Journal of Machine Learning Research 21 (2020) 1-37 Submitted 4/19; Revised 5/20; Published 6/20

Nesterov's Acceleration for Approximate Newton

Haishan Ye HSYE_CS@OUTLOOK.COM
Shenzhen Research Insititution of Big Data
The Chinese University of Hong Kong, Shenzen
2001 Longxiang Road, Shenzhen, 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
National Engineering Lab for Big Data Analysis and Applications
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing, China

Editor: Qiang Liu

Abstract

Optimization plays a key role in machine learning. Recently, stochastic second-order methods have attracted considerable attention because of their low computational cost in each iteration. However, these methods might suffer from poor performance when the Hessian is hard to be approximate well in a computation-efficient way. To overcome this dilemma, we resort to Nesterov's acceleration to improve the convergence performance of these second-order methods and propose accelerated approximate Newton. We give the theoretical convergence analysis of accelerated approximate Newton and show that Nesterov's acceleration can improve the convergence rate. Accordingly, we propose an accelerated regularized sub-sampled Newton (ARSSN) which performs much better than the conventional regularized sub-sampled Newton empirically and theoretically. Moreover, we show that ARSSN has better performance than classical first-order methods empirically.

Keywords: Nesterov's Acceleration, Approximate Newton, Stochastic Second-order

1. Introduction

Optimization has become an increasingly important issue in machine learning. Many machine learning models can be reformulated as the following optimization problem:

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

where each f_i is the loss with respect to (w.r.t.) the i-th training sample. There are many examples such as logistic regressions, smoothed support vector machines, neural networks, and graphical models.

[*] Corresponding author.

In the era of big data, large-scale optimization is an important challenge. The stochastic gradient descent (SGD) method has been widely employed to reduce the computational cost per iteration (Cotter et al., 2011; Li et al., 2014; Robbins and Monro, 1951). However, SGD has poor convergence property. Hence, many variants have been proposed to improve the convergence rate of SGD (Johnson and Zhang, 2013; Roux et al., 2012; Schmidt et al., 2017; Zhang et al., 2013). At the same time, Nesterov's acceleration technique has become a very effective tool for first-order methods (Nesterov, 1983). It greatly improves the convergence rate of gradient descent (Nesterov, 1983) and stochastic gradient descent with variance reduction (Allen-Zhu, 2017; Lan and Zhou, 2018).

Recently, second-order methods have also received great attention due to their high convergence rate. However, conventional second-order methods are very costly because they take heavy computation to obtain the Hessian matrix. To conquer this weakness, one proposed a sub-sampled Newton which only selects a subset of functions f_i randomly to construct a sub-sampled Hessian (Roosta-Khorasani and Mahoney, 2016; Byrd et al., 2011; Xu et al., 2016). Meanwhile, 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) applied the sketching technique to alleviate the computational burden of computing Hessian and brought up *sketch Newton*. In fact, for many machine learning problems which take $f_i(x) = \ell(b_i, a_i^T x)$, where $\ell(\cdot, \cdot)$ is any convex smooth loss function and a_i is a data point, the sub-sampled Newton method can be regarded as a kind of sketch Newton because the Hessian can be expressed as $\nabla^2 F(x) = A^T D A = [D^{1/2} A]^T [D^{1/2} A]$, where D is a diagonal matrix with $D_{i,i} = \nabla^2 \ell(b_i, a_i^T x)$. Hence, when we refer to sketch Newton, we also include sub-sample Newton methods. The sketch Newton has its advantages when the size of the training set is sufficiently larger than the data dimension. In this case, one can approximate the Hessian efficiently via the sketching technique. Most importantly, such an approximate Hessian guarantees that the sketch Newton keeps a constant linear convergence rate independent of the condition number of the objective function (Ye et al., 2017).

However, the sketch Newton can not function properly because its prerequisite is not satisfied in many applications, where the number of training data is close to or even smaller than the actual size of data dimensions. To compensate for this gap, regularized sketch Newton methods were proposed (Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2016; Li et al., 2020). However, though adding a regularizer is an effective way to reduce the sketching size, a small sketching size will lead to a slow convergence rate. Ye et al. (2017) demonstrated that if approximate Hessian $H^{(t)}$ satisfies

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

where $0 < \eta < 1$, then the approximate Newton converges linearly with the rate η. Hence, one can sub-sample a small set of samples and constructs an approximate Hessian very efficiently but suffers from a slow convergence rate.

In this paper, we aim to establish balance between the computational efficiency of constructed approximate Hessian and the convergence rate. We resort to Nesterov's acceleration technique and propose the accelerated approximate Newton that one can construct an approximate Hessian efficiently while keeping a fast convergence rate. We will show that using Nesterov's acceleration technique, the convergence rate can be promoted from η to $1-\sqrt{1-\eta}$. Accordingly, we propose accelerated regularized sub-sampled Newton (ARSSN) which applies random sampling to construct approximate Hessian. ARSSN has a fast convergence rate but a low computational cost per iteration.

We summarize contribution of our work as follows:

- We introduce Nesterov's acceleration technique to improve the convergence rate of second-order methods (approximate Newton). This acceleration is of significance especially when the number of samples n and the feature dimension d are close to each other in which, it causes difficulties to construct a proper approximate Hessian with a low computational cost.
- Our theoretical analysis shows that by the acceleration technique, the convergence rate of the approximate Newton can be improved to $1 - \sqrt{1-\eta}$ from η with $0 < \eta < 1$ when the initial point is close to the optimal point. This result admits that accelerated approximate Newton can construct an approximate Hessian by stochastic methods very efficiently while enjoying a fast convergence rate.
- We propose Accelerated Regularized Sub-sampled Newton. Compared with classical first-order methods, ARSSN presents a better performance which demonstrates the computational efficiency of accelerated second-order methods. This is verified by the empirical study indicating the ability of the acceleration technique to improve approximate Newton methods effectively. In addition, the experiments also reveal a fact that adding curvature information properly can always improve the algorithm's convergence rate.

Organization. In the remainder of this paper, we introduce notation and preliminaries in Section 2. In Section 3, we describe accelerated approximate Newton method in detail and provide its local convergence analysis. In Section 4, we implement accelerated approximate Newton with random sampling and propose accelerated regularized sub-sampled Newton method. In Section 5, we conduct empirical studies on the accelerated approximate Newton and compare ARSSN with baseline algorithms to validate its computational efficiency. Finally, we conclude our work in Section 6. All proofs are provided in the appendices with the order of their appearance.

1.1. Related Work

Since Byrd et al. (2011) first proposed the sub-sampled Newton, stochastic second order methods have become a hot research topic. Erdogdu and Montanari (2015) gave the quantitive convergence rate of sub-sampled Newton and proposed a regularized sub-sampled Newton called NewSamp. Roosta-Khorasani and Mahoney (2016) proposed several new variants of sub-sampled Newton, including the methods which use sub-sampled Hessian and mini-batch gradient. Pilanci and Wainwright (2017) first used sketching techniques within the context of Newton-like methods. The authors proposed a randomized second-order method which performs an approximate Newton's step using a randomly sketched Hessian. Their algorithm is specialized to the case that the Hessian matrix can be expressed as $\nabla^2 F(x) = B^T(x)B(x)$ where $B(x) \in \mathbb{R}^{n\times d}$ with $n \gg d$ is readily available. In addition, combining stochastic Hessian with Taylors expansion, Agarwal et al. (2017) conceived a novel method named LiSSA, which is also termed as Newton-SGI (semi-stochastic gradient iteration) method by Bollapragada et al. (2019).

On the other hand, Nesterov's technique has been used in second-order methods with cubic regularization and the exact Hessian to improve the convergence rate (Nesterov, 2008). Without the strong convexity assumption, this method converges at the rate of $O(1/t^3)$. Monteiro and Svaiter (2013) proposed A-NPE which also used the exact Hessian. A-NPE had a convergence rate $O(1/t^{7/2})$ when the objective function is only convex but its Hessian is Lipschitz continuous (Monteiro and Svaiter, 2013). Very recently, accelerated second-order methods with cubic regularization under inexact Hessian information have been proposed (Ghadimi et al., 2017; Chen et al., 2018). It is also notable that the convergence rate shown in (Ghadimi et al., 2017) is different from our work. Once the approximate Hessian $H^{(t)}$ satisfies that $\left\|H^{(t)} - \nabla^2 f(x)\right\| \le \tau$ and the objective function is μ-strongly convex, the convergence rate in Ghadimi et al. (2017) is $1-\sqrt{\tau/\mu}$ which is no better than our result. Note that even η is of small value such as 0.1 in Eqn. (2), the value of $\left\|H^{(t)} - \nabla^2 f(x)\right\|$ can still be large since Eqn. (2) can only derive that $\left\|H^{(t)} - \nabla^2 f(x)\right\| \le \eta \left\|\nabla^2 f(x)\right\|$ and $\left\|\nabla^2 f(x)\right\|$ is commonly of large value in practice. Be aware of this problem, our theoretical result provides a much tighter convergence rate bound. However, we only provide a local convergence rate while Ghadimi et al. (2017); Chen et al. (2018) gave global convergence rates. Furthermore, accelerated second-order methods with cubic regularization should solve a costly sub-problem for each iteration. In contrast, the algorithm in this paper is neat and remains a similar form to accelerated first-order methods.

Nesterov and Stich (2017) and Tu et al. (2017) devised an accelerated block Gauss-Seidel method by introducing the acceleration technique to block Gauss-Seidel. Nesterov and Stich (2017) chose a fixed partitioning of the coordinates in advance while Tu et al. (2017) selected coordinates randomly. Both algorithms can be regarded as combining the Nesterov's acceleration with the random coordinate Newton method.

2. Notation and Preliminaries

We first introduce notation that will be used in this paper. Then, we give some assumptions on the objective function that will be used.

2.1. Notation

Given a matrix $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ of rank ℓ, its condensed SVD is given as $A = \sum_{i=1}^{\ell} \sigma_i u_i v_i^T$, where u_i and v_i are left and right singular vectors of A related to i-th largest singular value $\sigma_i > 0$. Accordingly, $\|A\| \triangleq \sigma_1$ is the spectral norm and $\|A\|_F \triangleq \sqrt{\sum_{i=1}^{\ell} \sigma_\ell^2}$ is the Frobenius norm. Using the spectral norm and Frobenius norm, we can define the stable rank of A as $\mathrm{sr}(A) \triangleq \frac{\|A\|_F^2}{\|A\|^2}$. If A is symmetric positive semi-definite, then u_i equals to v_i and the singular value decomposition of A is the identical to its eigenvalue decomposition. It also holds that $\lambda_i(A) = \sigma_i(A)$, where $\lambda_i(A)$ is the i-th largest eigenvalue of A. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalue of A, respectively. If A is a symmetric positive semi-definite matrix, we can define A-norm as $\|x\|_A = \sqrt{x^T A x}$. Furthermore, if matrix B is also symmetric positive semi-definite, we say $B \preceq A$ when $A - B$ is positive semi-definite.

2.2. Assumptions

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

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

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

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

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

Assumption 3 The Hessian $\nabla^2 F(x)$ is γ-Lipschitz continuous, that is,

$$\|\nabla^2 F(x) - \nabla^2 F(y)\| \le \gamma \|y - x\|, \text{ for } \gamma > 0.$$

or equivalently

$$-\frac{\gamma}{6}\|x-y\|^3 \le F(y) - \left[F(x) + \langle \nabla F(x), y - x\rangle + \frac{1}{2}\langle \nabla^2 F(x)(y-x), y-x\rangle\right] \le \frac{\gamma}{6}\|y-x\|^3. \quad (3)$$

By Assumptions 1 and 2, we define the condition number of function $F(x)$ as: $\kappa \triangleq \frac{L}{\mu}$.

2.3. Row Norm Squares Sampling

The row norm squares sampling matrix $S = D\Omega \in \mathbb{R}^{s\times n}$ w.r.t. $A \in \mathbb{R}^{n\times d}$ is determined by sampling probability p_i, a sampling matrix $\Omega \in \mathbb{R}^{s\times n}$ and a diagonal rescaling matrix $D \in \mathbb{R}^{s\times s}$. The sampling probability satisfies

$$p_i \ge \beta \frac{\|A_{i,:}\|^2}{\|A\|_F^2} \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1,$$

where $A_{i,:}$ means the i-th row of A and $0 < \beta$ is a constant. We construct S as follows. For every $j = 1, \dots, s$, independently and with replacement, pick an index i from the set $\{1, 2 \dots, n\}$ with probability p_i and set $\Omega_{ji} = 1$ and $\Omega_{jk} = 0$ for $k \neq i$ as well as $D_{jj} = 1/\sqrt{p_i s}$.

The row norm squares sampling matrix has the following important property.

Theorem 1 (Tropp et al. (2015)) *Let $S \in \mathbb{R}^{s\times n}$ be a row norm squares sampling matrix w.r.t. $A \in \mathbb{R}^{n\times d}$. If $s = O\left(\frac{\mathrm{sr}(A)\log(d/\delta)}{\beta c^2}\right)$, then it holds that*

$$\|A^T S^T S A - A^T A\| \le c \cdot \|A\|^2$$

with probability at least $1 - \delta$.

3. Accelerated Approximate Newton

In practice, it is common that the concerned problem is ill-conditioned and the number of the training data n and data dimension d are close to each other. In this case, the conventional sketch Newton and its regularized variants can not construct a desirable approximate Hessian in a computation efficient way while keeping a fast convergence rate. To conquer this dilemma, we take advantage of Nesterov's acceleration technique and propose *accelerated approximate Newton*. We describe the algorithmic procedure of accelerated approximate Newton as follows.

First, we construct an approximate Hessian $H^{(t)}$ satisfying

$$\begin{cases} (1-\eta)\left(\mathbb{E}\left[[H^{(t)}]^{-1}\right]\right)^{-1} \preceq \nabla^2 F(y^{(t)}) \preceq \left(\mathbb{E}\left[[H^{(t)}]^{-1}\right]\right)^{-1} \\ \left(\mathbb{E}\left[[H^{(t)}]^{-1}\right]\right)^{-1} + \nabla^2 F(y^{(t)}) \preceq 2H^{(t)} \end{cases} \quad (4)$$

for $0 < \eta < 1$. The first condition of Eqn. (4) is similar to the left one of Eqn. (2), but the second condition is much different from the right one of Eqn. (2). We will see that Condition (4) can be easily satisfied in practice with high probability in the next section.

Algorithm 1 Accelerated Approximate Newton.

```
Input: $x^{(0)}$ and $x^{(1)}$ are initial points sufficiently close to $x^*$; $\theta$ is the acceleration parameter.
for $t = 1, \dots$ until termination do
    Construct an approximate Hessian $H^{(t)}$ satisfying Eqn. (4);
    $y^{(t)} = x^{(t)} - \frac{1-\theta}{1+\theta}\left(x^{(t)} - x^{(t-1)}\right)$;
    $x^{(t+1)} = y^{(t)} - [H^{(t)}]^{-1}\nabla F(y^{(t)})$.
end for
```

Second, we update sequence $x^{(t)}$ as follows:

$$\begin{cases} y^{(t)} = x^{(t)} - \dfrac{1-\theta}{1+\theta}\left(x^{(t)} - x^{(t-1)}\right), \\ x^{(t+1)} = y^{(t)} - [H^{(t)}]^{-1}\nabla F(y^{(t)}), \end{cases} \tag{5}$$

where θ is chosen in terms of the value of η. It could be observed that the iteration (5) mostly resembles the update procedure of Nesterov's accelerated gradient descent where $[H^{(t)}]^{-1}$ is the counterpart of step size. If we set $\theta = 1$, the scheme (5) reduces to the update of approximate Newton (Ye et al., 2017). Thus, we refer to a class of methods satisfying Eqn. (4) and (5) as the *accelerated approximate Newton*.

