PROXIMAL NEWTON-TYPE METHODS FOR MINIMIZING COMPOSITE FUNCTIONS*

JASON D. LEE†, YUEKAI SUN†, AND MICHAEL A. SAUNDERS‡

Abstract. We generalize Newton-type methods for minimizing smooth functions to handle a sum of two convex functions: a smooth function and a nonsmooth function with a simple proximal mapping. We show that the resulting proximal Newton-type methods inherit the desirable convergence behavior of Newton-type methods for minimizing smooth functions, even when search directions are computed inexactly. Many popular methods tailored to problems arising in bioinformatics, signal processing, and statistical learning are special cases of proximal Newton-type methods, and our analysis yields new convergence results for some of these methods.

1. Introduction. Many problems of relevance in bioinformatics, signal processing, and statistical learning can be formulated as minimizing a *composite function*:

$$(1.1) \qquad \underset{x \in \mathbf{R}^n}{\text{minimize}} \; f(x) := g(x) + h(x),$$

where g is a convex, continuously differentiable loss function, and h is a convex but not necessarily differentiable penalty function or regularizer. Such problems include the *lasso* [25], the *graphical lasso* [11], and trace-norm matrix completion [6].

We describe a family of Newton-type methods for minimizing composite functions that achieve superlinear rates of convergence subject to standard assumptions. The methods can be interpreted as generalizations of the classic proximal gradient method that account for the curvature of the function when selecting a search direction. Many popular methods for minimizing composite functions are special cases of these *proximal Newton-type methods*, and our analysis yields new convergence results for some of these methods.

In section 1 we review state-of-the-art methods for problem (1.1) and related work on projected Newton-type methods for constrained optimization. In sections 2 and 3 we describe proximal Newton-type methods and their convergence behavior, and in section 4 we discuss some applications of these methods and evaluate their performance.

†Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 (jdl17@stanford.edu, yuekai@stanford.edu).

‡Department of Management Science and Engineering, Stanford University, Stanford, CA (saunders@stanford.edu).

Notation. The methods we consider are *line search methods*, which produce a sequence of points $\{x_k\}$ according to

$$x_{k+1} = x_k + t_k \Delta x_k,$$

where t_k is a *step length* and Δx_k is a *search direction*. When we focus on one iteration of an algorithm, we drop the subscripts (e.g., $x_+ = x + t\Delta x$). All the methods we consider compute search directions by minimizing local models of the composite function f. We use an accent $\hat{\cdot}$ to denote these local models (e.g., $\hat{f}_k$ is a local model of f at the kth step).

1.1. First-order methods. The most popular methods for minimizing composite functions are *first-order methods* that use *proximal mappings* to handle the nonsmooth part h. SpaRSA [29] is a popular *spectral projected gradient* method that uses a *spectral step length* together with a *nonmonotone line search* to improve convergence. TRIP [14] also uses a spectral step length but selects search directions using a trust-region strategy.

We can accelerate the convergence of first-order methods using ideas due to Nesterov [17]. This yields *accelerated first-order methods*, which achieve ϵ-suboptimality within $O(1/\sqrt{\epsilon})$ iterations [26]. The most popular method in this family is the fast iterative shrinkage-thresholding algorithm (FISTA) [1]. These methods have been implemented in the package TFOCS [2] and used to solve problems that commonly arise in statistics, signal processing, and statistical learning.

1.2. Newton-type methods. There are two classes of methods that generalize Newton-type methods for minimizing smooth functions to handle composite functions (1.1). *Nonsmooth Newton-type methods* [30] successively minimize a local quadratic model of the composite function f:

$$\hat{f}_k(y) = f(x_k) + \sup_{z \in \partial f(x_k)} z^T(y - x_k) + \frac{1}{2}(y - x_k)^T H_k(y - x_k),$$

where H_k accounts for the curvature of f. (Although computing this Δx_k is generally not practical, we can exploit the special structure of f in many statistical learning problems.) *Proximal Newton-type methods* approximate only the smooth part g with a local quadratic model:

$$\hat{f}_k(y) = g(x_k) + \nabla g(x_k)^T(y - x_k) + \frac{1}{2}(y - x_k)^T H_k(y - x_k) + h(y),$$

where H_k is an approximation to $\nabla^2 g(x_k)$. This idea can be traced back to the *generalized proximal point method* of Fukushima and Mine [12].

Proximal Newton-type methods are a special case of cost approximation (Patriksson [19]). In particular, Theorem 4.1 (convergence under Rule E) and Theorem 4.6 (linear convergence) of [19] apply to proximal Newton-type methods. Patriksson shows superlinear convergence of the exact proximal Newton method, but does not analyze the quasi-Newton approximation, nor consider the adaptive stopping criterion of section 3.4. By focusing on Newton-type methods, we obtain stronger and more practical results.

Many popular methods for minimizing composite functions are special cases of proximal Newton-type methods. Methods tailored to a specific problem include FPC_AS [28], `glmnet` [10], `newglmnet` [31], QUIC [13], and the Newton-LASSO

method [18]. Generic methods include *projected Newton-type methods* [24, 23], proximal quasi-Newton methods [22, 3], and the method of Tseng and Yun [27, 16].

This article is the full version of our preliminary work [15], and section 3 includes a convergence analysis of inexact proximal Newton-type methods (i.e., when the subproblems are solved inexactly). Our main convergence results are

1. the proximal Newton and proximal quasi-Newton methods (with line search) converge superlinearly;
2. the inexact proximal Newton method (with unit step length) converges locally at a linear or superlinear rate depending on the forcing sequence.

We also describe an adaptive stopping condition to decide how exactly (or inexactly) to solve the subproblem, and we demonstrate the benefits empirically.

There is a rich literature on generalized equations, such as monotone inclusions and variational inequalities. Minimizing composite functions is a special case of solving generalized equations, and proximal Newton-type methods are special cases of Newton-type methods for solving them [19]. We refer to Patriksson [20] for a unified treatment of descent methods for solving a large class of generalized equations.

2. Proximal Newton-type methods. In problem (1.1) we assume g and h are closed, convex functions, with g continuously differentiable and its gradient ∇g Lipschitz continuous. The function h is not necessarily differentiable everywhere, but its *proximal mapping* (2.1) can be evaluated efficiently. We refer to g as "the smooth part" and h as "the nonsmooth part". We assume the optimal value is attained at some optimal solution $x^\star$, not necessarily unique.

2.1. The proximal gradient method. The proximal mapping of a convex function h at x is

$$(2.1) \qquad \operatorname{prox}_h(x) := \operatorname*{arg\,min}_{y \in \mathbf{R}^n} \; h(y) + \frac{1}{2} \left\| y - x \right\|^2 .$$

Proximal mappings can be interpreted as generalized projections because if h is the indicator function of a convex set, $\operatorname{prox}_h(x)$ is the projection of x onto the set. If h is the ℓ_1 norm and t is a step-length, then $\operatorname{prox}_{th}(x)$ is the *soft-threshold operation*:

$$\operatorname{prox}_{t\ell_1}(x) = \operatorname{sign}(x) \cdot \max\{|x| - t, 0\},$$

where sign and max are entrywise, and $\cdot$ denotes the entrywise product.

The *proximal gradient method* uses the proximal mapping of the nonsmooth part to minimize composite functions. For some step length t_k, the next iterate is $x_{k+1} = \operatorname{prox}_{t_k h}(x_k - t_k \nabla g(x_k))$. This is equivalent to

$$(2.2) \qquad\qquad\qquad x_{k+1} = x_k - t_k G_{t_k f}(x_k),$$

$$(2.3) \qquad\qquad G_{t_k f}(x_k) := \frac{1}{t_k} \left(x_k - \operatorname{prox}_{t_k h}(x_k - t_k \nabla g(x_k)) \right),$$

where $G_{t_k f}(x_k)$ is a *composite gradient step*. Most first-order methods, including SpaRSA and accelerated first-order methods, are variants of this simple method. We note three properties of the composite gradient step.

1. Let $\hat{g}$ be a simple quadratic model of g near x_k (with H_k a multiple of I):

$$\hat{g}_k(y) := g(x_k) + \nabla g(x_k)^T (y - x_k) + \frac{1}{2t_k} \left\| y - x_k \right\|^2 .$$

The composite gradient step moves to the minimum of $\hat{g}_k + h$:

$$(2.4) \qquad x_{k+1} = \text{prox}_{t_k h} \left(x_k - t_k \nabla g(x_k) \right)$$

$$(2.5) \qquad = \arg\min_y \, t_k h(y) + \frac{1}{2} \left\| y - x_k + t_k \nabla g(x_k) \right\|^2$$

$$(2.6) \qquad = \arg\min_y \, \nabla g(x_k)^T (y - x_k) + \frac{1}{2t_k} \left\| y - x_k \right\|^2 + h(y).$$

2. The composite gradient step is neither a gradient nor a subgradient of f at any point; rather it is the sum of an explicit gradient (at x) and an implicit subgradient (at $\text{prox}_h(x)$). The first-order optimality conditions of (2.6) are

$$\partial h(x_{k+1}) + \frac{1}{t_k}(x_{k+1} - x_k) = 0.$$

We express $x_{k+1} - x_k$ in terms of $G_{t_k f}$ and rearrange to obtain

$$G_{t_k f}(x_k) \in \nabla g(x_k) + \partial h(x_{k+1}).$$

3. The composite gradient step is zero if and only if x minimizes f, i.e., $G_f(x) = 0$ generalizes the usual (zero gradient) optimality condition to composite functions (where $G_f(x) = G_{tf}(x)$ when $t = 1$).

We shall use the length of $G_f(x)$ to measure the optimality of a point x. We show that G_f inherits the Lipschitz continuity of ∇g.

