A Unified Scheme to Accelerate Adaptive Cubic Regularization and Gradient Methods for Convex Optimization

Bo JIANG ∗ Tianyi LIN † Shuzhong ZHANG ‡

December 29, 2017

Abstract

In this paper we propose a unified two-phase scheme for convex optimization to accelerate: (1) the adaptive cubic regularization methods with exact/inexact Hessian matrices, and (2) the adaptive gradient method, without any knowledge of the Lipschitz constants for the gradient or the Hessian. This is achieved by tuning the parameters used in the algorithm adaptively in its process of progression, which can be viewed as a relaxation over the existing algorithms in the literature. Under the assumption that the sub-problems can be solved approximately, we establish overall iteration complexity bounds for three newly proposed algorithms to obtain an $\epsilon$-optimal solution. Specifically, we show that the adaptive cubic regularization methods with the exact/inexact Hessian matrix both achieve an iteration complexity in the order of $O\left(\frac{1}{\epsilon^{1/3}}\right)$, which matches that of the original accelerated cubic regularization method presented in [24] assuming the availability of the exact Hessian information and the Lipschitz constants, and the global solution of the sub-problems. Under the same two-phase adaptive acceleration framework, the gradient method achieves an iteration complexity in the order of $O\left(\frac{1}{\epsilon^{1/2}}\right)$, which is known to be best possible (cf. [26]). Our numerical experiment results show a clear effect of acceleration displayed in the adaptive Newton’s method with cubic regularization on a set of regularized logistic regression instances.

Keywords: convex optimization; acceleration; adaptive algorithm; cubic regularization; Newton’s method; gradient method; iteration complexity.

Mathematics Subject Classification: 90C06, 90C60, 90C53.

∗Research Center for Management Science and Data Analytics, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China. Email: isyebojiang@gmail.com.
†Department of Industrial Engineering and Operations Research, UC Berkeley, Berkeley, CA 94720, USA. Email: darren_lin@berkeley.edu
‡Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: zhangs@umn.edu.
1 Introduction

1.1 Motivations

We consider the following generic unconstrained optimization model:

$$f^* := \min_{x \in \mathbb{R}^d} f(x),$$  \hspace{1cm} (1)

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex, and $f^* > -\infty$. During the past decades, various classes of optimization algorithms for solving (1) have been developed and carefully analyzed; see [19, 28, 26] for detailed information and references. Two types of concerns often arise in the design of optimization algorithms. First, the high order information (such as the Hessian matrices) maybe expensive to acquire. Second, the problem parameters such as the first and the second order Lipschitz constants are usually hard to estimate. On the other hand, for an optimization algorithms to be effective and practical, they will need to be robust and less dependent on the knowledge of the structure of the problem at hand. In this context, schemes to adaptively adjust the parameters used in the algorithm are desirable, and are likely leading to improve its numerical performances. As an example, researchers in the area of deep learning tend to train their models with adaptive gradient method (see e.g. AdaGrad in [12]) due to its robustness and effectiveness (cf. [13]). In fact, Adam [14] and RMSProp [32] are recognized as the default solution methods in the deep learning setting.

Another fundamental issue in optimization (as well as in machine learning) is to understand how the classical algorithms (including both the first-order and second-order methods) can be accelerated. Nesterov [23] put forward the very first accelerated (optimal in its iteration counts) gradient-based algorithm for convex optimization. Recently, a number of adaptively accelerated gradient methods have been proposed; see [12, 25, 18, 21]. Unfortunately, none of these are fully parameter free. Comparing to their first-order counterpart, investigations on the second-order methods is relatively scarce, as acceleration with the second-order information is much more involved. To the best of our knowledge, [24, 22] are the only papers that are concerned with accelerating the second-order methods. However, these two algorithms do require the knowledge of some problem (Lipschitz) constants,

Indeed, algorithms exhibiting both traits of acceleration and adaptation have been largely missing in the literature. As a matter of fact, we are unaware of any prior accelerated second-order methods (or even any first-order methods) that are fully independent of the problem constants while maintaining superior theoretical iteration complexity bounds. For instance, the adaptive cubic regularized Newton’s method [8] merely achieves an iteration complexity bound of $O \left( 1/\epsilon^{1/2} \right)$ without acceleration. Thus, a natural question raises:

Can we develop an implementable accelerated cubic regularization method with an iteration complexity lower than $O \left( 1/\epsilon^{1/2} \right)$?

This paper sets out to present an affirmative answer to the above question. Moreover, the resulting accelerated adaptive cubic regularization algorithm displays an excellent numerical performance in solving a variety of large-scale machine learning models in our experiments.
1.2 Related Work

Nesterov’s seminal work [23] triggered a burst of research on accelerating first-order methods. There have been a good deal of recent efforts to understand its nature from other perspectives [2, 4, 9, 11, 13, 14, 16], or modify it to account for more general settings [3, 10, 11, 29, 17]. Parallel to this, the adaptive gradient methods with the optimal convergence rate have been proposed [12, 25, 18, 21], and widely used in training the deep neural networks [14, 32]. However, all of these algorithms are not fully parameter-independent. Specifically, Duchi et al. [12] needs to tune the step-size $\eta$ and the regularization parameter $\delta$; Lin and Xiao [18] and Nesterov [25] require a lower bound on the Lipschitz constant $L$ for the gradient; and Monteiro and Svaiter [21] need an upper bound of $L_g - \mu$, where $\mu$ is a strong convexity parameter.

In terms of the second-order methods (in particular Newton’s method), the literature regarding acceleration is quite limited. To the best of our knowledge, Nesterov [24] is the first along this direction, where the overall iteration complexity for convex optimization was improved from $O\left(1/\epsilon^{1/2}\right)$ to $O\left(1/\epsilon^{1/3}\right)$ for the cubic regularization for Newton’s method [27]. After that, Monteiro and Svaiter [22] managed to accelerate the Newton proximal extragradient method [20] with an improved iteration complexity of $O\left(1/\epsilon^{2/7}\right)$. Moreover, this approach allows a larger stepsize and can even accommodate a non-smooth objective function. Very recently, Shamir and Shiff [30] proved that $O\left(1/\epsilon^{2/7}\right)$ is actually a lower bound for the oracle complexity of the second-order methods for convex smooth optimization, which implies that the accelerated Newton proximal extragradient method is an optimal second-order method. However, viewed from an implementation perspective, the acceleration second-order scheme in [24, 22] are not easy to apply in practice. Indeed, Nesterov’s method assumes that all the parameters, including the Lipschitz constant for the Hessian, are known, and the sub-problems with cubic regularization are solved to global optimality; Monteiro and Svaiter’s method also assumes the knowledge of the Lipschitz constant of the Hessian. To alleviate this, Cartis et al. incorporated an adaptive strategy into Nesterov’s approach [24], and further relaxed the criterion for solving each sub-problem while maintaining the convergence properties for both convex [8] and non-convex [6, 7] cases. However, as mentioned earlier, the iteration complexity established in [8] for convex optimization is merely $O\left(1/\epsilon^{1/2}\right)$. Furthermore, in [9] the same authors also developed a way to construct an approximation for the Hessian, which significantly reduces the per-iteration cost. There are other recent works on approximate cubic regularization for Newton’s method. For instance, Carmon and Duchi [5] and Agarwal et al. [1] proposed some variants, where the sub-problem is approximately solved without resorting to Hessian matrix; Kohler and Lucchi [15] proposed a uniform sub-sampling strategy to approximate the Hessian in the cubic regularization for Newton’s method. However, the approximative Hessian and gradient are constructed based on a priori unknown step which can only be determined after such approximations are formed. Xu et al. [36, 35] fixed this issue by proposing appropriate uniform and non-uniform sub-sampling strategies to construct Hessian approximations in the trust region context, as well as the cubic regularization for Newton’s method.

1.3 Contributions

The contributions of this paper can be summarized as follows. We present a unified adaptive accelerating scheme that can be specialized to several optimization algorithms including cubic regularized Newton’s method with exact/inexact Hessian and gradient method. This can be considered complementary to the current stream of research in two aspects. First, all the accelerated algorithms developed in this paper
are parameter-free due to the new fully adaptive strategies, while only partially adaptive strategies are observed from other accelerated first-order methods in the literature [25] [18] [21]. Second, it is worth noting that the research efforts on accelerated algorithms have been rather unequally spread between the first-order and second-order methods, with the former receiving a lot more attention. Our results on the adaptive and accelerated cubic regularization for Newton’s method contribute as one step towards balancing the studies on the two methods.

In terms of the convergence rates of our algorithms, for the cubic regularized Newton’s method we show that a global convergence rate of $O \left( \frac{1}{\epsilon^{1/3}} \right)$ holds (Theorem 3.8) without assuming any knowledge of the problem parameters. We further prove that, even without the exact Hessian information, the same $O \left( \frac{1}{\epsilon^{1/3}} \right)$ rate of convergence (Theorem 4.3) is still achievable for the cubic regularized approximative Newton’s method. For the gradient descent method, our adaptive algorithm achieves a convergence rate of $O \left( \frac{1}{\epsilon^{1/2}} \right)$ (Theorem 5.2) which matches the optimal rate for the first order methods [26]. When the objective function is strongly convex, the convergence results are also established for these three algorithms accordingly.

For the subproblem in the cubic regularized Newton’s method with exact/inexact Hessian, we only require an approximative solution satisfying (7). Note that our approximation measure does not include the usual condition in the form of (8), and thus is weaker than the one used in [6]. This relaxation opens up possibilities for other approximation solution methods to solve the subproblem. For instance, Carmon and Duchi [5] proposed to use the gradient descent method, and they proved that it works well even when the cubic regularized subproblem is nonconvex. Moreover, such function in our case is strongly convex, and thus the gradient descent subroutine is expected to have a fast (linear) convergence.

1.4 Notations and Organization

Throughout the paper, we denote vectors by bold lower case letters, e.g., $\mathbf{x}$, and matrices by regular upper case letters, e.g., $X$. The transpose of a real vector $\mathbf{x}$ is denoted as $\mathbf{x}^\top$. For a vector $\mathbf{x}$, and a matrix $X$, $||\mathbf{x}||$ and $||X||$ denote the $\ell_2$ norm and the matrix spectral norm, respectively. $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ are respectively the gradient and the Hessian of $f$ at $\mathbf{x}$, and $I$ denotes the identity matrix. For two symmetric matrices $A$ and $B$, $A \succeq B$ indicates that $A - B$ is symmetric positive semi-definite. The subscript, e.g., $x_i$, denotes iteration counter. $\log(x)$ denotes the natural logarithm of $x$. The inexact Hessian is denoted by $H(\mathbf{x})$, but for notational simplicity, we also use $H_i$ to denote the inexact Hessian evaluated at the iterate $\mathbf{x}_i$ in iteration $i$, i.e., $H_i \triangleq H(\mathbf{x}_i)$.

The rest of the paper is organized as follows. In Section 2.1 we introduce notations and assumptions used throughout this paper, and present our general framework in Section 2.2. Then the specializations to cubic regularized Newton’s method with exact/inexact Hessian matrix and gradient descent method are presented in Sections 3, 4 and 5 respectively. In Section 6 we present some preliminary numerical results on solving Regularized Logistic Regression, where acceleration of the method based on the adaptive cubic regularization for Newton’s method is clearly observed. The details of all the proofs can be found in the appendix.
2 A Unified Adaptive Acceleration Framework

In this section, we first introduce the main definitions and assumptions used in the paper, and then present our unified adaptive acceleration framework.

2.1 Assumptions

Throughout this paper, we refer to the following definition of $\epsilon$-optimality.

**Definition 2.1** ($\epsilon$-optimality). Given $\epsilon \in (0, 1)$, $x \in \mathbb{R}^d$ is said to be an $\epsilon$-optimal solution to problem (1), if

$$f(x) - f(x^*) \leq \epsilon,$$

where $x^* \in \mathbb{R}^d$ is the global optimal solution to problem (1).

To proceed, we make the following standard assumption regarding the gradient and Hessian of the objective function $f$.

**Assumption 2.1** The objective function $f(x)$ in problem (1) is convex and twice differentiable with the gradient and the Hessian being both Lipschitz continuous, i.e., there are $0 < L_g, L_h < \infty$ such that for any $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L_g \|x - y\|,$$

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_h \|x - y\|.$$

We also study the problem with a strongly convex objective defined as follows:

**Definition 2.2** A function $f$ is said to be strongly convex if there is $\mu > 0$, such that for any $x, y \in \mathbb{R}^d$ we have

$$f(y) - f(x) - (y - x)^T \nabla f(x) \geq \frac{\mu}{2} \|y - x\|^2.$$

2.2 Framework

The adaptive acceleration framework is composed of two separate subroutines. Specifically, the framework starts with a Simple Adaptive Subroutine (SAS), which terminates as soon as one successful iteration is identified. Then, the output of SAS is used as an initial point to run Accelerated Adaptive Subroutine (AAS) until a sufficient number of successful iterations are recorded. The details of our framework are summarized in Table 1.

Note that certain adaptive strategies are adopted to tune the regularization parameters in both $m(x_i, s, \sigma_i)$ and $\psi_l(z, \varsigma_l)$ while the acceleration is only installed in AAS, where the tuple $(\bar{x}_l, y_l, z_l)$ is updated when a successful iteration is identified. In addition, the criteria for identifying the successful iteration in each subroutine are different. When specialized to cubic regularization for Newton’s method, SAS can
Begin Phase I: Simple Adaptive Subroutine (SAS)

for $i = 0, 1, \cdots$, do

Construct certain regularized function $m(x_i, s, \sigma_i)$ with a regularization parameter $\sigma_i$;
Compute $s_i$ by solving $m(x_i, s, \sigma_i)$ approximately or exactly;
if iteration $i$ is successful then

Set $x_{i+1} = x_i + s_i$ and update $\sigma_{i+1}$;
Record the total number of iterations for SAS: $T_1 = i + 1$;
break;
else

Set $x_{i+1} = x_i$, and update $\sigma_{i+1}$.
end if
end for

End Phase I (SAS)

Begin Phase II: Accelerated Adaptive Subroutine (AAS)

Set the count of successful iterations $l = 1$ and let $\bar{x}_1 = x_{T_1}$;
Construct auxiliary function $\psi_1(z, \varsigma_1)$ with some $\varsigma_1 > 0$, and let $z_1 = \arg\min_{z \in \mathbb{R}^d} \psi_1(z, \varsigma_1)$,
and choose $y_1 = \alpha_1 \bar{x}_1 + (1 - \alpha_1)z_1$;

for $j = 0, 1, \cdots$, do

Construct regularized function $m(y_{l}, s, \sigma_{T_1+j})$ with regularized parameter $\sigma_{T_1+j}$;
Compute $s_{T_1+j}$ by solving $m(y_{l}, s, \sigma_{T_1+j})$ approximately or exactly;
if iteration $T_1 + j$ is successful then

Update $\sigma_{T_1+j+1}$ and set $x_{T_1+j+1} = x_{T_1+j} + s_{T_1+j}$;
Update the count of successful iterations $l = l + 1$;
Update the auxiliary function $\psi_l(z, \varsigma_l)$ by choosing the regularization parameter $\varsigma_l$ automatically;
Solve $z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z)$, let $\bar{x}_l = x_{T_1+j+1}$ and $y_l = \alpha_l \bar{x}_l + (1 - \alpha_l)z_l$;
else

Set $x_{T_1+j+1} = x_{T_1+j}$ and update $\sigma_{T_1+j+1}$;
end if
end for

Record the total number of iterations for AAS: $T_2 = j + 1$.
End Phase II (AAS)

Table 1: Unified Adaptive Acceleration Framework

be interpreted as the initialization step based on a modification of adaptive cubic regularization method proposed in [6, 7].

For the three algorithms mentioned above, the specific forms of regularized function $m(x, s, \sigma)$ are presented in Table 2 and the iterative update rule for auxiliary function $\psi_l(z)$ and the accelerating coefficient $\alpha_l$ are presented in Table 3. In the rest of the paper, we shall analyze these three specialized algorithms within the framework just introduced.
3 Accelerated Adaptive Cubic Regularization with Exact Hessian

As illustrated in Table 2 we consider the following approximation of \( f \) evaluated at \( x_i \) with cubic regularization [6, 7]:

\[
m(x_i, s, \sigma) = f(x_i) + s^T \nabla f(x_i) + \frac{1}{2} s^T \nabla^2 f(x_i)s + \frac{1}{3} \sigma_i \|s\|^3,
\]

where \( \sigma_i > 0 \) is a regularized parameter adjusted by the algorithm in the process of iterating. Now we present the accelerated adaptive cubic regularization for Newton’s method with exact Hessian in Algorithm 1.

