Convergence analysis of inexact proximal Newton-type methods

Jason D. Lee and Yuekai Sun
Institute for Computational and Mathematical Engineering
Stanford University, Stanford, CA
{jdl17,yuekai}@stanford.edu

Michael A. Saunders
Department of Management Science and Engrineering
Stanford University, Stanford, CA
saunders@stanford.edu

Abstract

We study inexact proximal Newton-type methods to solve convex optimization problems in composite form:

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

where g is convex and continuously differentiable and $h : \mathbb{R}^n \to \mathbb{R}$ is a convex but not necessarily differentiable function whose proximal mapping can be evaluated efficiently. Proximal Newton-type methods require the solution of subproblems to obtain the search directions, and these subproblems are usually solved using first-order methods or coordinate descent methods, which converge to the solution linearly. In this paper, we analyze the convergence rate of inexact proximal Newton method, which solves the subproblems inexactly.

1 Introduction

Many problems of relevance in machine learning, signal processing, and high dimensional statistics can be posed in *composite form*:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x) := g(x) + h(x), \tag{1}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a convex, continuously differentiable loss function and $h : \mathbb{R}^n \to \mathbb{R}$ is a convex, continuous, but not necessarily differentiable penalty function. Such problems include: (i) the *lasso* [1] (ii) multitask learning [2] and (iii) trace-norm matrix completion [3].

We assume $g : \mathbb{R}^n \to \mathbb{R}$ is a closed and proper convex, continuously differentiable function and its gradient ∇g is Lipschitz continuous with constant L; *i.e.*

$$\|\nabla g(x) - \nabla g(y)\| \le L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$. $h : \mathbb{R}^n \to \mathbb{R}$ is a closed and proper convex but not necessarily everywhere differentiable function whose *proximal mapping* can be evaluated efficiently. We also assume the optimal value, f^*, is attained at some optimal solution $x^\star$, not necessarily unique.

1.1 Proximal Newton-type methods

Proximal Newton-type methods were previously studied in [4, 5, 6]. The Newton and quasi-Newton variants converges to an optimal solution Q-quadratically and Q-superlinearly subject to standard assumptions. We refer the reader to [6] for the convergence analysis of these methods.

At every iteration, proximal Newton-type methods must solve a subproblem to obtain the search direction Δx

$$\begin{aligned}\Delta x_k &= \arg\min_{\Delta x}\ \nabla g(x_k)^T \Delta x + \frac{1}{2}\Delta x^T H \Delta x + h(x_k + \Delta x)\\ &= \arg\min_{\Delta x}\ Q_k(x_k + \Delta x) + h(x_k + \Delta x).\end{aligned}$$

H_k is a positive definite matrix that approximates the Hessian $\nabla^2 g(x_k)$. To ensure the global convergence, a backtracking line search is typically used to select a step length t that satisfies a sufficient descent condition.

1.2 Inexact proximal Newton-type methods

In most cases, the proximal Newton subproblem is solved inexactly to reduce computational expense [7, 8, 9]. The popular methods GLMNET [8] (l_1-regularized multiclass logistic regression), LIBLINEAR [7] (l_1-regularized logistic regression), and QUIC [9] (sparse inverse covariance estimation) are special cases of proximal Newton-type methods. [6] and [5] both contain empirical results about the affect of solving the subproblem inexactly. They vary the number of iterations performed on the subproblem to see the affect on the overall convergence rate.

Inexact solutions to the subproblem yield viable descent directions empirically [6, 5]. In this paper, we analyze how these errors affect the convergence behavior of the proximal Newton iteration. We call these methods *inexact proximal Newton-type methods*.

To simplify notation, we shall drop the subscripts and say $x^+ = x + t\Delta x^\epsilon$ in lieu of $x_{k+1} = x_k + t_k \Delta x_k^{\epsilon_k}$ when discussing a single iteration.