We describe the accelerated approximate Newton in Algorithm 1 in detail.

3.1. Theoretical Analysis

In this section, we will give the local convergence properties of the accelerated approximate Newton (Algorithm 1). Let us denote

$$\bar{H}^{(t)} \triangleq \left(\mathbb{E}[(H^{(t)})^{-1}]\right)^{-1} \quad \text{and} \quad \bar{H}^* \triangleq \left(\mathbb{E}[(H^*)^{-1}]\right)^{-1},$$

where $H^{(t)}$ and H^* are the approximate Hessians constructed by stochastic methods at point $y^{(t)}$ and x^* respectively. We also denote

$$V^{(t)} \triangleq F(x^{(t)}) - F(x^*) + \frac{\theta^2}{2}\|x^* - v^{(t)}\|^2_{\bar{H}^\star}, \tag{6}$$

where

$$v^{(t)} = x^{(t-1)} + \frac{1}{\theta}(x^{(t)} - x^{(t-1)}), \quad \text{and} \quad \theta = \sqrt{1-\eta}.$$

We will prove that $V^{(t)}$ will almost decrease with rate $1 - \sqrt{1-\eta}$ in expectation. The following theorem gives a detailed description.

Theorem 2 *Let $F(x)$ be a convex function such that Assumptions 1 and 2 hold. Suppose that $\nabla^2 F(x)$ exists and Assumption 3 holds in a neighborhood of the minimizer x^*. The approximate Hessian $H^{(t)}$ satisfies Condition* (4). *Matrices $H^{(t+1)}$ and $H^{(t)}$ satisfy that $\|\bar{H}^{(t+1)} - \bar{H}^{(t)}\| \le \gamma\varsigma\|y^{(t+1)} - y^{(t)}\|$, where ς is a constant. Then, Algorithm 1 has the following convergence properties*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \gamma\varphi\left[V^{(t)}\right]^{3/2} \tag{7}$$

Method	Iterations to obtain ϵ-suboptimal	Condition	Reference
Approximate Newton	$O\left(\frac{1}{1-\eta} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	Eqn. (2)	Ye et al. (2017)
Accelerated Approximate Newton	$O\left(\frac{1}{\sqrt{1-\eta}} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	Eqn. (4)	Theorem 2

Table 1: Compare accelerated approximate Newton with approximate Newton

where expectation is taken with respect to $H^{(t)}$. And φ is defined as

$$\varphi = \frac{272\kappa^3 + 12(\theta^3 + \kappa^2)\varsigma}{(1+\theta)^3\mu^{3/2}} + \frac{32L(1+\theta^2)\varsigma}{(1+\theta)^3\mu^{5/2}}.$$

Remark 3 *From Eqn.* (7)*, we can observe that the accelerated approximate Newton method will converge super-linearly at the beginning. This is because the second term on the right hand of* (7) *dominates the convergence property at this phase. Once $V^{(t)}$ is small enough, then the accelerated approximate Newton method will turn into linear convergence with the rate $1-\theta$ because it holds that $\left[V^{(t)}\right]^{3/2} \ll V^{(t)}$ in this case. Our experiments validate such phenomenon. In fact, approximate Newton method has similar convergence properties while approximate Newton converges quadratically at the beginning (Pilanci and Wainwright, 2017; Xu et al., 2016; Erdogdu and Montanari, 2015) opposed to the superlinear rate of the accelerated approximate Newton.*

Remark 4 *We only provide a local convergence rate of accelerated approximate Newton in Theorem 2. To achieve a fast convergence rate, $x^{(t)}$ is required to enter into the local region close enough to the optima x^*. However, in real applications, this local region can be much large just as pointed out in Nesterov (2018): the region of quadratic convergence of the Newton method is almost the same as the region of the linear convergence of the gradient method. Therefore, we recommend to run stochastic gradient descent several iterations to find a good initial point $x^{(0)}$ just as LiSSA does (Agarwal et al., 2017), then use accelerated approximate Newton to obtain a high precision solution.*

Theorem 2 shows that Algorithm 1 converges Q-linearly with rate $1-\sqrt{1-\eta}$ in expectation, that is

$$\lim_{t\to\infty} \frac{\mathbb{E}\left[V^{(t+1)}\right]}{V^{(t)}} = (1-\theta) = 1-\sqrt{1-\eta}.$$

In contrast, with the approximate Hessian, the approximate Newton method can only achieve a η convergence rate. To obtain an ϵ-suboptimal solution, approximate Newton method needs $O\left(\frac{1}{1-\eta} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$ iterations while the complexity can be reduced to $O\left(\frac{1}{\sqrt{1-\eta}} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$ with the acceleration. This shows that the acceleration technique can effectively promote the convergence properties of the approximate Newton especially when η is close to one. The detailed comparisons between approximate Newton and accelerated approximate Newton are listed in Table 1.

3.2. Inexact Solution of Sub-problem

Algorithm 1 takes $[H^{(t)}]^{-1}\nabla F(y^{(t)})$ as the descent direction which is the solution of the following problem

$$\min_p \frac{1}{2}p^T H^{(t)} p - p^T \nabla F(y^{(t)}). \tag{8}$$

Algorithm 2 Accelerated Regularized Sub-sample Newton (ARSSN).

```
Input: initial points $x^{(0)}$ and $x^{(1)}$ sufficiently close to $x^*$, acceleration parameter $\theta$, and sample size $s$.
for $t = 1, \dots$ until termination do
    Select a sample set $\mathcal{S}$ of size $s$ by random sampling and construct $H^{(t)}$ of the form (10) satisfying Eqn. (4);
    $y^{(t)} = x^{(t)} - \frac{1-\theta}{1+\theta}\left(x^{(t)} - x^{(t-1)}\right)$;
    $x^{(t+1)} = y^{(t)} - [H^{(t)}]^{-1}\nabla F(y^{(t)})$
end for
```

In fact, an inexact solution of problem (8) can also work. If the direction vector $p^{(t)}$ satisfies

$$\|H^{(t)}p^{(t)} - \nabla F(y^{(t)})\| \le r\epsilon_0\|\nabla F(y^{(t)})\|, \tag{9}$$

where $0 \le \epsilon_0 < 1$ and r depends on the approximate Hessian, the inexactness of $p^{(t)}$ only affects the convergence rate at most with ϵ_0.

Theorem 5 *Let $F(x)$ and $H^{(t)}$ satisfy the properties described in Theorem 2. Assume the step 5 of Algorithm 1 to be replaced by $x^{(t+1)} = y^{(t)} - p^{(t)}$, where $p^{(t)}$ satisfies Equation* (9) *with $r \le \frac{\theta^2(1+\theta)^2}{8(1+2\theta^2)\kappa^2}$. Then, Algorithm 1 converges as*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta+\epsilon_0)V^{(t)} + \gamma\tilde{\varphi}\left[V^{(t)}\right]^{3/2}.$$

where expectation is taken with respect to $H^{(t)}$ and $\tilde{\varphi}$ is defined as

$$\tilde{\varphi} = \frac{1}{(1+\theta)^3\mu^{3/2}}\left(16\sqrt{2}(1+\theta^2)\kappa\varsigma + 16\sqrt{2}\kappa^2\varsigma + 12\theta^3\varsigma + 62\kappa^3\right).$$

Theorem 5 reveals that the precision of solution to problem (8) will affect the convergence rate directly. Hence, a high precision solution is preferred for the accelerated approximate Newton methods. In addition, Theorem 5 shows that when the approximate Hessian $H^{(t)}$ is of large size which causes difficulties to obtain the direct inversion of $H^{(t)}$, we only need to solve the problem (8) to get a descent vector by algorithms such as conjugate gradient method (Nocedal and Wright, 2006). Using this direction vector still guarantees that the accelerated approximate Newton methods will converge but the convergence rate would undergo minor perturbation.

4. Accelerated Regularized Sub-sampled Newton

Commonly, the number of training data n and the data dimension d are close to each other in real applications. Extant stochastic second-order methods such as the sub-sampled Newton and sketch Newton are not suitable because the sketching size or sampled size will be less than d, and the approximate Hessian is not invertible. Hence, adding a proper regularizer is a potential approach. Accordingly, regularized sub-sample Newton methods have been proposed (Roosta-Khorasani and Mahoney, 2016). To conquer the weakness of the low convergence rate of the regularized sub-sample Newton, we propose the accelerated regularized sub-sample Newton method (ARSSN) in Algorithm 2.

In Algorithm 2, $H^{(t)}$ is an approximation of $\nabla^2 F(y^{(t)})$ constructed via random sampling and has the following form

$$H^{(t)} = \hat{H}^{(t)} + \alpha^{(t)} I, \tag{10}$$

where $\hat{H}^{(t)}$ is the sub-sampled Hessian and $\alpha^{(t)} I$ is the regularizer. Note that $H^{(t)}$ specifies a way to construct the approximate Hessian by random sampling and gives rise to a kind of accelerated approximate Newton. Therefore, the convergence property of Algorithm 2 can be analyzed by Theorem 2.

We first focus on the explicit multiplication case where the Hessian $\nabla^2 F(x)$ satisfies the following structure

$$\nabla^2 F(x) = B(x)^T B(x),\ B(x) \in \mathbb{R}^{n\times d}. \tag{11}$$

Because the Hessian matrix can be represented by the multiplication of two explicit matrices, we call it explicit multiplication case.

Then we analyze `ARSSN` applied to the finite sum case with the Hessian of finite sum form, i.e.

$$\nabla^2 F(x) = \frac{1}{n}\nabla^2 f_i(x), \quad \text{with} \quad \nabla^2 f_i(x) \in \mathbb{R}^{d\times d}. \tag{12}$$

Note that there are many cases in which the finite sum case can also be formulated as an explicit multiplication case (11).

4.1. Explicit Multiplication Case

The explicit multiplication case (11) occurs frequently in machine learning applications who take $f_i(x) = \ell(b_i, a_i^T x)$, where $\ell(\cdot,\cdot)$ is any convex smooth loss function and a_i is a data point with b_i being the corresponding label. In this case, the Hessian can be expressed as

$$\nabla^2 F(x) = \frac{1}{n} A^T D(x) A = \underbrace{\left(\frac{1}{\sqrt{n}}\sqrt{D(x)}A\right)^T}_{B^T(x)} \underbrace{\left(\frac{1}{\sqrt{n}}\sqrt{D(x)}A\right)}_{B(x)}, \tag{13}$$

where $D(x) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $D_{i,i}(x) = \nabla^2 \ell(b_i, a_i^T x)$ and $A \in \mathbb{R}^{n\times d}$ is the data matrix with i-th row being a_i.

For the explicit multiplication case with the Hessian structure as Eqn. (11), we construct an approximate Hessian as

$$H^{(t)} = \underbrace{[S^{(t)} B^{(t)}]^T S^{(t)} B^{(t)}}_{\hat{H}^{(t)}} + \alpha^{(t)} I, \tag{14}$$

where $S^{(t)} \in \mathbb{R}^{s\times n}$ is a row norm squares random sampling matrix and $\alpha^{(t)}$ is a regularizer parameter.

Lemma 6 *Assume $\nabla^2 F(x)$ satisfies Eqn.* (11), *$S^{(t)} \in \mathbb{R}^{s\times n}$ is a row norm squares sampling matrix w.r.t. $B^{(t)}$ with $s = O(c^{-2}\cdot \mathrm{sr}(B^{(t)}) \log d/\delta)$, where $0 < c < 1$ is a constant and $0 < \delta < 1$ is the failure rate. Let approximate Hessian $H^{(t)}$ be constructed as Eqn.* (14) *with $\alpha^{(t)} = 2c\|B^{(t)}\|^2$. Then, Condition* (4) *holds with $\eta = \frac{3c\kappa}{1+3c\kappa}$ and probability at least $1-\delta$.*

Method	Time to reach ϵ-suboptimality	Applicable to n close to d ?	Reference
`Sketch Newton`	$O\left(nd + d^3\right)\log\left(\frac{1}{\epsilon}\right)$	No	Pilanci and Wainwright (2017)
`SSN` (leverage scores)	$\tilde{O}\left(nd + d^2\kappa^{3/2}\right)\log\left(\frac{1}{\epsilon}\right)$	No	Xu et al. (2016)
`SSN` (row norm squares)	$\tilde{O}\left(nd + \mathrm{sr}(B)d\kappa^{5/2}\right)\log\left(\frac{1}{\epsilon}\right)$	Yes	Xu et al. (2016)
`RSSN`	$\tilde{O}(\kappa n^{3/4}d(\mathrm{sr}(B))^2)\log\left(\frac{1}{\epsilon}\right)$	Yes	Derived from Ye et al. (2017)
`ARSSN`	$\tilde{O}(\sqrt{\kappa}n^{7/8}d(\mathrm{sr}(B))^2)\log\left(\frac{1}{\epsilon}\right)$	Yes	Theorem 8

Table 2: Compare `ARSSN` with previous works under explicit multiplication case (Eqn. 11). $\kappa = \frac{L}{\mu}$ is the condition number of the objective function. $\mathrm{sr}(B)$ is the stable rank of B with $\nabla^2 F(x) = B^T(x)B(x)$.

Remark 7 *Lemma 6 only analyzes the case that $S^{(t)}$ is the row norm squares sampling matrix. However, the same result still holds for the randomized sketching matrices such as count sketch matrix (Clarkson and Woodruff, 2013; Woodruff, 2014). The core property of row norm squares sampling is to make $\hat{H}^{(t)}$ satisfy $\left\|\hat{H}^{(t)} - \nabla^2 F(x^{(t)})\right\| \le c\left\|\nabla^2 F(x^{(t)})\right\|$ as described in Theorem 1. This result also holds for those randomized sketching matrices. In this paper, we focus on the row norm squares sampling matrix because it does not need to be explicitly constructed and the overhead of computing the sampling probabilities is only $\mathrm{nnz}(B(x))$ which is identical to constructing the gradient $\nabla F(x)$.*

Combining the properties of constructed approximate Hessian (14) described in Lemma 6 and the convergence properties of accelerated approximate Newton, we can obtain the convergence properties of `ARSSN` (Algorithm 2) under the explicit multiplication case.