DEFINITION 2.1. *A function F is Lipschitz continuous with constant L if*

$$(2.7) \qquad \left\| F(x) - F(y) \right\| \le L \left\| x - y \right\| \text{ for any } x, y.$$

LEMMA 2.2. *If ∇g is Lipschitz continuous with constant L_1, then*

$$\left\| G_f(x) \right\| \le (L_1 + 1) \left\| x - x^\star \right\|.$$

Proof. The composite gradient steps at x and the optimal solution $x^\star$ satisfy

$$G_f(x) \in \nabla g(x) + \partial h(x - G_f(x)),$$
$$G_f(x^\star) \in \nabla g(x^\star) + \partial h(x^\star).$$

We subtract these two expressions and rearrange to obtain

$$\partial h(x - G_f(x)) - \partial h(x^\star) \ni G_f(x) - (\nabla g(x) - \nabla g(x^\star)).$$

Since h is convex, ∂h is monotone and

$$0 \le (x - G_f(x) - x^\star)^T \partial h(x - G_f(x_k))$$
$$= -G_f(x)^T G_f(x) + (x - x^\star)^T G_f(x) + G_f(x)^T (\nabla g(x) - \nabla g(x^\star))$$
$$+ (x - x^\star)^T (\nabla g(x) - \nabla g(x^\star)).$$

We drop the last term because it is nonnegative (∇g is monotone) to obtain

$$0 \le - \left\| G_f(x) \right\|^2 + (x - x^\star)^T G_f(x) + G_f(x)^T (\nabla g(x) - \nabla g(x^\star))$$
$$\le - \left\| G_f(x) \right\|^2 + \left\| G_f(x) \right\| \left(\left\| x - x^\star \right\| + \left\| \nabla g(x) - \nabla g(x^\star) \right\| \right),$$

so that

$$(2.8) \qquad \left\| G_f(x) \right\| \le \left\| x - x^\star \right\| + \left\| \nabla g(x) - \nabla g(x^\star) \right\|.$$

Since ∇g is Lipschitz continuous, we have

$$\left\| G_f(x) \right\| \le (L_1 + 1) \left\| x - x^\star \right\|. \qquad \square$$

2.2. Proximal Newton-type methods. Proximal Newton-type methods use a symmetric positive definite matrix $H_k \approx \nabla^2 g(x_k)$ to model the curvature of g:

$$\hat{g}_k(y) = g(x_k) + \nabla g(x_k)^T(y - x_k) + \frac{1}{2}(y - x_k)^T H_k(y - x_k).$$

A proximal Newton-type search direction Δx_k solves the subproblem

$$(2.9) \qquad \Delta x_k = \arg\min_d \hat{f}_k(x_k + d) := \hat{g}_k(x_k + d) + h(x_k + d).$$

There are many strategies for choosing H_k. If we choose $H_k = \nabla^2 g(x_k)$, we obtain the *proximal Newton method*. If we build an approximation to $\nabla^2 g(x_k)$ according to a quasi-Newton strategy, we obtain a *proximal quasi-Newton method*. If the problem is large, we can use limited memory quasi-Newton updates to reduce memory usage. Generally speaking, most strategies for choosing Hessian approximations in Newton-type methods (for minimizing smooth functions) can be adapted to choosing H_k in proximal Newton-type methods.

When H_k is not positive definite, we can also adapt strategies for handling indefinite Hessian approximations in Newton-type methods. The simplest strategy is Hessian modification: we add a multiple of the identity to H_k when H_k is indefinite. This makes the subproblem strongly convex and damps the search direction. In a proximal quasi-Newton method, we can also do update skipping: if an update causes H_k to become indefinite, simply skip the update.

We can also express the proximal Newton-type search direction using *scaled proximal mappings*. This lets us interpret a proximal Newton-type search direction as a "composite Newton step" and reveals a connection with the composite gradient step.

DEFINITION 2.3. *Let h be a convex function and H be a positive definite matrix. Then the scaled proximal mapping of h at x is*

$$(2.10) \qquad \mathrm{prox}_h^H(x) := \arg\min_{y \in \mathbf{R}^n} h(y) + \frac{1}{2}\|y - x\|_H^2.$$

Scaled proximal mappings share many properties with (unscaled) proximal mappings.

1. The scaled proximal point $\mathrm{prox}_h^H(x)$ exists and is unique for $x \in \mathrm{dom}\,h$ because the proximity function is strongly convex if H is positive definite.
2. Let $\partial h(x)$ be the subdifferential of h at x. Then $\mathrm{prox}_h^H(x)$ satisfies

$$(2.11) \qquad H\left(x - \mathrm{prox}_h^H(x)\right) \in \partial h\left(\mathrm{prox}_h^H(x)\right).$$

3. Scaled proximal mappings are *firmly nonexpansive* in the H-norm. That is, if $u = \mathrm{prox}_h^H(x)$ and $v = \mathrm{prox}_h^H(y)$, then

$$(u - v)^T H(x - y) \geq \|u - v\|_H^2,$$

and the Cauchy–Schwarz inequality implies $\|u - v\|_H \leq \|x - y\|_H$.

We can express proximal Newton-type search directions as "composite Newton steps" using scaled proximal mappings:

$$(2.12) \qquad \Delta x = \mathrm{prox}_h^H\left(x - H^{-1}\nabla g(x)\right) - x.$$

We use (2.11) to deduce that proximal Newton search directions satisfy

$$H\left(H^{-1}\nabla g(x) - \Delta x\right) \in \partial h(x + \Delta x).$$

We simplify to obtain

$$(2.13) \qquad H\Delta x \in -\nabla g(x) - \partial h(x + \Delta x).$$

Thus proximal Newton-type search directions, like composite gradient steps, combine an explicit gradient with an implicit subgradient. This expression reduces to the Newton system when $h = 0$.

PROPOSITION 2.4 (search direction properties). *If H is positive definite, then Δx in (2.9) satisfies*

$$(2.14) \qquad f(x_+) \leq f(x) + t\left(\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x)\right) + O(t^2),$$

$$(2.15) \qquad \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \leq -\Delta x^T H \Delta x.$$

Proof. For $t \in (0, 1]$,

$$\begin{aligned}
f(x_+) - f(x) &= g(x_+) - g(x) + h(x_+) - h(x) \\
&\leq g(x_+) - g(x) + th(x + \Delta x) + (1 - t)h(x) - h(x) \\
&= g(x_+) - g(x) + t(h(x + \Delta x) - h(x)) \\
&= \nabla g(x)^T (t\Delta x) + t(h(x + \Delta x) - h(x)) + O(t^2),
\end{aligned}$$

which proves (2.14).

Since Δx steps to the minimizer of $\hat{f}$ (2.9), $t\Delta x$ satisfies

$$\nabla g(x)^T \Delta x + \frac{1}{2}\Delta x^T H \Delta x + h(x + \Delta x)$$

$$\leq \nabla g(x)^T (t\Delta x) + \frac{1}{2}t^2 \Delta x^T H \Delta x + h(x_+)$$

$$\leq t\nabla g(x)^T \Delta x + \frac{1}{2}t^2 \Delta x^T H \Delta x + th(x + \Delta x) + (1 - t)h(x).$$

We rearrange and then simplify:

$$(1 - t)\nabla g(x)^T \Delta x + \frac{1}{2}(1 - t^2)\Delta x^T H \Delta x + (1 - t)(h(x + \Delta x) - h(x)) \leq 0,$$

$$\nabla g(x)^T \Delta x + \frac{1}{2}(1 + t)\Delta x^T H \Delta x + h(x + \Delta x) - h(x) \leq 0,$$

$$\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \leq -\frac{1}{2}(1 + t)\Delta x^T H \Delta x.$$

Finally, we let $t \to 1$ and rearrange to obtain (2.15). $\quad\square$

Proposition 2.4 implies the search direction is a descent direction for f because we can substitute (2.15) into (2.14) to obtain

$$(2.16) \qquad f(x_+) \leq f(x) - t\Delta x^T H \Delta x + O(t^2).$$

PROPOSITION 2.5. *Suppose H is positive definite. Then $x^\star$ is an optimal solution if and only if at $x^\star$ the search direction Δx (2.9) is zero.*

Proof. If Δx at $x^\star$ is nonzero, then it is a descent direction for f at $x^\star$. Clearly $x^\star$ cannot be a minimizer of f.

If $\Delta x = 0$, then x is the minimizer of $\hat{f}$, so that

$$\nabla g(x)^T (td) + \frac{1}{2}t^2 d^T H d + h(x + td) - h(x) \geq 0$$

for all $t > 0$ and d. We rearrange to obtain

$$(2.17) \qquad h(x + td) - h(x) \geq -t\nabla g(x)^T d - \frac{1}{2}t^2 d^T H d.$$

Let $Df(x, d)$ be the directional derivative of f at x in the direction d:

$$
\begin{aligned}
Df(x, d) &= \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} \\
&= \lim_{t \to 0} \frac{g(x + td) - g(x) + h(x + td) - h(x)}{t} \\
(2.18) \qquad &= \lim_{t \to 0} \frac{t\nabla g(x)^T d + O(t^2) + h(x + td) - h(x)}{t}.
\end{aligned}
$$

We substitute (2.17) into (2.18) to obtain

$$
\begin{aligned}
Df(x, u) &\geq \lim_{t \to 0} \frac{t\nabla g(x)^T d + O(t^2) - \frac{1}{2}t^2 d^T H d - t\nabla g(x)^T d}{t} \\
&= \lim_{t \to 0} \frac{-\frac{1}{2}t^2 d^T H d + O(t^2)}{t} = 0.
\end{aligned}
$$

Since f is convex, x is an optimal solution if and only if $\Delta x = 0$. □

In a few special cases we can derive a closed-form expression for the proximal Newton search direction, but usually we must resort to an iterative method to solve the subproblem (2.9). The user should choose an iterative method that exploits the properties of h. For example, if h is the ℓ_1 norm, (block) coordinate descent methods combined with an active set strategy are known to be very efficient [10].

We use a line search procedure to select a step length t that satisfies a sufficient descent condition: the next iterate x_+ satisfies $f(x_+) \leq f(x) + \alpha t \lambda$, where

$$(2.19) \qquad \lambda := \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x).$$

The parameter $\alpha \in (0, 0.5)$ can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we will accept. A simple example is a *backtracking line search* [4]: backtrack along the search direction until a suitable step length is selected. Although simple, this procedure performs admirably in practice.

