LINE SEARCH METHODS FOR CONVEX-COMPOSITE OPTIMIZATION

J. V. BURKE AND A. ENGLE

ABSTRACT. We consider descent methods for solving infinite-valued nonsmooth convex-composite optimization problems that employ search directions derived from Gauss-Newton subproblems. The descent algorithms are based on either a weak Wolfe or a backtracking line search using a continuous approximation to the directional derivative that exploits the structure associated with convex-composite problems.

1. Introduction

This work considers descent methods for solving the convex-composite optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) := h(c(x)) + g(x),$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is convex, $g : \mathbb{R}^n \to \mathbb{R}$ is proper, strictly continuous relative to its domain and convex, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is $C^1$-smooth. Our focus is on methods that employ search directions $d^k \in \mathbb{R}^n$ that approximate solutions to Gauss-Newton subproblems of the form

$$\min_{d \in \mathbb{R}^n} \Delta f(x^k; d)$$

subject to $\|d\| \leq \eta_k,$

where $\{x^k\} \subset \text{dom } (g)$ are the iterates generated by the algorithm, $\{\eta_k\} \subset (0, \infty],$ and $\Delta f(x; d)$ (see (4)) is an approximation to the directional derivative $f'(x; d)$ as in [2]. By descent, we mean that the search direction $d^k$ satisfies $\Delta f(x^k; d^k) < 0$ at each iteration $k$. The two descent methods studied are a weak Wolfe and a backtracking line search employed at each iteration $k$. Algorithms for the problem $P$ have recently received renewed interest due to numerous modern applications in machine learning and nonlinear dynamics [1,8–11].

In past work, the backtracking line search was studied in the absence of the function $g$. That is, for finite-valued convex-composite problems. In recent work, Lewis and Wright [14] utilized the backtracking line search in the context of prox-regular composite optimization. Lewis and Overton [13] developed a weak Wolfe algorithm using directional derivatives for finite-valued nonsmooth functions $f$ that are absolutely continuous along the line segment of interest, with finite termination whenever the function $f$ is semi-algebraic. Their theory can be applied in the finite-valued convex-composite case where $g = 0$. Here, we develop a weak Wolfe algorithm for infinite-valued problems that uses an approximation...
to the directional derivative that exploits the structure associated with convex-composite problems.

The function $g$ in $\mathcal{P}$ is typically nonsmooth and is used to induce structure in the solution $\overline{x}$. For example, it can be used to introduce constraints or sparsity in the solution $\overline{x}$. Drusvyatskiy and Lewis [9] have established local and global convergence of proximal-based methods for solving $\mathcal{P}$, and Drusvyatskiy and Paquette [10] have established iteration complexity results for proximal methods to locate first-order stationary points for $\mathcal{P}$. Both works utilized similar assumptions on the functions $h$, $c$, and $g$.

While our assumptions are similar to [9, 10], our algorithmic approach differs significantly. In particular, we use either adaptive weak Wolfe or backtracking line search techniques to induce objective function descent at each iteration. In addition, we do not make use of proximal methods to generate search directions or employ the backtracking line search to estimate Lipschitz constants as in [10, 14]. Moreover, both line searches presented here make explicit use of the structure in $\mathcal{P}$, thereby differing from the method developed in [13].

2. Notation

This section records notation and tools from convex and variational analysis used throughout the paper. Unless otherwise stated, we follow the notation of [5, 15, 16, 20].

For any two points $x, x' \in \mathbb{R}^n$, denote the line segment connecting $x$ and $x'$ by $[x, x'] := \{(1 - \lambda)x + \lambda x' \mid 0 \leq \lambda \leq 1\}$. For a nonempty closed convex set $C \subset \mathbb{R}^m$ let $\text{aff} C$ denote its affine hull. Then the relative interior of $C$ is

$$\text{ri}(C) = \{x \in \text{aff} C \mid \exists (\epsilon > 0) (x + \epsilon B) \cap \text{aff} C \subset C\}.$$ 

The functions in this paper take values in the extended reals $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm \infty\}$. For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the domain of $f$ is $\text{dom}(f) := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$, and the epigraph of $f$ is $\text{epi} f := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$.

A function $f$ is closed if the level sets $\text{lev}_f(\alpha) := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ are closed for all $\alpha \in \mathbb{R}$, proper if $\text{dom}(f) \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$, and convex if $\text{epi} f$ is a convex subset of $\mathbb{R}^{n+1}$. For a set $X \subset \text{dom}(f)$ and $\overline{x} \in X$, the function $f$ is strictly continuous at $\overline{x}$ relative to $X$ if

$$\limsup_{x \to \overline{x}, x' \to \overline{x} \atop x \neq x'} \frac{\|f(x) - f(x')\|}{\|x - x'\|} < \infty,$$

where $x, x \to \overline{x} \iff x, x' \in X$ and $x, x' \to \overline{x}$ represents convergence within $X$. This finiteness property is equivalent to $f$ being locally Lipschitz at $\overline{x}$ relative to $X$ (see [16, Section 9.A]). By [15, Theorem 10.4], proper and convex functions $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ are strictly continuous relative to $\text{ri}(\text{dom}(g))$. To each nonempty closed convex set $C$, we associate the closed, proper, and convex indicator function defined by
Suppose $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is finite at $\bar{x}$ and $w \in \mathbb{R}^n$. The subderivative $\text{df}(\bar{x}) : \mathbb{R}^n \to \overline{\mathbb{R}}$ and one-sided directional derivative $f'(\bar{x}; \cdot)$ at $\bar{x}$ for $w$ are
\[
\text{df}(\bar{x})(w) := \liminf_{t \to 0^+} \frac{f(\bar{x} + tw) - f(\bar{x})}{t}, \quad f'(\bar{x}; w) := \lim_{t \to 0^+} \frac{f(\bar{x} + tw) - f(\bar{x})}{t}.
\]
The structure of $\mathcal{P}$ allows the classical one-sided directional derivative $f'(\bar{x}; \cdot)$ to capture the variational properties of its more general counterpart as discussed in the next section. Results in the following section also require the notion of set convergence from variational analysis, as in [16, Section 4.A]. For a sequence of sets $\{C_n\}_{n \in \mathbb{N}}$, with $C_n \subset \mathbb{R}^m$, the outer and inner limits are defined, respectively, as
\[
\limsup_{n \to \infty} C_n := \left\{ x \mid \exists (\text{infinite } K \subset \mathbb{N}, \ x^k \to \bar{x}) \ \forall (k \in K) \ x^k \in C_k \right\},
\liminf_{n \to \infty} C_n := \left\{ x \mid \exists (n_0 \in \mathbb{N}, \ x^n \to \bar{x}) \ \forall (n \geq n_0) \ x^n \in C_n \right\}.
\]
The sets $C_n$ converge to a set $C$ if the two limits agree and equal $C$:
\[
\limsup_{n \to \infty} C_n = \liminf_{n \to \infty} C_n = C.
\]
With this notion of convergence in mind, we apply it to the epigraphs of a sequence of functions and say that $f^k : \mathbb{R}^m \to \overline{\mathbb{R}}$ epigraphically converge to $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, written $f^k \rightarrow e f$, if and only if $\text{epi} f^k \rightarrow \text{epi} f$ (see [16, Section 7.B]).