Theorem 8 *Let $F(x)$ be a convex function under Assumptions 1 and 2. Suppose that $\nabla^2 F(x)$ exists and is of the form (11) in a neighborhood of the minimizer x^*. $S^{(t)} \in \mathbb{R}^{s\times n}$ is a row norm squares sampling matrix w.r.t. $B^{(t)}$ with $s = O(c^{-2}\cdot \mathrm{sr}(B^{(t)})\log d/\delta)$, where $0 < c < 1$ is sample size parameter and $0 < \delta < 1$ is the failure rate. Let us set regularizer $\alpha^{(t)} = 2c\|B^{(t)}\|^2$ and construct the approximate Hessian $H^{(t)}$ as Eqn. (14). Assume $\|B^{(t+1)} - B^{(t)}\| \le \frac{\gamma}{2\sqrt{L}}\|y^{(t+1)} - y^{(t)}\|$ for all $y^{(t+1)}$ and $y^{(t)}$. Letting $\eta = \frac{3c\kappa}{1+3c\kappa}$, where κ is the condition number of $\nabla^2 F(y^{(t)})$, then Algorithm 2 has the following convergence property*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \gamma\varphi\left[V^{(t)}\right]^{3/2}$$

with probability at least $1-\delta$. φ is defined as

$$\varphi = \frac{272\kappa^3 + 12(\theta^3 + \kappa^2)\varsigma}{(1+\theta)^3\mu^{3/2}} + \frac{32L(1+\theta^2)\varsigma}{(1+\theta)^3\mu^{5/2}}, \quad \textit{and}, \quad \varsigma = 9 + \frac{9n}{4cs}.$$

Theorem 8 shows that Algorithm 2 converges Q-linearly with rate $\left(1 - \frac{1}{\sqrt{1+3c\kappa}}\right)$ compared with $\left(1 - \frac{1}{1+3c\kappa}\right)$ of the regularized sub-sampled Newton (`RSSN`) with row norm squares sampling (Ye et al., 2017). Let us set $c = n^{-1/4}$, then the sample size s is about $n^{1/2}$ and Algorithm 2 takes about

$O(nd)$ time for each iteration. This matches the computational cost of computing gradient. In this case, the computational cost of Algorithm 2 needs

$$O(\sqrt{1+3c\kappa}nd)\log\left(\frac{1}{\epsilon}\right) \approx \tilde{O}(\sqrt{\kappa}n^{7/8}d\mathrm{sr}^2(B))\log\left(\frac{1}{\epsilon}\right),$$

while notation $\tilde{O}(\cdot)$ omits polynomial of $\log(d)$. In contrast, without acceleration, `RSSN` requires

$$O\left((1+3c\kappa)nd\right)\log\left(\frac{1}{\epsilon}\right) \approx \tilde{O}(\kappa n^{3/4}d\mathrm{sr}^2(B))\log\left(\frac{1}{\epsilon}\right)$$

times. Assuming that condition number κ is of the same order as number of the training sample which is common in practice, then `ARSSN` runs about $n^{3/8}$ times faster than `RSSN` because of

$$\frac{\kappa n^{3/4}d\mathrm{sr}^2(B)}{\sqrt{\kappa}n^{7/8}d\mathrm{sr}^2(B)} = \sqrt{\kappa}n^{-1/8} \approx n^{3/8}.$$

Therefore, `ARSSN` has considerable advantages over the regularized sub-sampled Newton especially when either the number of the training data or the condition number of the Hessian are large.

Furthermore, we compare `ARSSN` with previous work in Table 2. We find that `Sketch Newton` and sub-sampled Newton (`SSN`) with leverage score sampling can not be applied to the problem with n and d being close to each other. With well chosen sketching size s and regularization, they can be extended to solve the problem with n being close to d. However, the convergence of these regularized `Sketch Newton` and `SSN` with leverage score sampling will degrade similar to `RSSN` just as discussed in Remark 7. Then we compare `ARSSN` with `SSN` with row norm squares. The computational complexity of `SSN` with row norm squares sampling is linear to $O(d\kappa^{5/2})$. Once κ is of order n, we can observe that `SSN` with row norm squares sampling requires about $O(n^{9/8})$ times computational cost than `ARSSN`.

Furthermore, if the objective function is quadratic, that is the Lipschitz constant of the Hessian is zero, then the Hessian can be expressed as $\nabla^2 F(x) = B^T B,\ B \in \mathbb{R}^{n\times d}$. Let us construct an approximate Hessian as

$$H^{(t)} = [S^{(t)}B]^T S^{(t)}B + \alpha^{(t)}I, \tag{15}$$

where $S^{(t)}$ is a row norm squares random sampling matrix. Then we have the following corollary.

Corollary 9 *Let $F(x)$ be a convex and quadratic function subjected to Assumptions 1 and 2. $S^{(t)} \in \mathbb{R}^{s\times n}$ is a row norm squares sampling matrix w.r.t. B with $s = O(c^{-2}\cdot\mathrm{sr}(B)\log d/\delta)$, where $0 < c < 1$ is the sample size parameter and $0 < \delta < 1$ is the failure rate. Let us set regularizer $\alpha^{(t)} = 2c\|B\|^2$ and construct the approximate Hessian $H^{(t)}$ as Eqn.* (15)*. Letting $\eta = \frac{3c\kappa}{1+3c\kappa}$, where κ is the condition number of the Hessian, then Algorithm 2 has the following convergence property*

$$\mathbb{E}\left[V^{(t+1)}\right] \le \left(1 - \frac{1}{\sqrt{1+3c\kappa}}\right)V^{(t)},$$

with probability at least $1-\delta$.

Method	Time to reach ϵ-accurate solution	Reference
`NewSamp`	$\tilde{O}\left(\frac{\lambda_{k+1}}{K}n + \left(\frac{K}{\lambda_{k+1}}\right)^{3/2}\right) d\hat{\kappa}\log\left(\frac{1}{\epsilon}\right)$	Erdogdu and Montanari (2015)
`LiSSA`	$\tilde{O}\left(n + (\hat{\kappa}^{max})^2\hat{\kappa}\right) d\log\left(\frac{1}{\epsilon}\right)$	Agarwal et al. (2017)
`RSSN`	$\tilde{O}\left(\frac{\lambda_{k+1}}{K}n + \left(\frac{K}{\lambda_{k+1}}\right)^{3/2}\right) d\hat{\kappa}\log\left(\frac{1}{\epsilon}\right)$	Derived from Ye et al. (2017)
`ARSSN`	$\tilde{O}\left(\sqrt{\frac{\lambda_{k+1}}{K}}n + \left(\frac{K}{\lambda_{k+1}}\right)^{2}\right) d\sqrt{\hat{\kappa}}\log\left(\frac{1}{\epsilon}\right)$	Theorem 10

Table 3: Compare `ARSSN` with previous works under finite sum case (Eqn. 14). We assume that the gradient can be computed in $O(nd)$ and the Hessian-vector product $\nabla^2 f(x)v$ takes $O(d)$ computational cost. We also assume that $[H^{(t)}]^{-1}\nabla F(x)$ is computed by conjugated gradient. Condition number $\hat{\kappa}$ and $\hat{\kappa}^{max}$ are defined as $\hat{\kappa} = \frac{K}{\mu}$ and $\hat{\kappa}^{max} = \frac{K}{\min_i \lambda_{\min}(\nabla^2 f(x))}$, respectively. λ_{k+1} denotes the $(k+1)$-th largest eigenvalue of $\nabla^2 F(x)$. The notation $\tilde{O}(\cdot)$ omits the polynomial of $log(\cdot)$ terms.

The above corollary shows that Algorithm 2 converges linearly with rate $\left(1 - \frac{1}{\sqrt{1+3c\kappa}}\right)$ in expectation when the objective function is convex and quadratic. If we set sample size $s = \sqrt{n}$, then Algorithm 2 takes $O(nd)$ time for each iteration which is the same as accelerated gradient descent. By the above corollary, we can show that the convergence rate is

$$\rho = 1 - O\left(\frac{n^{1/8}}{\sqrt{\kappa}(\mathrm{sr}(B)\log d)^{1/4}}\right).$$

If the stable rank of B is small, then above convergence rate implies that Algorithm 2 is almost faster $n^{1/8}$ times than the accelerated gradient descent whose convergence rate is $\rho = 1 - O\left(\frac{1}{\sqrt{\kappa}}\right)$.

4.2. Finite-Sum Case

If the Hessian matrix satisfies the form (12), then we construct the approximate Hessian $H^{(t)}$ as follows

$$H^{(t)} = \underbrace{\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\nabla^2 f_i(x^{(t)})}_{\hat{H}^{(t)}} + \alpha^{(t)} I, \tag{16}$$

and we sample each $\nabla^2 f_i(x)$ uniformly. To analyze the convergence rate of `ARSSN` in finite sum case, we assume that each $f_i(x)$ and $F(x)$ in (1) have the following properties:

$$\max_{1\le i\le n} \|\nabla^2 f_i(x)\| \le K < \infty, \tag{17}$$

$$\lambda_{\min}(\nabla^2 F(x)) \ge \mu > 0. \tag{18}$$

In this case, we do not need the Hessian to be the specific form (14) but require each individual Hessian to be upper bounded.

Dataset	n	d	sparsity	source
gisette	$5,000$	$6,000$	dense	libsvm dataset
sido0	$12,678$	$4,932$	dense	Guyon
svhn	$19,082$	$3,072$	dense	libsvm dataset
rcv1	$20,242$	$47,236$	0.16%	libsvm dataset
real-sim	$72,309$	$20,958$	0.24%	libsvm dataset
avazu	$2,085,163$	$999,975$	0.0015%	libsvm dataset

Table 4: Datasets summary (sparsity= $\frac{\text{\#Non-Zero Entries}}{n\times d}$). We use the first $2,085,163$ training samples of the full avazu training set.

Similar to the explicit multiplication case, to make $H^{(t)}$ satisfy condition (4), we need to obtain $\left\|\hat{H}^{(t)} - \nabla^2 F(x^{(t)})\right\| \le c\left\|\nabla^2 F(x^{(t)})\right\|$ by the matrix Bernstein inequality. The following theorem describes the convergence properties of Algorithm 2 when the approximate Hessian is constructed as Eqn. (16).

Theorem 10 *Let $F(x)$ be a convex function under Assumptions 1 and 2. Suppose Eqn.* (17) *and* (18) *hold. We construct the approximate Hessian $H^{(t)}$ as Eqn.* (16) *by uniformly sampling with sample size $s = O(c^{-2}K^2\log(d/\delta))$ with $0 < c < 1$ and $0 < \delta < 1$. Let us set regularizer $\alpha^{(t)} = 2c$. Assume $\|\nabla^2 f_i(y^{(t+1)}) - \nabla^2 f_i(y^{(t)})\| \le \gamma\|y^{(t+1)} - y^{(t)}\|$ holds for all $i \in \{1,\dots,n\}$. Letting $\eta = \frac{2c}{\mu+2c}$, then Algorithm 2 has the following convergence property*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \gamma\varphi\left[V^{(t)}\right]^{3/2}$$

with probability at least $1-\delta$. φ is defined as

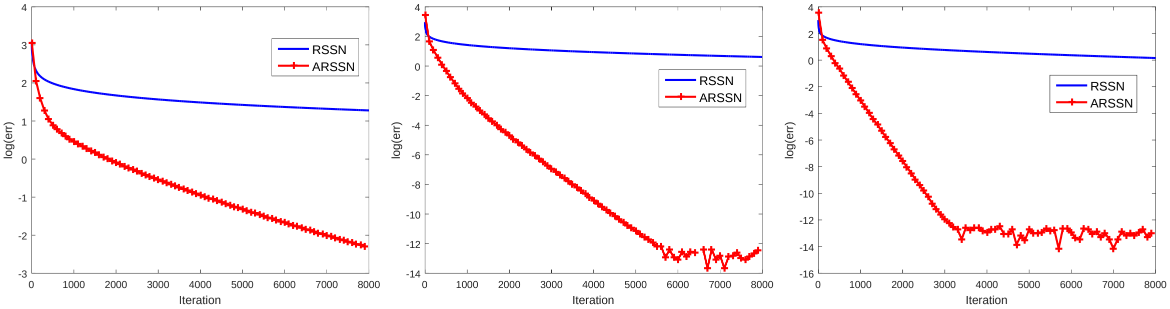
$$\varphi = \frac{272\kappa^3 + 12(\theta^3+\kappa^2)}{(1+\theta)^3\mu^{3/2}} + \frac{32L(1+\theta^2)}{(1+\theta)^3\mu^{5/2}}.$$

We compare ARSSN with previous work and list the detailed comparisons in Table 3. First, let us choose $c = \lambda_{k+1}$ and deploy conjugate gradient to solve $H^{(t)}\nabla F(x)$. Assuming that $\nabla^2 f_i(x)v$ can be computed in $O(d)$, then it takes $\tilde{O}\left(\left(\frac{K}{\lambda_{k+1}}\right)^{5/2} d\right)$ to approximate $H^{(t)}\nabla F(x)$ (This can be guaranteed by Theorem 5). Since Theorem 10 shows ARSSN converges at the rate of $1 - \sqrt{\frac{\mu}{\mu+2\lambda_{k+1}}}$, we can obtain the computational cost to achieve an ϵ-suboptimal solution as

$$\tilde{O}\left(nd + \left(\frac{K}{\lambda_{k+1}}\right)^{5/2} d\right)\sqrt{\frac{\lambda_{k+1}}{\mu}}\log\left(\frac{1}{\epsilon}\right) = \tilde{O}\left(\sqrt{\frac{\lambda_{k+1}}{K}}n + \left(\frac{K}{\lambda_{k+1}}\right)^2\right)d\sqrt{\hat{\kappa}}\log\left(\frac{1}{\epsilon}\right).$$

The computational cost of NewSamp and RSSN while with a convergence rate $1-\eta$ in contrast to $1-\sqrt{1-\eta}$ of ARSSN.

From the comparison in Table 3, we can observe that ARSSN obtain a faster convergence rate than NewSamp and RSSN. Furthermore, compared with LiSSA, we can observe that ARSSN outperforms LiSSA even $\hat{\kappa}$ is of order $n^{1/2}$ because LiSSA cubically depends on the $\hat{\kappa}$ ($\hat{\kappa}^{max} \ge \hat{\kappa}$).



(a) $|\mathcal{S}| = 1\%$

(b) $|\mathcal{S}| = 5\%$

(c) $|\mathcal{S}| = 10\%$

Figure 1: Experiment on the ridge regression with different sample sizes $|\mathcal{S}|$

5. Experiments

In this section, we will validate our theory empirically. We first compare accelerated regularized sub-sampled Newton (Algorithm 2) with regularized sub-sampled Newton (Algorithm 3 in Appendix called RSSN Roosta-Khorasani and Mahoney (2016)) on the ridge regression whose objective function is a quadratic function. Then we conduct more experiments on a popular machine learning problem called Ridge Logistic Regression, and compare accelerated regularized sub-sampled Newton with other classical methods.

All algorithms except SVRG are implemented in MATLAB. To achieve the best performance of SVRG (Johnson and Zhang, 2013), we implement it with C++ language. All experiments are conducted on a laptop with Intel i7-7700HQ CPU and 8GB RAM.

5.1. Experiments on the Ridge Regression

Th objective function of ridge regression is defined as

$$F(x) = \|Ax - b\|^2 + \lambda\|x\|^2, \tag{19}$$

where λ is a regularizer controlling the condition number of the Hessian. In our experiments, we choose dataset 'gisette' just as depicted in Table 4, and we set the regularizer $\lambda = 1$.

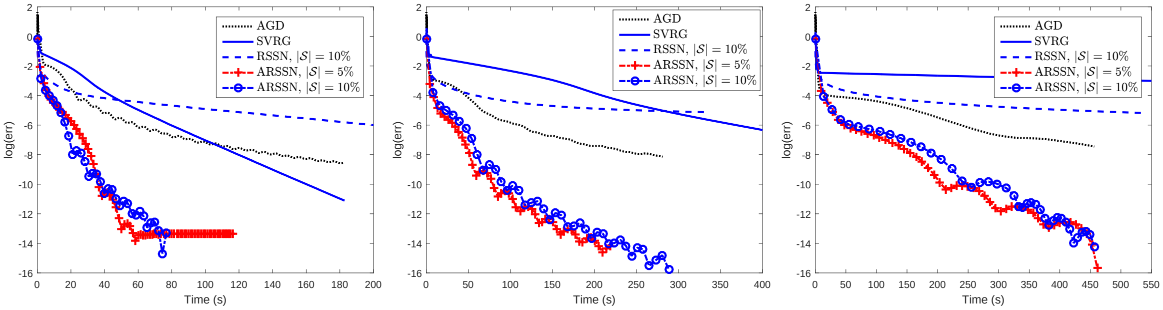
In the experiments, we set the sample size $|\mathcal{S}|$ to be $1\%n$, $5\%n$, and $10\%n$, respectively. The regularizer α of Algorithm 2 is properly chosen according to $|\mathcal{S}|$. ARSSN and RSSN share the same sample size $|\mathcal{S}|$. The acceleration parameter θ is appropriately selected and fixed. The experimental result is reported in Figure 1.

