Inexact proximal Newton methods and self-concordant functions

Lieven Vandenberghe
Department of Electrical Engineering
University of California, Los Angeles

Joint work with Jinchao Li and Martin S. Andersen

Workshop on Convex and Real-Time Optimization
Aalborg University
August 22–23, 2016

Proximal Newton method

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

- g convex, twice continuously differentiable
- h convex, possibly nondifferentiable, but 'simple' (for example, $h(x) = \|x\|_1$)

Proximal Newton step: step at x is defined as

$$v(x) = \operatorname*{argmin}_v \left(g(x) + \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x+v) \right)$$

- if h is zero, this is the standard Newton step $v(x) = -\nabla^2 g(x)^{-1} \nabla g(x)$
- also known as *successive quadratic approximation* step

(Lee, Sun, Saunders 2014, Byrd, Nocedal, Oztoprak 2015, Tran-Dinh, Kyrillidis, Cevher 2015 ...)

Inexact proximal Newton method

proximal Newton step $v(x)$ at x is the solution v of

$$\text{minimize} \quad \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x+v)$$

- a LASSO problem when $h(x) = \|x\|_1$
- in practice, solved *inexactly* by first order algorithm, *e.g.*, proximal gradient

$$v^+ = \mathrm{prox}_{th}\big(x + v - t(\nabla g(x) + \nabla^2 g(x) v)\big) - x$$

- analysis of proximal Newton method must account for error in $v(x)$

Forcing term in inexact Newton method

$$\text{minimize} \quad g(x)$$

v is accepted as approximate Newton step at iteration k if

$$\|\nabla g(x^k) + \nabla^2 g(x^k) v\| \leq \eta_k \|\nabla g(x^k)\|$$

- left-hand side is residual in Newton equation
- $\eta_k \in [0, 1)$ is the *forcing term,* constant or selected adaptively
- η_k controls speed of local convergence

$\eta_k = 0$	$\eta_k \searrow 0$	$\eta_k \geq \eta_{\min} > 0$
quadratic	superlinear	linear

(Dembo, Eisenstat, Steihaug 1982; Eisenstat & Walker 1994, 1996)

Inexact proximal Newton method (classical assumptions)

$$\text{minimize} \quad g(x) + h(x)$$

- g strongly convex with Lipschitz continuous gradient ($mI \preceq \nabla^2 g(x) \preceq LI$)
- Hessian of g is Lipschitz continuous

Forcing condition (Lee, Sun, Saunders 2014, Byrd, Nocedal, Oztoprak 2015)

$$\|\hat{F}_t^k(x^k + v)\| \leq \eta_k \|F_t(x^k)\|$$

- $F_t(x)$ and $\hat{F}_t^k(y)$ are *gradient maps* for $g + h$ and its 2nd order approximation
- if h is zero, gradient maps reduce to

$$\hat{F}_t^k(x^k + v) = \nabla g(x^k) + \nabla^2 g(x^k) v, \qquad F_t(x^k) = \nabla g(x^k)$$

- with this forcing condition, convergence is similar to inexact Newton method

Inexact proximal Newton method (self-concordant assumption)

$$\text{minimize} \quad g(x) + h(x)$$

g is a self-concordant function

Convergence results (Tran-Dinh, Kyrillidis, Cevher 2014, 2015, 2016)

- results for exact steps are similar to Newton method for self-concordant g
- results for inexact steps use the notion of δ-*solution* $\hat{v}$:

$$\tilde{f}(x + \hat{v}) - \inf_v \tilde{f}(x + v) \leq \frac{\delta^2}{2}$$

where

$$\tilde{f}(x + v) = g(x) + \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x + v)$$

Outline of the talk

$$\text{minimize} \quad g(x) + h(x), \qquad g \text{ self-concordant}$$

Analysis of proximal Newton method

- forcing condition

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

- theorems are direct extensions of results for exact (standard) Newton method

Application: restricted covariance selection

$$\text{minimize} \quad \underbrace{\mathbf{tr}(CX) - \log\det X}_{g(X)} + \underbrace{\gamma\|X\|_1}_{h(X)}$$