3. Convex-Composite Theory

The general convex-composite optimization problem [3] is of the form
\[
\min_{x \in \mathbb{R}^n} \ f(x) := \psi(\Phi(x)),
\]
where $\psi : \mathbb{R}^m \to \overline{\mathbb{R}}$ is closed, proper and convex, and $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ is sufficiently smooth. Allowing infinite-valued convex functions $\psi$ into the composition presents theoretical difficulties discussed in [5,7,12,14,16]. In this work, we assume $f$ takes the form given in $\mathcal{P}$ by setting $\psi(y, x) := h(y) + g(x)$ and $\Phi(x) = (c(x), x)$. In this case, the calculus simplifies dramatically. As in [3], we have $\text{dom} \ (f) = \text{dom} \ (g)$ and
\[
f(x + d) = h(c(x) + \nabla c(x)d) + g(x + d) + o(||d||).
\]
Consequently, at any $x \in \text{dom} \ (g)$ and $d \in \mathbb{R}^n$, $f$ is directionally differentiable, with
\[
\text{df}(x)(d) = f'(x; d) = h'(c(x); \nabla c(x)d) + g'(x; d).
\]
This motivates defining the subdifferential of $f$ at any $x \in \text{dom} \ (g)$ by setting
\[
\partial f(x) := \nabla c(x)^\top \partial h(c(x)) + \partial g(x).
\]
Within the context of variational analysis [16], we have that $f$ is subdifferentially regular on its domain and the subdifferential of $f$ as defined above agrees with the regular and limiting subdifferentials of variational analysis. In particular, $f'(x;d) = \sup_{v \in \partial f(x)} \langle v, d \rangle$.

Following [2], we define an approximation to the directional derivative that is key to our algorithmic development.

**Definition 3.1.** Let $f$ be as in $\mathcal{P}$ and $x \in \text{dom } (g)$. Define $\Delta f(x;\cdot) : \mathbb{R}^n \to \overline{\mathbb{R}}$ by

$$\Delta f(x;d) = h(c(x) + \nabla c(x)d) + g(x + d) - h(c(x)) - g(x).$$

The next lemma records the interplay between $\Delta f(x;d)$ and its infinitesimal counterpart $f'(x;d)$ and is a consequence of (2) and the definitions.

**Lemma 3.1.** Let $f$ be given as in $\mathcal{P}$ and let $x \in \text{dom } (g)$. Then

(a) the function $d \mapsto \Delta f(x;d)$ is convex;

(b) for any $d \in \mathbb{R}^n$, the difference quotients $\frac{\Delta f(x;td)}{t}$ are nondecreasing in $t > 0$, with

$$f'(x;d) = \inf_{t>0} \frac{\Delta f(x;td)}{t},$$

and in particular;

(c) for any $d \in \mathbb{R}^n$, $f'(x;d) \leq \Delta f(x;d)$;

(d) for any $d \in \mathbb{R}^n$, $t \in [0, 1]$, $\Delta f(x;td) \leq t\Delta f(x;d)$.

We now state equivalent first-order necessary conditions for a local minimizer $\overline{x}$ of $\mathcal{P}$, emphasizing that $f'(x;d)$ and $\Delta f(x;d)$ are interchangeable with respect to these conditions. The proof of this result parallels that given in [3] using (2) and (3).

**Theorem 3.1** (First-order necessary conditions for $\mathcal{P}$). [3, Theorem 2.6] Let $h$, $c$, and $g$ be as given in $\mathcal{P}$. If $\overline{x} \in \text{dom } (g)$ is a local minimizer of $\mathcal{P}$, then $f'(\overline{x};d) \geq 0$, for all $d \in \mathbb{R}^n$. Moreover, the following conditions are equivalent for any $x \in \text{dom } (g)$,

(a) $0 \in \partial f(x)$;

(b) for all $d \in \mathbb{R}^n$, $0 \leq f'(x;d)$;

(c) for all $d \in \mathbb{R}^n$, $0 \leq \Delta f(x;d)$;

(d) for all $\eta > 0$, $d = 0$ solves $\min \{\Delta f(x;d) \mid \|d\| \leq \eta\}$.

The next lemma shows that if the sequence $\{(x^k, d^k)\} \subset \mathbb{R}^n \times \mathbb{R}^n$ is such that $d^k$ is an approximate solution to $\mathcal{P}_k$ for all $k$ with $\Delta f(x^k;d^k) \to 0$, then cluster points of $\{x^k\}$ are first-order stationary for $\mathcal{P}$. 

**Lemma 3.2.** Let $h$, $c$, and $g$ be as in $\mathcal{P}$ and $\alpha \in \mathbb{R}$. Set $\mathcal{L} := \text{lev } f(\alpha)$. Let $\{x^k, \eta_k\} \subset \mathcal{L} \times \mathbb{R}_+$, with $(x^k, \eta_k) \to (\overline{x}, \overline{\eta}) \in \mathbb{R}^n \times \mathbb{R}_+$ and $0 < \overline{\eta} < \infty$. Define

$$\Delta_k f(d) := \Delta f(x^k;d) + \delta_{\eta_k} B(d), \quad \text{and}$$

$$\Delta_k f := \min_d \Delta_k f(d).$$

(5)
If, for each \( k \geq 1 \), \( d^k \in \eta_k \mathcal{B} \) satisfies
\[
\Delta f(x^k; d^k) \leq \beta \bar{\Delta} f \leq 0,
\]
with \( \Delta f(x^k; d^k) \to 0 \), then \( 0 \in \partial f(\bar{x}) \).

