota and Yu. Nesterov, Convergence of descent methods under Kurdyka-Łojasiewicz properties, arXiv:5701197, 2024.

[5] J. Bolté, A. Daniilidis, A. S. Lewis and M. Shiota, Clarke subgradients of stratifiable functions, *SIAM J. Optim.*, 18:556–572, 2007.

[6] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, *ACM Trans. Intel. Syst. Tech. (TIST)*, 2:1–27, 2011.

[7] N. H. Chieu, T. D. Chuong, J.-C. Yao and N. D. Yen, Characterizing convexity of a function by its Fréchet and limiting second-order subdifferentials, *Set-Valued Var. Anal.*, 19:75–96, 2011.

[8] N. H. Chieu, G. M. Lee and N. D. Yen, Second-order subdifferentials and optimality conditions for C^1-smooth optimization problems, *Appl. Anal. Optim*, 1:461–476, 2017.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, PA, 1990.

[10] D. Di Serafino, G. Toraldo, M. Viola and J. Barlow, A two-phase gradient method for quadratic programming problems with a single linear constraint and bounds on the variables, *SIAM J. Optim.*, 28:2809–2838, 2018.

[11] N. Doikov and Yu. Nesterov, Gradient regularization of Newton method with Bregman distances, *Math. Program*, 204:1–25, 2024.

[12] D. Drusvyatskiy, B. S. Mordukhovich and T. T. A. Nghia, Second-order growth, tilt stability, and metric regularity of the subdifferential, *J. Convex Anal.*, 21:1165–1192, 2014.

[13] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2003.

[14] K. Fountoulakis, J. Gondzio and P. Zhlobich, Matrix-free interior point method for compressed sensing problems, *Math. Program. Comput.*, 6:61–31, 2014.

[15] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. Appl.*, 2:17–40, 1976.

[16] H. Gfrerer, On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs, *Set-Valued Var. Anal.*, 21:151–176, 2013.

[17] H. Gfrerer and J. V. Outrata, On a semismooth* Newton method for solving generalized equations, *SIAM J. Optim.*, 31:489–517, 2021.

[18] H. Gfrerer and J. V. Outrata, On (local) analysis of multifunctions via subspaces contained in graphs of generalized derivatives, *J. Math. Anal. Appl.*, 508:125895, 2022.

[19] I. Ginchev and B. S. Mordukhovich, On directionally dependent subdifferentials, *C. R. Acad. Bulg. Sci.*, 64:497–508, 2011.

[20] R. Glowinski and A. Marroco, Sur lápproximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de Dirichlet non linéaires, *Rev. Franc. d'Automat. Infor. Rech. Opérat. Analyse Numér.*, 9:41–76, 1975.

[21] R. Henrion, B. S. Mordukhovich and N. M. Nam, Second-order analysis of polyhedral systems in finite and infinite dimensions with applications to robust stability of variational inequalities, *SIAM J. Optim.*, 20:2199–2227, 2010.

[22] A. F. Izmailov and M. V. Solodov, *Newton-Type Methods for Optimization and Variational Problems*, Springer, Berlin, 2014.

[23] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Applications to R*, 2nd edition, Springer, New York, 2013.

[24] H. Karimi, J. Nutini and M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition, *Machine Learning and Knowledge Discovery in Databases, Part 1*, pp. 795–811, Springer, Cham, Switzerland, 2016.

[25] P. D. Khanh, B. S. Mordukhovich and V. T. Phat, A generalized Newton method for subgradient systems, *Math. Oper. Res.*, 48:1811–1845, 2023.

[26] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran, Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials, *J. Global Optim.*, 86:93–122, 2023.

[27] P. D. Khanh, B. S. Mordukhovich, V. T. Phat and D. B. Tran, Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization, *Math. Program.*, 205:373–429, 2024.

[28] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization: Regularity, Calculus, and Application*, Kluwer, Boston, 2002.

[29] B. Kummer, Newton's method for nondifferentiable functions, in *Advances in Mathematical Optimization*, pp. 114–124, Akademi-Verlag, Berlin, 1988.

