Approximate Newton Methods and Their Local Convergence

Haishan Ye [1] **Luo Luo** [1] **Zhihua Zhang** [2]

Abstract

Many machine learning models are reformulated as optimization problems. Thus, it is important to solve a large-scale optimization problem in big data applications. Recently, stochastic second order methods have emerged to attract much attention for optimization due to their efficiency at each iteration, rectified a weakness in the ordinary Newton method of suffering a high cost in each iteration while commanding a high convergence rate. However, the convergence properties of these methods are still not well understood. There are also several important gaps between the current convergence theory and the performance in real applications. In this paper, we aim to fill these gaps. We propose a unifying framework to analyze local convergence properties of second order methods. Based on this framework, our theoretical analysis matches the performance in real applications.

1. Introduction

Mathematical optimization is an importance pillar of machine learning. We consider the following optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where the $f_i(x)$ are smooth functions. Many machine learning models can be expressed as (1) where each f_i is the loss with respect to (w.r.t.) the i-th training sample. There are many examples such as logistic regressions, smoothed support vector machines, neural networks, and graphical models.

Many optimization algorithms to solve the problem in (1) are based on the following iteration:

$$x^{(t+1)} = x^{(t)} - \eta_t Q_t \mathrm{g}(x^{(t)}),\ t = 0, 1, 2, \ldots,$$

where $\eta_t > 0$ is the step length. If Q_t is the identity matrix and $\mathrm{g}(x^{(t)}) = \nabla F(x^{(t)})$, the resulting procedure is called *Gradient Descent* (GD) which achieves sublinear convergence for a general smooth convex objective function and linear convergence for a smooth-strongly convex objective function. When n is large, the full gradient method is inefficient due to its iteration cost scaling linearly in n. Consequently, stochastic gradient descent (SGD) has been a typical alternative (Robbins & Monro, 1951; Li et al., 2014; Cotter et al., 2011). In order to achieve cheaper cost in each iteration, such a method constructs an approximate gradient on a small mini-batch of data. However, the convergence rate can be significantly slower than that of the full gradient methods (Nemirovski et al., 2009). Thus, a great deal of efforts have been made to devise modification to achieve the convergence rate of the full gradient while keeping low iteration cost (Johnson & Zhang, 2013; Roux et al., 2012; Schmidt et al., 2013; Zhang et al., 2013).

If Q_t is a $d\times d$ positive definite matrix containing the curvature information, this formulation leads us to *second-order* methods. It is well known that second order methods enjoy superior convergence rate in both theory and practice in contrast to *first-order* methods which only make use of the gradient information. The standard Newton method, where $Q_t = [\nabla^2 F(x^{(t)})]^{-1}$, $\mathrm{g}(x^{(t)}) = \nabla F(x^{(t)})$ and $\eta_t = 1$, achieves a quadratic convergence rate for smooth-strongly convex objective functions. However, the *Newton method* takes $\mathcal{O}(nd^2 + d^3)$ cost per iteration, so it becomes extremely expensive when n or d is very large. As a result, one tries to construct an approximation of the Hessian in which way the update is computationally feasible, and while keeping sufficient second order information. One class of such methods are quasi-Newton methods, which are generalizations of the secant methods to find the root of the first derivative for multidimensional problems. The celebrated Broyden-Fletcher-Goldfarb-Shanno (BFGS) and its limited memory version (L-BFGS) are the most popular and widely used (Nocedal & Wright, 2006). They take $\mathcal{O}(nd + d^2)$ cost per iteration.

Recently, when $n \gg d$, so-called *subsampled Newton* methods have been proposed, which define an approximate Hessian matrix with a small subset of samples. The most naive approach is to sample a subset of functions f_i randomly (Roosta-Khorasani & Mahoney, 2016; Byrd et al.,

[1]Shanghai Jiao Tong University, Shanghai, China [2]Peking University & Beijing Institute of Big Data Research, Beijing, China. Correspondence to: Zhihua Zhang <zhzhang@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

2011; Xu et al., 2016) to construct a subsampled Hessian. Erdogdu & Montanari (2015) proposed a regularized subsampled Newton method called NewSamp. When the Hessian can be written as $\nabla^2 F(x) = [B(x)]^T B(x)$ where $B(x)$ is an available $n \times d$ matrix, Pilanci & Wainwright (2015) used sketching techniques to approximate the Hessian and proposed a *sketch Newton* method. Similarly, Xu et al. (2016) proposed to sample rows of $B(x)$ with non-uniform probability distribution. Agarwal et al. (2016) brought up an algorithm called LiSSA to approximate the inversion of Hessian directly.

Although the convergence performance of stochastic second order methods has been analyzed, the convergence properties are still not well understood. There are several important gaps lying between the convergence theory and real application.

The first gap is the necessity of Lipschitz continuity of Hessian. In previous work, to achieve a linear-quadratic convergence rate, stochastic second order methods all assume that $\nabla^2 F(x)$ is Lipschitz continuous. However, in real applications without this assumption, they might also converge to the optimal point. For example, Erdogdu & Montanari (2015) used NewSamp to successfully train smoothed-SVM in which the Hessian is not Lipschitz continuous.

The second gap is about the sketched size of sketch Newton methods. To obtain a linear convergence, the sketched size is $\mathcal{O}(d\kappa^2)$ in (Pilanci & Wainwright, 2015) and then is improved to $\mathcal{O}(d\kappa)$ in (Xu et al., 2016) using Gaussian sketching matrices, where κ is the condition number of the Hessian matrix in question. However, the sketch Newton empirically performs well even when the Hessian matrix is ill-conditioned. Sketched size being several tens of times or even several times of d can achieve a linear convergence rate in unconstrained optimization. But the theoretical result of Pilanci & Wainwright (2015); Xu et al. (2016) implies that sketched size may be beyond n in ill-condition cases.

