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

Inexact proximal Newton methods for self-concordant functions

Jinchao Li · Martin S. Andersen · Lieven Vandenberghe

October 1, 2016

Abstract We analyze the proximal Newton method for minimizing a sum of a self-concordant function and a convex function with an inexpensive proximal operator. We present new results on the global and local convergence of the method when inexact search directions are used. The method is illustrated with an application to L1-regularized covariance selection, in which prior constraints on the sparsity pattern of the inverse covariance matrix are imposed. In the numerical experiments the proximal Newton steps are computed by an accelerated proximal gradient method, and multifrontal algorithms for positive definite matrices with chordal sparsity patterns are used to evaluate gradients and matrix-vector products with the Hessian of the smooth component of the objective.

Keywords Proximal Newton method, self-concordance, convex optimization, chordal sparsity, covariance selection

1 Introduction

The *proximal Newton algorithm* is a method for solving optimization problems

$$\text{minimize} \quad f(x) = g(x) + h(x) \tag{1}$$

Jinchao Li
Electrical Engineering Department, University of California, Los Angeles.
E-mail: lijinchao@ucla.edu

Martin Andersen
Department of Applied Mathematics and Computer Science, Technical University of Denmark.
E-mail: mskan@dtu.dk

Lieven Vandenberghe
Electrical Engineering Department, University of California, Los Angeles.
E-mail: vandenbe@ucla.edu

This work was supported by National Science Foundation Grants 1128817 and 1509789.

with g convex and twice continuously differentiable, and h convex and possibly nondifferentiable. At each iteration of the algorithm, an update $x := x+\alpha v(x)$ is made, where α is a positive stepsize and $v(x)$ is the *proximal Newton step* at x, defined as

$$v(x) = \arg\min_v \left(g(x) + \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x+v) \right). \tag{2}$$

The vector $x + v(x)$ minimizes an approximation

$$\hat{f}_x(y) = g(x) + \nabla g(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 g(x)(y-x) + h(y) \tag{3}$$

of the cost function f, obtained by replacing g with a second-order approximation around x. For this reason the algorithm is also called a *successive quadratic approximation method* [8]. When h is zero, the proximal Newton step is $v(x) = -\nabla^2 g(x)^{-1} \nabla g(x)$ and the proximal Newton method reduces to the standard Newton method for minimizing $g(x)$.

The proximal Newton method and some of its variants have recently been studied for applications in statistics and machine learning, in which $h(x)$ is an ℓ_1-norm penalty, added to a differentiable objective to promote sparsity in the solution [8, 15–17, 22, 27, 28]. This approach is motivated by the fact that the optimization problem in (2) is a 'lasso' problem (minimization of a convex quadratic function plus an ℓ_1-norm) that can be solved by efficient iterative algorithms. More generally, the proximal Newton method is interesting when h has an inexpensive proximal operator, so the subproblem in (2) can be solved by proximal gradient methods.

With exact steps $v(x)$, the proximal Newton algorithm is known to have the same excellent convergence properties as the Newton method for smooth unconstrained minimization: fast local convergence, and global convergence from any starting point if a proper stepsize selection is used. Moreover, in constrast to many other nonsmooth optimization algorithms, the same line search strategies can be adopted as for the unconstrained Newton method. These convergence properties are discussed in [17] under the assumptions that g is strongly convex with Lipschitz continuous gradient, and in [16, 27, 28] for self-concordant functions g.

In practice, it is expensive to compute the proximal Newton step accurately, since $v(x)$ is found by minimizing (3) numerically. This is a fundamental difference with the standard Newton method. It is therefore important to understand the convergence of the proximal Newton method with inexact steps [8, 22, 26]. Lee, Sun, and Saunders [17, page 1428] propose the following criterion for accepting an approximation v of (2). A vector v is accepted as an approximate proximal Newton step at x if it satisfies

$$\|\hat{F}_{x,t}(x+v)\| \leq \eta_{\mathrm{f}} \|F_t(x)\| \tag{4}$$

where $t \leq 1/\lambda_{\max}(\nabla^2 g(x))$, and F_t, $\hat{F}_{x,t}$ are the *gradient mappings* [20, section 2.2.3] of the cost function f and its local approximation $\hat{f}_x$, respectively, *i.e.*,

$$F_t(x) = \frac{1}{t}\left(x - \mathbf{prox}_{th}(x - t\nabla g(x))\right),$$
$$\hat{F}_{x,t}(x+v) = \frac{1}{t}\left(x + v - \mathbf{prox}_{th}\left(x + v - t(\nabla g(x) + \nabla^2 g(x)v)\right)\right),$$

where $\mathbf{prox}_{th}$ denotes the proximal operator (defined in (12)). When $h(x) = 0$, these definitions reduce to $F_t(x) = \nabla g(x)$ and $\hat{F}_{x,t}(x+v) = \nabla g(x) + \nabla^2 g(x)v$, and the inequality (4) to a classical condition in the literature on inexact Newton methods [10,12]. The *forcing term* η_{f} in (4) can be adjusted adaptively to obtain superlinear local convergence. Byrd, Nocedal, and Oztoprak [8] use a similar condition, but also impose the condition

$$\hat{f}_x(x+v) - f(x) \leq \beta\left(\nabla g(x)^T v + h(x+v) - h(x)\right)$$

with $\beta \in (0, 1/2)$ and show that this ensures global convergence when g is strongly convex with a Lipschitz continuous gradient. The papers [8, 17, 28] also analyze variable metric or quasi-Newton methods, in which approximate Hessians are used in the approximation (3). The effect of inexactness on the proximal Newton method with a self-concordant function g is discussed in [16, 27]. In this analysis, inexactness is measured by the suboptimality (in function value) of the approximate solution of (2).

In the first part of this paper (sections 2–4) we extend the results of [28] for the (exact) proximal Newton method for self-concordant functions g to the proximal Newton method with inexact steps. In the algorithms we analyze, the condition (4) is replaced by the following criterion: a step v is accepted as an approximation of $v(x)$ if a residual

$$r \in \nabla g(x) + \nabla^2 g(x)v + \partial h(x+v)$$

in the optimality conditions for (2) is known that satisfies the inequality

$$\|\nabla^2 g(x)^{-1/2} r\| \leq (1-\theta)\|\nabla^2 g(x)^{1/2} v\|.$$

We show that if g is self-concordant, then the inexact proximal Newton method converges globally if a damped stepsize or backtracking line search is used. The parameter $1 - \theta$ plays a role similar to the forcing term η_{f} in (4). We show that the local convergence is quadratic if $\theta = 1$, linear if θ constant and less than one, and superlinear if θ approaches one as the algorithm converges.

The composite optimization problem (1) with self-concordant functions g has important applications in machine learning [28]. The proximal Newton method that we develop in sections 2–4 is motivated by an application to sparse inverse covariance selection. In this problem, the smooth component g is self-concordant, but it is not strongly convex and its gradient is not Lipschitz continuous on its entire domain. Moreover, in the large sparse setting that we describe in section 5, matrix-vector products with the Hessian $\nabla^2 g(x)v$ or the inverse Hessian $\nabla^2 g(x)^{-1}w$ can be computed quite efficiently, at roughly the

same cost as the gradient $\nabla g(x)$. These properties make it possible to compute a sufficiently accurate approximate Newton step by applying a proximal gradient method to minimize (3). This is described in more detail in section 5.

The rest of the paper is organized as follows. In section 2 we first review the definition and key properties of self-concordant functions, and present a theorem that provides bounds on the optimum of (1) in terms of the magnitude of inexact proximal Newton steps. In sections 3 and 4 we discuss the proximal Newton method with a damped stepsize and a backtracking line search, respectively, and give global and local convergence results that account for inexactness of the search directions. In section 5 we discuss the application to covariance selection and present some numerical results.