An alternative strategy is to search along the *proximal arc*, i.e., the arc/curve

$$(2.20) \qquad \Delta x_k(t) := \arg\min_y \nabla g(x_k)^T (y - x_k) + \frac{1}{2t}(y - x_k)^T H_k (y - x_k) + h(y).$$

Arc search procedures have some benefits relative to line search procedures. First, the arc search step is the optimal solution to a subproblem. Second, when the optimal solution lies on a low-dimensional manifold of $\mathbf{R}^n$, an arc search strategy is likely to identify this manifold. The main drawback is the cost of obtaining trial points: a subproblem must be solved at each trial point.

LEMMA 2.6. *Suppose $H \succeq mI$ for some $m > 0$ and ∇g is Lipschitz continuous with constant L_1. Then the sufficient descent condition* (2.19) *is satisfied by*

$$(2.21) \qquad t \leq \min\left\{1, \frac{2m}{L_1}(1 - \alpha)\right\}.$$

Proof. We can bound the decrease at each iteration by

$$f(x_+) - f(x)$$
$$= g(x_+) - g(x) + h(x_+) - h(x)$$

$$\leq \int_0^1 \nabla g(x + s(t\Delta x))^T (t\Delta x) ds + t h(x + \Delta x) + (1 - t) h(x) - h(x)$$

$$= \nabla g(x)^T (t\Delta x) + t(h(x + \Delta x) - h(x)) + \int_0^1 (\nabla g(x + s(t\Delta x)) - \nabla g(x))^T (t\Delta x) ds$$

$$\leq t \left(\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) + \int_0^1 \|\nabla g(x + s(\Delta x)) - \nabla g(x)\| \, \|\Delta x\| \, ds \right).$$

Since ∇g is Lipschitz continuous with constant L_1,

$$f(x_+) - f(x) \leq t \left(\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) + \frac{L_1 t}{2} \|\Delta x\|^2 \right)$$

(2.22)
$$= t \left(\lambda + \frac{L_1 t}{2} \|\Delta x\|^2 \right).$$

If we choose $t \leq \frac{2m}{L_1}(1 - \alpha)$, then

$$\frac{L_1 t}{2} \|\Delta x\|^2 \leq m(1 - \alpha) \|\Delta x\|^2 \leq (1 - \alpha) \Delta x^T H \Delta x.$$

By (2.15), we have $\frac{L_1 t}{2} \|\Delta x\|^2 \leq -(1 - \alpha)\lambda$. We substitute this expression into (2.22) to obtain the desired expression:

$$f(x_+) - f(x) \leq t \left(\lambda - (1 - \alpha)\lambda \right) = t(\alpha \lambda). \qquad \square$$

Algorithm 1. A GENERIC PROXIMAL NEWTON-TYPE METHOD.

Require: starting point $x_0 \in \operatorname{dom} f$
1: **repeat**
2:　　Choose H_k, a positive definite approximation to the Hessian.
3:　　Solve the subproblem for a search direction:
　　　　$\Delta x_k \leftarrow \arg\min_d \nabla g(x_k)^T d + \frac{1}{2} d^T H_k d + h(x_k + d)$.
4:　　Select t_k with a backtracking line search.
5:　　Update: $x_{k+1} \leftarrow x_k + t_k \Delta x_k$.
6: **until** stopping conditions are satisfied.

2.3. Inexact proximal Newton-type methods. Inexact proximal Newton-type methods solve the subproblem (2.9) approximately to obtain inexact search directions. These methods can be more efficient than their exact counterparts because they require less computation per iteration. Indeed, many practical implementations of proximal Newton-type methods such as `glmnet`, `newGLMNET`, and QUIC solve the subproblem inexactly. In practice, how inexactly we solve the subproblem is critical to the efficiency and reliability of the method. The practical implementations just mentioned use a variety of heuristics to decide. Although these methods perform admirably in practice, there are few results on how the inexact subproblem solutions affect their convergence behavior.

We now describe an adaptive stopping condition for the subproblem. In section 3 we analyze the convergence behavior of inexact Newton-type methods, and in section 4 we conduct computational experiments to compare the performance of our stopping condition against commonly used heuristics.

Our adaptive stopping condition follows the one used by *inexact Newton-type methods* for minimizing smooth functions:

$$(2.23) \qquad \|\nabla \hat{g}_k(x_k + \Delta x_k)\| \leq \eta_k \|\nabla g(x_k)\|,$$

where η_k is a *forcing term* that requires the left-hand side to be small. We generalize the condition to composite functions by substituting composite gradients into (2.23): if $H_k \preceq MI$ for some $M > 0$, we require

$$(2.24) \qquad \|G_{\hat{f}_k/M}(x_k + \Delta x_k)\| \leq \eta_k \|G_{f/M}(x_k)\|.$$

We set η_k based on how well $\hat{G}_{k-1}$ approximates G near x_k: if $mI \preceq H_k$ for some $m > 0$, we require

$$(2.25) \qquad \eta_k = \min\left\{\frac{m}{2}, \frac{\|G_{\hat{f}_{k-1}/M}(x_k) - G_{f/M}(x_k)\|}{\|G_{f/M}(x_{k-1})\|}\right\}.$$

This choice due to Eisenstat and Walker [9] yields desirable convergence results and performs admirably in practice.

Intuitively, we should solve the subproblem exactly if (i) x_k is close to the optimal solution, and (ii) $\hat{f}_k$ is a good model of f near x_k. If (i), we seek to preserve the fast local convergence behavior of proximal Newton-type methods; if (ii), then minimizing $\hat{f}_k$ is a good surrogate for minimizing f. In these cases, (2.24) and (2.25) are small, so the subproblem is solved accurately.

We can derive an expression like (2.13) for an inexact search direction in terms of an explicit gradient, an implicit subgradient, and a residual term r_k. This reveals connections to the inexact Newton search direction in the case of smooth problems. The adaptive stopping condition (2.24) is equivalent to

$$\begin{aligned}
0 &\in G_{\hat{f}_k}(x_k + \Delta x_k) + r_k \\
&= \nabla \hat{g}_k(x_k + \Delta x_k) + \partial h(x_k + \Delta x_k + G_{\hat{f}_k}(x_k + \Delta x_k)) + r_k \\
&= \nabla g(x_k) + H_k \Delta x_k + \partial h(x_k + \Delta x_k + G_{\hat{f}_k}(x_k + \Delta x_k)) + r_k
\end{aligned}$$

for some r_k such that $\|r_k\| \leq \eta_k \|G_f(x_k)\|$. Thus an inexact search direction satisfies

$$(2.26) \qquad H_k \Delta x_k \in -\nabla g(x_k) - \partial h(x_k + \Delta x_k + G_{\hat{f}_k}(x_k + \Delta x_k)) + r_k.$$

Recently, Byrd, Nocedal, and Oztoprak [5] analyzed the inexact proximal Newton method with a more stringent adaptive stopping condition

$$(2.27) \quad \|G_{\hat{f}_k/M}(x_k + \Delta x_k)\| \leq \eta_k \|G_{f/M}(x_k)\| \text{ and } \hat{f}_k(x_k + \Delta x_k) - \hat{f}_k(x_k) \leq \beta \lambda_k$$

for some $\beta \in (0, \frac{1}{2})$. The second condition is a sufficient descent condition on the subproblem. When h is the ℓ_1 norm, they show that the inexact proximal Newton method with the stopping criterion (2.27)

1. converges globally,
2. eventually accepts the unit step length,
3. converges linearly or superlinearly depending on the forcing terms.

Although the first two results generalize readily to composite functions with a generic h, the third result depends on the separability of the ℓ_1 norm, and does not apply to generic composite functions. Since most practical implementations such as [13] and [31] more closely correspond to (2.24), we state our results for the adaptive stopping condition that does not impose sufficient descent. However our local convergence result combined with their first two results imply the inexact proximal Newton method with stopping condition (2.27) globally converges, and converges linearly or superlinearly (depending on the forcing term) for a generic h.

3. Convergence of proximal Newton-type methods. We now analyze the convergence behavior of proximal Newton-type methods. In section 3.1 we show that proximal Newton-type methods converge globally when the subproblems are solved exactly. In sections 3.2 and 3.3 we show that proximal Newton-type methods and proximal quasi-Newton methods converge q-quadratically and q-superlinearly subject to standard assumptions on the smooth part g. In section 3.4, we show that the inexact proximal Newton method converges

- q-linearly when the forcing terms η_k are uniformly smaller than the inverse of the Lipschitz constant of G_f;
- q-superlinearly when the forcing terms η_k are chosen according to (2.25).

Notation. G_f is the composite gradient step on the composite function f, and λ_k is the decrease in f predicted by linear extrapolation on g at x_k along the search direction Δx_k:

$$\lambda_k := \nabla g(x_k)^T \Delta x_k + h(x_k + \Delta x_k) - h(x_k).$$

L_1, L_2, and L_{G_f} are the Lipschitz constants of ∇g, $\nabla^2 g$, and G_f, respectively, while m and M are the (uniform) strong convexity and smoothness constants for the $\hat{g}_k$'s, i.e., $mI \preceq H_k \preceq MI$. If we set $H_k = \nabla^2 g(x_k)$, then m and M are also the strong convexity and strong smoothness constants of g.

3.1. Global convergence. Our first result shows proximal Newton-type methods converge globally to some optimal solution $x^\star$. There are many similar results; e.g., those in [20, section 4], and Theorem 3.1 is neither the first nor the most general. We include this result because the proof is simple and intuitive. We assume

1. f is a closed, convex function and $\inf_x\{f(x) \mid x \in \operatorname{dom} f\}$ is attained;
2. the H_k's are (uniformly) positive definite; i.e., $H_k \succeq mI$ for some $m > 0$.