The third gap is about the sample size in regularized subsampled Newton methods. In both (Erdogdu & Montanari, 2015) and (Roosta-Khorasani & Mahoney, 2016), their theoretical analysis shows that the sample size of regularized subsampled Newton methods should be set as the same as the conventional subsampled Newton method. In practice, however, adding a large regularizer can obviously reduce the sample size while keeping convergence. Thus, this contradicts the extant theoretical analysis (Erdogdu & Montanari, 2015; Roosta-Khorasani & Mahoney, 2016).

In this paper, we aim to fill these gaps between the current theory and empirical performance. More specifically, we first cast these second order methods into an algorithmic framework that we call *approximate Newton*. Then we propose a general result for analysis of local convergence properties of second order methods. Based on this framework, we give detailed theoretical analysis which matches the empirical performance very well. We summarize our contribution as follows:

- We propose a unifying framework (Theorem 3) to analyze local convergence properties of second order methods including stochastic and deterministic versions. The convergence performance of second order methods can be analyzed easily and systematically in this framework.
- We prove that the Lipschitz continuity condition of Hessian is not necessary for achieving linear and superlinear convergence in variants of subsampled Newton. But it is needed to obtain quadratic convergence. This explains the phenomenon that NewSamp (Erdogdu & Montanari, 2015) can be used to train smoothed SVM in which the Lipschitz continuity condition of Hessian is not satisfied. It also reveals the reason why previous stochastic second order methods, such as subsampled Newton, sketch Newton, LiSSA, etc., all achieve a linear-quadratic convergence rate.
- We prove that the sketched size is *independent* of the condition number of the Hessian matrix which explains that sketched Newton performs well even when the Hessian matrix is ill-conditioned.
- We provide a theoretical guarantee that adding a regularizer is an effective way to reduce the sample size in subsampled Newton methods while keeping converging. Our theoretical analysis also shows that adding a regularizer will lead to poor convergence behavior as the sample size decreases.

1.1. Organization

The remainder of the paper is organized as follows. In Section 2 we present notation and preliminaries. In Section 3 we present a unifying framework for local convergence analysis of second order methods. In Section 4 we analyze the local convergence properties of sketch Newton methods and prove that sketched size is independent of condition number of the Hessian. In Section 5 we give the local convergence behaviors of several variants of subsampled Newton method. Especially, we reveal the relationship among the sample size, regularizer and convergence rate. In Section 6, we derive the local convergence properties of inexact Newton methods from our framework. In Section 7, we validate our theoretical results experimentally. Finally, we conclude our work in Section 8. All the proofs are presented in the supplementary metarials.

2. Notation and Preliminaries

In this section, we introduce the notation and preliminaries that will be used in this paper.

2.1. Notation

Given a matrix $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ of rank ℓ and a positive integer $k \leq \ell$, its condensed SVD is given as $A = U\Sigma V^T = U_k\Sigma_k V_k^T + U_{\backslash k}\Sigma_{\backslash k}V_{\backslash k}^T$, where U_k and $U_{\backslash k}$ contain the left singular vectors of A, V_k and $V_{\backslash k}$ contain the right singular vectors of A, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_\ell)$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\ell > 0$ are the nonzero singular values of A. We use $\sigma_{\max}(A)$ to denote the largest singular value and $\sigma_{\min}(A)$ to denote the smallest non-zero singular value. Thus, the condition number of A is defined by $\kappa(A) \triangleq \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$. If A is positive semidefinite, then $U = V$ and the square root of A can be defined as $A^{1/2} = U\Sigma^{1/2}U^T$. It also holds that $\lambda_i(A) = \sigma_i(A)$, where $\lambda_i(A)$ is the i-th largest eigenvalue of A, $\lambda_{\max}(A) = \sigma_{\max}(A)$, and $\lambda_{\min}(A) = \sigma_{\min}(A)$.

Additionally, $\|A\| \triangleq \sigma_1$ is the spectral norm. Given a positive definite matrix M, $\|x\|_M \triangleq \|M^{1/2}x\|$ is called the M-norm of x. Give square matrices A and B with the same size, we denote $A \preceq B$ if $B - A$ is positive semidefinite.

2.2. Randomized sketching matrices

We first give an ϵ-subspace embedding property which will be used to sketch Hessian matrices. Then we list two most popular types of randomized sketching matrices.

Definition 1 *$S \in \mathbb{R}^{s\times m}$ is said to be an ϵ-subspace embedding matrix w.r.t. a fixed matrix $A \in \mathbb{R}^{m\times d}$ where $d < m$, if $\|SAx\|^2 = (1 \pm \epsilon)\|Ax\|^2$ (i.e., $(1-\epsilon)\|Ax\|^2 \leq \|SAx\|^2 \leq (1+\epsilon)\|Ax\|^2$) for all $x \in \mathbb{R}^d$.*

From the definition of the ϵ-subspace embedding matrix, we can derive the following property directly.