X is a symmetric matrix with given sparsity pattern

Self-concordant function

a function $g : \mathbf{R}^n \to \mathbf{R}$ is **self-concordant** if

- g is closed, convex, with open domain
- g is three times continuously differentiable and $\nabla^2 g(x) \succ 0$ on $\mathbf{dom}\, g$
- the Hessian satisfies the inequality

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

  for all $x \in \mathbf{dom}\, g$ and all $v \in \mathbf{R}^n$, where

$$\|v\|_x = \left(v^T \nabla^2 g(x) v\right)^{1/2}$$

(Nesterov & Nemirovski 1994, Renegar 2001, Nesterov 2004)

Examples: logarithmic barrier functions for interior-point methods

Bounds on Hessian and gradient

Bounds on Hessian: for $\|v\|_x < 1$

$$(1 - \|v\|_x)^2 \nabla^2 g(x) \preceq \nabla^2 g(x+v) \preceq \frac{1}{(1-\|v\|_x)^2} \nabla^2 g(x)$$

Bound on gradient: for $\|v\|_x < 1$

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

here $\|w\|_{x*} = \left(w^T \nabla^2 g(x)^{-1} w\right)^{1/2}$ is the dual norm of $\|w\|_x$

(Nesterov 2012)

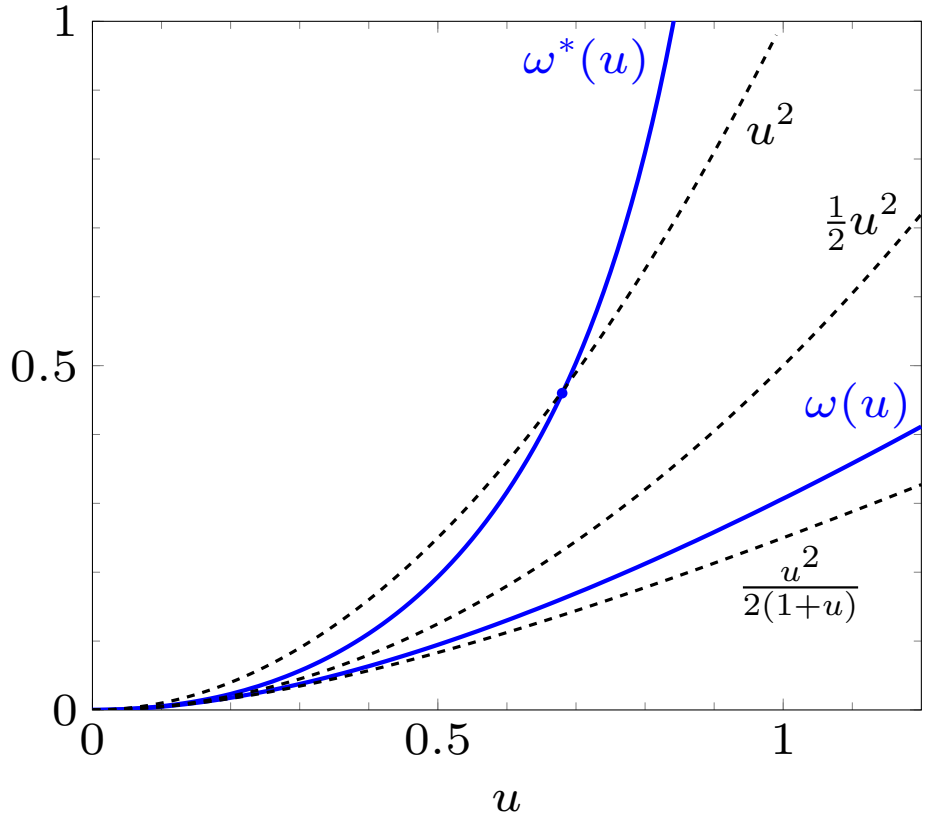
Bounds on function value

if $\|v\|_x < 1$, then

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

where

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



Dikin ellipsoid

the (open) *Dikin ellipsoid* centered at $x \in \mathbf{dom}\, g$ is defined

$$\begin{aligned} \mathcal{E}_x &= \{y \mid \|y - x\|_x < 1\} \\ &= \left\{y \mid (y-x)^T \nabla^2 g(x)(y-x) < 1\right\} \end{aligned}$$