**Proof.** Since \( f \) is closed, \( f(\bar{x}) \leq \alpha \), which implies \( \bar{x} \in \text{dom} \ (g) \). Define the functions
\[
h_k(d) := h(c(x^k) + \nabla c(x^k)d) - h(c(x^k)),
\]
\[
h_\infty(d) := h(c(\bar{x}) + \nabla c(\bar{x})d) - h(c(\bar{x})) ,
\]
\[
g_k(d) := g(x^k + d) - g(x^k), \quad \text{and}
\]
\[
g_\infty(d) := g(\bar{x} + d) - g(\bar{x}).
\]
Since \( 0 < \eta_k \to \bar{\eta} \), with \( \bar{\eta} > 0 \), and since \( \delta \left( d \mid \eta_k \mathcal{B} \right) = \delta \left( \frac{1}{\eta_k} d \mid \mathcal{B} \right) \), [16, Proposition 7.2] implies
\[
\delta \left( \cdot \mid \eta_k \mathcal{B} \right) \xrightarrow{\varepsilon} \delta \left( \cdot \mid \bar{\mathcal{B}} \right).
\]
By [16, Exercise 7.8(d)], \( g_k \xrightarrow{\varepsilon} g_\infty \), so [16, Exercise 7.47] implies \( g_k + \delta \left( \cdot \mid \eta_k \mathcal{B} \right) \xrightarrow{\varepsilon} g_\infty + \delta \left( \cdot \mid \bar{\mathcal{B}} \right) \), and applying [16, Exercise 7.47] again yields
\[
h_k + g_k + \delta \left( \cdot \mid \eta_k \mathcal{B} \right) \xrightarrow{\varepsilon} h_\infty + g_\infty + \delta \left( \cdot \mid \bar{\mathcal{B}} \right).
\]
Equivalently,
\[
\Delta f(x^k; \cdot) + \delta \left( \cdot \mid \eta_k \mathcal{B} \right) \xrightarrow{\varepsilon} \Delta f(\bar{x}; \cdot) + \delta \left( \cdot \mid \bar{\mathcal{B}} \right).
\]
By [16, Proposition 7.30] and (6),
\[
0 = \lim \sup_k \bar{\Delta} f_k \leq \min_{\|d\| \leq \bar{\eta}} \Delta f(\bar{x}; d) \leq 0,
\]
so Theorem 3.1 implies \( 0 \in \partial f(\bar{x}) \). \( \square \)

The approximate solution condition (6) is described in [2]. It can be satisfied by employing the trick described in [4, Remark 6, page 343]. Specifically, any solution technique solving the convex subproblems \( \mathcal{P}_k \) that also generates lower bounds \( \ell_{k,j} \in \mathbb{R} \) such that \( \ell_{k,j} \geq \bar{\Delta} f \) and \( \Delta f(x^k; d^{k,j}) \searrow \bar{\Delta} f \) as \( j \to \infty \). If \( \bar{\Delta} f_k < 0 \), then the condition
\[
\Delta f(x^k; d^{k,j}) \leq \beta \ell_{k,j}
\]
is finitely satisfied, and
\[
\Delta f(x^k; d^{k,j}) \leq \beta \bar{\Delta} f.
\]
We conclude this section with a mean-value theorem for \( \mathcal{P} \).

**Theorem 3.2** (Mean-Value for Convex-Composite). [16, Theorem 10.48] Let \( f \) be as in \( \mathcal{P} \) and \( x_0, x_1 \in \text{dom} \ (g) \). Then there exists \( t \in (0, 1) \), \( x_t := (1-t)x_0 + tx_1 \) and \( v \in \partial f(x_t) \) such that
\[
f(x_1) - f(x_0) = \langle v, x_1 - x_0 \rangle.
\]
Proof. Let $F(t) := (1-t)x_0 + tx_1$ and let $\varphi(t) = f(F(t)) - (1-t)f(x_0) - tf(x_1)$. Then

$$\varphi(t) = h(c(F(t))) + g(F(t)) - (1-t)f(x_0) - tf(x_1)$$

is an instance of $\mathcal{P}$, since $g \circ F$ is convex. Consequently, the chain rules for $\varphi$ and $-\varphi$ on $[0,1]$ are

$$\partial \varphi(t) = F'(t)^\top \nabla c(F(t))^\top \partial h(c(F(t))) + F'(t)^\top \partial g(F(t)) + f(x_0) - f(x_1)$$

$$\partial (-\varphi)(t) = F'(t)^\top \nabla c(F(t))^\top \partial (-h)(c(F(t))) + F'(t)^\top \partial (-g)(F(t)) + f(x_1) - f(x_0)$$

As $g$ is continuous on its domain, $\varphi$ is continuous on $[0,1]$ with $\varphi(0) = \varphi(1) = 0$. Therefore, $\varphi$ attains either its minimum or maximum value at some $\overline{t} \in (0,1)$, and $0 \in \partial \varphi(\overline{t})$ or $0 \in \partial (-\varphi)(\overline{t})$ respectively. \hfill $\Box$

4. Weak Wolfe for Convex-Composite Minimization

For any $x \in \text{dom} \ (g), 0 < \sigma_1 < 1$, and $d \in \mathbb{R}^n$ such that $\Delta f(x; d) < 0$, Lemma 3.1 implies that, for sufficiently small $t > 0$,

$$f(x + td) \leq f(x) + \sigma_1 t \Delta f(x; d),$$

which gives rise to the backtracking line search of [2, 6] based on $\Delta f(x; \cdot)$. This line search is employed in the recent work of [10, 14].

**Definition 4.1.** *Weak Wolfe in the convex composite case is defined at $x \in \text{dom} \ (g)$ with $\Delta f(x; d) < 0$ by choosing $0 < \sigma_1 < \sigma_2 < 1$ and $\mu > 0$ and requiring

$$f(x + td) \leq f(x) + \sigma_1 t \Delta f(x; d), \text{ and}$$

$$\sigma_2 \Delta f(x; d) \leq \frac{\Delta f(x + td; \mu d)}{\mu}. \tag{WWII}$$

**Remark 1.** The first condition (WWI) is sufficient decrease of $f$ along the ray $\{x + td \mid t > 0\}$, with $\Delta f$ acting as a surrogate for the directional derivative. The second condition (WWII) is a curvature condition that parallels the classical weak Wolfe [18, 19] curvature condition for smooth, unconstrained minimization:

$$\sigma_2 f'(x; d) \leq f'(x + td; d),$$

which prevents the line search early termination at “strongly negative” slopes [20, Section 3.1].

