Journal of Machine Learning Research 22 (2021) 1-41 Submitted 10/19; Revised 1/21; Published 3/21

Approximate Newton Methods

Haishan Ye YEHAISHAN@XJTU.EDU.CN
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University
Xi'an, China

Luo Luo LUOLUO@UST.HK
Department of Mathematics
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Zhihua Zhang* ZHZHANG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing, China

Editor: Qiang Liu

Abstract

Many machine learning models involve solving optimization problems. Thus, it is important to address a large-scale optimization problem in big data applications. Recently, subsampled Newton methods have emerged to attract much attention 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. Other efficient stochastic second order methods have been also proposed. 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 empirical performance in real applications. In this paper, we aim to fill these gaps. We propose a unifying framework to analyze both local and global convergence properties of second order methods. Accordingly, we present our theoretical results which match the empirical performance in real applications well.

Keywords: Approximate Newton, Stochastic Second-order, Hessian Approximation

1. Introduction

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

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

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 regression, smoothed support vector machines, neural networks, and graphical models.

$*$. Corresponding author.

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

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

where $s_t > 0$ 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 and Monro, 1951; Li et al., 2014; Cotter et al., 2011). 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, many efforts have been made to devise modification to achieve the convergence rate of the full gradient while keeping low iteration cost (Johnson and Zhang, 2013; Roux et al., 2012; Schmidt et al., 2017; Zhang et al., 2013).

If Q_t is a $d \times d$ positive definite matrix of 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 $s_t = 1$, 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 such that the update is computationally feasible while keeping sufficient second order information. One class of such methods is quasi-Newton methods, which are generalization 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 and Wright, 2006). They take $\mathcal{O}(nd + d^2)$ cost per iteration.

Recently, when $n \gg d$, a class of 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 and Mahoney, 2019; Byrd et al., 2011; Xu et al., 2016) to construct a subsampled Hessian. Erdogdu and 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 and Wainwright (2017) used sketching techniques to approximate the Hessian and proposed *sketch Newton* method. Similarly, Xu et al. (2016) proposed to sample rows of $B(x)$ with non-uniform probability distribution. Agarwal et al. (2017) brought up an algorithm called LiSSA to approximate the inverse of Hessian directly.

Although the convergence performance of stochastic second order methods has been analyzed, the convergence properties are still not well understood. Several important gaps are lying between the convergence theory and the practical performance of these algorithms in real applications.

First, it is about the necessity of Lipschitz continuity of the 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 some scenarios without this assumption, they might also converge to an optimal point. For example, Erdogdu and Montanari (2015) used NewSamp to successfully train the smoothed-SVM in which the ℓ_2-hinge loss is used, so the corresponding Hessian is not Lipschitz continuous.

Second, it involves the sketching size of sketch Newton methods. To obtain a linear convergence, the sketching size is $\mathcal{O}(d\kappa^2)$ in Pilanci and Wainwright (2017) and then improved to $\mathcal{O}(d\kappa)$ in Xu et al. (2016), where κ is the condition number of the Hessian matrix in question. Recently, Wang et al. (2018) extended the sketch Newton method to distributed optimization and achieved fast convergence rate and low communication cost. However, the sketch Newton empirically performs well even when the Hessian matrix is ill-conditioned. Sketching 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 and Wainwright (2017); Xu et al. (2016); Wang et al. (2018) implies that sketching size may be beyond n in ill-condition cases.

Third, it talks about the sample size in regularized subsampled Newton methods. In both Erdogdu and Montanari (2015) and Roosta-Khorasani and 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 does not agree with the extant theoretical analysis (Erdogdu and Montanari, 2015; Roosta-Khorasani and 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*. Accordingly, we propose a general result for analysis of both local and global convergence properties of second order methods. Based on this framework, we then give a detailed theoretical analysis that matches the empirical performance. We summarize our contribution as follows:

- We propose a unifying framework (Theorem 3 and Theorem 5) to analyze local and global 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 and Montanari, 2015) can be used to train the 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 sketching 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.

- Based on our analysis framework, we provide a much tighter bound of the sample size of subsampled Newton methods. To our best knowledge, it is the tightest bound of subsampled Newton methods in the extant results.

- We provide a theoretical guarantee for that adding a regularizer is an effective way to reduce sample size in subsampled Newton methods while keeping convergence. Our theoretical analysis also shows that adding a regularizer will lead to poor convergence behavior as the sample size decreases.

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 and global convergence analysis of second order methods. In Section 4 we analyze the convergence properties of sketch Newton methods and prove that sketching size is independent of the condition number of the Hessian matrix. In Section 5 we give the convergence behaviors of several variants of subsampled Newton methods. Especially, we reveal the relationship among sample size, regularizer, and convergence rate. In Section 6, we validate our theoretical results experimentally. Finally, we conclude our work in Section 7. The proofs of theorems are given in the appendices.

2. Notation and Preliminaries

Section 2.1 defines the notation used in this paper. Section 2.2 introduces matrix sketching techniques and their properties. Section 2.3 describes some important assumptions about objective functions.

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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\ell > 0$ contains the nonzero singular values of A. We let $\sigma_{\max}(A)$ denote the largest singular value and $\sigma_{\min}(A)$ 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 symmetric positive semi-definite, 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, $\|A\|_F \triangleq (\sum_{i,j} a_{ij}^2)^{1/2} = (\sum_i \sigma_i^2)^{1/2}$ is the Frobenius norm of A and $\|A\| \triangleq \sigma_1$ is the spectral norm. Given a symmetric positive definite matrix M, $\|x\|_M \triangleq \|M^{1/2}x\|$ is called the M-norm of x. Given 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_0-subspace embedding property which will be used to sketch Hessian matrices. Then we list some useful types of randomized sketching matrices including Gaussian projection (Halko et al., 2011; Johnson and Lindenstrauss, 1984), leverage score sampling (Drineas et al., 2006), count sketch (Clarkson and Woodruff, 2013; Nelson and Nguyên, 2013; Meng and Mahoney, 2013).

Definition 1 *$S \in \mathbb{R}^{\ell\times n}$ is said to be an ϵ_0-subspace embedding matrix w.r.t. a fixed matrix $A \in \mathbb{R}^{n\times d}$ where $d < n$, if $\|SAx\|^2 = (1 \pm \epsilon_0)\|Ax\|^2$ (i.e., $(1-\epsilon_0)\|Ax\|^2 \leq \|SAx\|^2 \leq (1+\epsilon_0)\|Ax\|^2$) for all $x \in \mathbb{R}^d$.*

From the definition of the ϵ_0-subspace embedding matrix, we can derive the following property directly.

Lemma 2 *$S \in \mathbb{R}^{\ell\times n}$ is an ϵ_0-subspace embedding matrix w.r.t. the matrix $A \in \mathbb{R}^{n\times d}$ if and only if*

$$(1-\epsilon_0)A^TA \preceq A^TS^TSA \preceq (1+\epsilon_0)A^TA.$$

Gaussian sketching matrix. The most classical sketching matrix is the Gaussian sketching matrix $S \in \mathbb{R}^{\ell \times n}$, whose entries are i.i.d. from the normal of mean 0 and variance $1/\ell$. Owing to the well-known concentration properties (Woodruff, 2014), Gaussian random matrices are very attractive. Note that $\ell = \mathcal{O}(d/\epsilon_0^2)$ is enough to guarantee the ϵ_0-subspace embedding property for any fixed matrix $A \in \mathbb{R}^{n \times d}$. Moreover, $\ell = \mathcal{O}(d/\epsilon_0^2)$ is the tightest bound among known types of sketching matrices. However, the Gaussian random matrix is usually dense, so it is costly to compute SA.

Leverage score sampling matrix. A leverage score sampling matrix $S = D\Omega \in \mathbb{R}^{\ell \times n}$ w.r.t. $A \in \mathbb{R}^{n \times d}$ is defined by sampling probabilities p_i, a sampling matrix $\Omega \in \mathbb{R}^{\ell \times n}$ and a diagonal rescaling matrix $D \in \mathbb{R}^{\ell \times \ell}$. Specifically, we construct S as follows. For every $j = 1, \ldots, \ell$, independently and with replacement pick an index i from the set $\{1, 2 \ldots, n\}$ with probability p_i, and set $\Omega_{ji} = 1$ and $\Omega_{jk} = 0$ for $k \neq i$ as well as $D_{jj} = 1/\sqrt{p_i \ell}$. The sampling probabilities p_i are the leverage scores of A defined as follows. Let $V \in \mathbb{R}^{n \times d}$ be the column orthonormal basis of A, and let $v_{i,*}$ denote the i-th row of V. Then $q_i \triangleq \|v_{i,*}\|^2/d$ for $i = 1, \ldots, n$ are the leverage scores of A. To achieve an ϵ_0-subspace embedding property w.r.t. A, $\ell = \mathcal{O}(d \log d/\epsilon_0^2)$ is sufficient.

Sparse embedding matrix. A sparse embedding matrix $S \in \mathbb{R}^{\ell \times n}$ is such a matrix in each column of which there is only one nonzero entry uniformly sampled from $\{1, -1\}$ (Clarkson and Woodruff, 2013). Hence, it is very efficient to compute SA, especially when A is sparse. To achieve an ϵ_0-subspace embedding property w.r.t. $A \in \mathbb{R}^{n \times d}$, $\ell = \mathcal{O}(d^2/\epsilon_0^2)$ is sufficient (Meng and Mahoney, 2013; Woodruff, 2014).

Other sketching matrices such as Subsampled Randomized Hadamard Transformation (Drineas et al., 2012; Halko et al., 2011) 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. We define the condition number of the objective function of $F(x)$ as

$$\kappa \triangleq \frac{L}{\mu}.$$

Note that κ is an upper bound of the condition number of the Hessian matrix $\nabla^2 F(x)$ for any x. Furthermore, if $\nabla^2 F(x)$ is Lipschitz continuous, then we have

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

Algorithm 1 Approximate Newton.

1: **Input:** $x^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_0 < 1$;
2: **for** $t = 0, 1, \ldots$ until termination **do**
3: Construct an approximate Hessian $H^{(t)}$ satisfying Condition (2);
4: Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$;
5: Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
6: **end for**

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 $M = \nabla^2 F(x^*)$ 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 $0 < \rho < 1$,

$$\limsup_{t \to \infty} \frac{\|x^{(t+1)} - x^*\|_M}{\|x^{(t)} - x^*\|_M} = \rho.$$

Similarly, superlinear convergence and quadratic convergence are respectively defined as

$$\limsup_{t \to \infty} \frac{\|x^{(t+1)} - x^*\|_M}{\|x^{(t)} - x^*\|_M} = 0, \qquad \limsup_{t \to \infty} \frac{\|x^{(t+1)} - x^*\|_M}{\|x^{(t)} - x^*\|_M^2} = \rho.$$

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

$$\|x^{(t+1)} - x^*\|_M \leq \rho_1 \|x^{(t)} - x^*\|_M + \rho_2 \|x^{(t)} - x^*\|_M^2,$$

where $0 < \rho_1 < 1$ and $0 \leq \rho_2$.

3. Main Results

The existing variants of stochastic second order methods share some important attributes. First, these methods such as NewSamp (Erdogdu and Montanari, 2015), LiSSA (Agarwal et al., 2017), 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

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

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

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

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. The detailed algorithmic description is listed in Algorithm 1.

3.1 Local Convergence Analysis

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

Theorem 3 *Let Assumptions 1 and 2 hold. Suppose that $\nabla^2 F(x)$ 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*

$$\|\nabla F(x^{(t)}) - H^{(t)} p^{(t)}\| \leq \frac{\epsilon_1}{\kappa^{3/2}} \|\nabla F(x^{(t)})\|, \tag{4}$$

where $0 < \epsilon_1 < 1$. Then Algorithm 1 has the following convergence properties.

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

$$\left\|x^{(t+1)} - x^*\right\|_M \leq \left(\epsilon_0 + \epsilon_1 + 2\nu\mu^{-1} + \frac{2\kappa\mu^{-1}\nu(1+\nu\mu^{-1})}{1-\kappa\mu^{-1}\nu}\right) \left\|x^{(t)} - x^*\right\|_M. \tag{5}$$

Moreover, ν will go to 0 as $x^{(t)}$ goes to x^*.

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

$$\left\|x^{(t)} - x^*\right\|_M \leq \frac{1}{2}\mu^{\frac{3}{2}}\kappa^{-1}\hat{L}^{-1}, \tag{6}$$

then it holds that

$$\left\|x^{(t+1)} - x^*\right\|_M \leq (\epsilon_0 + \epsilon_1)\left\|x^{(t)} - x^*\right\|_M + 4(\kappa+1)\mu^{-3/2}\hat{L}\left\|x^{(t)} - x^*\right\|_M^2. \tag{7}$$

Remark 4 *We give the convergence properties of the sequence $\left\{\left\|x^{(t)} - x^*\right\|_M\right\}$ in contrast to $\left\{\left\|x^{(t)} - x^*\right\|\right\}$ used in works (Erdogdu and Montanari, 2015; Pilanci and Wainwright, 2017; Roosta-Khorasani and Mahoney, 2019). Due to this difference, our unifying framework provides the foundation of many stronger theoretical results of approximate Newton methods which we will discuss in the next sections. For example, let us consider the case that $F(x)$ is quadratic which implies $\hat{L} = 0$ and the Hessian is $M = \nabla^2 F(x) = A^T A$ with $A \in \mathbb{R}^{n\times d}$. By setting $\epsilon_1 = 0$, then Eqn.* (7) *reduces to $\left\|x^{(t+1)} - x^*\right\|_M \leq \epsilon_0 \left\|x^{(t)} - x^*\right\|_M$. To achieve a linear convergence rate with $0 < \epsilon_0 < 1$, we can construct an approximate Hessian $H^{(t)} = A^T S^T S A$ with $S \in \mathbb{R}^{\ell\times n}$ and $\ell = \mathcal{O}(d/\epsilon_0^2)$ for Gaussian sketching matrix. In contrast, Eqn. (14) of Pilanci and Wainwright (2017) reduces to $\left\|x^{(t+1)} - x^*\right\| \leq \epsilon_0'\kappa \left\|x^{(t)} - x^*\right\|$. To achieve a linear convergence rate of $0 < \epsilon_0 < 1$, that is, $\epsilon_0'\kappa \leq \epsilon_0$, then it requires $\epsilon_0' \leq \epsilon_0\kappa^{-1}$. Combining with the properties of the Gaussian sketching matrix, the sketching size ℓ is required to be $\ell = \mathcal{O}(d\kappa^2/\epsilon_0^2)$. We can observe the sketching size obtained by our theory is κ^2 smaller than the one derived in Pilanci and Wainwright (2017).*

From Theorem 3, we can find some important insights. First, Theorem 3 provides sufficient conditions to get different convergence rates including linear, super-liner, and quadratic rates. If $(\epsilon_0 + \epsilon_1)$ is a constant less than 1, then sequence $\{x^{(t)}\}$ converges linearly with rate $(\epsilon_0 + \epsilon_1)$. Furthermore, 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 super-linearly. If $\nabla^2 F(x)$ is $\hat{L}$-Lipschitz and $\epsilon_0 + \epsilon_1 = 0$, then

Algorithm 2 Approximate Newton with backtracking line search.

