Mathematical Programming (2020) 184:35–70
https://doi.org/10.1007/s10107-019-01405-z

FULL LENGTH PAPER

Series A

Newton-type methods for non-convex optimization under inexact Hessian information

Peng Xu[1] · **Fred Roosta**[2,3] · **Michael W. Mahoney**[3,4]

Received: 27 November 2017 / Accepted: 16 May 2019 / Published online: 22 May 2019

Abstract
We consider variants of trust-region and adaptive cubic regularization methods for non-convex optimization, in which the Hessian matrix is approximated. Under certain condition on the inexact Hessian, and using approximate solution of the corresponding sub-problems, we provide iteration complexity to achieve ε-approximate second-order optimality which have been shown to be tight. Our Hessian approximation condition offers a range of advantages as compared with the prior works and allows for direct construction of the approximate Hessian with a priori guarantees through various techniques, including randomized sampling methods. In this light, we consider the canonical problem of finite-sum minimization, provide appropriate uniform and non-uniform sub-sampling strategies to construct such Hessian approximations, and obtain optimal iteration complexity for the corresponding sub-sampled trust-region and adaptive cubic regularization methods.

Keywords Non-convex optimization · Inexact Hessian · Trust region · Cubic regularization · Randomized numerical linear algebra

Mathematics Subject Classification 49M15 · 65K05 · 90C25 · 90C06

✉ Fred Roosta
fred.roosta@uq.edu.au

Peng Xu
pengxu@stanford.edu

Michael W. Mahoney
mmahoney@stat.berkeley.edu

[1] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, USA

[2] School of Mathematics and Physics, University of Queensland, Brisbane, Australia

[3] International Computer Science Institute, Berkeley, USA

[4] Department of Statistics, University of California at Berkeley, Berkeley, USA

1 Introduction

Consider the generic unconstrained optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}), \tag{P0}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is *smooth* and *non-convex*. Faced with the large-scale nature of modern "big-data" problems, many of the classical optimization algorithms might prove to be inefficient, if applicable at all. In this light, many of the recent research efforts have been centered around designing variants of classical algorithms which, by employing suitable *approximations* of the gradient and/or Hessian, improve upon the cost-per-iteration, while maintaining the original iteration complexity. In this light, we focus on trust-region (TR) [17] and cubic regularization (CR) [34], two algorithms which are considered as among the most elegant and theoretically sound general-purpose Newton-type methods for non-convex problems.

In doing so, we first consider (**P0**), and study the theoretical convergence properties of variants of these two algorithms in which, under favorable conditions, Hessian is suitably approximated. We show that our Hessian approximation conditions, in many cases, are weaker than the existing ones in the literature. In addition, and in contrast to some prior works, our conditions allow for efficient constructions of the inexact Hessian with a priori guarantees via various approximation methods, of which Randomized Numerical Linear Algebra (RandNLA), [22,42], techniques are shown to be highly effective.

Subsequently, to showcase the application of randomized techniques for construction of the approximate Hessian, we consider an important instance of (**P0**), i.e., large-scale *finite-sum* minimization, of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{P1}$$

and its special case

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{a}_i^T\mathbf{x}), \tag{P2}$$

where $n \gg 1$, each f_i is a smooth but possibly non-convex function, and $\mathbf{a}_i \in \mathbb{R}^d, i = 1, \ldots, n$, are given. Problems of the form (**P1**) and (**P2**) arise very often in machine learning, e.g., [51] as well as scientific computing, e.g., [47,48]. In big-data regime where $n \gg 1$, operations with the Hessian of F, e.g., matrix-vector products, typically constitute the main bottleneck of computations. Here, we show that our relaxed Hessian approximation conditions allow one to draw upon the *sub-sampling* ideas of [6,49,62], to design variants of TR and CR algorithms where the Hessian is *(non-)uniformly* sub-sampled. We then present the theoretical convergence properties of these variants for non-convex finite-sum problems of the form (**P1**) and (**P2**).

The rest of this paper is organized as follows. In Sect. 1.1, we first introduce the notation and definitions used throughout the paper. For completeness, in Sect. 1.2, we give a brief review of trust region (Sect. 1.2.1) and cubic regularization (Sect. 1.2.2) along with related prior works. Our main contributions are summarized in Sect. 1.3. Theoretical analysis of the proposed algorithms for solving generic non-convex problem (**P0**) are presented in Sect. 2. Various randomized sub-sampling strategies as well as theoretical properties of the proposed algorithms for finite-sum minimization problems (**P1**) and (**P2**) are given in Sect. 3. Conclusions and further thoughts are gathered in Sect. 4.

1.1 Notation and definitions

Throughout the paper, vectors are denoted by bold lowercase letters, e.g., $\mathbf{v}$, and matrices or random variables are denoted by bold upper case letters, e.g., $\mathbf{V}$. $\mathbf{v}^T$ denotes the transpose of a real vector $\mathbf{v}$. We use regular lower-case and upper-case letters to denote scalar constants, e.g., c or K. For two vectors, $\mathbf{v}$, $\mathbf{w}$, their inner-product is denoted as $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w}$. For a vector $\mathbf{v}$, and a matrix $\mathbf{V}$, $\|\mathbf{v}\|$ and $\|\mathbf{V}\|$ denote the vector ℓ_2 norm and the matrix spectral norm, respectively, while $\|\mathbf{V}\|_F$ is the matrix Frobenius norm. $\nabla F(\mathbf{x})$ and $\nabla^2 F(\mathbf{x})$ are the gradient and the Hessian of F at $\mathbf{x}$, respectively, and $\mathbb{I}$ denotes the identity matrix. For two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is symmetric positive semi-definite. The subscript, e.g., $\mathbf{x}_t$, denotes iteration counter and $\log(x)$ is the natural logarithm of x. The inexact Hessian is denoted by $\mathbf{H}(\mathbf{x})$, but for notational simplicity, we may use $\mathbf{H}_t$ to, instead, denote the approximate Hessian evaluated at the iterate $\mathbf{x}_t$ in iteration t, i.e., $\mathbf{H}_t \triangleq \mathbf{H}(\mathbf{x}_t)$. Throughout the paper, $\mathcal{S}$ denotes a collection of indices from $\{1, 2, \cdots, n\}$, with potentially repeated items and its cardinality is denoted by $|\mathcal{S}|$.

Unlike convex functions for which "local optimality" and "global optimality" are in fact the same, in non-convex settings, we are often left with designing algorithms that can guarantee convergence to approximate local optimality. In this light, throughout this paper, we make use of the following definition of $(\varepsilon_g, \varepsilon_H)$-Optimality:

Definition 1 ($(\varepsilon_g, \varepsilon_H)$-Optimality) Given $\varepsilon_g, \varepsilon_H \in (0, 1)$, $\mathbf{x} \in \mathbb{R}^d$ is an $(\varepsilon_g, \varepsilon_H)$-optimal solution to the problem (**P0**), if

$$\|\nabla F(\mathbf{x})\| \leq \varepsilon_g, \quad \lambda_{\min}(\nabla^2 F(\mathbf{x})) \geq -\varepsilon_H. \tag{1}$$

We note that $(\varepsilon_g, \varepsilon_H)$-Optimality (even with $\varepsilon_g = \varepsilon_H = 0$) does not necessarily imply closeness to any local minimum, neither in iterate nor in the objective value. However, if the saddle points satisfy the strict-saddle property [26,40], then an $(\varepsilon_g, \varepsilon_H)$-optimality guarantees vicinity to a local minimum for sufficiently small ε_g and ε_H.

1.2 Background and related work

Arguably, the most straightforward approach for *globalization* of many Newton-type algorithms is the application of line-search. However, near saddle points where the gradient magnitude can be small, traditional line search methods can be very ineffective

and in fact produce iterates that can get stuck at a saddle point [46]. Trust region and cubic regularization methods are two elegant globalization alternatives that, specially recently, have attracted much attention. The main advantage of these methods is that they are reliably able to take advantage of the direction of negative curvature and escape saddle points. In this section we briefly review these algorithms as they pertain to the present paper and mention the relevant prior works.

1.2.1 Trust region

TR methods [17,54] encompass a general class of iterative methods which specifically define a region around the current iterate within which they trust the model to be a reasonable approximation of the true objective function. The most widely used approximating model, which we consider here, is done via a quadratic function. More specifically, using the current iterate $\mathbf{x}_t$, the quadratic variant of TR algorithm finds the next iterate as $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{s}_t$ where $\mathbf{s}_t$ is a solution of the *constrained* sub-problem

$$\begin{aligned} \min \quad & m_t(\mathbf{s}) \triangleq \langle \mathbf{s}, \nabla F(\mathbf{x}_t) \rangle + \frac{1}{2} \langle \mathbf{s}, \nabla^2 F(\mathbf{x}_t) \mathbf{s} \rangle \\ \text{s.t.} \quad & \|\mathbf{s}\|_2 \leq \Delta_t. \end{aligned} \tag{2a}$$

Here, Δ_t is the region in which we "trust" our quadratic model to be an acceptable approximation of the true objective for the current iteration. The major bottleneck of computations in TR algorithm is the minimization of the constrained quadratic sub-problem (2a), for which numerous approaches have been proposed, e.g., [23,28,29, 36,41,43,53,56].

For a smooth non-convex objective and in order to obtain approximate first-order criticality, i.e., $\|\nabla F(\mathbf{x}_t)\| \leq \varepsilon_g$ for some $\varepsilon_g \in (0, 1)$, the complexity of an (inexact) trust-region method, which ensures at least a Cauchy (steepest-descent-like) decrease at each iteration, is shown to be of the same order as that of steepest descent, i.e., $\mathcal{O}(\varepsilon_g^{-2})$; e.g., [5,13,31–33]. Recent non-trivial modifications of the classical TR methods have also been proposed which improve upon the complexity to $\mathcal{O}(\varepsilon_g^{-3/2})$; see [20] and further extensions to a more general framework in [19]. These bounds can be shown to be tight [9] in the worst case. Under a more general algorithmic framework and in terms of objective function sub-optimality, i.e., $F(\mathbf{x}) - F^* \leq \varepsilon$, better complexity bounds, in the convex and strongly-convex settings, have been obtained which are of the orders of $\mathcal{O}(\varepsilon_g^{-1})$ and $\mathcal{O}(\log(1/\varepsilon_g))$, respectively [30].

For non-convex problems, however, it is more desired to obtain complexity bounds for achieving approximate second-order criticality, i.e., Definition 1. For this, bounds in the orders of $\mathcal{O}(\max\{\varepsilon_H^{-1}\varepsilon_g^{-2}, \varepsilon_H^{-3}\})$ and $\mathcal{O}(\max\{\varepsilon_g^{-3}, \varepsilon_H^{-3}\})$ have been obtained in [13] and [30], respectively. Similar bounds were also given in [32] under probabilistic model. Bounds of this order have shown to be optimal in certain cases [13].

More closely related to the present paper, there have been several results which study the role of derivative-free and probabilistic models in general, and Hessian approximation in particular, e.g., see [2,5,13,16,18,32,39,52] and references therein.

1.2.2 Cubic regularization

An alternative to the traditional line-search and TR for globalization of Newton-type methods is the application of cubic regularization. Such class of methods is characterized by generating iterates as $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{s}_t$ where $\mathbf{s}_t$ is a solution of the following *unconstrained* sub-problem

$$\min_{\mathbf{s}\in\mathbb{R}^d} \quad m_t(\mathbf{s}) \triangleq \langle \mathbf{s}, \nabla F(\mathbf{x}_t)\rangle + \frac{1}{2}\langle \mathbf{s}, \nabla^2 F(\mathbf{x}_t)\mathbf{s}\rangle + \frac{\sigma_t}{3}\|\mathbf{s}\|^3, \tag{2b}$$

where σ_t is the cubic regularization parameter chosen for the current iteration. As in the case of TR, the major bottleneck of CR involved solving the sub-problem (2b), for which various techniques have been proposed, e.g., [1,4,8,10].

To the best of our knowledge, the use of such regularization, was first introduced in the pioneering work of [34], and subsequently further studied in the seminal works of [10,11,45]. From the worst-case complexity point of view, CR has a better dependence on ε_g compared to TR. More specifically, [45] showed that, under global Lipschitz continuity assumption on the Hessian, if the sub-problem (2b) is solved exactly, then the resulting CR algorithm achieves the approximate first-order criticality with complexity of $\mathcal{O}(\varepsilon_g^{-3/2})$. These results were extended by the pioneering and seminal works of [10,11] to an adaptive variant, which is often referred to as ARC (Adaptive Regularization with Cubics). In particular, the authors showed that the worst case complexity of $\mathcal{O}(\varepsilon_g^{-3/2})$ can be achieved without requiring the knowledge of the Hessian's Lipschitz constant, access to the exact Hessian, or multi-dimensional global optimization of the sub-problem (2b). These results were further refined in [13] where it was shown that, not only, multi-dimensional global minimization of (2b) is unnecessary, but also the same complexity can be achieved with mere one or two dimensional search. This $\mathcal{O}(\varepsilon^{-3/2})$ bound has been shown to be tight [12]. As for the approximate second-order criticality, [13] showed that at least $\mathcal{O}(\max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\})$ is required. With further assumptions on the inexactness of sub-problem solution, [11,13] also show that one can achieve $\mathcal{O}(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\})$, which is shown to be tight [9]. Better dependence on ε_g can be obtained if one assumes additional structure, such as convexity, e.g., see [14,45] as well as the acceleration scheme of [44].