**Remark 2.** The strong Wolfe conditions require $|f'(x + td; d)| \leq -\sigma_2 f'(x; d)$, whenever $f$ is smooth. However, in nonsmooth minimization, kinks and upward cusps at local minimizers make this condition unworkable.

The following lemma shows that the set of points satisfying (WWI) and (WWII) has nonempty interior.
Lemma 4.1. Let $f$ be as in $\mathcal{P}$, $x \in \text{dom}(g)$, and $d$ chosen so that $\Delta f(x; d) < 0$. Suppose $f$ is bounded below on the ray $\{x + td : t > 0\}$, and $\mu \in \mathbb{R}$. Then, the set

$$ C(\mu) := \left\{ t > 0 \mid \begin{array}{l} f(x + td) \leq f(x) + \sigma_1 t \Delta f(x; d), \\ \sigma_2 \Delta f(x; d) \leq \frac{\Delta f(x + td; \mu d)}{\mu} \end{array} \right\} $$

has nonempty interior for any $\mu > 0$.

**Proof.** Define

$$ K(y, z, t) := h(y) + g(z) - [f(x) + \sigma_1 t \Delta f(x; d)], $$

$$ G(t) := \begin{pmatrix} c(x + td) \\ x + td \\ t \end{pmatrix}, \quad \text{with} \quad G'(t) = \begin{pmatrix} \nabla c(x + td) \\ d \\ 1 \end{pmatrix}, $$

and set $\phi(t) := K(G(t)) = f(x + td) - [f(x) + \sigma_1 t \Delta f(x; d)]$. Then, $\phi(t)$ is convex-composite,

$$ \Delta \phi(t; \mu) = K(G(t) + G'(t) \mu) - K(G(t)) $$

$$ = h(c(x + td) + \nabla c(x + td) \mu d) + g(x + (t + \mu) d) - [f(x) + \sigma_1 (t + \mu) \Delta f(x; d)] $$

$$ - (h(c(x + td)) + g(x + td) - [f(x) + \sigma_1 t \Delta f(x; d)]) $$

$$ = \Delta f(x + td; \mu d) - \mu \sigma_1 \Delta f(x; d), $$

and, by Lemma 3.1,

$$ \phi'(t; \mu) = f'(x + td; \mu d) - \mu \sigma_1 \Delta f(x; d) $$

$$ \leq \mu \Delta \phi(t; 1). $$

Consequently, $\phi'(0; 1) \leq (1 - \sigma_1) \Delta f(x; d) < 0$, so there exists $\tilde{t} > 0$ such that for all $t \in (0, \tilde{t})$, $\phi(t) < 0$. This is equivalent to (WWI) being satisfied on $(0, \tilde{t})$.

Since $\phi$ is bounded below on the ray, $\phi(t) \nearrow \infty$. Let $\tilde{t} := \sup \{ t > \tilde{t} : \phi(s) < 0 \text{ for all } s \in (0, t) \}$. Then, since $g$ is closed and $h$ is finite-valued, $\phi(\tilde{t}) = \liminf_{t \nearrow \tilde{t}} \phi(t)$, which implies

$$ h(c(x + \tilde{t}d)) + g(x + \tilde{t}d) = -[f(x) + \sigma_1 \tilde{t} \Delta f(x; d)] + \liminf_{t \nearrow \tilde{t}} \phi(t) $$

$$ \leq -[f(x) + \sigma_1 \tilde{t} \Delta f(x; d)] < \infty, $$

so $x + \tilde{t}d \in \text{dom}(g)$. Since $g$ is continuous relative to its domain, $\phi$ is continuous relative to its domain, so $\phi(\tilde{t}) \leq 0$. We now consider two cases on the value of $\phi(\tilde{t})$.

Suppose $\phi(\tilde{t}) < 0$. We aim to show that $f$ satisfies (WWI) and (WWII) on the interval $((\tilde{t} - \mu)_+, \tilde{t})$. To prove this, we show that if $\phi(\tilde{t}) < 0$, then $t > \tilde{t}$ implies $x + td \not\in \text{dom}(g)$ and, as a consequence, $\tilde{t} \geq 1$. Suppose to the contrary that there exists $t > \tilde{t}$ with $x + td \in \text{dom}(g)$. Then, the definition of $\tilde{t}$, convexity of dom $(\phi)$, and the intermediate value theorem imply there exists $\tilde{t}$ such that $\phi(\tilde{t}) = 0$ and $t \geq \tilde{t} > \tilde{t}$. But relative continuity of $\phi$ at $\tilde{t}$ with respect to dom $(\phi)$ means there exist points in $(\tilde{t}, \tilde{t})$ which contradict the definition of $\tilde{t}$. This proves the claim. Consequently, if $\phi(\tilde{t}) < 0$, then $f$ satisfies both (WWI) and (WWII) on the interval $((\tilde{t} - \mu)_+, \tilde{t})$, as the right-hand side of (WWII) is $+\infty$. 
Otherwise, $\phi(\hat{t}) = 0$. Let $\tilde{t} \in \arg \min_{t \in [0,\hat{t}]} \phi(t)$. Then, $\tilde{t} \in (0, \hat{t})$, with $0 \leq \phi'(\tilde{t}; \mu) \leq \Delta \phi(\tilde{t}, \mu)$ for all $\mu$, equivalently

$$\frac{\Delta f(x + \tilde{t}d; \mu d)}{\mu} \geq \sigma_1 \Delta f(x; d) > \sigma_2 \Delta f(x; d) \quad \forall \mu > 0,$$

so (WWI) and (WWII) hold with strict inequality at $\tilde{t}$. We now consider two cases based on whether $x + (\tilde{t} + \mu)d \in \text{dom} \ (g)$.

First, for all sufficiently small $\mu > 0$, $x + (\tilde{t} + \mu)d \in \text{dom} \ (g)$. Because the inequalities in (WWI) and (WWII) are strict at $\tilde{t}$, relative continuity of $f$ and of $t \mapsto \Delta f(x + td; d)$ at $t = \tilde{t}$ imply there exists an open interval $I$ with $\tilde{t} \in I$ and $x + I d \subset \text{dom} \ (g)$ where both (WWI) and (WWII) hold.