Note that in each iteration of Algorithm 1 we approximately solve

\[
s_i \approx \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i),
\]

where \( m(x_i, s, \sigma_i) \) is defined in [6] and the symbol “\( \approx \)” is quantified as follows:

**Condition 3.1** We call \( s_i \) to be an approximative solution – denoted as \( s_i \approx \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i) \) – for \( \min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i) \), if the following holds

\[
\| \nabla m(x_i, s_i, \sigma_i) \| \leq \kappa_\theta \min (1, \|s_i\|) \min (\|s_i\|, \|\nabla f(x_i)\|),
\]

where \( \kappa_\theta \in (0, 1) \) is a pre-specified constant.

Note that [7] is also used as one of the two stopping criteria for solving the subproblem in the original adaptive cubic regularization for Newton’s method in [8]. However, the other criterion

\[
s_i^T \nabla f(x_i) + s_i^T \nabla^2 f(x_i)s_i + \sigma_i \|s_i\|^3 = 0
\]

is not needed in Algorithm 1. Another difference is that both criteria for the successful iterations in SAS and AAS of Algorithm 1 are different than these used in [8].
Algorithm 1 Accelerated Adaptive Cubic Regularization for Newton’s Method with Exact Hessian

Given $\gamma_2 > \gamma_1 > 1$, $\gamma_3 > 1$, $\eta > 0$ and $\sigma_{min} > 0$. Specify $m(x_i, s, \sigma_i)$ as in Table 2. Choose $x_0 \in \mathbb{R}^d$, $\sigma_0 \geq \sigma_{min}$, and $c_1 > 0$.

Begin Phase I: Simple Adaptive Subroutine (SAS)

for $i = 0, 1, 2, \ldots$ do
  Compute $s_i \in \mathbb{R}^d$ such that $s_i = \text{argmin}_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i)$;
  Compute $\rho_i = f(x_i + s_i) - m(x_i, s_i, \sigma_i)$.
  if $\rho_i < 0$ [successful iteration] then
    $x_{i+1} = x_i + s_i$,
    $\sigma_{i+1} \in [\sigma_{min}, \sigma_i]$;
    Record the total number of iterations for SAS: $T_1 = i + 1$.
    break.
  else
    $x_{i+1} = x_i$,
    $\sigma_{i+1} \in [\gamma_1 \sigma_i, \gamma_2 \sigma_i]$.
  end if
end for

End Phase I (SAS).

Begin Phase II: Accelerated Adaptive Subroutine (AAS)

Set the count of successful iterations $l = 1$ and let $x_1 = x_{T_1}$;

Construct $\psi_1(z) = f(\bar{x}_1) + \frac{\eta}{2} \|z - \bar{x}_1\|^2$, and let $x_1 = \text{argmin}_{z \in \mathbb{R}^d} \psi_1(z)$, and choose $y_1 = \frac{\partial}{\partial z} \psi_1(z) + \frac{\partial}{\partial \sigma} \psi_1(z)$.

for $j = 0, 1, 2, \ldots$ do
  Compute $s_{T_1+j} \in \mathbb{R}^d$ such that $s_{T_1+j} = \text{argmin}_{s \in \mathbb{R}^d} m(y_j, s, \sigma_{T_1+j})$, and $\rho_{T_1+j} = \frac{s_{T_1+j}}{\|\sigma_{T_1+j}\|^2}$;  
  if $\rho_{T_1+j} \geq \eta$ [successful iteration] then
    $x_{T_1+j+1} = x_j + s_{T_1+j}$,
    $\sigma_{T_1+j+1} \in [\sigma_{min}, \sigma_{T_1+j}]$;
    Set $l = l + 1$ and $c = c_l$;
    Update $\psi_j(z)$ as illustrated in Table 3 by using $c_l = c$, and compute $x_j = \text{argmin}_{z \in \mathbb{R}^d} \psi_j(z)$;
    while $\psi_j(z) \geq f_{l+1}(z) + \frac{f_{l+1}(z)}{l+1}$ do
      Set $c = \gamma_3 c$,
      $\psi_{j+1} = \psi_j(z) + \frac{f_{l+1}(z)}{l+1}$;
      Compute $x_{j+1} = \text{argmin}_{z \in \mathbb{R}^d} \psi_{j+1}(z)$.
    end while
    Set $\bar{x}_1 = \frac{1}{l+1} x_j + \frac{3}{l+1} z_1$;
    else
      $x_{T_1+j+1} = x_{T_1+j}$,
      $\sigma_{T_1+j+1} \in [\gamma_1 \sigma_{T_1+j}, \gamma_2 \sigma_{T_1+j}]$;
    end if
end for

Record the total number of iterations for AAS: $T_2 = j + 1$.

End Phase II (AAS).

From the standpoint of acceleration, we shall show that Algorithm 1 will retain the same iteration complexity of $O\left(1/\epsilon^{1/3}\right)$ as for the nonadaptive version of [24] even when the subproblem is now only solved approximatively. On the surface, under the new scheme we need to solve an additional cubic subproblem:

$z_l = \text{argmin}_{z \in \mathbb{R}^d} \psi_l(z)$.

Fortunately, this problem admits a closed-form solution. In particular, recall that the objective function is obtained by using the updating rule in Table 3 and so

$\psi_l(z) = \ell_l(z) + \frac{1}{6} \|z - \bar{x}_1\|^3$, \quad $l = 1, 2, \ldots$,

where $\ell_l(z)$ is a certain linear function of $z$. By letting

$\nabla \psi_l(z) = \nabla \ell_l(z) + \frac{1}{2} \bar{x}_1 \|z - \bar{x}_1\| (z - \bar{x}_1) = 0$,
we have $\|z - \bar{x}_1\| = \sqrt{\frac{2}{\varsigma} \|\nabla \ell_i(z)\|}$. Since $\ell_i(z)$ is linear, $\nabla \ell_i(z)$ is independent of $z$. Therefore, we have

$$z_i = \bar{x}_1 - \sqrt{\frac{2}{\varsigma \|\nabla \ell_i(z)\|}} \nabla \ell_i(z).$$

3.1 The Convex Case

In this subsection, we aim to analyze the theoretical performance of Algorithm 1 when the objective function is convex.

3.1.1 Sketch of the Proof

To give a holistic picture of the proof, we sketch some major steps below.

Proof Outline:

1. We denote $T_1$ to be the total number of iterations in SAS. Note that the criterion for the successfully iteration in SAS will be satisfied when $\sigma_i$ is sufficiently large. Then $T_1$ is bounded above by some constant (Lemma 3.1).

2. We denote $T_2$ by the total number of iterations in AAS, and

$$S = \{ j \leq T_2 : T_1 + j \text{ successful iteration} \}$$

be the index set of all successful iterations in AAS. Then $T_2$ is bounded above by $|S|$ multiplied by some constant (Lemma 3.2).

3. We denote $T_3$ by the total number of counts successfully updating $\varsigma > 0$, and $T_3$ is upper bound by some constant (Lemma 3.6).

4. We relate the objective function to the count of successful iterations in AAS (Theorem 3.7).

5. Putting all the pieces together, we obtain an iteration complexity result (Theorem 3.8).

3.1.2 Bound the Iteration Numbers

Lemma 3.1 Letting $\bar{\sigma}_1 = \max \left\{ \sigma_0, \frac{\gamma_1 L}{2} \right\} > 0$, we have $T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\sigma_1}{\sigma_{\min}} \right)$. 

9
Proof. We have
\[
    f(x_i + s_i) = f(x_i) + s_i^T \nabla f(x_i) + \frac{1}{2}s_i^T \nabla^2 f(x_i) s_i + \int_0^1 (1 - \tau) s_i^T [\nabla^2 f(x_i + \tau s_i) - \nabla^2 f(x_i)] s_i \, d\tau
\]
\[
    \leq f(x_i) + s_i^T \nabla f(x_i) + \frac{1}{2}s_i^T \nabla^2 f(x_i) s_i + \frac{L_h}{6} ||s_i||^3
\]
\[
    = m(x_i, s_i, \sigma_i) + \left( \frac{L_h}{6} - \frac{\sigma_i}{3} \right) ||s_i||^3,
\]
where the inequality holds true due to Assumption 2.1. Therefore, we conclude that
\[
    \sigma_i \geq \frac{L_h}{2} \implies f(x_i + s_i) \leq m(x_i, s_i, \sigma_i),
\]
which further implies that \( \sigma_i < \frac{L_h}{2} \) for \( i \leq T_1 - 2 \). Hence,
\[
    \sigma_{1j} \leq \sigma_{1j-1} \leq \gamma_2 \sigma_{1j-2} \leq \frac{\gamma_2 L_h}{2}.
\]
By the definition that \( \bar{\sigma}_1 = \max \{ \sigma_0, \frac{\gamma_2 L_h}{2} \} \), it follows from the construction of Algorithm 1 that \( \sigma_{\min} \leq \sigma_i \) for all iterations, and \( \gamma_1 \sigma_i \leq \sigma_{i+1} \) for all unsuccessful iterations. Consequently, we have
\[
    \frac{\bar{\sigma}_1}{\sigma_{\min}} \geq \frac{\sigma_{1j}}{\sigma_0} = \frac{\sigma_{1j}}{\sigma_{1j-1}} \cdot \frac{\sigma_{1j-1}}{\sigma_{1j-2}} \cdots \frac{\sigma_{1j+1}}{\sigma_{1j}} \geq \gamma_1^{T_1-1} \left( \frac{\sigma_{\min}}{\bar{\sigma}_1} \right),
\]
and hence \( T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) \). \( \square \)

Lemma 3.2 Letting \( \bar{\sigma}_2 = \max \{ \bar{\sigma}_1, \frac{\gamma_2 L_h}{2} + \gamma_2 \kappa \theta + \gamma_2 \eta \} > 0 \), we have \( T_2 \leq \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) |S| \).

Proof. We have
\[
    s_{T_1+j}^T \nabla f(y_i + s_{T_1+j})
\]
\[
    = s_{T_1+j}^T [\nabla f(y_i + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j}] + s_{T_1+j}^T [\nabla f(y_i) + \nabla^2 f(y_i) s_{T_1+j}]
\]
\[
    \leq ||\nabla f(y_i + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j}|| ||s_{T_1+j}|| + s_{T_1+j}^T [\nabla m(y_i, s_{T_1+j}, \sigma_{T_1+j}) - \sigma_{T_1+j} ||s_{T_1+j}|| s_{T_1+j}]
\]
\[
    \leq ||\nabla f(y_i + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j}|| ||s_{T_1+j}|| - \sigma_{T_1+j} ||s_{T_1+j}||^3 + \kappa \theta ||s_{T_1+j}||^3
\]
\[
    = \left( \int_0^1 [\nabla^2 f(y_i + \tau \cdot s_{T_1+j}) - \nabla^2 f(y_i)] s_{T_1+j} \, d\tau \right) ||s_{T_1+j}||^3 + \kappa \theta ||s_{T_1+j}||^3
\]
\[
    \leq \left( \frac{L_h}{2} + \kappa \theta - \sigma_{T_1+j} \right) ||s_{T_1+j}||^3,
\]
where the last inequality is due to Assumption 2.1. Then it follows that
\[
    -\frac{s_{T_1+j}^T \nabla f(y_i + s_{T_1+j})}{||s_{T_1+j}||^3} \geq \sigma_{T_1+j} - \frac{L_h}{2} - \kappa \theta.
\]
Therefore, we have
\[ \sigma_{T_1+j} \geq \frac{L_h}{2} + \kappa \theta + \eta \implies -\frac{s_{T_1+j}^\top \nabla f(y_l + s_{T_1+j})}{\|s_{T_1+j}\|^3} \geq \eta, \]
which further implies that
\[ \sigma_{T_1+j+1} \leq \sigma_{T_1+j} \leq \gamma_2 \cdot \sigma_{T_1+j-1} \leq \gamma_2 \left( \frac{L_h}{2} + \kappa \theta + \eta \right), \quad \forall \ j \in S. \]

Therefore, we can define \( \bar{\sigma}_2 = \max \{ \bar{\sigma}_1, \gamma_2 L_h + \gamma_2 \kappa \theta + \gamma_2 \eta \} \), where the term \( \bar{\sigma}_1 \) accounts for an upper bound of \( \sigma_{T_1} \). In addition, it follows from the construction of Algorithm \( \square \) that \( \sigma_{\min} \leq \sigma_{T_1+j} \) for all iterations, and \( \gamma_1 \sigma_{T_1+j} \leq \sigma_{T_1+j+1} \) for all unsuccessful iterations. Therefore, we have
\[ \bar{\sigma}_2 \sigma_{\min} \geq \sigma_{T_1}, \]
and hence
\[ |S| \leq T_2 \leq |S| + \left( \frac{|S| + 1}{\log \gamma_1} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) |S|. \]

Before estimating the upper bound of \( T_3 \), i.e., \( \frac{1}{\zeta} \) the total number of the count of successfully updating \( z > 0 \), we need the following three technical lemmas.

**Lemma 3.3** For any \( s \in \mathbb{R}^d \) and \( g \in \mathbb{R}^d \), it holds that
\[ s^\top g + \frac{1}{3} \sigma \|s\|^3 \geq -\frac{2}{3\sqrt{\sigma}} \|g\|^2. \]

**Proof.** Denote \( s^* \) as the minimum of \( s^\top g + \frac{1}{3} \sigma \|s\|^3 \). The first-order optimality condition gives that
\[ g + \sigma \|s^*\| s^* = 0. \]
Therefore, we have \( (s^*)^\top g = -\sigma \|s^*\|^3 \) and \( \|g\| = \sigma \|s^*\|^2 \), and
\[ (s^*)^\top g + \frac{1}{3} \sigma \|s^*\|^3 = -\frac{2}{3} \sigma \|s^*\|^3 = -\frac{2}{3\sqrt{\sigma}} \|g\|^2. \]

**Lemma 3.4** Letting \( z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z) \), we have \( \psi_l(z) - \psi_l(z_l) \geq \frac{1}{12 \zeta} \|z - z_l\|^3 \).

**Proof.** It suffices to show that
\[ \psi_l(z) - \psi_l(z_l) - \nabla \psi_l(z_l)^\top (z - z_l) \geq \frac{1}{12 \zeta} \|z - z_l\|^3, \]
Lemma 3.5 For each iteration \( j \) in the subroutine AAS, if it is successful, we have

\[
(1 - \kappa_\theta) \| \nabla f(x_{j+1}) \| \leq \left( \frac{L_h}{2} + \bar{\sigma}_2 + \kappa_\theta L_g \right) \| s_j \|^2,
\]

where \( \kappa_\theta \in (0, 1) \) is used in Condition 3.1.

Proof. We denote \( j \)-th iteration to be the \( l \)-th successful iteration, and note \( \nabla s_m(y_l, s_j, \sigma_j) = \nabla f(y_l) + \nabla^2 f(y_l) s_j + \sigma_j \| s_j \| \cdot s_j \). Then we have

\[
\| \nabla f(x_{j+1}) \| \leq \| \nabla f(y_l + s_j) - \nabla s_m(y_l, s_j, \sigma_j) \| + \| \nabla s_m(y_l, s_j, \sigma_j) \|
\leq \| \nabla f(y_l + s_j) - \nabla s_m(y_l, s_j, \sigma_j) \| + \kappa_\theta \cdot \min (1, \| s_j \|) \cdot \| \nabla f(y_l) \|
\leq \left| \int_0^1 (\nabla^2 f(y_l + \tau s_j) - \nabla^2 f(y_l)) s_j \, d\tau \right| + \sigma_j \| s_j \|^2 + \kappa_\theta \cdot \min (1, \| s_j \|) \cdot \| \nabla f(y_l) \|
\leq \frac{L_h}{2} \| s_j \|^2 + \sigma_j \| s_j \|^2 + \kappa_\theta \| \nabla f(y_l) - \nabla f(y_l + s_j) \| + \kappa_\theta \| \nabla f(x_{j+1}) \|
\leq \frac{L_h}{2} \| s_j \|^2 + \bar{\sigma}_2 \| s_j \|^2 + \kappa_\theta L_g \| s_j \|^2 + \kappa_\theta \| \nabla f(x_{j+1}) \|,
\]

where the second inequality holds true due to Condition 3.1, and the last two inequality follow from Assumption 2.1. Rearranging the terms, the conclusion follows. \( \square \)

Now we are ready to estimate an upper bound of \( T_3 \), i.e., the total number of count of successfully updating \( \varsigma > 0 \).