Recently, for (strongly) convex problems, [27] obtained sub-optimal complexity for ARC and its accelerated variants using Hessian approximations. In the context of stochastic optimization problems, [57] considers cubic regularization with a priori chosen fixed regularization parameter using both approximations of the gradients and Hessian. Specific to the finite-sum problem (**P1**) , and by a direct application of the theoretical results of [10,11,37] presents a sub-sampled variant of ARC, in which the exact Hessian and the gradient are replaced by sub-samples. However, unfortunately, their analysis suffers from a rather vicious circle: the approximate Hessian and gradient are formed based on an *a priori unknown* step which can only be determined after such approximations are formed.

1.3 Contributions

In this section, we summarize the key aspects of our contributions. In Sect. 2, we consider (**P0**) and establish the worst-case iteration complexities for variants of trust-region and adaptive cubic regularization methods in which the Hessian is suitably approximated. More specifically, our entire analysis is based on the following key condition on the approximate Hessian $\mathbf{H}(\mathbf{x})$:

Condition 1 (Inexact Hessian regularity) *For some* $0 < K_H < \infty$, $\varepsilon > 0$, *the approximating Hessian,* $\mathbf{H}(\mathbf{x}_t)$, *satisfies*

$$\left\| \left(\mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t) \right) \mathbf{s}_t \right\| \leq \varepsilon \cdot \|\mathbf{s}_t\|, \tag{3a}$$

$$\|\mathbf{H}(\mathbf{x}_t)\| \leq K_H, \tag{3b}$$

where $\mathbf{x}_t$ *and* $\mathbf{s}_t$ *are, respectively, the iterate and the update at iteration* t.

Under Condition 1, we show that our proposed algorithms (Algorithms 1 and 2) achieve *the same worst-case iteration complexity* to obtain *approximate second order critical solution* as that of the exact variants (Theorems 1, 2, and 3).

In Sect. 3, we describe schemes for constructing $\mathbf{H}(\mathbf{x}_t)$ to satisfy Condition 1. Specifically, in the context of finite-sum optimization framework, i.e., problems (**P1**) and (**P2**), we present various *sub-sampling* schemes to probabilistically ensure Condition 1 (Lemmas 16 and 17). Our proposed randomized sub-sampling strategies guarantee, with high probability, a stronger condition than (3a), namely

$$\|\mathbf{H}(\mathbf{x}) - \nabla^2 F(\mathbf{x})\| \leq \varepsilon. \tag{4}$$

It is clear that (4) implies (3a). We then give *optimal* iteration complexities for Algorithms 1 and 2 for optimization of non-convex finite-sum problems where the Hessian is approximated by means of appropriate sub-sampling (Theorems 4, 5 and 6).

To establish optimal second-order iteration complexity, many previous works considered Hessian approximation conditions that, while enjoying many advantages, come with certain disadvantages. Our proposed Condition 1 aims to remedy some of these disadvantages. We first briefly review the conditions used in the prior works, and subsequently highlight the merits of Condition 1 in comparison.

1.3.1 Conditions used in prior works

For the analysis of trust-region, many authors have considered the following condition

$$\left\| \mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t + \mathbf{s}) \right\| \leq C_1 \Delta_t, \quad \forall \mathbf{s} \in \{\mathbf{s};\ \|\mathbf{s}\| \leq \Delta_t\}, \tag{5a}$$

for some $0 < C_1 < \infty$, where Δ_t is the current trust-region radius, e.g., [2,32]. In [5], condition (5a) is replaced with

$$\left\| \mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t) \right\| \leq C_2 \Delta_t, \tag{5b}$$

for some $0 < C_2 < \infty$. In fact, by assuming Lipschitz continuity of Hessian, it is easy to show that (5a) and (5b) are equivalent, in that one implies the other, albeit with modified constants. We also note that [2,5,32] study a more general framework under which the entire sub-problem model is probabilistically constructed and approximation extends beyond just the Hessian.

For cubic regularization, the condition imposed on the inexact Hessian is often considered as

$$\left\| \left(\mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t) \right) \mathbf{s}_t \right\| \leq C_3 \|\mathbf{s}_t\|^2, \tag{5c}$$

for some $0 < C_3 < \infty$, e.g., [10,11,13] and other follow-up works. In fact, [13] has also established optimal iteration complexity for trust-region algorithm under (5c). Both of (5a) and (5c), are stronger than the celebrated Dennis-Moré [21] condition, i.e.,

$$\lim_{t \to \infty} \frac{\left\| \left(\mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t) \right) \mathbf{s}_t \right\|}{\|\mathbf{s}_t\|} = 0.$$

Indeed, under certain assumptions, Dennis-Moré condition is satisfied by a number of quasi-Newton methods, although the same cannot be said about (5a) and (5c) [10].

1.3.2 Merits of Condition 1

For our trust-region analysis, we require Condition 1 with $\varepsilon \in \mathcal{O}(\max\{\varepsilon_H, \Delta_t\})$; see (11) in Theorem 1. Hence, when Δ_t is large, e.g., at the beginning of iterations, all the conditions (3a), (5a), and (5b) are equivalent, up to some constants. However, the constants in (5a) and (5b) can be larger than what is implied by (3a), amounting to cruder approximations in practice for when Δ_t is large. As iterations progress, the trust-region radius will get smaller, and in fact it is expected that Δ_t will eventually shrink to be $\Delta_t \in \Theta\left(\min\{\varepsilon_g, \varepsilon_H\}\right)$. In prior works, e.g., [5,32], the convergence analysis is derived using $\varepsilon_H = \varepsilon_g$, whereas here we allow $\varepsilon_H = \sqrt{\varepsilon_g}$. As a result, the requirements (5a) and (5b) can eventually amount to stricter conditions than (3a).

As for (5c), the main drawback lies in the difficulty of enforcing it. Despite the fact that for certain values of $\|\mathbf{s}_t\|$ and ε, e.g., $\varepsilon \ll \|\mathbf{s}_t\|$, (5c) can be less restrictive than (3a), a priori enforcing (5c) requires one to have already computed the search direction $\mathbf{s}_t$, which itself can be done only after $\mathbf{H}(\mathbf{x}_t)$ is constructed, hence creating a vicious circle. A posteriori guarantees can be given if one obtains a lower-bound estimate on the yet-to-be-computed step-size, i.e., to have $s_0 > 0$ such that $s_0 \leq \|\mathbf{s}_t\|$. This allows one to consider a stronger condition as $\left\| \left(\mathbf{H}(\mathbf{x}_t) - \nabla^2 F(\mathbf{x}_t) \right) \right\| \leq C_3 s_0$, which can be enforced using a variety of methods such as those described in Sect. 3. However, to obtain such a lower-bound estimate on the next step-size, one has to resort to a recursive procedure, which necessitates repeated constructions of the approximate Hessian and subsequent solutions of the corresponding subproblems. Consequently, this procedure may result in a significant computational overhead and will lead to undesirable theoretical complexities.

In sharp contrast to (5c), the condition (3a) allows for theoretically principled use of many practical techniques to construct $\mathbf{H}_t$. For example, under (3a), the use of quasi-Newton methods to approximate the Hessian is theoretically justified. Further, by considering the stronger condition (4), many *randomized matrix approximation* techniques can be readily applied, e.g., [42,58–60]; see Sect. 3. To the best of our knowledge, the only successful attempt at guaranteed a priori construction of $\mathbf{H}_t$ using (5c) is done in [15]. Specifically, by considering probabilistic models, which are "sufficiently accurate" in that they are partly based on (5c), [15] studies first-order complexity of a large class of methods, including ARC, and discusses ways to construct such probabilistic models as long as the gradient is large enough, i.e., before first-order approximate-optimality is achieved. Here, by considering (3a), we are able to provide an alternative analysis, which allows us to obtain second-order complexity results.

Requiring (4), as a way of enforcing (3a), offers a variety of other practical advantages, which are not readily available with other conditions. For example, consider distributed/parallel environments where the data is distributed across a network and the main bottleneck of computations is the communications across the nodes. In such settings, since (4) allows for the Hessian accuracy to be set a priori and to remain fixed across all iterations, the number of samples in each node can stay the same throughout iterations. This prevents unnecessary communications to re-distribute the data at every iteration.

Furthermore, in case of failed iterations, i.e., when the computed steps are rejected, the previous $\mathbf{H}_t$ may seamlessly be used in the next iteration, which avoids repeating many such, potentially expensive, computations throughout the iterations. For example, consider approximate solutions to the underlying sub-problems by means of dimensionality reduction, i.e., $\mathbf{H}_t$ is projected onto a lower dimensional sub-space as $\mathbf{U}^T\mathbf{H}_t\mathbf{U}$ for some $\mathbf{U} \in \mathbb{R}^{d\times p}$ with $p \ll d$, resulting in a smaller dimensional sub-problem. Now if the current iteration leads to a rejected step, the projection of the $\mathbf{H}_t$ from the previous iteration can be readily re-used in the next iteration. This naturally amounts to saving further Hessian computations.

2 Algorithms and convergence analysis

We are now ready to present our main algorithms for solving the generic non-convex optimization (**P0**) along with their corresponding iteration complexity results to obtain a $(\varepsilon_g, \varepsilon_H)$-optimal solution as in (1). More precisely, in Sects. 2.1 and 2.2 , respectively, we present modifications of the TR and ARC methods which incorporate inexact Hessian information, according to Condition 1.

We remind that, though not specifically mentioned in the statement of the theorems or the algorithms, when the computed steps are rejected and an iteration needs to be repeated with different Δ_t or σ_t, the previous $\mathbf{H}_t$ may seamlessly be used in the next iteration. This can be a desirable feature in many practical situations and is directly the result of enforcing (4); see also the discussion in Sect. 1.3.2.

For our analysis throughout the paper, we make the following standard assumption regarding the regularity of the exact Hessian of the objective function F.

Assumption 1 (*Hessian regularity*) $F(\mathbf{x})$ is twice differentiable and has bounded and Lipschitz continuous Hessian on the piece-wise linear path generated by the iterates, i.e. for some $0 < K, L < \infty$ and all iterations

$$\left\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{x}_t)\right\| \leq L\|\mathbf{x} - \mathbf{x}_t\|, \ \forall \mathbf{x} \in [\mathbf{x}_t, \mathbf{x}_t + \mathbf{s}_t], \tag{6a}$$

$$\left\|\nabla^2 F(\mathbf{x}_t)\right\| \leq K, \tag{6b}$$

where $\mathbf{x}_t$ and $\mathbf{s}_t$ are, respectively, the iterate and the update step at iteration t.

Although, we do not know of a particular way to, a priori, verify (6a), it is clear that Assumption (6a) is weaker than Lipschitz continuity of the Hessian for all $\mathbf{x}$, i.e.,

$$\left\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\right\| \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{7}$$

Despite the fact that theoretically (6a) is weaker than (7), to the best of our knowledge as of yet, (7) is the only practical sufficient condition for verifying (6a).

2.1 Trust region with inexact Hessian

Algorithm 1 depicts a trust-region algorithm where at each iteration t, instead of the true Hessian $\nabla^2 F(\mathbf{x}_t)$, only an inexact approximation, $\mathbf{H}_t$, is used. For Algorithm 1, the accuracy tolerance in (3a) is adaptively chosen as $\varepsilon_t \leq \max\{\varepsilon_0, \Delta_t\}$, where Δ_t is the trust region in the t-th iteration and $\varepsilon_0 \in \mathcal{O}(\varepsilon_H)$ is some fixed threshold. This allows for a very crude approximation at the beginning of iterations, when Δ_t is large.

Algorithm 1 Trust Region with Inexact Hessian

Input: Starting point $\mathbf{x}_0$, initial radius $0 < \Delta_0 < \infty$, hyper-parameters $\varepsilon_0, \varepsilon_g, \varepsilon_H, \eta \in (0, 1), \gamma > 1$
for $t = 0, 1, \ldots$ **do**
  Set the approximate Hessian, $\mathbf{H}_t$, as in (3) with $\varepsilon_t \leq \max\{\varepsilon_0, \Delta_t\}$
  if $\|\nabla F(\mathbf{x}_t)\| \leq \varepsilon_g, \lambda_{\min}(\mathbf{H}_t) \geq -\varepsilon_H$ **then**
    Return $\mathbf{x}_t$.
  end if
  Solve the sub-problem approximately

$$\mathbf{s}_t \approx \arg\min_{\|\mathbf{s}\| \leq \Delta_t} m_t(s) \triangleq \langle \nabla F(\mathbf{x}_t), \mathbf{s}\rangle + \frac{1}{2}\langle \mathbf{s}, \mathbf{H}_t \mathbf{s}\rangle \tag{8}$$

  Set $\rho_t \triangleq \dfrac{F(\mathbf{x}_t) - F(\mathbf{x}_t + \mathbf{s}_t)}{-m_t(\mathbf{s}_t)}$
  if $\rho_t \geq \eta$ **then**
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{s}_t$
    $\Delta_{t+1} = \gamma \Delta_t$
  else
    $\mathbf{x}_{t+1} = \mathbf{x}_t$
    $\Delta_{t+1} = \Delta_t / \gamma$
  end if
end for
Output: $\mathbf{x}_t$

As iterations progress towards optimality and Δ_t gets small, the threshold ε_0 can prevent ε from getting unnecessarily too small.

