# Proximal Newton-type methods for convex optimization

Yuekai Sun

Joint work with Jason Lee and Michael Saunders

October 10, 2012

## Convex optimization problems in composite form

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

- $g$ and $h$ are closed, proper convex functions.
- $g$ is continuously differentiable, and its gradient is Lipschitz continuous with constant $L_1$.
- $h$ is not necessarily everywhere differentiable but its *proximal mapping* can be evaluated efficiently.
- Example: the lasso:

$$\underset{\beta}{\text{minimize}}\ \frac{1}{2}||Y - X\beta||^2 + \lambda||\beta||_1.$$

## The proximal mapping

$$\text{prox}_h(x) = \underset{y}{\text{argmin}} \; h(y) + \frac{1}{2}||y - x||^2.$$

- $\text{prox}_h(x)$ exists and is unique for all $x \in \text{dom} h$.
- Proximal mappings generalize projections onto closed convex sets:
  - If $h$ is the indicator function of a convex set $C$ , then $\text{prox}_h(x)$ is the projection of $x$ onto the set.
- Example: soft-thresholding: If $h(x) = ||x||_1$, then

$$\text{prox}_{th}(x) = \begin{cases} x_i - t & x_i \geq t \\ 0 & -t \leq x_i \leq t \\ x_i + t & x_i \leq t \end{cases}.$$

## The proximal gradient iteration

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

The proximal gradient method: choose $x_0 \in \mathrm{dom} f$ and repeat:

$$x_{k+1} = \mathrm{prox}_{t_k h}(x_k - t_k \nabla g(x_k)).$$

The proximal gradient method step minimizes $h$ plus a *simple quadratic*:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{t_k h}\left(x_k - t_k \nabla g(x_k)\right) \\ &= \underset{y}{\mathrm{argmin}}\ \nabla g(x_k)^T (y - x_k) + \frac{1}{2t_k}\|y - x_k\|^2 + h(y), \end{aligned}$$

The proximal gradient step at $x$ is zero if and only if $x$ minimizes $f$.

## Proximal Newton-type methods

We use a *local quadratic approximation* to $g$ in lieu of the simple quadratic:

$$Q_k(d) = \nabla g(x_k)^T d + \frac{1}{2} d^T H_k d.$$

The proximal Newton step minimizes $h$ plus $Q_k$:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_d \; Q_k(d) + h(x_k + d) \\ &= \operatorname*{argmin}_y \; \nabla g(x_k)^T (y - x_k) + \frac{1}{2}(y - x_k)^T H_k (y - x_k) + h(y). \end{aligned}$$

There are many variants of proximal Newton-type methods:

- set $H_k = \nabla^2 g(x_k)$.
- update $H_k$ using a quasi-Newton strategy.
- solve the subproblem inexactly.

## Scaled proximal mappings

Let $h$ be a convex function and $H$, a positive definite matrix. Then the scaled proximal mapping of $h$ at $x$ is defined to be

$$\text{prox}_h^H(x) := \underset{y}{\text{argmin}} \; h(y) + \frac{1}{2}||y - x||_H^2.$$

Scaled proximal mappings share many properties with proximal mappings.

- $\text{prox}_h^H(x)$ exists and is unique for $x \in \text{dom} h$.
- Let $\partial h(y)$ denotes the subdifferential of $h$ at $y$. $\text{prox}_h^H(x)$ satisfies

$$y = \text{prox}_h^H(x) \iff H(x - y) \in \partial h(y).$$

## The proximal Newton-type iteration

The proximal Newton-type iteration: choose $x_0 \in \mathrm{dom} f$ and repeat:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_h^{H_k}(x_k - H_k^{-1}\nabla g(x_k)) \\ &= \underset{y}{\mathrm{argmin}} \ \nabla g(x_k)^T(y - x_k) + \frac{1}{2}||y - x_k||_{H_k}^2 + h(y). \end{aligned}$$