Lemma 3.6 We have

\[
\psi_l(z_l) \geq \frac{l(l + 1)(l + 2)}{6} f(\bar{x}_l)
\]

if \( \varsigma_l \geq \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right)^3 \frac{1}{\eta^2} \), which further implies that

\[
T_3 \leq \left[ \frac{1}{\log (\gamma_3)} \log \left[ \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right)^3 \frac{1}{\eta^2 \varsigma_1} \right] \right].
\]

Proof. When \( l = 1 \), it trivially holds true that \( \psi_l(z_l) \geq \frac{l(l + 1)(l + 2)}{6} f(\bar{x}_l) \) since \( \psi_1(z_1) = f(\bar{x}_1) \). As a result, it suffices to show that \( \varsigma_l \geq \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right)^3 \frac{1}{\eta^2} \) by mathematical induction. Without loss of
generality, we assume (10) holds true for some \( l - 1 \geq 1 \). Then, it follows from Lemma 3.3 and the construction of \( \psi_l(z) \) that

\[
\psi_{l-1}(z) \geq \psi_{l-1}(z_{l-1}) + \frac{1}{12} \varsigma_{l-1} \| z - z_{l-1} \|^3 \geq \frac{(l-1)l(l+1)}{6} f(\bar{x}_{l-1}) + \frac{1}{12} \varsigma_{l-1} \| z - z_{l-1} \|^3.
\]

As a result, we have

\[
\psi_l(z_l) = \min_{z \in \mathbb{R}^d} \left\{ \psi_{l-1}(z) + \frac{l(l+1)}{2} \left[ f(\bar{x}_l) + (z - \bar{x}_l)^\top \nabla f(\bar{x}_l) \right] + \frac{1}{6} (\varsigma_l - \varsigma_{l-1}) \| z - \bar{x}_1 \|^3 \right\}
\]

\[
\geq \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)l(l+1)}{6} f(\bar{x}_{l-1}) + \frac{1}{12} \varsigma_l \| z - z_{l-1} \|^3 + \frac{l(l+1)}{2} \left[ f(\bar{x}_l) + (z - \bar{x}_l)^\top \nabla f(\bar{x}_l) \right] \right\}
\]

\[
\geq \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)l(l+1)}{6} \left[ f(\bar{x}_l) + (\bar{x}_{l-1} - \bar{x}_l)^\top \nabla f(\bar{x}_l) \right] + \frac{1}{12} \varsigma_l \| z - z_{l-1} \|^3 \right. \\
\left. + \frac{l(l+1)}{2} \left[ f(\bar{x}_l) + (z - \bar{x}_l)^\top \nabla f(\bar{x}_l) \right] \right\}
\]

\[
= \frac{l(l+1)(l+2)}{6} f(\bar{x}_l) + \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)l(l+1)}{6} \left( \bar{x}_{l-1} - \bar{x}_l \right)^\top \nabla f(\bar{x}_l) + \frac{1}{12} \varsigma_l \| z - z_{l-1} \|^3 \right. \\
\left. + \frac{l(l+1)}{2} \left( z - \bar{x}_l \right)^\top \nabla f(\bar{x}_l) \right\}
\]

where the first equality holds since \( \varsigma_l \geq \varsigma_{l-1} \). By the construction of \( y_{l-1} \), we have

\[
\frac{(l-1)l(l+1)}{6} \bar{x}_{l-1} = \frac{l(l+1)(l+2)}{6} \bar{x}_l - \frac{l-1}{l+2} \bar{x}_{l-1}
\]

\[
= \frac{l(l+1)(l+2)}{6} \left( y_{l-1} - \frac{3}{l+2} z_{l-1} \right)
\]

\[
= \frac{l(l+1)(l+2)}{6} y_{l-1} - \frac{l(l+1)}{2} z_{l-1}.
\]

Combining the above two formulas yields

\[
\psi_l(z_l) \geq \frac{l(l+1)(l+2)}{6} f(\bar{x}_l) + \min_{z \in \mathbb{R}^d} \left\{ \frac{l(l+1)(l+2)}{6} \left( y_{l-1} - \bar{x}_l \right)^\top \nabla f(\bar{x}_l) + \frac{1}{12} \varsigma_l \| z - z_{l-1} \|^3 \\
+ \frac{l(l+1)}{2} \left( z - z_{l-1} \right)^\top \nabla f(\bar{x}_l) \right\}
\]

Then, by the criterion of successful iteration in AAS and Lemma 3.5, we have

\[
(y_{l-1} - \bar{x}_l)^\top \nabla f(\bar{x}_l) = -\bar{s}_{T_{l+1}}^\top \nabla f(y_{l-1} + s_{T_{l+1}}) \geq \eta \| s_{T_{l+1}} \|^3
\]

\[
\geq \eta \left( \frac{1 - \kappa_\theta}{L_h + \sigma_2 + \kappa_\theta L_g} \right)^2 \| \nabla f(\bar{x}_l) \| ^\frac{3}{2},
\]

13
where the $l$-th successful iteration count refers to the $(j-1)$-th iteration count in AAS. Hence, it suffices to establish
\[
\frac{l(l+1)(l+2)\eta}{6} \left( \frac{1 - \kappa_\theta}{\frac{L_h}{2} + \sigma_2 + \kappa_\theta L_g} \right)^{\frac{3}{2}} \|\nabla f(\mathbf{x}_l)\|^3 \geq \left( \frac{l(l+1)}{2} \right)^{\frac{3}{2}} (\mathbf{z} - \mathbf{z}_{l-1})^\top \nabla f(\mathbf{x}_l) \geq 0.
\]

Using Lemma 3.3 and setting $g = \frac{l(l+1)}{2} \nabla f(\mathbf{x}_l)$, $s = \mathbf{z} - \mathbf{z}_l$, and $\sigma = \frac{1}{\eta} \varsigma_l$, the above is implied by
\[
\frac{l(l+1)(l+2)\eta}{6} \left( \frac{1 - \kappa_\theta}{\frac{L_h}{2} + \sigma_2 + \kappa_\theta L_g} \right)^{\frac{3}{2}} \geq \frac{4}{3\sqrt{\varsigma_l}} \left( \frac{l(l+1)}{2} \right)^{\frac{3}{2}}. \tag{11}
\]
Therefore, the conclusion follows if
\[
\varsigma_l \geq \left( \frac{L_h + 2\sigma_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right)^3 \frac{1}{\eta^2}.
\]

\[\square\]

### 3.1.3 Iteration Complexity

Recall that $l = 1, 2, \ldots$ is the count of successful iterations, and the sequence $\{\mathbf{x}_l, \ l = 1, 2, \ldots\}$ is updated when a successful iteration is identified. The iteration complexity result is presented in Theorem 3.7 and Theorem 3.8.

**Theorem 3.7** The sequence $\{\mathbf{x}_l, \ l = 1, 2, \ldots\}$ generated by Algorithm 1 satisfies
\[
\frac{l(l+1)(l+2)\eta}{6} f(\mathbf{x}_l) \leq \psi_l(z_l) \leq \psi_l(z)
\]
\[
\leq \frac{l(l+1)(l+2)\eta}{6} f(z) + \frac{L_h + \sigma_1}{3} \|\mathbf{z} - \mathbf{x}_0\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{1}{6 \varsigma_l} \|\mathbf{z} - \mathbf{x}_1\|^3,
\]
where
\[
\sigma_1 = \max \left\{ \sigma_0, \frac{\gamma_2 L_h}{2} \right\} > 0.
\]

**Proof.** The proof is based on mathematical induction. We postpone the base case of $l = 1$ to Theorem 3.9. Suppose that the theorem is true for some $l \geq 1$. Let us consider the case of $l + 1$:
\[
\psi_{l+1}(z_{l+1}) \leq \psi_{l+1}(z)
\]
\[
\leq \frac{l(l+1)(l+2)\eta}{6} f(z) + \frac{L_h + \sigma_1}{3} \|\mathbf{z} - \mathbf{x}_0\|^3 + \frac{1}{6 \varsigma_l} \|\mathbf{z} - \mathbf{x}_1\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2
\]
\[\begin{align*}
\quad &+ \frac{(l + 1)}{2} \left[ f(x_l) + (x_l - x_0)^\top \nabla f(x_l) \right] + \frac{1}{6} (\varsigma_{l+1} - \varsigma_l) \|\mathbf{z} - \mathbf{x}_1\|^3 \\
\quad &\leq \frac{(l + 1)(l + 2)(l + 3)}{6} f(z) + \frac{L_h + \sigma_1}{3} \|\mathbf{z} - \mathbf{x}_0\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{1}{6 \varsigma_{l+1}} \|\mathbf{z} - \mathbf{x}_1\|^3,
\end{align*}
\]
where the last inequality is due to convexity of $f(z)$. On the other hand, it follows from the way that $\psi_{l+1}(z)$ is updated that $(l+1)(l+2)(l+3) f(\bar{x}_{l+1}) \leq \psi_{l+1}(\bar{x}_{l+1})$, and thus Theorem 3.7 is proven. 

After establishing Theorem 3.7, the iteration complexity of Algorithm 1 readily follows.

**Theorem 3.8** The sequence $\{\bar{x}_l, l = 1, 2, \ldots\}$ generated by Algorithm 1 satisfies that

$$f(\bar{x}_l) - f(x^*) \leq \frac{C_1}{l(l+1)(l+2)} \leq \frac{C_1}{l^3},$$

where

$$C_1 = (2L_h + 2\bar{\sigma}_1)\|x_0 - x^*\|^3 + \left(\frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta}\right) \frac{1}{\eta^2} \|x_1 - x^*\|^3 + \frac{12\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}}\|x_0 - x^*\|^2.$$

The total number of iterations required to find $\bar{x}_k$ such that $f(\bar{x}_k) - f(x^*) \leq \epsilon$ is

$$k \leq 1 + \frac{2}{\log(\gamma_1)} \log \left(\frac{\bar{\sigma}_1}{\sigma_{\min}}\right) + \left(1 + \frac{2}{\log(\gamma_1)} \log \left(\frac{\bar{\sigma}_2}{\sigma_{\min}}\right)\right) \left[\left(\frac{C_1}{\epsilon}\right) + 1\right] + \left[\frac{1}{\log(\gamma_3)} \log \left(\frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta}\right)^3 \frac{1}{\eta^2}\right],$$

where

$$\bar{\sigma}_2 = \max \left\{\bar{\sigma}_1, \frac{\gamma_2L_h}{2} + \gamma_2\kappa_\theta + \gamma_2\eta\right\} > 0.$$

**Proof.** By Theorem 3.7 and taking $z = x^*$ we have

$$\frac{l(l+1)(l+2)}{6} f(\bar{x}_l) \leq \frac{l(l+1)(l+2)}{6} f(x^*) + \frac{L_h + \bar{\sigma}_1}{3} \|x^* - x_0\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}}\|x_0 - x^*\|^2 + \frac{1}{6} \|\bar{x}_1 - x^*\|^3.$$ 

Rearranging the terms, and combining with Lemmas 3.1, 3.2, and 3.6 lead to the conclusions. 

Finally let us go back to prove the base case $(l = 1)$ of Theorem 3.7.

**Theorem 3.9** It holds that

$$f(\bar{x}_1) \leq \psi_1(z_1) \leq \psi_1(z) \leq f(z) + \frac{L_h + \bar{\sigma}_1}{3} \|z - x_0\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}}\|x_0 - x^*\|^2 + \frac{1}{6} \|z - \bar{x}_1\|^3.$$

**Proof.** By the definition of $\psi_1(z)$ and the fact that $\bar{x}_1 = x_{T_1}$, we have

$$f(\bar{x}_1) = f(x_{T_1}) = \psi_1(z_1).$$

Furthermore, by the criterion of successful iteration in SAS,

$$f(\bar{x}_1) = f(x_{T_1}) \leq m(x_{T_1 - 1}, s_{T_1 - 1}, \sigma_{T_1 - 1}) \leq m(x_{T_1 - 1}, s_{T_1 - 1}, \sigma_{T_1 - 1}) + m(x_{T_1 - 1}, s_{T_1 - 1}^m, \sigma_{T_1 - 1}),$$
where $s_{T_1-1}^m$ denotes the global minimizer of $m(x_{T_1-1}, s, \sigma_{T_1-1})$ over $\mathbb{R}^d$. Since $f$ is convex, so is $m(x_{T_1-1}, s, \sigma_{T_1-1})$. Therefore, we have

$$m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1}) - m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) \leq \nabla_s m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1})^\top (s_{T_1-1} - s_{T_1-1}^m) \leq \|\nabla_s m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1})\| \|s_{T_1-1} - s_{T_1-1}^m\| \leq \kappa_\theta \|\nabla f(x_{T_1-1})\| \|s_{T_1-1}\| \|s_{T_1-1} - s_{T_1-1}^m\|.$$ 

To bound $\|s_{T_1-1} - s_{T_1-1}^m\|$, we observe that

$$\sigma_{\min} \|s\|^3 \leq \sigma_{T_1-1} \|s\|^3 = s^\top \left[ \nabla m(x_{T_1-1}, s, \sigma_{T_1-1}) - \nabla f(x_{T_1-1}) - \nabla^2 f(x_{T_1-1}) s \right] \leq \|s\| \|\nabla f(x_{T_1-1})\| + \|\nabla m(x_{T_1-1}, s, \sigma_{T_1-1})\| \leq (1 + \kappa_\theta) \|s\| \|\nabla f(x_{T_1-1})\|,$$

where $s = s_{T_1-1}$ or $s = s_{T_1-1}^m$. Thus, we conclude that

$$\|s_{T_1-1} - s_{T_1-1}^m\| \leq \|s_{T_1-1}\| + \|s_{T_1-1}^m\| \leq 2 \sqrt{\frac{(1 + \kappa_\theta) \|\nabla f(x_{T_1-1})\|}{\sigma_{\min}}},$$

which combines with Assumption 2.1 implies that

$$m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) - m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) \leq \frac{2\kappa_\theta (1 + \kappa_\theta)}{\sigma_{\min}} \|\nabla f(x_{T_1-1})\|^2 \leq \frac{2\kappa_\theta (1 + \kappa_\theta) L_\theta^2}{\sigma_{\min}} \|x_{T_1-1} - x^\ast\|^2.$$

On the other hand, we have

$$m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) = f(x_{T_1-1}) + (s_{T_1-1}^m)^\top \nabla f(x_{T_1-1}) + \frac{1}{2} (s_{T_1-1}^m)^\top \nabla^2 f(x_{T_1-1}) s_{T_1-1}^m + \frac{1}{3} \sigma_{T_1-1} \|s_{T_1-1}^m\|^3 \leq f(x_{T_1-1}) + (z - x_{T_1-1})^\top \nabla f(x_{T_1-1}) + \frac{1}{2} (z - x_{T_1-1})^\top \nabla^2 f(x_{T_1-1})(z - x_{T_1-1}) + \frac{1}{3} \sigma_{T_1-1} \|z - x_{T_1-1}\|^3 \leq f(z) + \frac{L_h + \sigma_1}{3} \|z - x_{T_1-1}\|^3 \leq f(z) + \frac{L_h + \sigma_1}{3} \|z - x_0\|^3,$$

where the second inequality is due to (13) and Assumption 2.1. Therefore, we conclude that

$$\psi_1(z) = f(x_1) + \frac{1}{6} \|z - x_1\|^3 \leq f(z) + \frac{L_h + \sigma_1}{3} \|z - x_0\|^3 + \frac{2\kappa_\theta (1 + \kappa_\theta) L_\theta^2}{\sigma_{\min}} \|x_0 - x^\ast\|^2 + \frac{1}{6} \|z - x_1\|^3.$$
3.2 Strongly Convex Case

Next we extend the analysis to the case where the objective function is strongly convex (cf. Definition 2.2). We further assume the level set of \( f(x) \), \( \{x \in \mathbb{R}^d : f(x) \leq f(x_0)\} \), is bounded and is contained in \( \|x - x_*\| \leq D \). Then according to Lemma 3 in [24], we have

\[
\nabla^2 f(x) \succeq \mu I,
\]

(12)

and

\[
f(y) - f(x) - (y - x)^\top \nabla f(x) \leq \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)\|^2.
\]

(13)

We shall prove the improvement of the adaptive acceleration scheme in terms of the constant underlying the linear rate of convergence. To this end, denote \( A_m^1(x) \) \((m \geq 1)\) to be the point generated by running \( m \) iterations of Algorithm 1 with starting point \( x \). Then, generate sequence \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) through the following procedure

1. Define

\[
m = 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) + \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) \left[ 2 \left( \frac{\tau_1 D + \tau_2}{\mu} \right) \frac{1}{3} + 1 \right] + \left[ \frac{1}{\log(\gamma_3)} \log \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right) \right] \frac{1}{\eta^2 \varsigma_1},
\]

with

\[
\tau_1 = 2L_h + 2\bar{\sigma}_1 + \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right) \frac{1}{\eta^2} \quad \text{and} \quad \tau_2 = \frac{12\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}}.
\]

2. Set \( \hat{x}_0 \in \mathbb{R}^d \).

3. For \( k \geq 0 \), iterate \( \hat{x}_k = A_m^1(\hat{x}_{k-1}) \).

The linear convergence of \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) is presented in the following theorem.