2 Proximal Newton step for self-concordant functions

We consider unconstrained optimization problems of the form (1) with $g : \mathbf{R}^n \to \mathbf{R}$ a self-concordant function and $h : \mathbf{R}^n \to \mathbf{R}$ a closed, convex, and possibly nondifferentiable function. We assume the problem is feasible ($\mathbf{dom}\, f = \mathbf{dom}\, g \cap \mathbf{dom}\, h \neq \emptyset$). This implies that the sum $f = g + h$ is a closed function (see, for example, [14, page 158]).

2.1 Self-concordance

Specifically, we make the following assumptions about g.

- g is closed, convex, with open domain.
- g is three times continuously differentiable with $\nabla^2 g(x)$ positive definite on $\mathbf{dom}\, g$.
- The Hessian of g satisfies the matrix inequality

$$\frac{d}{d\alpha} \nabla^2 g(x + \alpha v)\Big|_{\alpha=0} \preceq 2\|v\|_x \nabla^2 g(x) \tag{5}$$

  for all $x \in \mathbf{dom}\, g$ and all v, where $\|v\|_x = (v^T \nabla^2 g(x) v)^{1/2}$. (The inequality $A \preceq B$ means $B - A$ is positive semidefinite.)

These properties characterize self-concordant functions as defined by Renegar [23] and Nesterov [20]. They define a subclass of the self-concordant functions introduced in [21]: in Nesterov and Nemirovski's book, closed self-concordant functions are called *strongly* self-concordant, self-concordant functions with nonsingular Hessians are called *nondegenerate*, and the fundamental inequality (5) includes a scaling parameter a that we take to be one. Nesterov [20, page 181] refers to self-concordant functions with $a = 1$ as *standard* self-concordant functions. We do not assume that g is a self-concordant *barrier* (*i.e.*, has the property that $\nabla g(x)^T \nabla^2 g(x)^{-1} \nabla g(x)$ is bounded on $\mathbf{dom}\, g$; see [21, definition 2.3.1]).

For future reference, we list the properties of self-concordant functions that are used in the paper.

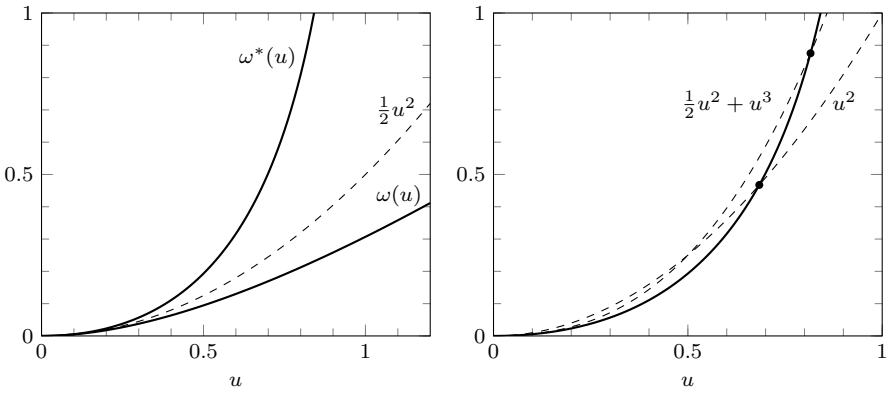

Fig. 1 *Left.* The functions $\omega(u) = u - \log(1+u)$ and $\omega^*(u) = -u - \log(1-u)$. *Right.* The function $\omega^*(u)$ in solid line, with two upper bounds $\omega^*(u) \leq u^2$ for $u \leq 0.68$ and $\omega^*(u) \leq u^2/2 + u^3$ for $u \leq 0.81$.

– *Bounds on Hessian* [21, theorem 2.1.1]. If $x, y \in \mathbf{dom}\, g$ and $\|y - x\|_x < 1$, then

$$(1 - \|y-x\|_x)^2 \nabla^2 g(x) \preceq \nabla^2 g(y) \preceq \frac{1}{(1-\|y-x\|_x)^2} \nabla^2 g(x). \tag{6}$$

– *Bounds on gradient* [19, lemma 1]. If $x, y \in \mathbf{dom}\, g$ and $\|x - y\|_x < 1$, then

$$\|\nabla g(y) - \nabla g(x) - \nabla^2 g(x)(y-x)\|_{x*} \leq \frac{\|y-x\|_x^2}{1 - \|y-x\|_x}. \tag{7}$$

Here $\|v\|_{x*} = (v^T \nabla^2 g(x)^{-1} v)^{1/2}$ denotes the dual norm of $\|\cdot\|_x$.

– *Bounds on function value* [20, theorems 4.1.7 and 4.1.8]. If $x, y \in \mathbf{dom}\, g$, then

$$\omega(\|y-x\|_x) \leq g(y) - g(x) - \nabla g(x)^T (y-x) \leq \omega^*(\|y-x\|_x), \tag{8}$$

where ω and ω^* denote the functions

$$\omega(u) = u - \log(1+u), \qquad \omega^*(u) = -u - \log(1-u).$$

The left-hand inequality in (8) holds for all $x, y \in \mathbf{dom}\, g$. The right-hand inequality holds for all $x, y \in \mathbf{dom}\, g$ with $\|x - y\|_x < 1$. Note that ω and ω^* are Fenchel conjugates (Legendre transforms). In particular, we will use the fact that

$$\inf_v (\omega(v) - uv) = -\omega^*(u), \qquad \inf_u (\omega^*(u) - uv) = -\omega(v). \tag{9}$$

Figure 1 shows the two functions and illustrates the inequalities $\omega(u) \leq u^2/2 \leq \omega^*(u)$ and

$$\omega^*(u) \leq u^2/2 + u^3 \quad \text{for } u \in [0, 0.81], \quad \omega^*(u) \leq u^2 \quad \text{for } u \in [0, 0.68]. \tag{10}$$

A useful lower bound on $\omega(u)$ is

$$\omega(u) \geq \frac{u^2}{2(1+u)} \quad \text{for } u \geq 0. \tag{11}$$

– *Dikin ellipsoid theorem* [21, theorem 2.1.1.b]. The (open) Dikin ellipsoid at $x \in \mathbf{dom}\, g$ is defined as

$$\mathcal{E}_x = \{y \mid \|y - x\|_x < 1\}.$$

The upper bound in (8) implies that $\mathcal{E}_x \subset \mathbf{dom}\, g$.

2.2 Scaled proximal operator

The proximal operator of a closed convex function h is defined as

$$\text{prox}_h(y) = \arg\min_u \left(h(u) + \frac{1}{2}\|u - y\|^2 \right), \tag{12}$$

where $\|\cdot\|$ denotes the Euclidean norm. It can be shown that the proximal operator $\text{prox}_h(y)$ is uniquely defined for all y [18].