Examples of proximal Newton-type methods:

- `glmnet`: the lasso, $\ell_1$ regularized logistic regression, elastic net penalty
- `LIBLINEAR`: $\ell_1$ regularized logistic regression,
- `QUIC`: sparse inverse covariance estimation

## A generic proximal Newton-type method

Choose $x_0 \in \mathrm{dom} f$ and repeat:

1. Choose a local approximation to the Hessian $H_k$.
2. Compute a search direction $\Delta x_k$ via the solution of

$$\Delta x_k = \underset{d}{\mathrm{argmin}} \ \nabla g(x_k)^T d + \frac{1}{2} d^T H_k d + h(x_k + d)$$

3. Select a step length $t_k$ using a line search procedure.
4. Set $x_{k+1} \leftarrow x_k + t_k \Delta x_k$.

## Global convergence

Suppose $H_k \succeq mI$, $k = 1, 2, \ldots$ for some $m > 0$. Then the sequence $\{x_k\}$ generated by a proximal Newton-type method converges to a minimizer of $f$.

**Key ingredients of proof:**

- The proximal Newton search direction satisfies.

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

- $H_k \succeq mI$, $k = 1, 2, \ldots$ ensure the step lengths $t_k$ are bounded away from zero, so the objective function decreases at every iteration.

## Forward-backward splitting

Let $y^\star$ denote $\text{prox}_h^H\left(x - H^{-1}\nabla g(x)\right)$, then

$$H\left(x - H^{-1}\nabla g(x) - y^\star\right) \in \partial h(u).$$

or equivalently

$$\left[H - \nabla g\right](x) \in \left[H + \partial h\right](y^\star).$$

We rearrange to obtain:

$$y^\star = \underbrace{\left[\frac{1}{m}(H + \partial h)\right]^{-1}}_{\text{backward}} \underbrace{\left[\frac{1}{m}(H - \nabla g)\right]}_{\text{forward}}(x) = R \circ S(x).$$

## Convergence rate: proximal Newton method

Suppose (i) $\nabla^2 g \succeq mI$ and (ii) $\nabla^2 g$ is Lipschitz continuous with constant $L_2$. If we let $H_k = \nabla^2 g(x_k)$, $k = 1, 2, \ldots$, then the sequence $\{x_k\}$ converges to $x^\star$ Q-quadratically; *i.e.*

$$\frac{||x_{k+1} - x^\star||}{||x_k - x^\star||^2} \to c.$$

**Sketch of proof:**

- The backward map is $R$ is a shrinking map so

$$||x_{k+1} - x^\star|| = ||R \circ S(x_k) - R \circ S(x^\star)|| \leq ||S(x_k) - S(x^\star)||.$$

- We can bound $||S(x_k) - S(x^\star)||$ using Taylor's theorem and the Lipschitz continuity of $\nabla^2 g$.

## Convergence rate: proximal quasi-Newton methods

**The *Dennis–Moré criterion* for quasi-Newton updates:**

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

Common quasi-Newton strategies (such as BFGS) satisfy this criterion.

**Convergence rate of proximal quasi-Newton methods:**

Suppose $g$ is twice-continuously differentiable and the eigenvalues of $H_k$, $k = 1, 2, \ldots$ are bounded. If $\{H_k\}$ satisfy the Dennis-Moré criterion, then the sequence $\{x_k\}$ converges to $x^\star$ Q-superlinearly; *i.e.*

$$\frac{\|x_{k+1} - x^\star\|}{\|x_k - x^\star\|} \to 0.$$

## Inexact solutions to the subproblem

**$\epsilon_k$ inexact search directions:**

Let $y_k^\epsilon$ denote an $\epsilon_k$ inexact solution to the $k$th subproblem; *i.e.*

$$\|\text{prox}_h(y_k^\epsilon - \nabla Q_k(y_k^\epsilon)) - y_k^\epsilon\| \leq \epsilon_k.$$

We say $\Delta x_k^{\epsilon_k} = y_k^\epsilon - x_k$ is an $\epsilon_k$ inexact search direction.