Input: $x^{(0)}$, $0 < \alpha < 0.5$, $0 < \beta < 1$;
for $t = 0, 1, \ldots$ until termination **do**
  Construct an approximate Hessian $H^{(t)}$ satisfying Condition (2);
  Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$;
  Line search:
  while $F(x^{(t)} + sp^{(t)}) > F(x^{(t)}) + \alpha s [\nabla F(x^{(t)})]^T p^{(t)}$ **do**
    $s = \beta s$
  end while
  Update $x^{(t+1)} = x^{(t)} - sp^{(t)}$;
end for

$\{x^{(t)}\}$ converges quadratically. In this case, approximate Newton reduces to the exact Newton method.

Second, Theorem 3 makes it clear that the Lipschitz continuity of the Hessian is *not necessary* for linear convergence and super-linear 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 (Erdogdu and Montanari, 2015). 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., 2017), NewSamp (Erdogdu and Montanari, 2015), Subsampled Newton with non-uniformly sampling (Xu et al., 2016), and Sketched Newton (Pilanci and Wainwright, 2017) 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 $\nabla^2 F(x)$ is not necessary to achieve a linear or superlinear convergence rate for inexact Newton methods. However, our work first proves this result still holds for the approximate Hessian.

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. We can observe that by different choices of $H^{(t)}$ and different ways to calculate $p^{(t)}$, one can obtain different kinds of second order methods.

3.2 Global Convergence Analysis

In the previous analysis, the theory is local and approximate Newton can achieve a fast convergence rate once the iterations enter a suitable basin of the optimal point. In this section, we are going to obtain global convergence results for *self-concordant* functions. The self-concordant assumption is widely studied in the global convergence analysis of Newton methods (Pilanci and Wainwright, 2017; Boyd and Vandenberghe, 2004).

Note that a closed, convex function F: $\mathbb{R}^d \to \mathbb{R}$ is called self-concordant if

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

for all x in the domain of $F(x)$ and $v \in \mathbb{R}^d$. Here $\|v\|_x = (v^T \nabla^2 F(x) v)^{1/2}$ is the local norm.

To achieve a global convergence, the approximate Newton method should combine with the line search. At the damped phase where $[\nabla F(x^{(t)})]^T p^{(t)}$ is large, a line search is applied to guarantee the

Algorithm 3 Sketch Newton.

1: **Input:** $x^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_0 < 1$;
2: **for** $t = 0, 1, \ldots$ until termination **do**
3: 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 $H^{(t)} = [B(x^{(t)})]^T S^T S B(x^{(t)})$;
4: Calculate $p^{(t)} \approx \operatorname{argmin}_p \frac{1}{2} p^T H^{(t)} p - p^T \nabla F(x^{(t)})$;
5: Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
6: **end for**

convergence of approximate Newton methods. Once $[\nabla F(x^{(t)})]^T p^{(t)}$ is sufficient small, then step size $s = 1$ can keep approximate Newton converging with a linear rate. The detailed algorithmic description of approximate Newton with backtracking line search is presented in Algorithm 2.

In the following theorem, we provide the iteration complexity of Algorithm 2 to achieve an ϵ-suboptimality.

Theorem 5 *Assume the objective function $F(x)$ is self-concordant, and $H^{(t)}$ is a positive definite matrix satisfying Eqn.* (2) *with $0 \leq \epsilon_0 < 1$. Let $p^{(t)}$ be a descent direction satisfying Eqn.* (4)*. To achieve $F(x^{(T)}) - F(x^*) \leq \epsilon$, the iteration complexity of the approximate Newton method with backtracking line search (Algorithm 2) is at most*

$$T = \frac{F(x^{(0)}) - F(x^*)}{\eta} + \frac{2}{1 - \epsilon_0 - 2\epsilon_1 \kappa^{-1}} \log \left(\frac{1 - \epsilon_0 - 2\epsilon_1 \kappa^{-1}}{12\epsilon} \right), \tag{8}$$

where η is defined as

$$\eta = \alpha\beta \frac{(1-\epsilon_0)\rho^2 (1 - \epsilon_0 - 2\epsilon_1\kappa^{-1})^2}{144 + 12\rho\sqrt{(1-\epsilon_0)}(1-\epsilon_0 - 2\epsilon_1\kappa^{-1})}, \text{ with } \quad \rho = \frac{\left(1 - \epsilon_1 \kappa^{-1} \cdot \left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)^{1/2}}{(1+\epsilon_0)^{1/2} \left(1 + \epsilon_1 \kappa^{-1} \cdot \left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)}.$$

Remark 6 *In the above theorem, the iteration complexity of approximate Newton with line search still depends on the condition number of the objective function even it is self-concordant. This dependency on the condition number is caused by the approximation to $H^{-1}\nabla F(x)$. If $\epsilon_1 = 0$ in Eqn.* (4)*, we can obtain that $\eta = \alpha\beta \frac{(1-\epsilon_0)^3}{144(1+\epsilon_0) + 12(1+\epsilon_0)^{1/2}(1-\epsilon_0)^{3/2}}$ which is independent on the condition number. Thus, the total complexity is independent on the condition number.*

4. The Sketch Newton Method

In this section, we use Theorem 3 to analyze the convergence properties of Sketch Newton which utilizes the sketching technique to approximate the Hessian which satisfies the form

$$\nabla^2 F(x) = B(x)^T B(x), \tag{9}$$

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.

Table 1: Comparison with previous work with leverage score sampling matrix.

Reference	Sketching size	Condition number free?
Pilanci and Wainwright (2017)	$\mathcal{O}\left(\frac{d\kappa^2 \log d}{\epsilon_0^2}\right)$	No
Xu et al. (2016)	$\mathcal{O}\left(\frac{d\kappa \log d}{\epsilon_0^2}\right)$	No
Our result (Theorem 7)	$\mathcal{O}\left(\frac{d \log d}{\epsilon_0^2}\right)$	Yes

The Sketch Newton method constructs the approximate Hessian matrix as follows:

$$H^{(t)} = [S^{(t)}B(x)]^T S^{(t)}B(x) \tag{10}$$

where $S^{(t)} \in \mathbb{R}^{\ell \times n}$ is a randomized sketching matrix. The approximate Newton method with such Hessian approximation is referred to as the Sketch Newton method. The detailed algorithmic description is given in Algorithm 3.

Theorem 7 *Let $F(x)$ satisfy the conditions described in Theorem 3. Assume the Hessian matrix is given as Eqn.* (9)*. Let $0 < \delta < 1$, $0 < \epsilon_0 < 1/2$ and $0 \leq \epsilon_1 < 1$ be given. $S^{(t)} \in \mathbb{R}^{\ell \times n}$ is an ϵ_0-subspace embedding matrix w.r.t. $B(x)$ with probability at least $1 - \delta$. Then sketch Newton (Algorithm 3) has the following convergence properties:*

(a) There exists a sufficient small value γ and $\nu = o(1)$ such that when $\|x^{(t)} - x^*\|_M \leq \gamma$, 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 $1 - \delta$.*

(c) If $F(x)$ is self-concordant, the iteration complexity of the sketch Newton with backtracking line search (Algorithm 2 with $H^{(t)}$ constructed as Eqn. (10)*) is upper bounded as Eqn.* (8)*.*

Theorem 7 directly provides a bound of the sketching size. Using the leverage score sampling matrix as an example, the sketching size $\ell = \mathcal{O}(d \log d/\epsilon_0^2)$ is sufficient. We compare our theoretical bound of the sketching size with the ones of Pilanci and Wainwright (2017) and Xu et al. (2016). We present Eqn. (14) of Pilanci and Wainwright (2017) as follows

$$\left\|x^{(t+1)} - x^*\right\| \leq \epsilon_0' \kappa \cdot \left\|x^{(t)} - x^*\right\| + \frac{4\hat{L}}{\mu}\left\|x^{(t)} - x^*\right\|^2.$$

To achieve a linear convergence with rate ϵ_0, it requires that $\epsilon_0'\kappa \leq \epsilon_0$, that is, $\epsilon_0' \leq \epsilon_0 \kappa^{-1}$. Combining with the properties of leverage score sampling, we have that the sketching size is required as $\ell = \mathcal{O}(d \log d/(\epsilon_0')^2) = \mathcal{O}(d\kappa^2 \log d/\epsilon_0^2)$. Similarly, we rewrite Eqn. (5) in Lemma 2 of Xu et al. (2016) as follows

$$\left\|x^{(t+1)} - x^*\right\| \leq \frac{3\epsilon_0'\sqrt{\kappa}}{1 - \epsilon_0'} \cdot \left\|x^{(t)} - x^*\right\| + \frac{2\hat{L}}{(1 - \epsilon_0')\mu}\left\|x^{(t)} - x^*\right\|^2.$$

To achieve a linear convergence with rate ϵ_0, it requires that $\frac{3\epsilon_0'\sqrt{\kappa}}{1-\epsilon_0'} \leq \epsilon_0$, that is, $\epsilon_0' \leq \epsilon_0/(3\sqrt{\kappa}+\epsilon_0)$. Combining with the properties of leverage score sampling, we obtain that the sketching size is required as $\ell = \mathcal{O}(d\log d/(\epsilon_0')^2) = \mathcal{O}(d\kappa \log d/\epsilon_0^2)$. We list the detailed comparison in Table 1.

As we can see, our sketching size is much smaller than the other two, especially when the Hessian matrix is ill-conditioned. Theorem 7 shows that the sketching 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 sketching size is only several times of d.

Furthermore, the iteration complexity of the sketch Newton with backtracking line search shares the similar result with that of Pilanci and Wainwright (2017). Especially when $\epsilon_1 = 0$, Eqn. (8) reduces to

$$T = \frac{F(x^{(0)}) - F(x^*)}{\eta} + 4\log\left(\frac{1}{24\epsilon}\right), \quad \text{with} \quad \eta = \frac{\alpha\beta(1-\epsilon_0)^3}{12\left[12(1+\epsilon_0)+(1+\epsilon_0)^{1/2}(1-\epsilon_0)^{3/2}\right]}.$$

We observe that T is independent on the condition number of the objective function. A similar result can be found in Theorem 2 of Pilanci and Wainwright (2017).

Theorem 7 also contains the possibility of achieving an asymptotically super-linear rate by using an iteration-dependent sketching accuracy $\epsilon_0 = \epsilon_0(t)$. In particular, we present the following corollary.

Corollary 8 *Assume $F(x)$ satisfies the properties described in Theorem 3. Consider the approximate Hessian $H^{(t)}$ constructed in Eqn.* (10) *with the iteration-dependent sketching accuracy given as $\epsilon_0(t) = \frac{1}{\log(1+t)}$, and $p^{(t)} = [H^{(t)}]^{-1}\nabla F(x)$. If the initial point $x^{(0)}$ is close enough to the optimal point x^*, then sequence $\{x^{(t)}\}$ of the Sketch Newton (Algorithm 1 with $H^{(t)}$ constructed in Eqn.* (10)*) converges superlinearly.*

5. The Subsampled Newton method and Variants

In this section we apply Theorem 3 to analyze subsampled Newton methods. Without the assumption that the Hessian can be presented as Eqn. (9), for subsampled Newton methods, we assume that the Hessian is the sum of individual Hessians:

$$\nabla^2 F(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(x), \quad \text{with} \quad \nabla^2 f_i(x) \in \mathbb{R}^{d\times d}. \tag{11}$$

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

$$\max_{1\leq i\leq n} \|\nabla^2 f_i(x)\| \leq K < \infty, \tag{12}$$

$$\lambda_{\min}(\nabla^2 F(x)) \geq \mu > 0. \tag{13}$$

Accordingly, we can define a new type of condition number $\hat{\kappa} = \frac{K}{\mu}$.

In our earlier version (Ye et al., 2017), we apply Theorem 3 to analyze the convergence properties of subsampled Newton methods. However, the sample sizes obtained in the earlier version is no better than the ones obtained in (Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2019). In this version, we give a much tighter bound on the sample sizes of subsampled Newton method and its variants. This improvement is because we use the matrix Chernoff inequality to bound the sample size instead of the matrix Bernstein inequality.

Algorithm 4 Subsampled Newton (SSN).

1: **Input:** $x^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_0 < 1$;
2: Set the sample size $|\mathcal{S}| \geq \frac{3K/\mu \log(2d/\delta)}{\epsilon_0^2}$.
3: **for** $t = 0, 1, \ldots$ until termination **do**
4: Select a sample set $\mathcal{S}$ of size $|\mathcal{S}|$, and $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$;
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**

5.1 The Subsampled Newton Method

The Subsampled Newton method is depicted in Algorithm 4 and the approximate Hessian is constructed by sampling:

$$H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)}), \text{ with } j \text{ being uniformly sampled from } 1, \ldots, n. \tag{14}$$

First, by the properties of uniform sampling and matrix Chernoff inequality. The approximate Hessian in Eqn. (14) has the following properties.

Lemma 9 *Assume Eqn.* (12) *and Eqn.* (13) *hold, and let* $0 < \delta < 1$ *and* $0 < \epsilon_0 < 1/2$ *be given. The sample size* $|\mathcal{S}|$ *satisfies* $|\mathcal{S}| \geq \frac{3K/\mu \log(2d/\delta)}{\epsilon_0^2}$. *With probability at* $1 - \delta$, *the approximate Hessian* $H^{(t)}$ *is constructed in Eqn.* (14) *satisfies*

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

Above lemma gives the upper bound of sample size to guarantee the approximate Hessian in Eqn. (14) to satisfy Eqn. (2). Since Eqn. (2) is satisfied, the local and global convergence properties of the Subsampled Newton can be derived by Theorem 3 and Theorem 5. We give its local and global convergence properties in the following theorem.

Theorem 10 *Let* $F(x)$ *satisfy the properties described in Theorem 3. Assume Eqn.* (12) *and Eqn.* (13) *hold, and let* $0 < \delta < 1$, $0 < \epsilon_0 < 1/2$ *and* $0 \leq \epsilon_1 < 1$ *be given. The sample size* $|\mathcal{S}|$ *satisfies* $|\mathcal{S}| \geq \frac{3K/\mu \log(2d/\delta)}{\epsilon_0^2}$. *The approximate Hessian* $H^{(t)}$ *is constructed in Eqn.* (14)*, and the direction vector* $p^{(t)}$ *satisfies Eqn.* (4). *Then for* $t = 1, \ldots, T$, *Algorithm 4 has the following convergence properties:*

(a) There exists a sufficient small value γ *and* $\nu = o(1)$ *such that when* $\|x^{(t)} - x^*\|_M \leq \gamma$, *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* $1 - \delta$.

(c) If $F(x)$ *is self-concordant, the iteration complexity of the sketch Newton with backtracking line search (Algorithm 2 with* $H^{(t)}$ *constructed in Eqn.* (14)*) is upper bounded by Eqn.* (8).