In Algorithm 1, we require that the sub-problem (8) is solved only approximately. Indeed, in large-scale problems, where the exact solution of the sub-problem is the main bottleneck of the computations, this is a very crucial relaxation. Such approximate solution of the sub-problem (8) has been adopted in many previous work. Here, we follow the inexactness conditions discussed in [17], which are widely known as Cauchy and Eigenpoint conditions. Recall that the Cauchy and Eigen directions correspond, respectively, to one dimensional minimization of the sub-problem (8) along the directions given by the gradient and negative curvature.

Condition 2 (Sufficient descent Cauchy and Eigen directions [17]) *Assume that we solve the sub-problem* (8) *approximately to find* $\mathbf{s}_t$ *such that*

$$-m_t(\mathbf{s}_t) \geq -m_t(\mathbf{s}_t^C) \geq \frac{1}{2}\|\nabla F(\mathbf{x}_t)\| \min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{1+\|\mathbf{H}_t\|}, \Delta_t\right\}, \tag{9a}$$

$$-m_t(\mathbf{s}_t) \geq -m_t(\mathbf{s}_t^E) \geq \frac{1}{2}\nu|\lambda_{\min}(\mathbf{H}_t)|\Delta_t^2, \quad \text{if} \quad \lambda_{\min}(\mathbf{H}_t) < 0. \tag{9b}$$

Here, $m_t(\cdot)$ *is defined in* (8), $\mathbf{s}_t^C$ *(Cauchy point) is along negative gradient direction and* $\mathbf{s}_t^E$ *is along approximate negative curvature direction such that* $\langle \mathbf{s}_t^E, \mathbf{H}_t\mathbf{s}_t^E\rangle \leq \nu\lambda_{\min}(\mathbf{H}_t)\|\mathbf{s}_t^E\|^2 < 0$*, for some* $\nu \in (0, 1]$ *(see Appendix B for a way to efficiently compute* $\mathbf{s}_t^E$*).*

One way to ensure that an approximate solution to the sub-problem (8) satisfies (9), is by replacing (8) with the following reduced-dimension problem, in which the search space is a two-dimensional sub-space containing vectors $\mathbf{s}_t^C$, and $\mathbf{s}_t^E$, i.e.,

$$\mathbf{s}_t = \arg \min_{\substack{\|\mathbf{s}\| \leq \Delta_t \\ \mathbf{s} \in \text{Span}\{\mathbf{s}_t^C, \mathbf{s}_t^E\}}} \langle \nabla F(\mathbf{x}_t), \mathbf{s}\rangle + \frac{1}{2}\langle \mathbf{s}, \mathbf{H}_t\mathbf{s}\rangle.$$

Of course, any larger dimensional sub-space $\mathscr{P}$ for which we have $\text{Span}\{\mathbf{s}_t^C, \mathbf{s}_t^E\} \subseteq \mathscr{P}$ would also guarantee (9). In fact, a larger dimensional sub-space implies a more accurate solution to our original sub-problem (8).

We now set out to provide iteration complexity for Algorithm 1. Our analysis follows similar line of reasoning as that in [10,11,13]. First, we show the discrepancy between the quadratic model and objective function in Lemma 1.

Lemma 1 *Given Assumption* 1 *and Condition* (3a) *with any* $\varepsilon_t > 0$*, we have*

$$|F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t)| \leq \frac{L}{2}\Delta_t^3 + \frac{\varepsilon_t}{2}\Delta_t^2. \tag{10}$$

Proof Applying Mean Value Theorem on F at $\mathbf{x}_t$ gives $F(\mathbf{x}_t + \mathbf{s}_t) = F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)^T \mathbf{s}_t + \frac{1}{2}\mathbf{s}_t^T \nabla^2 F(\xi_t)\mathbf{s}_t$, for some ξ_t in the segment of $[\mathbf{x}_t, \mathbf{x}_t + \mathbf{s}_t]$. We have

$$\begin{aligned}
|F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t)| &= \frac{1}{2}\left|\mathbf{s}_t^T(\nabla^2 F(\xi_t) - \mathbf{H}_t)\mathbf{s}_t\right| \\
&= \frac{1}{2}\left|\mathbf{s}_t^T(\nabla^2 F(\xi_t) - \nabla^2 F(\mathbf{x}_t) + \nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\mathbf{s}_t\right| \\
&\le \frac{1}{2}\left|\mathbf{s}_t^T(\nabla^2 F(\xi_t) - \nabla^2 F(\mathbf{x}_t))\mathbf{s}_t\right| + \frac{1}{2}\left|\mathbf{s}_t^T(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t)\mathbf{s}_t\right| \\
&\le \frac{L}{2}\|\mathbf{s}_t\|^3 + \frac{\varepsilon_t}{2}\|\mathbf{s}_t\|^2 \le \frac{L}{2}\Delta_t^3 + \frac{\varepsilon_t}{2}\Delta_t^2.
\end{aligned}$$

□

Combining with Conditions 1 and 2 , we get the following two lemmas that characterize sufficient conditions for successful iterations.

Lemma 2 *Consider any $\varepsilon_H > 0$, let $\varepsilon_0 \triangleq \alpha(1-\eta)\nu\varepsilon_H$ for some $\alpha \in (0,1)$, and suppose Condition 1 is satisfied with $\varepsilon_t \le \max\{\varepsilon_0, \Delta_t\}$, where Δ_t is the trust region at the t-th iteration. Given Assumption 1 and Condition 2, if $\lambda_{\min}(\mathbf{H}_t) < -\varepsilon_H$ and $\Delta_t \le (1-\alpha)(1-\eta)\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,/(L+1)$, then the t-th iteration is successful, i.e. $\Delta_{t+1} = \gamma\Delta_t$.*

Proof Suppose $\Delta_t \le \varepsilon_0$. From (9b) and (10), we have

$$\begin{aligned}
1 - \rho_t = \frac{F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t)}{-m_t(\mathbf{s}_t)} &\le \frac{L\Delta_t^3 + \varepsilon_t\Delta_t^2}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2} \le \frac{L\Delta_t^3 + \alpha(1-\eta)\nu\varepsilon_H\Delta_t^2}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2} \\
&\le \frac{L\Delta_t^3 + \alpha(1-\eta)\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2} \le \frac{L\Delta_t + \alpha(1-\eta)\nu\,|\lambda_{\min}(\mathbf{H}_t)|}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|}.
\end{aligned}$$

By the assumption on Δ_t, we get $\rho_t \ge \eta$ and the iteration is successful. Now consider $\Delta_t \ge \varepsilon_0$. Similar to the above, we have

$$\begin{aligned}
1 - \rho_t = \frac{F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t)}{-m_t(\mathbf{s}_t)} &\le \frac{L\Delta_t^3 + \varepsilon_t\Delta_t^2}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2} \\
&\le \frac{(L+1)\Delta_t^3}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2} \le \frac{(L+1)\Delta_t}{\nu\,|\lambda_{\min}(\mathbf{H}_t)|},
\end{aligned}$$

which again by assumption on Δ_t and noting $\alpha < 1$, we get $\rho_t \ge \eta$ and the iteration is successful. □

Lemma 3 *Suppose Condition 1 is satisfied with any $\varepsilon_t > 0$. Given Assumption 1 and Condition 2, if $\|\nabla F(\mathbf{x}_t)\| > \varepsilon_g$ and*

$$\Delta_t \le \min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{(1+K_H)}, \frac{\sqrt{\varepsilon_t^2 + 4L(1-\eta)\|\nabla F(\mathbf{x}_t)\|} - \varepsilon_t}{2L}\right\},$$

then, the t-th iteration is successful, i.e. $\Delta_{t+1} = \gamma \Delta_t$.

Proof By assumption on Δ_t, (9a), and since $\|\nabla F(\mathbf{x}_t)\| > \varepsilon_g$, we have

$$-m_t(\mathbf{s}_t) \geq \frac{1}{2}\|\nabla F(\mathbf{x}_t)\| \min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{1+\|\mathbf{H}_t\|}, \Delta_t\right\} \geq \frac{1}{2}\|\nabla F(\mathbf{x}_t)\|\Delta_t.$$

Therefore,

$$1-\rho_t = \frac{F(\mathbf{x}_t+\mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t)}{-m_t(\mathbf{s}_t)} \leq \frac{L\Delta_t^3 + \varepsilon_t\Delta_t^2}{\|\nabla F(\mathbf{x}_t)\|\Delta_t} \leq \frac{L\Delta_t^2 + \varepsilon_t\Delta_t}{\|\nabla F(\mathbf{x}_t)\|} \leq 1-\eta,$$

where the last inequality follows by assumption on Δ_t. So $\rho_t \geq \eta$, which means the iteration is successful. □

Lemma 4 gives a lower bound for the trust region radius before the algorithm terminates, i.e., this ensures that the trust region never shrinks to become too small.

Lemma 4 *Consider any* $\varepsilon_g, \varepsilon_H > 0$ *such that* $\varepsilon_H \leq \sqrt{\varepsilon_g}$ *and let* $\varepsilon_0 \triangleq \alpha(1-\eta)\nu\varepsilon_H$ *for some* $\alpha \in (0,1)$. *Further, suppose Condition 1 is satisfied with* $\varepsilon_t \leq \max\{\varepsilon_0, \Delta_t\}$, *where* Δ_t *is the trust region at the t-th iteration. For Algorithm 1, under Assumption 1 and Condition 2, we have* $\Delta_t \geq \kappa_\Delta \min\{\varepsilon_g, \varepsilon_H\}, \forall t \geq 0$, *where*

$$\kappa_\Delta \triangleq \min\{\kappa_1, \kappa_2, \kappa_3, \kappa_4\}/\gamma, \quad \kappa_1 \triangleq (1-\alpha)(1-\eta)\nu/(L+1), \quad \kappa_2 \triangleq \alpha(1-\eta)\nu,$$
$$\kappa_3 \triangleq 1/(1+K_H), \quad \kappa_4 \triangleq \sqrt{(\alpha(1-\eta)\nu)^2 + 4L(1-\eta)} - \alpha(1-\eta)\nu/(2L).$$

Proof We prove by contradiction. Assume that the t-th iteration is the first unsuccessful iteration such that $\Delta_{t+1} = \Delta_t/\gamma \leq \kappa_\Delta \min\{\varepsilon_g, \varepsilon_H\}$, i.e., we have

$$\Delta_t \leq \min\{\kappa_1, \kappa_2, \kappa_3, \kappa_4\} \cdot \min\{\varepsilon_g, \varepsilon_H\}.$$

Suppose $\lambda_{\min}(\mathbf{H}_t) < -\varepsilon_H$. By Lemma 2, since $\Delta_t \leq (1-\alpha)(1-\eta)\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,/\,(L+1)$, iteration t must have been accepted and we must have $\Delta_{t+1} = \gamma\Delta_t > \Delta_t$, which is a contradiction. Now suppose $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$. By assumption on Δ_t, we have that $\Delta_t \leq \kappa_2\varepsilon_H = \varepsilon_0$, which implies that $\varepsilon_t \leq \varepsilon_0$. Since the function $h(a,b) \triangleq -a + \sqrt{a^2+b}$, for any fixed $b > 0$, is decreasing in a, and for any fixed a, is increasing in $b \geq 0$, we have

$$\begin{aligned} h(\varepsilon_t, 4L(1-\eta)\|\nabla F(\mathbf{x}_t)\|) &\geq h(\varepsilon_0, 4L(1-\eta)\|\nabla F(\mathbf{x}_t)\|) \\ &\geq h(\varepsilon_0, 4L(1-\eta)\varepsilon_g) \geq h(\varepsilon_0, 4L(1-\eta)\varepsilon_H^2), \end{aligned}$$

which implies

$$\frac{\sqrt{\varepsilon_t^2 + 4L(1-\eta)\|\nabla F(\mathbf{x}_t)\|} - \varepsilon_t}{2L} \geq \kappa_4\varepsilon_H.$$

As a result, since $\Delta_t \leq \min\{\kappa_3\varepsilon_g, \kappa_4\varepsilon_H\}$, it must satisfy the condition of Lemma 3. This implies that iteration t must have been accepted, which is a contradiction. □

The following lemma follows closely the line of reasoning in [13, Lemma 4.5].

Lemma 5 (Successful iterations) *Consider any $\varepsilon_g, \varepsilon_H > 0$ such that $\varepsilon_H \leq \sqrt{\varepsilon_g}$ and let $\varepsilon_0 \triangleq \alpha(1-\eta)\nu\varepsilon_H$ for some $\alpha \in (0,1)$. Further, suppose Condition 1 is satisfied with $\varepsilon_t \leq \max\{\varepsilon_0, \Delta_t\}$, where Δ_t is the trust region at the t-th iteration. Let $\mathcal{T}_{succ}$ denote the set of all the successful iterations before Algorithm 1 stops. Then, under Assumption 1, Condition 2, the number of successful iterations is upper bounded by,*

$$|\mathcal{T}_{succ}| \leq \frac{(F(\mathbf{x}_0) - F_{\min})}{\eta \min\{\widehat{\kappa}_\Delta, \widetilde{\kappa}_\Delta\}} \cdot \max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\}$$

where $\widehat{\kappa}_\Delta \triangleq \kappa_\Delta/2, \widetilde{\kappa}_\Delta \triangleq \nu\kappa_\Delta^2/2$, and κ_Δ is as defined in Lemma 4.