Dikin ellipsoid theorem

$$\mathcal{E}_x \subseteq \mathbf{dom}\, g$$

this follows from the upper bound on $g(y)$ and the fact that g is a closed function

Inexact proximal Newton step

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

- the exact proximal Newton step $v(x)$ at x minimizes

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

- the optimality condition for this problem is

$$0 \in \nabla g(x) + \nabla^2 g(x) v + \partial h(x+v)$$

- for $\eta \in [0, 1)$, we will call v an η**-inexact step** if

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

Bounds on suboptimality

Exact Newton method: if $v(x) = -\nabla^2 g(x)^{-1} \nabla g(x)$ satisfies

$$\|v(x)\|_x < 1$$

then g has a unique minimizer $x^\star$ and

$$g(x) - g(x^\star) \leq \omega^*(\|v(x)\|_x)$$

Inexact proximal Newton method: if v is η-inexact proximal Newton step and

$$\|v\|_x < \frac{1}{1+\eta}$$

then $f = g + h$ has a unique minimizer $x^\star$ and

$$f(x+v) - f(x^\star) \leq \omega^*(\|v\|_x) + \omega^*((1+\eta)\|v\|_x) - (1-\eta)\|v\|_x^2$$

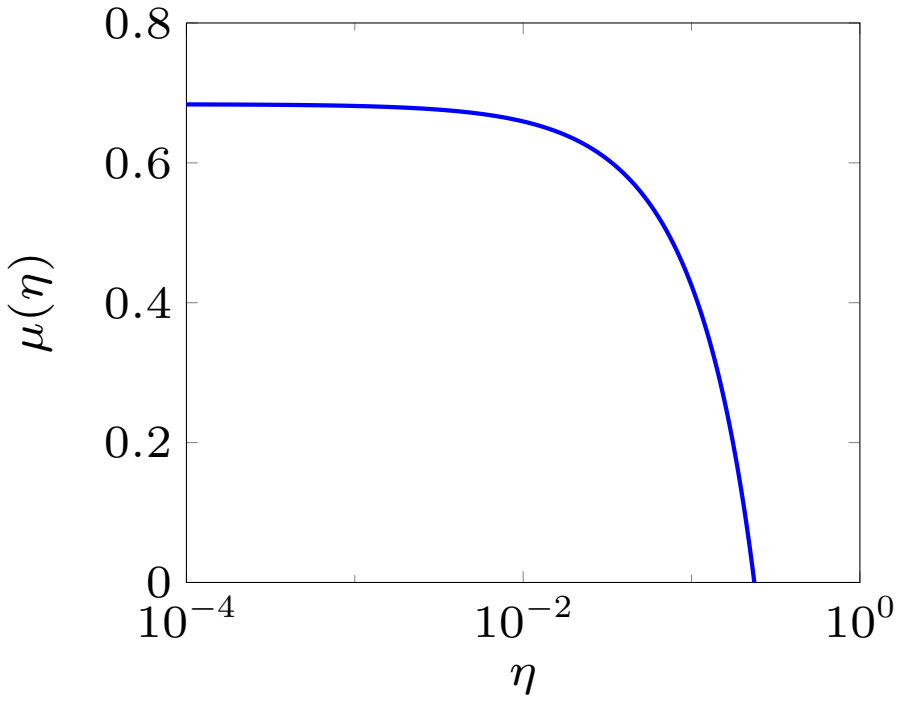
Simpler bounds on suboptimality

Exact Newton method: if $\|v(x)\|_x < 0.68$ a simpler bound holds

$$g(x) - g(x^\star) \leq \|v(x)\|_x^2$$

Inexact proximal Newton method: if v is η-inexact and $\|v\|_x \leq \mu(\eta)$

$$f(x+v) - f(x^\star) \leq \|v\|_x^2$$



Damped proximal Newton method

select $\theta \in (0, 1/4]$, $\eta_{\max} \in [0, 1)$, and a starting point $x \in \mathbf{dom}\, g$

repeat:

1. choose $\eta \in [0, \eta_{\max}]$ and compute η-inexact proximal Newton step v
2. if $f(x+v) - f(x^\star)$ is sufficiently small, return $x + v$
3. otherwise, set $x := x + \alpha v$ with

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

- bounds on suboptimality from $\|v\|_x$ are used in stopping criterion in step 2
- for $\eta = 0$, damped step size in step 3 is the standard $\alpha = 1/(1 + \|v(x)\|_x)$

Local convergence

Exact Newton method: if $\|v(x)\|_x < 1$, then

$$x^+ = x + v(x) \in \mathbf{dom}\, g, \qquad \|v(x^+)\|_{x^+} \leq \left(\frac{\|v(x)\|_x}{1 - \|v(x)\|_x}\right)^2$$

Inexact proximal Newton method

if $\|v\|_x < 1$, where v is η-inexact proximal Newton step at x, then

- $x^+ = x + v \in \mathbf{dom}\, f$
- if v^+ is η^+-inexact proximal Newton step at x^+, then

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

Simpler bounds for local convergence

Exact Newton method: if $\|v(x)\|_x \leq 0.29$, then

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

implies **quadratic** local convergence

Inexact proximal Newton method: if $\|v\|_x \leq 0.29$, then

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

- quadratic convergence if $\eta^k = 0$ for all k
- linear convergence if $\eta^k = \eta > 0$
- superlinear convergence if η^k decreases to zero

Rate of local convergence

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

Quadratic convergence: suppose $\eta^k = 0$ for all k

if $\|v(x^0)\|_{x^0} \leq 1/4$ then

$$\|v^k\|_{x^k} \leq (0.5)^{2^k+1}$$

Linear convergence: suppose $\eta^k = \eta > 0$

if $\|v(x^0)\|_{x^0} \leq 1/4$ then

$$\|v^k\|_{x^k} \leq 0.71^k \, \|v^0\|_{x^0}$$

Global convergence

Exact Newton method: damped Newton update $x^+ = x + \alpha v(x)$ with

$$\alpha = \frac{1}{1 + \|v(x)\|_x}$$

gives nonzero reduction in function value:

$$g(x + \alpha v(x)) \leq g(x) - \omega(\|v(x)\|_x)$$

Inexact proximal Newton method: if v is η-inexact step, then $x^+ = x + \alpha v$ with

$$\alpha = \frac{1 - \eta}{1 + (1 - \eta)\|v\|_x}$$

gives nonzero reduction in function value:

$$f(x + \alpha v) \leq f(x) - \omega((1 - \eta)\|v\|_x)$$

implies convergence from any starting point if the problem is bounded below

Proximal Newton method with line search

select $\eta_{\max} \in [0, 1)$, $\gamma \in (0, (1 - \eta_{\max})/2)$, and a starting point $x \in \mathbf{dom}\, g$

repeat:

1. choose $\eta \in [0, \eta_{\max}]$ and compute η-inexact proximal Newton step v
2. if $f(x + v) - f(x^\star)$ is sufficiently small, return $x + v$
3. otherwise, find largest $\alpha \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\}$ such that

$$x + \alpha v \in \mathbf{dom}\, g, \qquad f(x + \alpha v) \leq f(x) - \alpha\gamma(1 - \eta)\|v\|_x^2$$