Algorithm 5 Regularized Subsample Newton (RegSSN).

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

Remark 11 *We see that subsampled Newton with the sample size $\mathcal{O}\left(K/\mu \cdot \epsilon_0^{-2}\right)$ can achieve a linear convergence rate ϵ_0. In contrast, previous works require sample sizes at least to be $\mathcal{O}\left(K^2/\mu^2 \cdot \epsilon_0^{-2}\right)$ (Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2019). This difference comes from the way to derive the precision of the approximate Hessian in Lemma 9. First, previous works have to bound the approximation error as $\left\|\nabla^2 F(x^{(t)}) - H^{(t)}\right\| \leq \epsilon_0 \mu$ by the matrix Bernstein inequality. This will lead to the sample size bound $\mathcal{O}\left(K^2/\mu^2 \cdot \epsilon_0^{-2}\right)$ (Lemma 2 of Roosta-Khorasani and Mahoney (2019)). In contrast, our method only requires Condition* (2)*. Thus, we can use the matrix Chernoff inequality (Lemma 17) which only depends on K/μ linearly. Detailed proof of Theorem 10 can be found in Appendix E. Similarly, we can obtain tighter bounds of sample size for other variants of subsampled Newton methods.*

As we can see, Algorithm 4 almost has the same convergence properties as Algorithm 3 except several minor differences. The main difference is the construction manner of $H^{(t)}$ which should satisfy Eqn. (2). Algorithm 4 relies on the assumption that each $\|\nabla^2 f_i(x)\|$ is upper bounded (i.e., Eqn. (12) holds), while Algorithm 3 is built on the setting of the Hessian matrix as in Eqn. (9).

5.2 Regularized Subsampled Newton

In the ill-conditioned case (i.e., $\hat{\kappa} = \frac{K}{\mu}$ is large), the subsampled Newton method in Algorithm 4 should take a lot of samples because the sample size $|\mathcal{S}|$ depends on $\frac{K}{\mu}$ linearly. To overcome this problem, one resorts to a regularized subsampled Newton method which adds a regularizer to the original subsampled Hessian:

$$H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)}) + \xi \cdot I \tag{15}$$

where $\xi > 0$ is the regularization parameter. The detailed algorithmic procedure of the regularized subsampled Newton is described in Algorithm 5. In the following analysis, we prove that adding a regularizer is an effective way to reduce the sample size while keeping convergence in theory. First, we prove that the sample size $|\mathcal{S}|$ is properly chosen, the approximate Hessian in Eqn. (15) satisfies condition (2).

Lemma 12 *Assume Eqn.* (12) *and* (13) *hold, and let $0 < \delta < 1$, and $0 < \xi$ be given. Assume the sample size $|\mathcal{S}|$ satisfies $|\mathcal{S}| = \frac{18(K+\xi)\log(2d/\delta)}{\mu+\xi}$, and $H^{(t)}$ is constructed as in Algorithm 5. With probability at $1-\delta$, the approximate Hessian $H^{(t)}$ is constructed in Eqn.* (15) *satisfies*

$$(1-\epsilon_0)H^{(t)} \preceq \nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0)H^{(t)}, \text{ with } \epsilon_0 = \max\left(\frac{3\xi+\mu}{3\xi+3\mu}, \frac{L-2\xi}{2(L+\xi)}\right).$$

Combining above lemma with Theorem 3 and Theorem 5, we can obtain the local and global convergence properties of Algorithm 5 in the following theorem.

Theorem 13 *Let $F(x)$ satisfy the properties described in Theorem 3. Assume Eqn.* (12) *and* (13) *hold, and let $0 < \delta < 1$, $0 \leq \epsilon_1 < 1$ and $0 < \xi$ be given. Assume the sample size $|\mathcal{S}|$ satisfies $|\mathcal{S}| = \frac{18(K+\xi)\log(2d/\delta)}{\mu+\xi}$, and $H^{(t)}$ is constructed as in Algorithm 5. Define*

$$\epsilon_0 = \max\left(\frac{3\xi+\mu}{3\xi+3\mu}, \frac{L-2\xi}{2(L+\xi)}\right), \tag{16}$$

which implies that $0 < \epsilon_0 < 1$. Moreover, the direction vector $p^{(t)}$ satisfies Eqn. (4)*. Then Algorithm 5 has the following convergence properties:*

(a) There exists a sufficient small value γ and $\nu = o(1)$ such that when $\|x^{(t)} - x^*\|_M \leq \gamma$, 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 $1 - \delta$.*

(c) If $F(x)$ is self-concordant, the iteration complexity of the sketch Newton with backtracking line search (Algorithm 2 with $H^{(t)}$ constructed in Eqn. (15)*) is upper bounded by Eqn.* (8)*.*

In Theorem 13 the parameter ϵ_0 mainly decides convergence properties of Algorithm 5. It is determined by two terms just as shown in Eqn. (16). These two terms depict the relationship among the sample size, regularizer $\xi \cdot I$, and convergence rate.

We can observe that the sample size $|\mathcal{S}| = \frac{18(K+\xi)\log(2d/\delta)}{\mu+\xi}$ decreases as ξ increases. Hence Theorem 13 gives a theoretical guarantee that adding the regularizer $\xi \cdot I$ is an effective approach for reducing the sample size when K/μ is large. Conversely, if we want to sample a small part of f_i's, we should choose a large ξ.

Although a large ξ can reduce the sample size, it is at the expense of a slower convergence rate. As we can see, $\frac{3\xi+\mu}{3\xi+3\mu}$ goes to 1 as ξ increases. At the same time, ϵ_1 also has to decrease. Otherwise, $\epsilon_0 + \epsilon_1$ may be beyond 1 which means that Algorithm 5 will not converge.

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

5.3 NewSamp

Erdogdu and Montanari (2015) proposed NewSamp which is another regularized subsampled Newton method. NewSamp constructs its approximate Hessian as follows:

$$H^{(t)} = H^{(t)}_{|\mathcal{S}|} + U_{\backslash r}(\hat{\lambda}^{(t)}_{r+1} I - \hat{\Lambda}_{\backslash r})U^T_{\backslash r}, \tag{17}$$

where

$$H^{(t)}_{|\mathcal{S}|} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)}),$$

Algorithm 6 NewSamp.

1: **Input:** $x^{(0)}$, $0 < \delta < 1$, r, sample size $|\mathcal{S}|$;
2: **for** $t = 0, 1, \ldots$ until termination **do**
3: Select a sample set $\mathcal{S}$ of size $|\mathcal{S}|$, and get $H_{|\mathcal{S}|}^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$;
4: Compute rank $r+1$ truncated SVD deompostion of $H_{|\mathcal{S}|}^{(t)}$ to get U_{r+1} and $\hat{\Lambda}_{r+1}$. Construct $H^{(t)} = H_{|\mathcal{S}|}^{(t)} + U_{\backslash r}(\hat{\lambda}_{r+1}^{(t)} I - \hat{\Lambda}_{\backslash r})U_{\backslash r}^T$
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**

and its SVD decomposition is

$$H_{|\mathcal{S}|}^{(t)} = U\hat{\Lambda}U^T = U_r\hat{\Lambda}_r U_r^T + U_{\backslash r}\hat{\Lambda}_{\backslash r}U_{\backslash r}^T.$$

The detailed algorithm is depicted in Algorithm 6. $H^{(t)}$ constructed above sets $\lambda_i(H^{(t)})$ with $i \geq r+1$ to $\lambda_r(H_{\mathcal{S}}^{(t)})$ which aims to achieve a small condition number and a smaller sample size. Note that, the above approach for constructing the approximate Hessian is only for convenience of convergence analysis. Erdogdu and Montanari (2015) provided a computation efficient implementation of NewSamp.

We now give a novel analysis of convergence properties of NewSamp (Algorithm 6) based on our framework in Theorem 3 and Theorem 5. Our convergence analysis provides a tighter bound on the samples size than the one of Erdogdu and Montanari (2015). We first give how the approximate Hessian in Eqn. (17) approximates the Hessian.

Lemma 14 *Assume Eqn.* (12) *and Eqn.* (13) *hold, and let* $0 < \delta < 1$ *and target rank* r *be given. Let* λ_{r+1} *be the* $(r+1)$*-th eigenvalue of* $\nabla^2 F(x^{(t)})$*. Set the sample size* $|\mathcal{S}| \geq \frac{18K\log(2d/\delta)}{\lambda_{r+1}}$*. Then, with probability at* $1-\delta$*, the approximate Hessian* $H^{(t)}$ *is constructed in Eqn.* (17) *satisfies*

$$(1-\epsilon_0)H^{(t)} \preceq \nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0)H^{(t)}, \text{ with } \epsilon_0 = \max\left(\frac{5\lambda_{r+1}+\mu}{5\lambda_{r+1}+3\mu}, \frac{1}{2}\right).$$

Combining above lemma with Theorem 3 and Theorem 5, we can obtain the local and global convergence properties of Algorithm 6 in the following theorem.

Theorem 15 *Let* $F(x)$ *satisfy the properties described in Theorem 3. Assume Eqn.* (12) *and Eqn.* (13) *hold, and let* $0 < \delta < 1$ *and target rank* r *be given. Let* λ_{r+1} *be the* $(r+1)$*-th eigenvalue of* $\nabla^2 F(x^{(t)})$*. Set the sample size* $|\mathcal{S}| \geq \frac{18K\log(2d/\delta)}{\lambda_{r+1}}$*, and define*

$$\epsilon_0 = \max\left(\frac{5\lambda_{r+1}+\mu}{5\lambda_{r+1}+3\mu}, \frac{1}{2}\right), \tag{18}$$

which implies $0 < \epsilon_0 < 1$*. Assume the direction vector* $p^{(t)}$ *satisfies Eqn.* (4)*. Then for* $t = 1, \ldots, T$*, Algorithm 6 has the following convergence properties:*

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

Table 2: Comparison with previous work. We use (Reg)SSN to denote the (regularized) subsampled Newton method. For all algorithms, we set $\epsilon_1 = 0$. The notation $\tilde{\mathcal{O}}(\cdot)$ hides the polynomial of $\log(d/\delta)$. We obtain the iteration complexities of algorithms by their local convergence rates.

Method	Reference	Sample Size	Iterations Complexity
SSN	Theorem 5 of Roosta-Khorasani and Mahoney (2019)	$\tilde{\mathcal{O}}(K^2/\mu^2)$	$\tilde{\mathcal{O}}(\log(1/\epsilon))$
	Theorem 10	$\tilde{\mathcal{O}}(K/\mu)$	$\tilde{\mathcal{O}}(\log(1/\epsilon))$
RegSSN	**Theorem 13**	$\tilde{\mathcal{O}}\left(K/\xi\right)$	$\tilde{\mathcal{O}}\left(\frac{\xi}{\mu}\log(1/\epsilon)\right)$
NewSamp	Theorem 3.2 of Erdogdu and Montanari (2015)	$\tilde{\mathcal{O}}\left(K^2/\mu^2\right)$	$\tilde{\mathcal{O}}(\log(1/\epsilon))$
	Theorem 15	$\tilde{\mathcal{O}}\left(K/\lambda_{r+1}\right)$	$\tilde{\mathcal{O}}\left(\frac{\lambda_{r+1}}{\mu}\log(1/\epsilon)\right)$

(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 $1-\delta$.*

(c) If $F(x)$ is self-concordant, the iteration complexity of the sketch Newton with backtracking line search (Algorithm 2 with $H^{(t)}$ constructed in Eqn. (17)*) is upper bounded by Eqn.* (8)*.*

The first term of the right hand of Eqn. (18) reveals the relationship between the target rank r and sample size. We can observe the sample size is linear in $1/\lambda_{r+1}$. Hence, a small r means that small sample size is sufficient. Conversely, if we want to sample a small portion of f_i's, we should choose a small r. Eqn. (18) shows that small sample size will lead to a poor convergence rate. If we set $r = 0$, then ϵ_0 will be $1 - \frac{2\mu}{5\lambda_1+3\mu}$. Consequently, the convergence rate of NewSamp is almost the same as gradient descent.

It is worth pointing out that Theorem 15 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 and Montanari, 2015).

5.4 Comparison with Previous Work

We compare our results in this section with previous work. Although many variants of subsampled Newton methods have been proposed recently, they share a similar proof procedure. Thus, these algorithms have almost the same sample size and convergence rate. For example, the subsampled Newton method (Roosta-Khorasani and Mahoney, 2019) and NewSamp (Erdogdu and Montanari, 2015) have the same order of sample size and convergence rate (referring to Table 2). Thus, we only compare our results with the recent work of Roosta-Khorasani and Mahoney (2019) and NewSamp (Erdogdu and Montanari, 2015). The detailed comparison is listed in Table 2.

First, comparing our analysis of subsampled Newton with the one of Roosta-Khorasani and Mahoney (2019), we can see that to achieve the same convergence rate, our result only needs $\tilde{\mathcal{O}}(K/\mu)$ in contrast to $\tilde{\mathcal{O}}(K^2/\mu^2)$ of Roosta-Khorasani and Mahoney (2019). Hence, our result is substantially much tighter than previous work.

Then we compare our theoretical analysis of NewSamp with that of Erdogdu and Montanari (2015). We can observe that although NewSamp is a kind of regularized subsampled Newton, it still takes $\tilde{\mathcal{O}}(K^2/\mu^2)$ samples which are the same with subsampled Newton. In contrast, our analysis (Theorem 15) describes how regularization reduces the sample number and convergence speed. This

Table 3: Datasets Description

Dataset	n	d	source
mushrooms	$8,124$	112	UCI
a9a	$32,561$	123	UCI
Covertype	$581,012$	54	UCI

theory matches the empirical study that a small r (implying a large λ_{r+1}) will reduce the sample number and convergence speed (Erdogdu and Montanari, 2015).

Finally, we compare NewSamp with regularized subsampled Newton (Algorithm 5). We mainly focus on the parameter ϵ_0 in Theorems 13 and 15 which mainly determines convergence properties of Algorithms 5 and 6. Specifically, if we set $\xi = \lambda_{r+1}$ in Eqn. (16), then $\epsilon_0 = \frac{3\lambda_{r+1}+\mu}{3\lambda_{r+1}+3\mu}$ which is of the same order of the first term of the right hand of Eqn. (18). Hence, we can regard NewSamp as a special case of Algorithm 5. However, NewSamp provides an approach for automatic choice of ξ. Recall that NewSamp includes another parameter: the target rank r. Thus, NewSamp and Algorithm 5 have the same number of free parameters. If r is not properly chosen, NewSamp will still have poor performance. Therefore, Algorithm 5 is preferred because NewSamp needs extra cost to perform SVD.

6. Empirical Analysis

In this section, we experimentally validate our theoretical results about the unnecessity of the Lipschitz continuity condition of $\nabla^2 F(x)$, sketching size of the sketch Newton, and how the regularization affects the sample number and convergence rate of regularized Newton.