The second assumption ensures that the methods are executable, i.e., there exist step lengths that satisfy the sufficient descent condition (cf. Lemma 2.6).

THEOREM 3.1. *Suppose f is a closed convex function, and $\inf_x\{f(x) \mid x \in \operatorname{dom} f\}$ is attained at some $x^\star$. If $H_k \succeq mI$ for some $m > 0$ and the subproblems (2.9) are solved exactly, then x_k converges to an optimal solution starting at any $x_0 \in \operatorname{dom} f$.*

Proof. The sequence $\{f(x_k)\}$ is decreasing because Δx_k are descent directions (2.16) and there exist step lengths satisfying sufficient descent (2.19) (cf. Lemma 2.6):

$$f(x_k) - f(x_{k+1}) \leq \alpha t_k \lambda_k \leq 0.$$

The sequence $\{f(x_k)\}$ must converge to some limit because f is closed and the optimal value is attained. Thus $t_k \lambda_k$ must decay to zero. The step lengths t_k are bounded away from zero because sufficiently small step lengths attain sufficient descent. Thus λ_k must decay to zero. By (2.15), we deduce that Δx_k also converges to zero:

$$\|\Delta x_k\|^2 \leq \frac{1}{m} \Delta x_k^T H_k \Delta x_k \leq -\frac{1}{m} \lambda_k.$$

Since the search direction Δx_k is zero if and only if x is an optimal solution (cf. Proposition 2.5), x_k must converge to some optimal solution $x^\star$.   □

3.2. Local convergence of the proximal Newton method. In this section and section 3.3 we study the convergence rate of the proximal Newton and proximal quasi-Newton methods when the subproblems are solved exactly. First, we state our assumptions on the problem.

DEFINITION 3.2. *A function f is strongly convex with constant m if*

$$(3.1) \qquad f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|^2 \ \text{for any } x, y.$$

We assume the smooth part g is strongly convex with constant m. This is a standard assumption in the analysis of Newton-type methods for minimizing smooth functions. If f is twice continuously differentiable, then this assumption is equivalent to $\nabla^2 f(x) \succeq mI$ for any x. For our purposes, this assumption can usually be relaxed by only requiring (3.1) for any x and y close to $x^\star$.

We also assume the gradient of the smooth part ∇g and Hessian $\nabla^2 g$ are Lipschitz continuous with constants L_1 and L_2. The assumption on $\nabla^2 g$ is standard in the analysis of Newton-type methods for minimizing smooth functions. For our purposes, this assumption can be relaxed by only requiring (2.7) for any x and y close to $x^\star$.

The proximal Newton method uses the exact Hessian $H_k = \nabla^2 g(x_k)$ in the second-order model of f. This method converges q-quadratically,

$$\|x_{k+1} - x^\star\| = O\left(\|x_k - x^\star\|^2\right),$$

subject to standard assumptions on the smooth part: that g is twice continuously differentiable and strongly convex with constant m, and ∇g and $\nabla^2 g$ are Lipschitz continuous with constants L_1 and L_2. We first prove an auxiliary result.

LEMMA 3.3. *If $H_k = \nabla^2 g(x_k)$, the unit step length satisfies the sufficient decrease condition* (2.19) *for k sufficiently large.*

Proof. Since $\nabla^2 g$ is Lipschitz continuous,

$$g(x + \Delta x) \leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 .$$

We add $h(x + \Delta x)$ to both sides to obtain

$$f(x + \Delta x) \leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 + h(x + \Delta x).$$

We then add and subtract h from the right-hand side to obtain

$$\begin{aligned}
f(x + \Delta x) &\leq g(x) + h(x) + \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \\
&\quad + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 \\
&\leq f(x) + \lambda + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 .
\end{aligned}$$

Since g is strongly convex and Δx satisfies (2.15), we have

$$f(x + \Delta x) \leq f(x) + \lambda - \frac{1}{2}\lambda + \frac{L_2}{6m} \|\Delta x\| \lambda.$$

We rearrange to obtain

$$f(x + \Delta x) - f(x) \le \frac{1}{2}\lambda - \frac{L_2}{6m}\lambda \|\Delta x\| \le \left(\frac{1}{2} - \frac{L_2}{6m}\|\Delta x\|\right)\lambda.$$

We can show that Δx_k decays to zero via the argument we used to prove Theorem 3.1. Hence, if k is sufficiently large, $f(x_k + \Delta x_k) - f(x_k) < \frac{1}{2}\lambda_k$. $\square$

THEOREM 3.4. *The proximal Newton method converges q-quadratically to $x^\star$.*

Proof. Since the assumptions of Lemma 3.3 are satisfied, unit step lengths satisfy the sufficient descent condition:

$$x_{k+1} = x_k + \Delta x_k = \text{prox}_h^{\nabla^2 g(x_k)}\left(x_k - \nabla^2 g(x_k)^{-1}\nabla g(x_k)\right).$$

Since scaled proximal mappings are firmly nonexpansive in the scaled norm, we have

$$\begin{aligned}
\|x_{k+1} - x^\star\|_{\nabla^2 g(x_k)} &= \Big\| \text{prox}_h^{\nabla^2 g(x_k)}(x_k - \nabla^2 g(x_k)^{-1}\nabla g(x_k)) \\
&\quad - \text{prox}_h^{\nabla^2 g(x_k)}(x^\star - \nabla^2 g(x_k)^{-1}\nabla g(x^\star))\Big\|_{\nabla^2 g(x_k)} \\
&\le \left\|x_k - x^\star + \nabla^2 g(x_k)^{-1}(\nabla g(x^\star) - \nabla g(x_k))\right\|_{\nabla^2 g(x_k)} \\
&\le \frac{1}{\sqrt{m}}\left\|\nabla^2 g(x_k)(x_k - x^\star) - \nabla g(x_k) + \nabla g(x^\star)\right\|.
\end{aligned}$$

Since $\nabla^2 g$ is Lipschitz continuous, we have

$$\left\|\nabla^2 g(x_k)(x_k - x^\star) - \nabla g(x_k) + \nabla g(x^\star)\right\| \le \frac{L_2}{2}\|x_k - x^\star\|^2$$

and we deduce that x_k converges to $x^\star$ quadratically:

$$\|x_{k+1} - x^\star\| \le \frac{1}{\sqrt{m}}\|x_{k+1} - x^\star\|_{\nabla^2 g(x_k)} \le \frac{L_2}{2m}\|x_k - x^\star\|^2. \qquad \square$$

3.3. Local convergence of proximal quasi-Newton methods. If the sequence $\{H_k\}$ satisfies the Dennis–Moré criterion [8], namely,

$$(3.2) \qquad \frac{\left\|\left(H_k - \nabla^2 g(x^\star)\right)(x_{k+1} - x_k)\right\|}{\|x_{k+1} - x_k\|} \to 0,$$

we can prove that a proximal quasi-Newton method converges q-superlinearly:

$$\|x_{k+1} - x^\star\| \le o(\|x_k - x^\star\|).$$

Again we assume that g is twice continuously differentiable and strongly convex with constant m, and ∇g and $\nabla^2 g$ are Lipschitz continuous with constants L_1 and L_2. These are the assumptions required to prove that quasi-Newton methods for minimizing smooth functions converge superlinearly.

First, we prove two auxiliary results: that (i) step lengths of unity satisfy the sufficient descent condition after sufficiently many iterations, and (ii) the proximal quasi-Newton step is close to the proximal Newton step.

LEMMA 3.5. *If $\{H_k\}$ satisfy the Dennis–Moré criterion and $mI \preceq H_k \preceq MI$ for some $0 < m \le M$, then the unit step length satisfies the sufficient descent condition* (2.19) *after sufficiently many iterations.*

Proof. The proof is very similar to the proof of Lemma 3.3, and we defer the details to Appendix A. □

The proof of the next result mimics the analysis of Tseng and Yun [27].

PROPOSITION 3.6. *Suppose H_1 and H_2 are positive definite matrices with bounded eigenvalues: $mI \preceq H_1 \preceq MI$ and $m_2 I \preceq H_2 \preceq M_2 I$. Let Δx_1 and Δx_2 be the search directions generated using H_1 and H_2 respectively:*

$$\Delta x_1 = \mathrm{prox}_h^{H_1}\left(x - H_1^{-1}\nabla g(x)\right) - x,$$
$$\Delta x_2 = \mathrm{prox}_h^{H_2}\left(x - H_2^{-1}\nabla g(x)\right) - x.$$

Then there is some $\bar{\theta} > 0$ such that these two search directions satisfy

$$\|\Delta x_1 - \Delta x_2\| \leq \sqrt{\frac{1 + \bar{\theta}}{m_1}}\left\|(H_2 - H_1)\Delta x_1\right\|^{1/2}\|\Delta x_1\|^{1/2}.$$

Proof. By (2.9) and Fermat's rule, Δx_1 and Δx_2 are also the solutions to

$$\Delta x_1 = \arg\min_d \nabla g(x)^T d + \Delta x_1^T H_1 d + h(x + d),$$
$$\Delta x_2 = \arg\min_d \nabla g(x)^T d + \Delta x_2^T H_2 d + h(x + d).$$

Thus Δx_1 and Δx_2 satisfy

$$\nabla g(x)^T \Delta x_1 + \Delta x_1^T H_1 \Delta x_1 + h(x + \Delta x_1)$$
$$\leq \nabla g(x)^T \Delta x_2 + \Delta x_2^T H_1 \Delta x_2 + h(x + \Delta x_2)$$

and

$$\nabla g(x)^T \Delta x_2 + \Delta x_2^T H_2 \Delta x_2 + h(x + \Delta x_2)$$
$$\leq \nabla g(x)^T \Delta x_1 + \Delta x_1^T H_2 \Delta x_1 + h(x + \Delta x_1).$$

We sum these two inequalities and rearrange to obtain

$$\Delta x_1^T H_1 \Delta x_1 - \Delta x_1^T (H_1 + H_2)\Delta x_2 + \Delta x_2^T H_2 \Delta x_2 \leq 0.$$