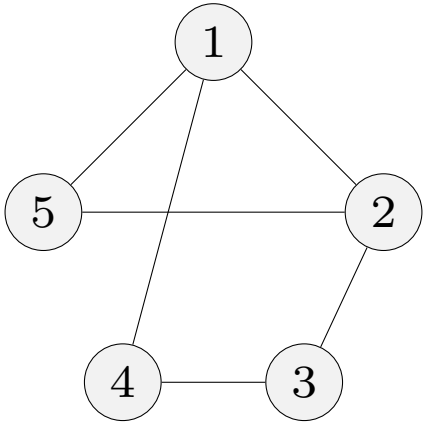
and take $x := x + \alpha v$

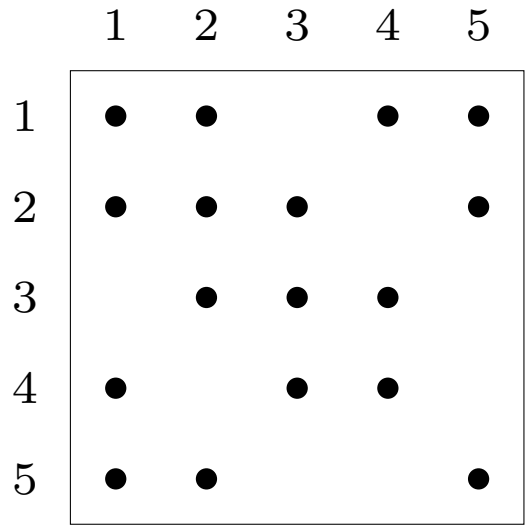
- converges globally from any starting point
- switches automatically to unit step size if $\|v\|_x$ is sufficiently small

Outline

- Analysis of proximal Newton method
- Application: restricted covariance selection

Covariance selection





Gaussian graphical model (Dempster 1972)

for $x \sim N(0, \Sigma)$, sparsity of Σ^{-1} indicates conditional independence relations

Least squares interpretation

- x is random variable (not necessarily Gaussian) with covariance Σ
- for each i, define linear least squares estimator $\hat{x}_i = \sum_{j \neq i} A_{ij} x_j$
- sparsity of coefficients A_{ij} is sparsity pattern of Σ^{-1}

Sparse covariance selection

$$\text{minimize} \quad \mathbf{tr}(CX) - \log\det X + \gamma \sum_{i>j} |X_{ij}|$$

- solution X is a sparse positive definite approximation of C^{-1}
- regularized maximum likelihood estimation of $N(0, \Sigma)$
- C is sample covariance; X is estimate of inverse of covariance Σ

(Dahl *et al.* 2005, Banerjee *et al.* 2008, Hastie & Tibshirani 2008, . . .)

Restricted covariance selection

$$\text{minimize} \quad \mathbf{tr}(CX) - \log\det X + \gamma \sum_{\{i,j\}\in E} |X_{ij}|$$

- variable $X \in \mathbf{S}^n_E$ (symmetric sparse matrices with sparsity pattern E)
- sparsity pattern E represents prior knowledge of certain zeros (*e.g.*, banded)
- penalized ML estimation is used to discover zeros within E

Proximal Newton method

$$\text{minimize} \quad \underbrace{\mathbf{tr}(CX) - \log\det X}_{g(X)} + \underbrace{\gamma\|X\|_1}_{h(X)}$$