With every $x \in \mathbf{dom}\, g$ we can associate a *scaled proximal operator* $\text{prox}_{h,x}$, defined in a similar way as the standard proximal operator, but using the local quadratic norm $\|v\|_x = (v^T \nabla^2 g(x) v)^{1/2}$ instead of the Euclidean norm:

$$\text{prox}_{h,x}(y) = \arg\min_u \left(h(u) + \frac{1}{2}\|u - y\|_x^2 \right). \tag{13}$$

This scaled proximal operator can be expressed in terms of the standard (unscaled) proximal operator of the function $\tilde{h}(y) = h(\nabla^2 g(x)^{-1/2} y)$:

$$\text{prox}_{h,x}(y) = \nabla^2 g(x)^{1/2} \text{prox}_{\tilde{h}}(\nabla^2 g(x)^{1/2} y).$$

It can be shown (directly from the definition (13) or by reduction to the unscaled proximal operator) that $u = \text{prox}_{h,x}(y)$ exists and is unique for all $x \in \mathbf{dom}\, g$ and all y, and that it is the unique solution of the monotone inclusion problem

$$0 \in \partial h(u) + \nabla^2 g(x)(u - y). \tag{14}$$

As an immediate consequence we note that if $x^\star$ minimizes $f(x)$, *i.e.*, $0 \in \nabla g(x^\star) + \partial h(x^\star)$, then

$$x^\star = \text{prox}_{h,x}\left(x^\star - \nabla^2 g(x)^{-1} \nabla g(x^\star)\right) \tag{15}$$

for all $x \in \mathbf{dom}\, g$. Conversely, if $x^\star$ satisfies (15) for some $x \in \mathbf{dom}\, g$, then $x^\star$ minimizes f.

2.3 Proximal Newton step

The *proximal Newton step* $v(x)$ at x is defined as

$$\begin{aligned} v(x) &= \mathrm{prox}_{h,x}\left(x - \nabla^2 g(x)^{-1}\nabla g(x)\right) - x \\ &= \arg\min_v \left(g(x) + \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x+v) \right). \end{aligned}$$

From the second expression, or from the first expression and (14), we see that $v(x)$ is characterized by the condition

$$0 \in \nabla g(x) + \nabla^2 g(x) v(x) + \partial h(x + v(x)), \tag{16}$$

and that x is optimal if and only if $v(x) = 0$.

The magnitude $\|v(x)\|_x$ of the Newton step in the local norm $\|\cdot\|_x$ plays an important role in the analysis of Newton's method for minimizing self-concordant functions (*i.e.*, problem (1) with $h(x) = 0$) [20,21]. In [21] $\|v(x)\|_x$ is called the *Newton decrement* of f at x.

When $h(x)$ is nonzero, it is generally not possible to compute $v(x)$ very accurately, and it is important to allow for inexact proximal Newton steps. In the algorithms discussed in the next sections, the following criterion will be used for accepting a vector v as an inexact proximal Newton step at x: there exists an r such that

$$r \in \nabla g(x) + \nabla^2 g(x) v + \partial h(x + v), \qquad \|r\|_{x*} \leq (1-\theta)\|v\|_x, \tag{17}$$

where $\theta \in (0, 1]$ is an algorithm parameter. We can interpret $1 - \theta$ as a bound on the relative error in the conditions (16) that characterize the exact proximal Newton step. With $\theta = 1$, the condition requires $r = 0$ and therefore $v = v(x)$, the exact proximal Newton step.

The next theorem shows that if v satisfies (17) for some r, and $\|v\|_x$ is sufficiently small, then x is close to optimal for (1). The theorem is an extension of theorem 4.1.11 in [20], which characterizes the distance to the minimum of a self-concordant function in terms of the norm $\|v(x)\|_x$ of the Newton step when $\|v(x)\|_x < 1$.

Theorem 1 *Suppose $x \in \mathbf{dom}\, g$, $x + v \in \mathbf{dom}\, h$, and v and r satisfy (17) with $\theta \in (0, 1]$. If*

$$\|v\|_x < \frac{1}{2-\theta}. \tag{18}$$

then the following properties hold.

- *f is bounded below and*

$$\inf_y f(y) \geq f(x+v) + \theta\|v\|_x^2 - \omega^*(\|v\|_x) - \omega^*((2-\theta)\|v\|_x). \tag{19}$$

– *The sublevel set* $\mathcal{S}_x = \{y \mid f(y) \leq f(x+v)\}$ *is bounded:* $\mathcal{S}_x \subseteq \{y \mid \|y-x\|_x \leq \hat{\rho}\}$ *where* $\hat{\rho}$ *is the positive root of the nonlinear equation*

$$\omega(\rho) - \rho(2-\theta)\|v\|_x = \max\{0, \omega^*(\|v\|_x) - \theta\|v\|_x^2\} \tag{20}$$

if $\|v\|_x > 0$*, and* $\hat{\rho} = 0$ *if* $\|v\|_x = 0$.
– f *has a unique minimizer* $x^\star$ *and* $\|x - x^\star\|_x \leq \hat{\rho}$.

Proof We first note that, by the Dikin ellipsoid theorem, $x+v \in \mathbf{dom}\, g$, since $\|v\|_x < 1$. Therefore $x + v \in \mathbf{dom}\, f = \mathbf{dom}\, g \cap \mathbf{dom}\, h$, and the right-hand side of (19) and the sublevel set $\mathcal{S}_x$ are well defined.

To show (19) we consider an arbitrary $y \in \mathbf{dom}\, f$. We combine the lower bound on $g(y)$ from (8) and the upper bound on $g(x + v)$ from (8), to get

$$\begin{aligned} g(y) &\geq g(x) + \nabla g(x)^T(y - x) + \omega(\|y - x\|_x) \\ &\geq g(x + v) + \nabla g(x)^T(y - x - v) - \omega^*(\|v\|_x) + \omega(\|y - x\|_x). \end{aligned}$$

A lower bound on $h(y)$ follows from from the subgradient in (17):

$$h(y) \geq h(x + v) + (r - \nabla g(x) - \nabla^2 g(x)v)^T(y - x - v).$$

Adding the lower bounds on $g(y)$ and $h(y)$ gives a lower bound on $f(y)$:

$$\begin{aligned} &f(y) - f(x + v) \\ &\geq (r - \nabla^2 g(x)v)^T(y - x) - r^T v + \|v\|_x^2 - \omega^*(\|v\|_x) + \omega(\|y - x\|_x) \\ &\geq (r - \nabla^2 g(x)v)^T(y - x) - \|r\|_{x*}\|v\|_x + \|v\|_x^2 - \omega^*(\|v\|_x) + \omega(\|y - x\|_x) \\ &\geq (r - \nabla^2 g(x)v)^T(y - x) + \theta\|v\|_x^2 - \omega^*(\|v\|_x) + \omega(\|y - x\|_x). \end{aligned} \tag{21}$$

Next, we find a lower bound for the right-hand side of (21). We express y as $y = x + tw$ with $\|w\|_x = 1$ and $t \geq 0$ and write (21) as

$$f(x + tw) \geq f(x + v) + t(r - \nabla^2 g(x)v)^T w + \theta\|v\|_x^2 - \omega^*(\|v\|_x) + \omega(t).$$

We first consider the minimum of the right-hand side over w. Using the Cauchy-Schwarz inequality, the triangle inequality, and the condition (17) we get

$$\begin{aligned} f(x + tw) &\geq f(x + v) - t\|r - \nabla^2 g(x)v\|_{x*} + \theta\|v\|_x^2 - \omega^*(\|v\|_x) + \omega(t) \\ &\geq f(x + v) - t(\|r\|_{x*} + \|v\|_x) + \theta\|v\|_x^2 - \omega^*(\|v\|_x) + \omega(t) \\ &\geq f(x + v) - t(2 - \theta)\|v\|_x + \theta\|v\|_x^2 - \omega^*(\|v\|_x) + \omega(t). \end{aligned} \tag{22}$$

The lower bound (19) now follows if we use the conjugacy relation (9) to minimize the right-hand side of (22) over t.

To show the bound on the sublevel set, we note that (22) implies that $f(x + tw) > f(x + v)$ when

$$\omega(t) - t(2 - \theta)\|v\|_x > \omega^*(\|v\|_x) - \theta\|v\|_x^2.$$

When $v = 0$, this holds for any $t > 0$. For nonzero v, it holds if t is greater than the positive root of the nonlinear equation (20).