Lemma 2 *$S \in \mathbb{R}^{s\times m}$ is an ϵ-subspace embedding matrix w.r.t. the matrix $A \in \mathbb{R}^{m\times d}$ if and only if*

$$(1-\epsilon)A^TA \preceq A^TS^TSA \preceq (1+\epsilon)A^TA.$$

Leverage score sketching matrix. A leverage score sketching matrix $S = D\Omega \in \mathbb{R}^{s\times m}$ w.r.t. $A \in \mathbb{R}^{m\times d}$ is defined by sampling probabilities p_i, a sampling matrix $\Omega \in \mathbb{R}^{m\times s}$ and a diagonal rescaling matrix $D \in \mathbb{R}^{s\times s}$. Specifically, we construct S as follows. For every $j = 1, \ldots, s$, independently and with replacement, pick an index i from the set $\{1, 2 \ldots, m\}$ with probability p_i, and set $\Omega_{ji} = 1$ and $\Omega_{jk} = 0$ for $k \neq i$ as well as $D_{jj} = 1/\sqrt{p_i s}$. The sampling probabilities p_i are the leverage scores of A defined as follows. Let $V \in \mathbb{R}^{m\times d}$ be the column orthonormal basis of A, and let $v_{i,*}$ denote the i-th row of V. Then $\ell_i \triangleq \|v_{i,*}\|^2/d$ for $i = 1, \ldots, m$ are the leverage scores of A. To achieve an ϵ-subspace embedding property w.r.t. A, $s = \mathcal{O}(d\log d/\epsilon^2)$ is sufficient.

Sparse embedding matrix. A sparse embedding matrix $S \in \mathbb{R}^{s\times m}$ is such a matrix in each column of which there is only one nonzero entry uniformly sampled from $\{1, -1\}$ (Clarkson & Woodruff, 2013). Hence, it is very efficient to compute SA, especially when A is sparse. To achieve an ϵ-subspace embedding property w.r.t. $A \in \mathbb{R}^{m\times d}$, $s = \mathcal{O}(d^2/\epsilon^2)$ is sufficient (Meng & Mahoney, 2013; Woodruff, 2014).

Other sketching matrices such as Gaussian Random Projection and Subsampled Randomized Hadamard Transformation as well as their properties can be found in the survey (Woodruff, 2014).

2.3. Assumptions and Notions

In this paper, we focus on the problem described in Eqn. (1). Moreover, we will make the following two assumptions.

Assumption 1 The objective function F is μ-strongly convex, that is,

$$F(y) \geq F(x) + [\nabla F(x)]^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \text{ for } \mu > 0.$$

Assumption 2 $\nabla F(x)$ is L-Lipschitz continuous, that is,

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|y-x\|, \text{ for } L > 0.$$

Assumptions 1 and 2 imply that for any $x \in \mathbb{R}^d$, we have

$$\mu I \preceq \nabla^2 F(x) \preceq LI,$$

where I is the identity matrix of appropriate size. With a little confusion, we define $\kappa \triangleq \frac{L}{\mu}$. In fact, κ is an upper bound of condition number of the Hessian matrix $\nabla^2 F(x)$ for any x.

Besides, if $\nabla^2 F(x)$ is Lipschitz continuous, then we have

$$\|\nabla^2 F(x) - \nabla^2 F(y)\| \leq \hat{L}\|x-y\|,$$

where $\hat{L} > 0$ is the Lipschitz constant of $\nabla^2 F(x)$.

Throughout this paper, we use notions of linear convergence rate, superlinear convergence rate and quadratic convergence rate. In our paper, the convergence rates we will use are defined w.r.t. $\|\cdot\|_{M^*}$, where $M^* = [\nabla^2 F(x^*)]^{-1}$ and x^* is the optimal solution to Problem (1). A sequence of vectors $\{x^{(t)}\}$ is said to converge linearly to a limit point x^*, if for some $0 < \rho < 1$,

$$\limsup_{t\to\infty} \frac{\|\nabla F(x^{(t+1)})\|_{M^*}}{\|\nabla F(x^{(t)})\|_{M^*}} = \rho.$$

Similarly, superlinear convergence and quadratic convergence are respectively defined as

$$\limsup_{t\to\infty}\frac{\|\nabla F(x^{(t+1)})\|_{M^*}}{\|\nabla F(x^{(t)})\|_{M^*}}=0, \limsup_{t\to\infty}\frac{\|\nabla F(x^{(t+1)})\|_{M^*}}{\|\nabla F(x^{(t)})\|^2_{M^*}}=\rho.$$

We call it the linear-quadratic convergence rate if the following condition holds:

$$\|\nabla F(x^{(t+1)})\|_{M^*}\leq\rho_1\|\nabla F(x^{(t)})\|_{M^*}+\rho_2\|\nabla F(x^{(t)})\|^2_{M^*},$$

where $0<\rho_1<1$.

3. Approximate Newton Methods and Local Convergence Analysis

The existing variants of stochastic second order methods share some important attributes. First, these methods such as NewSamp (Erdogdu & Montanari, 2015), LiSSA (Agarwal et al., 2016), subsampled Newton with conjugate gradient (Byrd et al., 2011), and subsampled Newton with non-uniformly sampling (Xu et al., 2016), all have the same convergence properties; that is, they have a linear-quadratic convergence rate.

Second, they also enjoy the same algorithm procedure summarized as follows. In each iteration, they first construct an approximate Hessian matrix $H^{(t)}$ such that

$$(1-\epsilon_0)H^{(t)}\preceq\nabla^2F(x^{(t)})\preceq(1+\epsilon_0)H^{(t)}, \tag{2}$$

where $0\leq\epsilon_0<1$. Then they solve the following optimization problem

$$\min_p\frac{1}{2}p^TH^{(t)}p-p^T\nabla F(x^{(t)}) \tag{3}$$

approximately or exactly to obtain the direction vector $p^{(t)}$. Finally, their update equation is given as $x^{(t+1)}=x^{(t)}-p^{(t)}$. With this procedure, we regard these stochastic second order methods as *approximate Newton* methods.

In the following theorem, we propose a unifying framework which describes the convergence properties of the second order optimization procedure depicted above.