1.3 Related work

Rockafellar analyzed the convergence behavior inexact proximal point algorithm [10]. Recently there has been renewed interest in the convergence behavior the inexact proximal gradient method and accelerated proximal gradient methods [11, 12, 13]. These papers establish the fact that inexact algorithms maintain the convergence rates of $O(1/k)$ and $O(1/k^2)$ in the non-strongly convex case, subject to certain summability conditions on the sequence of errors $\{\epsilon_k\}$. If the objective function is strongly convex, then these inexact methods attain the same linear convergence as the exact method.

Patriksson proves the inexact proximal Newton-type methods are globally convergent [4]. However, this analysis does not generalize to proximal quasi-Newton methods and does not reveal the rates of convergence. We present a classical analysis similar to [4], but yields convergence rates.

If the objective function is smooth, the seminal work of [14] analyzes the rates of convergence for inexact Newton methods. Subsequent work [15] studies various choices of stopping criterion for the Newton system and how they affects the convergence behavior of inexact Newton methods.

2 Convergence analysis

We assume the inexact solutions to the proximal Newton subproblems are accurate enough to ensure global convergence; *i.e.* the sequence $\{x_k\}$ eventually enters a neighborhood of the optimal solution $x^\star$. These assumptions are also made by Dembo et. al. in their classic analysis of inexact Newton methods for smooth optimization [14].

Definition 1. *Let y_k^ϵ denote an ϵ_k inexact solution to the kth subproblem; i.e.*

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

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

We first prove a lemma that quantifies the difference between the exact and inexact proximal Newton steps.

Lemma 2. *Let Δx^ϵ denote an ϵ search direction and Δx denote the exact search direction. Then these two search directions satisfy*

$$\|\Delta x^\epsilon - \Delta x\| \le \left(1 + \frac{1+L}{m}\right)\epsilon.$$

Proof. Let $y^\star$ denote the optimal solution to the subproblem; *i.e.* $\Delta x = y^\star - x$. We invoke a result due to Nesterov (Lemma 2 in [16]):

$$\|\operatorname{prox}_h(y^\epsilon - \nabla Q(y^\epsilon)) - y^\star\| \leq \frac{1+L}{m}\|\operatorname{prox}_h(y^\epsilon - \nabla Q(y^\epsilon)) - y^\epsilon\|.$$

The difference $\|\Delta x^\epsilon - \Delta x\|$ can be bounded in terms of $\|\operatorname{prox}_h(y^\epsilon - \nabla Q(y^\epsilon)) - y^\star\|$:

$$\begin{aligned}\|\Delta x^\epsilon - \Delta x\| &= \|y^\epsilon - y^\star\| \\ &= \|y^\epsilon - \operatorname{prox}_h(y^\epsilon - \nabla Q_k(y_k^\epsilon)) + \operatorname{prox}_h(y^\epsilon - \nabla Q_k(y_k^\epsilon)) - y^\star\| \\ &\leq \|y^\star - \operatorname{prox}_h(y^\epsilon - \nabla Q_k(y_k^\epsilon))\| + \|\operatorname{prox}_h(y^\epsilon - \nabla Q_k(y_k^\epsilon)) - y^\epsilon\|.\end{aligned}$$

We combine these two inequalities to obtain

$$\|\Delta x^\epsilon - \Delta x\| \leq \left(1 + \frac{1+L}{m}\right)\|\operatorname{prox}_h(y^\epsilon - \nabla Q(y^\epsilon)) - y^\epsilon\|.$$

Δx^ϵ is an ϵ search direction so

$$\|\Delta x^\epsilon - \Delta x\| \leq \left(1 + \frac{1+L}{m}\right)\epsilon.$$

□

We now establish the local convergence of inexact proximal Newton methods

Theorem 3. *Suppose $\{x_k\}$ eventually enters a neighborhood of the optimal solution $x^\star$, where an exact proximal Newton-type method accepts step length one and $\{\epsilon_k\} \to 0$. Then the pure inexact proximal Newton iteration*

$$x_{k+1} = x_k + \Delta x_k^{\epsilon_k}, \tag{2}$$

$$\Delta x_k^{\epsilon_k} := \operatorname{prox}_h^{H_k}\left(x_k - H_k^{-1}\nabla g(x_k)\right) - x_k, \tag{3}$$

converges to $x^\star$.