In the following experiments, we choose $x^{(0)} = 0$ as the initial point. Furthermore, we use $p^{(t)} = [H^{(t)}]^{-1}\nabla F(x^{(t)})$ as the descent vector which implies $\epsilon_1 = 0$. We will obtain different convergence rates ϵ_0 by choosing different sketching or sample sizes.

6.1 Unnecessity of Lipschitz Continuity of Hessian

We conduct the experiment on the primal problem for the linear SVM as follows

$$\min_x F(x) = \frac{1}{2}\|x\|^2 + \frac{C}{2n}\sum_{i=1}^{n} \ell(b_i, \langle x, a_i\rangle),$$

where $C > 0$, the (a_i, b_i) denote the training data, x is the separating hyperplane, and $\ell(\cdot)$ is the loss function. In our experiment, we use Hinge-2 loss as our loss function. That is,

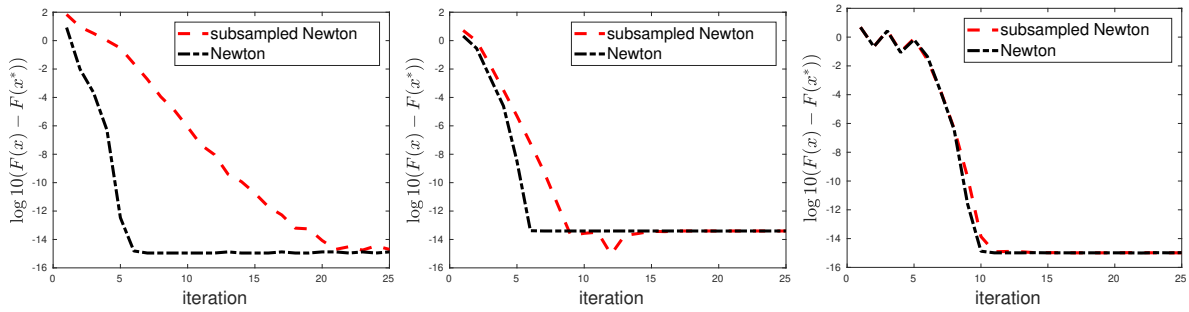
$$\ell(b, \langle x, a\rangle) = \max(0, 1 - b\langle x, a\rangle)^2.$$

Let $SV^{(t)}$ denote the set of indices of all the support vectors at iteration t, namely,

$$SV^{(t)} = \{i : b_i\langle x^{(t)}, a_i\rangle < 1\}.$$

Then the Hessian matrix of $F(x^{(t)})$ can be written as

$$\nabla^2 F(x^{(t)}) = I + \frac{1}{n}\sum_{i\in SV^{(t)}} a_i a_i^T.$$



(a) mushrooms. (b) a9a. (c) covtype.

Figure 1: Convergence properties on different datasets.

From the above equation, we can see that $\nabla^2 F(x)$ is not Lipschitz continuous.

Without loss of generality, we use the Subsampled Newton method (Algorithm 4) in our experiment. We sample 5% support vectors in each iteration. Our experiments are implemented on three datasets whose detailed description is given in Table 3 and the results are reported in Figure 1.

From Figure 1, we see that Subsampled Newton converges linearly and the Newton method converges superlinearly. This matches our theory that the Lipschitz continuity of $\nabla^2 F(x)$ is not necessary to achieve a linear or superlinear convergence rate.

6.2 Sketching Size of Sketch Newton

Now we validate our theoretical result that sketching size is independent on the condition number of the Hessian matrix in Sketch Newton. To control the condition number of the Hessian conveniently, we conduct the experiment on the least squares regression problem:

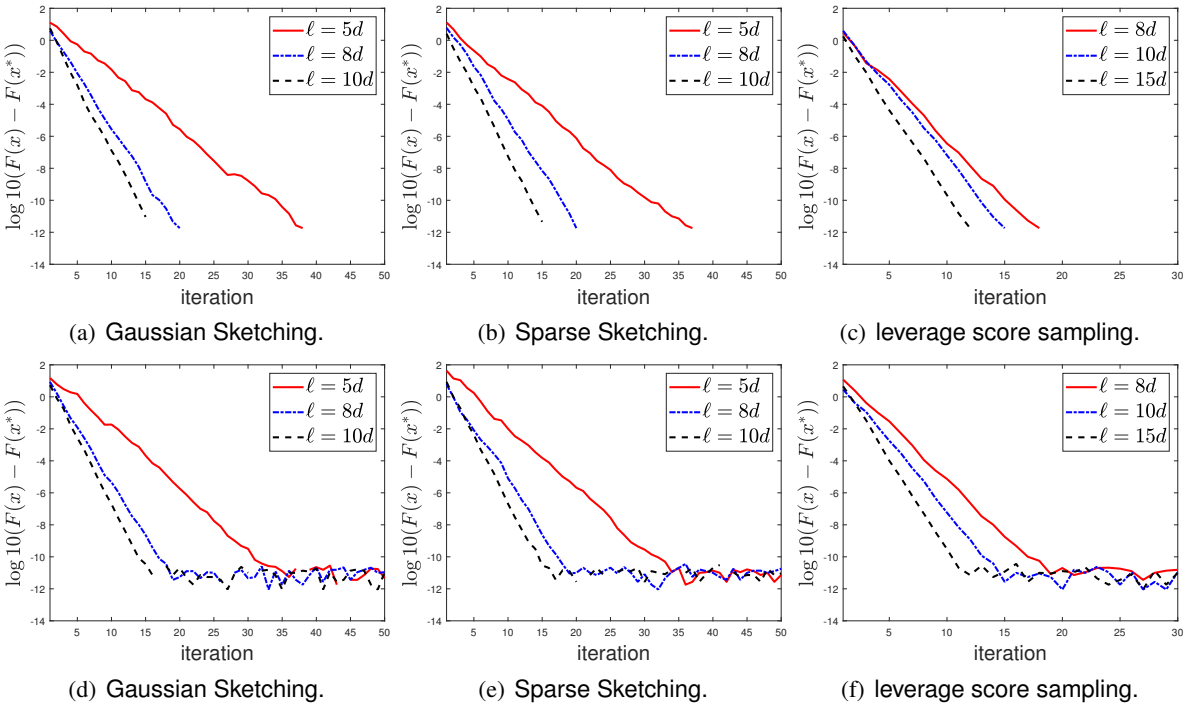
$$\min_x \frac{1}{2}\|Ax - b\|^2. \tag{19}$$

In each iteration, the Hessian matrix is $A^T A$. In our experiment, A is a 10000×54 matrix of the form $A = U\Sigma V^\top$. U and V are orthonormal matrices and Σ is a diagonal matrix whose diagonal entries are singular values of A. Vector b is a d dimension Gaussian vector. We set the singular values σ_i of A as:

$$\sigma_i = 1.2^{-i} \quad \text{and} \quad \sigma_i = 1.1^{-i}.$$

Then the condition numbers of A are $\kappa(A) = 1.2^{54} = 1.8741 \times 10^4$ and $\kappa(A) = 1.1^{54} = 172$, respectively. We use different sketching matrices in Sketch Newton (Algorithm 3). We control the value of ϵ_0 by choosing different sketching sizes ℓ. We report our empirical results in Figure 2.

Figure 2 shows 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. Furthermore, we see that the convergence rate of Sketch Newton is independent on the condition number of the function but depends on both the sketching size and sketching matrix. This matches our theory that sketching size is independent on the condition number of the Hessian matrix to achieve a linear convergence rate.



(a) Gaussian Sketching. (b) Sparse Sketching. (c) leverage score sampling.

(d) Gaussian Sketching. (e) Sparse Sketching. (f) leverage score sampling.

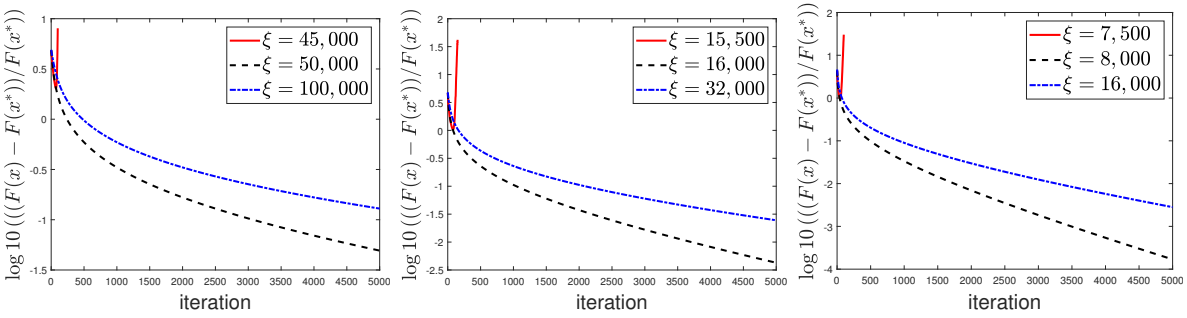
Figure 2: Convergence properties of different sketching sizes. In the top row, the condition number of the Hessian is $\kappa = \kappa^2(A) = 172^2$. In the bottom row, the condition number of the Hessian is $\kappa = \kappa^2(A) = 1.87^2 \times 10^8$.

6.3 Sample Size of Regularized Subsampled Newton

We also consider the least squares regression defined in Eqn. (19) 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 $A \in \mathbb{R}^{n \times d}$ be a random matrix whose entries are drawn from $[0, 1]$ uniformly with $n = 6,000$ and $d = 5,000$. In this case, Sketch Newton can not be used because n and d are close to each other. Thus, we conduct experiments on the regularized subsampled Newton (Algorithm 5).

By Theorem 13, we know that ϵ_0 and sample size $|\mathcal{S}|$ depend on the value of regularizer ξ. In our experiment, we study how $|\mathcal{S}|$ and ξ effect the convergence rate of ϵ_0. We set different sample sizes $|\mathcal{S}|$ and set different regularizer terms ξ for each $|\mathcal{S}|$. We report our results in Figures 3.

As we can see, if the sample size $|\mathcal{S}|$ is small, we have to choose a large ξ in Algorithm 5; otherwise the algorithm will diverge. However, if the regularizer term ξ is too large, it will lead to a slow convergence rate. Furthermore, increasing the sample size and choosing a proper regularizer will improve the convergence rate obviously. When $|\mathcal{S}| = 300$, we can obtain a solution with precision of 10^{-4} with proper ξ in contrast to $10^{-1.5}$ for $|\mathcal{S}| = 50$. Thus, our empirical study matches the theoretical analysis in Subsection 5.2 very well.



(a) Sample Size $|\mathcal{S}| = 50$. (b) Sample size $|\mathcal{S}| = 150$. (c) Sample size $|\mathcal{S}| = 300$.

Figure 3: Convergence properties of Regularized Subsampled Newton

7. Conclusion

In this paper we have proposed a framework to analyze both local and global 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 were unknown before. The most important thing is in 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 the future.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. Zhihua Zhang has been supported by the National Natural Science Foundation of China (No. 11771002), Beijing Natural Science Foundation (Z190001), and Beijing Academy of Artificial Intelligence (BAAI).

Appendix A. Some Important Lemmas

In this section, we give several important lemmas which will be used in the proof of the theorems in this paper.

Lemma 16 *If A and B are $d \times d$ symmetric positive matrices, and $(1-\epsilon_0)B \preceq A \preceq (1+\epsilon_0)B$ where $0 < \epsilon_0 < 1$, then we have*

$$(1-\epsilon_0) \cdot I \preceq A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} \preceq (1+\epsilon_0) \cdot I$$

where I is the identity matrix.

Proof Because $A \preceq (1+\epsilon_0)B$, we have $z^T[A-(1+\epsilon_0)B]z \leq 0$ for any nonzero $z \in \mathbb{R}^d$. This implies $\frac{z^TAz}{z^TBz} \leq 1+\epsilon_0$ for any $z \neq 0$. Subsequently,

$$\begin{aligned}
\lambda_{\max}(B^{-1}A) =& \lambda_{\max}(B^{-1/2}AB^{-1/2}) \\
=& \max_{u \neq 0} \frac{u^T B^{-1/2} A B^{-1/2} u}{u^T u} \\
=& \max_{z \neq 0} \frac{z^T A z}{z^T B z} \\
\leq& 1+\epsilon_0,
\end{aligned}$$

where the last equality is obtained by setting $z = B^{-1/2}u$. Similarly, we have $\lambda_{\min}(B^{-1}A) \geq 1-\epsilon_0$. Since $B^{-1}A$ and $A^{1/2}B^{-1}A^{1/2}$ are similar, the eigenvalues of $A^{1/2}B^{-1}A^{1/2}$ are all between $1-\epsilon_0$ and $1+\epsilon_0$. ■

Lemma 17 (Matrix Chernoff Inequality Tropp et al. (2015)) *Let $X_1, X_2, \ldots, X_k$ be independent, random, symmetric, real matrices of size $d \times d$ with $0 \preceq X_i \preceq LI$, where I is the $d \times d$ identity matrix. Let $Y = \sum_{i=1}^k X_i$, $\mu_{\min} = \lambda_{\min}(\mathbb{E}[Y])$ and $\mu_{\max} = \lambda_{\max}(\mathbb{E}[Y])$. Then, we have*

$$\mathbb{P}\left(\lambda_{\min}(Y) \leq (1-\epsilon)\mu_{\min}\right) \leq d \cdot e^{-\epsilon^2 \mu_{\min}/2L},$$

and

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq (1+\epsilon)\mu_{\max}\right) \leq d \cdot e^{-\epsilon^2 \mu_{\min}/3L}.$$

Appendix B. Proofs of Theorem 3

The proof Theorem 3 consists of the following lemmas. First, by Lemma 18, we upper bound $\left\|x^{(t+1)} - x^*\right\|_M$ by three terms. The first term dominates the convergence property. The second term depicts how the approximate descent direction affects convergence. The third term is the high order term.

In Lemma 19, we prove that the first term of right hand of Eqn. (20) is upper bounded by $\epsilon_0 \left\|x^{(t)} - x^*\right\|$ and a high order term. Lemma 20 shows that the second term affect the convergence rate at most ϵ_1. In Lemma 21, we complete the convergence analysis when the Hessian is continuous near the optimal point but the Hessian is not Lipschitz continuous. If the Hessian is $\hat{L}$-Lipschitz continuous, Lemma 22 provides the detailed convergence analysis.

Lemma 18 *Letting sequence $\{x^{(t)}\}$ update as Algorithm 1, then it satisfies*

$$\begin{aligned}\left\|x^{(t+1)}-x^*\right\|_M \leq &\left\|I-M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\|\cdot\left\|x^{(t)}-x^*\right\|_M+\left\|[H^{(t)}]^{-1}\nabla F(x^{(t)})-p^{(t)}\right\|_M\\ &+\left\|M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)})-\nabla^2F(x^*)(x^{(t)}-x^*)\right)\right\|,\end{aligned}\tag{20}$$

where $M=\nabla^2F(x^*)$.