[30] K. Kurdyka, On gradients of functions definable in o-minimal structures, *Ann. Inst. Fourier*, 48:769–783, 1998.

[31] X. Li, D. Sun and K.-C. Toh, A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems, *SIAM J. Optim.*, 28:433–458, 2018.

[32] X. Li, D. Sun and K.-C. Toh, On efficiently solving the subproblems of a level-set method for fused Lasso problems, *SIAM J. Optim.*, 28:1842–1866, 2018.

[33] L. Liang, X. Li, D. Sun and K.-C. Toh, QPPAL: A two-phase proximal augmented Lagrangian method for high-dimensional convex quadratic programming problems, *ACM Trans. Math. Softw.* (*TOMS*), 48:1–27, 2022.

[34] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, *Coll. du CNRS, Les équations aux dérivées partielles*, 87–89, 1963.

[35] R. Mifflin, Semismooth and semiconvex functions in constrained optimization, *SIAM J. Control Optim.*, 15:959–972, 1977.

[36] A. Mohammadi, B. S. Mordukhovich ans M. E. Sarabi, Parabolic regularity in geometric variational analysis, *Trans. Amer. Math. Soc.*, 374: 1711–1763, 2021.

[37] A. Mohammadi, B. S. Mordukhovich and M. E. Sarabi, Variational analysis of composite models with applications to continuous optimization, *Math. Oper. Res.*, 47: 397–426, 2022.

[38] A. Mohammadi and M. E. Sarabi, Twice epi-differentiability of extended-real-valued functions with applications in composite optimization, *SIAM J. Optim.*, 30: 2379–2409, 2020.

[39] B. S. Mordukhovich, Sensitivity analysis in nonsmooth optimization, in *Theoretical Aspects of Industrial Design* (D. A. Field and V. Komkov, eds.), *SIAM Proc.* 58:32–46, 1992.

[40] B. S. Mordukhovich, Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions, *Trans. Amer. Math. Soc.*, 340:1–35, 1993.

[41] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation, I: Basic Theory, II: Applications*, Springer, Berlin, 2006.

[42] B. S. Mordukhovich, *Variational Analysis and Applications*, Springer, Cham, Switzerland, 2018.

[43] B. S. Mordukhovich, *Second-Order Variational Analysis in Optimization, Variational Stability, and Control: Theory, Algorithms, Applications*, Springer, Cham, Switzerland, 2024.

[44] B. S. Mordukhovich and T. T. A. Nghia, Second-order characterizations of tilt stability with applications to nonlinear programming, *Math. Program.*, 149:83–104, 2015.
[45] B. S. Mordukhovich and M. E. Sarabi, Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization, *SIAM J. Optim.*, 31:1184–1214, 2021.
[46] Yu. Nesterov, A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$, *Dokl. Akad. Nauk SSSR*, 269:543, 1983.
[47] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[48] Y. Qian and S. Pan, A superlinear convergence framework for Kurdyka-Łojasiewicz optimization, arXiv:2210.12449, 2023.
[49] P. Patrinos and A. Bemporad, Proximal Newton methods for convex composite optimization, in the *52nd IEEE Conference on Decision and Control*, pp. 2358–2363. IEEE, 2013.
[50] R. A. Poliquin and R. T. Rockafellar, Tilt stability of a local minimum, *SIAM J. Optim.*, 8:287–299, 1998.
[51] B. T. Polyak, Gradient methods for the minimization of functionals, *USSR Comput. Math. Math. Phys.* 3: 864–878, 1963.
[52] R. A. Polyak, Regularized Newton method for unconstrained convex optimization, *Math. Program.*, 120:125–145, 2009.
[53] L. Qi and J. Sun, A nonsmooth version of Newton's method, *Math. Program.*, 58:353–367, 1993.
[54] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Math. Oper. Res.*, 1:97–116, 1976.
[55] R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer, Berlin, 1998.
[56] L. Stella, A. Themelis and P. Patrinos, Forward–backward quasi-Newton methods for nonsmooth optimization problems, *Comput. Optim. Appl.*, 67:443–487, 2017.
[57] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. Royal Stat. Soc, Ser, B: Statistical Methodology*, 58:267–288, 1996.