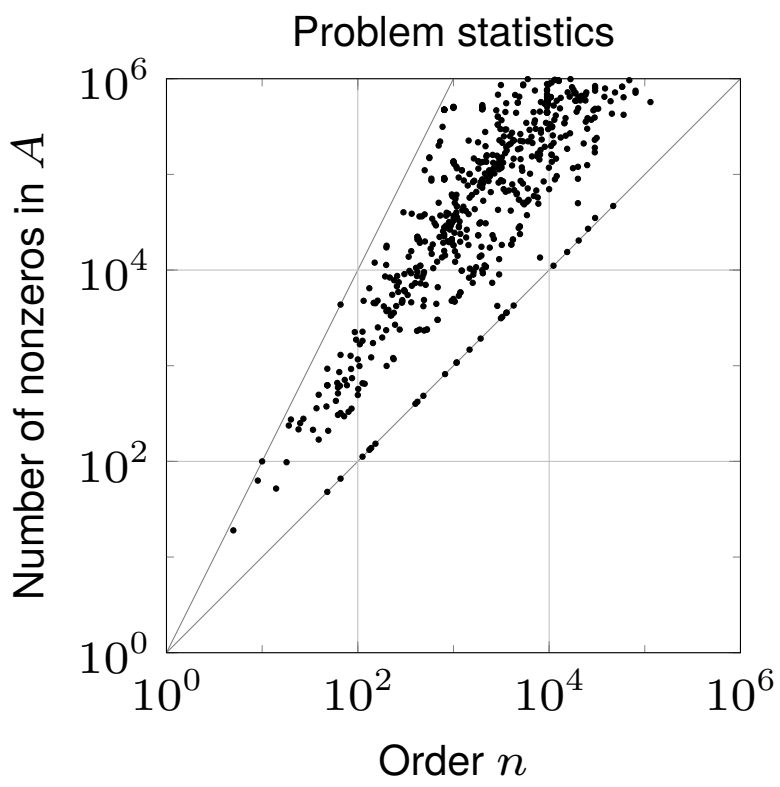
- proximal operator of $h : \mathbf{S}_E^n \to \mathbf{R}$ is component-wise soft-thresholding
- $g : \mathbf{S}_E^n \to \mathbf{R}$ is self-concordant
- gradient of g is projection of inverse on sparsity pattern

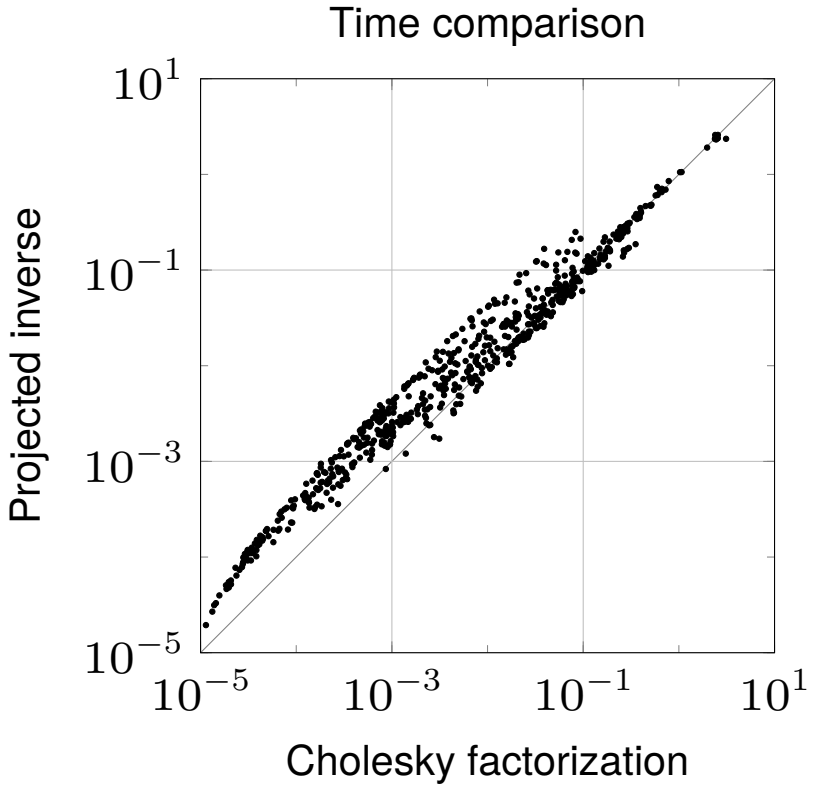
$$\nabla g(X) = \Pi_E(C - X^{-1}), \qquad (\Pi_E(A))_{ij} = \begin{cases} A_{ij} & \{i,j\} \in E \text{ or } i = j \\ 0 & \text{otherwise} \end{cases}$$