Theorem 3 *Let Assumptions 1 and 2 hold. Suppose that $\nabla^2F(x)$ exists and is continuous in a neighborhood of a minimizer x^*. $H^{(t)}$ is a positive definite matrix that satisfies Eqn. (2) with $0\leq\epsilon_0<1$. Let $p^{(t)}$ be an approximate solution of Problem (3) such that*

$$\|\nabla F(x^{(t)})-H^{(t)}p^{(t)}\|\leq\frac{\epsilon_1}{\kappa}\|\nabla F(x^{(t)})\|, \tag{4}$$

where $0<\epsilon_1<1$. Define the iteration $x^{(t+1)}=x^{(t)}-p^{(t)}$.

(a) There exists a sufficient small value γ, $0<\nu(t)<1$, and $0<\eta(t)<1$ such that when $\|x^{(t)}-x^*\|\leq\gamma$, we have that

$$\begin{aligned}&\|\nabla F(x^{(t+1)})\|_{M^*}\\&\leq\left(\epsilon_0+\frac{\epsilon_1}{1-\epsilon_0}+\frac{2\eta(t)}{1-\epsilon_0}\right)\frac{1+\nu(t)}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*}.\end{aligned} \tag{5}$$

Besides, $\nu(t)$ and $\eta(t)$ will go to 0 as $x^{(t)}$ goes to x^*.

(b) Furthermore, if $\nabla^2F(x)$ is Lipschitz continuous with parameter $\hat{L}$, and $x^{(t)}$ satisfies

$$\|x^{(t)}-x^*\|\leq\frac{\mu}{\hat{L}\kappa}\nu(t), \tag{6}$$

where $0<\nu(t)<1$, then it holds that

$$\begin{aligned}&\|\nabla F(x^{(t+1)})\|_{M^*}\\&\leq\left(\epsilon_0+\frac{\epsilon_1}{1-\epsilon_0}\right)\frac{1+\nu(t)}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*}\\&\quad+\frac{2}{(1-\epsilon_0)^2}\frac{\hat{L}\kappa}{\mu\sqrt{\mu}}\frac{(1+\nu(t))^2}{1-\nu(t)}\|\nabla F(x^{(t)})\|^2_{M^*}.\end{aligned} \tag{7}$$

From Theorem 3, we can find some important insights. First, Theorem 3 provides sufficient conditions to get different convergence rates including super-liner and quadratic convergence rates. If $\left(\epsilon_0+\frac{\epsilon_1}{1-\epsilon_0}\right)$ is a constant, then sequence $\{x^{(t)}\}$ converges linearly because $\nu(t)$ and $\eta(t)$ will go to 0 as t goes to infinity. If we set $\epsilon_0=\epsilon_0(t)$ and $\epsilon_1=\epsilon_1(t)$ such that $\epsilon_0(t)$ and $\epsilon_1(t)$ decrease to 0 as t increases, then sequence $\{x^{(t)}\}$ will converge superlinearly. Similarly, if $\epsilon_0(t)=\mathcal{O}(\|\nabla F(x^{(t)})\|_{M^*})$, $\epsilon_1(t)=\mathcal{O}(\|\nabla F(x^{(t)})\|_{M^*})$, and $\nabla^2F(x)$ is Lipschitz continuous, then sequence $\{x^{(t)}\}$ will converge quadratically.

Second, Theorem 3 makes it clear that the Lipschitz continuity of $\nabla^2F(x)$ is *not necessary* for linear convergence and superlinear convergence of stochastic second order methods including Subsampled Newton method, Sketch Newton, NewSamp, etc. This reveals the reason why NewSamp can be used to train the smoothed SVM where the Lipschitz continuity of the Hessian matrix is not satisfied. The Lipschitz continuity condition is only needed to get a quadratic convergence or linear-quadratic convergence. This explains the phenomena that LiSSA(Agarwal et al., 2016), NewSamp (Erdogdu & Montanari, 2015), subsampled Newton with non-uniformly sampling (Xu et al., 2016), Sketched Newton (Pilanci & Wainwright, 2015) have linear-quadratic convergence rate because they all assume that the Hessian is Lipschitz continuous. In fact, it is well known that the Lipschitz continuity condition of $\nabla^2F(x)$ is not necessary to achieve a linear or superlinear convergence rate for inexact Newton methods.

Algorithm 1 Sketch Newton.

Input: $x^{(0)}, 0 < \delta < 1, 0 < \epsilon_0 < 1$;
for $t = 0, 1, \ldots$ until termination **do**
  Construct an ϵ_0-subspace embedding matrix S for $B(x^{(t)})$ and where $\nabla^2 F(x)$ is of the form $\nabla^2 F(x) = (B(x^{(t)}))^T B(x^{(t)})$, and calculate $H^{(t)} = [B(x^{(t)})]^T S^T S B(x^{(t)})$;
  Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$;
  Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
end for

Third, the unifying framework of Theorem 3 contains not only stochastic second order methods, but also the deterministic versions. For example, letting $H^{(t)} = \nabla^2 F(x^{(t)})$ and using conjugate gradient to get $p^{(t)}$, we obtain the famous "Newton-CG" method. In fact, different choice of $H^{(t)}$ and different way to calculate $p^{(t)}$ lead us to different second order methods. In the following sections, we will use this framework to analyze the local convergence performance of these second order methods in detail.

4. Sketch Newton Method

In this section, we use Theorem 3 to analyze the local convergence properties of Sketch Newton (Algorithm 1). We mainly focus on the case that the Hessian matrix is of the form

$$\nabla^2 F(x) = B(x)^T B(x) \tag{8}$$

where $B(x)$ is an explicitly available $n \times d$ matrix. Our result can be easily extended to the case that

$$\nabla^2 F(x) = B(x)^T B(x) + Q(x),$$

where $Q(x)$ is a positive semi-definite matrix related to the Hessian of regularizer.