Proof By the update procedure of $x^{(t)}$, we have

$$\begin{aligned}x^{(t+1)}-x^* =&x^{(t)}-x^*-p^{(t)}\\ =&x^{(t)}-x^*-[H^{(t)}]^{-1}\nabla F(x^{(t)})+[H^{(t)}]^{-1}\nabla F(x^{(t)})-p^{(t)}\\ =&x^{(t)}-x^*+[H^{(t)}]^{-1}\nabla F(x^{(t)})-p^{(t)}\\ &-[H^{(t)}]^{-1}\left(\nabla^2F(x^*)(x^{(t)}-x^*)+\nabla F(x^{(t)})-\nabla^2F(x^*)(x^{(t)}-x^*)\right).\end{aligned}$$

Letting us denote $M=\nabla^2F(x^*)$, and multiplying $M^{1/2}$ to the left and right hands of above equality, we can obtain that

$$\begin{aligned}M^{1/2}(x^{(t+1)}-x^*) =&M^{1/2}(x^{(t)}-x^*)-M^{1/2}[H^{(t)}]^{-1}M^{1/2}\cdot M^{1/2}(x^{(t)}-x^*)\\ &+M^{1/2}\left([H^{(t)}]^{-1}\nabla F(x^{(t)})-p^{(t)}\right)\\ &-M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)})-\nabla^2F(x^*)(x^{(t)}-x^*)\right).\end{aligned}$$

Thus, we can obtain that

$$\begin{aligned}\left\|x^{(t+1)}-x^*\right\|_M \leq &\left\|I-M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\|\cdot\left\|x^{(t)}-x^*\right\|_M+\left\|[H^{(t)}]^{-1}\nabla F(x^{(t)})-p^{(t)}\right\|_M\\ &+\left\|M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)})-\nabla^2F(x^*)(x^{(t)}-x^*)\right)\right\|.\end{aligned}$$

■

Lemma 19 *Assume that the objective function $F(x)$ satisfies Assumption 1 and 2. Let M denote $\nabla^2F(x^*)$, and the approximate Hessian $H^{(t)}$ satisfy Condition* (2)*. Then if $\|\Delta\|$ is sufficient small with*

$$\Delta=\nabla^2F(x^*)-\nabla^2F(x^{(t)}),\tag{21}$$

we have

$$\left\|I-M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\|\leq\epsilon_0+\frac{2\kappa\mu^{-1}\|\Delta\|}{1-\kappa\mu^{-1}\|\Delta\|}.$$

Proof If $\|\Delta\|$ is sufficient small (which implies that $\nabla^2F(x^*)$ and $\nabla^2F(x^{(t)})$ are close enough), then we have

$$\begin{aligned}\lambda_{\max}\left([\nabla^2F(x^*)]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^*)]^{\frac{1}{2}}\right)=&1+\epsilon_0'\\ \lambda_{\min}\left([\nabla^2F(x^*)]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^*)]^{\frac{1}{2}}\right)=&1-\epsilon_0'',\end{aligned}$$

with $0 < \epsilon_0' < 1, 0 < \epsilon_0'' < 1$. This will imply that

$$\left\| I - M^{1/2}[H^{(t)}]^{-1}M^{1/2} \right\| \leq \max\{\epsilon_0', \epsilon_0''\}.$$

Now we begin to bound the value of ϵ_0'. First, by the definition of $\lambda_{\max}(\cdot)$ for the positive definite matrix, we have

$$\begin{aligned}
&\lambda_{\max}\left([\nabla^2 F(x^*)]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^*)]^{\frac{1}{2}}\right) \\
=&\max_{x\neq 0} \frac{x^T M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\|x\|^2} \\
=&\max_{x\neq 0} \left(\frac{x^T[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}x}{\|x\|^2} \right. \\
&\left. + \frac{x^T\left(M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}} - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right)x}{\|x\|^2} \right) \\
\leq& 1 + \epsilon_0 + \max_{x\neq 0} \frac{x^T\left(M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}} - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right)x}{\|x\|^2} \\
\leq& 1 + \epsilon_0 + \max_{x\neq 0} \frac{x^T M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\|x\|^2} + \max_{x\neq 0} \frac{-x^T[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}x}{\|x\|^2}, \quad (22)
\end{aligned}$$

where the first inequality is because of Condition (2) and Lemma 16. Furthermore, by setting $y = M^{\frac{1}{2}}x$, we have

$$\max_{x\neq 0} \frac{x^T M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\|x\|^2} = \max_{y\neq 0} \frac{y^T[H^{(t)}]^{-1}y}{y^T M^{-1}y}. \quad (23)$$

We also have

$$\left\| M^{-1} - [\nabla^2 F(x^{(t)})]^{-1} \right\| \leq \left\| M^{-1} \right\| \left\| [\nabla^2 F(x^{(t)})]^{-1} \right\| \|\Delta\|, \quad (24)$$

which implies that for any $u \neq 0$, it holds that

$$\begin{aligned}
&-\left\| M^{-1} \right\| \left\| [\nabla^2 F(x^{(t)})]^{-1} \right\| \|\Delta\| \cdot \|u\|^2 \leq u^T\left(M^{-1} - [\nabla^2 F(x^{(t)})]^{-1}\right)u \\
\Rightarrow& \left(1 - \left\| M^{-1} \right\| \left\| [\nabla^2 F(x^{(t)})]^{-1} \right\| \|\Delta\| \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1})\right) \cdot u^T[\nabla^2 F(x^{(t)})]^{-1}u \leq u^T M^{-1}u \\
\Rightarrow& \left(1 - \left\| M^{-1} \right\| \left\| [\nabla^2 F(x^{(t)})]^{-1} \right\| \|\Delta\| \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1})\right) [\nabla^2 F(x^{(t)})]^{-1} \preceq M^{-1}.
\end{aligned}$$

Note that it holds that $\left\|[\nabla^2 F(x^{(t)})]^{-1}\right\| \cdot \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1}) \leq \kappa$, we can obtain that

$$\left(1 - \kappa \|\Delta\| \left\| M^{-1} \right\|\right) [\nabla^2 F(x^{(t)})]^{-1} \preceq M^{-1}.$$

Combining with Eqn. (23), we have

$$\max_{y\neq 0} \frac{y^T[H^{(t)}]^{-1}y}{y^T M^{-1}y} \leq \frac{1}{1 - \kappa \|\Delta\| \|M^{-1}\|} \cdot \max_{y\neq 0} \frac{y^T[H^{(t)}]^{-1}y}{y^T[\nabla^2 F(x^{(t)})]^{-1}y}$$

$$=\frac{1}{1-\kappa\mu^{-1}\left\|\Delta\right\|}\cdot\max_{x\neq 0}\frac{x^T[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}x}{\left\|x\right\|^2},$$

where the last equality is using $x=[\nabla^2F(x^{(t)})]^{-\frac{1}{2}}y$ and $\left\|M^{-1}\right\|\leq\mu^{-1}$. Combining with Eqn. (22) and (23), we can obtain

$$\begin{aligned}
&\lambda_{\max}\left([\nabla^2F(x^*)]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^*)]^{\frac{1}{2}}\right)\\
\leq&1+\epsilon_0+\left(\frac{1}{1-\kappa\mu^{-1}\left\|\Delta\right\|}-1\right)\cdot\max_{x\neq 0}\frac{x^T[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}x}{\left\|x\right\|^2}\\
\leq&1+\epsilon_0+\frac{\kappa\mu^{-1}\left\|\Delta\right\|}{1-\kappa\mu^{-1}\left\|\Delta\right\|}\cdot(1+\epsilon_0),
\end{aligned}$$

where the last inequality is due to Condition (2) and Lemma 16. By $\epsilon_0<1$, we can obtain that

$$\epsilon_0'\leq\epsilon_0+\frac{2\kappa\mu^{-1}\left\|\Delta\right\|}{1-\kappa\mu^{-1}\left\|\Delta\right\|}.$$

Now we are going to bound the value of ϵ_0''. By the definition of $\lambda_{\min}(\cdot)$, we have

$$\begin{aligned}
&\lambda_{\min}\left([\nabla^2F(x^*)]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^*)]^{\frac{1}{2}}\right)\\
=&\min_{x\neq 0}\frac{x^TM^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\left\|x\right\|^2}\\
=&\min_{x\neq 0}\left(\frac{x^T[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}x}{\left\|x\right\|^2}\right.\\
&\left.+\frac{x^T\left(M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}-[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}\right)x}{\left\|x\right\|^2}\right)\\
\geq&1-\epsilon_0+\min_{x\neq 0}\frac{x^T\left(M^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}-[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}\right)x}{\left\|x\right\|^2}\\
\geq&1-\epsilon_0+\min_{x\neq 0}\frac{x^TM^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\left\|x\right\|^2}+\min_{x\neq 0}\frac{-x^T[\nabla^2F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2F(x^{(t)})]^{\frac{1}{2}}x}{\left\|x\right\|^2},
\end{aligned}\tag{25}$$

where the first inequality is because of Condition (2).

By Eqn. (24), we have

$$\begin{aligned}
&u^T\left(M^{-1}-[\nabla^2F(x^{(t)})]^{-1}\right)u\leq\left\|M^{-1}\right\|\left\|[\nabla^2F(x^{(t)})]^{-1}\right\|\left\|\Delta\right\|\cdot\left\|u\right\|^2\\
\Rightarrow&u^TM^{-1}u\leq\left(1+\left\|M^{-1}\right\|\left\|[\nabla^2F(x^{(t)})]^{-1}\right\|\left\|\Delta\right\|\lambda_{\min}([\nabla^2F(x^{(t)})]^{-1})\right)\cdot u^T[\nabla^2F(x^{(t)})]^{-1}u\\
\Rightarrow&M^{-1}\preceq\left(1+\kappa\left\|M^{-1}\right\|\left\|\Delta\right\|\right)[\nabla^2F(x^{(t)})]^{-1}.
\end{aligned}$$

Thus, we can obtain that

$$\min_{x\neq 0}\frac{x^TM^{\frac{1}{2}}[H^{(t)}]^{-1}M^{\frac{1}{2}}x}{\left\|x\right\|^2}=\min_{y\neq 0}\frac{y^T[H^{(t)}]^{-1}y}{y^TM^{-1}y}$$

$$\begin{aligned}
\geq & \frac{1}{1+\kappa \left\|M^{-1}\right\| \|\Delta\|} \cdot \min_{y \neq 0} \frac{y^T [H^{(t)}]^{-1} y}{y^T [\nabla^2 F(x^{(t)})]^{-1} y} \\
\geq & \frac{1}{1+\kappa\mu^{-1} \|\Delta\|} \cdot \min_{x \neq 0} \frac{x^T [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} x}{\|x\|^2},
\end{aligned}$$

where the first equality is because of $y = M^{\frac{1}{2}} x$ and the last equality is due to $x = [\nabla^2 F(x^{(t)})]^{-\frac{1}{2}} y$.

Combining with Eqn. (25), we can obtain

$$\begin{aligned}
& \lambda_{\min} \left([\nabla^2 F(x^*)]^{\frac{1}{2}} [H^{(t)}]^{-1} [\nabla^2 F(x^*)]^{\frac{1}{2}} \right) \\
\geq & 1 - \epsilon_0 + \min_{x \neq 0} \left(\left(\frac{1}{1+\kappa\mu^{-1} \|\Delta\|} - 1 \right) \cdot \frac{x^T [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} x}{\|x\|^2} \right) \\
\geq & 1 - \epsilon_0 - \frac{2\kappa\mu^{-1} \|\Delta\|}{1+\kappa\mu^{-1} \|\Delta\|}.
\end{aligned}$$

Thus, we can obtain that

$$\epsilon_0'' \leq \epsilon_0 + \frac{2\kappa\mu^{-1} \|\Delta\|}{1+\kappa\mu^{-1} \|\Delta\|}.$$

Therefore, we have

$$\left\| I - M^{1/2} [H^{(t)}]^{-1} M^{1/2} \right\| \leq \max\{\epsilon_0', \epsilon_0''\} \leq \epsilon_0 + \frac{2\kappa\mu^{-1} \|\Delta\|}{1-\kappa\mu^{-1} \|\Delta\|}.$$

■

Lemma 20 *Let $p^{(t)}$ satisfy Condition* (4) *and $F(x)$ satisfy Assumption 1 and 2, then we have*

$$\left\| [H^{(t)}]^{-1} \nabla F(x^{(t)}) - p^{(t)} \right\|_M \leq \epsilon_1 \left\| x^{(t)} - x^* \right\|_M. \tag{26}$$

Proof First, we have

$$\begin{aligned}
\left\| [H^{(t)}]^{-1} \nabla F(x^{(t)}) - p^{(t)} \right\|_M = & \left\| M^{1/2} [H^{(t)}]^{-1} \left(\nabla F(x^{(t)}) - H^{(t)} p^{(t)} \right) \right\| \\
\overset{(4)}{\leq} & \epsilon_1 (1+\epsilon_0)^{-1} \kappa^{-3/2} \left\| M^{1/2} \right\| \left\| [H^{(t)}]^{-1} \right\| \left\| \nabla F(x^{(t)}) \right\| \\
\overset{(2)}{\leq} & \epsilon_1 \kappa^{-3/2} \left\| M^{1/2} \right\| \left\| [\nabla^2 F(x^{(t)})]^{-1} \right\| \left\| \nabla F(x^{(t)}) \right\| \\
\leq & \epsilon_1 \kappa^{-1/2} \left\| M^{1/2} \right\| \left\| x^{(t)} - x^* \right\| \\
\leq & \epsilon_1 \left\| x^{(t)} - x^* \right\|_M,
\end{aligned}$$

where the last two inequalities follow from the assumptions that $F(x)$ is L-smooth and μ-strongly convex. ■

In following lemma, we will prove the result of Theorem 3 (a).