Proof Suppose Algorithm 1 doesn't terminate at the t-th iteration. Then we have either $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$ or $\lambda_{\min}(\Delta^2 F(\mathbf{x}_t)) \leq -\varepsilon_H$. In the first case, from (9a), we have

$$\begin{aligned} -m_t(\mathbf{s}_t) &\geq \frac{\varepsilon_g}{2}\min\left\{\frac{\varepsilon_g}{1+K_H}, \Delta_t\right\} \geq \frac{\varepsilon_g}{2}\min\left\{\frac{\varepsilon_g}{1+K_H}, \kappa_\Delta\varepsilon_g, \kappa_\Delta\varepsilon_H\right\} \\ &\geq \widehat{\kappa}_\Delta\varepsilon_g \min\{\varepsilon_g, \varepsilon_H\}, \end{aligned}$$

where κ_Δ is as defined in Lemma 4. Similarly, in the second case, from (9b), we obtain

$$-m_t(\mathbf{s}_t) \geq \frac{1}{2}\nu\,|\lambda_{\min}(\mathbf{H}_t)|\,\Delta_t^2 \geq \frac{1}{2}\nu\kappa_\Delta^2\varepsilon_H \min\{\varepsilon_g^2, \varepsilon_H^2\} = \widetilde{\kappa}_\Delta\varepsilon_H \min\{\varepsilon_g^2, \varepsilon_H^2\}.$$

Since $F(\mathbf{x}_t)$ is monotonically decreasing, we have

$$\begin{aligned} F(\mathbf{x}_0) - F_{\min} &\geq \sum_{t=0}^{\infty} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \geq \sum_{t\in\mathcal{T}_{succ}} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \\ &\geq \eta \sum_{t\in\mathcal{T}_{succ}} \min\left\{\widehat{\kappa}_\Delta\varepsilon_g \min\{\varepsilon_g, \varepsilon_H\}, \widetilde{\kappa}_\Delta\varepsilon_H \min\{\varepsilon_g^2, \varepsilon_H^2\}\right\} \\ &\geq |\mathcal{T}_{succ}|\,\eta \min\{\widehat{\kappa}_\Delta, \widetilde{\kappa}_\Delta\} \min\{\varepsilon_g^2\varepsilon_H, \varepsilon_H^3\}. \end{aligned}$$

Hence, we have $|\mathcal{T}_{succ}| \leq (F(\mathbf{x}_0) - F_{\min}) \max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\}/(\eta \min\{\widehat{\kappa}_\Delta, \widetilde{\kappa}_\Delta\})$. □

Now we are ready to present the final complexity in Theorem 1.

Theorem 1 (Optimal complexity of Algorithm 1) *Consider any $\varepsilon_g, \varepsilon_H > 0$ such that $\varepsilon_H \leq \sqrt{\varepsilon_g}$ and let $\varepsilon_0 \triangleq \alpha(1-\eta)\nu\varepsilon_H$ for some $\alpha \in (0,1)$ where η is a hyper-parameter in Algorithm 1, and ν is as in (9b). Suppose the inexact Hessian, $\mathbf{H}(\mathbf{x})$, satisfies Condition 1 with the approximation tolerance, ε_t, in (3a) as*

$$\varepsilon_t \leq \max\{\varepsilon_0, \Delta_t\}, \tag{11}$$

where Δ_t is the trust region at the t-th iteration. For Problem (**P0**)*, under Assumption* 1 *and Condition* 2*, Algorithm* 1 *terminates after at most* $T \in \mathcal{O}\left(\max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\}\right)$ *iterations.*

Proof Suppose Algorithm 1 terminates at the t-th iteration. Let $\mathcal{T}_{succ}$ and $\mathcal{T}_{fail}$ denote the sets of all the successful and unsuccessful iterations, respectively. Then $T = |\mathcal{T}_{succ}| + |\mathcal{T}_{fail}|$ and $\Delta_T = \Delta_0\gamma^{|\mathcal{T}_{succ}|-|\mathcal{T}_{fail}|}$, where γ is a hyper-parameter of Algorithm 1. From Lemma 4, we have $\Delta_T \geq \kappa_\Delta \min\{\varepsilon_g, \varepsilon_H\}$. Hence, $\left(|\mathcal{T}_{succ}| - |\mathcal{T}_{fail}|\right)\log\gamma \geq \log\left(\kappa_\Delta \cdot \min\{\varepsilon_g, \varepsilon_H\}/\Delta_0\right)$, which implies $|\mathcal{T}_{fail}| \leq \log\left(\Delta_0/\left(\kappa_\Delta \cdot \min\{\varepsilon_g, \varepsilon_H\}\right)\right)/\log\gamma + |\mathcal{T}_{succ}|$. Combine the result from Lemma 5, we have the total iteration complexity as

$$T \leq \frac{1}{\log\gamma}\log\left(\frac{\Delta_0}{\kappa_\Delta \cdot \min\{\varepsilon_g, \varepsilon_H\}}\right) + \frac{2(F(\mathbf{x}_0) - F_{\min})}{\eta \min\{\widehat{\kappa}_\Delta, \widetilde{\kappa}_\Delta\}} \cdot \max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\}$$
$$\in \mathcal{O}\left(\max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\}\right),$$

where $\kappa_\Delta, \widehat{\kappa}_\Delta, \widetilde{\kappa}_\Delta$ are defined in the proofs of Lemmas 4 and 5, respectively. □

As it can be seen, the worst-case total number of iterations required by Algorithm 1 before termination, matches the optimal iteration complexity obtained in [13]. Furthermore, from (3a), it follows that upon termination of Algorithm 1 after T iterations, in addition to $\|\nabla F(\mathbf{x}_T)\| \leq \varepsilon_g$, we have $\lambda_{\min}\left(\nabla^2 F(\mathbf{x}_T)\right) \geq -(\varepsilon_H + \varepsilon_T)$, i.e., the obtained solution satisfies $(\varepsilon_g, \varepsilon_T + \varepsilon_H)$-Optimality as in (1).

For Algorithm 1, the Hessian approximation tolerance ε_t is allowed to be chosen per-iteration as $\varepsilon_t \leq \mathcal{O}(\max\{\varepsilon_H, \Delta_t\})$. This way, when Δ_t is large (e.g., at the beginning of iterations), one can employ crude Hessian approximations. As iterations progress towards optimality, Δ_t can get very small, in which case Hessian accuracy is set in the order of ε_H. Note that by Lemma 4, we are always guaranteed to have $\Delta_t \in \Omega\left(\min\left\{\varepsilon_g, \varepsilon_H\right\}\right)$. As a result, when $\varepsilon_g \ll \varepsilon_H$, e.g., $\varepsilon_H^2 = \varepsilon_g = \varepsilon$, we can have that $\Delta_t \ll \varepsilon_H$. In such cases, the choice $\varepsilon_t \leq \mathcal{O}(\max\{\varepsilon_H, \Delta_t\})$ ensures that the Hessian approximation tolerance never gets unnecessarily too small.

2.2 Adaptive cubic regularization with inexact Hessian

Similar to Sect. 2.1, in this section, we present the algorithm and its corresponding convergence results for the case of adaptive cubic regularization with inexact Hessian. In particular, Algorithm 2 depicts a variant of ARC algorithm where at each iteration t, the inexact approximation, $\mathbf{H}_t$, is constructed according to Condition 1. Here, unlike Sect. 2.1, we were unable to provide convergence guarantees with adaptive tolerance in (3a) and as result, ε is set fixed a priori to a sufficiently small value, i.e., $\varepsilon \in \mathcal{O}(\sqrt{\varepsilon_g}, \varepsilon_H)$ to guarantee $(\varepsilon_g, \varepsilon_H)$-optimality.

Similar to Algorithm 1, here we also require that the sub-problem (12) in Algorithm 2 is solved only approximately. Although similar inexact solutions to the sub-problem (12) by using Cauchy and Eigenpoint has been considered in several

Algorithm 2 Adaptive Cubic Regularization with Inexact Hessian

Input: Starting point $\mathbf{x}_0$, initial regularization $0 < \sigma_0 < \infty$, hyper-parameters $\varepsilon_g, \varepsilon_H, \eta \in (0, 1), \gamma > 1$

for $t = 0, 1, \ldots$ **do**
  Set the approximating Hessian, $\mathbf{H}_t$, as in (3)
  if $\|\nabla F(\mathbf{x}_t)\| \leq \varepsilon_g, \lambda_{\min}(\mathbf{H}_t) \geq -\varepsilon_H$ **then**
    Return $\mathbf{x}_t$.
  end if
  Solve the sub-problem approximately

$$\mathbf{s}_t \approx \arg\min_{\mathbf{s}\in\mathbb{R}^d} \; m_t(s) \triangleq \langle \nabla F(\mathbf{x}_t), \mathbf{s}\rangle + \frac{1}{2}\langle \mathbf{s}, \mathbf{H}_t\mathbf{s}\rangle + \frac{\sigma_t}{3}\|\mathbf{s}\|^3 \tag{12}$$

  Set $\rho_t \triangleq \dfrac{F(\mathbf{x}_t) - F(\mathbf{x}_t + \mathbf{s}_t)}{-m_t(\mathbf{s}_t)}$
  if $\rho_t \geq \eta$ **then**
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{s}_t$
    $\sigma_{t+1} = \sigma_t/\gamma$
  else
    $\mathbf{x}_{t+1} = \mathbf{x}_t$
    $\sigma_{t+1} = \gamma\sigma_t$
  end if
end for
Output: $\mathbf{x}_t$

previous work, e.g., [13], here we provide refined conditions which prove to be instrumental in obtaining iteration complexities with the relaxed Hessian approximation (3a), as opposed to the stronger Condition (5c).

Condition 3 (Sufficient descent Cauchy and Eigen directions) *Assume that we solve the sub-problem* (12) *approximately to find* $\mathbf{s}_t$ *such that*

$$-m_t(\mathbf{s}_t) \geq -m_t(\mathbf{s}_t^C) \geq \max\left\{\frac{1}{12}\|\mathbf{s}_t^C\|^2\left(\sqrt{K_H^2 + 4\sigma_t\|\nabla F(\mathbf{x}_t)\|} - K_H\right),\right.$$
$$\left.\frac{\|\nabla F(\mathbf{x}_t)\|}{2\sqrt{3}}\min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{K_H}, \sqrt{\frac{\|\nabla F(\mathbf{x}_t)\|}{\sigma_t}}\right\}\right\}, \tag{13a}$$

$$-m_t(\mathbf{s}_t) \geq -m_t(\mathbf{s}_t^E) \geq \frac{\nu|\lambda_{\min}(\mathbf{H}_t)|}{6}\max\left\{\|\mathbf{s}_t^E\|^2, \frac{\nu^2|\lambda_{\min}(\mathbf{H}_t)|^2}{\sigma_t^2}\right\}, \;\; \textit{if} \;\; \lambda_{\min}(\mathbf{H}_t) < 0. \tag{13b}$$

Here $m_t(\cdot)$ *is defined in* (12), $\mathbf{s}_t^C$ *(Cauchy point) is along negative gradient direction and* $\mathbf{s}_t^E$ *is along approximate negative curvature direction such that* $\langle \mathbf{s}_t^E, \mathbf{H}_t\mathbf{s}_t^E\rangle \leq \nu\lambda_{\min}(\mathbf{H}_t)\|\mathbf{s}_t^E\|^2 < 0$ *for some* $\nu \in (0, 1]$ *(see Appendix B for a way to efficiently compute* $\mathbf{s}_t^E$*).*

Note that Condition (13) describes the quality of the descent obtained by Cauchy and Eigen directions more accurately than is usually found in similar literature. A natural way to ensure that the approximate solution to the sub-problem (12) satisfies (13), is

by replacing the unconstrained high-dimensional sub-problem (12) with the following constrained but lower-dimensional problem, in which the search space is reduced to a two-dimensional sub-space containing vectors $\mathbf{s}_t^C$, and $\mathbf{s}_t^E$, i.e.,

$$\mathbf{s}_t = \arg \min_{\mathbf{s} \in \text{Span}\{\mathbf{s}_t^C, \mathbf{s}_t^E\}} \langle \nabla F(\mathbf{x}_t), \mathbf{s} \rangle + \frac{1}{2} \langle \mathbf{s}, \mathbf{H}_t \mathbf{s} \rangle + \frac{\sigma_t}{3} \|\mathbf{s}\|^3.$$

Note that, if $\mathbf{U} \in \mathbb{R}^{d \times p}$ is an orthogonal basis for the sub-space "Span$\{\mathbf{s}_t^C, \mathbf{s}_t^E\}$", by a linear transformation, we can turn the above sub-problem into an unconstrained problem as

$$\mathbf{v}_t = \arg \min_{\mathbf{v} \in \mathbb{R}^p} \langle U^T \nabla F(\mathbf{x}_t), \mathbf{v} \rangle + \frac{1}{2} \langle \mathbf{v}, \mathbf{U}^T \mathbf{H}_t \mathbf{U} \mathbf{v} \rangle + \frac{\sigma_t}{3} \|\mathbf{v}\|^3,$$

and set $\mathbf{s}_t = \mathbf{U}\mathbf{v}_t$. As before, any larger dimensional sub-space $\mathscr{P}$ for which we have Span$\{\mathbf{s}_t^C, \mathbf{s}_t^E\} \subseteq \mathscr{P}$ would also ensure (13), and, indeed, implies a more accurate solution to our original sub-problem (12).

