Proximal Newton-type methods for minimizing composite function

Jason Lee, Yuekai Sun, Michael Saunders

Institute for Computational and Mathematical Engineering, Stanford University

Minimizing composite functions

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

- g and h are closed, convex functions
- g is continuously differentiable, and its gradient ∇g is Lipschitz continuous
- h is not necessarily everywhere differentiable, but its *proximal mapping* can be evaluated efficiently

Sparse inverse covariance estimation

- $x^{(1)}, \ldots, x^{(m)} \in \mathbf{R}^n$ are *i.i.d.* samples from a Gaussian MRF:
$$\Pr(x; \Theta) \propto \exp(x^T \Theta x/2 - \log\det(\Theta))$$
- We form the sample covariance matrix $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m x^{(i)} x^{(i)T}$ and seek a sparse maximum likelihood estimate of Θ:
$$\underset{\Theta \in \mathbf{R}^{n\times n}}{\text{minimize}}\ -\text{logdet}(\Theta) + \text{tr}(\hat{\Sigma}\Theta) + \lambda \|\text{vec}(\Theta)\|_1$$

Proximal Newton-type methods

Main idea: use a local quadratic model (in lieu of a simple quadratic model) to account for the curvature of g:

$$\Delta x_k := \arg\min_d\ \nabla g(x_k)^T d + \underbrace{\frac{1}{2} d^T H_k d}_{\text{approx. Hessian term}} + h(x_k + d).$$

Δx_k can be expressed as

$$H_k \Delta x_k \in -\underbrace{\nabla g(x_k)}_{\text{forward/explicit gradient}} - \underbrace{\partial h(x_k + \Delta x_k)}_{\text{backward/implicit subgradient}}.$$

A generic proximal Newton-type method

Algorithm 1 A generic proximal Newton-type method

Require: starting point $x_0 \in \mathbf{dom}\, f$

1: **repeat**

2: Choose an approximation to the Hessian H_k.

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.

Convergence of proximal Newton-type methods

Global convergence:
- smallest eigenvalue of H_k's bounded away from zero

Local quadratic convergence (prox-Newton method):
- g is locally strongly convex
- $\nabla^2 g$ is locally Lipschitz continuous

Local superlinear convergence (prox-quasi-Newton methods):
- assumptions for quadratic convergence
- eigenvalues of H_k's bounded and H_k's satisfy:

$$\lim_{k\to\infty} \frac{\left\| \left(H_k - \nabla^2 g(x^\star)\right)(x_{k+1} - x_k) \right\|_2}{\|x_{k+1} - x_k\|_2} = 0.$$

Inexact proximal Newton-type methods

Main idea: no need to solve the subproblem exactly only need a good enough search direction.

$$\Delta x_k \approx \arg\min_d\ \nabla g(x_k)^T d + \frac{1}{2} d^T H_k d + h(x_k + d).$$

- We solve the subproblem approximately with an iterative method, terminating (sometimes very) early
- number of iterations may increase, but computational expense per iteration is smaller
- many practical implementations use inexact search directions

Another idea: choose H_k so the subproblem is easy to solve.

Early stopping conditions

Intuition: solve the subproblem almost exactly when
- x_k is close to the optimal solution
- H_{k-1} captures the curvature of g

Typical stopping condition:

$$\underbrace{\|\nabla g_k(x_k) + H_k \Delta x_k + \partial h(x_k + \Delta x_k)\|}_{\text{optimality of subproblem solution}} \le \eta_k \underbrace{\|G_f(x_k)\|}_{\text{optimality of } x_k}$$

choose η_k based on how good the quadratic model is:

$$\eta_k \sim \frac{\|\nabla g_{k-1}(x_{k-1}) + H_{k-1}\Delta x_{k-1} - \nabla g(x_k)\|_2}{\|\nabla g(x_{k-1})\|_2}$$

Convergence of the inexact prox-Newton method

Local linear convergence (inexact prox-Newton method):
- g is locally strongly convex
- $\nabla^2 g$ is locally Lipschitz continuous
- η_k is smaller than some $\bar{\eta}$

Local superlinear convergence (...):
- assumptions for linear convergence
- η_k decays to zero (*e.g.* under our choice of forcing term)