Theorem 4 *Let $F(x)$ satisfy the conditions described in Theorem 3. Assume the Hessian matrix is given as Eqn. (8). Let $0 < \delta < 1$, $0 < \epsilon_0 < 1/2$ and $0 \leq \epsilon_1 < 1$ be given. $S \in \mathbb{R}^{\ell \times n}$ is an ϵ_0-subspace embedding matrix w.r.t. $B(x)$ with probability at least $1 - \delta$, and direction vector $p^{(t)}$ satisfies Eqn. (4). Then Algorithm 1 has the following convergence properties:*

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, and $0 < \eta(t) < 1$ such that when $\|x^{(t)} - x^*\| \leq \gamma$, then each iteration satisfies Eqn. (5) with probability at least $1 - \delta$.

(b) If $\nabla^2 F(x^{(t)})$ is also Lipschitz continuous and $\{x^{(t)}\}$ satisfies Eqn. (6), then each iteration satisfies Eqn. (7) with probability at least $1 - \delta$.

Table 1. Comparison with previous work

Reference	Sketched Size
Pilanci & Wainwright (2015)	$\mathcal{O}\left(\frac{d\kappa^2 \log d}{\epsilon_0^2}\right)$
Xu et al. (2016)	$\mathcal{O}\left(\frac{d\kappa \log d}{\epsilon_0^2}\right)$
Our result(Theorem 4)	$\mathcal{O}\left(\frac{d \log d}{\epsilon_0^2}\right)$

Theorem 4 directly provides a bound of the sketched size. Using the leverage score sketching matrix as an example, the sketched size $\ell = \mathcal{O}(d \log d/\epsilon_0^2)$ is sufficient. We compare our theoretical bound of the sketched size with the ones of Pilanci & Wainwright (2015) and Xu et al. (2016) in Table 1. As we can see, our sketched size is much smaller than the other two, especially when the Hessian matrix is ill-conditioned.

Theorem 4 shows that the sketched size ℓ is *independent* on the condition number of the Hessian matrix $\nabla^2 F(x)$ just as shown in Table 1. This explains the phenomena that when the Hessian matrix is ill-conditioned, Sketch Newton performs well even when the sketched size is only several times of d. For a large condition number, the theoretical bounds of both Xu et al. (2016) and Pilanci & Wainwright (2015) may be beyond the number of samples n. Note that the theoretical results of (Xu et al., 2016) and (Pilanci & Wainwright, 2015) still hold in the constrained optimization problem. However, our result proves the effectiveness of the sketch Newton method for the unconstrained optimization problem in the ill-conditioned case.

5. The Subsampled Newton method and Variants

In this section, we apply Theorem 3 to analyze Subsampled Newton and regularized subsampled Newton method.

First, we make the assumption that each $f_i(x)$ and $F(x)$ have the following properties:

$$\max_{1 \leq i \leq n} \|\nabla^2 f_i(x)\| \leq K < \infty, \tag{9}$$

$$\lambda_{\min}(\nabla^2 F(x)) \geq \sigma > 0. \tag{10}$$

Accordingly, if $\nabla^2 F(x)$ is ill-conditioned, then the value $\frac{K}{\sigma}$ is large.

5.1. The Subsampled Newton method

The Subsampled Newton method is depicted in Algorithm 2, and we now give its local convergence properties in the following theorem.

Theorem 5 *Let $F(x)$ satisfy the properties described in Theorem 3. Assume Eqn. (9) and Eqn. (10) hold and let $0 < \delta < 1$, $0 < \epsilon_0 < 1/2$ and $0 \leq \epsilon_1 < 1$ be given. $|\mathcal{S}|$ and $H^{(t)}$ are set as in Algorithm 2, and the direction vector $p^{(t)}$ satisfies Eqn. (4). Then Algorithm 2 has the following*

Algorithm 2 Subsampled Newton.

1: **Input:** $x^{(0)}, 0 < \delta < 1, 0 < \epsilon_0 < 1$;
2: Set the sample size $|\mathcal{S}| \geq \frac{16K^2 \log(2d/\delta)}{\sigma^2 \epsilon_0^2}$.
3: **for** $t = 0, 1, \ldots$ until termination **do**
4: Select a sample set $\mathcal{S}$, of size $|\mathcal{S}|$ and construct $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$;
5: Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$;
6: Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
7: **end for**

convergence properties:

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, and $0 < \eta(t) < 1$ such that when $\|x^{(t)} - x^*\| \leq \gamma$, then each iteration satisfies Eqn. (5) with probability at least $1 - \delta$.

(b) If $\nabla^2 F(x^{(t)})$ is also Lipschitz continuous and $\{x^{(t)}\}$ satisfies Eqn. (6), then each iteration satisfies Eqn. (7) with probability at least $1 - \delta$.

As we can see, Algorithm 2 almost has the same convergence properties as Algorithm 1 except several minor differences. The main difference is the construction manner of $H^{(t)}$ which should satisfy Eqn. (2). Algorithm 2 relies on the assumption that each $\|\nabla^2 f_i(x)\|$ is upper bounded (i.e., Eqn. (9) holds), while Algorithm 1 is built on the setting of the Hessian matrix as in Eqn. (8).

5.2. Regularized Subsampled Newton

In ill-conditioned cases (i.e., $\frac{K}{\sigma}$ is large), the subsampled Newton method in Algorithm 2 should take a lot of samples because the sample size $|\mathcal{S}|$ depends on $\frac{K}{\sigma}$ quadratically. To overcome this problem, one resorts to a regularized subsampled Newton method. The key idea is to add αI to the original subsampled Hessian just as described in Algorithm 3. Erdogdu & Montanari (2015) proposed NewSamp which is another regularized subsampled Newton method depicted in Algorithm 4. In the following analysis, we prove that adding a regularizer is an effective way to reduce the sample size while keeping converging in theory.

We first give the theoretical analysis of local convergence properties of Algorithm 3.