**Theorem 3.10** Suppose the sequence \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) is generated by the procedure above. For \( k \geq O \left( \log \left( \frac{1}{\epsilon} \right) \right) \) we have \( f(\hat{x}_k) - f(x^*) \leq \epsilon \). Specifically, the total number of iterations required to find such solution is \( O \left( \sqrt[3]{\max \left\{ \frac{L_g}{\mu}, \frac{L_h}{\mu} \right\} \log \left( \frac{1}{\epsilon} \right)} \right) \).

**Proof.** By Theorem 3.7 we have

\[
f(\hat{x}_{k+1}) - f(x^*) \\
\leq \frac{1}{m^3} \left[ 2L_h + 2\bar{\sigma}_1 + \left( \frac{L_h + 2\bar{\sigma}_2 + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right) \frac{1}{\eta^2} \right] \|\hat{x}_k - x^*\|^3 + \frac{12\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}} \|\hat{x}_k - x^*\|^2
\]

17
where the number of successful iteration \( m = \left( \frac{\tau_1 D + \frac{\tau_2}{\sigma_{\min}}}{{\mu}^2} \right)^{1/3} \geq \left( \frac{\tau_1 \|x_k - x^*\| + \frac{\tau_2}{\sigma_{\min}}}{{\mu}^2} \right)^{1/3} \).

Combining this with \([5]\) implies that

\[
 f(\hat{x}_{k+1}) - f(x^*) \leq \frac{\mu}{8} \|\hat{x}_k - x^*\|^2 \leq \frac{1}{4} (f(\hat{x}_k) - f(x^*)) ,
\]

which proves the first part of the conclusion. Then the total iteration is

\[
 m \cdot \log \left( \frac{1}{\epsilon} \right) = O \left( \frac{\sqrt{\max \{ L_g, L_h \} \log \left( \frac{1}{\epsilon} \right)}}{\mu} \right),
\]

where we want to explore how the iteration complexity dependent on the conditional number \( \frac{L_u}{\mu} \) and \( \frac{L_h}{\mu} \), and the Lipschitz parameters \( L_g \) and \( L_h \) that are not coupling with \( \mu \) are treated as constants. \( \square \)

**Remark 3.11** Remark that comparing to \([6]\), the accelerated scheme has improved the dependence of the conditional number from \( O \left( \sqrt{\gamma} \right) \) to \( O \left( \sqrt[3]{\gamma} \right) \).

Furthermore, when the objective function is strongly convex, the local quadratic convergence is retained by our adaptive scheme even without solving the cubic sub-problem exactly if we set \( 0 < \kappa_\theta \leq \frac{\mu}{2} \). Indeed, we can construct sequence \( \{w_l, \; l = 1, 2, \ldots\} \) such that \( w_{l+1} = w_l + \tilde{s}_l \) and \( \tilde{s}_l \) is obtained by running the subroutine SAS of Algorithm \([4]\) with starting point \( w_l \). Then it holds that

\[
 f(w_l) - f(w_{l+1}) \geq f(w_l) - m(w_l, \tilde{s}_l, \tilde{\sigma}_l) \\
 = -\tilde{s}_l^T \nabla f(w_l) - \frac{1}{2} \tilde{s}_l \nabla^2 f(w_l) \tilde{s}_l - \frac{\tilde{\sigma}_l}{3} \|\tilde{s}_l\|^3 \\
 = -\tilde{s}_l^T \nabla m(w_l, \tilde{s}_l, \tilde{\sigma}_l) + \frac{1}{2} \tilde{s}_l \nabla^2 f(w_l) \tilde{s}_l + \frac{2\tilde{\sigma}_l}{3} \|\tilde{s}_l\|^3 \geq \frac{\mu - 2\kappa_\theta}{2} \|\tilde{s}_l\|^2 \\
 \geq \frac{(\mu - 2\kappa_\theta) (1 - \kappa_\theta)}{2(l_\mu \frac{\mu}{2} + \bar{\sigma}_2 + \kappa_\theta L_g)} \|\nabla f(w_{l+1})\| \quad \text{Lemma 3.3} \\
 \geq \frac{(\mu - 2\kappa_\theta) (1 - \kappa_\theta)}{2(l_\mu \frac{\mu}{2} + \bar{\sigma}_2 + \kappa_\theta L_g)} \sqrt{2\mu (f(w_{l+1}) - f(x^*))} \quad \text{Lemma 3.3}
\]

Hence,

\[
 f(w_{l+1}) - f(x^*) \leq \frac{2(l_\mu \frac{\mu}{2} + \bar{\sigma}_2 + \kappa_\theta L_g)^2}{\mu(\mu - 2\kappa_\theta)^2 (1 - \kappa_\theta)^2} (f(w_l) - f(w_{l+1}))^2 \leq \frac{2(l_\mu \frac{\mu}{2} + \bar{\sigma}_2 + \kappa_\theta L_g)^2}{\mu(\mu - 2\kappa_\theta)^2 (1 - \kappa_\theta)^2} (f(w_l) - f(x^*))^2 ,
\]

and the region of quadratic convergence is given by

\[
 Q = \left\{ w \in \mathbb{R}^d : f(w) - f(x^*) \leq \frac{\mu(\mu - 2\kappa_\theta)^2 (1 - \kappa_\theta)^2}{2(l_\mu \frac{\mu}{2} + \bar{\sigma}_2 + \kappa_\theta L_g)^2} \right\}.
\]

The above discussion suggests that we can first run Algorithm \([4]\) until the generated sequence fall into the local quadratic convergence region \( Q \), and then switch back and stick to SAS by allowing performing multiple successful iterations. This way, one would still benefit from the accelerated global convergence rate before local quadratic convergence becomes effective.
4 Accelerated Adaptive Cubic Regularization with Inexact Hessian

In this section, we study the scenario where the Hessian information is not available; instead, an approximation is used, based on the gradient information. Indeed, as illustrated in Table 2 we consider the following approximation of $f$ evaluated at $x_i$ with cubic regularization:

$$m(x_i, s, \sigma) = f(x_i) + s^T \nabla f(x_i) + \frac{1}{2} s^T H(x_i) s + \frac{1}{3} \sigma \|s\|^3,$$

(14)

where $\sigma > 0$ is a regularized parameter, and $H(x_i)$ is an approximation of the Hessian $\nabla^2 f(x_i)$, i.e., the inexact Hessian. In particular, the inexact Hessian $H(x_i)$ can be computed by first computing $d$ forward gradient differences at $x_i$ with stepsize $h_i \in \mathbb{R}$,

$$A_i = \begin{bmatrix} \frac{\nabla f(x_i + h_i e_1) - \nabla f(x_i)}{h_i} & \cdots & \frac{\nabla f(x_i + h_i e_d) - \nabla f(x_i)}{h_i} \end{bmatrix},$$

symmetrizing the resulting matrix: $\hat{H}(x_i) = \frac{1}{2} \left( A_i + A_i^T \right)$ and then further adding an constant multiple of identity matrix to $\hat{H}(x_i)$: $H(x_i) = \hat{H}(x_i) + \kappa_i h_i I$, where $e_j$ is the $j$-th vector of the canonical basis. It is well known in [28] that, for some constant $\kappa_e > 0$, we have $\|\hat{H}(x_i) - \nabla^2 f(x_i)\| \leq \kappa_e h_i$. Consequently, it holds that

$$\|H(x_i) - \nabla^2 f(x_i)\| \leq (\kappa_e + \kappa_c) h_i. \quad (15)$$

That is to say, the gap between exact and inexact Hessian can be bounded by a multiple of the stepsize $h_i$. This together with Algorithm 4.1 in [9] inspires us to design a procedure to search a pair of $(h_i, s_i)$ such that, for some $\kappa_{hs} > 0$,

$$h_i \leq \kappa_{hs} \|s_i\|. \quad (16)$$

Combining (15) and (16) yields that

$$\|H(x_i) - \nabla^2 f(x_i)\| \leq (\kappa_e + \kappa_c) \kappa_{hs} \|s_i\|. \quad (17)$$

Moreover, since $f$ is convex, we set $\kappa_e \geq \kappa_c$ such that

$$H(x_i) = \hat{H}(x_i) + \kappa_e h_i I \succeq \nabla^2 f(x_i) - \kappa_e h_i I + \kappa_c h_i I \succeq 0. \quad (18)$$

Now we propose the accelerated adaptive cubic regularization of Newton’s method with inexact Hessian in Algorithm 4.1. In each iteration we instead approximately solve

$$s_i \approx \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i),$$

where $m(x_i, s, \sigma_i)$ is defined in (14) and the symbol “$\approx$” is quantified in Condition 3.1. and (17) is a key property that will be used in the iteration complexity analysis for Algorithm 4.

4.1 The Convex Case

In this subsection, we aim to analyze the theoretical performance of Algorithm 4. The main difference between Algorithm 2 and Algorithm 1 is an extra inner loop to update $\{h_{i,k}, i = 0, 1, 2, \ldots\}$. We
Algorithm 2 Accelerated Adaptive Cubic Regularization for Newton’s Method with Inexact Hessian

Given $\gamma_2 > \gamma_1 > 1$, $\gamma_1 > 1$, $\gamma_4 \in (0, 1)$, and $\sigma_{\min} > 0$. Specify $m(x_i, s, \sigma_i)$ as in Table 2.

Choose $x_0 \in \mathbb{R}^d$, $\sigma_0 \geq \sigma_{\min}$, $h_0, 0 \in (0, 1)$, and $\varsigma_1 > 0$.

Begin Phase I: Simple Adaptive Subroutine (SAS)

for $i = 0, 1, 2, \ldots$
do
\hspace{1em}for $k = 0, 1, 2, \ldots$
do
\hspace{2em}Compute $H_k(x_i)$ using the finite difference with stepsize $h_{i,k}$ and the iterate $x_i$.
\hspace{2em}Compute $s_{i,k} \in \mathbb{R}^d$ such that $s_{i,k} \approx \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i)$ with the inexact Hessian $H_k(x_i)$.
\hspace{2em}if $h_{i,k} > \kappa_h \left\| s_{i,k} \right\|$ then
\hspace{3em}Let $h_{i,k+1} = \gamma_i h_{i,k}$;
\hspace{3em}else
\hspace{4em}Set $s_i = s_{i,k}$ and $h_i = h_{i,k}$;
\hspace{3em}break.
end if
\hspace{2em}end if
\hspace{2em}end for
\hspace{1em}end for

End Phase I: Simple Adaptive Subroutine

Begin Phase II: Accelerated Adaptive Subroutine (AAS)

Set the count of successful iterations $l = 1$ and let $x_1 = x_{T_1}$.

Construct $\psi(z) = f(x_1) + \frac{1}{2\varsigma_1}(z - x_1)^3$, and let $z_1 = \arg\min_{z \in \mathbb{R}^d} \psi_1(z)$, and compute $y_1 = \frac{1}{\varsigma_1}x_1 + \frac{3}{4}z_1$.

for $j = 0, 1, 2, \ldots$
do
\hspace{1em}for $k = 0, 1, 2, \ldots$
do
\hspace{2em}Compute $H_k(y_l)$ using the finite difference with stepsize $h_{T_1+j,k}$ and the iterate $y_l$.
\hspace{2em}Compute $s_{T_1+j,k} \in \mathbb{R}^d$ such that $s_{T_1+j,k} \approx \arg\min_{s \in \mathbb{R}^d} m(y_l, s, \sigma_{T_1+j})$ with the inexact Hessian $H_k(y_l)$.
\hspace{2em}if $h_{T_1+j,k} > \kappa_h \left\| s_{T_1+j,k} \right\|$ then
\hspace{3em}Let $h_{T_1+j,k+1} = \gamma_i h_{T_1+j,k}$;
\hspace{3em}else
\hspace{4em}Set $s_{T_1+j,k} = s_{T_1+j,k}$ and $h_{T_1+j,k} = h_{T_1+j,k}$;
\hspace{4em}break.
\hspace{4em}end if
\hspace{2em}end if
\hspace{2em}end for
\hspace{1em}end for

Set $h_{T_1+j+1,0} = h_{T_1+j}$, and compute $\rho_{T_1+j} = -\frac{s_{T_1+j}^T \nabla f(y_{l+j} + s_{T_1+j})}{\left\| s_{T_1+j} \right\|^2}$.

if $\rho_{T_1+j} \geq \eta$ [successful iteration] then
\hspace{1em}Let $x_{T_1+j+1} = y_l + s_{T_1+j}$ and choose $\sigma_{T_1+j+1} \in [\sigma_{\min}, \sigma_{T_1+j}]$.
\hspace{1em}Set $l = l + 1$ and $\varsigma = \varsigma_{l-1}$;
\hspace{1em}Update $\psi(z)$ as illustrated in Table 3 by using $\varsigma = \varsigma$, and compute $z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z)$.
\hspace{1em}while $\psi_l(z) \geq \frac{l(l+1)}{4} f(y_l)$ do
\hspace{2em}Let $\varsigma = \gamma_1 \varsigma$, and $\psi(z) = \psi_{l-1}(z) + \frac{l(l+1)}{2} \left[ f(x_{T_1+j+1}) + (z - x_{T_1+j+1})^T \nabla f(x_{T_1+j+1}) \right] + \frac{\varsigma}{4} (\varsigma - \varsigma_{l-1}) \left\| z - z_l \right\|^3$;
\hspace{2em}Compute $z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z)$.
\hspace{2em}end while
\hspace{1em}end if
\hspace{1em}end if

Record the total number of iterations of AAS: $T_2 = j + 1$.