We then complete the square on the left side and rearrange to obtain

$$\Delta x_1^T H_1 \Delta x_1 - 2\Delta x_1^T H_1 \Delta x_2 + \Delta x_2^T H_1 \Delta x_2$$
$$\leq \Delta x_1^T (H_2 - H_1)\Delta x_2 + \Delta x_2^T (H_1 - H_2)\Delta x_2.$$

The left side is $\|\Delta x_1 - \Delta x_2\|_{H_1}^2$ and the eigenvalues of H_1 are bounded. Thus

$$\|\Delta x_1 - \Delta x_2\| \leq \frac{1}{\sqrt{m_1}}\left(\Delta x_1^T (H_2 - H_1)\Delta x_1 + \Delta x_2^T (H_1 - H_2)\Delta x_2\right)^{1/2}$$

$$(3.3) \qquad \leq \frac{1}{\sqrt{m_1}}\left\|(H_2 - H_1)\Delta x_2\right\|^{1/2}(\|\Delta x_1\| + \|\Delta x_2\|)^{1/2}.$$

We use a result due to Tseng and Yun (cf. Lemma 3 in [27]) to bound the term $(\|\Delta x_1\| + \|\Delta x_2\|)$. Let P denote $H_2^{-1/2}H_1 H_2^{-1/2}$. Then $\|\Delta x_1\|$ and $\|\Delta x_2\|$ satisfy

$$\|\Delta x_1\| \leq \left(\frac{M_2\left(1 + \lambda_{\max}(P) + \sqrt{1 - 2\lambda_{\min}(P) + \lambda_{\max}(P)^2}\right)}{2m}\right)\|\Delta x_2\|.$$

We denote the constant in parentheses by $\bar{\theta}$ and conclude that

$$(3.4) \qquad \|\Delta x_1\| + \|\Delta x_2\| \leq (1 + \bar{\theta}) \|\Delta x_2\|.$$

We substitute (3.4) into (3.3) to obtain

$$\|\Delta x_1 - \Delta x_2\|^2 \leq \sqrt{\frac{1 + \bar{\theta}}{m_1}} \left\|(H_2 - H_1)\Delta x_2\right\|^{1/2} \|\Delta x_2\|^{1/2}. \qquad \square$$

We use these two results to show proximal quasi-Newton methods converge superlinearly to $x^\star$ subject to standard assumptions on g and H_k.

THEOREM 3.7. *If $\{H_k\}$ satisfy the Dennis–Moré criterion and $mI \preceq H_k \preceq MI$ for some $0 < m \leq M$, then a proximal quasi-Newton method converges q-superlinearly to $x^\star$.*

Proof. Since the assumptions of Lemma 3.5 are satisfied, unit step lengths satisfy the sufficient descent condition after sufficiently many iterations:

$$x_{k+1} = x_k + \Delta x_k.$$

Since the proximal Newton method converges q-quadratically (cf. Theorem 3.4),

$$\|x_{k+1} - x^\star\| \leq \left\|x_k + \Delta x_k^{\mathrm{nt}} - x^\star\right\| + \left\|\Delta x_k - \Delta x_k^{\mathrm{nt}}\right\|$$
$$(3.5) \qquad \leq \frac{L_2}{m} \left\|x_k^{\mathrm{nt}} - x^\star\right\|^2 + \left\|\Delta x_k - \Delta x_k^{\mathrm{nt}}\right\|,$$

where Δx_k^{nt} denotes the proximal-Newton search direction and $x^{\mathrm{nt}} = x_k + \Delta x_k^{\mathrm{nt}}$. We use Proposition 3.6 to bound the second term:

$$(3.6) \qquad \left\|\Delta x_k - \Delta x_k^{\mathrm{nt}}\right\| \leq \sqrt{\frac{1 + \bar{\theta}}{m}} \left\|(\nabla^2 g(x_k) - H_k)\Delta x_k\right\|^{1/2} \|\Delta x_k\|^{1/2}.$$

Since the Hessian $\nabla^2 g$ is Lipschitz continuous and Δx_k satisfies the Dennis–Moré criterion, we have

$$\left\|\left(\nabla^2 g(x_k) - H_k\right)\Delta x_k\right\| \leq \left\|\left(\nabla^2 g(x_k) - \nabla^2 g(x^\star)\right)\Delta x_k\right\| + \left\|\left(\nabla^2 g(x^\star) - H_k\right)\Delta x_k\right\|$$
$$\leq L_2 \|x_k - x^\star\| \|\Delta x_k\| + o(\|\Delta x_k\|).$$

We know $\|\Delta x_k\|$ is within some constant $\bar{\theta}_k$ of $\|\Delta x_k^{\mathrm{nt}}\|$ (cf. Lemma 3 in [27]). We also know the proximal Newton method converges q-quadratically. Thus

$$\|\Delta x_k\| \leq \bar{\theta}_k \left\|\Delta x_k^{\mathrm{nt}}\right\| = \bar{\theta}_k \left\|x_{k+1}^{\mathrm{nt}} - x_k\right\|$$
$$\leq \bar{\theta}_k \left(\left\|x_{k+1}^{\mathrm{nt}} - x^\star\right\| + \|x_k - x^\star\|\right)$$
$$\leq O\left(\|x_k - x^\star\|^2\right) + \bar{\theta}_k \|x_k - x^\star\|.$$

We substitute these expressions into (3.6) to obtain

$$\left\|\Delta x_k - \Delta x_k^{\mathrm{nt}}\right\| = o(\|x_k - x^\star\|).$$

We substitute this expression into (3.5) to obtain

$$\|x_{k+1} - x^\star\| \leq \frac{L_2}{m} \left\|x_k^{\mathrm{nt}} - x^\star\right\|^2 + o(\|x_k - x^\star\|),$$

and we deduce that x_k converges to $x^\star$ superlinearly. $\quad \square$

3.4. Local convergence of the inexact proximal Newton method. Because subproblem (2.9) is rarely solved exactly, we now analyze the adaptive stopping criterion (2.24):

$$\|G_{\hat{f}_k/M}(x_k + \Delta x_k)\| \leq \eta_k \left\|G_{f/M}(x_k)\right\|.$$

We show that the inexact proximal Newton method with unit step length (i) converges q-linearly if the forcing terms η_k are smaller than some $\bar{\eta}$, and (ii) converges q-superlinearly if the forcing terms decay to zero.

As before, we assume (i) g is twice continuously differentiable and strongly convex with constant m, and (ii) g and $\nabla^2 g$ are Lipschitz continuous with constants L_1 and L_2. We also assume (iii) x_k is close to $x^\star$, and (iv) the unit step length is eventually accepted. These are the assumptions made by Dembo, Eisenstat, and Steihaug [7] and Eisenstat and Walker [9] in their analysis of *inexact Newton methods* for minimizing smooth functions.

First, we prove two auxiliary results that show (i) $G_{\hat{f}_k}$ is a good approximation to G_f, and (ii) $G_{\hat{f}_k}$ inherits the strong monotonicity of $\nabla \hat{g}$.

LEMMA 3.8. *We have* $\|G_f(x) - G_{\hat{f}_k}(x)\| \leq \frac{L_2}{2} \|x - x_k\|^2$.

Proof. The proximal mapping is nonexpansive:

$$\|G_f(x) - G_{\hat{f}_k}(x)\| \leq \|\mathrm{prox}_h(x - \nabla g(x)) - \mathrm{prox}_h(x - \nabla \hat{g}_k(x))\| \leq \|\nabla g(x) - \nabla \hat{g}_k(x)\|.$$

Since $\nabla g(x)$ and $\nabla^2 g(x_k)$ are Lipschitz continuous,

$$\|\nabla g(x) - \nabla \hat{g}_k(x)\| \leq \left\|\nabla g(x) - \nabla g(x_k) - \nabla^2 g(x_k)(x - x_k)\right\| \leq \frac{L_2}{2} \|x - x_k\|^2.$$

Combining the two inequalities gives the desired result. ▯

The proof of the next result mimics the analysis of Byrd, Nocedal, and Ozto-prak [5].

LEMMA 3.9. $G_{tf}(x)$ *with* $t \leq \frac{1}{L_1}$ *is strongly monotone with constant* $\frac{m}{2}$, *i.e.,*

$$(3.7) \qquad (x - y)^T (G_{tf}(x) - G_{tf}(y)) \geq \frac{m}{2} \|x - y\|^2 \ \text{ for } t \leq \frac{1}{L_1}.$$

Proof. The composite gradient step on f has the form

$$G_{tf}(x) = \frac{1}{t}(x - \mathrm{prox}_{th}(x - t\nabla g(x)))$$

(cf. (2.3)). We decompose $\mathrm{prox}_{th}(x - t\nabla g(x))$ (by Moreau's decomposition) to obtain

$$G_{tf}(x) = \nabla g(x) + \frac{1}{t} \mathrm{prox}_{(th)^*}(x - t\nabla g(x)).$$

Thus $G_{tf}(x) - G_{tf}(y)$ has the form

$$G_{tf}(x) - G_{tf}(y)$$
$$(3.8) \qquad = \nabla g(x) - \nabla g(y) + \frac{1}{t}\left(\mathrm{prox}_{(th)^*}(x - t\nabla g(x)) - \mathrm{prox}_{(th)^*}(y - t\nabla g(y))\right).$$

Let $w = \mathrm{prox}_{(th)^*}(x - t\nabla g(x)) - \mathrm{prox}_{(th)^*}(y - t\nabla g(y))$ and

$$d = x - t\nabla g(x) - (y - t\nabla g(y)) = (x - y) - t(\nabla g(x) - t\nabla g(y)).$$

We express (3.8) in terms of $W = \frac{ww^T}{w^T d}$ to obtain

$$G_{tf}(x) - G_{tf}(y) = \nabla g(x) - \nabla g(y) + \frac{w}{t} = \nabla g(x) - \nabla g(y) + \frac{1}{t}Wd.$$