Theorem 6 *Let $F(x)$ satisfy the properties described in Theorem 3. Assume Eqns. (9) and (10) hold, and let $0 < \delta < 1$, $0 \leq \epsilon_1 < 1$ and $0 < \alpha$ be given. Assume β is a constant such that $0 < \beta < \alpha + \frac{\sigma}{2}$, the subsampled size $|\mathcal{S}|$ satisfies $|\mathcal{S}| \geq \frac{16K^2 \log(2d/\delta)}{\beta^2}$, and $H^{(t)}$ is constructed as in Algorithm 3. Define*

$$\epsilon_0 = \max\left(\frac{\beta - \alpha}{\sigma + \alpha - \beta}, \frac{\alpha + \beta}{\sigma + \alpha + \beta}\right), \tag{11}$$

which implies that $0 < \epsilon_0 < 1$. Besides, the direction vector $p^{(t)}$ satisfies Eqn. (4). Then Algorithm 3 has the following convergence properties:

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, and $0 < \eta(t) < 1$ such that when $\|x^{(t)} - x^*\| \leq \gamma$, each iteration satisfies Eqn. (5) with probability at least $1 - \delta$.

(b) If $\nabla^2 F(x^{(t)})$ is also Lipschitz continuous and $\{x^{(t)}\}$ satisfies Eqn. (6), then each iteration satisfies Eqn. (7) with probability at least $1 - \delta$.

In Theorem 6 the parameter ϵ_0 mainly decides convergence properties of Algorithm 3. It is determined by two terms just as shown in Eqn. (11). These two terms depict the relationship among the sample size, regularizer αI, and convergence rate.

The first term describes the relationship between the regularizer αI and sample size. Without loss of generality, we set $\beta = \alpha$ which satisfies $0 < \beta < \alpha + \sigma/2$. Then the sample size $|\mathcal{S}| = \frac{16K^2 \log(2d/\delta)}{\alpha^2}$ decreases as α increases. Hence Theorem 6 gives a theoretical guarantee that adding the regularizer αI is an effective approach for reducing the sample size when K/σ is large. Conversely, if we want to sample a small part of f_i's, then we should choose a large α. Otherwise, β will go to $\alpha + \sigma/2$ which means $\epsilon_0 = 1$, i.e., the sequence $\{x^{(t)}\}$ does not converge.

Though a large α can reduce the sample size, it is at the expense of slower convergence rate just as the second term shows. As we can see, $\frac{\alpha+\beta}{\sigma+\alpha+\beta}$ goes to 1 as α increases. Besides, ϵ_1 also has to decrease. Otherwise, $\epsilon_0 + \frac{\epsilon_1}{1-\epsilon_0}$ may be beyond 1 which means that Algorithm 3 will not converge.

In fact, slower convergence rate via adding a regularizer is because the sample size becomes small, which implies less curvature information is obtained. However, a small sample size implies low computational cost in each iteration. Therefore, a proper regularizer which balances the cost of each iteration and convergence rate is the key in the regularized subsampled Newton algorithm.

Next, we give the theoretical analysis of local convergence properties of NewSamp (Algorithm 4).

Theorem 7 *Let $F(x)$ satisfy the properties described in Theorem 3. Assume Eqn. (9) and Eqn. (10) hold and let $0 < \delta < 1$ and target rank r be given. Let β be a constant such that $0 < \beta < \frac{\lambda_{r+1}^{(t)}}{2}$, where $\lambda_{r+1}^{(t)}$ is the $(r+1)$-th*

Algorithm 3 Regularized Subsample Newton.

Input: $x^{(0)}$, $0 < \delta < 1$, regularizer parameter α, sample size $|\mathcal{S}|$;
for $t = 0, 1, \ldots$ until termination **do**
  Select a sample set $\mathcal{S}$, of size $|\mathcal{S}|$ and construct $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)}) + \alpha I$;
  Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$
  Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
end for

Algorithm 4 NewSamp.

Input: $x^{(0)}$, $0 < \delta < 1$, r, sample size $|\mathcal{S}|$;
for $t = 0, 1, \ldots$ until termination **do**
  Select a sample set $\mathcal{S}$, of size $|\mathcal{S}|$ and get $H^{(t)}_{|\mathcal{S}|} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$;
  Compute rank $r+1$ truncated SVD deompostion of $H^{(t)}_{|\mathcal{S}|}$ to get U_{r+1} and $\hat{\Lambda}_{r+1}$. Construct $H^{(t)} = H^{(t)}_{|\mathcal{S}|} + U_{\setminus r}(\hat{\lambda}^{(t)}_{r+1} I - \hat{\Lambda}_{\setminus r})U^T_{\setminus r}$
  Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$
  Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
end for

eigenvalue of $\nabla^2 F(x^{(t)})$. *Set the subsampled size* $|\mathcal{S}|$ *such that* $|\mathcal{S}| \geq \frac{16K^2 \log(2d/\delta)}{\beta^2}$, *and define*

$$\epsilon_0 = \max\left(\frac{\beta}{\lambda^{(t)}_{r+1} - \beta}, \frac{2\beta + \lambda^{(t)}_{r+1}}{\sigma + 2\beta + \lambda^{(t)}_{r+1}}\right), \tag{12}$$

which implies $0 < \epsilon_0 < 1$. *Assume the direction vector* $p^{(t)}$ *satisfies Eqn.* (4). *Then Algorithm 4 has the following convergence properties:*

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, *and* $0 < \eta(t) < 1$ *such that when* $\|x^{(t)} - x^*\| \leq \gamma$, *each iteration satisfies Eqn.* (5) *with probability at least* $1 - \delta$.

(b) If $\nabla^2 F(x^{(t)})$ *is also Lipschitz continuous and* $\{x^{(t)}\}$ *satisfies Eqn.* (6), *then each iteration satisfies Eqn.* (7) *with probability at least* $1 - \delta$.