Proof. In this neighborhood of $x^\star$, the exact proximal Newton-type method accepts step length one. Let Δx_k^{ex} denote the iterates and search directions generated by an exact proximal Newton-type method. By assumption, the iteration

$$x_{k+1}^{ex} = x_k^{ex} + \Delta x_k^{ex}$$

converges if x_0 is within this neighborhood. Therefore, the sequence of exact search directions $\{\Delta x^{ex}\}$ must approach zero.

$$\|x_{k+1} - x^\star\| = \|x_k + \Delta x^{ex} - x^\star\| + \|\Delta x^{\epsilon_k} - \Delta x^{ex}\|$$

The sequence of errors also approaches zero by assumption so $\|\Delta x^{\epsilon_k} - \Delta x^{ex}\| \to 0$. Therefore, $\|x_{k+1} - x^\star\| = \|x_{k+1} - x^\star\|$ so the iterates generated by the inexact proximal Newton method also converges. □

We now analyze how the sequence of errors affect the rate of convergence of the proximal Newton method.

Theorem 4. *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 $\{\epsilon_k\} = O(\|x_{k+1}^{ex} - x^\star\|)$.*

Proof. We can split $\|x_{k+1} - x^\star\|$ into two terms:

$$\begin{aligned}\|x_{k+1} - x^\star\| &\leq \|x_k + \Delta x_k^{ex} - x^\star\| + \|\Delta x_k - \Delta x_k^{ex}\| \\ &= \|x_{k+1}^{ex} - x^\star\| + \left(1 + \frac{1+L}{m}\right)\epsilon_k.\end{aligned}$$

If $\{\epsilon_k\} = O(\|x_{k+1}^{ex} - x^\star\|)$, then convergence rate of the exact method is attained. □

Theorem 4 says the sequence of errors $\{\epsilon_k\}$ must decay quadratically and superlinearly to recover the quadratic and superlinear convergence rates of proximal Newton and proximal quasi-Newton methods.

Corollary 5. *The inexact proximal Newton method with $H_k = \nabla^2 g(x_k)$ converges quadratically if $\epsilon_k = O(\|x_k - x^*\|^2)$.*

Corollary 6. *The inexact quasi-proximal Newton method with H_k satisfying*

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

converges superlinearly if $\epsilon_k = o(\|x_k - x^*\|)$.

3 Future work

In this paper, we analyze the convergence behavior of inexact proximal Newton-type methods. We prove that such methods are locally convergent and that they attain the same convergence rates as their exact counterparts, subject to reasonable assumptions about the sequence of errors. In the future, we hope to generalize our analysis to the global setting via a line search strategy and establish global convergence and rates of convergence.

References

[1] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[2] G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[3] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[4] M. Patriksson. *Nonlinear Programming and Variational Inequality Problems*. Kluwer, 1999.

[5] M. Schmidt. *Graphical Model Structure Learning with l1-Regularization*. PhD thesis, University of British Columbia, 2010.

[6] J.D. Lee, Y. Sun, and M.A. Saunders. Proximal newton-type methods for minimizing convex objective functions in composite form. *Neural Information Processing Systems*, 2012.

[7] G.X. Yuan, C.H. Ho, and C.J. Lin. An improved glmnet for l1-regularized logistic regression. *The Journal of Machine Learning Research*, 98888:1999–2030, 2012.

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

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

[10] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[11] M. Schmidt, N.L. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415*, 2011.

[12] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *Optimization Online, E-Print*, 8(3132):2011, 2011.

[13] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *J. Convex Anal*, 19(4), 2012.

[14] R.S. Dembo, S.C. Eisenstat, and T. Steihaug. Inexact newton methods. *SIAM Journal on Numerical analysis*, 19(2):400–408, 1982.

[15] 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.

[16] Y. Nesterov. Gradient methods for minimizing composite objective function. 2007.