For those $\mu > 0$ for which $x + (\tilde{t} + \mu)d \notin \text{dom} \ (g)$, (WWI) and (WWII) hold for all $t \in ((\tilde{t} - \mu), \hat{t})$ as argued in the previous case where $\phi(\hat{t}) < 0$. □

Next, we prove finite termination of a bisection algorithm to point $\tilde{t} \geq 0$ satisfying the weak Wolfe conditions. The algorithm is analogous to the weak Wolfe bisection method for finite-valued nonsmooth minimization in [13].

**Algorithm 1 Weak Wolfe Bisection Method**

| Input: $x \in \text{dom} \ (g), d \in \mathbb{R}^n$ with $\Delta f(x; d) < 0$, and $0 < \sigma_1 < \sigma_2 < 1$, $\mu > 0$. |
|---|
| 1: **procedure** WWBisect($x, d, \sigma_1, \sigma_2$) |
| 2: $\alpha \leftarrow 0$; |
| 3: $t \leftarrow 1$; |
| 4: $\beta \leftarrow \infty$; |
| 5: **while** (WWI) and (WWII) fail **do** |
| 6: **if** $f(x + td) > f(x) + \sigma_1 t \Delta f(x; d)$ **then** \quad $\triangleright$ If not sufficient decrease |
| 7: $\beta \leftarrow t$ |
| 8: **else if** $\sigma_2 \Delta f(x; d) > \frac{\Delta f(x + td; \mu d)}{\mu}$ **then** \quad $\triangleright$ Else if not curvature |
| 9: $\alpha \leftarrow t$ |
| 10: **else** |
| 11: return $t$ |
| 12: **end if** |
| 13: **if** $\beta = \infty$ **then** \quad $\triangleright$ Doubling Phase |
| 14: $t \leftarrow 2t$ |
| 15: **else** |
| 16: $t \leftarrow \frac{1}{2}(\alpha + \beta)$ \quad $\triangleright$ Bisection Phase |
| 17: **end if** |
| 18: **end while** |
| 19: **end procedure** |

**Lemma 4.2.** Let $f$ be given as in $\mathcal{P}$, $x \in \text{dom} \ (g)$, and $d$ chosen such that $\Delta f(x; d) < 0$. Then, one of the following must occur in Algorithm 1:
(a) the doubling phase does not terminate finitely, with the parameter $\beta$ never set to a finite value, the parameter $\alpha$ becoming positive on the first iteration and doubling every iteration thereafter, with $f(x + t_k d) \searrow -\infty$,

(b) both the doubling phase and the bisection phase terminate finitely to a $\bar{t} \geq 0$ for which the weak Wolfe conditions are satisfied.

Proof. Suppose the procedure does not terminate finitely. If the parameter $\beta$ is never set to a finite value, then the doubling phase does not terminate. Then the parameter $\alpha$ becomes positive on the first iteration and doubles on each subsequent iteration $k$, with $t_k$ satisfying

$$f(x + t_k d) \leq f(x) + \sigma_1 t_k \Delta f(x; d), \quad \forall k \geq 1.$$ 

Therefore, since $\Delta f(x; d) < 0$, the function values $f(x + t_k d) \searrow -\infty$, so the first option occurs.

Otherwise, the procedure does not terminate finitely, and $\beta$ is eventually finite. Therefore, the doubling phase terminates finitely, but the bisection phase does not terminate finitely. This implies there exists $\bar{t} \geq 0$ such that

\begin{equation}
\alpha_k \nearrow \bar{t}, \quad t_k \to \bar{t}, \quad \beta_k \searrow \bar{t}.
\end{equation}

We now consider two cases. First, suppose that the parameter $\alpha$ is never set to a positive number. Then, $\alpha_k = 0$ for all $k \geq 1$, and $t_k, \beta_k \to 0$, so the first if statement is entered in each iteration. This implies

$$\sigma_1 \Delta f(x; d) < \frac{f(x + t_k d) - f(x)}{t_k}, \quad \forall k \geq 1.$$ 

Since $[x, x + d] \subset \text{dom}(g)$, Lemma 3.1 yields the chain of inequalities

$$\sigma_1 \Delta f(x; d) \leq f'(x; d) \leq \Delta f(x; d) < 0,$$

which contradicts $\sigma_1 \in (0, 1)$.

Otherwise, the parameter $\alpha$ is eventually positive. Then, the bisection phase does not terminate, and the algorithm generates infinite sequences $\{\alpha_k\}, \{t_k\},$ and $\{\beta_k\}$ satisfying (7) such that, for all $k$ large, $0 < \alpha_k < t_k < \beta_k < \infty$, and

\begin{align*}
(8) & \quad f(x + \alpha_k d) \leq f(x) + \sigma_1 \alpha_k \Delta f(x; d), \\
(9) & \quad f(x + \beta_k d) > f(x) + \sigma_1 \beta_k \Delta f(x; d), \\
(10) & \quad \sigma_2 \Delta f(x; d) > \frac{\Delta f(x + \beta_k d; \mu d)}{\mu}, \\
(11) & \quad [x, x + \max \{\alpha_k + \mu, \beta_k\} d] \subset \text{dom}(g).
\end{align*}

Letting $k \to \infty$ in (10) and using lower semicontinuity of $g$ gives

\begin{equation}
\sigma_2 \Delta f(x; d) \geq \frac{\Delta f(x + \bar{t} d; \mu d)}{\mu}.
\end{equation}
By Theorem 3.2, for sufficiently large \( k \) there exists \( \tau_k \in (0,1) \) so that the vectors

\[
x^k := (1 - \tau_k)(x + \alpha_k d) + \tau_k(x + \beta_k d) = x + [(1 - \tau_k)\alpha_k + \tau_k \beta_k]d,
\]

\( v^k \in \partial f(x^k) \)

yield an extended form of the mean-value theorem

\[
(13) \quad f(x + \beta_k d) - f(x + \alpha_k d) = \langle v^k, (\beta_k - \alpha_k)d \rangle.
\]

Let \( \gamma_k := (1 - \tau_k)\alpha_k + \tau_k \beta_k \in (\alpha_k, \beta_k) \), so that \( x^k \) = \( x + \gamma_k d \). Then, \( \gamma_k \to T \) as \( k \to \infty \). Combining (8) and (9) and using (13) gives

\[
\sigma_1(\beta_k - \alpha_k)\Delta f(x; d) < f(x + \beta_k d) - f(x + \alpha_k d) = \langle v^k, (\beta_k - \alpha_k)d \rangle.
\]