Similar to Theorem 6, parameter ϵ_0 in NewSamp is also determined by two terms. The first term reveals the the relationship between the target rank r and sample size. Without loss of generality, we can set $\beta = \lambda^{(t)}_{r+1}/4$. Then the sample size is linear in $1/[\lambda^{(t)}_{r+1}]^2$. Hence, a small r means that a small sample size is sufficient. Conversely, if we want to sample a small portion of f_i's, then we should choose a small r. Otherwise, β will go to $\lambda^{(t)}_{r+1}/2$ which means $\epsilon_0 = 1$, i.e., the sequence $\{x^{(t)}\}$ does not converge. The second term shows that a small sample size will lead to a poor convergence rate. If we set $r = 0$ and $\beta = \lambda_1/2$, then ϵ_0 will be $1 - \frac{1}{1+2\lambda_1/\sigma}$. Consequently, the convergence rate of NewSamp is almost the same as gradient descent. Similar to Algorithm 3, a small r means a precise solution to Problem (3) and the initial point $x^{(0)}$ being close to the optimal point x^*.

It is worth pointing out that Theorem 7 explains the empirical results that NewSamp is applicable in training SVM in which the Lipschitz continuity condition of $\nabla^2 F(x)$ is not satisfied (Erdogdu & Montanari, 2015).

We now conduct comparison between Theorem 6 and Theorem 7. We mainly focus on the parameter ϵ_0 in these two theorems which mainly determines convergence properties of Algorithm 3 and Algorithm 4. Specifically, if we set $\alpha = \beta + \lambda^{(t)}_{r+1}$ in Eqn. (11), then $\epsilon_0 = \frac{2\beta + \lambda^{(t)}_{r+1}}{\sigma + 2\beta + \lambda^{(t)}_{r+1}}$ which equals to the second term on the right-hand side in Eqn. (12). Hence, we can regard NewSamp as a special case of Algorithm 3. However, NewSamp provides an approach for automatical choice of α.

Recall that NewSamp includes another parameter: the target rank r. Thus, NewSamp and Algorithm 3 have the same number of free parameters. If r is not properly chosen, NewSamp will still have poor performance. Therefore, Algorithm 3 is theoretically preferred because NewSamp needs extra cost to perform SVDs.

6. Inexact Newton Methods

Let $H^{(t)} = \nabla^2 F(x^{(t)})$, that is, $\epsilon_0 = 0$. Then Theorem 3 depicts the convergence properties of inexact Newton methods.

Theorem 8 *Let* $F(x)$ *satisfy the properties described in Theorem 3, and* $p^{(t)}$ *be a direction vector such that*

$$\|\nabla F(x^{(t)}) - \nabla^2 F(x^{(t)}) p^{(t)}\| \leq \frac{\epsilon_1}{\kappa} \|\nabla F(x^{(t)})\|,$$

where $0 < \epsilon_1 < 1$. *Consider the iteration* $x^{(t+1)} = x^{(t)} - p^{(t)}$.

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, *and* $0 < \eta(t) < 1$ *such that when* $\|x^{(t)} - x^*\| \leq \gamma$, *then it holds that*

$$\|\nabla F(x^{(t+1)})\|_{M^*} \leq (\epsilon_1 + 2\eta(t)) \frac{1 + \nu(t)}{1 - \nu(t)} \|\nabla F(x^{(t)})\|_{M^*}.$$

(b) If $\nabla^2 F(x)$ *is also Lipschitz continuous with parameter* $\hat{L}$, *and* $\{x^{(t)}\}$ *satisfies Eqn.* (6), *then it holds that*

$$\|\nabla F(x^{(t+1)})\|_{M^*} \leq \epsilon_1 \frac{1 + \nu(t)}{1 - \nu(t)} \|\nabla F(x^{(t)})\|_{M^*} + \frac{2\hat{L}\kappa}{\mu\sqrt{\mu}} \frac{(1 + \nu(t))^2}{1 - \nu(t)} \|\nabla F(x^{(t)})\|^2_{M^*}.$$

7. Empirical Study

In this section, we validate our theoretical results about sketched size of the sketch Newton, and sample size of regularized Newton, experimentally. Experiments for validating unnecessity of the Lipschitz continuity condition of $\nabla^2 F(x)$ are given in the supplementary materials.

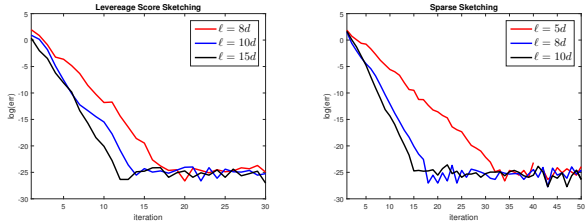
7.1. Sketched Size of Sketch Newton

Now we validate that our theoretical result that sketched size is independent of the condition number of the Hessian in Sketch Newton. To control the condition number of the Hessian conveniently, we conduct the experiment on least squares regression which is defined as

$$\min_x \frac{1}{2}\|Ax - b\|^2. \tag{13}$$

In each iteration, the Hessian matrix is $A^T A$. In our experiment, A is a 10000×54 matrix. We set the singular values σ_i of A as: $\sigma_i = 1.2^{-i}$. Then the condition number of A is $\kappa(A) = 1.2^{54} = 1.8741 \times 10^4$. We use different sketch matrices in Sketch Newton (Algorithm 1) and set different values of the sketched size ℓ. We report our empirical results in Figure 1.