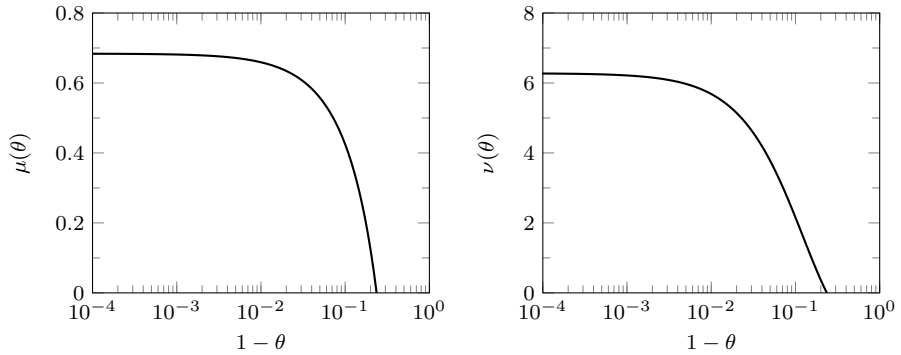

Fig. 2 *Left.* $\mu(\theta)$ is the solution u of the nonlinear equation $\omega^*((2-\theta)u) = \theta u^2$ for $3-\sqrt{5} \leq \theta \leq 1$. We have $\mu(1) = 0.68$ and $\mu(3-\sqrt{5}) = 0$. *Right.* The function $\nu(\theta)$ defined in (25). We have $\nu(1) = 6.28$ and $\nu(3-\sqrt{(5)}) = 0$.

Finally, since f is a closed function, it attains its minimum if the sublevel sets are bounded (by the Weierstrass theorem [6, page 119]). Since f is also strictly convex (the sum of a strictly convex function g and a convex function h), the minimizer is unique. □

The bounds on $f(x^\star)$ and $\|x - x^\star\|_x$ in theorem 1 can be simplified by restricting $\|v\|_x$ to a smaller interval than allowed by (18). We mentioned in section 2.1, that $\omega^*(u) \approx u^2/2$ for small u and $\omega^*(u) \leq u^2$ for $u \in [0, 0.68]$. More generally, for each $\theta \in (3-\sqrt{5}, 1] = (0.764, 1]$ there exists a positive $\mu(\theta)$ such that

$$\omega^*((2-\theta)u) \leq \theta u^2 \quad \text{for } u \in [0, \mu(\theta)] \tag{23}$$

(see figure 2). If $\theta \in (3-\sqrt{5}, 1]$, we can use the inequality (23) to simplify the lower bound (19) as follows: if $\|v\|_x \leq \mu(\theta)$, then

$$\begin{aligned} \inf_y f(y) &\geq f(x+v) + \theta\|v\|_x^2 - 2\omega^*((2-\theta)\|v\|_x) \\ &\geq f(x+v) - \theta\|v\|_x^2. \end{aligned} \tag{24}$$

Hence, for sufficiently small $\|v\|_x$, the quantity $\theta\|v\|_x^2$ gives an upper bound on $f(x+v) - \inf_y f(y)$.

We can also derive a simple upper bound on $\hat{\rho}$. For $0 < \|v\|_x \leq \mu(\theta)$ and $\theta \in (3-\sqrt{5}, 1]$, the right-hand side of (20) is zero because of (23), and $\hat{\rho}$ is the positive root of the equation

$$\log(1+\rho) = \rho(1 - (2-\theta)\|v\|_x).$$

In other words, $\hat{\rho} = \phi^{-1}(1-(2-\theta)\|v\|_x)$ where $\phi(t) = \log(1+t)/t$. Since ϕ^{-1} is a convex function and $\phi^{-1}(1) = 0$, Jensen's inequality gives

$$\hat{\rho} \leq \left(1 - \frac{\|v\|_x}{\mu(\theta)}\right)\phi^{-1}(1) + \frac{\|v\|_x}{\mu(\theta)}\phi^{-1}(1-(2-\theta)\mu(\theta)) = \frac{\nu(\theta)}{\mu(\theta)}\|v\|_x$$

where

$$\nu(\theta) = \phi^{-1}\left(1 - (2-\theta)\mu(\theta)\right). \tag{25}$$

This function is shown in figure 2. It follows that when $\|v(x)\|_x \leq \mu(\theta)$, the sublevel set $\mathcal{S}_x$ is bounded by a ball with radius $(\nu(\theta)/\mu(\theta))\|v(x)\|_x$ around x. In particular,

$$\|x - x^\star\|_x \leq \frac{\nu(\theta)}{\mu(\theta)}\,\|v\|_x. \tag{26}$$

For $\theta = 1$ and $v = v(x)$, the bounds (24) and (26) are

$$\inf_y f(y) \geq f(x + v(x)) - \|v(x)\|_x^2, \qquad \|x - x^\star\|_x \leq 9.18\,\|v(x)\|_x, \tag{27}$$

and these are valid if $\|v(x)\|_x \leq 0.68$. In the following section we will be interested in values of θ close to one, and it will be useful to note that $\mu(\theta) = 1/4$ for $\theta = 0.84$. In particular, if $\theta \geq 0.84$, then the bound (24) holds for $\|v\|_x \leq 1/4$.

3 Damped proximal Newton method

In this section we analyze the following version of the proximal Newton method with inexact proximal Newton steps.

Algorithm 3.1. *Proximal Newton algorithm with damped stepsize.*
Input: A starting point $x \in \mathbf{dom}\, g$ and three parameters $\theta_{\min} \in [0.9, 1]$, $\eta \in (0, 1/4]$, and $\delta \in (0, 1)$.
Repeat:

1. Compute a step v that satisfies (17) for some r and $\theta \geq \theta_{\min}$.
2. If $\|v\|_x \leq 0.25$ and $\theta\|v\|_x^2 \leq \delta$, return $x + v$.
3. Otherwise, set $x := x + \alpha v$ with

$$\alpha = \frac{\theta}{1 + \theta\|v\|_x} \quad \text{if } \|v\|_x \geq \eta, \qquad \alpha = 1 \quad \text{otherwise.}$$

The exit condition guarantees that $f(x+v) - \inf_y f(y) \leq \delta$. This follows from the fact that (24) holds if $\theta \geq 0.84$ and $\|v\|_x \leq 1/4$, as we saw at the end of the previous section. The lower bound $\theta_{\min} \geq 0.9$ is imposed only to simplify this stopping criterion. Alternatively, one can take any $\theta_{\min} \in (0, 1]$ and use (19) to bound $f(x + v) - \inf_y f(y)$.

Note that the starting point x is not required to be in $\mathbf{dom}\, h$. However, the Dikin ellipsoid theorem guarantees that $x \in \mathbf{dom}\, f$ after the first iteration.

3.1 Local convergence

The following theorem extends a quadratic convergence result for Newton's method applied to a self-concordant function [20, theorem 4.1.14]. A related result is [28, theorem 7] on the local convergence of the exact proximal Newton method with self-concordant g. For $\theta = 1$, theorem 2 gives an improvement over [28, theorem 7], which requires the condition $\|v(x)\|_x < 1 - 1/\sqrt{2}$; see also [28, remark 10]. Theorem 2 further generalizes these results by allowing inexact proximal Newton steps.