We can see that ARSSN runs much faster than RSSN from Figure 1. This agrees with our theoretical analysis. Furthermore, we can also observe that ARSSN converges faster as the sample size $|\mathcal{S}|$ increases. When $|\mathcal{S}| = 10\%n$, ARSSN takes only about 3000 iterations to achieve an 10^{-14} error while it needs about 6000 iterations to achieve the same precision when $|\mathcal{S}| = 5\%n$.

5.2. Experiments on the Ridge Logistic Regression

We conduct experiments on the Ridge Logistic Regression problem whose objective function is

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} \log[1 + \exp(-b_i\langle a_i, x\rangle)] + \frac{\lambda}{2}\|x\|^2, \tag{20}$$

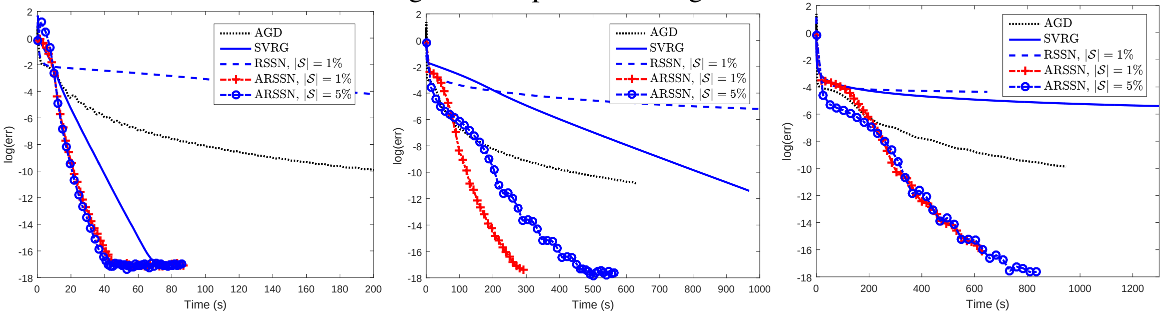


(a) $\lambda = 1/n$

(b) $\lambda = 10^{-1}/n$

(c) $\lambda = 10^{-2}/n$

Figure 2: Experiment on 'gisette'



(a) $\lambda = 1/n$

(b) $\lambda = 10^{-1}/n$

(c) $\lambda = 10^{-2}/n$

Figure 3: Experiment on 'sido0'

where $a_i \in \mathbb{R}^d$ is the i-th input vector, and $b_i \in \{-1, 1\}$ is the corresponding label.

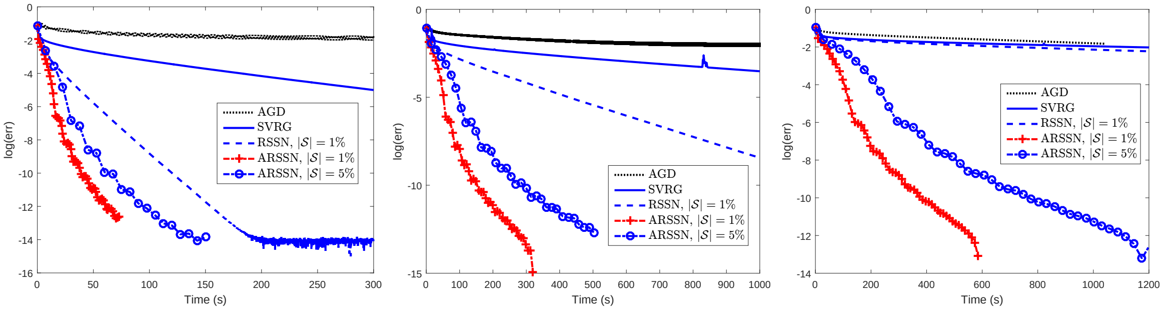
We conduct our experiments on six datasets: 'gisette', 'sido0', 'svhn', 'rcv1', 'real-sim', and 'avazu'. The first three datasets are dense and the last three ones are sparse. We give a detailed description of the datasets in Table 4. Notice that the size and dimension of the dataset are close to each other, so the sketch Newton method (Pilanci and Wainwright, 2017; Xu et al., 2016) can not be utilized. In our experiments, we try different settings of the regularizer λ, as $1/n$, $10^{-1}/n$, and $10^{-2}/n$ to represent different levels of regularization. Furthermore, in our experiments, we choose zero vector $x = [0, \dots, 0]^T \in \mathbb{R}^d$ as the initial point.

We compare `ARSSN` with `RSSN` (Algorithm 3 in appendix), `AGD` (Nesterov (1983)) and `SVRG` (Johnson and Zhang (2013)) which are classical and popular optimization methods in machine learning. In our experiments, the sample size $|\mathcal{S}|$ of `ARSSN` are chosen close to the square root of training sample size. Such $|\mathcal{S}|$ guarantees that the time complexity of computing the inverse of the approximate Hessian is similar to that of computing a full gradient. The regularizer $\alpha^{(t)}$ is chosen according to the sample size $|\mathcal{S}|$ and the norm of $B(x^{(t)})$ in Eqn. (11) by Theorem 8. In practice, the norm of $B(x^{(t)})$ is close to each other for different iterations. Hence, we pick a fixed α properly.

In our experiment, the sub-sampled Hessian $H^{(t)}$ constructed in Algorithm 2 can be written as

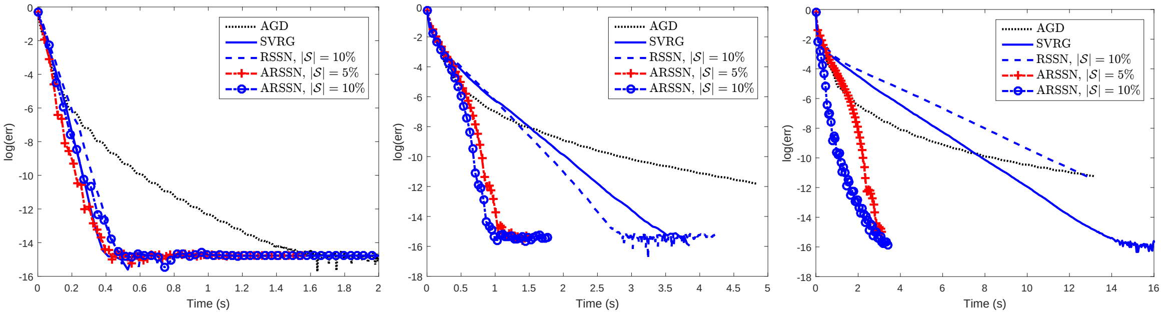
$$H^{(t)} = \tilde{A}^T \tilde{A} + (\alpha + \lambda) I,$$

where $\tilde{A} \in \mathbb{R}^{\ell \times d}$ and $\ell < n$. We can resort to Woodbury's identity to compute the inverse of $H^{(t)}$ at cost of $O(d\ell^2)$. In our experiments on sparse datasets, we use conjugate gradient to obtain an approximation of $[H^{(t)}]^{-1} \nabla F(x^{(t)})$ which exploits the sparsity of $\tilde{A}$.



(a) $\lambda = 1/n$ (b) $\lambda = 10^{-1}/n$ (c) $\lambda = 10^{-2}/n$

Figure 4: Experiment on ‘svhn’



(a) $\lambda = 1/n$ (b) $\lambda = 10^{-1}/n$ (c) $\lambda = 10^{-2}/n$

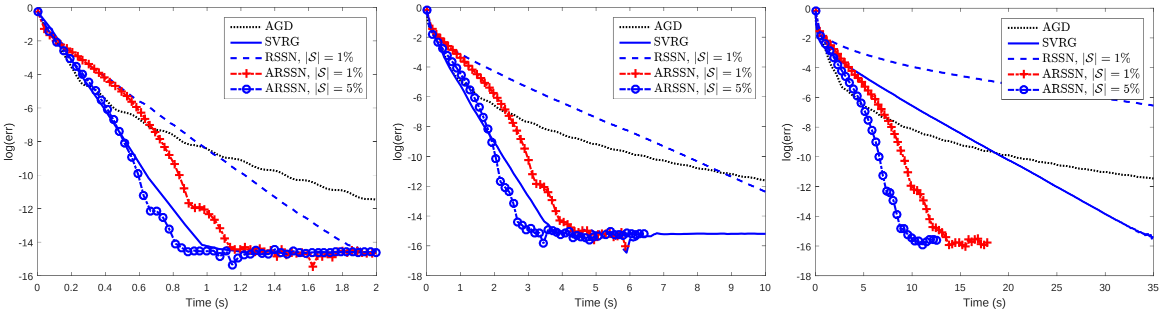
Figure 5: Experiment on ‘rcv1’

For the acceleration parameter θ, it is hard to get the best value for ARSSN just like AGD. However, our theoretical analysis implies that for large sample size $|\mathcal{S}|$, a small θ should be chosen. In our experiments, we empirically choose a proper θ.

We report our results in Figures 2 - 7 which illustrate that ARSSN converges significantly faster than RSSN when these two algorithms have the same sample size. This shows Nesterov’s acceleration technique can promote the performance of regularized sub-sampled Newton effectively. We can also observe that ARSSN outperforms AGD significantly even when the sample size $\mathcal{S}$ is $1\%n$ or even less. This reveals such a fact that adding some curvature information is an effective approach for improving the performance of accelerated gradient descent. It can also be observed that ARSSN converges superlinear at first then turns into a linear convergence. This verifies that the discussion in Remark 3.

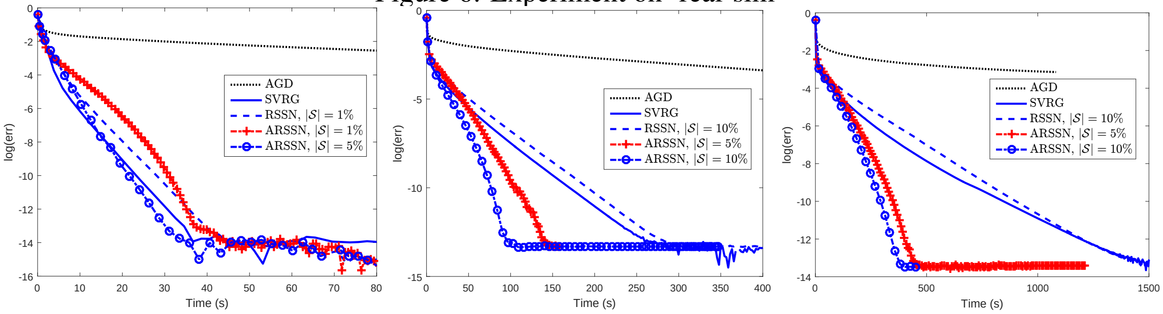
Furthermore, ARSSN outperforms SVRG, especially when λ is small like $\lambda = 10^{-2}/n$. When $\lambda = 10^{-1}/n$, ARSSN has comparable performance with SVRG on ‘sido0’, ‘rcv1’, ‘real-sim’, and ‘avazu’ while ARSSN has a much better performance on ‘gisette’ and ‘svhn’.

Moreover, the experiments reveal the fact that ARSSN has great advantages over other algorithms when the objective function is ill-conditioned. The advantages of ARSSN increase as the ragularizer λ decreasing and a small λ implies a large condition number of the objective function. Furthermore, on ‘gisette’, ‘sido0’, ‘svhn’, other algorithms have relatively poor performance in the case that $\lambda = 10^{-2}/n$. But ARSSN shows desirable convergence property. This is another evidence that ARSSN has advantages especially when the objective function is ill-conditioned.



(a) $\lambda = 1/n$ (b) $\lambda = 10^{-1}/n$ (c) $\lambda = 10^{-2}/n$

Figure 6: Experiment on 'real-sim'



(a) $\lambda = 1/n$ (b) $\lambda = 10^{-1}/n$ (c) $\lambda = 10^{-2}/n$

Figure 7: Experiment on 'avazu'

6. Conclusion

In this paper, we have exploited the acceleration technique to promote convergence rate of second order methods and proposed an framework named accelerated approximate Newton. We showed that accelerated approximate Newton has a much better convergence behavior when the approximate Hessian has low accuracy. We have also developed ARSSN algorithm based on the theory of accelerated approximate Newton, which enjoys a fast convergence rate than existing stochastic second-order optimization methods. Our experiments have shown that ARSSN performs much better than conventional RSSN, which meets our theory well. ARSSN also has several advantages over other classical algorithms which demonstrates the efficiency of accelerated second-order methods.

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). Haishan Ye has been supported by the project of Shenzhen Research Institute of Big Data (named 'Automated Machine Learning'). Luo Luo has been supported by GRF 16201320.

Algorithm 3 Regularized Sub-sample Newton (RSSN).

```
Input: $x^{(0)}$, $0 < \delta < 1$, regularizer parameter $\alpha$, sample size $|\mathcal{S}|$ ;
for $t = 0, 1, \dots$ until termination do
    Select a sample set $\mathcal{S}$, of size $|\mathcal{S}|$ and $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)}) + \alpha I$;
    Update $x^{(t+1)} = x^{(t)} - \left[H^{(t)}\right]^{-1} \nabla F(x^{(t)})$;
end for
```

Appendix A. Regularized Sub-sampled Newton

The regularized Sub-sampled Newton method is described in Algorithm 3. Now we give its local convergence properties in the following theorem.

Theorem 11 (Ye et al. (2017)) *Let $F(x)$ satisfy Assumption 1 and 2. Assume Eqns.* (17) *and* (18) *hold, and let $0 < \delta < 1$, $0 \le \epsilon_1 < 1$ and $0 < \alpha$ be given. Assume β is a constant such that $0 < \beta < \alpha + \frac{\sigma}{2}$, the subsampled size $|\mathcal{S}|$ satisfies $|\mathcal{S}| \ge \frac{16K^2 \log(2d/\delta)}{\beta^2}$, and $H^{(t)}$ is constructed as in Algorithm 3. Define*

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

which implies that $0 < \epsilon_0 < 1$. We define $\|x\|_{M^*} = \|[\nabla^2 F(x^*)]^{-\frac{1}{2}} x\|$. If $\nabla^2 F(x^{(t)})$ is γ-Lipschitz continuous and $x^{(t)}$ satisfies

$$\|x^{(t)} - x^*\| \le \frac{\mu}{\gamma\kappa}\nu(t),$$

where $0 < \nu(t) < 1$, then Algorithm 3 has the following convergence properties

$$\|\nabla F(x^{(t+1)})\|_{M^*} \le \epsilon_0 \frac{1 + \nu(t)}{1 - \nu(t)} \|\nabla F(x^{(t)})\|_{M^*} + \frac{2}{(1 - \epsilon_0)^2} \frac{\gamma\kappa}{\mu\sqrt{\mu}} \frac{(1 + \nu(t))^2}{1 - \nu(t)} \|\nabla F(x^{(t)})\|_{M^*}^2.$$

Appendix B. Proof of Theorem 2

Proof of Theorem 2 By the update procedure of algorithm, we can prove that the energy function $V^{(t+1)}$ (defined in Eqn. 6) will decrease with rate $1 - \theta$ compared with $V^{(t)}$ but with some perturbations denoted as $\Delta_1 + \Delta_2 + \Delta_3$, that is,

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1 - \theta)V^{(t)} + \Delta_1 + \Delta_2 + \Delta_3.$$

This result is proved in Lemma 12.