End Phase II: Accelerated Adaptive Subroutine
Lemma 4.1 When $\epsilon$ is sufficiently small, the total number of iterations $T_4$ in the inner loop can not exceed
\[
\left\lceil -\frac{1}{\log(\gamma_4)} \log \left[ \frac{(L_g + (\kappa_c + \kappa_e)\kappa_{hs} + \bar{\sigma}_2)h_{0,0}}{(1 - \kappa_0)\kappa_{hs}} \cdot \frac{1}{\epsilon} \right] \right\rceil.
\]

Proof. Note that before Algorithm 2 terminates, we always have $\|\nabla f(x_t)\| \geq \epsilon$ for any iterate $x_t$ in the process except the last one and $\{h_{i,k}, i, k = 0, 1, 2, \ldots\}$ is not updated for the last iterate. We let $x_j$ be the second last iterate before termination with $s_j = s_{j,k}$ and stepsize $h_{j,k}$ for Hessian approximation. According to Condition 3.1, we have
\[
\|\nabla f(x_j) + H_j s_{j,k} + \sigma_j \|= s_j \| s_{j,k} \| \leq \kappa_0 \min(1, \| s_{j,k} \|) \cdot \| f(x_j) \|,
\]
which implies that
\[
\kappa_0 \| \nabla f(x_i) \| \geq \| \nabla f(x_j) \| - \| H_j s_{j,k} + \sigma_j \|= s_j \| s_{j,k} \|.
\]
As a result,
\[
(1 - \kappa_0) \| \nabla f(x_i) \| \leq \| H_j s_{j,k} + \sigma_j \|= s_j \| s_{j,k} \| \leq \| H_j - \nabla^2 f(x_j) \| \cdot \| s_{j,k} \| + \| \nabla^2 f(x_j) \| \cdot \| s_{j,k} \| + \| s_{j,k} \|^2 \leq (\kappa_c + \kappa_e)\kappa_{hs} \| s_{j,k} \|^2 + L_g \| s_{j,k} \| + \sigma_j \| s_{j,k} \|^2.
\]
where the third inequality is due to mean value theorem and (17). Consequently,
\[
\epsilon \leq \| \nabla f(x_j) \| \leq \frac{L_g + (\kappa_c + \kappa_e)\kappa_{hs} + \bar{\sigma}_2}{1 - \kappa_0} \| s_{j,k} \|
\]
with $\bar{\sigma}_2$ is defined in Lemma A.2, and thus we have
\[
\min \left\{ 1, \frac{\epsilon(1 - \kappa_0)}{L_g + (\kappa_c + \kappa_e)\kappa_{hs} + \bar{\sigma}_2} \right\} \leq \| s_{j,k} \|
\]
That is $\| s_{j,k} \|$ has a constant lower bound. Since $\{h_{i,k}, i, k = 0, 1, 2, \ldots\}$ is a monotonically decreasing sequence, $h_{i,k}$ will not be updated as long as $h_{0,0}\gamma_4^{T_4} \leq \kappa_{hs} \| s_{j,k} \|$, and according to (19) with a sufficiently small $\epsilon$ this can be achieved by letting
\[
T_4 = \left\lceil -\frac{1}{\log(\gamma_4)} \log \left[ \frac{(L_g + (\kappa_c + \kappa_e)\kappa_{hs} + \bar{\sigma}_2)h_{0,0}}{(1 - \kappa_0)\kappa_{hs}} \cdot \frac{1}{\epsilon} \right] \right\rceil.
\]

Recall that $l = 1, 2, \ldots$ is the count of successful iterations, and the sequence $\{x_l, l = 1, 2, \ldots\}$ is updated when a successful iteration is identified. The iteration complexity result is presented in Theorem 4.2 and Theorem 4.3.
Theorem 4.2 The sequence \( \{\tilde{x}_l, l = 1, 2, \ldots\} \) generated by Algorithm \( \mathcal{A} \) satisfies

\[
\frac{l(l+1)(l+2)}{6} f(x_l) \leq \psi_l(z_l) \leq \psi_l(z)
\]

\[
\leq \frac{l(l+1)(l+2)}{6} f(z) + \frac{L_h + \tilde{\sigma}_1 + (\kappa_e + \kappa_c)\kappa_{hs}}{2} \|z - x_0\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min} - (\kappa_e + \kappa_c)\kappa_{hs}} \|x_0 - x^*\|^2 + \frac{1}{6} \sigma \|z - x_1\|^3,
\]

where

\[
\tilde{\sigma}_1 = \max \left\{ \sigma_0, \frac{3\gamma_2L_h + \gamma_2(\kappa_e + \kappa_c)\kappa_{hs}}{2} \right\} > 0.
\]

Proof. The proof is based on mathematical induction. The base case of \( l = 1 \) can be found in Theorem A.3. Suppose that the theorem is true for some \( l \geq 1 \). Let us consider the case of \( l + 1 \):

\[
\psi_{l+1}(z_{l+1}) \leq \psi_{l+1}(z)
\]

\[
\leq \frac{l(l+1)(l+2)}{6} f(z) + \frac{L_h + \tilde{\sigma}_1 + (\kappa_e + \kappa_c)\kappa_{hs}}{2} \|z - x_0\|^3 + \frac{1}{6} \sigma \|z - x_1\|^3 + \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min} - (\kappa_e + \kappa_c)\kappa_{hs}} \|x_0 - x^*\|^2
\]

\[
+ \frac{(l+1)(l+2)}{2} \left[ f(x_l) + (z - x_l)^\top \nabla f(x_l) \right] + \frac{1}{6} (\sigma_{l+1} - \sigma) \|z - x_1\|^3
\]

\[
\leq \frac{(l+1)(l+2)(l+3)}{6} f(z) + \frac{L_h + \tilde{\sigma}_1 + (\kappa_e + \kappa_c)\kappa_{hs}}{2} \|z - x_0\|^3
\]

\[
+ \frac{2\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min}} \|x_0 - x^*\|^2 + \frac{1}{6} \sigma_{l+1} \|z - x_1\|^3,
\]

where the last inequality is due to convexity of \( f(z) \). On the other hand, it follows from the way that \( \psi_{l+1}(z) \) is updated that \( \frac{(l+1)(l+2)(l+3)}{6} f(x_{l+1}) \leq \psi_{l+1}(z_{l+1}) \), and thus Theorem 4.2 is proven. \( \square \)

The established Theorem 4.2 implies the following main result on iteration complexity of Algorithm \( \mathcal{A} \).

Theorem 4.3 The sequence \( \{\tilde{x}_l, l = 1, 2, \ldots\} \) generated by Algorithm \( \mathcal{A} \) satisfies that

\[
f(\tilde{x}_l) - f(x^*) \leq \frac{C_2}{l(l+1)(l+2)} \leq \frac{C_2}{l^3},
\]

where

\[
C_2 = \left( 3L_h + 3\tilde{\sigma}_1 + 3(\kappa_e + \kappa_c)\kappa_{hs} \right) \|x_0 - x^*\|^3 + \left( \frac{L_h + 2\tilde{\sigma}_2 + 2(\kappa_e + \kappa_c)\kappa_{hs} + 2\kappa_\theta L_g}{1 - \kappa_\theta} \right)^3 \frac{1}{\eta^3} \|x_1 - x^*\|^3
\]

\[
+ \frac{12\kappa_\theta(1 + \kappa_\theta)L_g^2}{\sigma_{\min} - (\kappa_e + \kappa_c)\kappa_{hs}} \|x_0 - x^*\|^2.
\]

When \( \epsilon \) is sufficiently small, the total number of iterations required to find \( \tilde{x}_k \) such that \( f(x_k) - f(x^*) \leq \)
max\{\epsilon, \epsilon D\} \text{ is}

\[ k \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) + \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) \left[ \left( \frac{C_2}{\epsilon} \right)^{\frac{1}{\delta}} + 1 \right]
+ \frac{1}{\log(\gamma_3)} \log \left( \frac{L_{\bar{x}} + 2\bar{\sigma}_2 + 2(\kappa_c + \kappa_c)\kappa_{hs} + 2\kappa_{\theta} L_g}{1 - \kappa_{\theta}} \right)^{\frac{3}{\gamma}} \frac{1}{\eta^2 \kappa_1} \right]
+ \left[ \left( \frac{L_{\bar{x}} + (\kappa_c + \kappa_c)\kappa_{hs} + \bar{\sigma}_2)h_{0,0}}{(1 - \kappa_{\theta})\kappa_{hs}} \right) \frac{1}{\epsilon} \right],
\]

where

\[ \bar{\sigma}_2 = \max \left\{ \frac{\gamma_2 L_h}{2} + \gamma_2 \kappa_{\theta} + \gamma_2 (\kappa_c + \kappa_c)\kappa_{hs} + \gamma_2 \eta \right\} > 0. \]

\textbf{Proof.} By Theorem 4.2 and taking \( z = x^* \) we have

\[ \frac{l(l + 1)(l + 2)}{6} f(\bar{x}_l) \leq \frac{l(l + 1)(l + 2)}{6} f(x^*) + \frac{L_{\bar{x}} + \bar{\sigma}_1 + (\kappa_c + \kappa_c)\kappa_{hs}}{2} \|x^* - x_0\|^3 + \frac{2\kappa_{\theta}(1 + \kappa_c) L^2}{\sigma_{\min}} \|x_0 - x^*\|^2 + \frac{1}{6} \|x^* - \bar{x}_l\|^3. \]

Rearranging the terms, and combining with Lemmas A.1, A.2 and A.4 yields the conclusions. \( \square \)

\subsection{Strongly Convex Case}

Next we extend the analysis to the case where the objective function is strongly convex. We further assume the level set of \( f(x) \), \( \{x \in \mathbb{R}^d : f(x) \leq f(x_0)\} \), is bounded and is contained in \( \|x - x_*\| \leq D \). We denote \( A_m^2(x) \), \( m \geq 1 \), as the point generated by running \( m \) outer loop iterations of Algorithm 2. Assume that \( 0 < \kappa_{\theta} < \mu \), we show that the accelerated adaptive cubic regularization for Newton’s method has a linear convergence rate. In particular, we can generate sequence \( \{\bar{x}_k, k = 0, 1, 2, \ldots\} \) through the following procedure:

1. Define

\[ m = 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) + \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) \left[ \left( \frac{1}{\mu} \right)^{\frac{1}{\delta}} + 1 \right]
+ \frac{1}{\log(\gamma_3)} \log \left( \frac{L_{\bar{x}} + 2\bar{\sigma}_2 + 2(\kappa_c + \kappa_c)\kappa_{hs} + 2\kappa_{\theta} L_g}{1 - \kappa_{\theta}} \right)^{\frac{3}{\gamma}} \frac{1}{\eta^2 \kappa_1} \right]
+ \left[ \left( \frac{L_{\bar{x}} + (\kappa_c + \kappa_c)\kappa_{hs} + \bar{\sigma}_2)h_{0,0}}{(1 - \kappa_{\theta})\kappa_{hs}} \right) \frac{1}{\epsilon} \right],
\]

with

\[ \tau_1 = 3L_2 + 3\bar{\sigma}_1 + 3(\kappa_c + \kappa_c)\kappa_{hs} + \left( \frac{L_2 + 2\bar{\sigma}_2 + 2\kappa_{\theta} L_1}{1 - \kappa_{\theta}} \right) \frac{1}{\eta^2} \]
and \( \tau_2 = \frac{12\kappa_{\theta}(1 + \kappa_c) L^2}{\sigma_{\min}}. \)

2. Set \( \bar{x}_0 \in \mathbb{R}^d \).
3. For \( k \geq 0 \), iterate \( \mathbf{x}_k = A^2_m(\mathbf{x}_{k-1}) \).

The theoretical guarantee of the above procedure can be described by the following theorem, whose proof is identical to that of Theorem 3.10 and thus omitted.

**Theorem 4.4** Suppose the sequence \( \{x_k, k = 0, 1, 2, \ldots\} \) is generated by the procedure above. For \( k \geq O(\log(\frac{1}{\epsilon})) \) we have \( f(x_k) - f(x^*) \leq \epsilon. \) Specifically, the total number of iterations required to find such solution is \( \tilde{O}(\sqrt{\max\left\{ \frac{L_{u}}{\mu}, \frac{L_{b}}{\mu} \right\} \log(\frac{1}{\epsilon})}) \).

**Remark 4.5** Theorem 4.4 implies a surprising result that, in view of the order of iteration complexity, the accelerated adaptive cubic regularization method for Newton’s method remains even with inexact Hessians estimated from the gradients. Specifically, it still has an \( O\left( \sqrt{T} \right) \) dependence on the conditional numbers \( \frac{L_u}{\mu} \) and \( \frac{L_b}{\mu} \). However, we need to set \( 0 < \kappa_\theta < \mu \) where \( \mu \) is unknown in practice.

Furthermore, we can construct sequence \( \{x_l, l = 1, 2, \ldots\} \) such that \( z_{l+1} = z_l + \mathbf{s}_l \) and \( z_l \) is obtained by running SAS of Algorithm 2 with initial point \( z_l \). Recall that \( \sigma_{\text{min}} \geq \kappa_c \kappa_{hs} \), and so

\[
\begin{align*}
f(z_l) - f(z_{l+1}) & \geq f(z_l) - m(z_l, s_l, \sigma_l) \\
& = -s_l^\top \nabla f(z_l) - \frac{1}{2} s_l^\top H(z_l) s_l - \frac{\sigma_l}{3} \|s_l\|^3 \\
& = -s_l^\top m(z_l, s_l, \sigma_l) + \frac{1}{2} s_l^\top H(z_l) s_l + \frac{2\sigma_l}{3} \|s_l\|^3 \\
& \geq \frac{1}{2} s_l^\top H(z_l) s_l - \kappa_\theta \|s_l\|^2 + \frac{2\sigma_{\text{min}}}{3} \|s_l\|^3 \\
& \geq \frac{1}{2} s_l^\top \nabla^2 f(z_l) s_l - \kappa_\theta \|s_l\|^2 + \frac{1}{2} s_l^\top (H(z_l) - \nabla^2 f(z_l)) s_l + \frac{2\sigma_{\text{min}}}{3} \|s_l\|^3 \\
& \geq \frac{\mu - 2\kappa_\theta}{2} \|s_l\|^2 + \left( \frac{2\sigma_{\text{min}}}{3} - \frac{\kappa_c \kappa_{hs}}{2} \right) \|s_l\|^3
\end{align*}
\]

Lemma 4.3

\[
\begin{align*}
& \geq \frac{(\mu - 2\kappa_\theta)(1 - \kappa_\theta)}{2(\frac{L_u}{\mu} + \sigma_2 + (\kappa_c + \kappa_c) \kappa_{hs} + \kappa_\theta L_g)} \| \nabla f(z_{l+1}) \| \\
& \geq \frac{(\mu - 2\kappa_\theta)(1 - \kappa_\theta)}{2(\frac{L_b}{\mu} + \sigma_2 + (\kappa_c + \kappa_c) \kappa_{hs} + \kappa_\theta L_g)} \sqrt{2\mu(f(z_{l+1}) - f(x^*))}
\end{align*}
\]

where \( \kappa_\theta \in (0, 1) \) is defined in Condition 3.11. Hence, we have

\[
f(z_{l+1}) - f(x^*) \leq \frac{2(\frac{L_u}{\mu} + \sigma_2 + (\kappa_c + \kappa_c) \kappa_{hs} + \kappa_\theta L_g)^2}{\mu(\mu - 2\kappa_\theta)^2(1 - \kappa_\theta)^2} (f(z_l) - f(z_{l+1}))^2 \leq \frac{2(\frac{L_b}{\mu} + (\kappa_c + \kappa_c) \kappa_{hs} + \sigma_2 + \kappa_\theta L_g)^2}{\mu(\mu - 2\kappa_\theta)^2(1 - \kappa_\theta)^2} (f(z_l) - f(x^*))^2
\]

and the region of quadratic convergence is given by

\[
Q = \left\{ z \in \mathbb{R}^d : f(z) - f(x^*) \leq \frac{\mu(\mu - 2\kappa_\theta)^2(1 - \kappa_\theta)^2}{2(\frac{L_b}{\mu} + (\kappa_c + \kappa_c) \kappa_{hs} + \sigma_2 + \kappa_\theta L_g)^2} \right\}.
\]
5 Accelerated Adaptive Gradient Method

In this section, we present an accelerated adaptive gradient method that is fully Lipschitz-constant-free. In particular, we consider the following standard approximation of $f$ evaluated at $x_i$ with quadratic regularization:

$$m(x_i, s, \sigma_i) = f(x_i) + s^T \nabla f(x_i) + \frac{1}{2} \sigma_i \|s\|^2,$$

where $\sigma_i > 0$ is a regularized parameter. Then our algorithms are described in Algorithm 3.

**Algorithm 3 Accelerated Gradient Method with Adaptive Quadratic Regularization**

Given $\gamma_1 \geq 1$, $\gamma_2 \geq 1$, $\eta > 0$, and $\sigma_{\text{min}} > 0$. Choose $x_0 \in \mathbb{R}^d$, $\sigma_0 \geq \sigma_{\text{min}}$, and $c_1 > 0$.