Lemmas 6 and 7 describe the model reduction obtained by Cauchy and eigen points as required by Condition (3).

Lemma 6 (Descent with Cauchy direction) *Consider the Cauchy direction as* $\mathbf{s}_t^C = -\alpha \nabla F(\mathbf{x}_t)$ *where* $\alpha = \arg\min_{\widehat{\alpha} \geq 0} m_t(-\widehat{\alpha} \nabla F(\mathbf{x}_t))$. *We have*

$$-m_t(\mathbf{s}_t^C) \geq \max \left\{ \frac{1}{12} \|\mathbf{s}_t^C\|^2 \left(\sqrt{K_H^2 + 4\sigma_t \|\nabla F(\mathbf{x}_t)\|} - K_H \right), \right.$$
$$\left. \frac{\|\nabla F(\mathbf{x}_t)\|}{2\sqrt{3}} \min \left\{ \frac{\|\nabla F(\mathbf{x}_t)\|}{K_H}, \sqrt{\frac{\|\nabla F(\mathbf{x}_t)\|}{\sigma_t}} \right\} \right\}.$$

Proof For any $\widehat{\alpha} \geq 0$, we have $m_t(-\widehat{\alpha} \nabla F(\mathbf{x}_t)) \leq m_t(\widehat{\alpha} \nabla F(\mathbf{x}_t))$, which implies $\alpha = \arg\min_{\widehat{\alpha} \in \mathbb{R}} m_t(-\widehat{\alpha} \nabla F(\mathbf{x}_t))$. Hence, we have $-\|\nabla F(\mathbf{x}_t)\|^2 + \alpha \langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle + \sigma_t \alpha^2 \|\nabla F(\mathbf{x}_t)\|^3 = 0$. We can find explicit formula for such α by finding the roots of the quadratic function $r(\alpha) = \sigma_t \|\nabla F(\mathbf{x}_t)\|^3 \alpha^2 + \langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle \alpha - \|\nabla F(\mathbf{x}_t)\|^2$. Hence, we must have

$$\alpha = \frac{-\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle + \sqrt{\left(\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle \right)^2 + 4\sigma_t \|\nabla F(\mathbf{x}_t)\|^5}}{2\sigma_t \|\nabla F(\mathbf{x}_t)\|^3} \geq 0.$$

It follows that

$$2\alpha\sigma_t \|\nabla F(\mathbf{x}_t)\| = \sqrt{\left(\frac{\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle}{\|\nabla F(\mathbf{x}_t)\|^2} \right)^2 + 4\sigma_t \|\nabla F(\mathbf{x}_t)\|} - \frac{\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t) \rangle}{\|\nabla F(\mathbf{x}_t)\|^2}.$$

Consider the function $h(x;\beta) = \sqrt{x^2+\beta} - x$. It is easy to verify that, for $\beta \geq 0$, $h(x)$ is decreasing function of x. Now since $\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t)\rangle \leq K_H \|\nabla F(\mathbf{x}_t)\|^2$, we get

$$\|\mathbf{s}_t^C\| = \alpha \|\nabla F(\mathbf{x}_t)\| \geq \frac{1}{2\sigma_t}\left[\sqrt{K_H^2 + 4\sigma_t \|\nabla F(\mathbf{x}_t)\|} - K_H\right]. \tag{14}$$

Now, from [13, Lemma 2.1], we get

$$\begin{aligned} -m_t(s_t^C) &\geq \frac{\sigma_t \|\mathbf{s}_t^C\|^3}{6} = \frac{\|\mathbf{s}_t^C\|^2}{6}\alpha\sigma_t\|\nabla F(\mathbf{x}_t)\| \\ &\geq \frac{\|\mathbf{s}_t^C\|^2}{12}\left(\sqrt{K_H^2 + 4\sigma_t\|\nabla F(\mathbf{x}_t)\|} - K_H\right). \end{aligned}$$

Alternatively, following the proof of [10, Lemma 2.1], for any $\alpha \geq 0$, we get

$$\begin{aligned} m_t(s_t^C) &\leq m_t(-\alpha \nabla F(\mathbf{x}_t)) \\ &= -\alpha\|\nabla F(\mathbf{x}_t)\|^2 + \frac{1}{2}\alpha^2\langle \nabla F(\mathbf{x}_t), \mathbf{H}_t \nabla F(\mathbf{x}_t)\rangle + \frac{\alpha^3}{3}\sigma_t\|\nabla F(\mathbf{x}_t)\|^3 \\ &\leq \frac{\alpha\|\nabla F(\mathbf{x}_t)\|^2}{6}\left(-6 + 3\alpha K_H + 2\alpha^2\sigma_t\|\nabla F(\mathbf{x}_t)\|\right). \end{aligned}$$

Consider the quadratic polynomial $r(\alpha) = 2\alpha^2\sigma_t\|\nabla F(\mathbf{x}_t)\| + 3\alpha K_H - 6$. We have $r(\alpha) \leq 0$ for $\alpha \in [0, \bar{\alpha}]$, where

$$\bar{\alpha} = \frac{-3K_H + \sqrt{9K_H^2 + 48\sigma_t\|\nabla F(\mathbf{x}_t)\|}}{4\sigma_t\|\nabla F(\mathbf{x}_t)\|} = \frac{12}{\left(3K_H + \sqrt{9K_H^2 + 48\sigma_t\|\nabla F(\mathbf{x}_t)\|}\right)}.$$

Note that $\sqrt{9K_H^2 + 48\sigma_t\|\nabla F(\mathbf{x}_t)\|} \leq 8\sqrt{3}\max\left\{K_H, \sqrt{\sigma_t\|\nabla F(\mathbf{x}_t)\|}\right\}$ and trivially $3K_H \leq 4\sqrt{3}\max\left\{K_H, \sqrt{\sigma_t\|\nabla F(\mathbf{x}_t)\|}\right\}$. Hence, defining $\alpha_0 \triangleq 1/(\sqrt{3}\max\{K_H, \sqrt{\sigma_t\|\nabla F(\mathbf{x}_t)\|}\})$, it is easy to see that $0 < \alpha_0 \leq \bar{\alpha}$. With this α_0, we get $r(\alpha_0) \leq 2/9 + 3/\sqrt{3} - 6 \leq -3$. Therefore

$$\begin{aligned} m_t(s_t) &\leq \frac{-3\|\nabla F(\mathbf{x}_t)\|^2}{6\sqrt{3}\max\left\{K_H, \sqrt{\sigma_t\|\nabla F(\mathbf{x}_t)\|}\right\}} \\ &= \frac{-\|\nabla F(\mathbf{x}_t)\|}{2\sqrt{3}}\min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{K_H}, \sqrt{\frac{\|\nabla F(\mathbf{x}_t)\|}{\sigma_t}}\right\}. \end{aligned}$$

□

Lemma 7 (Descent with negative curvature) *Suppose* $\lambda_{\min}(\mathbf{H}_t) < 0$. *For some* $\nu \in (0, 1]$, *define* $\mathbf{s}_t^E = \alpha \mathbf{u}_t$, *where* $\alpha = \arg\min_{\widehat{\alpha} \in \mathbb{R}} m_t(\widehat{\alpha}\mathbf{u}_t)$, *and* $\langle \mathbf{u}_t, \mathbf{H}_t \mathbf{u}_t \rangle \leq \nu \lambda_{\min}(\mathbf{H}_t) \|\mathbf{u}_t\|^2 < 0$.

We have

$$-m_t(\mathbf{s}_t^E) \geq \frac{\nu |\lambda_{\min}(\mathbf{H}_t)|}{6} \max \left\{ \|\mathbf{s}_t^E\|^2, \frac{\nu^2 |\lambda_{\min}(\mathbf{H}_t)|^2}{\sigma_t^2} \right\}.$$

Proof By the first-order necessary optimality condition of α, we get $\langle \nabla F(\mathbf{x}_t), \mathbf{u}_t \rangle + \alpha \langle \mathbf{u}_t, \mathbf{H}_t \mathbf{u}_t \rangle + \sigma_t \alpha^2 \|\mathbf{u}_t\|^3 = 0$, which implies $\langle \nabla F(\mathbf{x}_t), \mathbf{s}_t^E \rangle + \langle \mathbf{s}_t^E, \mathbf{H}_t \mathbf{s}_t^E \rangle + \sigma_t \|\mathbf{s}_t^E\|^3 = 0$. Next, since α is a minimizer of $m_t(\widehat{\alpha}\mathbf{u}_t)$, we have $m_t(\alpha \mathbf{u}_t) \leq m_t(-\alpha \mathbf{u}_t)$, which implies $\langle \nabla F(\mathbf{x}_t), \mathbf{s}_t^E \rangle \leq 0$. Hence, we also obtain $\langle \mathbf{s}_t^E, \mathbf{H}_t \mathbf{s}_t^E \rangle + \sigma_t \|\mathbf{s}_t^E\|^3 \geq 0$. From [13, Lemma 2.1], we get $-m_t(s_t^E) \geq \sigma_t \|\mathbf{s}_t^E\|^3/6 = \left(-\langle \nabla F(\mathbf{x}_t), \mathbf{s}_t^E \rangle - \langle \mathbf{s}_t^E, \mathbf{H}_t \mathbf{s}_t^E \rangle\right)/6 \geq \nu |\lambda_{\min}(\mathbf{H}_t)| \|\mathbf{s}_t^E\|^2/6$. Now, we have

$$\sigma_t \|\mathbf{s}_t^E\| \geq -\frac{\langle \mathbf{s}_t^E, \mathbf{H}_t \mathbf{s}_t^E \rangle}{\|\mathbf{s}_t^E\|^2} \geq \nu |\lambda_{\min}(\mathbf{H}_t)|, \tag{15}$$

which gives $\sigma_t \|\mathbf{s}_t^E\|^3 \geq \nu |\lambda_{\min}(\mathbf{H}_t)| \|\mathbf{s}_t^E\|^2$ and $\sigma_t \|\mathbf{s}_t^E\|^3 \geq \nu^3 \sigma_t^{-2} |\lambda_{\min}(\mathbf{H}_t)|^3$. Hence, we have $-m_t(\mathbf{s}_t^E) \geq \sigma_t \|\mathbf{s}_t^E\|^3/6 \geq \nu |\lambda_{\min}(\mathbf{H}_t)| \|\mathbf{s}_t^E\|^2/6$ and $-m_t(\mathbf{s}_t^E) \geq \sigma_t \|\mathbf{s}_t^E\|^3/6 \geq \nu^3 \sigma_t^{-2} |\lambda_{\min}(\mathbf{H}_t)|^3/6$. □

The next lemma is used to show sufficient decrease in the objective function using the approximate solution of the sub-problem (12).

Lemma 8 *Given Assumption 1 and Condition 1, we have*

$$F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t) \leq \left(\frac{L}{2} - \frac{\sigma_t}{3} \right) \|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2} \|\mathbf{s}_t\|^2.$$

Proof Apply Mean Value Theorem on F at $\mathbf{x}_t$ gives $F(\mathbf{x}_t + \mathbf{s}_t) = F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t)^T \mathbf{s}_t + \frac{1}{2} \mathbf{s}_t^T \nabla^2 F(\xi_t) \mathbf{s}_t$, for some ξ_t in the segment of $[\mathbf{x}_t, \mathbf{x}_t + \mathbf{s}_t]$. Now, it follows that

$$\begin{aligned}
F(\mathbf{x}_t + \mathbf{s}_t) - F(\mathbf{x}_t) - m_t(\mathbf{s}_t) &= \frac{1}{2} \mathbf{s}_t^T (\nabla^2 F(\xi_t) - \mathbf{H}_t) \mathbf{s}_t - \frac{\sigma_t}{3} \|\mathbf{s}_t\|^3 \\
&= \frac{1}{2} \mathbf{s}_t^T (\nabla^2 F(\xi_t) - \nabla^2 F(\mathbf{x}_t) + \nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t) \mathbf{s}_t - \frac{\sigma_t}{3} \|\mathbf{s}_t\|^3 \\
&\leq \frac{1}{2} \mathbf{s}_t^T (\nabla^2 F(\xi_t) - \nabla^2 F(\mathbf{x}_t)) \mathbf{s}_t + \frac{1}{2} \mathbf{s}_t^T (\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t) \mathbf{s}_t - \frac{\sigma_t}{3} \|\mathbf{s}_t\|^3 \\
&\leq \frac{L}{2} \|\mathbf{s}_t\|^3 + \frac{1}{2} \varepsilon \|\mathbf{s}_t\|^2 - \frac{\sigma_t}{3} \|\mathbf{s}_t\|^3 \leq \left(\frac{L}{2} - \frac{\sigma_t}{3} \right) \|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2} \|\mathbf{s}_t\|^2.
\end{aligned}$$

□

Lemma 9 *Given Assumption 1, Conditions 1 and 3, suppose*

$$\sigma_t \geq 2L, \quad \varepsilon \leq \min\left\{\frac{1}{12}\left(\sqrt{K_H^2 + 8L\varepsilon_g} - K_H\right), \frac{\nu\varepsilon_H}{6\gamma}\right\}.$$

Then, we have

$$\left(\frac{L}{2} - \frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \begin{cases} \frac{\varepsilon}{2}\|\mathbf{s}_t^C\|^2, \\ \frac{\varepsilon}{2}\|\mathbf{s}_t^E\|^2, \quad \textit{If } \lambda_{\min}(\mathbf{H}_t) \geq -\varepsilon_H \end{cases}.$$

Proof First consider $\|\mathbf{s}_t^C\|$ for which we have two cases.