Next, we will show that these perturbations are high-order terms compared with the energy function $V^{(t)}$. In Lemma 13, we first give the upper bound of $\left\|x^{(t)} - x^*\right\|$ and $\left\|y^{(t)} - x^*\right\|$. Using these two bounds, in Lemma 14, we show that $\Delta_1 + \Delta_2 + \Delta_3$ is upper bounded as

$$\Delta_1 + \Delta_2 + \Delta_3 \le \frac{\gamma}{(1 + \theta)^3}\left(\frac{272\kappa^3 + 12(\theta^3 + \kappa^2)\varsigma}{\mu^{3/2}} + \frac{32L(1 + \theta^2)\varsigma}{\mu^{5/2}}\right)\left[V^{(t)}\right]^{3/2}.$$

Combining the results of two lemmas, we can obtain that

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \frac{\gamma}{(1+\theta)^3}\left(\frac{272\kappa^3 + 12(\theta^3+\kappa^2)\varsigma}{\mu^{3/2}} + \frac{32L(1+\theta^2)\varsigma}{\mu^{5/2}}\right)\left[V^{(t)}\right]^{3/2}.$$

■

In the rest of this section, we will give the detailed descriptions and proofs of Lemma 12, Lemma 13 and Lemma 14.

Lemma 12 *Let $F(x)$ be a convex function such that Assumptions 1 and 2 hold. Suppose that $\nabla^2 F(x)$ exists and is γ-Lipschitz continuous in a neighborhood of the minimizer x^*. Let $H^{(t)}$ be a random matrix approximating $\nabla^2 F(y^{(t)})$ satisfying Eqn.* (4)*. We set $\theta = \sqrt{1-\eta}$. Then, sequence $\{x^{(t)}\}$ of Algorithm 1 has the following property*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \Delta_1 + \Delta_2 + \Delta_3,$$

where expectation is taken with respect to $H^{(t)}$. Δ_1, Δ_2, and Δ_3 are defined as

$$\begin{aligned}
\Delta_1 =& \frac{\gamma}{6}\left(\|[H^{(t)}]^{-1}\nabla F(y^{(t)})\|^3 + (1-\theta)\|x^{(t)} - y^{(t)}\|^3 + \theta\|x^* - y^{(t)}\|^3\right),\\
\Delta_2 =& \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}^*} - \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}^{(t)}} + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\bar{H}^*} - \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\bar{H}^{(t)}},\\
\Delta_3 =& \langle\left(\bar{H}^* - \bar{H}^{(t)}\right)[\bar{H}^{(t)}]^{-1}\nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle.
\end{aligned}$$

Proof For notation convenience, we denote $H = H^{(t)}$. We also denote $\hat{H} = \nabla^2 F(y^{(t)})$ and $\bar{H} = \left(\mathbb{E}[H^{-1}]\right)^{-1}$. By the update procedure of algorithm, we have

$$\begin{aligned}
&F(y^{(t)} - H^{-1}\nabla F(y^{(t)}))\\
\overset{(3)}{\le}& F(y^{(t)}) - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\hat{H}} + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3\\
\overset{(3)}{\le}& F(z) - \langle\nabla F(y^{(t)}), z - y^{(t)}\rangle - \frac{1}{2}\left\|z - y^{(t)}\right\|^2_{\hat{H}} - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\\
&+ \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\hat{H}} + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3 + \frac{\gamma}{6}\|z - y^{(t)}\|^3\\
\le& F(z) - \langle\nabla F(y^{(t)}), z - y^{(t)}\rangle - \frac{1-\eta}{2}\|z - y^{(t)}\|^2_{\bar{H}} - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\\
&+ \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\hat{H}} + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3 + \frac{\gamma}{6}\|z - y^{(t)}\|^3. \qquad (21)
\end{aligned}$$

Inequality (21) is due to the condition $(1-\eta)\bar{H} \preceq \hat{H}$ in Eqn. (4). Let us denote $\theta = \sqrt{1-\eta}$. Then, we have the following inequality

$$\begin{aligned}
F(x^{(t+1)}) \le& F(z) - \langle\nabla F(y^{(t)}), z - y^{(t)}\rangle + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2_{\hat{H}} - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\\
&- \frac{\theta^2}{2}\|z - y^{(t)}\|^2_{\bar{H}} + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3 + \frac{\gamma}{6}\|z - y^{(t)}\|^3.
\end{aligned}$$

Setting $z = x^{(t)}$, $z = x^*$ respectively in above inequality, we obtain

$$
\begin{aligned}
&(1-\theta)F(x^{(t+1)})\\
\leq&(1-\theta)\bigg(F(x^{(t)}) - \langle \nabla F(y^{(t)}), x^{(t)} - y^{(t)}\rangle + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 - \frac{\theta^2}{2}\|x^{(t)} - y^{(t)}\|_{\bar{H}}^2\\
&- \langle \nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3 + \frac{\gamma}{6}\|x^{(t)} - y^{(t)}\|^3\bigg),
\end{aligned}
$$

and

$$
\begin{aligned}
\theta F(x^{(t+1)}) \leq&\theta\bigg(F(x^*) - \langle \nabla F(y^{(t)}), x^* - y^{(t)}\rangle + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 - \frac{\theta^2}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2\\
&- \langle \nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle + \frac{\gamma}{6}\|H^{-1}\nabla F(y^{(t)})\|^3 + \frac{\gamma}{6}\|x^* - y^{(t)}\|^3\bigg).
\end{aligned}
$$

Adding the above two inequalities and definition of Δ_1, we get

$$
\begin{aligned}
&F(x^{(t+1)}) - F(x^*)\\
\leq&(1-\theta)(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 - \langle \nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\\
&- \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2 - \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle - \frac{(1-\theta)\theta^2}{2}\|x^{(t)} - y^{(t)}\|_{\bar{H}}^2 + \Delta_1\\
\leq&(1-\theta)(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 - \langle \nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\\
&- \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle - \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2 + \Delta_1. \qquad (22)
\end{aligned}
$$

By the update iteration in Algorithm 2, we have

$$
v^{(t+1)} = x^{(t)} + \frac{1}{\theta}(x^{(t+1)} - x^{(t)}),
$$

and

$$
y^{(t)} = \frac{1}{1+\theta}(x^{(t)} + \theta v^{(t)}). \qquad (23)
$$

Thus, it holds that

$$
\theta^2 v^{(t+1)} = (1-\theta)\theta^2 v^{(t)} + \theta^3 y^{(t)} - \theta H^{-1}\nabla F(y^{(t)}). \qquad (24)
$$

Now, we build the relation between $\|x^\star - v^{(t+1)}\|_{\bar{H}^*}$ and $\|x^\star - v^{(t)}\|_{\bar{H}^*}$. First, we have

$$
\begin{aligned}
&\frac{\theta^2}{2}(\|x^\star - v^{(t+1)}\|_{\bar{H}^*}^2 - \|y^{(t)} - v^{(t+1)}\|_{\bar{H}^*}^2)\\
=&\frac{\theta^2}{2}(\|x^\star\|_{\bar{H}^*}^2 - \|y^{(t)}\|_{\bar{H}^*}^2 - 2\langle x^\star - y^{(t)}, \bar{H}^* v^{(t+1)}\rangle)\\
\overset{(24)}{=}&\frac{(1-\theta)\theta^2 + \theta^3}{2}(\|x^\star\|_{\bar{H}^*}^2 - \|y^{(t)}\|_{\bar{H}^*}^2) - \langle x^\star - y^{(t)}, \bar{H}^*[(1-\theta)\theta^2 v^{(t)} + \theta^3 y^{(t)} - \theta H^{-1}\nabla F(y^{(t)})]\rangle
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{(1-\theta)\theta^2}{2}(\|x^\star - v^{(t)}\|_{\bar{H}^*}^2 - \|y^{(t)} - v^{(t)}\|_{\bar{H}^*}^2) + \frac{\theta^3}{2}\|x^\star - y^{(t)}\|_{\bar{H}^*}^2 \\
&+ \theta\langle x^* - y^{(t)}, \bar{H}^* H^{-1}\nabla F(y^{(t)})\rangle.
\end{aligned}
$$

Then, we expand $\|y^{(t)} - v^{(t+1)}\|_{\bar{H}^*}^2$ as follows

$$
\begin{aligned}
&\frac{\theta^2}{2}\|y^{(t)} - v^{(t+1)}\|_{\bar{H}^*}^2 \\
=&\frac{1}{2\theta^2}\|\theta^2 y^{(t)} - ((1-\theta)\theta^2 v^{(t)} + \theta^3 y^{(t)} - \theta H^{-1}\nabla F(y^{(t)}))\|_{\bar{H}^*}^2 \\
\overset{(24)}{=}&\frac{1}{2}\|(1-\theta)\theta(y^{(t)} - v^{(t)}) - H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 \\
=&\frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 + \frac{\theta^2(1-\theta)^2}{2}\|y^{(t)} - v^{(t)}\|_{\bar{H}^*}^2 - (1-\theta)\theta\langle y^{(t)} - v^{(t)}, \bar{H}^* H^{-1}\nabla F(y^{(t)})\rangle \\
=&\frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 + \frac{\theta^2(1-\theta)^2}{2}\|y^{(t)} - v^{(t)}\|_{\bar{H}^*}^2 + (1-\theta)\langle x^{(t)} - y^{(t)}, \bar{H}^* H^{-1}\nabla F(y^{(t)})\rangle.
\end{aligned}
$$

Therefore, using two above equations, we have

$$
\begin{aligned}
&\frac{\theta^2}{2}\|x^* - v^{(t+1)}\|_{\bar{H}^*}^2 \\
=&\frac{\theta^2}{2}(\|x^\star - v^{(t+1)}\|_{\bar{H}^*}^2 - \|y^{(t)} - v^{(t+1)}\|_{\bar{H}^*}^2) + \frac{\theta^2}{2}\|y^{(t)} - v^{(t+1)}\|_{\bar{H}^*}^2 \\
=&\frac{(1-\theta)\theta^2}{2}\|x^\star - v^{(t)}\|_{\bar{H}^*}^2 + \frac{\theta^3}{2}\|x^\star - y^{(t)}\|_{\bar{H}^*}^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 \\
&+ \langle \bar{H}^* H^{-1}\nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle - \frac{\theta^3(1-\theta)}{2}\|y^{(t)} - v^{(t)}\|_{\bar{H}^*}^2.
\end{aligned}
$$

By the definition of $V^{(t)}$ (defined in Eqn. (6)) and Eqn. (22), taking expectation with respect to H, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[V^{(t+1)}\right] \\
\leq&(1-\theta)V^{(t+1)} + \Delta_1 - \langle\nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle \\
&+ \langle\bar{H}^*\mathbb{E}[H^{-1}]\nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle + \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}^*}^2 - \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2 \\
&+ \mathbb{E}\left[\frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\right] \\
=&(1-\theta)V^{(t)} + \Delta_1 + \langle\left(\bar{H}^* - \bar{H}\right)\bar{H}^{-1}\nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle \\
&+ \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}^*}^2 - \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}^*}^2 - \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}}^2 \\
&+ \mathbb{E}\left[\frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}}^2 - \langle\nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle\right] \\
\leq&(1-\theta)V^{(t)} + \Delta_1 + \Delta_2 + \Delta_3,
\end{aligned}
$$

where the last inequality is because of the condition (4) that

$$
\hat{H} + \bar{H} - 2H \preceq 0,
$$

and

$$\begin{aligned}
&\frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\hat{H}}^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|_{\bar{H}}^2 - \langle \nabla F(y^{(t)}), H^{-1}\nabla F(y^{(t)})\rangle \\
=&\frac{1}{2}\left(H^{-1}\nabla F(y^{(t)})\right)^T\left(\hat{H}+\bar{H}-2H\right)\left(H^{-1}\nabla F(y^{(t)})\right) \\
\leq&0.
\end{aligned}$$

■

Now, we begin to show that the perturbation terms in Lemma 12 are high-order terms compared with $V^{(t)}$. First, we list some important results in the following lemma.

Lemma 13 *Let $F(x)$ satisfy the properties described in Lemma 12. Sequences $\{x^{(t)}\}$ and $\{y^{(t)}\}$ satisfy that*

$$\|x^{(t)} - x^*\|^2 \leq \frac{2}{\mu}\left(F(x^{(t)}) - F(x^*)\right) \tag{25}$$

and

$$\|y^{(t)} - x^*\| \leq \frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}}. \tag{26}$$

Proof By the definition of μ-strongly convex, we have

$$F(x^{(t)}) \geq F(x^*) + \left\langle \nabla F(x^*), x^{(t)} - x^*\right\rangle + \frac{\mu}{2}\left\|x^{(t)} - x^*\right\|^2.$$

Combining the fact that $\nabla F(x^*) = 0$, we can obtain Eqn. (25).

By Eqn. (23), we have

$$\|y^{(t)} - x^*\| = \|\frac{1}{1+\theta}\left(x^{(t)} + \theta v^{(t)}\right) - x^*\| \leq \frac{1}{1+\theta}\|x^{(t)} - x^*\| + \frac{\theta}{1+\theta}\|v^{(t)} - x^*\|.$$

We also have

$$\|v^{(t)} - x^*\|^2 \leq \frac{1}{\lambda_{\min}(\bar{H}^*)}\|x^* - v^{(t)}\|_{\bar{H}^*}^2.$$

Due to the condition (4) which implies $\hat{H}^* \preceq \bar{H}^*$, where $\hat{H}^* = \nabla^2 F(x^*)$, and the problem is μ-strongly convex, we have

$$\frac{1}{\lambda_{\min}(\bar{H}^*)} \leq \frac{1}{\lambda_{\min}(\hat{H}^*)} \leq \frac{1}{\mu} < \frac{2}{\mu}.$$

Combining above results, we can bound $\|y^{(t)} - x^*\|$ as

$$\begin{aligned}
\|y^{(t)} - x^*\| \leq&\frac{\sqrt{2}}{(1+\theta)\sqrt{\mu}}\left(\sqrt{F(x^{(t)}) - F(x^*)} + \theta\|x^* - v^{(t)}\|_{\bar{H}^*}\right) \\
\leq&\frac{2}{(1+\theta)\sqrt{\mu}}\sqrt{F(x^{(t)}) - F(x^*) + \theta^2\|x^* - v^{(t)}\|_{\bar{H}^*}^2} \\
\leq&\frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}},
\end{aligned}$$

where the second inequality is because $(a+b)^2 \leq 2a^2 + 2b^2$. ■

Lemma 14 *Let $F(x)$ satisfy the properties described in Lemma 12. Matrices $H^{(t+1)}$ and $H^{(t)}$ satisfy that $\|\bar{H}^{(t+1)} - \bar{H}^{(t)}\| \le \gamma\varsigma\|y^{(t+1)} - y^{(t)}\|$, where ς is a constant. Then, we have*

$$\Delta_1 + \Delta_2 + \Delta_3 \le \frac{\gamma}{(1+\theta)^3}\left(\frac{272\kappa^3 + 12(\theta^3+\kappa^2)\varsigma}{\mu^{3/2}} + \frac{32L(1+\theta^2)\varsigma}{\mu^{5/2}}\right)\left[V^{(t)}\right]^{3/2}.$$

Proof We will bound the value of Δ_1, Δ_2 and Δ_3 sequentially. First, we are going to bound Δ_1.

By the L-smoothness property, we can bound $\|\nabla F(y^{(t)})\|$ as follows

$$\|\nabla F(y^{(t)})\| = \left\|\nabla F(y^{(t)}) - \nabla F(x^*)\right\| \le L\|y^{(t)} - x^*\| \le \frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}}. \tag{27}$$

We also have

$$\begin{aligned}\left\|x^{(t)} - y^{(t)}\right\| =& \left\|x^{(t)} - x^* + \left(x^* - y^{(t)}\right)\right\| \le \left\|x^{(t)} - x^*\right\| + \left\|x^* - y^{(t)}\right\| \\ \overset{(25)}{\le}& \sqrt{\frac{2}{\mu}\left(F(x^{(t)}) - F(x^*)\right)} + \left\|x^* - y^{(t)}\right\| \\ \overset{(26)}{\le}& \sqrt{\frac{2}{\mu}\left(F(x^{(t)}) - F(x^*)\right)} + \frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \\ \le& \frac{4\sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}}, \end{aligned} \tag{28}$$

where the last inequality follows from the fact that $F(x^{(t)}) - F(x^*) \le V^{(t)}$.