Theorem 2 (Unit steps) *Suppose $x \in \mathbf{dom}\, g$, $x + v \in \mathbf{dom}\, h$, $\|v\|_x < 1$, and (17) is satisfied for some r and $\theta \in (0, 1]$. Define $x^+ = x + v$. Suppose $x^+ + v^+ \in \mathbf{dom}\, h$ and*

$$r^+ \in \nabla g(x^+) + \nabla^2 g(x^+)v^+ + \partial h(x^+ + v^+), \qquad \|r^+\|_{x^+ *} \leq (1 - \theta^+)\|v^+\|_{x^+}$$

holds for some r^+ and $\theta^+ \in (0, 1]$. Then

$$\|v^+\|_{x^+} \leq \frac{\|v\|_x}{\theta^+(1 - \|v\|_x)} \left(1 - \theta + \frac{\|v\|_x}{1 - \|v\|_x}\right).$$

If $\|v\|_x \leq 1 - 1/\sqrt{2} = 0.293$, we have the simpler bound

$$\|v^+\|_{x^+} \leq \frac{\sqrt{2}\|v\|_x}{\theta^+} \left(1 - \theta + \sqrt{2}\|v\|_x\right). \tag{28}$$

Proof We first note that $x^+ = x + v \in \mathbf{dom}\, g$ as a consequence of the Dikin ellipsoid theorem. Define

$$w = r - \nabla g(x) - \nabla^2 g(x)v, \qquad w^+ = r^+ - \nabla g(x^+) - \nabla^2 g(x^+)v^+.$$

We have $w \in \partial h(x + v)$ and $w^+ \in \partial h(x^+ + v^+)$, by definition of r and r^+. Monotonicity of the subdifferential ∂h implies that

$$(w^+ - w)^T v^+ = (w^+ - w)^T (x^+ + v^+ - x - v) \geq 0.$$

This observation is used in the first inequality of the following derivation:

$$\begin{aligned}
\|v^+\|_{x^+} &\leq \|v^+ + \nabla^2 g(x^+)^{-1}(w^+ - w)\|_{x^+} \\
&= \|\nabla^2 g(x^+)^{-1}(r^+ - \nabla g(x^+) - w)\|_{x^+} \\
&= \|r^+ - \nabla g(x^+) - w\|_{x^+ *} \\
&\leq \|r^+\|_{x^+ *} + \|\nabla g(x^+) + w\|_{x^+ *} \\
&= \|r^+\|_{x^+ *} + \|r + \nabla g(x^+) - \nabla g(x) - \nabla^2 g(x)v\|_{x^+ *} \\
&\leq (1 - \theta^+)\|v^+\|_{x^+} + \|r\|_{x^+ *} + \|\nabla g(x^+) - \nabla g(x) - \nabla^2 g(x)v\|_{x^+ *} \\
\theta^+\|v^+\|_{x^+} &\leq \frac{1}{1 - \|v\|_x} \left(\|r\|_{x*} + \|\nabla g(x + v) - \nabla g(x) - \nabla^2 g(x)v\|_{x*}\right) \\
&\leq \frac{\|v\|_x}{1 - \|v\|_x} \left(1 - \theta + \frac{\|v\|_x}{1 - \|v\|_x}\right).
\end{aligned}$$

On the second line we use the definition of w^+, and on the fifth line the definition of w. Line 7 follows from (6), which implies that

$$\|z\|_{x+v,*}^2 = z^T\nabla^2 g(x+v)^{-1}z \leq \frac{1}{(1-\|v\|_x)^2} z^T\nabla^2 g(x)^{-1}z = \frac{\|z\|_{x*}^2}{(1-\|v\|_x)^2}.$$

The last step follows from (7). □

Theorem 2 can be used to establish local convergence of algorithm 3.1.

Exact proximal Newton method. Suppose the starting point x satisfies $\|v\|_x < \eta$ and we take $\theta_{\min} = 1$, so $v = v(x)$. The inequality (28) reduces to

$$\|v(x^+)\|_{x^+} \leq 2\|v(x)\|_x^2 \tag{29}$$

and, since $\eta \leq 1/4$, we have $\|v(x^+)\|_{x^+} < \eta$. All subsequent iterates therefore satisfy $\|v(x)\|_x < \eta$. It then follows from (29) that after k iterations

$$2\|v(x)\|_x \leq (2\eta)^{2^k} \leq \left(\frac{1}{2}\right)^{2^k}.$$

This shows that algorithm 3.1 converges quadratically when started at a point with $\|v(x)\|_x < \eta$. Since $\|v(x)\|_x^2 \leq (1/2)^{2^{k+1}}$, the exit condition $\|v\|_x^2 \leq \delta$ is satisfied after less than $\log_2\log_2(1/\delta)$ iterations.

Inexact proximal Newton method. Suppose the starting point x satisfies $\|v\|_x < \eta$ and we take θ constant. From (28),

$$\begin{aligned}\|v^+\|_{x^+} &\leq \sqrt{2}\left(\frac{1+\sqrt{2}\eta}{\theta} - 1\right)\|v\|_x \\ &\leq \sqrt{2}\left(\frac{1+\sqrt{2}/4}{0.9} - 1\right)\|v\|_x \\ &= 0.713\,\|v\|_x.\end{aligned}$$

Therefore $\|v\|_x$ converges to zero linearly. If we let $\theta \to 1$, then the inequality (28) shows superlinear convergence.

3.2 Global convergence

The next theorem is an extension of a global convergence result for the standard damped Newton method for self-concordant functions [20, theorem 4.1.12]. When $\theta = 1$, the result is identical to [28, theorem 6].

Theorem 3 (Damped steps) *Suppose $x \in \mathbf{dom}\, f$, $x+v \in \mathbf{dom}\, h$, and (17) is satisfied for some r and $\theta \in (0,1]$. If $\alpha = \theta/(1+\theta\|v\|_x)$, then*

$$f(x+\alpha v) \leq f(x) - \omega(\theta\|v\|_x).$$

Proof First note that $\alpha\|v\|_x < 1$. Hence $x + \alpha v \in \mathbf{dom}\, f$ as a consequence of the Dikin ellipsoid theorem. To show the upper bound on $f(x + \alpha v)$ we apply the upper bound (8) with $y = x + \alpha v$:

$$g(x + \alpha v) \leq g(x) + \alpha \nabla g(x)^T v + \omega^*(\alpha\|v\|_x).$$

An upper bound on $h(x + \alpha v)$ follows from Jensen's inequality and the subgradient of h at $x + v$ from (17):

$$\begin{aligned} h(x + \alpha v) &\leq h(x) + \alpha(h(x + v) - h(x)) \\ &\leq h(x) + \alpha(r - \nabla g(x) - \nabla^2 g(x)v)^T v \\ &= h(x) + \alpha(r - \nabla g(x))^T v - \alpha\|v\|_x^2. \end{aligned}$$

Adding the upper bounds on g and h gives

$$\begin{aligned} f(x + \alpha v) &\leq f(x) + \alpha(r^T v - \|v\|_x^2) + \omega^*(\alpha\|v\|_x) \\ &\leq f(x) + \alpha(\|r\|_{x*}\|v\|_x - \|v\|_x^2) + \omega^*(\alpha\|v\|_x) \\ &\leq f(x) - \alpha\theta\|v\|_x^2 + \omega^*(\alpha\|v\|_x). \end{aligned} \tag{30}$$

This bound holds when $\alpha\|v\|_x < 1$. The right-hand side is minimized at $\alpha = \theta/(1 + \theta\|v\|_x)$, with minimum value $f(x) - \omega(\theta\|v\|_x)$. □

Theorem 3 implies that if $\|v\|_x \geq \eta$ in algorithm 3.1, then

$$f(x + \alpha v) \leq f(x) - \omega(\theta\eta),$$

so the cost function is decreased by at least a positive amount $\omega(\theta\eta)$. If the function is bounded below, we must reach $\|v\|_x < \eta$ after a finite number of iterations. Hence algorithm 3.1 converges from any starting point if the problem is bounded below.