- Hessian applied to sparse matrix V is projection of $X^{-1}VX^{-1}$

$$\nabla^2 g(X)[V] = \frac{d}{d\alpha} g(X + \alpha V)\Big|_{\alpha=0} = \Pi_E(X^{-1}VX^{-1})$$

Complexity of evaluating gradient and Hessian

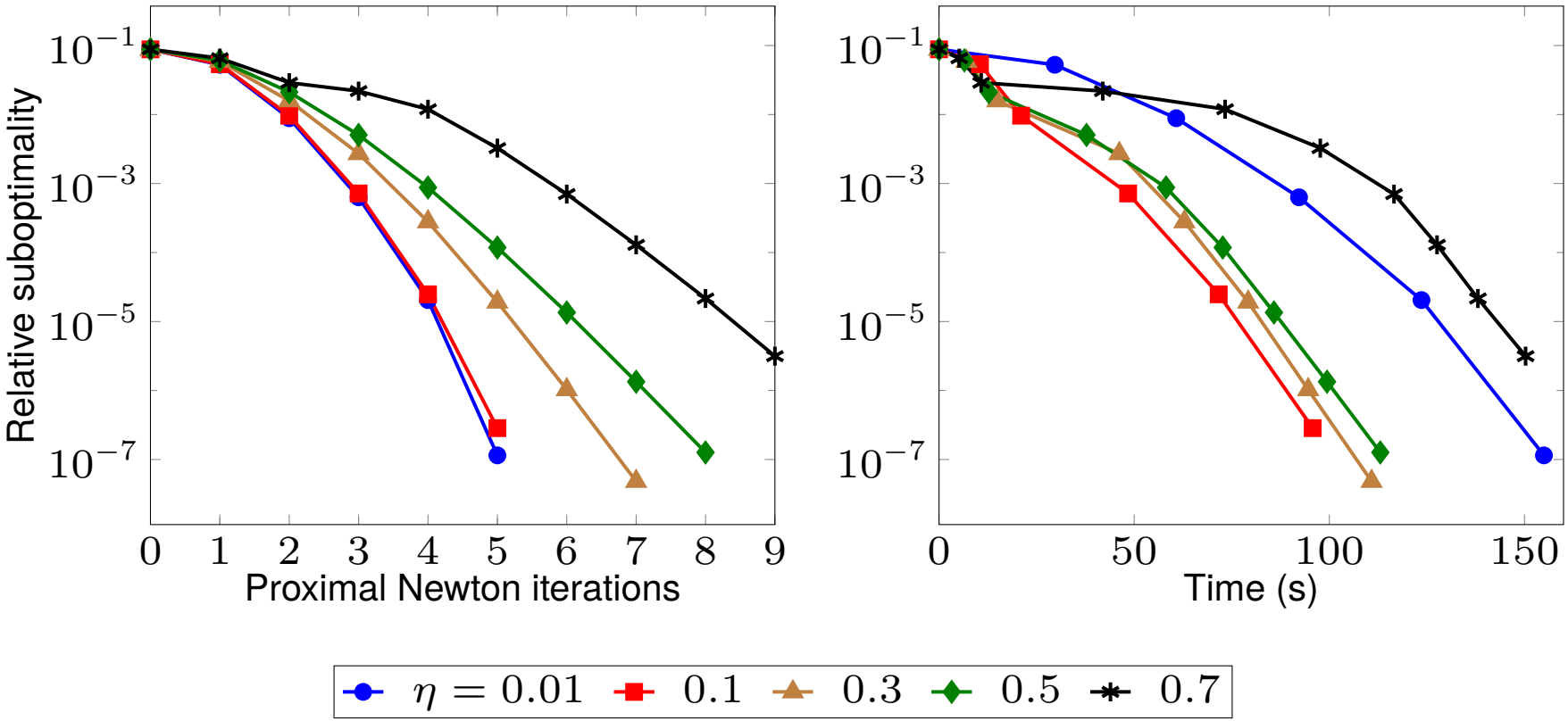




- 667 patterns from University of Florida Sparse Matrix Collection
- time in seconds for gradient $\Pi_E(X^{-1})$ and Cholesky factorization
- similar results for Hessian $\nabla^2 g(X)[V]$ and inverse Hessian $\nabla^2 g(X)^{-1}[V]$
- package for chordal matrix computations `cvxopt.github.io/chompack`

Band patterns

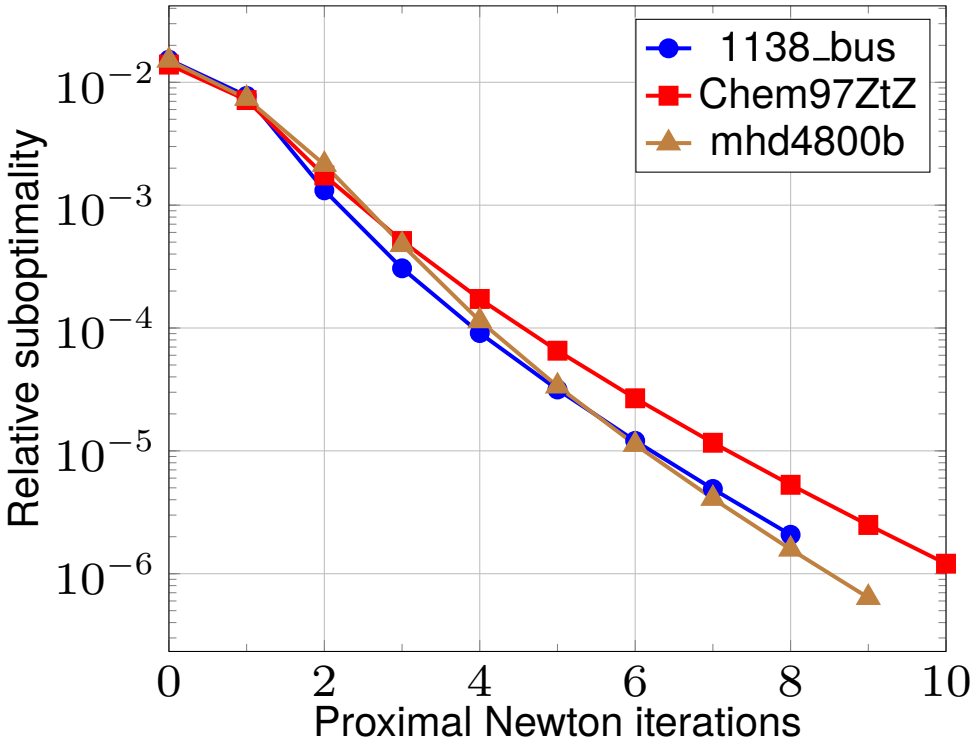
- band matrices of size 1000 with half-bandwith 20; Σ^{-1} has 80% zeros in E
- C generated from 10,000 samples
- proximal Newton steps computed by FISTA, with different levels of accuracy η



Sparsity patterns from University of Florida collection

	1138_bus	Chem97ZtZ	mhd4800b
matrix size n	1138	2541	4800
#nonzeros	4054	7361	27520

- Σ^{-1} has 70% zeros in E
- proximal Newton steps computed by FISTA ($\eta = 0.5$)



Summary

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

Proximal Newton step: v is approximate solution of

$$\text{minimize} \quad \nabla g(x)^T v + \frac{1}{2} v^T \nabla^2 g(x) v + h(x+v)$$

Analysis for self-concordant functions

extends theory for Newton method to proximal Newton method, inexact steps

Application to restricted covariance selection

exploits efficient algorithms for evaluating gradient, Hessian, inverse Hessian