Combining above results, we can bound Δ_1 as follows.

$$\begin{aligned}\Delta_1 =& \frac{\gamma}{6}\left(\|H^{-1}\nabla F(y^{(t)})\|^3 + (1-\theta)\|x^{(t)} - y^{(t)}\|^3 + \theta\|x^* - y^{(t)}\|^3\right) \\ \le& \frac{\gamma}{6}\left(\frac{1}{\lambda_{\min}^3(H^{(t)})}\left\|\nabla F(y^{(t)})\right\|^3 + (1-\theta)\|x^{(t)} - y^{(t)}\|^3 + \theta\|x^* - y^{(t)}\|^3\right) \\ \overset{(27),(28),(26)}{\le}& \frac{\gamma[V^{(t)}]^{3/2}}{6}\left(\frac{1}{\lambda_{\min}^3(H^{(t)})}\left(\frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\right)^3 + (1-\theta)\left(\frac{4\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3 + \theta\left(\frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3\right) \\ \le& \frac{\gamma[V^{(t)}]^{3/2}}{6}\left(\frac{L^3}{\lambda_{\min}^3(H^{(t)})}\cdot\left(\frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3 + \left(\frac{4\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3\right) \\ =& \frac{(2\sqrt{2})^3\gamma[V^{(t)}]^{3/2}}{6(1+\theta)^3\mu^{3/2}}\left(\frac{L^3}{\lambda_{\min}^3(H^{(t)})} + 8\right), \end{aligned}$$

where the last inequality is because it holds for all $0 \le \theta \le 1$ that $(1-\theta)a + \theta b \le \max\{a, b\}$. Using the second condition of (4), we have $\nabla^2 F(y^{(t)}) \preceq 2H^{(t)}$. Combining $\mu \le \lambda_{\min}(\nabla^2 F(y^{(t)}))$ and the fact $L/\lambda_{\min}(H^{(t)}) \ge 1$, we can obtain that

$$\begin{aligned}\Delta_1 \le& \frac{(2\sqrt{2})^3\gamma[V^{(t)}]^{3/2}}{6(1+\theta)^3\mu^{3/2}}\left(\frac{L^3}{\lambda_{\min}^3(H^{(t)})} + 8\right) \le \frac{9(2\sqrt{2})^3\gamma[V^{(t)}]^{3/2}}{6(1+\theta)^3\mu^{3/2}}\cdot\frac{2^3L^3}{\mu^3} \\ =& \frac{192\sqrt{2}\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}}\cdot[V^{(t)}]^{3/2} \le \frac{272\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}}\cdot[V^{(t)}]^{3/2}. \end{aligned}$$

Next, we begin to bound Δ_2. By the condition $\|\bar{H}^{(t+1)} - \bar{H}^{(t)}\| \le \gamma\varsigma\|y^{(t+1)} - y^{(t)}\|$, we have

$$\begin{aligned}
\Delta_2 \le& \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2\|\bar{H} - \bar{H}^*\| + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2\|\bar{H} - \bar{H}^*\| \\
\le& \gamma\varsigma\left(\frac{\theta^3}{2}\|x^* - y^{(t)}\|^2 + \frac{1}{2}\|H^{-1}\nabla F(y^{(t)})\|^2\right)\|x^* - y^{(t)}\| \\
\le& \gamma\varsigma\left(\frac{\theta^3}{2} + \frac{L^2\|H^{-1}\|^2}{2}\right)\|x^* - y^{(t)}\|^3 \\
\overset{(26)}{\le}& \frac{12(\theta^3 + \kappa^2)\gamma\varsigma}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

Finally, we begin to bound Δ_3. We have

$$\begin{aligned}
\left\|(1-\theta)x^{(t)} + \theta x^* - y^{(t)}\right\| \overset{(23)}{=}& \left\|x^{(t)} - \frac{1}{1+\theta}(x^{(t)} + \theta v^{(t)}) - \theta(x^{(t)} - x^*)\right\| \\
=& \left\|\frac{\theta^2}{1+\theta}(x^{(t)} - x^*) - \frac{\theta}{1+\theta}(v^{(t)} - x^*)\right\| \\
\le& \frac{\theta^2}{1+\theta}\cdot\|x^{(t)} - x^*\| + \frac{\theta}{1+\theta}\cdot\|v^{(t)} - x^*\| \\
\le& \frac{\sqrt{2}\theta^2}{(1+\theta)\sqrt{\mu}}\sqrt{F(x^{(t)}) - F(x^*)} + \frac{\theta}{1+\theta}\sqrt{\frac{1}{\mu}}\|v^{(t)} - x^*\|_{\bar{H}^*} && (29) \\
\le& \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}}. && (30)
\end{aligned}$$

Inequality (29) is because of $\|v^{(t)} - x^*\| \le \lambda_{\min}^{-1/2}(\bar{H}^*)\|v^{(t)} - x^*\|_{\bar{H}^*}$ and $1/\lambda_{\min}(\bar{H}^*) \le 1/\mu$. Thus, we have

$$\begin{aligned}
\Delta_3 \le& \|\left(\bar{H}^* - \bar{H}\right)\bar{H}^{-1}\nabla F(y^{(t)})\| \cdot \left\|\left((1-\theta)x^{(t)} + \theta x^* - y^{(t)}\right)\right\| \\
\overset{(30)}{\le}& \|\bar{H}^{-1}\| \cdot \|\bar{H}^* - \bar{H}\| \cdot \|\nabla F(y^{(t)})\| \cdot \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \\
\overset{(27)}{\le}& \|\bar{H}^{-1}\| \cdot \frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \gamma\varsigma\|y^{(t)} - x^*\| \\
\overset{(26)}{\le}& \frac{32L(1+\theta^2)\gamma\varsigma}{(1+\theta)^3\mu^{5/2}}\left[V^{(t)}\right]^{3/2},
\end{aligned}$$

where the first inequality is due to Cauchy's inequality and the last inequality also uses the fact $\|\bar{H}^{-1}\| = 1/\lambda_{\min}(\bar{H}) \le 2/\mu$.

Therefore, we obtain

$$\begin{aligned}
\Delta_1 + \Delta_2 + \Delta_3 \le& \left(\frac{272\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}} + \frac{12(\theta^3 + \kappa^2)\gamma\varsigma}{(1+\theta)^3\mu^{3/2}} + \frac{32L(1+\theta^2)\gamma\varsigma}{(1+\theta)^3\mu^{5/2}}\right)\left[V^{(t)}\right]^{3/2} \\
=& \gamma\left(\frac{272\kappa^3 + 12(\theta^3 + \kappa^2)\varsigma}{(1+\theta)^3\mu^{3/2}} + \frac{32L(1+\theta^2)\varsigma}{(1+\theta)^3\mu^{5/2}}\right)\left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

■

Appendix C. Proof of Theorem 5

The proof of Theorem 5 is close to the one of Theorem 2, so we will omit some detailed steps in the following proof.

Lemma 15 *Let $F(x)$ and $H^{(t)}$ satisfy the properties described in Theorem 2. Assume the step 5 of Algorithm 1 be replaced by $x^{(t+1)} = y^{(t+1)} - p^{(t+1)}$. $p^{(t)}$ satisfies Equation* (9)*. Then, sequence $\{x^{(t)}\}$ of Algorithm 1 has the following property*

$$\mathbb{E}\left[V^{(t+1)}\right] \le (1-\theta)V^{(t)} + \Delta_1 + \Delta_2 + \Delta_3,$$

where expectation is taken with respect to $H^{(t)}$. Δ_1, Δ_2, and Δ_3 are defined as

$$\begin{aligned}
\Delta_1 =& \frac{\gamma}{6}\left(\|p^{(t)}\|^3 + (1-\theta)\left\|x^{(t)} - y^{(t)}\right\|^3 + \theta\left\|x^* - y^{(t)}\right\|^3\right),\\
\Delta_2 =& \mathbb{E}\langle \bar{H}^* p^{(t)}, (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle - \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle,\\
\Delta_3 =& \mathbb{E}\left[\frac{1}{2}\|p^{(t)}\|^2_{\nabla^2 F(y^{(t)})} + \frac{1}{2}\|p^{(t)}\|^2_{\bar{H}^*} - \langle \nabla F(y^{(t)}), p^{(t)}\rangle\right] + \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}^*} - \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}^{(t)}}.
\end{aligned}$$

Proof For notation convenience, we denote $H = H^{(t)}$. We also denote $\hat{H} = \nabla^2 F(y^{(t)})$ and $\bar{H} = \left(\mathbb{E}[H^{-1}]\right)^{-1}$. Now, we have

$$\begin{aligned}
&F(y^{(t)} - p^{(t)})\\
\le& F(y^{(t)}) - \langle \nabla F(y^{(t)}), p^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} + \frac{\gamma}{6}\|p^{(t)}\|^3\\
\le& F(z) - \langle \nabla F(y^{(t)}), z - y^{(t)}\rangle - \frac{1-\eta}{2}\|z - y^{(t)}\|^2_{\bar{H}} - \langle \nabla F(y^{(t)}), p^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}}\\
&+ \frac{\gamma}{6}\left(\|p^{(t)}\|^3 + \left\|z - y^{(t)}\right\|^3\right).
\end{aligned}$$

Then, we have the following inequality

$$\begin{aligned}
F(x^{(t+1)}) \le& F(z) - \langle \nabla F(y^{(t)}), z - y^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} - \langle \nabla F(y^{(t)}), p^{(t)}\rangle - \frac{\theta^2}{2}\|z - y^{(t)}\|^2_{\bar{H}}\\
&+ \frac{\gamma}{6}\left(\|p^{(t)}\|^3 + \left\|z - y^{(t)}\right\|^3\right).
\end{aligned}$$

Setting $z = x^{(t)}$, $z = x^*$ respectively in above inequality, we obtain

$$\begin{aligned}
(1-\theta)F(x^{(t+1)}) \le (1-\theta)\Big(& F(x^{(t)}) - \langle \nabla F(y^{(t)}), x^{(t)} - y^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} - \frac{\theta^2}{2}\|x^{(t)} - y^{(t)}\|^2_{\bar{H}}\\
&- \langle \nabla F(y^{(t)}), p^{(t)}\rangle + \frac{\gamma}{6}\left(\|p^{(t)}\|^3 + \left\|x^{(t)} - y^{(t)}\right\|\right)\Big)
\end{aligned}$$

and

$$\theta F(x^{(t+1)}) \le \theta\Big(F(x^*) - \langle \nabla F(y^{(t)}), x^* - y^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} - \frac{\theta^2}{2}\|x^* - y^{(t)}\|^2_{\bar{H}}$$

$$- \langle \nabla F(y^{(t)}), p^{(t)} \rangle + \frac{\gamma}{6}\left(\|p^{(t)}\|^3 + \left\| x^* - y^{(t)} \right\|^3 \right) \Bigg).$$

Adding the above two inequalities and using the notation of Δ_1, we get

$$\begin{aligned}
&F(x^{(t+1)}) - F(x^*) \\
\leq &(1-\theta)(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} - \langle \nabla F(y^{(t)}), p^{(t)} \rangle \\
&- \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)} \rangle - \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}} + \Delta_1.
\end{aligned}$$

By the update iteration in Algorithm 2, we have

$$\theta^2 v^{(t+1)} = (1-\theta)\theta^2 v^{(t)} + \theta^3 y^{(t)} - \theta p^{(t)}. \tag{31}$$

Now, we build the relation between $\|x^\star - v^{(t+1)}\|_{\bar{H}^*}$ and $\|x^\star - v^{(t)}\|_{\bar{H}^*}$.

First, we have

$$\begin{aligned}
&\frac{\theta^2}{2}(\|x^\star - v^{(t+1)}\|^2_{\bar{H}^*} - \|y^{(t)} - v^{(t+1)}\|^2_{\bar{H}^*}) \\
=&\frac{(1-\theta)\theta^2}{2}(\|x^\star - v^{(t)}\|^2_{\bar{H}^*} - \|y^{(t)} - v^{(t)}\|^2_{\bar{H}^*}) + \frac{\theta^3}{2}\|x^\star - y^{(t)}\|^2_{\bar{H}} + \theta\langle x^* - y^{(t)}, \bar{H}^* p^{(t)} \rangle.
\end{aligned}$$

Then, we expand $\frac{\theta^2}{2}\|y^{(t)} - v^{(t+1)}\|^2_{\bar{H}^*}$ as follows

$$\begin{aligned}
&\frac{\theta^2}{2}\|y^{(t)} - v^{(t+1)}\|^2_{\bar{H}^*} \\
=&\frac{1}{2\theta^2}\|\theta^2 y^{(t)} - ((1-\theta)\theta^2 v^{(t)} + \theta^3 y^{(t)} - \theta p^{(t)})\|^2_{\bar{H}^*} \\
=&\frac{1}{2}\|p^{(t)}\|^2_{\bar{H}^*} + \frac{\theta^2(1-\theta)^2}{2}\|y^{(t)} - v^{(t)}\|^2_{\bar{H}^*} + (1-\theta)\langle x^{(t)} - y^{(t)}, \bar{H}^* p^{(t)} \rangle,
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\frac{\theta^2}{2}\|x^* - v^{(t+1)}\|^2_{\bar{H}^*} =&\frac{(1-\theta)\theta^2}{2}\|x^\star - v^{(t)}\|^2_{\bar{H}^*} + \frac{\theta^3}{2}\|x^\star - y^{(t)}\|^2_{\bar{H}^*} + \frac{1}{2}\|p^{(t)}\|^2_{\bar{H}^*} \\
&+ \langle \bar{H}^* p^{(t)}, (1-\theta)x^{(t)} + \theta x^* - y^{(t)} \rangle - \frac{\theta^3(1-\theta)}{2}\|y^{(t)} - v^{(t)}\|^2_{\bar{H}^*}.
\end{aligned}$$

Hence, taking expectation with respect to H, we obtain

$$\begin{aligned}
&\mathbb{E}\left[F(x^{(t+1)}) - F(x^*) + \frac{\theta^2}{2}\|x^* - v^{(t+1)}\|^2_{\bar{H}^*} \right] \\
\leq&(1-\theta)\left((F(x^{(t)}) - F(x^*) + \frac{\theta^2}{2}\|x^\star - v^{(t)}\|^2_{\bar{H}^*} \right) + \Delta_1 \\
&+ \mathbb{E}\langle \bar{H}^* p^{(t)}, (1-\theta)x^{(t)} + \theta x^* - y^{(t)} \rangle - \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)} \rangle \\
&+ \mathbb{E}\left[\frac{1}{2}\|p^{(t)}\|^2_{\hat{H}} + \frac{1}{2}\|p^{(t)}\|^2_{\bar{H}^*} - \langle \nabla F(y^{(t)}), p^{(t)} \rangle \right] + \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}^*} - \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2_{\bar{H}}
\end{aligned}$$

$$\leq (1-\theta)\left((F(x^{(t)}) - F(x^*) + \frac{\theta^2}{2}\|x^\star - v^{(t)}\|^2_{\bar{H}^*}\right) + \Delta_1 + \Delta_2 + \Delta_3,$$

where the last equality follows from the definition of Δ_2 and Δ_3. ■

Now, we begin to bound the value of Δ_1, Δ_2, and Δ_3.