4 Proximal Newton method with backtracking line search

Although backtracking line searches are typically used in smooth optimization algorithms, the proximal Newton algorithm is readily modified to include a backtracking line search of the type used with the Newton algorithm in [7, chapter 9]; see [17]. We will analyze the following algorithm and use it in the experiments of section 5.

Algorithm 4.1. *Proximal Newton algorithm with line search.*
Input: A starting point $x \in \mathbf{dom}\, f$, and parameters $\theta_{\min} \in (0, 1]$, $\beta \in (0, 1)$, and $\gamma \in (0, \theta_{\min}/2)$.
Repeat:

1. Compute a step v that satisfies (17) for some r and $\theta \geq \theta_{\min}$.
2. If $\|v\|_x$ is sufficiently small, return $x + v$.

3. Otherwise, set $x := x + \alpha v(x)$ where α is the largest number in $\{1, \beta, \beta^2, \beta^3, \ldots\}$ for which

$$x + \alpha v \in \mathbf{dom}\, f, \qquad f(x + \alpha v) \leq f(x) - \alpha\gamma\theta\|v\|_x^2. \quad (31)$$

To formulate a rigorous stopping condition that guarantees a bound on $f(x+v) - \inf_y f(y)$ one can use the inequality (19) in theorem 1, which is valid for any $\theta \in (0, 1)$, or the simpler inequality (24), which assumes $\theta > 0.764$.

We refer to the condition (31) as the *condition of sufficient decrease*. Note that the starting point of algorithm 4.1 is required to be in $\mathbf{dom}\, f$, so the right-hand side in the condition of sufficient decrease is well defined in the first iteration. Alternatively, one can start at $x \in \mathbf{dom}\, g$ and use a damped Newton step in the first iteration.

The following observation extends a result for the standard Newton method with backtracking line search applied to self-concordant functions [7, section 9.6.4].

Theorem 4 *The stepsize selected by the backtracking line search satisfies*

$$\frac{\beta\theta}{1 + \theta\|v\|_x} < \alpha \leq 1.$$

A unit stepsize is selected if $\|v\|_x \leq \theta(1 - \gamma) - 1/2$.

Proof We first note that the step size $\hat{\alpha} = \theta/(1+\theta\|v\|_x)$ satisfies the condition of sufficient decrease. This can be seen from the upper bound (30):

$$\begin{aligned} f(x + \hat{\alpha}v) &\leq f(x) - \hat{\alpha}\theta\|v\|_x^2 + \omega^*(\hat{\alpha}\|v\|_x) \\ &= f(x) - \omega(\theta\|v\|_x) \\ &\leq f(x) - \frac{\theta^2\|v\|_x^2}{2(1 + \theta\|v\|_x)} \\ &= f(x) - \hat{\alpha}\theta\|v\|_x^2/2 \\ &\leq f(x) - \hat{\alpha}\gamma\|v\|_x^2. \end{aligned}$$

Line 3 follows from the inequality (11). The last step follows because $\gamma \leq \theta/2$. Since $\hat{\alpha}$ satisfies the condition of sufficient decrease, the stepsize α selected by the line search can not be less than or equal to

$$\beta\hat{\alpha} = \frac{\beta\theta}{1 + \theta\|v\|_x}.$$

For the second part of the theorem, note that if $\|v\|_x \leq \theta(1 - \gamma) - 1/2$ then, again using (30),

$$\begin{aligned} f(x + v) &\leq f(x) - \theta\|v\|_x^2 + \omega^*(\|v\|_x) \\ &\leq f(x) - \theta\|v\|_x^2 + \frac{1}{2}\|v\|_x^2 + \|v\|_x^3 \\ &= f(x) - (\theta - 1/2 - \|v\|_x)\|v\|_x^2 \\ &\leq f(x) - \gamma\theta\|v\|_x^2. \end{aligned}$$

Line 2 follows from the first inequality in (10). □

Theorem 4 can be combined with the analysis of section 3 to show that algorithm 4.1 has the same convergence properties as algorithm 3.1. Choose any positive η. If $\|v\|_x > \eta$, the condition of sufficient decrease and the lower bound on the stepsize from theorem 4 guarantees

$$\begin{aligned} f(x + \alpha v) &\leq f(x) - \alpha\gamma\theta\|v\|_x^2 \\ &\leq f(x) - \beta\gamma\frac{\theta^2\|v\|_x^2}{1 + \theta\|v\|_x} \\ &\leq f(x) - \beta\gamma\frac{\theta_{\min}^2\eta^2}{1 + \theta_{\min}\eta}. \end{aligned}$$

(The last step follows from monotonicity of the function $t^2/(1 + t)$.) If the problem is bounded below, the algorithm reaches a stopping condition $\|v\|_x \leq \eta$, for any positive η, after a finite number of iterations.

Moreover, if we choose $\theta_{\min} > 1/2$ and $\gamma < 1 - 1/(2\theta_{\min})$ then theorem 4 guarantees that for sufficiently small $\|v\|_x$, a unit stepsize is chosen and the local convergence results of section 3.1 apply.

5 Restricted sparse inverse covariance selection

In this section we illustrate the convergence properties of the inexact proximal Newton method with an application to the covariance selection problem from statistics.

5.1 Covariance selection

The *covariance selection* problem was introduced by Dempster in 1972 [11]. Dempster considered the problem of computing the maximum likelihood estimate of the covariance matrix Σ of a normal distribution $N(0, \Sigma)$, subject to a constraint on the sparsity pattern of Σ^{-1}. Zeros in the inverse covariance Σ^{-1} indicate pairs of conditionally independent components of the random variable. If we assume the random vector has dimension p, then the maximum likelihood estimation problem can be shown to be equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(C\Sigma^{-1}) + \log\det\Sigma \\ \text{subject to} & (\Sigma^{-1})_{ij} = 0 \quad \text{for } (i,j) \in \bar{E}, \end{array} \tag{32}$$

where C is the sample covariance matrix, and the sets

$$\begin{aligned} E &\subseteq \{(i,j) \mid i,j \in \{1,2,\ldots,p\},\, i > j\}, \\ \bar{E} &= \{(i,j) \mid i,j \in \{1,2,\ldots,p\},\, i > j\} \setminus E \end{aligned}$$

are a subset of the off-diagonal index pairs and its complement. We refer to the set E, which contains the positions of the possibly nonzero entries in Σ^{-1}, as

the *sparsity pattern* of Σ^{-1}. Throughout this section, we assume that $\log\det X$ is only defined for positive definite X, *i.e.*, the problem (32) also includes an implicit constraint that the variable Σ is positive definite. Dempster observed that the problem is convex if $X = \Sigma^{-1}$ is used as optimization variable. After this change of variables, the covariance selection problem can be written as a convex optimization problem

$$\text{minimize} \quad \mathbf{tr}(CX) - \log\det X + \psi(X), \tag{33}$$

with variable $X \in \mathbf{S}^p$ (the set of symmetric $p \times p$ matrices), where ψ is the 'indicator' function of the sparsity pattern:

$$\psi(X) = \sum_{(i,j)\in\bar{E}} \delta(X_{ij}), \qquad \delta(u) = \begin{cases} 0 & u = 0 \\ \infty & u \neq 0. \end{cases} \tag{34}$$

A popular approach for estimating a sparse inverse covariance matrix $X = \Sigma^{-1}$ when the sparsity pattern is not known, is to add an ℓ_1-norm penalty to the log-likelihood objective, *i.e.*, solve (33) with

$$\psi(X) = \lambda \sum_{i>j} |X_{ij}|. \tag{35}$$

The solution is sometimes referred as the *graphical lasso* solution, and several specialized algorithms have been developed for computing it; see the surveys in [13, chapter 9] and [25].