Dividing by \( \beta_k - \alpha_k > 0 \) gives

\[
\sigma_1 \Delta f(x; d) \leq f'(x + \gamma_k d; d) \leq \frac{\Delta f(x + \gamma_k d; \mu d)}{\mu}.
\]

As \( k \to \infty \), using (12), we obtain the string of inequalities

\[
\frac{\Delta f(x + T d; \mu d)}{\mu} \leq \sigma_2 \Delta f(x; d) < \sigma_1 \Delta f(x; d) \leq \frac{\Delta f(x + T d; \mu d)}{\mu},
\]

which is a contradiction. Therefore, either the doubling phase never terminates or the procedure terminates finitely at some \( T \) at which \( f \) satisfies both weak Wolfe conditions. \( \square \)

A global convergence result for the weak Wolfe line search that parallels [2, Theorem 2.4] now follows under standard Lipschitz assumptions, which hold, in particular, if the initial set \( \text{lev}_f(f(x^0)) \) is compact.

**Algorithm 2** Global Weak Wolfe

1: **procedure** WeakWolfeGlobal\((x^0, \sigma_1, \sigma_2)\)
2: \( k \leftarrow 0 \)
3: repeat
4: \( \text{Find } d^k \in \mathbb{R}^n \text{ such that } \Delta f(x^k; d^k) < 0 \)
5: \( \text{if no such } d^k \text{ then} \)
6: \( \quad 0 \in \partial f(x^k) \text{ return} \)
7: \( \text{end if} \)
8: \( \text{Let } t_k \text{ be a step size satisfying (WWI) and (WWII)} \)
9: \( \text{if no such } t_k \text{ then} \)
10: \( \quad f \text{ unbounded below. return} \)
11: \( \text{end if} \)
12: \( x^k \leftarrow x^k + t_k d^k \)
13: \( k \leftarrow k + 1 \)
14: until
15: **end procedure**
Theorem 4.1. Let $f$ be as in $\mathcal{P}$, $x^0 \in \text{dom}(g)$, $0 < \sigma_1 < \sigma_2 < 1$, and $0 < \mu < 1$. Set $\mathcal{L} := \text{lev}_f(f(x^0))$. Suppose there exists $M, \tilde{M} > 0$ such that $\left\| d^k \right\| \leq M$ for all $k \geq 0$, $\sup_{x \in \mathcal{L}} \left\| \nabla c(x) \right\| \leq \tilde{M}$, and

(i) $c$ is $L_c$-Lipschitz on $\mathcal{L}$;

(ii) $\nabla c$ is $L_{\nabla c}$-Lipschitz on $\mathcal{L}$;

(iii) $g$ is $L_g$-Lipschitz on $(\mathcal{L} + M\mu B) \cap \text{dom}(g)$;

(iv) $h$ is $L_h$-Lipschitz on $c(\mathcal{L}) + M\tilde{M}\mu B$.

Let $\{x^k\}$ be a sequence initialized at $x^0$ and generated by Algorithm 2: Then at least one of the following must occur:

(a) the algorithm terminates finitely at a first-order stationary point for $f$;

(b) for some $k$ the step size selection procedure generates a sequence of trial step sizes $t_{k_n} \to \infty$ such that $f(x^k + t_{k_n}d^k) \to -\infty$;

(c) $f(x^k) \searrow -\infty$;

(d) $\sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\left\| d^k \right\| + \left\| d^k \right\|^2} < \infty$, in particular, $\Delta f(x^k; d^k) \to 0$.

Proof. We assume (a) - (c) do not occur and show (d) occurs. Since (a) does not occur, the sequence $\{x^k\}$ is infinite, and $\Delta f(x^k; d^k) < 0$ for all $k \geq 0$. Since (b) does not occur, Lemma 4.2 implies that the weak Wolfe bisection method terminates finitely at every iteration $k \geq 0$. The sufficient decrease condition (WWI) gives a strict descent method, so the function values $\{f(x^k)\}$ are strictly decreasing, with $\{x^k\} \subset \mathcal{L}$ for all $k \geq 0$. By the nonoccurrence of (c), $f(x^k) \searrow f > -\infty$.

We first show that for each $k \geq 0$, the step size $t_k$ satisfies

$$t_k \geq \min \left\{ 1 - \mu, \frac{\mu(1 - \sigma_2)|\Delta f(x^k; d^k)|}{K\left(\left\| d^k \right\| + \left\| d^k \right\|^2\right)} \right\},$$

by considering two cases.

First, suppose $\Delta f(x^{k+1}; \mu d^k) = \infty$. Then $x^{k+1} + \mu d^k = x^k + (t_k + \mu)d^k \notin \text{dom}(g)$. Since $x^k + d^k \in \text{dom}(g), t_k + \mu > 1$, and by assumption $0 < \mu < 1$. Therefore, $t_k \geq 1 - \mu$. 

Otherwise, $\Delta f(x^{k+1}; \mu d^k) < \infty$. Then

$$\Delta f(x^{k+1}; \mu d^k) - \Delta f(x^k; \mu d^k) = h(c(x^{k+1}) + \nabla c(x^{k+1})\mu d^k) - h(c(x^k)) + g(x^{k+1} + \mu d^k) - g(x^k)$$

$$= h(c(x^k)) - h(c(x^{k+1})) + h(c(x^{k+1}) + \nabla c(x^{k+1})\mu d^k) - h(c(x^k) + \nabla c(x^k)\mu d^k) + g(x^{k+1}) + g(x^{k+1} + \mu d^k) - g(x^k)$$

$$\leq 2L_h L_c t_k \|d^k\| + L_h L_c \mu t_k \|d^k\|^2 + 2L g t_k \|d^k\|$$

$$\leq K t_k \left( \|d^k\| + \|d^k\|^2 \right),$$

for some $K \geq 0$. Adding and subtracting in (WWII) gives

$$\sigma_2 \Delta f(x^k; d^k) \leq \frac{\Delta f(x^k + t_k d^k; \mu d^k)}{\mu}$$

$$= \frac{\Delta f(x^k; \mu d^k)}{\mu} + \left[ \frac{\Delta f(x^k + t_k d^k; \mu d^k)}{\mu} - \frac{\Delta f(x^k; \mu d^k)}{\mu} \right]$$

$$\leq \Delta f(x^k; d^k) + \frac{K}{\mu} t_k \left( \|d^k\| + \|d^k\|^2 \right) \operatorname{since} 0 < \mu < 1,$$

which rearranges to

$$0 < \frac{\mu(1 - \sigma_2) \|\Delta f(x^k; d^k)\|}{K \left( \|d^k\| + \|d^k\|^2 \right)} \leq t_k,$$

so (14) holds. Next, (WWI) and (14) imply

$$\sigma_1 \min \left\{ 1 - \mu, \frac{\mu(1 - \sigma_2) \|\Delta f(x^k; d^k)\|}{K \left( \|d^k\| + \|d^k\|^2 \right)} \right\} \|\Delta f(x^k; d^k)\| \leq \sigma_1 t_k \|\Delta f(x^k; d^k)\| \leq f(x^k) - f(x^{k+1}).$$