Lemma 21 *There exists a sufficient small value γ, $\nu = o(1)$, such that when $\|x^{(t)} - x^*\|_M \leq \gamma$, the sequence $\{x^{(t)}\}$ of Algorithm 1 satisfies*

$$\left\|x^{(t+1)} - x^*\right\|_M \leq \left(\epsilon_0 + \epsilon_1 + 2\nu\mu^{-1} + \frac{2\kappa\mu^{-1}\nu(1+\nu\mu^{-1})}{1-\kappa\mu^{-1}\nu}\right)\left\|x^{(t)} - x^*\right\|_M.$$

Proof Because $\nabla^2 F(x)$ is continuous around x^*, then existing a sufficient small value γ such that if $\left\|x^{(t)} - x^*\right\|_M \leq \gamma$, then it holds that (Ortega and Rheinboldt, 1970)

$$\left\|\nabla^2 F(x^*) - \nabla^2 F(x^{(t)})\right\| \leq \nu, \tag{27}$$

and

$$\left\|\nabla F(x^{(t)}) - \nabla F(x^*) - \nabla^2 F(x^*)(x^{(t)} - x^*)\right\|_M \leq \nu\left\|x^{(t)} - x^*\right\|_M. \tag{28}$$

By Lemma 19, we have

$$\left\|M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\| \leq 1 + \epsilon_0 + \frac{2\kappa\mu^{-1}\|\Delta\|}{1-\kappa\mu^{-1}\|\Delta\|} \overset{(27)}{\leq} 1 + \epsilon_0 + \frac{2\kappa\mu^{-1}\nu}{1-\kappa\mu^{-1}\nu}.$$

Combining with Lemma 18, 19 and 20, we can obtain that

$$\begin{aligned}
\left\|x^{(t+1)} - x^*\right\|_M &\leq \left(\epsilon_0 + \epsilon_1 + \frac{2\kappa\mu^{-1}\nu}{1-\kappa\mu^{-1}\nu}\right)\left\|x^{(t)} - x^*\right\|_M \\
&\quad + \left\|M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)}) - \nabla^2 F(x^*)(x^{(t)} - x^*)\right)\right\| \\
&\overset{(28)}{\leq} \left(\epsilon_0 + \epsilon_1 + \frac{2\kappa\mu^{-1}\nu}{1-\kappa\mu^{-1}\nu}\right)\left\|x^{(t)} - x^*\right\|_M \\
&\quad + \nu\left\|M^{-1}\right\|\left\|M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\|\left\|x^{(t)} - x^*\right\|_M \\
&\leq \left(\epsilon_0 + \epsilon_1 + 2\nu\mu^{-1} + \frac{2\kappa\mu^{-1}\nu(1+\nu\mu^{-1})}{1-\kappa\mu^{-1}\nu}\right)\left\|x^{(t)} - x^*\right\|_M.
\end{aligned}$$

From above equation, we can observe that if $\epsilon_0 + \epsilon_1 < 1$ and ν is sufficiently small which can be guaranteed by choosing proper γ, then we have $\left\|x^{(t+1)} - x^*\right\|_M \leq \left\|x^{(t)} - x^*\right\|_M \leq \gamma$. ■

In following lemma, we will prove the result of Theorem 3 (b).

Lemma 22 *Let the Hessian of $F(x)$ be $\hat{L}$-Lipschitz continuous and $x^{(t)}$ satisfy $\left\|x^{(t)} - x^*\right\|_M \leq \mu^{3/2}\hat{L}^{-1}$. Then the sequence $\{x^{(t)}\}$ of Algorithm 1 satisfies*

$$\left\|x^{(t+1)} - x^*\right\|_M \leq (\epsilon_0 + \epsilon_1)\left\|x^{(t)} - x^*\right\|_M + 4(\kappa+1)\mu^{-3/2}\hat{L}\left\|x^{(t)} - x^*\right\|_M^2.$$

Proof By Taylor's expansion at x^*, we have

$$\nabla F(x^{(t)}) - \nabla^2 F(x^*)(x^{(t)} - x^*) = \int_0^1 \nabla^2 F\left(x^* + s(x^{(t)} - x^*)\right) - \nabla^2 F(x^*)\, ds \cdot (x^{(t)} - x^*).$$

Thus, we can obtain that

$$\begin{aligned}
&\left\|M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)})-\nabla^2 F(x^*)(x^{(t)}-x^*)\right)\right\| \\
=&\left\|M^{1/2}[H^{(t)}]^{-1}M^{1/2}\int_0^1 M^{-1/2}\left(\nabla^2 F\left(x^*+s(x^{(t)}-x^*)\right)-\nabla^2 F(x^*)\right)M^{-1/2}\,ds\cdot M^{1/2}(x^{(t)}-x^*)\right\| \\
\leq&\underbrace{\left\|M^{1/2}[H^{(t)}]^{-1}M^{1/2}\right\|}_{T_1}\cdot\underbrace{\left\|\int_1^1\left(M^{-1/2}\left(\nabla^2 F(x^*+s(x^{(t)}-x^*))\right)M^{-1/2}-I\right)\,ds\right\|}_{T_2}\cdot\left\|x^{(t)}-x^*\right\|_M.
\end{aligned}$$

Next, we will bound the value of T_1 and T_2. By Lemma 19, we have

$$T_1\leq 1+\epsilon_0+\frac{2\kappa\mu^{-1}\left\|\Delta\right\|}{1-\kappa\mu^{-1}\left\|\Delta\right\|}\leq 2+\frac{2\kappa\mu^{-1}\hat{L}\left\|x^{(t)}-x^*\right\|}{1-\kappa\mu^{-1}\hat{L}\left\|x^{(t)}-x^*\right\|}\leq 4,$$

where the last inequality is because of the condition $\left\|x^{(t)}-x^*\right\|_M\leq\frac{1}{2}\mu^{\frac{3}{2}}\kappa^{-1}\hat{L}^{-1}$ and $\left\|x^{(t)}-x^*\right\|\leq\mu^{-1/2}\left\|x^{(t)}-x^*\right\|_M$.

Letting us represent that

$$\nabla^2 F(x^*+s(x^{(t)}-x^*))=M+\Delta',$$

then we have

$$\begin{aligned}
T_2=&\left\|\int_0^1\left(M^{-1/2}(M+\Delta')M^{-1/2}-I\right)\,ds\right\| \\
=&\left\|\int_0^1\left(M^{-1/2}\Delta' M^{-1/2}\right)\,ds\right\| \\
\leq&\left\|M^{-1}\right\|\int_0^1\left\|\Delta'\right\|\,ds \\
\leq&\mu^{-1}\hat{L}\int_0^1\left\|s(x^{(t)}-x^*)\right\|\,ds \\
\leq&\frac{\mu^{-3/2}\hat{L}}{2}\left\|x^{(t)}-x^*\right\|_M.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\left\|M^{1/2}[H^{(t)}]^{-1}\left(\nabla F(x^{(t)})-\nabla^2 F(x^*)(x^{(t)}-x^*)\right)\right\| \\
\leq&T_1\cdot T_2\left\|x^{(t)}-x^*\right\|_M \\
\leq&4\cdot\frac{\mu^{-3/2}\hat{L}}{2}\left\|x^{(t)}-x^*\right\|_M \\
\leq&2\mu^{-3/2}\hat{L}\left\|x^{(t)}-x^*\right\|_M.
\end{aligned}$$

Combining with Lemma 18, 19 and 20, we can obtain that

$$\left\|x^{(t+1)}-x^*\right\|_M\leq\left(\epsilon_0+\epsilon_1+\frac{2\kappa\mu^{-1}\left\|\Delta\right\|}{1-\kappa\mu^{-1}\left\|\Delta\right\|}\right)\left\|x^{(t)}-x^*\right\|_M$$

$$
\begin{aligned}
&+ 2\mu^{-3/2}\hat{L}\left\|x^{(t)}-x^*\right\|_M\left\|x^{(t)}-x^*\right\|_M^2\\
\le&(\epsilon_0+\epsilon_1)\left\|x^{(t)}-x^*\right\|_M + 4\kappa\mu^{-3/2}\hat{L}\left\|x^{(t)}-x^*\right\|_M^2\\
&+ 4\mu^{-3/2}\hat{L}\left\|x^{(t)}-x^*\right\|_M^2\\
=&(\epsilon_0+\epsilon_1)\left\|x^{(t)}-x^*\right\|_M + 4(\kappa+1)\mu^{-3/2}\hat{L}\left\|x^{(t)}-x^*\right\|_M^{3/2},
\end{aligned}
$$

where the last inequality follows from the conditions $\left\|x^{(t)}-x^*\right\|_M \le \frac{1}{2}\mu^{\frac{3}{2}}\kappa^{-1}\hat{L}^{-1}$ and $\left\|x^{(t)}-x^*\right\| \le \mu^{-1/2}\left\|x^{(t)}-x^*\right\|_M$. ■

Appendix C. Proof of Theorem 5

For a self-concordant function $F(x)$, if two points x, y satisfy $\|x-y\|_x < 1$, where $\|v\|_x = \left\|[\nabla^2 F(x)]^{1/2}v\right\|$, we have some useful inequalities:

1. Hessian bound:

$$(1-\|x-y\|_x)^2\nabla^2 F(y) \preceq \nabla^2 F(x) \preceq \frac{1}{(1-\|x-y\|_x)^2}\nabla^2 F(y). \tag{29}$$

2. Function value bound:

$$\zeta(\|y-x\|_x) \le F(y) - F(x) - \nabla F(x)^T(y-x) \le \zeta^*(\|y-x\|_x), \tag{30}$$

where $\zeta(\alpha) = \alpha - \log(1+\alpha)$ and $\zeta^*(\alpha) = -\alpha - \log(1-\alpha)$.

This section, we will prove the convergence rate of the damped approximate Newton method. First, we will show the case that $V(x)$ is smaller than a threshold which is mainly determined by how well the Hessian is approximated. In this case, the step size $s = 1$ will satisfy the exit condition of the line search. Then, we will provide the convergence analysis when $V(x)$ is larger than the threshold where the step size s should be chosen by the line search.

Before proving the convergence analysis, we first define some new notation and clarify their relationship. Let us denote

$$V(x^{(t)}) = \left\|[\nabla^2 F(x^{(t)})]^{-1/2}\nabla F(x^{(t)})\right\|, \tag{31}$$

$$\tilde{V}(x^{(t)}) = \left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|, \tag{32}$$

and

$$\hat{V}(x^{(t)}) = \left(\nabla^T F(x^{(t)})p^{(t)}\right)^{1/2}. \tag{33}$$

Lemma 23 *Let the approximate Hessian satisfy Eqn.* (2) *and the descent direction $p^{(t)}$ satisfy Eqn.* (4)*. Then it holds that*

$$\hat{V}^2(x^{(t)}) \ge \left(1-\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\cdot\tilde{V}^2(x^{(t)}),$$

and

$$\left\|p^{(t)}\right\|_{x^{(t)}}^2 \leq (1+\epsilon_0)\left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)^2\cdot\tilde{V}^2(x^{(t)}).$$

Proof First, we have

$$\begin{aligned}\nabla^T F(x^{(t)})p^{(t)} =&\nabla^T F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)}) + \nabla^T F(x^{(t)})[H^{(t)}]^{-1}\left([H^{(t)}]p^{(t)} - \nabla F(x^{(t)})\right)\\ \overset{(4)}{\geq}&\nabla^T F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)})\\ &-\left(\nabla^T F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)})\right)^{1/2}\left\|H^{-1/2}\right\|\kappa^{-3/2}\epsilon_1\left\|\nabla F(x^{(t)})\right\|\\ \geq&\nabla^T F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)})\\ &-\kappa^{-3/2}\epsilon_1\nabla^T F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)})\left\|H^{-1/2}\right\|\left\|H^{1/2}\right\|\\ \overset{(2)}{\geq}&\left(1-\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\cdot\tilde{V}^2(x^{(t)}).\end{aligned}$$

Similarly, we can obtain that

$$\nabla^T F(x^{(t)})p^{(t)} \leq \left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\cdot\tilde{V}^2(x^{(t)}). \tag{34}$$

By the condition (2), we can obtain that

$$\left\|p^{(t)}\right\|_{x^{(t)}}^2 \leq (1+\epsilon_0)[p^{(t)}]^T[H^{(t)}]p^{(t)}. \tag{35}$$

Furthermore, we have

$$\begin{aligned}[p^{(t)}]^T[H^{(t)}]p^{(t)} =&[p^{(t)}]^T\left(\nabla F(x^{(t)}) + [H^{(t)}]p^{(t)} - \nabla F(x^{(t)})\right)\\ \leq&[p^{(t)}]^T\nabla F(x^{(t)}) + \left\|p^{(t)}\right\|\left\|[H^{(t)}]p^{(t)} - \nabla F(x^{(t)})\right\|\\ \overset{(4)}{\leq}&[p^{(t)}]^T\nabla F(x^{(t)}) + \epsilon_1\kappa^{-3/2}\left\|p^{(t)}\right\|\left\|\nabla F(x^{(t)})\right\|.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\left\|p^{(t)}\right\| \leq&\left\|p^{(t)} - [H^{(t)}]^{-1}\nabla F(x^{(t)})\right\| + \left\|[H^{(t)}]^{-1}\nabla F(x^{(t)})\right\|\\ \overset{(4)}{\leq}&\epsilon_1\kappa^{-3/2}\left\|[H^{(t)}]^{-1}\right\|\left\|\nabla F(x^{(t)})\right\| + \left\|[H^{(t)}]^{-1/2}\right\|\left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|\\ \leq&\left(\epsilon_1\kappa^{-3/2}\left\|[H^{(t)}]^{-1}\right\|\left\|[H^{(t)}]^{1/2}\right\| + \left\|[H^{(t)}]^{-1/2}\right\|\right)\left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|\\ \leq&\left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\left\|[H^{(t)}]^{-1/2}\right\|\left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|,\end{aligned}$$

and

$$\left\|\nabla F(x^{(t)})\right\| \leq \left\|[H^{(t)}]^{1/2}\right\|\left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|.$$

Thus, we can obtain that

$$\begin{aligned}\left\|p^{(t)}\right\|\left\|\nabla F(x^{(t)})\right\| \leq & \left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\left\|[H^{(t)}]^{-1/2}\right\|\left\|[H^{(t)}]^{1/2}\right\|\left\|[H^{(t)}]^{-1/2}\nabla F(x^{(t)})\right\|^2\\ \leq & \kappa^{1/2}\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\tilde{V}^2(x^{(t)}).\end{aligned}$$

Therefore, we can obtain that

$$\begin{aligned}[p^{(t)}]^T[H^{(t)}]p^{(t)} \leq & [p^{(t)}]^T\nabla F(x^{(t)})+\epsilon_1\kappa^{-3/2}\left\|p^{(t)}\right\|\left\|\nabla F(x^{(t)})\right\|\\ \leq & [p^{(t)}]^T\nabla F(x^{(t)})+\epsilon_1\kappa^{-1}\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)\tilde{V}^2(x^{(t)})\\ \overset{(34)}{\leq} & \left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)^2\tilde{V}^2(x^{(t)}).\end{aligned}$$

Combining Eqn. (35), we can obtain

$$\left\|p^{(t)}\right\|_{x^{(t)}}^2 \leq (1+\epsilon_0)\left(1+\epsilon_1\kappa^{-1}\cdot\left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}\right)^2\cdot\tilde{V}^2(x^{(t)}).$$

■

Now, we begin to prove the case that $V(x^{(t)}) \leq \frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}$ and the step size $s=1$ is sufficient.