Lemma 16 *Let $F(x)$ and $H^{(t)}$ satisfy the properties described in Theorem 2. Assume the step 5 of Algorithm 1 be replaced by $x^{(t+1)} = y^{(t)} - p^{(t)}$ and $p^{(t)}$ satisfies Equation* (9) *with $r \leq \frac{\theta^2(1+\theta)^2}{8(1+2\theta^2)\kappa^2}$. Furthermore, matrices $H^{(t+1)}$ and $H^{(t)}$ satisfy that $\|\bar{H}^{(t+1)} - \bar{H}^{(t)}\| \leq \gamma\varsigma\|y^{(t+1)} - y^{(t)}\|$, where ς is a constant. Then, we have*

$$\Delta_1 + \Delta_2 + \Delta_3 \leq \epsilon_0 \cdot V^{(t)} + \gamma \cdot \tilde{\varphi} \cdot \left[V^{(t)}\right]^{3/2},$$

where $\tilde{\varphi}$ is defined as

$$\tilde{\varphi} = \frac{1}{(1+\theta)^3\mu^{3/2}}\left(16\sqrt{2}(1+\theta^2)\kappa\varsigma + 16\sqrt{2}\kappa^2\varsigma + 12\theta^3\varsigma + 31\kappa^3\right).$$

Proof We first bound the value of Δ_1. We have

$$\begin{aligned}
\Delta_2 =& \mathbb{E}\langle \bar{H}^* p^{(t)}, (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle - \langle \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\rangle \\
\leq& \mathbb{E}\left\langle \bar{H}^*\left(p^{(t)} - H^{-1}\nabla F(y^{(t)}\right) + \bar{H}^* H^{-1}\nabla F(y^{(t)}) - \nabla F(y^{(t)}), (1-\theta)x^{(t)} + \theta x^* - y^{(t)}\right\rangle \\
\leq& \|(1-\theta)x^{(t)} + \theta x^* - y^{(t)}\| \cdot \left(\mathbb{E}\|p^{(t)} - H^{-1}\nabla F(y^{(t)}\| \cdot \|\bar{H}^*\| + \|\bar{H}^*\bar{H}\nabla F(y^{(t)}) - \nabla F(y^{(t)})\|\right) \\
\overset{(30),(9)}{\leq}& \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \left(r\epsilon_0 \cdot (\mathbb{E}\|H^{-1}\|) \cdot \|\bar{H}^*\| \cdot \|\nabla F(y^{(t)}\| + \|\bar{H}^*\bar{H}\nabla F(y^{(t)}) - \nabla F(y^{(t)})\|\right) \\
\leq& \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \left(\frac{2r\epsilon_0 L}{\theta^2\mu}\|\nabla F(y^{(t)}\| + \|\bar{H}^*\bar{H}\nabla F(y^{(t)}) - \nabla F(y^{(t)})\|\right),
\end{aligned}$$

where the last inequality is because of condition (4) that $\nabla^2 F(y^{(t)}) \preceq 2H^{(t)}$ and $\bar{H}^* \preceq \frac{1}{1-\eta}\nabla^2 F(x^*)$. Furthermore, we have

$$\begin{aligned}
\|\bar{H}^*\bar{H}\nabla F(y^{(t)}) - \nabla F(y^{(t)})\| \leq& \|\bar{H}^{-1}\| \cdot \|\bar{H}^* - \bar{H}\| \cdot \|\nabla F(y^{(t)})\| \\
\overset{(27)}{\leq}& \frac{2}{\mu} \cdot \frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \gamma\varsigma\|y^{(t)} - x^*\| \\
\overset{(26)}{\leq}& \frac{16\kappa\gamma\varsigma}{(1+\theta)^2\mu}V^{(t)}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\Delta_2 \leq& \frac{\sqrt{2}\theta^2 + \sqrt{2}}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} \cdot \left(\frac{2r\epsilon_0\kappa}{\theta^2} \cdot \frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\sqrt{V^{(t)}} + \frac{16\kappa\gamma\varsigma}{(1+\theta)^2\mu}V^{(t)}\right) \\
=& \frac{8r\epsilon_0(1+\theta^2)\kappa^2}{\theta^2(1+\theta)^2} \cdot V^{(t)} + \frac{16\sqrt{2}\gamma(1+\theta^2)\kappa\varsigma}{(1+\theta)^3\mu^{3/2}} \cdot \left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

Now, we begin to bound the value of $\|p^{(t)}\|$ which will be used in bounding the value of Δ_2 and Δ_3. We give the bound of $\|p^{(t)}\|$ as follows

$$\begin{aligned}\|p^{(t)}\| =&\|H^{-1}\nabla F(y^{(t)}) - (p^{(t)} - H^{-1}\nabla F(y^{(t)}))\|\\ \le&\|H^{-1}\nabla F(y^{(t)})\| + \|(p^{(t)} - H^{-1}\nabla F(y^{(t)}))\|\\ \le&\|H^{-1}\| \left(\|\nabla F(y^{(t)})\| + \|H^{(t)}p^{(t)} - \nabla F(y^{(t)})\|\right)\\ \le&(1+r\epsilon_0)\|H^{-1}\| \cdot \|\nabla F(y^{(t)})\|\\ \le&\frac{4\sqrt{2}\kappa}{(1+\theta)\sqrt{\mu}}\left[V^{(t)}\right]^{1/2},\end{aligned} \tag{32}$$

where the last inequality is because of Eqn. (26) and condition (4) that $\nabla^2 F(y^{(t)}) \preceq 2H^{(t)}$ and $\mu \le \lambda_{\min}(\nabla^2 F(y^{(t)}))$.

For the value of Δ_3, we bound its first term as follows

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\|p^{(t)}\|_{\hat{H}}^2 + \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2 - \langle\nabla F(y^{(t)}), p^{(t)}\rangle\right]\\ =&\mathbb{E}\left[\frac{1}{2}\|p^{(t)}\|_{\hat{H}}^2 + \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2 - \|p^{(t)}\|_H^2 + \|p^{(t)}\|_H^2 - \langle\nabla F(y^{(t)}), p^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|_{\bar{H}^*}^2 - \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2\right]\\ \le&\mathbb{E}\left[\|p^{(t)}\|_H^2 - \langle\nabla F(y^{(t)}), p^{(t)}\rangle + \frac{1}{2}\|p^{(t)}\|_{\bar{H}^*}^2 - \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2\right]\\ =&\mathbb{E}\left[\langle p^{(t)}, Hp^{(t)} - \nabla F(y^{(t)}) + \frac{1}{2}\|p^{(t)}\|_{\bar{H}^*}^2 - \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2\rangle\right]\\ \le&\mathbb{E}\left[r\epsilon_0 \cdot \|p^{(t)}\| \cdot \|\nabla F(y^{(t)})\| + \frac{1}{2}\|p^{(t)}\|_{\bar{H}^*}^2 - \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2\right],\end{aligned}$$

where the first inequality is because of the condition (4) that $\hat{H} + \bar{H} - 2H \preceq 0$ and

$$\frac{1}{2}\|p^{(t)}\|_{\hat{H}}^2 + \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2 - \|p^{(t)}\|_H^2 = \frac{1}{2}\left(p^{(t)}\right)^T\left(\hat{H} + \bar{H} - 2H\right)\left(p^{(t)}\right) \le 0.$$

Hence, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\|p^{(t)}\|_{\hat{H}}^2 + \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2 - \langle\nabla F(y^{(t)}), p^{(t)}\rangle\right]\\ \le&\mathbb{E}\left[r\epsilon_0 \cdot \|p^{(t)}\| \cdot \|\nabla F(y^{(t)})\| + \frac{1}{2}\|p^{(t)}\|_{\bar{H}^*}^2 - \frac{1}{2}\|p^{(t)}\|_{\bar{H}}^2\right]\\ \overset{(32)(27)}{\le}&\mathbb{E}\left[r\epsilon_0\frac{4\sqrt{2}\kappa}{(1+\theta)\sqrt{\mu}}\left[V^{(t)}\right]^{1/2} \cdot \frac{2\sqrt{2}L}{(1+\theta)\sqrt{\mu}}\left[V^{(t)}\right]^{1/2} + \frac{1}{2}\|p^{(t)}\|^2 \cdot \|\bar{H}^* - \bar{H}\|\right]\\ \overset{(32)}{\le}&\frac{16r\epsilon_0\kappa^2}{(1+\theta)^2}V^{(t)} + \frac{8\kappa^2}{(1+\theta)^2\mu}V^{(t)} \cdot \gamma\varsigma\|y^{(t)} - x^*\|\\ \overset{(26)}{\le}&\frac{16r\epsilon_0\kappa^2}{(1+\theta)^2}V^{(t)} + \frac{16\sqrt{2}\gamma\varsigma\kappa^2}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2}.\end{aligned}$$

For the second term of Δ_3, we have

$$\begin{aligned}
&\frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}^*}^2 - \frac{\theta^3}{2}\|x^* - y^{(t)}\|_{\bar{H}}^2 \le \frac{\theta^3}{2}\|x^* - y^{(t)}\|^2 \|\bar{H} - \bar{H}^*\| \\
\le&\frac{\theta^3\gamma\varsigma}{2}\|x^* - y^{(t)}\|^3 \overset{(26)}{\le} \frac{12\theta^3\gamma\varsigma}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

Therefore, Δ_3 is upper bounded by

$$\Delta_3 \le \frac{16r\epsilon_0\kappa^2}{(1+\theta)^2}V^{(t)} + \frac{16\sqrt{2}\gamma\varsigma\kappa^2 + 12\theta^3\gamma\varsigma}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2}.$$

Finally, we bound the value of Δ_a as follows

$$\begin{aligned}
\Delta_1 =&\frac{\gamma}{6}\left(\|p^{(t)}\|^3 + (1-\theta)\left\|x^{(t)} - y^{(t)}\right\|^3 + \theta\left\|x^* - y^{(t)}\right\|^3\right) \\
\overset{(32)(28)(26)}{\le}&\frac{\gamma}{6}\left(\left(\frac{4\sqrt{2}\kappa}{(1+\theta)\sqrt{\mu}}\left[V^{(t)}\right]^{1/2}\right)^3 + (1-\theta)\left(\frac{4\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3 + \theta\left(\frac{2\sqrt{2}}{(1+\theta)\sqrt{\mu}}\right)^3\right) \\
\le&\frac{31\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2} + \frac{31\gamma}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2} \\
\le&\frac{62\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}}\left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\Delta_1 + \Delta_2 + \Delta_3 \\
\le&\frac{8r\epsilon_0(1+\theta^2)\kappa^2}{\theta^2(1+\theta)^2}\cdot V^{(t)} + \frac{16\sqrt{2}\gamma(1+\theta^2)\kappa\varsigma}{(1+\theta)^3\mu^{3/2}}\cdot\left[V^{(t)}\right]^{3/2} + \frac{16r\epsilon_0\kappa^2}{(1+\theta)^2}\cdot V^{(t)} \\
&+\frac{16\sqrt{2}\gamma\varsigma\kappa^2 + 12\theta^3\gamma\varsigma}{(1+\theta)^3\mu^{3/2}}\cdot\left[V^{(t)}\right]^{3/2} + \frac{62\gamma\kappa^3}{(1+\theta)^3\mu^{3/2}}\cdot\left[V^{(t)}\right]^{3/2} \\
=&\frac{8r\epsilon_0(1+3\theta^2)\kappa^2}{\theta^2(1+\theta)^2}\cdot V^{(t)} + \gamma\tilde{\varphi}\cdot\left[V^{(t)}\right]^{3/2} \\
\le&\epsilon_0\cdot V^{(t)} + \gamma\tilde{\varphi}\cdot\left[V^{(t)}\right]^{3/2},
\end{aligned}$$

where the last inequality is because of the condition about r.

■

Proof of Theorem 5 Combining Lemma 15 and Lemma 16, we have

$$\begin{aligned}
\mathbb{E}\left[V^{(t+1)}\right] \le&(1-\theta)V^{(t)} + \Delta_1 + \Delta_2 + \Delta_3 \\
\le&(1-\theta)V^{(t)} + \epsilon_0\cdot V^{(t)} + \gamma\cdot\tilde{\varphi}\cdot\left[V^{(t)}\right]^{3/2} \\
=&(1-\theta+\epsilon_0)V^{(t)} + \gamma\cdot\tilde{\varphi}\cdot\left[V^{(t)}\right]^{3/2}.
\end{aligned}$$

■

Appendix D. Proof of Theorem 8

The proof of Theorem 8 mainly lies on Theorem 2. First, we are going to prove Lemma 6 which show that the conditions in Theorem 2 will be satisfied when row norm squares sampling strategy is used to construct the approximate Hessian. Then we will compute the parameter ς in Theorem 2.

Proof of Lemma 6 For notation convenience, we will omit superscript and just use S, B and α instead of $S^{(t)}$, $B^{(t)}$ and $\alpha^{(t)}$. Let us denote $\bar{H} = \left(\mathbb{E}[H^{-1}]\right)^{-1}$ and $\hat{H} = \nabla^2 F(y^{(t)})$.

First, we will prove the first condition of (4). By Jensen's inequality, we have

$$\left(\mathbb{E}[H^{-1}]\right)^{-1} \preceq \mathbb{E}[H].$$

Thus, we can obtain

$$(1-\eta)\bar{H} \preceq (1-\eta)\mathbb{E}[H] = (1-\eta)(\hat{H} + 2\alpha\|B\|^2 \cdot I),$$

where the equality follows from the fact that $\mathbb{E}[H] = \hat{H} + 2\alpha\|B\|^2 \cdot I$. We also have

$$(1-\eta)\left(\hat{H} + 2\alpha\|B\|^2 \cdot I\right) - \hat{H} = \frac{2c\kappa\sigma_{\min}^2(B)}{1+3c\kappa} \cdot I - \frac{3c\kappa}{1+3c\kappa}\hat{H} \preceq 0.$$

Therefore, we get

$$(1-\eta)\bar{H} \preceq \nabla^2 F(y^{(t)}),$$

By Theorem 1, we have

$$\|B^T B - B^T S^T S B\| \leq \alpha/2, \tag{33}$$

with probability at least $1-\delta$. This implies $\hat{H} \preceq H$ and we can obtain $\hat{H} \preceq \left(\mathbb{E}[H^{-1}]\right)^{-1}$. Thus, the first condition of Eqn. (4) is satisfied.

Then, we begin to prove that the second condition of Eqn. (4) is satisfied. Using Eqn. (33), we can obtain that

$$\begin{aligned}
&y^T\left(\hat{H} - [SB]^T SB\right)y \leq \alpha\|y\|^2 \\
\Rightarrow &y^T\left(\hat{H} + \alpha I + \hat{H} - 2([SB]^T SB + \alpha I)\right)y \leq 0 \\
\Rightarrow &\hat{H} + \mathbb{E}[H] - 2H \preceq 0.
\end{aligned}$$

Furthermore, by the fact that $(\mathbb{E}[H^{-1}])^{-1} \preceq \mathbb{E}[H]$, we obtain

$$\hat{H} + \bar{H} - 2H \preceq \hat{H} + \mathbb{E}[H] - 2H \preceq 0.$$

That is the second condition of Eqn. (4) is satisfied. ■

Now, we determine the parameter ς in Theorem 2.