**Begin Phase I: Simple Adaptive Subroutine (SAS)**

for $i = 1, 2, \ldots$ do

Compute $s_i \in \mathbb{R}^d$ such that $s_i = \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i)$;

Compute $\rho_i = f(x_i + s_i) - m(x_i, s_i, \sigma_i)$;

if $\rho_i < 0$ [successful iteration] then

$x_{i+1} = x_i + s_i$ and $\sigma_{i+1} \in [\sigma_{\text{min}}, \sigma_i]$;

break.

Record the total number of iterations in SAA: $T_1 = i + 1$;

else

$x_{i+1} = x_i$ and $\sigma_{i+1} \in [\gamma_1 \sigma_i, \gamma_2 \sigma_i]$;

end if

end for

End Phase I: Simple Adaptive Subroutine

**Begin Phase II: Accelerated Adaptive Subroutine (AAS)**

Set the count of successful iterations $l = 1$ and let $\bar{x}_1 = x_{T_1}$;

Construct $\psi(z) = f(x_1) + \frac{1}{\eta} \|z - x_1\|^2$, and let $z_1 = \arg\min_{z \in \mathbb{R}^d} \psi_1(z)$, and choose $y_1 = \frac{1}{2} \bar{x}_1 + \sigma_1 z_1$;

for $j = 1, 2, \ldots$ do

Compute $s_{T_1 + j} = \arg\min_{s \in \mathbb{R}^d} m(y_1, s, \sigma_{T_1 + j})$, and $\rho_{T_1 + j} = \frac{\langle \nabla f(y_1 + s_{T_1 + j}) \rangle}{\|y_{T_1 + j} + s_{T_1 + j}\|}$;

if $\rho_{T_1 + j} \geq \eta$ [successful iteration] then

$x_{T_1 + j + 1} = y_1 + s_{T_1 + j}$, $\sigma_{T_1 + j + 1} \in [\sigma_{\text{min}}, \sigma_{T_1 + j}]$;

Set $l = l + 1$ and $\zeta = \zeta - 1$;

Update $\psi(z)$ as illustrated in Table 3 by using $\zeta = \zeta$, and compute $z_l = \arg\min_{z \in \mathbb{R}^d} \psi(z)$;

while $\psi(z_l) \geq \frac{l(l+1)}{2} f(x_l)$ do

Set $\zeta = \gamma_1 \zeta$, and $\psi_l(z_l) = \psi_{l-1}(z_l) + \frac{l(l+1)}{2} \left[ f(x_{T_1 + j + 1}) + \langle z - x_{T_1 + j + 1} \rangle^T \nabla f(x_{T_1 + j + 1}) \right] + \frac{1}{2} \left( \eta - \zeta \right) \|z - \bar{x}_1\|^2$;

Compute $z_l = \arg\min_{z \in \mathbb{R}^d} \psi(z)$;

end while

$\zeta = \zeta$;

Let $x_l = x_{T_1 + j + 1}$, $y_l = \frac{1}{\gamma_2} x_l + \frac{1}{\gamma_2} z_l$;

else

$x_{T_1 + j + 1} = x_{T_1 + j}$, $\sigma_{T_1 + j + 1} \in [\gamma_1 \sigma_{T_1 + j}, \gamma_2 \sigma_{T_1 + j}]$;

end if

end for

Record the total number of iterations in AAS: $T_2 = j + 1$.

End Phase II: Accelerated Adaptive Subroutine

Different from the accelerated adaptive cubic regularization for Newton’s method with exact/inexact Hessian, the subproblem in each iteration of Algorithm 3

$$s_i = \arg\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i) = \frac{1}{\sigma_i} \nabla f(x_i).$$
where \( m(x_i, s, \sigma_i) \) is defined in (20). Similarly, according to Table 3 the subproblem
\[
\mathbf{z}_l = \arg\min_{\mathbf{z} \in \mathbb{R}^d} \psi_l(\mathbf{z}) = \ell_l(\mathbf{z}) + \frac{1}{4} \xi \| \mathbf{z} - \bar{x}_1 \|^2, \quad l = 1, 2, \ldots,
\]
for the acceleration admits a closed-form solution as well, where \( \ell_l(\mathbf{z}) \) is a certain linear function of \( \mathbf{z} \). In particular, by letting
\[
\nabla \psi_l(\mathbf{z}) = \nabla \ell_l(\mathbf{z}) + \frac{1}{2} \xi (\mathbf{z} - \bar{x}_1) = 0,
\]
and using the fact that \( \nabla \ell_l(\mathbf{z}) \) is independent of \( \mathbf{z} \), we have
\[
\mathbf{z}_l = \bar{x}_1 - \frac{2}{\xi_l} \nabla \ell_l(\mathbf{z}).
\]

5.1 The Convex Case

In this subsection, we aim to analyze the theoretical performance of Algorithm 3. The proof sketch is similar to that of Algorithm 1. Thus, we shall move the details to the appendix, and only present two main results here. Recall that \( l = 1, 2, \ldots \) is the count of successful iterations, and the sequence \( \{\bar{x}_l, l = 1, 2, \ldots\} \) is updated when a successful iteration is identified. The iteration complexity result is presented in Theorem 5.1 and Theorem 5.2.

**Theorem 5.1** The sequence \( \{\bar{x}_l, l = 1, 2, \ldots\} \) generated by Algorithm 3 satisfies
\[
\frac{l(l + 1)}{2} f(\bar{x}_l) \leq \psi_l(\mathbf{z}_l) \leq \psi_l(\mathbf{z}) \leq \frac{l(l + 1)}{2} f(\mathbf{z}) + \frac{L_g + \bar{\sigma}_1}{2} \| \mathbf{z} - \mathbf{x}_0 \|^2 + \frac{1}{4} \xi \| \mathbf{z} - \bar{x}_1 \|^2,
\]
where
\[
\bar{\sigma}_1 = \max \{\sigma_0, \gamma_2 L_g\} > 0.
\]

**Proof.** As before, the proof is based on mathematical induction. The base case of \( l = 1 \) is precisely the result of Theorem A.12. Suppose that the theorem is true for some \( l \geq 1 \). Let us consider the case of \( l + 1 \):
\[
\psi_{l+1}(\mathbf{z}_{l+1}) \leq \psi_{l+1}(\mathbf{z}) \leq \frac{l(l + 1)}{2} f(\mathbf{z}) + \frac{L_g + \bar{\sigma}_1}{2} \| \mathbf{z} - \mathbf{x}_0 \|^2 + \frac{1}{4} \xi \| \mathbf{z} - \bar{x}_1 \|^2 + (l + 1) \left[ f(\bar{x}_l) + (\mathbf{z} - \bar{x}_l)^\top \nabla f(\bar{x}_l) \right] + \frac{1}{4} (\xi_{l+1} - \xi_l) \| \mathbf{z} - \bar{x}_1 \|^2,
\]
where the last inequality is due to the convexity of \( f(\mathbf{z}) \). On the other hand, it follows from the way that \( \psi_{l+1}(\mathbf{z}) \) is updated that \( \frac{l(l+1)}{2} f(\bar{x}_{l+1}) \leq \psi_{l+1}(\mathbf{z}_{l+1}) \), and thus Theorem 5.1 is proven. \( \square \)

Now Theorem 5.1 leads to the following main result on iteration complexity of Algorithm 3.
Theorem 5.2 The sequence \( \{\bar{x}_l, l = 1, 2, \ldots\} \) generated by Algorithm 3 satisfies that
\[
\frac{f(\bar{x}_l) - f(x^*)}{l(l+1)} \leq \frac{C_3}{l^2},
\]
where
\[
C_3 = (L_g + \bar{\sigma}_1)\|x_0 - x^*\|^2 + 2(L_g + \bar{\sigma}_2)^2\|x_1 - x^*\|^2.
\]
The total iteration number required to reach \( \bar{x}_k \) satisfying \( f(\bar{x}_k) - f(x^*) \leq \epsilon \) is bounded as follows:
\[
k \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) + \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) \left[ \left( \frac{C_3}{\epsilon} \right)^\frac{1}{2} + 1 \right] + \left[ \frac{1}{\log(\gamma_3)} \log \left( \frac{(L_g + \bar{\sigma}_2)^2 \frac{4}{\eta \varsigma_1}}{\mu} \right) \right],
\]
where
\[
\bar{\sigma}_2 = \max \{ \bar{\sigma}_1, \gamma_2 L_g + \gamma_2 \eta \} > 0.
\]

Proof. By Theorem 5.1 and taking \( z = x^* \) we have
\[
\frac{l(l+1)}{2} f(\bar{x}_l) \leq \frac{l(l+1)}{2} f(x^*) + \frac{L_g + \bar{\sigma}_1}{2} \|x^* - x_0\|^2 + \frac{1}{4 \varsigma_1} \|x^* - x_1\|^2.
\]
Rearranging the terms, and combining with Lemmas A.6, A.7 and A.11 yields the conclusions. \( \square \)

5.2 Strongly Convex Case

Next we extend the analysis to the case where the objective function is strongly convex. We denote \( A_3^m(x), m \geq 1 \), as the point generated by running \( m \) outer loop iterations of Algorithm 3. In particular, we can generate sequence \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) through the following procedure:

1. Define
\[
m = 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) + \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) \left[ 2 \left( \frac{L_1 + \bar{\sigma}_1 + 2(L_g + \bar{\sigma}_2)^2}{\mu} \right)^\frac{1}{2} + 1 \right]
\]
\[
+ \left[ \frac{1}{\log(\gamma_3)} \log \left( \frac{(L_g + \bar{\sigma}_2)^2 \frac{4}{\eta \varsigma_1}}{\mu} \right) \right].
\]

2. Set \( \hat{x}_0 \in \mathbb{R}^d \).

3. For \( k \geq 0 \), iterate \( \hat{x}_k = A_3^m(\hat{x}_{k-1}) \).

The linear convergence of the sequence \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) is presented in the following theorem.

Theorem 5.3 Suppose the sequence \( \{\hat{x}_k, k = 0, 1, 2, \ldots\} \) is generated by the procedure above. For \( k \geq O(\log(\frac{1}{\epsilon})) \) we have \( f(\hat{x}_k) - f(x^*) \leq \epsilon \). Specifically, the total number of iterations required to find such solution is \( O \left( \sqrt{\frac{L_g}{\mu}} \log(\frac{1}{\epsilon}) \right) \).
Proof. Because
\[
f(x_{k+1}) - f(x^*) \leq \mu \|x_k - x^*\|^2 \leq \frac{1}{2} (f(x_k) - f(x^*)) ,
\]
the total number of iterations to find an \(\epsilon\)-solution is \(O\left(\sqrt{\frac{L}{\mu} \log\left(\frac{1}{\epsilon}\right)}\right)\). \(\square\)

6 Numerical Experiments

In this section, we implement a variant of Algorithm 1, referred to as \textit{Adaptively Accelerated & Cubic Regularized (AARC) Newton’s method}. In this variant we first run Algorithm 1. After 10 successful iterations of \textit{Accelerated Adaptive Subroutine} are performed, we check the progress made by each iteration. In particular, when
\[
|f(x_{k+1}) - f(x_k)| \leq 0.1 ,
\]
which indicates that it is getting close to the global optimum, we switch to the adaptive cubic regularization phase of Newton’s method (ARC) in [6, 7] with stopping criterion \(\|\nabla f(x)\| \leq 10^{-9}\). In the implementation, we apply the so-called Lanczos process to approximately solve the subproblem \(\min_{s \in \mathbb{R}^d} m(x_i, s, \sigma_i)\). In addition to (7), the approximate solution \(s\) is also made to satisfy
\[
s^\top \nabla f(x_i) + s^\top \nabla^2 f(x_i) s + \sigma \|s\|^3 = 0 \tag{21}
\]
for given \(x\) and \(\sigma\). Note that (21) is a consequence of the first order necessary condition, and as shown in Lemma 3.2 [6], the global minimizer of \(m(x_i, s, \sigma_i)\) when restricted to a Krylov subspace
\[
\mathcal{K} := \text{span}\{\nabla f(x_i), \nabla^2 f(x_i) \nabla f(x_i), (\nabla^2 f(x_i))^2 \nabla f(x_i), \ldots\}
\]
satisfies (21) independent of the subspace dimension. Moreover, minimizing \(m(x_i, s, \sigma_i)\) in the Krylov subspace only involve factorizing a tri-diagonal matrix, which can be done at the cost of \(O(d)\). Thus, the associated approximate solution can be found through the so-called Lanczos process, where the dimension of \(\mathcal{K}\) is gradually increased and an orthogonal basis of each subspace \(\mathcal{K}\) is built up which typically involves one matrix-vector product. Condition (7) can be used as the termination criterion for the Lanczos process in the hope to find a suitable trial step before the dimension of \(\mathcal{K}\) approaches \(d\).

We test the performance of the algorithms by evaluating the following regularized logistic regression problem
\[
\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n \ln \left( 1 + \exp \left( -b_i \cdot a_i^\top x \right) \right) + \frac{\lambda}{2} \|x\|^2 \tag{22}
\]
where \((a_i, b_i)_{i=1}^n\) is the samples in the data set, and the regularization parameter is set as \(\lambda = 10^{-5}\). To observe the acceleration, the starting point is randomly generated from a Gaussian random variable with zero mean and a large variance (say 5000). In this way, initial solutions are likely to be far away from the global solution.

We compare the new AARC method with 5 other methods, including the adaptive cubic regularization of Newton’s method (ARC), the trust region method (TR), the limited memory Broyden-Fletcher-Goldfarb-Shanno method (L-BFGS) that is implemented in SCIPY Solvers\(^1\). Algorithm 3 referred to as \textit{adaptive accelerated gradient descent (AAGD)} and the standard Nesterov’s accelerated gradient descent

\(^1\)https://docs.scipy.org/doc/scipy/reference/optimize.html#module-scipy.optimize
The experiments are conducted on 6 LIBSVM Sets\(^2\) for binary classification, and the summary of those datasets are shown in Table 4.

Table 4: Statistics of datasets.

| Dataset     | Number of Samples | Dimension |
|-------------|-------------------|-----------|
| sonar       | 208               | 60        |
| splice      | 1,000             | 60        |
| svmguide1   | 3,089             | 4         |
| svmguide3   | 1,243             | 22        |
| w8a         | 49,749            | 300       |
| SUSY        | 5,000,000         | 18        |

The results in Figure 1 and Figure 2 confirm that AARC indeed accelerates ARC, especially when the current iterates has not entered the local region of quadratic convergence yet. Moreover, AARC outperforms other methods in both computational time and iterations numbers in most cases.

Acknowledgement

We would like to express our deep gratitude toward Professor Xi Chen of Stern School of Business at New York University for the fruitful discussions at various stages of this project.

References

[1] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima for nonconvex optimization in linear time. ArXiv Preprint: 1611.01146, 2016.

[2] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. ArXiv Preprint: 1407.1537, 2014.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[4] S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to nesterov’s accelerated gradient descent. ArXiv Preprint: 1506.08187, 2015.

[5] Y. Carmon and J. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex newton step. ArXiv Preprint: 1612.00547v2, 2016.

[6] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: Motivation, convergence and numerical results. Mathematical Programming, 127(2):245–295, 2011.

\(^2\)https://www.csie.ntu.edu.tw/~cjlin/libsvm/
Figure 1: Performance of AARC and all benchmark methods on the task of regularized logistic regression (loss vs. time)
Figure 2: Performance of AARC and all benchmark methods on the task of regularized logistic regression (loss vs. iterations)
[7] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: Worst-case function-and derivative-evaluation complexity. Mathematical Programming, 130(2):295–319, 2011.

[8] C. Cartis, N. I. M. Gould, and P. L. Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. Optimization Methods and Software, 27(2):197–219, 2012.

[9] C. Cartis, N. I. M. Gould, and P. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. SIAM Journal on Optimization, 22(1):66–86, 2012.

[10] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In NIPS, pages 1647–1655, 2011.

[11] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. Mathematical Programming, 145(1-2):451–482, 2014.

[12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(7):2121–2159, 2011.

[13] A. Karpathy. A peak at trends in machine learning. https://medium.com/@karpathy/a-peek-at-trends-in-machine-learning-ab8a1085a106, 2017.