We multiply by $x - y$ to obtain

$$(x - y)^T (G_{tf}(x) - G_{tf}(y))$$
$$= (x - y)^T (\nabla g(x) - \nabla g(y)) + \frac{1}{t}(x - y)^T W d$$

$$(3.9) \qquad = (x - y)^T (\nabla g(x) - \nabla g(y)) + \frac{1}{t}(x - y)^T W(x - y - t(\nabla g(x) - \nabla g(y))).$$

Let $H(\alpha) = \nabla^2 g(x + \alpha(x - y))$. By the mean value theorem, we have

$$(x - y)^T (G_{tf}(x) - G_{tf}(y))$$
$$= \int_0^1 (x - y)^T \left(H(\alpha) - WH(\alpha) + \frac{1}{t}W \right) (x - y)\, d\alpha$$

$$(3.10) \qquad = \int_0^1 (x - y)^T \left(H(\alpha) - \frac{1}{2}(WH(\alpha) + H(\alpha)W) + \frac{1}{t}W \right) (x - y)\, d\alpha.$$

To show (3.7), we must show that $H(\alpha) + \frac{1}{t}W - \frac{1}{2}(WH(\alpha) + H(\alpha)W)$ is positive definite for $t \leq \frac{1}{L_1}$. We rearrange $(\sqrt{t}H(\alpha) - \frac{1}{\sqrt{t}}W)(\sqrt{t}H(\alpha) - \frac{1}{\sqrt{t}}W) \succeq 0$ to obtain

$$tH(\alpha)^2 + \frac{1}{t}W^2 \succeq WH(\alpha) + H(\alpha)W,$$

and we substitute this expression into (3.10) to obtain

$$(x - y)^T (G_{tf}(x) - G_{tf}(y))$$
$$\geq \int_0^1 (x - y)^T \left(H(\alpha) - \frac{t}{2}H(\alpha)^2 + \frac{1}{t}\left(W - \frac{1}{2}W^2 \right) \right) (x - y)\, d\alpha.$$

Since $\text{prox}_{(th)^*}$ is firmly nonexpansive, we have $\|w\|^2 \leq d^T w$ and

$$W = \frac{ww^T}{w^T d} = \frac{\|w\|^2}{w^T d}\frac{ww^T}{\|w\|^2} \preceq I.$$

Since W is positive semidefinite and $W \preceq I$, $W - W^2$ is positive semidefinite and

$$(x - y)^T (G_{tf}(x) - G_{tf}(y)) \geq \int_0^1 (x - y)^T \left(H(\alpha) - \frac{t}{2}H(\alpha)^2 \right) (x - y)\, d\alpha.$$

If we set $t \leq \frac{1}{L_1}$, the eigenvalues of $H(\alpha) - \frac{t}{2}H(\alpha)^2$ are

$$\lambda_i(\alpha) - \frac{t}{2}\lambda_i(\alpha)^2 \geq \lambda_i(\alpha) - \frac{\lambda_i(\alpha)^2}{2L_1} \geq \frac{\lambda_i(\alpha)}{2} > \frac{m}{2},$$

where $\lambda_i(\alpha), i = 1, \ldots, n$ are the eigenvalues of $H(\alpha)$. We deduce that

$$(x - y)^T (G_{tf}(x) - G_{tf}(y)) \geq \frac{m}{2}\|x - y\|^2. \qquad \square$$

We use these two results to show that the inexact proximal Newton method with unit step lengths converges locally linearly or superlinearly depending on the forcing terms.

THEOREM 3.10. *Suppose x_0 is sufficiently close to $x^\star$.*

1. *If η_k is smaller than some $\bar{\eta} < \frac{m}{2}$, an inexact proximal Newton method with unit step lengths converges q-linearly to $x^\star$.*
2. *If η_k decays to zero, an inexact proximal Newton method with unit step lengths converges q-superlinearly to $x^\star$.*

Proof. The local model $\hat{f}_k$ is strongly convex with constant m. According to Lemma 3.9, $G_{\hat{f}_k/L_1}$ is strongly monotone with constant $\frac{m}{2}$:

$$(x - y)^T \left(G_{\hat{f}_k/L_1}(x) - G_{\hat{f}_k/L_1}(y) \right) \geq \frac{m}{2} \|x - y\|^2 .$$

By the Cauchy–Schwarz inequality, we have

$$\|G_{\hat{f}_k/L_1}(x) - G_{\hat{f}_k/L_1}(y)\| \geq \frac{m}{2} \|x - y\| .$$

We apply this result to $x_k + \Delta x_k$ and $x^\star$ to obtain

$$(3.11) \quad \|x_{k+1} - x^\star\| = \|x_k + \Delta x_k - x^\star\| \leq \frac{2}{m} \|G_{\hat{f}_k/L_1}(x_k + \Delta x_k) - G_{\hat{f}_k/L_1}(x^\star)\|.$$

Let r_k be the residual $-G_{\hat{f}_k/L_1}(x_k + \Delta x_k)$. The adaptive stopping condition (2.24) requires $\|r_k\| \leq \eta_k \|G_{f/L_1}(x_k)\|$. We substitute this expression into (3.11) to obtain

$$
\begin{aligned}
\|x_{k+1} - x^\star\| &\leq \frac{2}{m} \| - G_{\hat{f}_k/L_1}(x^\star) - r_k\| \\
&\leq \frac{2}{m} (\|G_{\hat{f}_k/L_1}(x^\star)\| + \|r_k\|) \\
(3.12) \qquad &\leq \frac{2}{m} (\|G_{\hat{f}_k/L_1}(x^\star)\| + \eta_k \|G_{f/L_1}(x_k)\|) .
\end{aligned}
$$

Applying Lemma 3.8 to f/L_1 and $\hat{f}_k/L_1$ gives

$$\|G_{\hat{f}_k/L_1}(x^\star)\| \leq \frac{1}{2} \frac{L_2}{L_1} \|x_k - x^\star\|^2 + \|G_{f/L_1}(x^\star)\| = \frac{1}{2} \frac{L_2}{L_1} \|x_k - x^\star\|^2 .$$

We substitute this bound into (3.12) to obtain

$$
\begin{aligned}
\|x_{k+1} - x^\star\| &\leq \frac{2}{m} \left(\frac{L_2}{2L_1} \|x_k - x^\star\|^2 + \eta_k \|G_{f/L_1}(x_k)\| \right) \\
&\leq \frac{L_2}{mL_1} \|x_k - x^\star\|^2 + \frac{2\eta_k}{m} \|x_k - x^\star\| .
\end{aligned}
$$

We deduce that (i) x_k converges q-linearly to $x^\star$ if x_0 is sufficiently close to $x^\star$ and $\eta_k \leq \bar{\eta}$ for some $\bar{\eta} < \frac{m}{2}$, and (ii) x_k converges q-superlinearly to $x^\star$ if x_0 is sufficiently close to $x^\star$ and η_k decays to zero. ☐

Finally, we justify our choice of forcing terms: if we choose η_k according to (2.25), then the inexact proximal Newton method converges q-superlinearly. When minimizing smooth functions, we recover the result of Eisenstat and Walker on choosing forcing terms in an inexact Newton method [9].

THEOREM 3.11. *Suppose x_0 is sufficiently close to $x^\star$. If we choose η_k according to (2.25), then the inexact proximal Newton method with unit step lengths converges q-superlinearly.*

Proof. To show superlinear convergence, we must show

$$(3.13) \qquad \frac{\|G_{\hat{f}_{k-1}/L_1}(x_k) - G_{f/L_1}(x_k)\|}{\|G_{f/L_1}(x_{k-1})\|} \to 0.$$

By Lemma 2.2, we have

$$\|G_{\hat{f}_{k-1}/L_1}(x_k) - G_{f/L_1}(x_k)\| \le \frac{1}{2}\frac{L_2}{L_1}\|x_k - x_{k-1}\|^2$$

$$\le \frac{1}{2}\frac{L_2}{L_1}\left(\|x_k - x^\star\| + \|x^\star - x_{k-1}\|\right)^2.$$

By Lemma 3.9, we also have

$$\left\|G_{f/L_1}(x_{k-1})\right\| = \left\|G_{f/L_1}(x_{k-1}) - G_{f/L_1}(x^\star)\right\| \ge \frac{m}{2}\|x_{k-1} - x^\star\|.$$

We substitute these expressions into (3.13) to obtain

$$\frac{\|G_{\hat{f}_{k-1}/L_1}(x_k) - G_{f/L_1}(x_k)\|}{\|G_{f/L_1}(x_{k-1})\|}$$

$$\le \frac{\frac{1}{2}\frac{L_2}{L_1}\left(\|x_k - x^\star\| + \|x^\star - x_{k-1}\|\right)^2}{\frac{m}{2}\|x_{k-1} - x^\star\|}$$

$$= \frac{1}{m}\frac{L_2}{L_1}\frac{\|x_k - x^\star\| + \|x_{k-1} - x^\star\|}{\|x_{k-1} - x^\star\|}\left(\|x_k - x^\star\| + \|x_{k-1} - x^\star\|\right)$$

$$= \frac{1}{m}\frac{L_2}{L_1}\left(1 + \frac{\|x_k - x^\star\|}{\|x_{k-1} - x^\star\|}\right)\left(\|x_k - x^\star\| + \|x_{k-1} - x^\star\|\right).$$

By Theorem 3.10, we have $\frac{\|x_k - x^\star\|}{\|x_{k-1} - x^\star\|} < 1$ and

$$\frac{\|G_{\hat{f}_{k-1}/M}(x_k) - G_{f/M}(x_k)\|}{\|G_{f/M}(x_{k-1})\|} \le \frac{2}{m}\frac{L_2}{M}\left(\|x_k - x^\star\| + \|x_{k-1} - x^\star\|\right).$$

We deduce (with Theorem 3.10) that the inexact proximal Newton method with adaptive stopping condition (2.24) converges *q*-superlinearly. □

4. Computational experiments. First we explore how inexact search directions affect the convergence behavior of proximal Newton-type methods on a problem in bioinfomatics. We show that choosing the forcing terms according to (2.25) avoids "oversolving" the subproblem. Then we demonstrate the performance of proximal Newton-type methods using a problem in statistical learning. We show that the methods are suited to problems with expensive smooth function evaluations.