(i) If $\|\mathbf{s}_t\| \leq \|\mathbf{s}_t^C\|$, then from assumption on σ_t, it immediately follows that

$$\left(\frac{L}{2} - \frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \frac{\varepsilon}{2}\|\mathbf{s}_t^C\|^2.$$

(ii) If $\|\mathbf{s}_t\| \geq \|\mathbf{s}_t^C\|$, since $L \leq \sigma_t/2$, then

$$\begin{aligned}\left(\frac{L}{2} - \frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 &\leq -\frac{\sigma_t}{12}\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \left(-\frac{\sigma_t}{12}\left\|\mathbf{s}_t^C\right\| + \frac{\varepsilon}{2}\right)\|\mathbf{s}_t\|^2 \\ &\leq \left(-\frac{\sqrt{K_H^2 + 8L\varepsilon_g} - K_H}{24} + \frac{\varepsilon}{2}\right)\|\mathbf{s}_t\|^2 \leq 0 \leq \frac{\varepsilon}{2}\left\|\mathbf{s}_t^C\right\|^2.\end{aligned}$$

The second last inequality follows from (14).

Similarly, for $\|\mathbf{s}_t^E\|$, we have two cases.

i. If $\|\mathbf{s}_t\| \leq \|\mathbf{s}_t^E\|$, then from assumption on σ_t, it immediately follows that

$$\left(\frac{L}{2} - \frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq \frac{\varepsilon}{2}\|\mathbf{s}_t^E\|^2.$$

ii. If $\|\mathbf{s}_t\| \geq \|\mathbf{s}_t^E\|$, since $L \leq \sigma_t/2$, then

$$\begin{aligned}\left(\frac{L}{2} - \frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 &\leq -\frac{\sigma_t}{12}\|\mathbf{s}_t\|^3 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq -\frac{\sigma_t}{12}\left\|\mathbf{s}_t^E\right\|\|\mathbf{s}_t\|^2 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \\ &\leq -\frac{\nu\varepsilon_H}{12}\|\mathbf{s}_t\|^2 + \frac{\varepsilon}{2}\|\mathbf{s}_t\|^2 \leq 0 < \frac{\varepsilon}{2}\left\|\mathbf{s}_t^E\right\|^2.\end{aligned}$$

The second last inequality follows from (15) and the last line follows from $\varepsilon \leq \frac{\nu\varepsilon_H}{6}$.

□

Lemma 10 *Given Assumption 1, Conditions 1 and 3, suppose at the t-th iteration,* $\lambda_{\min}(\mathbf{H}_t) < -\varepsilon_H$, $\sigma_t \geq 2L$, *and* $\varepsilon \leq \min\{1/6, (1-\eta)/3\}\nu\varepsilon_H$. *Then, the t-th iteration is successful, i.e.* $\sigma_{t+1} = \sigma_t/\gamma$.

Proof From (13b), Lemmas 8 and 9, as well as assumptions on σ_t and ε, we have

$$1-\rho_t = \frac{F(\mathbf{x}_t+\mathbf{s}_t)-F(\mathbf{x}_t)-m_t(\mathbf{s}_t)}{-m_t(\mathbf{s}_t)} \leq \frac{(L/2-\sigma_t/3)\,\|\mathbf{s}_t\|^3+\varepsilon\|\mathbf{s}_t\|^2/2}{\nu|\lambda_{\min}(\mathbf{H}_t)|\|\mathbf{s}_t^E\|^2/6}$$
$$\leq \frac{3\varepsilon\|\mathbf{s}_t^E\|^2}{\nu|\lambda_{\min}(\mathbf{H}_t)|\|\mathbf{s}_t^E\|^2} \leq \frac{3\varepsilon}{\nu\varepsilon_H} \leq 1-\eta.$$

Hence, $\rho_t \geq \eta$, and the iteration is successful. □

Lemma 11 *Given Assumption 1, Conditions 1 and 3, suppose at the t-th iteration,* $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$, $\sigma_t \geq 2L$, *and*

$$\varepsilon \leq \min\left\{\frac{1}{12}, \frac{1-\eta}{6}\right\}\left(\sqrt{K_H^2+8L\varepsilon_g}-K_H\right).$$

Then, the t-th iteration is successful, i.e. $\sigma_{t+1}=\sigma_t/\gamma$.

Proof First note that, from (13a), we have

$$-m_t(\mathbf{s}_t) \geq -m_t(\mathbf{s}_t^C) \geq \frac{1}{12}\|\mathbf{s}_t^C\|^2\left(\sqrt{K_H^2+4\sigma_t\|\nabla F(\mathbf{x}_t)\|}-K_H\right).$$

Hence, again, by (13a), Lemmas 8 and 9 , it follows that

$$1-\rho_t = \frac{F(\mathbf{x}_t+\mathbf{s}_t)-F(\mathbf{x}_t)-m_t(\mathbf{s}_t)}{-m_t(\mathbf{s}_t)} \leq \frac{\left(\frac{L}{2}-\frac{\sigma_t}{3}\right)\|\mathbf{s}_t\|^3+\frac{\varepsilon}{2}\|\mathbf{s}_t\|^2}{-m_t(\mathbf{s}_t^C)}$$
$$\leq \frac{\frac{\varepsilon}{2}\|\mathbf{s}_t^C\|^2}{\frac{1}{12}\|\mathbf{s}_t^C\|^2\left(\sqrt{K_H^2+4\sigma_t\|\nabla F(\mathbf{x}_t)\|}-K_H\right)}$$
$$\leq \frac{6\varepsilon}{\left(\sqrt{K_H^2+4\sigma_t\|\nabla F(\mathbf{x}_t)\|}-K_H\right)}$$
$$\leq \frac{6\varepsilon}{\left(\sqrt{K_H^2+8L\varepsilon_g}-K_H\right)} \leq 1-\eta.$$

Hence, $\rho_t \geq \eta$, and the iteration is successful. □

Now we can upper bound the cubic regularization parameter before the algorithm terminates, as in Lemma 12.

Lemma 12 *Consider Assumption 1, Conditions 1 and 3, and*

$$\varepsilon \leq \min\left\{\min\left\{\frac{1}{12}, \frac{1-\eta}{6}\right\}\left(\sqrt{K_H^2+8L\varepsilon_g}-K_H\right), \min\left\{\frac{1}{6}, \frac{1-\eta}{3}\right\}\nu\varepsilon_H\right\}, \tag{16}$$

where ν, L, K_H are, respectively, defined as in (13b), (6a), (3b), and η is a hyper-parameter of Algorithm 2. For Algorithm 2 we have for all t, $\sigma_t \leq \max\{\sigma_0, 2\gamma L\}$.

Proof We prove by contradiction. Assume the t-th iteration is the first unsuccessful iteration such that $\sigma_{t+1} = \gamma\sigma_t \geq 2\gamma L$, which implies that $\sigma_t \geq 2L$. However, according to Lemmas 10 and 11 , respectively, if $\lambda_{\min}(H_t) < -\varepsilon_H$ or $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$, then the iteration is successful and hence we must have $\sigma_{t+1} = \sigma_t/\gamma \leq \sigma_t$, which is a contradiction. □

Now, similar to [13, Lemma 2.8], we can get the following result about the estimate of the total number of successful iterations before algorithm terminates.

Lemma 13 (Success iterations) *Given Assumption 1, Conditions 1 and 3, let $\mathcal{T}_{succ}$ denote the set of all the successful iterations before Algorithm 2 stops. The number of successful iterations is upper bounded by,*

$$|\mathcal{T}_{succ}| \leq \frac{(F(\mathbf{x}_0) - F_{\min})}{\eta\kappa_\sigma} \cdot \max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\},$$

where $\kappa_\sigma \triangleq \min\left\{\nu^3/(24\gamma^2L^2), \min\left\{1/K_H, \sqrt{1/(2\gamma L)}\right\}/(2\sqrt{3})\right\}$.

Proof Suppose Algorithm 2 doesn't terminate at the t-th iteration. Then either we have $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$ or $\lambda_{\min}(\nabla^2\mathbf{H}_t) \leq -\varepsilon_H$. In the first case, (13a) and Lemma 12 gives

$$-m_t(\mathbf{s}_t) \geq \frac{\|\nabla F(\mathbf{x}_t)\|}{2\sqrt{3}} \min\left\{\frac{\|\nabla F(\mathbf{x}_t)\|}{K_H}, \sqrt{\frac{\|\nabla F(\mathbf{x}_t)\|}{\sigma_t}}\right\} \geq \frac{\varepsilon_g^2}{2\sqrt{3}} \min\left\{\frac{1}{K_H}, \sqrt{\frac{1}{2\gamma L}}\right\}.$$

Similarly, in the case where $\lambda_{\min}(\nabla^2\mathbf{H}_t) \leq -\varepsilon_H$, from (13b) and Lemma 12, we obtain $-m_t(\mathbf{s}_t) \geq \nu^3|\lambda_{\min}(\mathbf{H}_t)|^3/(6\sigma_t^2) \geq \nu^3\varepsilon_H^3/(24\gamma^2L^2)$.

Since $F(\mathbf{x}_t)$ is monotonically decreasing, we have

$$\begin{aligned} F(\mathbf{x}_0) - F_{\min} &\geq \sum_{t=0}^{\infty} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \geq \sum_{t\in\mathcal{T}_{succ}} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \geq -\eta \sum_{t\in\mathcal{T}_{succ}} m_t(\mathbf{s}_t) \\ &\geq \eta\,|\mathcal{T}_{succ}| \min\left\{\frac{\nu^3\varepsilon_H^3}{24\gamma^2L^2}, \frac{\varepsilon_g^2}{2\sqrt{3}} \min\left\{\frac{1}{K_H}, \sqrt{\frac{1}{2\gamma L}}\right\}\right\} \\ &\geq |\mathcal{T}_{succ}|\,\eta\kappa_\sigma \min\{\varepsilon_g^2, \varepsilon_H^3\}. \end{aligned}$$

□

Now we show the final complexity bounds of Algorithm 2 in Theorem 2.

Theorem 2 (Complexity of Algorithm 2) *Consider any $0 < \varepsilon_g, \varepsilon_H < 1$. Suppose the inexact Hessian, $\mathbf{H}(\mathbf{x})$, satisfies Condition 1 with the approximation tolerance, ε, in (3a) as (16). For Problem (**P0**), under Assumption 1 and Condition 3, Algorithm 2 terminates after at most $T \in \mathcal{O}\left(\max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\}\right)$ iterations.*

Proof Suppose Algorithm 2 terminates at the t-th iteration. Let $\mathcal{T}_{succ}$ and $\mathcal{T}_{fail}$ denote the sets of all the successful and unsuccessful iterations, respectively. Then $T = |\mathcal{T}_{succ}| + |\mathcal{T}_{fail}|$ and $\sigma_T = \sigma_0 \gamma^{|\mathcal{T}_{fail}| - |\mathcal{T}_{succ}|}$. From Lemma 12, we have $\sigma_T \leq 2\gamma L$. Hence, $|\mathcal{T}_{fail}| \leq \log(2\gamma L/\sigma_0)/\log\gamma + |\mathcal{T}_{succ}|$, which, using Lemma 13 gives the total iteration complexity as

$$T \leq \log(2\gamma L/\sigma_0)/\log\gamma + 2(F(\mathbf{x}_0) - F_{\min}) \cdot \max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\}/(\eta\kappa_\sigma),$$

where κ_σ is defined in Lemma 13. □

In Theorem 2 (as well as Theorem 3 below), we require $\varepsilon \in \mathcal{O}(\sqrt{\varepsilon_g}, \varepsilon_H)$. This can be rather strict and computationally unattractive, unless either crude solutions are required (e.g., in most machine learning applications very rough solutions are encouraged to avoid over-fitting), or the inexact Hessian is formed from a sub-set of data that is significantly smaller than the original dataset (e.g., see Sect. 3 in the context of big-data regimes where $n \gg 1$ and $|\mathcal{S}| \ll n$). Nonetheless, the theoretical existence of such tolerance, though small, implies a certain level of robustness of the algorithm, i.e., the complexity of the algorithm is not adversely affected by small errors in Hessian computations.

We note that, for iterations where $\varepsilon \ll \|\mathbf{s}_t\|$, (3a) is indeed a more stringent condition than (5c). As iterations progress towards optimality, step-size can become small, in which case (3a) might be theoretically more preferable. Nonetheless, beyond a direct theoretical comparison among various Hessian approximation bounds in terms of their tightness, the main advantage of (3a) should be regarded in light of its simplicity, which allows for direct constructions of $\mathbf{H}_t$ with a priori guarantees.

Condition 3 seems to be the bare minimum required to guarantee convergence to an approximate second-order criticality. Intuitively, however, if an approximate solution to the sub-problem (12) satisfies more than (13), i.e., if we solve (12) more exactly than just requiring (13), one could expect to be able to improve upon the iteration complexity of Theorem 2. Indeed, suppose we solve the reduced sub-problem on progressively embedded sub-spaces with increasingly higher dimensions, all of which including "Span$\{\mathbf{s}_t^C, \mathbf{s}_t^E\}$", and stop when the corresponding solution $\mathbf{s}_t$ satisfies the following conditions.