Sparse inverse covariance estimation

- We have samples $x^{(1)}, \ldots, x^{(m)} \in \mathbf{R}^n$ from a Gaussian MRF.
- We form the sample covariance matrix $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m x^{(i)} x^{(i)T}$ and seek a sparse maximum likelihood estimate of Θ:

$$\underset{\Theta \in \mathbf{R}^{n\times n}}{\text{minimize}}\ -\text{logdet}(\Theta) + \text{tr}(\hat{\Sigma}\Theta) + \lambda \|\text{vec}(\Theta)\|_1$$

Datasets:
- Estrogen: a gene expression dataset consisting of 682 probe sets collected from 158 patients
- Leukemia: another gene expression dataset consisting of 1255 genes from 72 patients

Q: How do inexact search directions affect convergence?

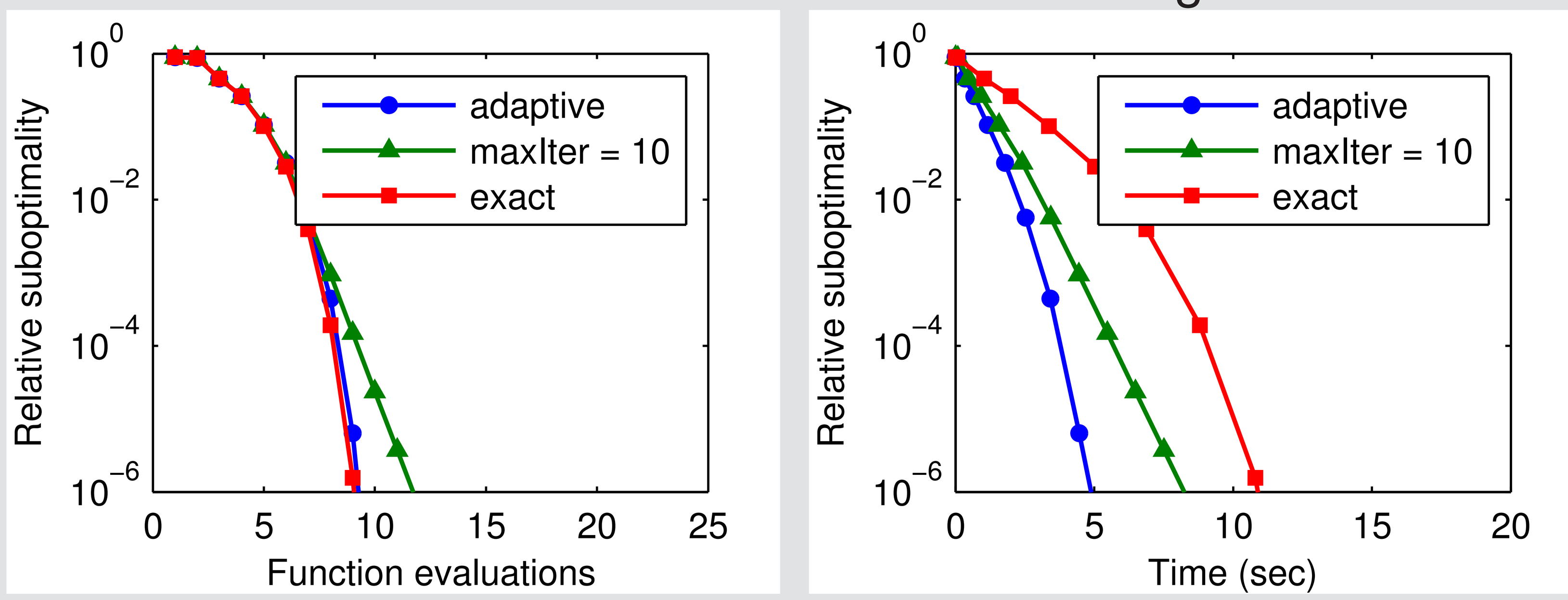

Figure : Estrogen dataset

Summary

Proximal Newton-type methods
- converge rapidly near the optimal solution, and can produce a solution of high accuracy.
- are insensitive to the condition number of the sublevel sets of the objective.
- are suited to problems where g, ∇g is expensive to evaluate compared to h, prox_h.

2013 MSM Consortium Meeting, Bethesda, MD, Oct 2–3, 2013 (canceled by US Government shutdown)

{jdl17,yuekai,saunders}@stanford.edu