Lemma 24 *Let the descent direction $p^{(t)}$ satisfy Eqn.* (4) *and $V(x^{(t)})$ satisfy*

$$V(x^{(t)}) \leq \frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}.$$

Then the approximate Newton with backtrack line search (Algorithm 2) has the following convergence property

$$V(x^{(t+1)}) \leq \frac{1+\epsilon_0+2\epsilon_1\kappa^{-1}}{2}V(x^{(t)}).$$

Proof Then we have

$$\begin{aligned}V(x^{(t+1)}) = & \left\|[\nabla^2F(x^{(t+1)})]^{-1/2}\nabla F(x^{(t+1)})\right\|\\ \overset{(29)}{\leq} & \frac{1}{1-\left\|p^{(t)}\right\|_x}\left\|[\nabla^2F(x^{(t)})]^{-1/2}\nabla F(x^{(t+1)})\right\|.\end{aligned}$$

By Taylor's expansion of $\nabla F(x^{(t+1)})$ at point $x^{(t)}$, we have

$$\begin{aligned}&\left\|[\nabla^2F(x^{(t)})]^{-1/2}\nabla F(x^{(t+1)})\right\|\\ =&\left\|[\nabla^2F(x^{(t)})]^{-1/2}\left(\nabla F(x^{(t)})+\nabla^2F(x^{(t)})(-p^{(t)})+\int_0^1[\nabla^2F(x^{(t)}+sp^{(t)})-\nabla^2F(x^{(t)})](-p^{(t)})ds\right)\right\|\end{aligned}$$

$$\leq \underbrace{\left\| \left(I - [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right) [\nabla^2 F(x^{(t)})]^{-1/2} \nabla F(x^{(t)}) \right\|}_{T_1}$$
$$+ \underbrace{\left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right\| \cdot \left\| [\nabla^2 F(x^{(t)})]^{-1/2} \right\| \cdot \left\| \nabla F(x^{(t)}) - H^{(t)} p^{(t)} \right\|}_{T_2}$$
$$+ \underbrace{\left\| \int_0^1 \left([\nabla^2 F(x^{(t)})]^{-1/2} \nabla^2 F(x^{(t)} - s p^{(t)}) [\nabla^2 F(x^{(t)})]^{-1/2} - I \right) ds \cdot [\nabla^2 F(x^{(t)})]^{1/2} p^{(t)} \right\|}_{T_3}.$$

We are going to bound the above terms. First, by the assumption (2), we have

$$\left\| I - [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right\| \leq \epsilon_0.$$

Combining the definition of $V(x)$, we can obtain

$$\begin{aligned} T_1 \leq & \left\| I - [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right\| \cdot \left\| [\nabla^2 F(x^{(t)})]^{-1/2} \nabla F(x^{(t)}) \right\| \\ \leq & \epsilon_0 V(x^{(t)}). \end{aligned}$$

Also by the condition (2), we have

$$\left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right\| \leq (1 + \epsilon_0).$$

Combining the condition (4) and the definition of $V^{(t)}$, we can obtain that

$$T_2 \leq (1 + \epsilon_0) \mu^{-1/2} \frac{\epsilon_1}{\kappa^{3/2}} \left\| \nabla F(x^{(t)}) \right\| \leq \frac{(1 + \epsilon_0) \epsilon_1}{\kappa} V(x^{(t)}) \leq \frac{2\epsilon_1}{\kappa} V(x^{(t)}).$$

We also have

$$\begin{aligned} T_3 \leq & \left\| \int_0^1 \left([\nabla^2 F(x^{(t)})]^{-1/2} \nabla^2 F(x^{(t)} - s p^{(t)}) [\nabla^2 F(x^{(t)})]^{-1/2} - I \right) ds \right\| \cdot \left\| p^{(t)} \right\|_x \\ \overset{(29)}{\leq} & \left| \int_0^1 \left(\frac{1}{(1 - s \left\| p^{(t)} \right\|_x)^2} - 1 \right) ds \right| \cdot \left\| p^{(t)} \right\|_x \\ = & \frac{\left\| p^{(t)} \right\|_x}{1 - \left\| p^{(t)} \right\|_x} \cdot \left\| p^{(t)} \right\|_x. \end{aligned}$$

Next, we will bound the value of $\left\| p^{(t)} \right\|_x$. We have

$$\begin{aligned} \left\| p^{(t)} \right\|_x = & \left\| [\nabla^2 F(x^{(t)})]^{1/2} p^{(t)} \right\| \\ = & \left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} \nabla F(x^{(t)}) - [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} \left(\nabla F(x^{(t)}) - H^{(t)} p^{(t)} \right) \right\| \\ \leq & \left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \cdot [\nabla^2 F(x^{(t)})]^{-1/2} \nabla F(x^{(t)}) \right\| \\ & + \left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \right\| \cdot \left\| [\nabla^2 F(x^{(t)})]^{-1/2} \right\| \cdot \left\| \nabla F(x^{(t)}) - H^{(t)} p^{(t)} \right\| \end{aligned}$$

$$
\begin{aligned}
&= \left\| [\nabla^2 F(x^{(t)})]^{1/2} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{1/2} \cdot [\nabla^2 F(x^{(t)})]^{-1/2} \nabla F(x^{(t)}) \right\| + T_2 \\
&\leq (1+\epsilon_0) V(x^{(t)}) + \frac{2\epsilon_1}{\kappa} V(x^{(t)}).
\end{aligned}
$$

Combining above results, we can obtain that

$$
\begin{aligned}
V(x^{(t+1)}) \leq & \frac{1}{1 - \left\| p^{(t)} \right\|_x} (T_1 + T_2 + T_3) \\
\leq & \frac{(\epsilon_0 + 2\epsilon_1 \kappa^{-1}) V(x^{(t)})}{1 - (1 + \epsilon_0 + 2\epsilon_1 \kappa^{-1}) V(x^{(t)})} + \frac{(1 + \epsilon_0 + 2\epsilon_1 \kappa^{-1})^2 V^2(x^{(t)})}{(1 - (1 + \epsilon_0 + 2\epsilon_1 \kappa^{-1}) V(x^{(t)}))^2}.
\end{aligned}
$$

If $V(x^{(t)})$ satisfies that

$$
\begin{aligned}
V(x^{(t)}) \leq & \frac{1 - (\epsilon_0 + 2\epsilon_1 \kappa^{-1})^2}{(1 + \epsilon_0 + 2\epsilon_1 \kappa^{-1})^2 \left(2 + \epsilon_0 + 2\epsilon_1 \kappa^{-1} + \sqrt{(2 + \epsilon_0 + 2\epsilon_1 \kappa^{-1})^2 - 1 + (\epsilon_0 + 2\epsilon_1 \kappa^{-1})^2} \right)} \\
\leq & \frac{1 - \epsilon_0 - 2\epsilon_1 \kappa^{-1}}{12}, \qquad (36)
\end{aligned}
$$

we have

$$
V(x^{(t+1)}) \leq \frac{1 + \epsilon_0 + 2\epsilon_1 \kappa^{-1}}{2} V(x^{(t)}).
$$

■

Now we begin to analyze the phase that line search should be applied to find a step size $s < 1$. This phase is commonly commonly referred as *damped phase*.

Lemma 25 *Let the approximate Hessian satisfy Eqn.* (2) *and the descent direction* $p^{(t)}$ *satisfy Eqn.* (4)*. If it holds that*

$$
\tilde{V}(x) \geq \frac{\sqrt{(1-\epsilon_0)}(1 - \epsilon_0 - 2\epsilon_1 \kappa^{-1})}{12},
$$

then Algorithm 2 has the following convergence property

$$
F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha\beta \cdot \frac{\rho^2 \tilde{V}^2(x^{(t)})}{1 + \rho \tilde{V}(x^{(t)})},
$$

where ρ *is defined as*

$$
\rho = \frac{(1-\varphi)^{1/2}}{(1+\epsilon_0)^{1/2}(1+\varphi)}, \quad \textit{with} \quad \varphi = \epsilon_1 \kappa^{-1} \cdot \left(\frac{1+\epsilon_0}{1-\epsilon_0} \right)^{1/2}.
$$

Proof By the update rule, we can obtain that

$$
\begin{aligned}
F(x^{(t+1)}) \overset{(30)}{\leq} & F(x^{(t)}) - s \nabla F(x^{(t)})^T p^{(t)} + \zeta^* \left(s \left\| p^{(t)} \right\|_{x^{(t)}} \right) \\
= & F(x^{(t)}) - s \hat{V}^2(x^{(t)}) - s \left\| p^{(t)} \right\|_{x^{(t)}} - \log \left(1 - s \left\| p^{(t)} \right\|_{x^{(t)}}) \right),
\end{aligned}
$$

with $0 \leq s < 1/\tilde{V}(x^{(t)})$.

Letting us define $\hat{s}$ as

$$\hat{s} = \frac{\hat{V}^2(x^{(t)})}{\left(\hat{V}^2(x^{(t)}) + \left\|p^{(t)}\right\|_{x^{(t)}}\right)\left\|p^{(t)}\right\|_{x^{(t)}}}.$$

We can use this bound to show the backtracking line search always results in a step size $s \geq \beta\hat{s}$. Furthermore, we can obtain that

$$\begin{aligned}
F(x^{(t+1)}) \leq & F(x^{(t)}) - \frac{\hat{V}^2(x^{(t)})}{\left\|p^{(t)}\right\|_{x^{(t)}}} - \log\left(\frac{\left\|p^{(t)}\right\|_{x^{(t)}}}{\hat{V}^2(x^{(t)}) + \left\|p^{(t)}\right\|_{x^{(t)}}}\right) \\
= & F(x^{(t)}) - \frac{\hat{V}^2(x^{(t)})}{\left\|p^{(t)}\right\|_{x^{(t)}}} + \log\left(1 + \frac{\hat{V}^2(x^{(t)})}{\left\|p^{(t)}\right\|_{x^{(t)}}}\right) \\
\leq & F(x^{(t)}) - \frac{\left(\hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}})\right)^2}{2\left(1 + \hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}}\right)}, \\
= & F(x^{(t)}) - \frac{1}{2} \cdot \hat{s}\hat{V}^2(x^{(t)}) \\
\leq & F(x^{(t)}) - \alpha \cdot \hat{s}\hat{V}^2(x^{(t)}),
\end{aligned}$$

where the second inequality follows form the fact that it holds for $a > 0$ that

$$-a + \log(1+a) + \frac{a^2}{2(1+a)} \leq 0.$$

The last inequality is because $\alpha < 1/2$. Since we obtain that $F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha \cdot \hat{s}\hat{V}^2(x^{(t)})$, we show the exit condition of the line search has satisfied. Furthermore, the exit condition holds when the step size satisfies $s \geq \beta\hat{s}$. Thus, we can obtain that

$$F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha\beta \cdot \hat{s}\hat{V}^2(x^{(t)}).$$

Next, we will bound the value of $\hat{s}\hat{V}^2(x^{(t)})$. By the definition of $\hat{s}$, we can obtain that

$$\hat{s}\hat{V}^2(x^{(t)}) = \frac{\left(\hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}}\right)^2}{\left(1 + \hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}}\right)}.$$

By Lemma 23, we have

$$\frac{\hat{V}(x^{(t)})}{\left\|p^{(t)}\right\|_{x^{(t)}}} \geq \frac{(1-\varphi)^{1/2}\tilde{V}(x^{(t)})}{(1+\epsilon_0)^{1/2}(1+\varphi)\tilde{V}(x^{(t)})} = \frac{(1-\varphi)^{1/2}}{(1+\epsilon_0)^{1/2}(1+\varphi)},$$

where $\varphi = \epsilon_1\kappa^{-1} \cdot \left(\frac{1+\epsilon_0}{1-\epsilon_0}\right)^{1/2}$. Furthermore, we have

$$\hat{s}\hat{V}^2(x^{(t)}) = \frac{\left(\hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}}\right)^2}{\left(1 + \hat{V}^2(x^{(t)})/\left\|p^{(t)}\right\|_{x^{(t)}}\right)}$$

$$\geq \frac{\left(\frac{(1-\varphi)^{1/2}}{(1+\epsilon_0)^{1/2}(1+\varphi)}\hat{V}(x^{(t)})\right)^2}{1+\frac{(1-\varphi)^{1/2}}{(1+\epsilon_0)^{1/2}(1+\varphi)}\hat{V}(x^{(t)})}$$
$$\geq \frac{\left(\frac{(1-\varphi)}{(1+\epsilon_0)^{1/2}(1+\varphi)}\tilde{V}(x^{(t)})\right)^2}{1+\frac{(1-\varphi)}{(1+\epsilon_0)^{1/2}(1+\varphi)}\tilde{V}(x^{(t)})},$$

where the last inequality follows from Lemma 23.

Letting us denote $\rho = \frac{(1-\varphi)^{1/2}}{(1+\epsilon_0)^{1/2}(1+\varphi)}$, then we have

$$F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha\beta \cdot \hat{s}\hat{V}^2(x^{(t)}) \leq F(x^{(t)}) - \alpha\beta \cdot \frac{\rho^2\tilde{V}^2(x^{(t)})}{1+\rho\tilde{V}(x^{(t)})}.$$

By the Condition (2), we have

$$\frac{1}{1-\epsilon_0}\tilde{V}^2(x^{(t)}) \geq V^2(x^{(t)}). \tag{37}$$

Thus, we can obtain that if $\tilde{V}(x) \leq \frac{(1-\epsilon_0)^{1/2}1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}$, then it holds that $V(x) \leq \frac{1-\epsilon_0}{12}$. Therefore, we can obtain that when $\tilde{V}(x) \geq \frac{(1-\epsilon_0)^{1/2}1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}$, it holds that

$$F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha\beta \cdot \frac{\rho^2\tilde{V}^2(x^{(t)})}{1+\rho\tilde{V}(x^{(t)})}.$$

■

Combining Lemma 24 and 25, we can obtain the global convergence rate of approximate Newton with backtracking line search.

Proof of Theorem 5 Let us denote

$$\eta = \alpha\beta \cdot \frac{\rho^2\left(\frac{\sqrt{(1-\epsilon_0)}(1-\epsilon_0-2\epsilon_1\kappa^{-1})}{12}\right)^2}{1+\rho\frac{\sqrt{(1-\epsilon_0)}(1-\epsilon_0-2\epsilon_1\kappa^{-1})}{12}} = \alpha\beta\frac{(1-\epsilon_0)\rho^2(1-\epsilon_0-2\epsilon_1\kappa^{-1})^2}{144+12\rho\sqrt{(1-\epsilon_0)}(1-\epsilon_0-2\epsilon_1\kappa^{-1})}.$$

By Lemma 25, we can obtain that it takes at most

$$\frac{F(x^{(0)}) - F(x^*)}{\eta}$$

steps in the damped phase because of $F(x^{(t+1)}) - F(x^{(t)}) \leq -\eta$ when $\tilde{V}(x) \geq \frac{\sqrt{(1-\epsilon_0)}(1-\epsilon_0-2\epsilon_1\kappa^{-1})}{12}$.

If it holds that $\tilde{V}(x) \leq \frac{\sqrt{(1-\epsilon_0)}(1-\epsilon_0-2\epsilon_1\kappa^{-1})}{12}$, then we have $V(x^{(t)}) \leq \frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}$. By Lemma 24, we have

$$V(x^{(t+k)}) \leq \left(\frac{1+\epsilon_0+2\epsilon_1\kappa^{-1}}{2}\right)^k \frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}.$$

Furthermore, the self-concordance of $F(x)$ implies that

$$F(x^{(t+k)}) - F(x^*) \leq V(x^{(t+k)}) \leq \left(\frac{1+\epsilon_0+2\epsilon_1\kappa^{-1}}{2}\right)^k \frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12}.$$

To make the right hand of above equation less than ϵ, then it will take no more than

$$k = \frac{2}{1-\epsilon_0-2\epsilon_1\kappa^{-1}} \log\left(\frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12\epsilon}\right)$$

iterations.