Condition 4 (Sufficient descent for optimal complexity) *Assume that we solve the sub-problem* (12) *approximately to find* $\mathbf{s}_t$ *such that, in addition to* (13)*, we have*

$$\|\nabla m_t(\mathbf{s}_t)\| \leq \zeta \max\left\{\|\mathbf{s}_t\|^2, \theta_t\|\nabla F(\mathbf{x}_t)\|\right\}, \quad \theta_t \triangleq \min\{1, \|\mathbf{s}_t\|\}, \tag{17}$$

for some prescribed $\zeta \in (0, 1)$. *Here,* $m_t(\cdot)$ *is defined in* (12).

Conditions on the inexactness of the sub-problems were initially pioneered in [10,11,13]. However, the main drawback for these conditions is that the inexactness tolerance is closely tied with the magnitude of the gradient. More specifically, when gradient is small, e.g., near saddle points, the sub-problems are required to be solved exceedingly more accurately. In fact, at a saddle point where $\|\nabla F(\mathbf{x}_t)\| = 0$, these

conditions imply an exact solution to the sub-problem. To the best of our knowledge, Condition 4 represents a novel criterion, which offers the best of both worlds: when gradient is large, we allow for crude solutions to the sub-problem, but near saddle-points where the gradient is small, inexactness will be determined by the step length, which can be significantly larger than the gradient. Using Condition 4, we can obtain the optimal iteration complexity for Algorithm 2, as shown in Theorem 3. First, we prove the following two lemmas which will be used later for the proof of Theorem 3.

Lemma 14 *Suppose* $\|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g$. *Given Assumption 1 and Condition 3, let* (3a) *hold with* $\varepsilon_t = \min\{\varepsilon, \zeta\|\nabla F(\mathbf{x}_t)\|\}$ *where* ε *is as in* (16) *and* $\zeta \in (0, 1/2)$. *Furthermore, suppose* (12) *is solved such that Condition 4 eventually holds. Then, we have* $\|\mathbf{s}_t\| \geq \kappa_g \sqrt{\|\nabla F(\mathbf{x}_{t+1})\|}$, *where*

$$\kappa_g \triangleq \frac{2(1-2\zeta)}{((1+4\gamma)L + 2\max\{(\varepsilon + \zeta\max\{1, K\}), 2\zeta\max\{1, K\}\})}.$$

Proof First, suppose $\|\mathbf{s}_t\|^2 \leq \theta_t\|\nabla F(\mathbf{x}_t)\|$. Using Condition 4, we get $\|\nabla F(\mathbf{x}_{t+1})\| \leq \|\nabla F(\mathbf{x}_{t+1}) - \nabla m_t(\mathbf{s}_t)\| + \|\nabla m_t(\mathbf{s}_t)\| \leq \|\nabla F(\mathbf{x}_{t+1}) - \nabla m_t(\mathbf{s}_t)\| + \theta_t\|\nabla F(\mathbf{x}_t)\|$. Noting that $\nabla m_t(\mathbf{s}_t) = \nabla F(\mathbf{x}_t) + \mathbf{H}_t\mathbf{s}_t + \sigma_t\|\mathbf{s}_t\|\mathbf{s}_t$, and using Mean Value Theorem for vector-valued functions, (6a) and (3a), we get

$$\begin{aligned}
\|\nabla F(\mathbf{x}_{t+1}) - \nabla m_t(\mathbf{s}_t)\| &\leq \left\| \int_0^1 \nabla^2 F(\mathbf{x}_t + \tau\mathbf{s}_t)\mathbf{s}_t d\tau - \mathbf{H}_t\mathbf{s}_t \right\| + \sigma_t\|\mathbf{s}_t\|^2 \\
&\leq \left\| \int_0^1 \left(\nabla^2 F(\mathbf{x}_t + \tau\mathbf{s}_t) - \nabla^2 F(\mathbf{x}_t)\right)\mathbf{s}_t d\tau + \left(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\right)\mathbf{s}_t \right\| + \sigma_t\|\mathbf{s}_t\|^2 \\
&\leq \left\|\mathbf{s}_t\| \int_0^1 \|\nabla^2 F(\mathbf{x}_t + \tau\mathbf{s}_t) - \nabla^2 F(\mathbf{x}_t)\| d\tau + \| \left(\nabla^2 F(\mathbf{x}_t) - \mathbf{H}_t\right)\mathbf{s}_t \right\| + \sigma_t\|\mathbf{s}_t\|^2 \\
&\leq L\|\mathbf{s}_t\|^2 \int_0^1 \tau d\tau + \varepsilon_t\|\mathbf{s}_t\| + \sigma_t\|\mathbf{s}_t\|^2 \leq \left(\frac{L}{2} + 2\gamma L\right)\|\mathbf{s}_t\|^2 + \varepsilon_t\|\mathbf{s}_t\|,
\end{aligned}$$

where the last equality follows from Lemma 12. From (6b), it follows that

$$\|\nabla F(\mathbf{x}_t)\| \leq K\|\mathbf{s}_t\| + \|\nabla F(\mathbf{x}_{t+1})\|. \tag{18}$$

As such, using $\theta_t \leq \zeta$ from Condition 4 as well as the assumption on ε_t, we get

$$\begin{aligned}
\|\nabla F(\mathbf{x}_{t+1})\| &\leq \left(\frac{L}{2} + 2\gamma L\right)\|\mathbf{s}_t\|^2 + \varepsilon_t\|\mathbf{s}_t\| + \theta_t K\|\mathbf{s}_t\| + \theta_t\|\nabla F(\mathbf{x}_{t+1})\| \\
&\leq \left(\frac{L}{2} + 2\gamma L\right)\|\mathbf{s}_t\|^2 + \varepsilon_t\|\mathbf{s}_t\| + \theta_t K\|\mathbf{s}_t\| + \zeta\|\nabla F(\mathbf{x}_{t+1})\|,
\end{aligned}$$

which implies that $(1-\zeta)\|\nabla F(\mathbf{x}_{t+1})\| \leq (L/2 + 2\gamma L)\|\mathbf{s}_t\|^2 + (\varepsilon_t + \theta_t K)\|\mathbf{s}_t\|$. Now using Condition 4, we consider two cases:

(i) If $\|\mathbf{s}_t\| \geq 1$, then we get $(\varepsilon_t + \theta_t K)\|\mathbf{s}_t\| \leq (\varepsilon_t + \theta_t K)\|\mathbf{s}_t\|^2 \leq (\varepsilon + \zeta K)\|\mathbf{s}_t\|^2$. Hence, it follows that $(1-\zeta)\|\nabla F(\mathbf{x}_{t+1})\| \leq (L/2 + 2\gamma L + (\varepsilon + \zeta K))\|\mathbf{s}_t\|^2$.

(ii) If $\|\mathbf{s}_t\| \leq 1$, then from assumption on ε_t and (18), we have $\varepsilon_t\|\mathbf{s}_t\| \leq \zeta\|\nabla F(\mathbf{x}_t)\|\|\mathbf{s}_t\| \leq \zeta(K\|\mathbf{s}_t\|^2 + \|\nabla F(\mathbf{x}_{t+1})\|\|\mathbf{s}_t\|) \leq \zeta(K\|\mathbf{s}_t\|^2 + \|\nabla F(\mathbf{x}_{t+1})\|)$. Now by assumption on θ_t, we get $(\varepsilon_t + \theta_t K)\|\mathbf{s}_t\| = \varepsilon_t\|\mathbf{s}_t\| + \theta_t K\|\mathbf{s}_t\| \leq 2\zeta K\|\mathbf{s}_t\|^2 + \zeta\|\nabla F(\mathbf{x}_{t+1})\|$, which, in turn, implies that $(1-2\zeta)\|\nabla F(\mathbf{x}_{t+1})\| \leq (L/2 + 2\gamma L + 2\zeta K)\|\mathbf{s}_t\|^2$.

Now suppose, $\|\mathbf{s}_t\|^2 \geq \theta_t\|\nabla F(\mathbf{x}_t)\|$. As above, we have $\|\nabla F(\mathbf{x}_{t+1})\| \leq \|\nabla F(\mathbf{x}_{t+1}) - \nabla m_t(\mathbf{s}_t)\| + \|\nabla m_t(\mathbf{s}_t)\| \leq (L/2 + 2\gamma L + \zeta)\|\mathbf{s}_t\|^2 + \varepsilon_t\|\mathbf{s}_t\|$. If $\|\mathbf{s}_t\| \geq 1$, we have $\varepsilon_t\|\mathbf{s}_t\| \leq \varepsilon\|\mathbf{s}_t\|^2$, which gives $\|\nabla F(\mathbf{x}_{t+1})\| \leq (L/2 + 2\gamma L + \zeta + \varepsilon)\|\mathbf{s}_t\|^2$. Otherwise, if $\|\mathbf{s}_t\| \leq 1$, then $\|\mathbf{s}_t\|^2 \geq \theta_t\|\nabla F(\mathbf{x}_t)\|$ implies that $\|\mathbf{s}_t\| \geq \|\nabla F(\mathbf{x}_t)\|$. From assumption on ε_t, it follows that $\varepsilon_t\|\mathbf{s}_t\| \leq \zeta\|\nabla F(\mathbf{x}_t)\|\|\mathbf{s}_t\| \leq \zeta\|\mathbf{s}_t\|^2$, which in turn gives $\|\nabla F(\mathbf{x}_{t+1})\| \leq (L/2 + 2\gamma L + 2\zeta)\|\mathbf{s}_t\|^2$. □

Lemma 15 (Success iterations: optimal case) *Let*

$$\mathscr{T}_{succ} \triangleq \{t;\ \|\nabla F(\mathbf{x}_t)\| \geq \varepsilon_g\ \vee\ \lambda_{\min}(\mathbf{H}_t) \leq -\varepsilon_H\},$$

be the set of all successful iterations, before Algorithm 2 terminates. Under the conditions of Lemma 14, we must have $|\mathscr{T}_{succ}| \in \mathscr{O}(\max\{\varepsilon_H^{-3}, \varepsilon_g^{-3/2}\})$.

Proof From (13b) and Lemma 12, if $\lambda_{\min}(\nabla^2\mathbf{H}_t) \leq -\varepsilon_H$, it follows that $-m_t(\mathbf{s}_t) \geq \nu^3|\lambda_{\min}(\mathbf{H}_t)|^3/(6\sigma_t^2) \geq \nu^3\varepsilon_H^3/(24\gamma^2L^2)$. Note that $\mathscr{T}_{\text{succ}} = \mathscr{T}_{\text{succ}}^1 \bigcup \mathscr{T}_{\text{succ}}^2 \bigcup \mathscr{T}_{\text{succ}}^3$, where

$$\begin{aligned}
\mathscr{T}_{\text{succ}}^1 &\triangleq \left\{t \in \mathscr{T}_{\text{succ}};\ \|\nabla F(\mathbf{x}_{t+1})\| \geq \varepsilon_g\right\},\\
\mathscr{T}_{\text{succ}}^2 &\triangleq \left\{t \in \mathscr{T}_{\text{succ}};\ \|\nabla F(\mathbf{x}_{t+1})\| \leq \varepsilon_g \ \text{ and } \ \lambda_{\min}(H_{t+1}) \leq -\varepsilon_H\right\}\\
\mathscr{T}_{\text{succ}}^3 &\triangleq \left\{t \in \mathscr{T}_{\text{succ}};\ \|\nabla F(\mathbf{x}_{t+1})\| \leq \varepsilon_g \ \text{ and } \ \lambda_{\min}(H_{t+1}) \geq -\varepsilon_H\right\}.
\end{aligned}$$

We bound each of these sets individually. Since $F(\mathbf{x}_t)$ is monotonically decreasing, from [10, Lemma 3.3], $\sigma_t \geq \sigma_{\min}$, and Lemmas 12 and 14 , we have

$$\begin{aligned}
F(\mathbf{x}_0) - F_{\min} &\geq \sum_{t=0}^{\infty} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \geq \sum_{t\in\mathscr{T}_{\text{succ}}^1} F(\mathbf{x}_t) - F(\mathbf{x}_{t+1}) \geq -\eta \sum_{t\in\mathscr{T}_{\text{succ}}^1} m_t(\mathbf{s}_t)\\
&\geq \eta \sum_{t\in\mathscr{T}_{\text{succ}}^1} \min\left\{\frac{\nu^3\varepsilon_H^3}{24\gamma^2L^2}, \frac{\sigma_{\min}}{6}\|\mathbf{s}_t\|^3\right\}\\
&\geq \eta \sum_{t\in\mathscr{T}_{\text{succ}}^1} \min\left\{\frac{\nu^3\varepsilon_H^3}{24\gamma^2L^2}, \frac{\sigma_{\min}\kappa_g^3}{6}\|\nabla F(\mathbf{x}_{t+1})\|^{3/2}\right\}\\
&\geq \eta \sum_{t\in\mathscr{T}_{\text{succ}}^1} \min\left\{\frac{\nu^3\varepsilon_H^3}{24\gamma^2L^2}, \frac{\sigma_{\min}\kappa_g^3}{6}\varepsilon_g^{3/2}\right\}
\end{aligned}$$

$$\geq \eta \sum_{t \in \mathscr{T}^1_{\text{succ}}} \min\left\{\frac{\nu^3}{24\gamma^2 L^2}, \frac{\sigma_{\min}\kappa_g^3}{6}\right\} \min\{\varepsilon_H^3, \varepsilon_g^{3/2}\}.$$