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. ArXiv Preprint: 1412.6980, 2014.

[15] J. M. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. ArXiv Preprint: 1705.05933, 2017.

[16] G. Lan. An optimal method for stochastic composite optimization. Mathematical Programming, 133(1):365–397, 2012.

[17] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In NIPS, pages 3384–3392, 2015.

[18] Q. Lin and L. Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. Computational Optimization and Applications, 60(3):633–674, 2014.

[19] D. G. Luenberger and Y. Ye. Linear and nonlinear programming, volume 2. Springer, 1984.

[20] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of a newton proximal extragradient method for monotone variational inequalities and inclusion problems. SIAM Journal on Optimization, 22(3):914–935, 2012.

[21] Renato D. C. Monteiro, C. Ortiz, and B. F. Svaiter. An adaptive accelerated first-order method for convex optimization. Computational Optimization and Applications, 64(1):31–73, 2016.
A Proofs in Section 4 and Section 5

A.1 Proofs in Section 4

Lemma A.1 Letting \( \bar{\sigma}_1 = \max \left\{ \sigma_0, \frac{\gamma_2 L_h + 3\gamma_3 (\kappa_e + \kappa_c) \kappa_{\text{hk}}}{2} \sigma^2_{\min} \right\} > 0 \), we have \( T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) \).
Proof. We have
\[
 f(x_i + s_i) = f(x_i) + s_i^T \nabla f(x_i) + \frac{1}{2} s_i^T \nabla^2 f(x_i) s_i + \int_0^1 (1 - \tau) s_i^T \left[ \nabla^2 f(x_i + \tau s_i) - \nabla^2 f(x_i) \right] s_i \, d\tau \\
\leq f(x_i) + s_i^T \nabla f(x_i) + \frac{1}{2} s_i^T \nabla^2 f(x_i) s_i + \frac{L_h}{6} \|s_i\|^3 \\
= m(x_i, s_i, \sigma_i) + \frac{1}{2} s_i^T \left( \nabla^2 f(x_i) - H(x_i) \right) s_i + \left( \frac{L_h}{6} - \frac{\sigma_i}{3} \right) \|s_i\|^3, \\
\leq m(x_i, s_i, \sigma_i) + \left( \frac{L_h}{6} + \frac{(\kappa_e + \kappa_c)\kappa_{hs}}{2} - \frac{\sigma_i}{3} \right) \|s_i\|^3,
\]
where the inequalities hold true due to Assumption 2.1 and (17). Therefore, we conclude that
\[
 \sigma_i \geq \frac{L_h + 3(\kappa_e + \kappa_c)\kappa_{hs}}{2} \implies f(x_i + s_i) \leq m(x_i, s_i, \sigma_i),
\]
which further implies that \( \sigma_i < \frac{L_h + 3(\kappa_e + \kappa_c)\kappa_{hs}}{2} \) for \( i \leq T - 2 \). Hence,
\[
 \sigma_{T_1} \leq \sigma_{T_1 - 1} \leq \sigma_{T_1 - 2} \leq \frac{\gamma_2 L_h + 3\gamma_2 (\kappa_e + \kappa_c)\kappa_{hs}}{2}.
\]
Because \( \bar{\sigma}_1 = \max \left\{ \sigma_0, \frac{\gamma_2 L_h + 3\gamma_2 (\kappa_e + \kappa_c)\kappa_{hs}}{2} \right\} \), it follows from the construction of Algorithm 2 that \( \sigma_{\min} \leq \sigma_i \) for all iterations, and \( \gamma_1 \sigma_i \leq \sigma_{i+1} \) for all unsuccessful iterations. Consequently, we have
\[
 \frac{\bar{\sigma}_1}{\sigma_{\min}} \geq \frac{\sigma_{T_1}}{\sigma_0} = \frac{\sigma_{T_1}}{\sigma_{T_1 - 1}} \cdot \prod_{j=0}^{T_1 - 2} \frac{\sigma_{j+1}}{\sigma_j} \geq \frac{\gamma_1 - 1}{\bar{\sigma}_1} \left( \frac{\sigma_{\min}}{\sigma_0} \right),
\]
and hence \( T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_1}{\sigma_{\min}} \right) \).

**Lemma A.2** Letting \( \bar{\sigma}_2 = \max \left\{ \bar{\sigma}_1, \frac{\gamma_2 L_h}{2} + \gamma_2 \kappa_\theta + \gamma_2 (\kappa_e + \kappa_c)\kappa_{hs} + \gamma_2 \eta \right\} > 0 \), we have
\[
 T_2 \leq \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) |S|.
\]

**Proof.** We have
\[
 s_{T_1+j}^T \nabla f(y_{i} + s_{T_1+j}) \\
= s_{T_1+j}^T \left[ \nabla f(y_{i} + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j} \right] + s_{T_1+j}^T \left[ \nabla f(y_i) + \nabla^2 f(y_i) s_{T_1+j} \right] \\
\leq \left\| \nabla f(y_{i} + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j} \right\| \|s_{T_1+j}\| + s_{T_1+j}^T \nabla m(y_i, s_{T_1+j}, \sigma_{T_1+j}) \\
+ s_{T_1+j}^T \left( \nabla^2 f(y_i) - H(y_i) \right) s_{T_1+j} - \sigma_{T_1+j} \|s_{T_1+j}\|^2 \\
\leq \left\| \nabla f(y_{i} + s_{T_1+j}) - \nabla f(y_i) - \nabla^2 f(y_i) s_{T_1+j} \right\| \|s_{T_1+j}\| - \sigma_{T_1+j} \|s_{T_1+j}\|^2 + (\kappa_\theta + (\kappa_e + \kappa_c)\kappa_{hs}) \|s_{T_1+j}\|^3 \\
= \left\| \int_0^1 \left[ \nabla^2 f(y_i + \tau \cdot s_{T_1+j}) - \nabla^2 f(y_i) \right] s_{T_1+j} \, d\tau \right\| \|s_{T_1+j}\| - \sigma_{T_1+j} \|s_{T_1+j}\|^2 + (\kappa_\theta + (\kappa_e + \kappa_c)\kappa_{hs}) \|s_{T_1+j}\|^3 \\
\leq \left( \frac{L_h}{2} + \kappa_\theta + (\kappa_e + \kappa_c)\kappa_{hs} - \sigma_{T_1+j} \right) \|s_{T_1+j}\|^3,
\]
where the last inequality is due to Assumption 2.1. Then it follows that
\[-\frac{s_{T_{i}+j}^T \nabla f(y_i + s_{T_{i}+j})}{\|s_{T_{i}+j}\|^3} \geq \sigma_{T_{i}+j} - \frac{L_h}{2} - \kappa_\theta - (\kappa_e + \kappa_c)\kappa_{h_s},\]

Therefore, we have
\[\sigma_{T_{i}+j} \geq \frac{L_h}{2} + \kappa_\theta + (\kappa_e + \kappa_c)\kappa_{h_s} + \eta \implies -\frac{s_{T_{i}+j}^T \nabla f(y_i + s_{T_{i}+j})}{\|s_{T_{i}+j}\|^3} \geq \eta,
\]
which further implies that
\[\sigma_{T_{i}+j+1} \leq \sigma_{T_{i}+j} \leq \gamma_2 \cdot \sigma_{T_{i}+j-1} \leq \gamma_2 \left(\frac{L_h}{2} + \kappa_\theta + (\kappa_e + \kappa_c)\kappa_{h_s} + \eta\right), \quad \forall j \in S.
\]
Therefore, the above quantity can be bounded by \(\tilde{\sigma}_2 = \max\left\{\tilde{\sigma}_1, \frac{\gamma_2 L_h}{2} + \gamma_2 \kappa_\theta + \gamma_2 (\kappa_e + \kappa_c)\kappa_{h_s} + \gamma_2 \eta\right\}\), where \(\tilde{\sigma}_1\) is responsible for an upper bound of \(\sigma_{T_1}\). In addition, it follows from the construction of Algorithm 2 that \(\sigma_{\min} \leq \sigma_{T_1}\) for all iterations, and \(\gamma_1 \sigma_{T_1} \leq \sigma_{T_1+1}\) for all unsuccessful iterations. Therefore, we have
\[\frac{\tilde{\sigma}_2}{\sigma_{\min}} \geq \frac{\sigma_{T_{i}+T_2}}{\sigma_{T_1}} = \prod_{j \in S} \frac{\sigma_{T_{i}+j+1}}{\sigma_{T_{i}+j}} \cdot \prod_{j \notin S} \frac{\sigma_{T_{i}+j+1}}{\sigma_{T_{i}+j}} \geq T_2^{-|S|} \left(\frac{\sigma_{\min}}{\tilde{\sigma}_2}\right)^{|S|},
\]
hence
\[|S| \leq T_2 \leq |S| + \left(\frac{|S| + 1}{\log \gamma_1}\right) \log \left(\frac{\tilde{\sigma}_2}{\sigma_{\min}}\right) \leq \left(1 + \frac{2}{\log \gamma_1}\right) \log \left(\frac{\tilde{\sigma}_2}{\sigma_{\min}}\right) |S|.
\]
\[\square\]

Before estimating an upper bound for \(T_3\), i.e., the total number of times of successfully updating \(\zeta > 0\), we need to extend Lemma 3.3 in Algorithm 1 to the following lemma.

**Lemma A.3** For each iteration \(j\) in the subroutine AAS, if it is successful, we have

\[(1 - \kappa_\theta) \|\nabla f(x_{j+1})\| \leq \left(\frac{L_h}{2} + \tilde{\sigma}_2 + (\kappa_e + \kappa_c)\kappa_{h_s} + \kappa_\theta L_g\right) \|s_j\|^2,
\]

where \(\kappa_\theta \in (0, 1)\) is used in Condition 5.1.

**Proof.** We denote \(j\)-th iteration is the \(l\)-th successful iteration, and note \(\nabla s m(y_l, s_j, \sigma_j) = \nabla f(y_l) + H(y_l)s_j + \sigma_j \|s_j\| \cdot s_j\). Then we have
\[
\|\nabla f(x_{j+1})\| \\
\leq \|\nabla f(y_l + s_j) - \nabla s m(y_l, s_j, \sigma_j)\| + \|\nabla s m(y_l, s_j, \sigma_j)\| \\
\leq \|\nabla f(y_l + s_j) - \nabla s m(y_l, s_j, \sigma_j)\| + \kappa_\theta \cdot \min\left(1, \|s_j\|\right) \cdot \|\nabla f(y_l)\| \\
\leq \left[\int_0^1 (\nabla^2 f(y_l + \tau s_j) - \nabla^2 f(y_l)) s_j d\tau\right] + \|\nabla^2 f(y_l) - H(y_l)\| \|s_j\| + \sigma_j \|s_j\|^2 + \kappa_\theta \cdot \min\left(1, \|s_j\|\right) \cdot \|\nabla f(y_l)\| \\
\leq \frac{L_h}{2} \|s_j\|^2 + (\kappa_e + \kappa_c)\kappa_{h_s} \|s_j\|^2 + \sigma_j \|s_j\|^2 + \kappa_\theta \cdot \|s_j\| \cdot \|\nabla f(y_l) - \nabla f(y_l + s_j)\| + \kappa_\theta \|\nabla f(x_{j+1})\| \\
\leq \frac{L_h}{2} \|s_j\|^2 + (\kappa_e + \kappa_c)\kappa_{h_s} \|s_j\|^2 + \tilde{\sigma}_2 \|s_j\|^2 + \kappa_\theta L_g \|s_j\|^2 + \kappa_\theta \|\nabla f(x_{j+1})\|.
\]

35
where the second inequality holds true due to Condition 3.1, and the last two inequality follow from Assumption 2.1. Rearranging the terms, the conclusion follows.

Now we are ready to estimate the upper bound of $T_3$, i.e., the total number of count of successfully updating $\varsigma > 0$.

**Lemma A.4** We must have

$$\psi_1(z_l) \geq \frac{1}{6} f(\bar{x}_l)$$

if $\varsigma_l \geq \left(\frac{L_h + 2\bar{s}_2 + 2\kappa_e L_g}{1 - \kappa_\theta}\right)^3 \frac{1}{\eta^2}$, which further implies that

$$T_3 \leq \left[\frac{1}{\log (\gamma_3)} \log \left(\frac{L_h + 2\bar{s}_2 + 2(\kappa_e + \kappa_c)\kappa_h s + 2\kappa_\theta L_g}{1 - \kappa_\theta}\right)^3 \frac{1}{\eta^2 \varsigma_1}\right].$$

**Proof.** The proof is similar to that of Lemma 3.6 except replacing Lemma 3.5 with Lemma A.3.

Now we are able to prove the base case of $l = 1$ for Theorem 4.2.

**Theorem A.5** It holds that

$$f(\bar{x}_1) \leq \psi_1(z_1) \leq \psi_1(z) \leq f(z) + \frac{L_h + \bar{s}_1 + (\kappa_e + \kappa_c)\kappa_h s + 2\kappa_\theta L_g}{2} \|z - x_0\|^3 + \frac{2\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}} \|x_0 - x^*\|^2 + \frac{1}{6} \varsigma_1 \|z - \bar{x}_1\|^3.$$

**Proof.** By the definition of $\psi_1(z)$ and the fact that $\bar{x}_1 = x_{T_1}$, we have

$$f(\bar{x}_1) = f(x_{T_1}) = \psi_1(z_1).$$

Furthermore, by the criterion of successful iteration in SAS,

$$f(\bar{x}_1) = f(x_{T_1}) \leq m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1})$$

$$= m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1}) - m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) + m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}),$$

where $s_{T_1-1}^m$ denotes the global minimizer of $m(x_{T_1-1}, s, \sigma_{T_1-1})$ over $\mathbb{R}^d$. Since $H(x_{T_1-1}) \succeq 0$ due to (118), $m(x_{T_1-1}, s, \sigma_{T_1-1})$ is convex as well. Indeed, we have

$$\nabla_s^2 m(x_{T_1-1}, s, \sigma_{T_1-1}) \succeq H(x_{T_1-1}) + \sigma_{T_1-1} \|s\| \cdot I + \sigma_{T_1-1} \frac{ss^\top}{\|s\|^2} \succeq 0.$$

Therefore, we have

$$m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1}) - m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) \leq \nabla_s m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1})^\top (s_{T_1-1} - s_{T_1-1}^m)$$

$$\leq \|\nabla_s m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1})\| \|s_{T_1-1} - s_{T_1-1}^m\|$$

$$\leq \kappa_\theta \|\nabla f(x_{T_1-1})\| \|s_{T_1-1} - s_{T_1-1}^m\|. $$

36
To bound $\|s_{T_1-1} - s_{T_1-1}^m\|$, since $H(x_{T_1-1}) \succeq 0$ we have that

$$
\sigma_{\min} \|s\|^3 \leq \sigma_{T_1-1} \|s\|^3 = s^T [\nabla m(x_{T_1-1}, s, \sigma_{T_1-1}) - \nabla f(x_{T_1-1}) - H(x_{T_1-1}) s] \\
\leq s^T [\nabla m(x_{T_1-1}, s, \sigma_{T_1-1}) - \nabla f(x_{T_1-1})] \\
\leq \|s\| \|\nabla f(x_{T_1-1})\| + \|\nabla m(x_{T_1-1}, s, \sigma_{T_1-1})\| \\
\leq (1 + \kappa_\theta) \|s\| \|\nabla f(x_{T_1-1})\|
$$

where $s = s_{T_1-1}$ or $s = s_{T_1-1}^m$. Thus, we conclude that

$$
\|s_{T_1-1} - s_{T_1-1}^m\| \leq \|s_{T_1-1}\| + \|s_{T_1-1}^m\| \leq 2 \sqrt{\frac{(1 + \kappa_\theta) \|\nabla f(x_{T_1-1})\|}{\sigma_{\min}}},
$$

which combines with Assumption 2.1 yields that

$$
m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1}) - m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) \leq \frac{2\kappa_\theta (1 + \kappa_\theta)}{\sigma_{\min}} \|\nabla f(x_{T_1-1})\|^2 \\
= \frac{2\kappa_\theta (1 + \kappa_\theta)}{\sigma_{\min}} \|\nabla f(x_{T_1-1}) - \nabla f(x^*)\|^2 \\
\leq \frac{2\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}} \|x_{T_1-1} - x_*\|^2 \\
= \frac{2\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}} \|x_0 - x_*\|^2.
$$