4.1. Inverse covariance estimation. Suppose independently and identically distributed samples $x^{(1)}, \ldots, x^{(m)}$ are from a Gaussian Markov random field with mean zero and unknown inverse covariance matrix $\bar{\Theta}$:

$$\mathbf{Pr}(x; \bar{\Theta}) \propto \exp(x^T\bar{\Theta}x/2 - \log\det(\bar{\Theta})).$$

We seek a sparse maximum likelihood estimate of the inverse covariance matrix:

$$(4.1) \qquad \hat{\Theta} := \underset{\Theta \in \mathbf{R}^{n \times n}}{\arg\min}\ \mathrm{tr}\left(\hat{\Sigma}\Theta\right) - \log\det(\Theta) + \lambda\left\|\mathrm{vec}(\Theta)\right\|_1,$$

where $\hat{\Sigma}$ denotes the sample covariance matrix. We regularize using an entrywise ℓ_1 norm to avoid overfitting the data and to promote sparse estimates. The parameter λ balances goodness of fit and sparsity.

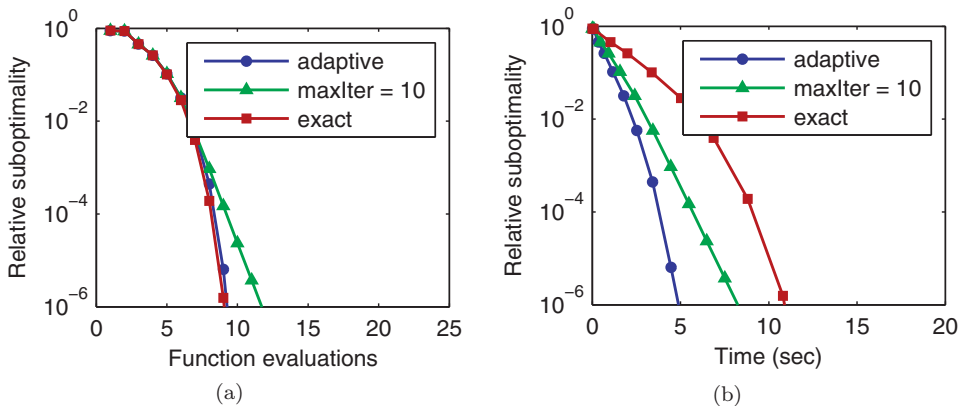(a)   (b)

FIG. 1. *Inverse covariance estimation problem (Estrogen dataset). Convergence behavior of proximal BFGS method with three subproblem stopping conditions.*
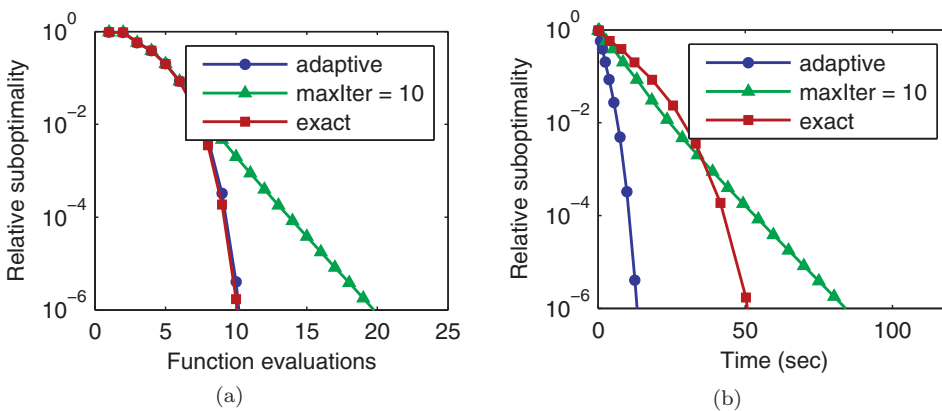

(a)   (b)

FIG. 2. *Inverse covariance estimation problem (Leukemia dataset). Convergence behavior of proximal BFGS method with three subproblem stopping conditions.*

We use two datasets: (i) Estrogen, a gene expression dataset consisting of 682 probe sets collected from 158 patients, and (ii) Leukemia, another gene expression dataset consisting of 1255 genes from 72 patients.[1] The features of Estrogen were converted to log scale and normalized to have zero mean and unit variance. The regularization parameter λ was chosen to match the values used in [21].

We solve the inverse covariance estimation problem (4.1) using a proximal BFGS method, i.e., H_k is updated according to the BFGS updating formula. (The proximal Newton method would be computationally very expensive on these large datasets.) To explore how inexact search directions affect the convergence behavior, we use three rules to decide how accurately to solve subproblem (2.9).

1. Adaptive: stop when the adaptive stopping condition (2.24) is satisfied;
2. exact: solve the subproblem accurately ("exactly");
3. stop after 10 iterations.

We use the TFOCS implementation of FISTA to solve the subproblem. We plot relative suboptimality versus function evaluations and time on the Estrogen dataset in Figure 1 and on the Leukemia dataset in Figure 2.

---

[1]These datasets are available from http://www.math.nus.edu.sg/∼mattohkc/ with the SPINCOVSE package.

Although the conditions for superlinear convergence (cf. Theorem 3.7) are not met ($\log \det$ is not strongly convex), we empirically observe in Figures 1 and 2 that a proximal BFGS method transitions from linear to superlinear convergence. This transition is characteristic of BFGS and other quasi-Newton methods with superlinear convergence.

On both datasets, the exact stopping condition yields the fastest convergence (ignoring computational expense per step), followed closely by the adaptive stopping condition (see Figures 1 and 2). If we account for time per step, then the adaptive stopping condition yields the fastest convergence. Note that the adaptive stopping condition yields superlinear convergence (like the exact proximal BFGS method). The third condition (stop after 10 iterations) yields only linear convergence (like a first-order method), and its convergence rate is affected by the condition number of $\hat{\Theta}$. On the Leukemia dataset, the condition number is worse and the convergence is slower.

4.2. Logistic regression. Suppose we are given samples $x^{(1)}, \ldots, x^{(m)}$ with labels $y^{(1)}, \ldots, y^{(m)} \in \{-1, 1\}$. We fit a logit model to our data:

$$(4.2) \qquad \underset{w \in \mathbf{R}^n}{\text{minimize}} \ \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i w^T x_i)) + \lambda \left\| w \right\|_1 .$$

Again, the regularization term $\left\| w \right\|_1$ promotes sparse solutions and λ balances sparsity with goodness of fit.

We use two datasets: (i) `gisette`, a handwritten digits dataset from the NIPS 2003 feature selection challenge ($n = 5000$), and (ii) `rcv1`, an archive of categorized news stories from Reuters ($n = 47000$).[2] The features of `gisette` have been scaled to be within the interval $[-1, 1]$, and those of `rcv1` have been scaled to be unit vectors. λ matched the value reported in [31], where it was chosen by fivefold cross validation on the training set.

We compare a proximal L-BFGS method with SpaRSA and the TFOCS implementation of FISTA (also Nesterov's 1983 method) on problem (4.2). We plot relative suboptimality versus function evaluations and time on the `gisette` dataset in Figure 3 and on the `rcv1` dataset in Figure 4.

The smooth part of the function requires many expensive exp/log evaluations. On the dense `gisette` dataset (30 million nonzero entries in a 6000×5000 design matrix), evaluating g dominates the computational cost. The proximal L-BFGS method clearly outperforms the other methods because the computational expense is shifted to solving the subproblems, whose objective functions are cheap to evaluate (see Figure 3). On the sparse `rcv1` dataset (40 million nonzero entries in a 542000×47000 design matrix), the evaluation of g makes up a smaller portion of the total cost, and the proximal L-BFGS method barely outperforms SpaRSA (see Figure 4).

4.3. Software. PNOPT. The methods described have been incorporated into a MATLAB package PNOPT (Proximal Newton OPTimizer, pronounced pee en opt) and are publicly available from the Systems Optimization Laboratory (SOL).[3] PNOPT shares an interface with the software package TFOCS [2] and is compatible with the function generators included with TFOCS. We refer to the SOL website for details about PNOPT.
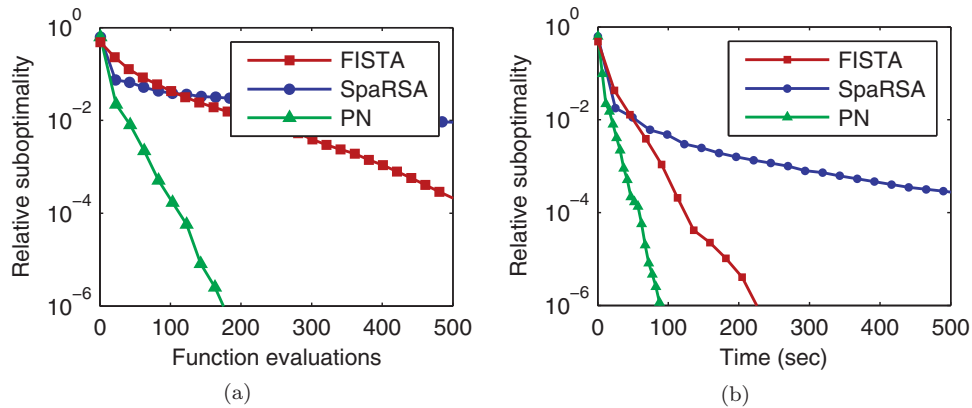