We aim to show that the bound (15) holds for all large $k$ by showing $\Delta f(x^k; d^k) \to 0$ and using boundedness of the search directions $\{d^k\}$. Suppose there exists a subsequence $J_1 \subset \mathbb{N}$ for which $\Delta f(x^k; d^k) \not\to 0$. Let $\gamma > 0$ be such that $\sup_{k \in J_1} \Delta f(x^k; d^k) \leq -\gamma < 0$.

Then, since $\{d^k\} \subset \mathcal{MB}$,

$$\frac{\mu(1 - \sigma_2) \|\Delta f(x^k; d^k)\|}{K \left( \|d^k\| + \|d^k\|^2 \right)} \not\to 0.$$

If there exists a further subsequence $J_2 \subset J_1$ with

$$\frac{\mu(1 - \sigma_2) \|\Delta f(x^k; d^k)\|}{K \left( \|d^k\| + \|d^k\|^2 \right)} \geq 1 - \mu, \quad \forall k \in J_2,$$
then by expanding the recurrence given by (WWI), and writing $J_2 = \{k_1, k_2, \ldots\}$, we have

\begin{equation}
\begin{aligned}
f(x^{k_n}) \leq f(x^{k_{n-1}}) - \sigma_1 (1 - \mu) \gamma \\
\leq f(x^{k_1}) - C(k_n) \sigma_1 (1 - \mu) \gamma
\end{aligned}
\end{equation}

with $C(k_n) \to \infty$ as $n \to \infty$. This contradicts the nonoccurrence of (c). By (17), there exists a subsequence $J_2 \subset J_1$ and $\delta > 0$ so that

\begin{equation}
0 < \delta \leq \frac{\mu (1 - \sigma_2) |\Delta f(x^k; d^k)|}{K \left( \|d^k\| + \|d^k\|^2 \right)} < 1 - \mu
\end{equation}

for all large $k \in J_2$. Repeating the argument at (18) with $\delta$ in place of $1 - \mu$, we conclude

\begin{equation}
\frac{\mu (1 - \sigma_2) |\Delta f(x^k; d^k)|}{K \left( \|d^k\| + \|d^k\|^2 \right)} \to 0,
\end{equation}

and consequently $\Delta f(x^k; d^k) \to 0$, which is a contradiction. Therefore, (15) holds for all $k \geq k_0$. Summing over $k \in \mathbb{N}$ in (16) gives

\begin{equation}
0 < \sum_{k \geq k_0} \frac{\sigma_1 \mu (1 - \sigma_2) |\Delta f(x^k; d^k)|^2}{K \left( \|d^k\| + \|d^k\|^2 \right)} < f(x^0) - \lim_{k \to \infty} f(x^k).
\end{equation}

Since (c) does not occur, $\lim_{k \to \infty} f(x^k) > -\infty$, so (d) must occur. □

**Remark 3.** When $h$ is the identity on $\mathbb{R}$ and $g = 0$, we recover the convergence analysis of weak Wolfe for smooth minimization given in [20, Theorem 3.2].

**Remark 4.** The hypotheses of Theorem 4.1 simplify if $h$ is globally Lipschitz. In that case, the boundedness condition on $\{\|\nabla c(x)\| : x \in \mathcal{L}\}$ is not necessary. Alternatively, if $\|\nabla c(x)\|$ is bounded on the closed convex hull of $\mathcal{L}$, then the Lipschitz condition of $c$ on $\mathcal{L}$ is immediate.

The following corollary is an immediate consequence of Lemma 3.2.

**Corollary 4.1.** Let the hypotheses of Theorem 4.1 hold. If $0 < \beta < 1$ and the directions $\{d^k\}$ are chosen to satisfy

\begin{equation}
\Delta f(x^k; d^k) \leq \beta \Delta_k f < 0,
\end{equation}

then the occurrence of (d) in Theorem 4.1 implies that cluster points of $\{x^k\}$ are first-order stationary for $\mathcal{P}$.

5. BACKTRACKING FOR CONVEX-COMPOSITE MINIMIZATION

A similar result holds for the backtracking line search given in [2], which enforces the step-size condition of (WWI) at each iteration. The method of proof adapts the step-size arguments given in Royer and Wright [17] to the convex-composite setting. Similar ideas on convex majorants for the composite $\mathcal{P}$ are employed in [10,14].
Algorithm 3 Global Backtracking

1: procedure BACKTRACKINGGLOBAL($x^0, \sigma_1, \theta$)
2: \( k \leftarrow 0 \)
3: repeat
4: Find \( d^k \in \mathbb{R}^n \) such that \( \Delta f(x^k; d^k) < 0 \)
5: if no such \( d^k \) then
6: \( 0 \in \partial f(x^k) \) return
7: end if
8: \( t \leftarrow 1 \)
9: while \( f(x^k + td^k) > f(x^k) + \sigma_1 t \Delta f(x^k; d^k) \) do
10: \( t \leftarrow \theta t \)
11: end while
12: \( t_k \leftarrow t \)
13: \( x^k \leftarrow x^k + t_k d^k \)
14: \( k \leftarrow k + 1 \)
15: until
16: end procedure

Theorem 5.1. Let \( f \) be as in \( \mathcal{P} \), \( x^0 \in \text{dom}(g) \), \( 0 < \sigma_1 < 1 \), and \( 0 < \theta < 1 \). Set \( \mathcal{L} := \text{lev}_f(f(x^0)) \).
Suppose there exists \( M > 0 \) and \( \tilde{M} > 0 \) such that \( \|d^k\| \leq M \), \( \sup_{x \in \mathcal{L}} \|\nabla c(x)\| \leq \tilde{M} \), and that

(i) \( \nabla c \) is \( L_{\nabla c} \)-Lipschitz on \( \mathcal{L} + MB_n \);

(ii) \( h \) is \( L_h \)-Lipschitz on \( c(\mathcal{L} + M\mathbb{B}) + \tilde{M}M\mathbb{B}_m \).