Therefore, the total complexity of approximate Newton method with backtracking line search to achieve an ϵ-suboptimality is at most

$$\frac{F(x^{(0)}) - F(x^*)}{\eta} + \frac{2}{1-\epsilon_0-2\epsilon_1\kappa^{-1}} \log\left(\frac{1-\epsilon_0-2\epsilon_1\kappa^{-1}}{12\epsilon}\right).$$

■

Appendix D. Proofs of Section 4

Proof of Theorem 7 If S is an ϵ_0-subspace embedding matrix w.r.t. $B(x^{(t)})$, then we have

$$(1-\epsilon_0)\nabla^2 F(x^{(t)}) \preceq [B(x^{(t)})]^T S^T S B(x^{(t)}) \preceq (1+\epsilon_0)\nabla^2 F(x^{(t)}). \tag{38}$$

By simple transformation and omitting ϵ_0^2, Eqn. (38) can be transformed into

$$(1-\epsilon_0)[B(x^{(t)})]^T S^T S \nabla^2 B(x^{(t)}) \preceq \nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0)[B(x^{(t)})]^T S^T S B(x^{(t)}).$$

The convergence rate can be derived directly from Theorem 3 and 5. ■

Proof of Corollary 8 If $\nabla^2 F(x)$ is not Lipschitz continuous, then we have

$$\begin{aligned}
\limsup_{t\to\infty} \frac{\|x^{(t+1)} - x^*\|_M}{\|x^{(t)} - x^*\|_M} &= \limsup_{t\to\infty} \left(\epsilon_0(t) + 2\nu(t)\mu^{-1} + \frac{2\kappa\mu^{-1}\nu(t)(1+\nu(t)\mu^{-1})}{1-\kappa\mu^{-1}\nu(t)}\right) \\
&= \limsup_{t\to\infty} \left(\frac{1}{\log(1+t)} + 2\nu(t)\mu^{-1} + \frac{2\kappa\mu^{-1}\nu(t)(1+\nu(t)\mu^{-1})}{1-\kappa\mu^{-1}\nu(t)}\right) \\
&= 0,
\end{aligned}$$

where $\nu(t) \to 0$ is because $\|\nabla^2 F(x^{(t)}) - \nabla^2 F(x^*)\| \to 0$ as $x^{(t)}$ approaches x^*.

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

$$\begin{aligned}
\limsup_{t\to\infty} \frac{\|x^{(t+1)} - x^*\|_M}{\|x^{(t)} - x^*\|_M} &\leq \limsup_{t\to\infty} \left(\epsilon_0(t) + +4(\kappa+1)\mu^{-3/2}\hat{L}\left\|x^{(t)} - x^*\right\|_M\right) \\
&= \limsup_{t\to\infty} \left(\frac{1}{\log(1+t)} + +4(\kappa+1)\mu^{-3/2}\hat{L}\left\|x^{(t)} - x^*\right\|_M\right) \\
&= 0.
\end{aligned}$$

■

Appendix E. Proofs of theorems of Section 5

Proof of Lemma 9 Let us denote that

$$X_i = [\nabla^2 F(x^{(t)})]^{-1/2} \nabla^2 f_i(x) [\nabla^2 F(x^{(t)})]^{-1/2}, \quad \text{and} \quad Y = \sum_{i \in \mathcal{S}} X_i.$$

Because $\nabla^2 f_i(x)$ is chosen uniformly, then we have $\mathbb{E}[Y] = \sum_{i \in \mathcal{S}} \mathbb{E}[X_i] = \mathcal{S}I$. Furthermore, by the Condition (12) and (13), we can obtain that

$$\|X_i\| \leq \frac{K}{\mu} \quad \text{and} \quad \lambda_{\max}(\mathbb{E}[Y]) = \lambda_{\min}(\mathbb{E}[Y]) = |\mathcal{S}|. \tag{39}$$

By Lemma 17, we have

$$\mathbb{P}\left(\lambda_{\min}(Y) \leq (1-\epsilon_0)|\mathcal{S}|\right) \leq d \exp\left(-\frac{\epsilon_0^2 |\mathcal{S}|}{2K/\mu}\right).$$

Letting us choose $|\mathcal{S}| = \frac{2K/\mu \log(d/\delta)}{\epsilon_0^2}$, then it holds with probability at least $1-\delta$ that

$$\lambda_{\min}(Y) \geq (1-\epsilon_0)|\mathcal{S}|,$$

which implies that

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} \frac{x^T [\nabla^2 F(x^{(t)})]^{-1/2} \left(\sum_{i \in \mathcal{S}} \nabla^2 f_i(x)\right) [\nabla^2 F(x^{(t)})]^{-1/2} x}{\|x\|^2} \geq (1-\epsilon_0)|\mathcal{S}| \\ \Rightarrow &\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla^2 f_i(x) \succeq (1-\epsilon_0) \nabla^2 F(x^{(t)}). \end{aligned}$$

By simple transformation and omitting ϵ_0^2, the above equation can be represented as

$$\nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0) H^{(t)}. \tag{40}$$

Also by Lemma 17, we have

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq (1+\epsilon_0)|\mathcal{S}|\right) \leq d \exp\left(-\frac{\epsilon_0^2 |\mathcal{S}|}{3K/\mu}\right).$$

By the similar proof of above, we can obtain that if we choose $|\mathcal{S}| = \frac{3K/\mu \log(d/\delta)}{\epsilon_0^2}$, it holds with probability at least $1-\delta$ that

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

Combining with Eqn. (40) and by the union bound of probability, we can obtain that if we choose $|\mathcal{S}| = \frac{3K/\mu \log(2d/\delta)}{\epsilon_0^2}$, it holds that

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

with probability at least $1-\delta$. ■

Proof of Theorem 10 Once the sample size $\mathcal{S}$ is properly chosen, the approximate Hessian in Eqn. (14) satisfies the condition (2). Then the local and global convergence properties of Algorithm 4 can be obtained by Theorem 3 and Theorem 5, respectively. ■

Proof of Lemma 12 Let us denote that

$$X_i = [\nabla^2 F(x^{(t)}) + \xi I]^{-1/2} \left(\nabla^2 f_i(x) + \xi I\right) [\nabla^2 F(x^{(t)}) + \xi I]^{-1/2}, \quad \text{and} \quad Y = \sum_{i \in \mathcal{S}} X_i.$$

Then we can obtain that

$$\|X_i\| \leq \frac{K + \xi}{\mu + \xi}$$

Because $\nabla^2 f_i(x)$ is chosen uniformly, then we have $\mathbb{E}[Y] = \sum_{i \in \mathcal{S}} \mathbb{E}[X_i] = \mathcal{S}I$. Hence, we can obtain that

$$\lambda_{\max}(\mathbb{E}[Y]) = \lambda_{\min}(\mathbb{E}[Y]) = \mathcal{S}. \tag{41}$$

By Lemma 17, we have

$$\mathbb{P}\left(\lambda_{\min}(Y) \leq \frac{2}{3}|\mathcal{S}|\right) \leq d \exp\left(-\frac{|\mathcal{S}|}{18(K+\xi)/(\mu+\xi)}\right).$$

Letting us choose $|\mathcal{S}| = \frac{18(K+\xi)\log(d/\delta)}{\mu+\xi}$, then it holds with probability at least $1-\delta$ that

$$\frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} \nabla^2 f_i(x) + \xi I \succeq \frac{2}{3}\left(\nabla^2 F(x^{(t)}) + \xi I\right) \succeq \frac{2}{3}\left(1 + \frac{\xi}{L}\right)\nabla^2 F(x^{(t)}), \tag{42}$$

which implies that

$$\nabla^2 F(x^{(t)}) \preceq \left(1 + \frac{L - 2\xi}{2(L+\xi)}\right) H^{(t)}.$$

Also by Lemma 17, we have

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq \frac{3}{2}|\mathcal{S}|\right) \leq d \exp\left(-\frac{|\mathcal{S}|}{12(K+\xi)/(\mu+\xi)}\right).$$

By the similar proof of above, we can obtain that if we choose $|\mathcal{S}| = \frac{12(K+\xi)\log(d/\delta)}{\mu+\xi}$, it holds with probability at least $1-\delta$ that

$$\frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} \nabla^2 f_i(x) + \xi I \preceq \frac{3}{2}\left(\nabla^2 F(x^{(t)}) + \xi I\right) \preceq \frac{3}{2}\left(1 + \frac{\xi}{\mu}\right)\nabla^2 F(x^{(t)}), \tag{43}$$

which implies that

$$\left(1 - \frac{3\xi + \mu}{3\alpha + 3\mu}\right) H^{(t)} \preceq \nabla^2 F(x^{(t)}).$$

Therefore, by choosing $|\mathcal{S}| = \frac{18(K+\xi)\log(2d/\delta)}{\mu+\xi}$, then it holds with probability at least $1-\delta$ that

$$\left(1 - \frac{3\xi + \mu}{3\xi + 3\mu}\right) H^{(t)} \preceq \nabla^2 F(x^{(t)}) \preceq \left(1 + \frac{L - 2\xi}{2(L+\xi)}\right) H^{(t)}.$$

■

Proof of Theorem 13 Once the sample size $\mathcal{S}$ is properly chosen, the approximate Hessian in Eqn. (15) satisfies the condition (2) with $\epsilon_0 = \max\left(\frac{3\xi+\mu}{3\xi+3\mu}, \frac{L-2\xi}{2(L+\xi)}\right)$. Then the local and global convergence properties of Algorithm 5 can be obtained by Theorem 3 and Theorem 5, respectively. ■

Proof of Lemma 14 Let us denote

$$H_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \nabla^2 f_i(x), \quad \text{and} \quad \tilde{H} = H_{\mathcal{S}} + \lambda_{r+1} I,$$

where λ_{r+1} is the $(r+1)$-th largest eigenvalue of $\nabla^2 F(x^{(t)})$. By the proof of Theorem 13 and Eqn. (42), if we choose $|\mathcal{S}| = \frac{18K\log(d/\delta)}{\lambda_{r+1}}$, then we have

$$H_{\mathcal{S}} \succeq \frac{2}{3}\nabla^2 F(x^{(t)}) - \frac{\lambda_{r+1}}{3} I. \tag{44}$$

Moreover, by Eqn. (43) and choosing $|\mathcal{S}| = \frac{12K\log(d/\delta)}{\lambda_{r+1}}$, we can obtain that

$$H_{\mathcal{S}} \preceq \frac{3}{2}\nabla^2 F(x^{(t)}) + \frac{\lambda_{r+1}}{2} I. \tag{45}$$

By Corollary 7.7.4 (c) of Horn and Johnson (2012), Eqn. (44) and (45) imply that

$$\frac{1}{3}\lambda_{r+1} \leq \lambda_{r+1}(H_{\mathcal{S}}) \leq 2\lambda_{r+1}. \tag{46}$$

Let us express the SVD of $H_{\mathcal{S}}^{(t)}$ as follows

$$H_{\mathcal{S}}^{(t)} = U\hat{\Lambda}U^T = U_r\hat{\Lambda}_r U_r^T + U_{\setminus r}\hat{\Lambda}_{\setminus r}U_{\setminus r}^T.$$

Then $H^{(t)}$ can be represented as

$$H^{(t)} = H_{\mathcal{S}} + U\begin{bmatrix} 0 & 0 \\ 0 & \lambda_{r+1}(H_{\mathcal{S}})I - \hat{\Lambda}_{\setminus r} \end{bmatrix} U^T.$$

By Eqn. (44) and $\frac{1}{3}\lambda_{r+1} \leq \lambda_{r+1}(H_{\mathcal{S}})$ (Eqn. (46)), we have

$$H^{(t)} \succeq \frac{2}{3}\nabla^2 F(x^{(t)}) - \frac{\lambda_{r+1}}{3} I + U\begin{bmatrix} 0 & 0 \\ 0 & \lambda_{r+1}(H_{\mathcal{S}})\cdot I - \hat{\Lambda}_{\setminus r} \end{bmatrix} U^T \succeq \frac{2}{3}\nabla^2 F(x^{(t)})$$

which implies that

$$\nabla^2 F(x^{(t)}) \preceq \left(1+\frac{1}{2}\right) H^{(t)}.$$

By Eqn. (45) and (46), we have

$$H^{(t)} \preceq \frac{3}{2}\nabla^2 F(x^{(t)}) + \frac{\lambda_{r+1}}{2} I + U\begin{bmatrix} 0 & 0 \\ 0 & \lambda_{r+1}(H_{\mathcal{S}})I - \hat{\Lambda}_{\setminus r} \end{bmatrix} U^T$$

$$\preceq \frac{3}{2}\nabla^2 F(x^{(t)}) + \frac{5}{2}\lambda_{r+1} I$$
$$\preceq \left(\frac{3}{2} + \frac{5\lambda_{r+1}}{2\mu}\right)\nabla^2 F(x^{(t)}),$$

which implies that

$$\left(1 - \frac{5\lambda_{r+1} + \mu}{5\lambda_{r+1} + 3\mu}\right) H^{(t)} \preceq \nabla^2 F(x^{(t)}).$$

Therefore, if choosing $|\mathcal{S}| = \frac{18K\log(2d/\delta)}{\lambda_{r+1}}$, we can obtain that

$$\left(1 - \frac{5\lambda_{r+1} + \mu}{5\lambda_{r+1} + 3\mu}\right) H^{(t)} \preceq \nabla^2 F(x^{(t)}) \preceq \left(1 + \frac{1}{2}\right) H^{(t)},$$

which concludes the proof. ■

Proof of Theorem 15 Once the sample size $\mathcal{S}$ is properly chosen, the approximate Hessian in Eqn. (17) satisfies the condition (2) with $\epsilon_0 = \max\left(\frac{5\lambda_{r+1}+\mu}{5\lambda_{r+1}+3\mu}, \frac{1}{2}\right)$. Then the local and global convergence properties of Algorithm 6 can be obtained by Theorem 3 and Theorem 5, respectively. ■

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

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

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

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (Dec):3475–3506, 2012.

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

N Halko, P G Martinsson, and J A Tropp. Finding Structure with Randomness : Probabilistic Algorithms for Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206), 1984.

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

Xiangrui Meng and Michael W Mahoney. 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*, pages 91–100. ACM, 2013.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 117–126. IEEE, 2013.

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

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

James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30. Siam, 1970.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

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

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

Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods. *Mathematical Programming*, 174(1-2):293–326, 2019.

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

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In *Advances in Neural Information Processing Systems*, pages 2332–2342, 2018.

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

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

Haishan Ye, Luo Luo, and Zhihua Zhang. Approximate newton methods and their local convergence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3931–3939. JMLR. org, 2017.

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