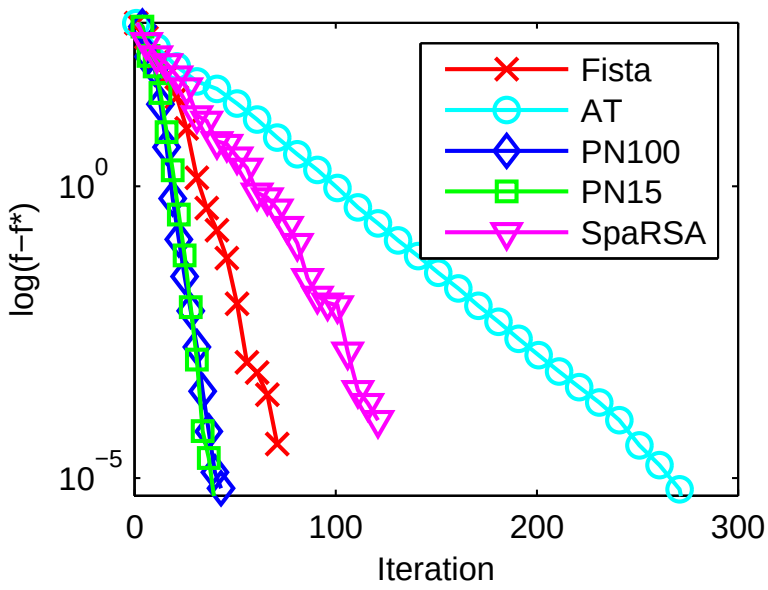
**Convergence rate of inexact proximal Newton-type methods:**

Suppose $\{x_k^{ex}\}$ are the iterates generated by an exact proximal Newton-type method that converges to an optimal solution $x^\star$. The inexact proximal Newton-type method achieves the same convergence rate if

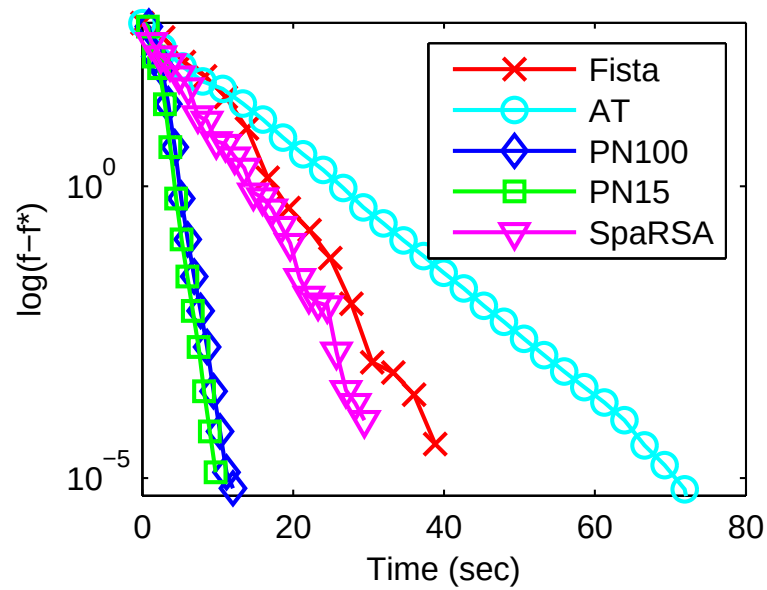
$$\{\epsilon_k\} = O(\|x_{k+1}^{ex} - x^\star\|).$$

# Computational results 1: Markov random field structure learning

$$\underset{\theta}{\text{minimize}} \; - \sum_{(r,j)\in E} \theta_{rj}(x_r, x_j) + \log Z(\theta) + \sum_{(r,j)\in E} \left(\lambda_1 \|\theta_{rj}|_2 + \lambda_2 \|\theta_{rj}|_F^2\right).$$