Let \( \{x^k\} \) be a sequence initialized at \( x^0 \) and generated by Algorithm 3: Then, at least one of the following must occur:

(a) the algorithm terminates finitely at a first-order stationary point for \( f \),

(b) \( f(x^k) \searrow -\infty \),

(c) \( \sum_{k=0}^{\infty} \frac{\Delta f(x^k; d^k)^2}{\|d^k\|_2^2} < \infty \), in particular, \( \Delta f(x^k; d^k) \to 0 \).

Proof. We assume (a) - (b) do not occur and show (c) occurs. Since (a) does not occur, the sequence \( \{x^k\} \) is infinite, and \( \Delta f(x^k; d^k) < 0 \) for all \( k \geq 0 \). The sufficient decrease (WWI) obtained by the backtracking subroutine gives a strict descent method, so the function values \( \{f(x^k)\} \) are strictly decreasing, with \( \{x^k\} \subset \mathcal{L} \) for all \( k \geq 0 \). In particular, \( f(x^k) \searrow \overline{f} > -\infty \).

We first show that for each \( k \geq 0 \), the step size \( 0 < t_k \leq 1 \) satisfies

(19) \( t_k \geq \min \left\{ 1, \frac{\mu(1 - \sigma_2)|\Delta f(x^k; d^k)|}{L_{\nabla c}L_h\|d^k\|_2^2} \right\} \).
by considering two cases.

If the unit step \( t_k = 1 \) is accepted, the bound is immediate. Following [17], suppose now that the unit step length is not accepted. Then \( \tilde{t} := \theta^j \in (0, 1] \) does not satisfy the decrease condition for some \( j \geq 0 \). Using the Lipschitz condition on \( h \), the quadratic bound lemma, and Lemma 3.1, we obtain

\[
\sigma_1 \tilde{t} \Delta f(x^k; d^k) < f(x^k + \tilde{t}d^k) - f(x) \leq \Delta f(x^k; \tilde{t}d^k) + \frac{L_{\nabla c}L_h}{2} \| \tilde{t}d^k \|^2_2 \\
\leq \tilde{t} \Delta f(x^k; d^k) + (\tilde{t})^2 \frac{L_{\nabla c}L_h}{2} \| d^k \|^2_2
\]

After dividing both sides by \( \tilde{t} > 0 \) and rearranging,

\[
(20) \quad \tilde{t} \geq \frac{2(1 - \sigma_1) |\Delta f(x^k; d^k)|}{L_{\nabla c}L_h \| d^k \|^2_2},
\]

Consequently, when the backtracking algorithm terminates at \( t_k > 0 \),

\[
(21) \quad t_k \geq \frac{2\theta(1 - \sigma_1) |\Delta f(x^k; d^k)|}{L_{\nabla c}L_h \| d^k \|^2_2}.
\]

Therefore, \( t_k \) satisfying (WWI) implies

\[
\sigma_1 \min \left\{ 1, \theta \frac{2(1 - \sigma_1) |\Delta f(x; d)|}{L_{\nabla c}L_h \| d \|^2_2} \right\} |\Delta f(x; d)| \leq \sigma_1 t_k |\Delta f(x^k; d^k)| \leq f(x^k) - f(x^{k+1}).
\]

Using the boundedness of the search directions and arguing as in the proof of Theorem 4.1, the bound (21) holds for all \( k \geq k_0 \). Summing the previous display,

\[
0 < \sum_{k \geq k_0} \theta \frac{2\sigma_1 (1 - \sigma_1) \Delta f(x^k; d^k)^2}{L_{\nabla c}L_h \| d^k \|^2_2} < f(x^0) - \lim_{k \to \infty} f(x^k).
\]

Since (b) does not occur, \( \lim_{k \to \infty} f(x^k) > -\infty \), so (c) must occur. \( \square \)

**Remark 5.** When \( h \) is the identity on \( \mathbb{R} \) and \( g = 0 \), we recover the convergence analysis of backtracking for smooth minimization.

The following corollary is an immediate consequence of Lemma 3.2.

**Corollary 5.1.** Let the hypotheses of Theorem 5.1 hold. If \( 0 < \beta < 1 \) and the directions \( \{d^k\} \) are chosen to satisfy

\[
\Delta f(x^k; d^k) < \beta \bar{\Delta}^k f < 0,
\]

then the occurrence of (c) in Theorem 5.1 implies that cluster points of \( \{x^k\} \) are first-order stationary for \( \mathcal{P} \).
References

[1] A. Aravkin, J. Burke, and G. Pillonetto. Sparse/robust estimation and kalman smoothing with non-smooth log-concave densities: Modeling, computation, and theory. *The Journal of Machine Learning Research*, 14(1):2689–2728, 2013.

[2] J. V. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33(3):260–279, 1985.

[3] J. V. Burke. Second order necessary and sufficient conditions for convex composite ndo. *Mathematical Programming*, 38(3):287–302, 1987.

[4] J. V. Burke. A sequential quadratic programming method for potentially infeasible mathematical programs. *Journal of Mathematical Analysis and Applications*, 139(2):319–351, 1989.

[5] J. V. Burke and A. Engle. Strong metric (sub) regularity of kkt mappings for piecewise linear-quadratic convex-composite optimization. *arXiv preprint arXiv:1805.01073*, 2018.

[6] J. V. Burke and M. C. Ferris. A gauss newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[7] J. V. Burke and R. Poliquin. Optimality conditions for non-finite valued convex composite functions. *Mathematical Programming*, 57(1):103–120, 1992.

[8] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[9] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 2018.

[10] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *ArXiv e-prints*, Apr. 2016.

[11] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *ArXiv e-prints*, May 2017.

[12] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

[13] A. S. Lewis and M. L. Overton. Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1-2):135–163, 2013.

[14] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[15] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[16] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer, 1998.

[17] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[18] P. Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

[19] P. Wolfe. Convergence conditions for ascent methods. ii: Some corrections. *SIAM review*, 13(2):185–188, 1971.

[20] S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.