From Figure 1, we can see that Sketch Newton performs well when the sketch size ℓ is several times of d for all different sketching matrices. Moreover, the corresponding algorithms converge linearly. This matches our theory that the sketched size is independent of the condition number of the Hessian matrix to achieve a linear convergence rate. In contrast, the theoretical result of (Xu et al., 2016) shows that the sketched size is $\ell = d * \kappa(A) = 1.02 \times 10^6$ bigger than $n = 10^4$.



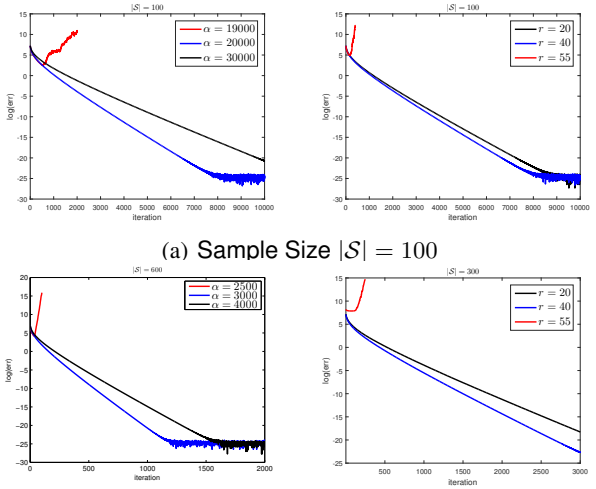
(a) Leverage Score Sampling. (b) Sparse Sketching.

Figure 1. Convergence properties of different sketched sizes

7.2. Sample Size of Regularized Subsampled Newton

We also choose least squares regression defined in Eqn. (13) in our experiment to validate the theory that adding a regularizer is an effective approach to reducing the sample size while keeping convergence in Subsampled Newton. Let $A \in \mathbb{R}^{n \times d}$ where $n = 8000$ and $d = 5000$. Hence Sketch Newton can not be used in this case because n and d are close to each other. In our experiment, we set different sample sizes $|\mathcal{S}|$. For each $|\mathcal{S}|$ we choose different regularizer terms α and different target ranks r. We report our results in Figure 2.

As we can see, if the sample size $|\mathcal{S}|$ is small, then we should choose a large α; otherwise, the algorithm will diverge. However, if the regularizer term α is too large, then the algorithm will converge slowly. Increasing the sample size and choosing a proper regularizer will improve convergence properties obviously. When $|\mathcal{S}| = 600$, it only needs about 1200 iterations to obtain a precise solution while it needs about 8000 iterations when $|\mathcal{S}| = 100$. Similarly, if the sample size $|\mathcal{S}|$ is small, then we should choose a small target rank. Otherwise NewSamp may diverge. Also, if the target rank is not chosen properly, NewSamp will have poor convergence properties. Furthermore, from Figure 2, we can see that the two algorithms have similar convergence properties. This validates the theoretical result that NewSamp provides a method to choose α automatically. Our empirical analysis matches the theoretical analysis in Subsection 5.2 very well.



(a) Sample Size $|\mathcal{S}| = 100$

(b) Sample Size $|\mathcal{S}| = 600$

Figure 2. Convergence properties of Regularized Subsampled Newton and NewSamp

8. Conclusion

In this paper, we have proposed a framework to analyze the local convergence properties of second order methods including stochastic and deterministic versions. This framework reveals some important convergence properties of the subsampled Newton method and sketch Newton method, which are unknown before. The most important thing is that our analysis lays the theoretical foundation of several important stochastic second order methods.

We believe that this framework might also provide some useful insights for developing new subsampled Newton-type algorithms. We would like to address this issue in future.

Acknowledgements

Ye has been supported by the National Natural Science Foundation of China (Grant No. 11426026, 61632017, 61173011) and a Project 985 grant of Shanghai Jiao Tong University. Luo and Zhang have been supported by he National Natural Science Foundation of China (No. 61572017), Natural Science Foundation of Shanghai City (No. 15ZR1424200), and Microsoft Research Asia Collaborative Research Award.

References

Agarwal, Naman, Bullins, Brian, and Hazan, Elad. Second order stochastic optimization in linear time. *arXiv preprint arXiv:1602.03943*, 2016.

Byrd, Richard H, Chin, Gillian M, Neveitt, Will, and Nocedal, Jorge. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

Clarkson, Kenneth L and Woodruff, David P. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 81–90. ACM, 2013.

Cotter, Andrew, Shamir, Ohad, Srebro, Nati, and Sridharan, Karthik. Better mini-batch algorithms via accelerated gradient methods. In *Advances in neural information processing systems*, pp. 1647–1655, 2011.

Erdogdu, Murat A and Montanari, Andrea. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems*, pp. 3034–3042, 2015.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Li, Mu, Zhang, Tong, Chen, Yuqiang, and Smola, Alexander J. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 661–670. ACM, 2014.

Meng, Xiangrui and Mahoney, Michael W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 91–100. ACM, 2013.

Nemirovski, Arkadi, Juditsky, Anatoli, Lan, Guanghui, and Shapiro, Alexander. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nocedal, Jorge and Wright, Stephen. *Numerical optimization*. Springer Science & Business Media, 2006.

Pilanci, Mert and Wainwright, Martin J. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Roosta-Khorasani, Farbod and Mahoney, Michael W. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016.

Roux, Nicolas L, Schmidt, Mark, and Bach, Francis R. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Schmidt, Mark, Roux, Nicolas Le, and Bach, Francis. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

Woodruff, David P. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Xu, Peng, Yang, Jiyan, Roosta-Khorasani, Farbod, Ré, Christopher, and Mahoney, Michael W. Sub-sampled newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pp. 3000–3008, 2016.

Zhang, Lijun, Mahdavi, Mehrdad, and Jin, Rong. Linear convergence with condition number independent access of full gradients. In *Advance in Neural Information Processing Systems 26 (NIPS)*, pp. 980–988, 2013.