FIG. 3. *Logistic regression problem (`gisette` dataset). Proximal L-BFGS method ($L = 50$) versus FISTA and SpaRSA.*
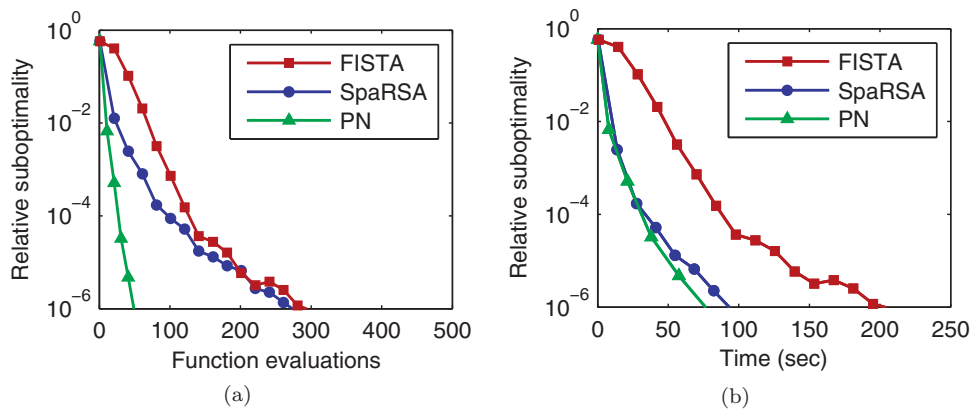

FIG. 4. *Logistic regression problem (`rcv1` dataset). Proximal L-BFGS method ($L = 50$) versus FISTA and SpaRSA.*

5. Conclusion. Given the popularity of first-order methods for minimizing composite functions, there has been a flurry of activity around the development of Newton-type methods for minimizing composite functions [13, 3, 18]. We analyze proximal Newton-type methods for such functions and show that they have several benefits over first-order methods:

1. they converge rapidly near the optimal solution, and can produce a solution of high accuracy;
2. they scale well with problem size;
3. the proximal Newton method is insensitive to the choice of coordinate system and to the condition number of the level sets of the objective.

Proximal Newton-type methods can readily handle composite functions where g is not convex, although care must be taken to ensure $\hat{g}_k$ remains strongly convex. The convergence analysis could be modified to give global convergence (to stationary points) and convergence rates near stationary points. We defer these extensions to future work.

The main disadvantage of proximal Newton-type methods is the cost of solving the subproblems. We have shown that it is possible to reduce the cost and retain

the fast convergence rate by solving the subproblems inexactly. We hope our results will kindle further interest in proximal Newton-type methods as an alternative to first-order methods and interior point methods for minimizing composite functions.

Appendix A. Proof of Lemma 3.5.

LEMMA A.1. *Suppose* (i) g *is twice continuously differentiable and strongly convex with constant* m, *and* (ii) $\nabla^2 g$ *is Lipschitz continuous with constant* L_2. *If the sequence* $\{H_k\}$ *satisfies the Dennis–Moré criterion and* $mI \preceq H_k \preceq MI$ *for some* $0 < m \le M$, *then the unit step length satisfies the sufficient descent condition* (2.19) *after sufficiently many iterations.*

Proof. Since $\nabla^2 g$ is Lipschitz,

$$g(x + \Delta x) \le g(x) + \nabla g(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 g(x)\Delta x + \frac{L_2}{6}\|\Delta x\|^3.$$

We add $h(x + \Delta x)$ to both sides to obtain

$$f(x + \Delta x) \le g(x) + \nabla g(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 g(x)\Delta x$$
$$+ \frac{L_2}{6}\|\Delta x\|^3 + h(x + \Delta x).$$

We then add and subtract $h(x)$ from the right-hand side to obtain

$$f(x + \Delta x) \le g(x) + h(x) + \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x)$$
$$+ \frac{1}{2}\Delta x^T \nabla^2 g(x)\Delta x + \frac{L_2}{6}\|\Delta x\|^3$$
$$\le f(x) + \lambda + \frac{1}{2}\Delta x^T \nabla^2 g(x)\Delta x + \frac{L_2}{6}\|\Delta x\|^3$$
$$\le f(x) + \lambda + \frac{1}{2}\Delta x^T \nabla^2 g(x)\Delta x + \frac{L_2}{6m}\|\Delta x\|\lambda,$$

where we use (2.15). We add and subtract $\frac{1}{2}\Delta x^T H \Delta x$ to yield

$$f(x + \Delta x) \le f(x) + \lambda + \frac{1}{2}\Delta x^T \left(\nabla^2 g(x) - H\right)\Delta x + \frac{1}{2}\Delta x^T H \Delta x + \frac{L_2}{6m}\|\Delta x\|\lambda$$

(A.1)
$$\le f(x) + \lambda + \frac{1}{2}\Delta x^T \left(\nabla^2 g(x) - H\right)\Delta x - \frac{1}{2}\lambda + \frac{L_2}{6m}\|\Delta x\|\lambda,$$

where we again use (2.15). Since $\nabla^2 g$ is Lipschitz continuous and the search direction Δx satisfies the Dennis–Moré criterion,

$$\frac{1}{2}\Delta x^T \left(\nabla^2 g(x) - H\right)\Delta x$$
$$= \frac{1}{2}\Delta x^T \left(\nabla^2 g(x) - \nabla^2 g(x^\star)\right)\Delta x + \frac{1}{2}\Delta x^T \left(\nabla^2 g(x^\star) - H\right)\Delta x$$
$$\le \frac{1}{2}\left\|\nabla^2 g(x) - \nabla^2 g(x^\star)\right\|\|\Delta x\|^2 + \frac{1}{2}\left\|\left(\nabla^2 g(x^\star) - H\right)\Delta x\right\|\|\Delta x\|$$
$$\le \frac{L_2}{2}\|x - x^\star\|\|\Delta x\|^2 + o\left(\|\Delta x\|^2\right).$$

We substitute this expression into (A.1) and rearrange to obtain

$$f(x + \Delta x) \le f(x) + \frac{1}{2}\lambda + o\left(\|\Delta x\|^2\right) + \frac{L_2}{6m}\|\Delta x\|\lambda.$$

We can show Δx_k converges to zero via the argument used in the proof of Theorem 3.1. Hence, for k sufficiently large, $f(x_k + \Delta x_k) - f(x_k) \le \frac{1}{2}\lambda_k$. $\square$

REFERENCES

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

[2] S.R. Becker, E.J. Candès, and M.C. Grant, *Templates for convex cone problems with applications to sparse signal recovery*, Math. Program. Comput., 3 (2011), pp. 165–218.

[3] S. Becker and M.J. Fadili, *A quasi-Newton proximal splitting method*, in Advances in Neural Information Processing Systems 25, Lake Tahoe, NY, 2012.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[5] R.H. Byrd, J. Nocedal, and F. Oztoprak, *An Inexact Successive Quadratic Approximation Method for Convex l-1 Regularized Optimization*, preprint, arXiv:1309.3529, 2013.

[6] E.J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[7] R.S. Dembo, S.C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[8] J.E. Dennis and J.J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[9] S.C. Eisenstat and H.F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[10] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, *Pathwise coordinate optimization*, Ann. Appl. Statist., 1 (2007), pp. 302–332.

[11] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostat., 9 (2008), pp. 432–441.

[12] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain non-convex minimization problems*, Internat. J. Syst. Sci., 12 (1981), pp. 989–1000.

[13] C.J. Hsieh, M.A. Sustik, I.S. Dhillon, and P. Ravikumar, *Sparse inverse covariance matrix estimation using quadratic approximation*, in Advances in Neural Information Processing Systems 24, Granada, Spain, 2011.

[14] D. Kim, S. Sra, and I.S. Dhillon, *A scalable trust-region algorithm with application to mixed-norm regression*, in Proceedings of 27th International Converence on Machine Learning, Haifa, Israel, Omnipress, Madison, WI, 2010, pp. 519–526.

[15] J.D. Lee, Y. Sun, and M.A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, in Advances in Neural Information Processing Systems 25, Lake Tahoe, NV, 2012, pp. 827–835.

[16] Z. Lu and Y. Zhang, *An augmented Lagrangian approach for sparse principal component analysis*, Math. Program. Ser. A, 135 (2011), pp. 149–193.

[17] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer, Boston, 2003.

[18] P.A. Olsen, F. Oztoprak, J. Nocedal, and S.J. Rennie, *Newton-like methods for sparse inverse covariance estimation*, in Advances in Neural Information Processing Systems 25, Lake Tahoe, NV, 2012.

[19] M. Patriksson, *Cost approximation: a unified framework of descent algorithms for nonlinear programs*, SIAM J. Optim., 8 (1998), pp. 561–582.

[20] M. Patriksson, Programming and Variational Inequality Problems: A Unified Approach, Kluwer Academic, Dordrecht, 1999.

[21] B. Rolfs, B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki, *Iterative thresholding algorithm for sparse inverse covariance estimation*, in Advances in Neural Information Processing Systems 25, Lake Tahoe, NV, 2012, pp. 1583–1591.

[22] M. Schmidt, *Graphical Model Structure Learning with ℓ_1-Regularization*, Ph.D. thesis, University of British Columbia, Vancouver, Canada, 2010.

[23] M. Schmidt, D. Kim, and S. Sra, *Projected Newton-type methods in machine learning*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S.J. Wright, eds., MIT Press, Cambridge, MA, 2011.

[24] M. Schmidt, E. Van Den Berg, M.P. Friedlander, and K. Murphy, *Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm*, in J. Mach. Learn. Res., 5 (2009), pp. 456–463.

[25] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 58 (1996), pp. 267–288.

[26] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, manuscript, 2009.

[27] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Prog. Ser. B, 117 (2009), pp. 387–423.

[28] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation*, SIAM J. Sci. Comput., 32 (2010), pp. 1832–1857.

[29] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.

[30] J. Yu, S.V.N. Vishwanathan, S. Günter, and N.N. Schraudolph, *A quasi-Newton approach to nonsmooth convex optimization problems in machine learning*, J. Mach. Learn. Res., 11 (2010), pp. 1145–1200.

[31] G.X. Yuan, C.H. Ho, and C.J. Lin, *An improved glmnet for ℓ_1-regularized logistic regression*, J. Mach. Learn. Res., 13 (2012), pp. 1999–2030.