An interesting combination of the functions (34) and (35) is

$$\psi(X) = \sum_{(i,j)\in\bar{E}} \delta(X_{ij}) + \lambda \sum_{(i,j)\in E} |X_{ij}|. \tag{36}$$

With this choice of ψ, the off-diagonal entries of X indexed by $\bar{E}$ are constrained to be zero; the remaining entries are penalized by an ℓ_1-norm penalty. This formulation is useful when the sparsity pattern of Σ^{-1} is partially known. The constraints on the entries in $\bar{E}$ then represent the prior information about the sparsity pattern. For example, if the random variable contains consecutive values of a vector autoregressive process with lag r, then the inverse covariance matrix is block-banded with half-bandwidth r. Incorporating prior information of this kind reduces the number of parameters to be estimated in the maximum-likelihood problem, and hence the number of samples needed for a good estimate. We will refer to problem (33) with the penalty function (36) as a *restricted* sparse inverse covariance selection.

The proximal Newton method is an attractive algorithm for the restricted covariance selection problem because the key computations in the algorithm can be implemented using efficient sparse matrix techniques. The starting point is to reformulate the problem as follows. We first compute a *triangulation* or *chordal extension* E' of the sparsity pattern E, *i.e.*, a sparsity pattern E' that contains E and is also *chordal* [29]. Instead of optimizing over $X \in \mathbf{S}^p$, as in (33), we can then restrict X, without loss of generality, to $\mathbf{S}^p_{E'}$, the space

of symmetric $p \times p$ matrices with sparsity pattern E'. Thus the problem can be written equivalently as

$$\text{minimize} \quad \phi(X) + \psi(X) \tag{37}$$

with a *sparse* matrix variable $X \in \mathbf{S}^p_{E'}$, and functions $\phi, \psi : \mathbf{S}^p_{E'} \to \mathbf{R}$ defined as

$$\phi(X) = \mathbf{tr}(CX) - \log\det X, \qquad \psi(X) = \sum_{(i,j)\in E'\setminus E} \delta(X_{ij}) + \gamma \sum_{(i,j)\in E} |X_{ij}|.$$

As mentioned we define $\mathbf{dom}\,\phi = \{X \in \mathbf{S}^p_{E'} \mid X \succ 0\}$. Problem (37) is a composite convex optimization problem that can be expressed as (1) if we represent the matrices X as vectors x of length $n = |E'| + p$. The second term ψ is separable and its proximal operator reduces to simple component-wise operations (soft-thresholding for entries in positions $(i,j) \in E$; substituting zero for entries in positions $(i,j) \in E' \setminus E$). The first term ϕ is self-concordant [21]. Moreover the chordal structure of E' allows us to apply specialized algorithms for computing ϕ and its derivatives. To evaluate ϕ at a given $X \succ 0$, we compute a sparse Cholesky factorization

$$X = P^T L L^T P,$$

where P is a permutation matrix and L is lower triangular. Adding the logarithms of the diagonal elements of L gives $\phi(X) = -2\sum_i \log L_{ii}$. Given the Cholesky factorization, the gradient and Hessian are also readily computed by algorithms that are closely related to the multifrontal algorithm for sparse Cholesky factorization and use similar recursions on an elimination tree or supernodal elimination tree [2, 29]. The gradient of ϕ, as a function from $\mathbf{S}^p_{E'}$ to $\mathbf{R}$, is given by

$$\nabla\phi(X) = \Pi_{E'}(C - X^{-1}),$$

where $\Pi_{E'}$ denotes projection on $\mathbf{S}^p_{E'}$. Computing the gradient therefore requires computing the entries of X^{-1} on the diagonal and in positions $(i,j) \in E'$, but not any of the other entries. For a chordal pattern, this projected inverse can be computed by a recursion on the elimination tree. The Hessian $\mathcal{H}_X$ of ϕ at $X \in \mathbf{dom}\,\phi$ is a linear mapping from $\mathbf{S}^p_{E'}$ to $\mathbf{S}^p_{E'}$ defined by

$$\mathcal{H}_X(V) = \nabla^2\phi(X)[V] = \frac{d}{d\alpha}\nabla\phi(X + \alpha V)\Big|_{\alpha=0} = \Pi_{E'}(X^{-1}VX^{-1}V).$$

For a chordal pattern E', the evaluations of $\mathcal{H}_X(V)$ or $\mathcal{H}_X^{-1}(V)$ that are required by the proximal Newton method, can be computed by two recursions on the elimination tree. The complexity of each of these operations is roughly the same as the cost of a sparse Cholesky factorization with sparsity pattern E'. We refer the interested reader to [29] for details and historical background on these techniques.

5.2 Subproblem

In the experiments described in the next section a basic version of the FISTA algorithm [4] was used to minimize the function (3) in the subproblems. At iteration k of FISTA a new estimate v^k of the solution of the subproblem is computed, by making a proximal gradient update

$$v^k = \text{prox}_{th}\left(x + w - t(\nabla g(x) + \nabla^2 g(x)w)\right) - x$$

where w is the previous value v^{k-1} plus an an extrapolation term,

$$w = v^{k-1} + \frac{k-2}{k+1}\left(v^{k-1} - v^{k-2}\right).$$

From the definition of the proximal operator prox_{th}, the following relation between these variables holds:

$$\frac{1}{t}(w - v^k) \in \nabla g(x) + \nabla^2 g(x)w + \partial h(x + v^k).$$

This shows that the vector

$$r = \frac{1}{t}(w - v^k) + \nabla^2 g(x)(v^k - w) = (\frac{1}{t}I - \nabla^2 g(x))(w - v^k)$$

satisfies $r \in \nabla g(x) + \nabla^2 g(x)v^k + \partial h(x + v^k)$. In our implementation, r was used in the condition $\|r\|_{x*} \leq (1-\theta)\|v^k\|_x$ to determine whether to accept v^k as an inexact proximal Newton step v.

To select the FISTA stepsize t, we used the simple backtracking strategy suggested in [4]. More sophisticated variants of FISTA, such as N83 in the TFOCS package [5], or methods that use different strategies for selecting t [24], are likely to lead to substantial improvements over our results. We also note that several first-order methods could be used as alternatives to FISTA, including the coordinate descent method [15] and the orthant-based method [8].

5.3 Experiments

In this section we present some results for the proximal Newton method applied to (37). We use the Python packages CHOMPACK [3] and CVXOPT [1] for the sparse matrix computations (evaluation of ϕ and its gradient, Hessian, and inverse Hessian). The main purpose of the experiments is to compare the convergence properties with the theoretical results in sections 3–4. Our implementation is not optimized, because it requires several conversions between different sparse matrix formats. Moreover the proximal Newton algorithm itself, and some key functions of CHOMPACK (such as the symbolic factorization), are implemented in Python and would be faster when implemented directly in C. This must be kept in mind when comparing the computation times for different parameter values in the experiments.

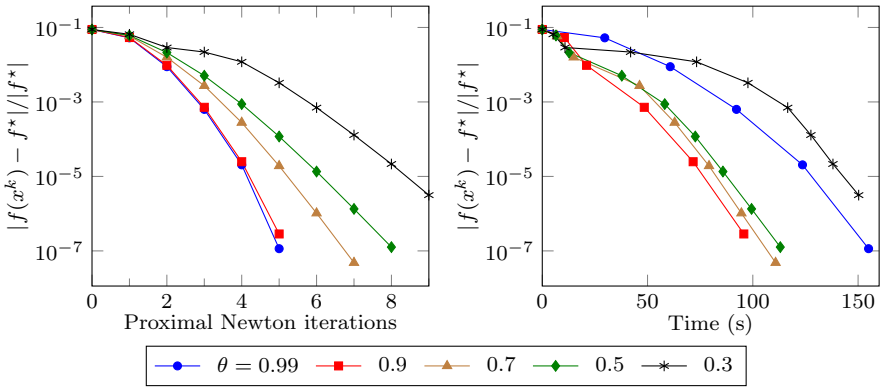

Fig. 3 Convergence of the proximal Newton method in the first experiment, for different values of θ.