On the other hand, we have

$$
m(x_{T_1-1}, s_{T_1-1}^m, \sigma_{T_1-1}) \\
= f(x_{T_1-1}) + (s_{T_1-1}^m)^T \nabla f(x_{T_1-1}) + \frac{1}{2} (s_{T_1-1}^m)^T H(x_{T_1-1}) s_{T_1-1}^m + \frac{1}{3} \sigma_{T_1-1} \|s_{T_1-1}^m\|^3 \\
\leq f(x_{T_1-1}) + (z - x_{T_1-1})^T \nabla f(x_{T_1-1}) + \frac{1}{2} (z - x_{T_1-1})^T H(x_{T_1-1})(z - x_{T_1-1}) + \frac{1}{3} \sigma_{T_1-1} \|z - x_{T_1-1}\|^3 \\
\leq f(z) + \frac{L_h}{6} \|z - x_{T_1-1}\|^3 + \frac{1}{3} \sigma_{T_1-1} \|z - x_{T_1-1}\|^3 + \frac{1}{2} (\kappa_c + \kappa_c) \kappa_{hs} \|z - x_{T_1-1}\|^3 \\
\leq f(z) + \frac{L_h + \bar{\sigma}_1 + (\kappa_c + \kappa_c) \kappa_{hs}}{2} \|z - x_{T_1-1}\|^3 \\
= f(z) + \frac{L_h + \bar{\sigma}_1 + (\kappa_c + \kappa_c) \kappa_{hs}}{2} \|z - x_0\|^3,
$$

where the second inequality is due to (23) and Assumption 2.1. Therefore, we conclude that

$$
\psi_1(z) = f(\bar{x}_1) + \frac{1}{6} \varsigma_1 \|z - \bar{x}_1\|^3 \\
\leq f(z) + \frac{L_h + \bar{\sigma}_1 + (\kappa_c + \kappa_c) \kappa_{hs}}{2} \|z - x_0\|^3 + \frac{2\kappa_\theta (1 + \kappa_\theta) L_g^2}{\sigma_{\min}} \|x_0 - x^*\|^2 + \frac{1}{6} \varsigma_1 \|z - \bar{x}_1\|^3.
$$

\qed
A.2 Proofs in Section 5

Lemma A.6 Letting $\tilde{\sigma}_1 = \max \{ \sigma_0, \gamma_2 L_g \} > 0$, we have $T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\tilde{\sigma}_1}{\sigma_{\min}} \right)$.

Proof. We have

$$
\begin{align*}
\sigma_i &\geq L_g \implies f(x_i + s_i) \leq m(x_i, s_i, \sigma_i), \\
\text{which further implies that } \sigma_i &< L_g \text{ for } i \leq T_1 - 2. \text{ Hence,}
\end{align*}
$$

because $\bar{\sigma}_1 = \max \{ \sigma_0, \gamma_2 L_g \}$, it follows from the construction of Algorithm 1 that $\sigma_{\min} \leq \sigma_i$ for all iterations, and $\gamma_1 \sigma_i \leq \sigma_{i+1}$ for all unsuccessful iterations. Consequently, we have

$$
\frac{\tilde{\sigma}_1}{\sigma_{\min}} \geq \frac{\sigma_{T_1}}{\sigma_0} = \frac{\sigma_{T_1}}{\sigma_{T_1-1}} \cdot \frac{T_1 - 2}{\prod_{j=0}^{T_1-2} \frac{\sigma_{j+1}}{\sigma_j}} \geq \gamma_1^{T_1-1} \left( \frac{\sigma_{\min}}{\tilde{\sigma}_1} \right),
$$

and hence $T_1 \leq 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\tilde{\sigma}_1}{\sigma_{\min}} \right)$. \qed

Lemma A.7 Letting $\tilde{\sigma}_2 = \max \{ \bar{\sigma}_1, \gamma_2 L_g + \gamma_2 \eta \} > 0$, we have $T_2 \leq \left( 1 + \frac{2}{\log(\gamma_1)} \log \left( \frac{\tilde{\sigma}_2}{\sigma_{\min}} \right) \right) |S|.$

Proof. We have

$$
\begin{align*}
s_{T_1+j}^T \nabla f(y_l + s_{T_1+j}) &= s_{T_1+j}^T \nabla f(y_l + s_{T_1}) - \nabla f(y_l) + s_{T_1+j}^T \nabla f(y_l) \\
&\leq \| \nabla f(y_l + s_{T_1+j}) - \nabla f(y_l) \| \| s_{T_1+j} \| - \sigma_{T_1+j} \| s_{T_1+j} \|^2 \\
&\leq (L_g - \sigma_{T_1+j}) \| s_{T_1+j} \|^2,
\end{align*}
$$

where the last inequality is due to Assumption 2.1. Then it follows that

$$
\frac{s_{T_1+j}^T \nabla f(y_l + s_{T_1+j})}{\| s_{T_1+j} \|^2} \geq \sigma_{T_1+j} - L_g.
$$

Therefore, we have

$$
\sigma_{T_1+j} \geq L_g + \eta \implies -\frac{s_{T_1+j}^T \nabla f(y_l + s_{T_1+j})}{\| s_{T_1+j} \|^2} \geq \eta,
$$

38
which further implies that

\[ \sigma_{T_1+j+1} \leq \sigma_{T_1+j} \leq \gamma_2 \cdot \sigma_{T_1+j-1} \leq \gamma_2 (L_g + \eta), \quad \forall \ j \in S. \]

Therefore, the above quantity is bounded by \( \bar{\sigma}_2 = \max \{ \bar{\sigma}_1, \gamma_2 L_g + \gamma_2 \eta \} \), where \( \bar{\sigma}_1 \) represents an upper bound on \( \sigma_{T_1} \). In addition, it follows from the construction of Algorithm 2 that \( \sigma_{\min} \leq \sigma_{T_1+j} \) for all iterations, and \( \gamma_1 \sigma_{T_1+j} \leq \sigma_{T_1+j+1} \) for all unsuccessful iterations. Therefore, we have

\[
\frac{\bar{\sigma}_2}{\sigma_{\min}} \geq \prod_{j \in S} \frac{\sigma_{T_1+j+1}}{\sigma_{T_1+j}} \cdot \prod_{j \notin S} \frac{\sigma_{T_1+j+1}}{\sigma_{T_1+j}} \geq \frac{\gamma_2 - |S|}{\gamma_1} \left( \frac{\sigma_{\min}}{\bar{\sigma}_2} \right)^{|S|},
\]

and hence

\[
|S| \leq T_2 \leq |S| + \left( \frac{|S| + 1}{\log \gamma_1} \right) \left( 1 + \frac{2}{\log \gamma_1} \log \left( \frac{\bar{\sigma}_2}{\sigma_{\min}} \right) \right) |S|.
\]

Before estimating the upper bound of \( T_3 \), i.e., the total number of the count of successfully updating \( \varsigma > 0 \), we need to prove a few technical lemmas.

**Lemma A.8** Let \( z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z) \), then we have \( \psi_l(z) - \psi_l(z_l) \geq \frac{1}{8\varsigma_l} \| z - z_l \|^2 \).

**Proof.** It suffices to show that

\[ \psi_l(z) - \psi_l(z_l) - \nabla \psi_l(z_l) \top (z - z_l) \geq \frac{1}{8\varsigma_l} \| z - z_l \|^2. \]

By using the fact that \( z_l = \arg\min_{z \in \mathbb{R}^d} \psi_l(z) \) and \( \nabla \psi_l(z_l) = 0 \), and the strongly convexity of \( \psi_l \), we obtain the desired result. \( \square \)

**Lemma A.9** For any \( s \in \mathbb{R}^d \) and \( g \in \mathbb{R}^d \), we have

\[ s \top g + \frac{1}{2} \sigma \| s \|^2 \geq - \frac{1}{2\sigma} \| g \|^2. \]

**Proof.** Denote \( s^* \) to be the minimum of \( s \top g + \frac{1}{2} \sigma \| s \|^2 \). Hence, \( g + \sigma s^* = 0 \). Therefore, \( (s^*) \top g = - \sigma \| s^* \|^2 \) and \( \| g \| = \sigma \| s^* \| \), and so

\[ (s^*) \top g + \frac{1}{2} \sigma \| s^* \|^2 = - \frac{1}{2} \sigma \| s^* \|^2 = - \frac{1}{2\sigma} \| g \|^2. \]

\( \square \)

**Lemma A.10** For each iteration \( j \) in the subroutine AAS, if it is a successful iteration, then we have

\[ \| \nabla f(x_{j+1}) \| \leq (L_g + \bar{\sigma}_2) \| s_j \|. \]

39
Lemma A.11 We have

\[ \psi_I(z_I) \geq \frac{l(l+1)}{2} f(\bar{x}_I) \quad (25) \]

if \( \var_1 \geq \frac{(2L_g + 2\var_2)^2}{\eta_1} \), which further implies that

\[ T_3 \leq \left[ \frac{1}{\log(\gamma_3)} \log \left( \frac{(2L_g + 2\var_2)^2}{\eta \var_1} \right) \right]. \]

Proof. When \( l = 1 \), it trivially holds true that \( \psi_I(z_I) \geq \frac{(l+1)}{2} f(\bar{x}_I) \) since \( \psi_1(z_I) = f(\bar{x}_1) \). As a result, it suffices to show that \( \var_1 \geq \frac{(2L_g + 2\var_2)^2}{\eta_1} \) by mathematical induction. Without loss of generality, we assume \( (25) \) holds true for some \( l - 1 \geq 1 \). Then, it follows from Lemma A.8 the construction of \( \psi_I(z) \) and our induction that

\[ \psi_{l-1}(z) \geq \psi_{l-1}(z_{l-1}) + \frac{1}{8} \var_{l-1} \| z - z_I \|^2 \geq \frac{(l-1)}{2} f(\bar{x}_{l-1}) + \frac{1}{8} \var_{l-1} \| z - z_{l-1} \|^2. \]

As a result, we have

\[ \psi_I(z_I) = \min_{z \in \mathbb{R}^d} \left\{ \psi_I(z) + l \left[ f(\bar{x}_I) + (z - \bar{x}_I)^T \nabla f(\bar{x}_I) \right] + \frac{1}{4} (\var_1 - \var_{l-1}) \| z - \bar{x}_1 \|^2 \right\} \]

\[ \geq \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)}{2} f(\bar{x}_{l-1}) + \frac{1}{8} \var_1 \| z - z_{l-1} \|^2 + l \left[ f(\bar{x}_I) + (z - \bar{x}_I)^T \nabla f(\bar{x}_I) \right] \right\} \]

\[ \geq \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)}{2} f(\bar{x}_I) + (\bar{x}_{l-1} - \bar{x}_I)^T \nabla f(\bar{x}_I) \right\} + \frac{1}{8} \var_1 \| z - z_{l-1} \|^2 \]

\[ + l \left[ f(\bar{x}_I) + (z - \bar{x}_I)^T \nabla f(\bar{x}_I) \right] \]

\[ = \frac{l(l+1)}{2} f(\bar{x}_I) + \min_{z \in \mathbb{R}^d} \left\{ \frac{(l-1)}{2} (\bar{x}_{l-1} - \bar{x}_I)^T \nabla f(\bar{x}_I) + \frac{1}{8} \var_1 \| z - z_{l-1} \|^2 \right\} \]

\[ + l (z - \bar{x}_I)^T \nabla f(\bar{x}_I) \].
where the first inequality holds true because \( \varsigma_l \geq \varsigma_{l-1} \). By the construction of \( y_{l-1} \), one has

\[
\frac{(l - 1)l}{2} x_{l-1} = \frac{l(l + 1)}{2} \frac{l - 1}{l + 1} x_{l-1} \\
= \frac{l(l + 1)}{2} \left( y_{l-1} - \frac{2}{l + 1} z_{l-1} \right) \\
= \frac{l(l + 1)}{2} y_{l-1} - l z_{l-1}.
\]

Combining the above two formulas yields

\[
\psi_l(z_l) \geq \frac{l(l + 1)}{2} f(\bar{x}_l) + \min_{u \in \mathbb{R}^d} \left\{ \frac{l(l + 1)}{2} (y_{l-1} - \bar{x}_l) \top \nabla f(\bar{x}_l) + \frac{1}{8} \varsigma_l \| z - z_{l-1} \|^2 + l (z - z_{l-1}) \top \nabla f(\bar{x}_l) \right\}.
\]

Then, by the criterion of successful iteration in AAS and Lemma A.10, we have

\[
(y_{l-1} - \bar{x}_l) \top \nabla f(\bar{x}_l) = -s_{T_1 + j} \top \nabla f(y_{l-1} + s_{T_1 + j}) \geq \eta \| s_{T_1 + j} \|^2 \geq \eta \left( \frac{1}{L_g + \sigma_2} \right)^2 \| \nabla f(\bar{x}_l) \|^2,
\]

where the \( l \)-th successful iteration count refers to the \((j - 1)\)-th iteration count in AAS. Hence, it suffices to establish

\[
\frac{l(l + 1) \eta}{2} \left( \frac{1}{L_g + \sigma_2} \right)^2 \| \nabla f(\bar{x}_l) \|^2 + \frac{1}{8} \varsigma_l \| z - z_{l-1} \|^2 + l (z - z_{l-1}) \top \nabla f(\bar{x}_l) \geq 0.
\]

Using Lemma A.9 and setting \( g = l \nabla f(\bar{x}_l) \) and \( \sigma = \frac{1}{4} \varsigma_l \), the above is implied by

\[
\frac{l(l + 1) \eta}{2} \left( \frac{1}{L_g + \sigma_2} \right)^2 \geq \frac{2}{\varsigma_l} \left( \frac{1}{L_g + \sigma_2} \right)^2 \geq \frac{2}{\varsigma_l} l^2.
\]

Therefore, the conclusion follows if \( \varsigma_l \geq (2L_g + 2\sigma_2)^2 \frac{1}{\eta} \). \( \square \)

Finally we are in a position to prove the base case of \( l = 1 \) for Theorem 5.1.

**Theorem A.12** It holds that

\[
f(\bar{x}_1) \leq \psi_1(z_1) \leq \psi_1(z) \leq f(z) + \frac{L_g + \sigma_1}{2} \| z - x_0 \|^2 + \frac{1}{8} \varsigma_1 \| z - \bar{x}_1 \|^2.
\]

**Proof.** By the definition of \( \psi_1(z) \) and the fact that \( \bar{x}_1 = x_{T_1} \), we have

\[
f(\bar{x}_1) = f(x_{T_1}) = \psi_1(z_1).
\]
Furthermore, by the criterion of successful iteration in SAS,

\[
\begin{align*}
    f(\bar{x}_1) &= f(x_{T_1}) \\
    &\leq m(x_{T_1-1}, s_{T_1-1}, \sigma_{T_1-1}) \\
    &= f(x_{T_1-1}) + s_{T_1-1}^\top \nabla f(x_{T_1-1}) + \frac{\sigma_{T_1-1}}{2} ||s_{T_1-1}||^2 \\
    &\leq f(x_{T_1-1}) + (z - x_{T_1-1})^\top \nabla f(x_{T_1-1}) + \frac{1}{2} \sigma_{T_1-1} ||z - x_{T_1-1}||^2 \\
    &\leq f(z) + \frac{L_g}{2} ||z - x_{T_1-1}||^2 + \frac{1}{2} \sigma_{T_1-1} ||z - x_{T_1-1}||^2 \\
    &\leq f(z) + \frac{L_g + \sigma_{\bar{\sigma}}}{2} ||z - x_0||^2,
\end{align*}
\]

where the third inequality is due to Assumption 2.1. Therefore, we conclude that

\[
\begin{align*}
    \psi_1(z) &= f(\bar{x}_1) + \frac{1}{4} \varsigma_{\bar{x}} ||z - \bar{x}_1||^2 \\
    &\leq f(z) + \frac{L_g + \sigma_{\bar{\sigma}}}{2} ||z - x_0||^3 + \frac{1}{4} \varsigma_{\bar{x}} ||z - \bar{x}_1||^2.
\end{align*}
\]