Lemma 17 *Assume $\nabla^2 F(x)$ satisfies Eqn.* (11). *$S^{(t)} \in \mathbb{R}^{s\times n}$ is a row norm squares sampling matrix w.r.t. $B^{(t)}$ with $s = O(c^{-2}\cdot \mathrm{sr}(B^{(t)})\log d/\delta)$, where $0 < c < 1$ is sample size parameter and $0 < \delta < 1$ is the failure rate. Let us set regularizer $\alpha^{(t)} = 2c\|B^{(t)}\|^2$ and construct the approximate Hessian $H^{(t)}$ as Eqn.* (14). *Assume $\|B^{(t+1)} - B^{(t)}\| \leq \frac{\gamma}{2\sqrt{L}}\|y^{(t+1)} - y^{(t)}\|$. Then, we have*

$$\|\bar{H}^{(t+1)} - \bar{H}^{(t)}\| \leq \left(9 + \frac{9n}{4cs}\right)\cdot \gamma \cdot \|y^{(t+1)} - y^{(t)}\|.$$

Proof First, we have

$$\bar{H}^{(t)} - \bar{H}^{(t+1)} = \bar{H}^{(t)} \left(\mathbb{E}(H^{(t+1)})^{-1} - \mathbb{E}(H^{(t)})^{-1} \right) \bar{H}^{(t+1)}$$

and

$$\mathbb{E}(H^{(t)})^{-1} = \sum p_S \left(\alpha^{(t)} + [S^{(t)} B^{(t)}]^T S^{(t)} B^{(t)} \right)^{-1},$$

where p_S is the probability of choosing such sampling matrix $S^{(t)}$.

Now, we assume that the same sampling probability is the same for different t. This assumption is reasonable because when $x^{(t)}$ is close to the optimal point x^*, the sampling probability is close to each other, and Theorem 1 also shows that a slight disturbance of sampling probability will not affect approximation precision severely.

Under this assumption, we just S instead of $S^{(t)}$ and $S^{(t+1)}$. Now, we have

$$\begin{aligned}
&\|\mathbb{E}(H^{(t+1)})^{-1} - \mathbb{E}(H^{(t)})^{-1}\| \\
=&\left\| \sum p_S \left(\left(\alpha^{(t+1)} + [SB^{(t+1)}]^T SB^{(t+1)} \right)^{-1} - \left(\alpha^{(t)} + [SB^{(t)}]^T SB^{(t)} \right)^{-1} \right) \right\| \\
=&\left\| \sum p_S \cdot [H^{(t+1)}]^{-1} \left(\alpha^{(t)} - \alpha^{(t+1)} + [SB^{(t)}]^T SB^{(t)} - [SB^{(t+1)}]^T SB^{(t+1)} \right) [H^{(t)}]^{-1} \right\| \\
\le& \sum p_S [\alpha^{(t)}]^{-1} [\alpha^{(t+1)}]^{-1} \left(\|\alpha^{(t)} - \alpha^{(t+1)}\| + \|[SB^{(t)}]^T SB^{(t)} - [SB^{(t+1)}]^T SB^{(t+1)}\| \right).
\end{aligned}$$

First, we have

$$\begin{aligned}
\|\alpha^{(t)} - \alpha^{(t+1)}\| =& 2c \left| \|B^{(t)}\|^2 - \|B^{(t+1)}\|^2 \right| \\
=& 2c \left| \|B^{(t)}\| - \|B^{(t+1)}\| \right| \left(\|B^{(t)}\| + \|B^{(t+1)}\| \right) \\
\le& \frac{2c\gamma}{2\sqrt{L}} \left(\|B^{(t)}\| + \|B^{(t+1)}\| \right) \|y^{(t)} - y^{(t+1)}\| \\
\le& 2c\gamma \|y^{(t)} - y^{(t+1)}\|,
\end{aligned}$$

where the last inequality is because of $\|B^{(t)}\|^2 = \|\nabla^2 F(y^{(t)})\| \le L$.

Let denote $\Delta = B^{(t+1)} - B^{(t)}$, we also have

$$\begin{aligned}
&\sum p_S \|[SB^{(t)}]^T SB^{(t)} - [SB^{(t+1)}]^T SB^{(t+1)}\| \\
\le& \sum p_S \|[SB^{(t)}]^T SB^{(t)} - [S(B^{(t)} + \Delta)]^T S(B^{(t)+\Delta})\| \\
\le& \sum p_S \|\Delta\| \|S^T S\| \|B^{(t)}\| \\
\le& \frac{\gamma}{2} \left(\sum p_S \|S^T S\| \right) \|y^{(t+1)} - y^{(t)}\|.
\end{aligned}$$

Now, we need to bound $\sum p_S \|S^T S\|$. By the property of sampling matrix, we have

$$\sum p_S \|S^T S\| = \sum_{i_1 \dots i_s} p_{i_1} \dots p_{i_s} \frac{1}{s} \|\mathrm{diag}(\frac{1}{p_{i_1}}, \dots, \frac{1}{p_{i_s}})\| \le \frac{n}{s}.$$

Therefore, we have

$$\begin{aligned}\|\mathbb{E}(H^{(t+1)})^{-1} - \mathbb{E}(H^{(t)})^{-1}\| \leq & [\alpha^{(t)}]^{-1}[\alpha^{(t+1)}]^{-1}\left(2c + \frac{n}{2s}\right)\gamma\|y^{(t+1)} - y^{(t)}\| \\ \leq & \frac{2c + \frac{n}{2s}}{2c\|\nabla^2 F(y^{(t)})\|\|\nabla^2 F(y^{(t+1)})\|}\gamma\|y^{(t+1)} - y^{(t)}\|.\end{aligned}$$

By the condition (4), we bound the norm of $\bar{H}^{(t)}$ as follows

$$\|\bar{H}^{(t)}\| \leq \|\nabla^2 F(y^{(t)})\| + \|H^{(t)}\| \leq 3\|\nabla^2 F(y^{(t)})\|.$$

Similarly, we also have

$$\|\bar{H}^{(t+1)}\| \leq 3\|\nabla^2 F(y^{(t+1)})\|.$$

Combining above results, we have

$$\begin{aligned}\|\bar{H}^{(t)} - \bar{H}^{(t+1)}\| \leq & \|\bar{H}^{(t)}\| \cdot \|\bar{H}^{(t+1)}\| \cdot \|\mathbb{E}(H^{(t+1)})^{-1} - \mathbb{E}(H^{(t)})^{-1}\| \\ \leq & 9\|\nabla^2 F(y^{(t)})\|\|\nabla^2 F(y^{(t+1)})\| \cdot \frac{c + \frac{n}{4s}}{c\|\nabla^2 F(y^{(t)})\|\|\nabla F(y^{(t+1)})\|}\gamma\|y^{(t+1)} - y^{(t)}\| \\ = & 9\left(1 + \frac{n}{4cs}\right)\gamma\|y^{(t+1)} - y^{(t)}\|.\end{aligned}$$

Therefore, we obtain that $\varsigma = 9 + \frac{9n}{4cs}$.

■

Combining above two lemmas, we can easily prove Theorem 8.

Proof of Theorem 8 First, we will prove that $\nabla^2 F(x)$ is γ-Lipschitz continuous. By the assumption $\left\|B^{(t+1)} - B^{(t)}\right\| \leq \frac{\gamma}{2\sqrt{L}}\left\|y^{(t+1)} - y^{(t)}\right\|$, we have

$$\begin{aligned}\left\|\nabla^2 F(y^{(t+1)}) - \nabla^2 F(y^{(t)})\right\| = & \left\|[B^{(t+1)}]^T B^{(t+1)} - [B^{(t+1)}]^T B^{(t+1)}\right\| \\ \leq & \left(\left\|B^{(t+1)}\right\| + \left\|B^{(t)}\right\|\right) \cdot \left\|B^{(t+1)} - B^{(t)}\right\| \\ \leq & 2\sqrt{L} \cdot \frac{1}{2\sqrt{L}}\left\|y^{(t+1)} - y^{(t)}\right\| = \left\|y^{(t+1)} - y^{(t)}\right\|,\end{aligned}$$

where the last inequality is because because of $\|B^{(t)}\|^2 = \|\nabla^2 F(y^{(t)})\| \leq L$.

Next, we will use Theorem 2 to prove the theorem. By Lemma 6, we can see that Condition 4 holds with probability at least $1 - \delta$. By Lemma 17, we obtain the value of ς in Theorem 2. Therefore, we can obtain the result. ■

Appendix E. Proof of Theorem 10

The proof of Theorem 10 is similar to the one of Theorem 8. First, we show that Condition (4) holds with high probability. We calculate the value of ς in Theorem 2.

Proof of Theorem 10 First, by assumption $\left\|\nabla^2 f_i(y^{(t+1)}) - \nabla^2 f_i(y^{(t)})\right\| \le \gamma \left\|y^{(t+1)} - y^{(t)}\right\|$ for all $i \in \{1, \dots, n\}$ and $y^{(t)}$, we can obtain that

$$\left\|\nabla^2 F(y^{(t+1)}) - \nabla^2 F(y^{(t)})\right\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 f_i(y^{(t+1)}) - \nabla^2 f_i(y^{(t)})\right)\right\|$$

$$\le \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla^2 f_i(y^{(t+1)}) - \nabla^2 f_i(y^{(t)})\right\| \le \gamma \left\|y^{(t+1)} - y^{(t)}\right\|,$$

that is, $\nabla^2 F(x)$ is γ-Lipschitz continuous. Next, we will use Theorem 2 to prove our theorem.

For notation convenience, we will omit superscript. We denote and $\bar{H} = \left(\mathbb{E}[H^{-1}]\right)^{-1}$ and $\hat{H} = \nabla^2 F(y^{(t)})$. By Jensen's inequality, we have

$$\left(\mathbb{E}[H^{-1}]\right)^{-1} \preceq \mathbb{E}[H].$$

Thus, we can obtain

$$(1-\eta)\bar{H} \preceq (1-\eta)\mathbb{E}[H] = (1-\eta)(\hat{H} + 2\alpha\|B\|^2 \cdot I).$$

We also have

$$(1-\eta)\left(\hat{H} + 2c \cdot I\right) - \hat{H} \preceq \frac{2c\mu}{\mu + 2c} \cdot I - \frac{2c}{\mu + 2c}\hat{H} \preceq 0.$$

where the last inequality is because of Eqn. (18). Therefore, we get

$$(1-\eta)\bar{H} \preceq \nabla^2 F(y^{(t)}). \tag{34}$$

Consider random matrices $H_j^{(t)} = \nabla^2 f_j(x^{(t)}), j = 1, \dots, s$ being sampled uniformly. Then, we have $\mathbb{E}[H_j^{(t)}] = \nabla^2 F(x^{(t)})$ for all $j = 1, \dots, s$. By (17) and the positive semi-definite property of $H_j^{(t)}$, we have $\lambda_{\max}(H_j^{(t)}) \le K$ and $\lambda_{\min}(H_j^{(t)}) \ge 0$.

We define random matrices $X_j = H_j^{(t)} - \nabla^2 F(x^{(t)})$ for all $j = 1, \dots, |\mathcal{S}|$. We have $\mathbb{E}[X_j] = 0$, $\|X_j\| \le 2K$ and $\|X_j\|^2 \le 4K^2$. By the matrix Bernstein inequality (Tropp et al., 2015), we have

$$\|H^{(t)} - \nabla^2 F(x^{(t)})\| \le \sqrt{\frac{2 \cdot 4K^2 \log 2d/\delta}{s}} + \frac{2K \log 2d/\delta}{3s},$$

with probability at least $1 - \delta$. When the sample size $|\mathcal{S}| = O(c^{-2}K^2 \log d/\delta)$, we have

$$\|H^{(t)} - \nabla^2 F(x^{(t)})\| \le \frac{\alpha}{2} \tag{35}$$

holds with probability $1 - \delta$. This implies $\hat{H} \preceq H$ and we can obtain that $\hat{H} \preceq \left(\mathbb{E}[H^{-1}]\right)^{-1}$. Combining Eqn. (34), we can obtain that the first condition of Eqn. (4) is satisfied. Eqn. (35) also implies that

$$y^T\left(\hat{H} - [SB]^T SB\right)y \le \alpha\|y\|^2$$

$$\begin{aligned}
&\Rightarrow y^T\left(\hat{H}+\alpha I+\hat{H}-2([SB]^TSB+\alpha I)\right)y\le 0\\
&\Rightarrow \hat{H}+\mathbb{E}[H]-2H\preceq 0.
\end{aligned}$$

Furthermore, by the fact that $(\mathbb{E}[H^{-1}])^{-1}\preceq\mathbb{E}[H]$, we obtain

$$\hat{H}+\bar{H}-2H\preceq\hat{H}+\mathbb{E}[H]-2H\preceq 0.$$

That is the second condition of Eqn. (4) is satisfied.

Similar to the proof of Theorem 8, we need to calculate the value of ς. First, we have

$$\bar{H}^{(t)}-\bar{H}^{(t+1)}=\bar{H}^{(t)}\left(\mathbb{E}(H^{(t+1)})^{-1}-\mathbb{E}(H^{(t)})^{-1}\right)\bar{H}^{(t+1)}$$

and

$$\mathbb{E}(H^{(t)})^{-1}=\sum p_S\left(\alpha^{(t)}+\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})\right)^{-1},$$

where p_S is the probability of choosing such a subset $\mathcal{S}$. Hence, we have

$$\begin{aligned}
&\|\mathbb{E}(H^{(t+1)})^{-1}-\mathbb{E}(H^{(t)})^{-1}\|\\
=&\left\|\sum p_S\left(\left(\alpha^{(t)}+\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t+1)})\right)^{-1}-\left(\alpha^{(t)}+\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})\right)^{-1}\right)\right\|\\
=&\left\|\sum p_S\cdot[H^{(t+1)}]^{-1}\left(\alpha^{(t)}-\alpha^{(t+1)}+\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})-\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t+1)})\right)[H^{(t)}]^{-1}\right\|\\
\le&\sum p_S[\alpha^{(t)}]^{-1}[\alpha^{(t+1)}]^{-1}\left(\|\alpha^{(t)}-\alpha^{(t+1)}\|+\|\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})-\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t+1)})\|\right)\\
\le&\frac{1}{4c^2}\sum p_S\left\|\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})-\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t+1)})\right\|,
\end{aligned}$$

where the last inequality is because $\alpha^{(t)}=2c$. We also have

$$\begin{aligned}
&\sum p_S\left\|\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t)})-\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\nabla^2 f_j(y^{(t+1)})\right\|\\
\le&\frac{1}{|\mathcal{S}|}\sum p_S\sum_{j\in\mathcal{S}}\|\nabla^2 f_j(y^{(t)})-\nabla^2 f_j(y^{(t+1)})\|\\
\le&\gamma\sum p_S\|y^{(t+1)}-y^{(t)}\|=\gamma\cdot\|y^{(t+1)}-y^{(t)}\|.
\end{aligned}$$

Hence, in this case, $\varsigma=1$. We can obtain the convergence property by Theorem 2. ■

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.

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.

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.

Xi Chen, Bo Jiang, Tianyi Lin, and Shuzhong Zhang. On adaptive cubic regularized newton's methods for convex optimization via random sampling. *arXiv preprint arXiv:1802.05426*, 2018.

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.

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

Saeed Ghadimi, Han Liu, and Tong Zhang. Second-order methods with cubic regularization under inexact information. *arXiv preprint arXiv:1710.05782*, 2017.

I. Guyon. Sido: A phamacology dataset. URL `http://www.causality.inf.ethz.ch/data/SIDO.html`.

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

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2):167–215, 2018.

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.

Xiang Li, Shusen Wang, and Zhihua Zhang. Do subsampled newton methods work for high-dimensional data? In *AAAI*, pages 4723–4730, 2020.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Yu Nesterov. Accelerating the cubic regularization of newtons method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov and Sebastian U Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

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

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.

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.

Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, and Benjamin Recht. Breaking locality accelerates block gauss-seidel. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3482–3491, 2017.

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 *International Conference on Machine Learning*, pages 3931–3939, 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.