Band patterns. In the first experiment we use a band pattern E of size $p = 1000$ with half-bandwidth 20. Band patterns are chordal, so $E' = E$ in this experiment. To generate a sample covariance matrix C we first create a sparse matrix Σ^{-1} as follows. We randomly select 80% of the entries within the band E, and set them to zero. For the remaining entries in E, we randomly generate values following a normal distribution $N(0, 1)$. A multiple of the identity is added to the matrix Σ^{-1} if it is not positive definite. We then generate $N = 10p$ samples from the distribution $N(0, \Sigma)$ and form the sample covariance matrix C. The regularization parameter in (37) was set to $\lambda = 0.02$.

Figure 3 shows the convergence of algorithm 4.1 with different, constant values of the parameter θ, and backtracking parameters $\gamma = 0.01$, $\beta = 1/2$. The first figure confirms the conclusions about the effect of θ in the theoretical analysis of section 4. It also shows that the proximal Newton method can reach a high accuracy, even with very inaccurate solutions of the subproblems (low values of θ). The second figure shows the convergence versus elapsed time (on a machine with a 2.5GHz Intel Core i7 processor). The plots suggest there is a value of θ that gives the fastest convergence. Although the best value of θ and the overall solution times are likely to be quite different in a more optimized implementation of the algorithm, the figure shows the benefits that can be expected from improvements in the algorithm for the subproblem, and from strategies for adapting θ during the algorithm, as suggested in [17].

Sparsity patterns from University of Florida collection. In the second experiment we use three patterns from the UF collection [9]. Table 1 gives the dimension and the number of nonzeros $2|E| + p$ for each pattern, and the number of nonzeros in a chordal extension (the second and third patterns are chordal, so $E = E'$). We generate a sample covariance matrix as in the first experiment. We first generate a sparse matrix $\Sigma^{-1} \in \mathbf{S}^p_E$. A randomly selected subset of 30% of the entries in E are set to zero. The values of the remaining entries in E are chosen from $N(0, 1)$. A multiple of the identity is added to

Name	p	nnz	nnz after extension
1138_bus	1138	4054	5392
Chem97ZtZ	2541	7361	7361
mhd4800b	4800	27520	27520

Table 1 Three sparsity patterns from the University of Florida collection.

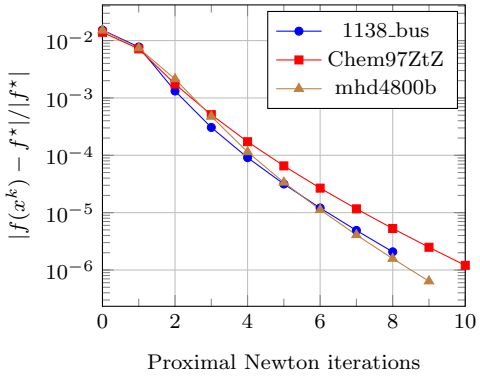

Fig. 4 Convergence of the proximal Newton method for the three test problems in the second experiment.

make the matrix positive definite. We then use Σ to generate $N = 10p$ samples and form the sample covariance C.

Figure 4 shows the convergence of algorithm 4.1 for the three problems. We use $\theta = 0.5$, $\gamma = 0.01$, and $\beta = 1/2$. Even though the dimensions of the three problems are quite different, the method converges in roughly the same, small number of iterations, as is typical for the standard Newton method.

6 Conclusion

We presented an analysis of the proximal Newton method for minimizing a sum of a self-concordant function and a function with an inexpensive proximal mapping. The analysis extends results from [28] by taking into account inexactness of the computation of the proximal Newton steps, and differs from [16,27] in the conditions used to describe inexactness of the Newton steps. The conclusions are similar to the results reached in [8,17] under different assumptions on the smooth component of the cost function.

The analysis presented in this paper is motivated by applications to the sparse covariance selection problem from statistics, in which we impose prior constraints on the sparsity pattern of the inverse covariance matrix. The log-det term in the cost function of this problem is self-concordant, and efficient methods exist for evaluating the matrix-vector products with its Hessian and inverse Hessian needed in the proximal Newton method.

Preliminary numerical results indicate that the method can reach a high accuracy, even with inexact computation of the proximal Newton steps. Important questions for further research include the choice of algorithm for solving the subproblems, and the formulation of good strategies for adaptive control of the accuracy with which the subproblems are solved. As was pointed out by a reviewer of this paper, path-following methods offer an alternative for minimizing self-concordant functions and have a lower computational complexity than the damped Newton method. It would be of great interest to formulate path-following methods for the composite problem (1), for example, by extending the algorithm of [20, page 205].

References

1. M. Andersen, J. Dahl, and L. Vandenberghe. *CVXOPT: A Python Package for Convex Optimization*. `www.cvxopt.org`, 2015.
2. M. S. Andersen, J. Dahl, and L. Vandenberghe. Logarithmic barriers for sparse matrix cones. *Optimization Methods and Software*, 28(3):396–423, 2013.
3. M. S. Andersen and L. Vandenberghe. *CHOMPACK: A Python Package for Chordal Matrix Computations, Version 2.2.1*, 2015. `cvxopt.github.io/chompack`.
4. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
5. S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.
6. D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.
7. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
8. R. H. Byrd, J. Nocedal, and F. Oztoprak. An inexact successive quadratic approximation method for L-1 regularized optimization. *Mathematical Programming*, pages 1–22, 2015.
9. T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software*, 38:1–25, 2011.
10. R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact newton methods. *SIAM J. on Numerical Analysis*, 19(2):400–408, April 1982.
11. A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.
12. S. C. Eisenstat and H. F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996.
13. T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity. The Lasso and Generalizations*. CRC Press, 2015.
14. J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1993.
15. 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 (NIPS)*, volume 24, pages 2330–2338, 2011.
16. A. Kyrillidis, R. Karimi-Mahabadi, Q. Tran-Dinh, and V. Cevher. Scalable sparse covariance estimation via self-concordance. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1946–1952, 2014.
17. J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.
18. J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Math. Soc. France*, 93:273–299, 1965.
19. Y. Nesterov. Towards non-symmetric conic optimization. *Optimization Methods and Software*, 27(4-5):893–917, 2012.

20. Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
21. Yu. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Methods in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.
22. 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 (NIPS)*, volume 25, pages 764–772, 2012.
23. J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. SIAM, 2001.
24. K. Scheinberg, D. Goldfarb, and X. Bai. Fast first-order methods for composite convex optimization with backtracking. *Foundations of Computational Mathematics*, 14:389–417, 2014.
25. K. Scheinberg and S. Ma. Optimization methods for sparse inverse covariance selection. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 455–477. MIT Press, 2012.
26. K. Scheinberg and X. Tang. Complexity of inexact proximal Newton methods. Technical Report 13T-02-R1, COR@L, Lehigh University, 2013.
27. Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. An inexact proximal path-following algorithm for constrained convex optimization. *SIAM Journal on Optimization*, 24(4):1718–1745, 2014.
28. Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. Composite self-concordant minimization. *Journal of Machine Learning Research*, 16:371–416, 2015.
29. L. Vandenberghe and M. S. Andersen. Chordal graphs and semidefinite optimization. *Foundations and Trends in Optimization*, 1(4):241–433, 2014.