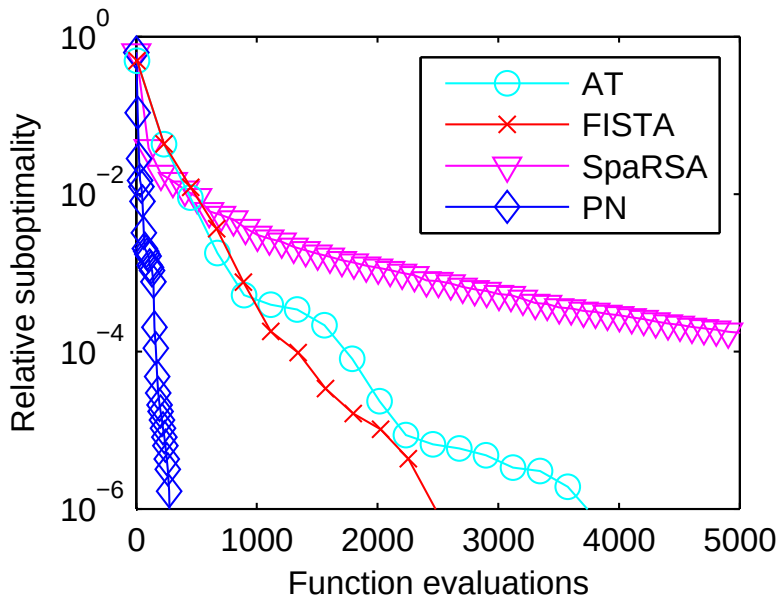

(a)



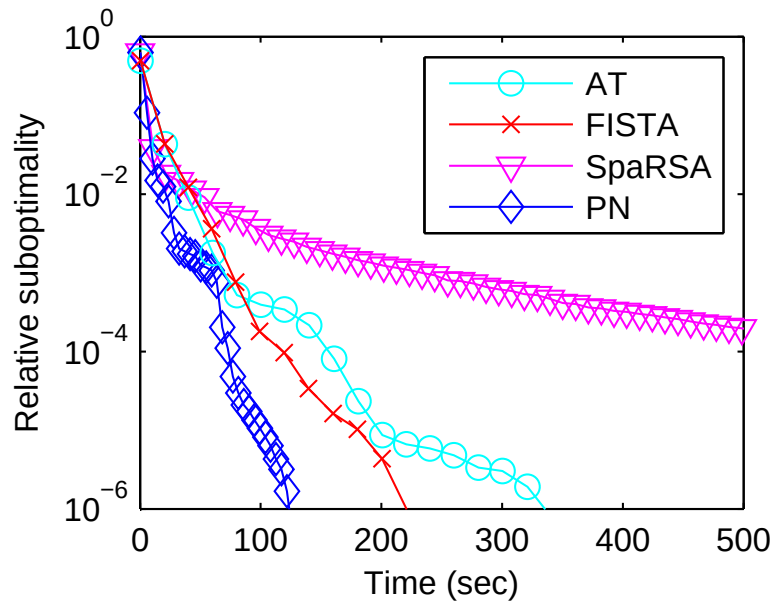

(b)

# Computational results 2: $\ell_1$ regularized logistic regression

$$\underset{w \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T x_i)) + \lambda ||w||_1.$$




(c)




(d)

## When are proximal Newton-type methods appropriate?

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

Proximal Newton-type methods are appropriate when $g$ and $\nabla g$ are expensive to evaluate compared to $\mathrm{prox}_h$.

- The computational cost is shifted to solving the subproblem, whose objective function is cheap to evaluate.
- We can interprete these methods as a proximal gradient method that uses a periodically updated quadratic surrogate of $g$ in lieu of $g$.

# Thanks for listening!

Paper on arXiv: Proximal Newton-type methods for convex optimization

Software available at `yuekai.github.com/PNOPT`.