Hence, $\left|\mathscr{T}^1_{\text{succ}}\right| \leq \kappa^1_{\mathscr{T}_{\text{succ}}} \max\{\varepsilon_H^{-3}, \varepsilon_g^{-3/2}\}$, where

$$\kappa^1_{\mathscr{T}_{\text{succ}}} \triangleq (F(\mathbf{x}_0) - F_{\min}) \max\{24\gamma^2 L^2/\nu^3, 6/(\sigma_{\min}\kappa_g^3)\}/\eta.$$

As for $\mathscr{T}^2_{\text{succ}}$, we have

$$\begin{aligned} F(\mathbf{x}_0) - F_{\min} &\geq F(\mathbf{x}_0) - F(\mathbf{x}_1) + \sum_{t=0}^{\infty} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_{t+2}) \\ &\geq F(\mathbf{x}_0) - F(\mathbf{x}_1) + \sum_{t \in \mathscr{T}^2_{\text{succ}}} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_{t+2}) \\ &\geq F(\mathbf{x}_0) - F(\mathbf{x}_1) - \eta \sum_{t \in \mathscr{T}^2_{\text{succ}}} m_{t+1}(\mathbf{s}_{t+1}) \\ &\geq F(\mathbf{x}_0) - F(\mathbf{x}_1) + \eta \sum_{t \in \mathscr{T}^2_{\text{succ}}} \frac{\nu^3 \varepsilon_H^3}{24\gamma^2 L^2}. \end{aligned}$$

Hence, $\left|\mathscr{T}^2_{\text{succ}}\right| \leq \kappa^2_{\mathscr{T}_{\text{succ}}} \varepsilon_H^{-3}$, where $\kappa^2_{\mathscr{T}_{\text{succ}}} \triangleq (F(\mathbf{x}_1) - F_{\min}) 24\gamma^2 L^2/(\eta\nu^3)$. Finally, we have $\left|\mathscr{T}^3_{\text{succ}}\right| = 1$, because in such a case, the algorithm stops in one iteration. Putting these bounds all together, we get $|\mathscr{T}_{\text{succ}}| \leq \max\{1, \kappa^1_{\mathscr{T}_{\text{succ}}}, \kappa^2_{\mathscr{T}_{\text{succ}}}\} \max\{\varepsilon_H^{-3}, \varepsilon_g^{-3/2}\}$. □

Now we can obtain the optimal complexity bound of Algorithm 2 in Theorem 3. The proof follows similarly as that of Theorem 2, and hence is omitted here.

Theorem 3 (Optimal complexity of Algorithm 2) *Consider any* $0 < \varepsilon_g, \varepsilon_H < 1$. *Suppose the inexact Hessian,* $\mathbf{H}(\mathbf{x})$, *satisfies Conditions (3) with the approximation tolerance,* ε, *in* (3a) *as* $\varepsilon = \min\{\varepsilon_0, \zeta\varepsilon_g\}$ *where* ε_0 *is as in* (16), *and* $\zeta \in (0, 1/2)$. *For Problem* (**P0**) *and under Assumption 1, if the approximate solution to the sub-problem* (12) *satisfies Conditions 3 and 4, then Algorithm 2 terminates after at most* $T \in \mathscr{O}\left(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\}\right)$ *iterations.*

From (3a), upon termination of Algorithm 2, the obtained solution satisfies $(\varepsilon_g, \varepsilon + \varepsilon_H)$-Optimality as in (1), i.e., $\|\nabla F(\mathbf{x}_T)\| \leq \varepsilon_g$ and $\lambda_{\min}\left(\nabla^2 F(\mathbf{x}_T)\right) \geq -(\varepsilon_H + \varepsilon)$.

3 Finite-sum minimization

In this section, we give concrete and practical examples to demonstrate ways to construct the approximate Hessian, which satisfies Condition 1. By considering *finite-sum*

minimization, a ubiquitous problem arising frequently in machine learning, we showcase the practical benefits of the proposed relaxed requirement (3a) for approximating Hessian, compared to the stronger alternative (5c). In Sect. 3.1, we describe randomized techniques to appropriately construct the approximate Hessian, followed by the convergence analysis of Algorithms 1 and 2 with such Hessian approximations in Sect. 3.2.

3.1 Randomized sub-sampling

Indeed, a major advantage of (3a) over (5c) is that there are many approximation techniques that can produce an inexact Hessian satisfying (3a). Of particular interest in our present paper is the application of randomized matrix approximation techniques, which have recently shown great success in the area of RandNLA at solving various numerical linear algebra tasks [22,42,60]. For this, we consider the highly prevalent finite-sum minimization problem (**P1**) and employ random sampling as a way to construct approximations to the exact Hessian, which are, probabilistically, ensured to satisfy (3a). Many machine learning and scientific computing applications involve finite-sum optimization problems of the form (**P1**) where each f_i is a loss (or misfit) function corresponding to ith observation (or measurement), e.g., see [7,24,47,48,50, 55] and references therein.

Here, we consider (**P1**) in large-scale regime where $n, d \gg 1$. In such settings, the mere evaluations of the Hessian and the gradient increase linearly in n. Indeed, for big-data problems, the operations with the Hessian, e.g., matrix-vector products involved in the (approximate) solution of the sub-problems (8) and (12), typically constitute the main bottleneck of computations, and in particular when $n \gg 1$, are computationally prohibitive. For the special case of (**P1**) in which each f_i is convex, randomized sub-sampling has shown to be effective in reducing such costs, e.g., [6,49,62]. We now show that such randomized approximation techniques can indeed be effectively employed for the non-convex settings considered in this paper.

In this light, suppose we have a probability distribution, $\mathbf{p} = \{p_i\}_{i=1}^n$, over the set $\{1, 2, \ldots, n\}$, such that for each index $i = 1, 2 \ldots, n$, we have $\Pr(i) = p_i > 0$ and $\sum_{i=1}^n p_i = 1$. Consider picking a sample of indices from $\{1, 2, \ldots, n\}$, at each iteration, randomly according to the distribution $\mathbf{p}$. Let $\mathcal{S}$ and $|\mathcal{S}|$ denote the sample collection and its cardinality, respectively and define

$$\mathbf{H}(\mathbf{x}) \triangleq \frac{1}{n|\mathcal{S}|} \sum_{j \in \mathcal{S}} \frac{1}{p_j} \nabla^2 f_j(\mathbf{x}), \tag{19}$$

to be the sub-sampled Hessian. In big-data regime when $n \gg 1$, if $|\mathcal{S}| \ll n$, such sub-sampling can offer significant computational savings.

Now, suppose

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \|\nabla^2 f_i(\mathbf{x})\| \leq K_i, \quad i = 1, 2, \ldots, n, \tag{20a}$$

and define

$$K_{\max} \triangleq \max_{i=1,\ldots,n} K_i. \tag{20b}$$

$$\widehat{K} \triangleq \frac{1}{n}\sum_{i=1}^{n} K_i. \tag{20c}$$

In this case, we can naturally consider uniform distribution over $\{1, 2, \ldots, n\}$, i.e., $p_i = 1/n, ; \forall i$. Lemma 16 gives the sample size required for the inexact Hessian, $\mathbf{H}(\mathbf{x})$, to probabilistically satisfy (3), for when the indices are picked uniformly at random *with* or *without* replacement.

Lemma 16 (Complexity of uniform sampling) *Given* (20a), (20b) , *and* $0 < \varepsilon, \delta < 1$, *let*

$$|\mathcal{S}| \geq \frac{16K_{\max}^2}{\varepsilon^2} \log \frac{2d}{\delta}, \tag{21}$$

where $K_{\max}$ *is defined as in* (20b). *At any* $\mathbf{x} \in \mathbb{R}^d$, *suppose picking the elements of* $\mathcal{S}$ *uniformly at random with or without replacement, and forming* $\mathbf{H}(\mathbf{x})$ *as in* (19) *with* $p_i = 1/n, ; \forall i$. *We have*

$$\Pr\Big(\|\mathbf{H}(\mathbf{x}) - \nabla^2 F(\mathbf{x})\| \leq \varepsilon\Big) \geq 1 - \delta. \tag{22}$$

Proof Consider $|\mathcal{S}|$ random matrices $\mathbf{H}_j(\mathbf{x})$, $j = 1, \ldots, |\mathcal{S}|$ s.t. $\Pr\big(\mathbf{H}_j(\mathbf{x}) = \nabla^2 f_i(\mathbf{x})\big) = 1/n;\ \forall i = 1, 2, \ldots, n$. Define $\mathbf{X}_j \triangleq \big(\mathbf{H}_j - \nabla^2 F(\mathbf{x})\big)$, $\mathbf{H} \triangleq \sum_{j\in\mathcal{S}} \mathbf{H}_j/|\mathcal{S}|$, and $\mathbf{X} \triangleq \sum_{j\in\mathcal{S}} \mathbf{X}_j = |\mathcal{S}|\big(\mathbf{H} - \nabla^2 F(\mathbf{x})\big)$. Note that $\mathbb{E}(\mathbf{X}_j) = 0$ and for $\mathbf{H}_j = \nabla^2 f_1(\mathbf{x})$ we have

$$\|\mathbf{X}_j\|^2 = \left\|\frac{n-1}{n}\nabla^2 f_1(\mathbf{x}) - \sum_{i=2}^{n}\frac{1}{n}\nabla^2 f_i(\mathbf{x})\right\|^2 \leq 4\left(\frac{n-1}{n}\right)^2 K_{\max}^2 \leq 4K_{\max}^2.$$

Hence, we can apply Operator-Bernstein inequality [35, Theorem 1] to get

$$\Pr\Big(\|\mathbf{H} - \nabla^2 F(\mathbf{x})\| \geq \varepsilon\Big) = \Pr\Big(\|\mathbf{X}\| \geq \varepsilon|\mathcal{S}|\Big) \leq 2d\exp\{-\varepsilon^2|\mathcal{S}|/(16K_{\max}^2)\}.$$

Now (21) ensure that $2d\exp\{-\varepsilon^2|\mathcal{S}|/(16K_{\max}^2)\} \leq \delta$, which gives (22). □

Indeed, if (22) holds, then (3a) follows with the same probability. In addition, if H is constructed according to Lemma 16, it is easy to see that (3b) is satisfied with $K_H = K_{\max}$ (in fact this is a deterministic statement). These two, together, imply that H satisfies Condition 1, with probability $1 - \delta$.

A Special Case: In certain settings, one might be able to construct a more "informative" distribution, $\mathbf{p}$, over the indices in the set $\{1, 2, \ldots, n\}$, as opposed to oblivious uniform

sampling. In particular, it might be advantageous to bias the probability distribution towards picking indices corresponding to those f_i's which are more *relevant*, in certain sense, in forming the Hessian. If this is possible, then we can only expect to require smaller sample size as compared with oblivious uniform sampling. One such setting where this is possible is the finite-sum optimization of the form (**P2**), which is indeed a special case of (**P1**) and arise often in many machine learning problems [51].

It is easy to see that, the Hessian of F in this case can be written as $\nabla^2 F(\mathbf{x}) = \mathbf{A}^T\mathbf{B}\mathbf{A} = \sum_{i=1}^n f_i''(\mathbf{a}_i^T\mathbf{x})\mathbf{a}_i\mathbf{a}_i^T/n$, where

$$\mathbf{A}^T = \begin{bmatrix} | & | & \dots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & \dots & | \end{bmatrix}_{d\times n} \quad \text{and}$$

$$\mathbf{B} = \frac{1}{n}\begin{bmatrix} f_1''(\mathbf{a}_1^T\mathbf{x}) & 0 & \dots & 0 \\ 0 & f_2''(\mathbf{a}_2^T\mathbf{x}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_n''(\mathbf{a}_n^T\mathbf{x}) \end{bmatrix}_{n\times n}.$$

Now let $\mathbf{S} \in \mathbb{R}^{n\times|\mathcal{S}|}$ be the sampling matrix and define the approximate Hessian as $\mathbf{H} \triangleq \mathbf{A}^T\mathbf{S}\mathbf{S}^T\mathbf{B}\mathbf{A}$. It can be seen that approximating the Hessian matrix $\nabla^2 F(\mathbf{x}) = \mathbf{A}^T\mathbf{B}\mathbf{A}$ can be regarded as approximating matrix-matrix multiplication from RandNLA [42,60]. For this, consider the sampling distribution $\mathbf{p}$ as

$$p_i = \frac{|f_i''(\mathbf{a}_i^T\mathbf{x})|\|\mathbf{a}_i\|_2^2}{\sum_{j=1}^n |f_j''(\mathbf{a}_j^T\mathbf{x})|\|\mathbf{a}_j\|_2^2}. \tag{23}$$

Note that the absolute values are needed since for non-convex f_i, we might have $f_j''(\mathbf{a}_j^T\mathbf{x}) < 0$ (for the convex case where all $f_j''(\mathbf{a}_j^T\mathbf{x}) \geq 0$, one can obtain stronger guarantees than Lemmas 16 and 17; see [62]). Using non-uniform sampling distribution (23), Lemma 17 gives sampling complexity for the approximate Hessian of (**P2**) to, probabilistically, satisfy (3).

Lemma 17 (Complexity of non-uniform sampling) *Given* (20a), (20c) *and* $0 < \varepsilon, \delta < 1$, *let*

$$|\mathcal{S}| \geq \frac{4\widehat{K}^2}{\varepsilon^2}\log\frac{2d}{\delta}, \tag{24}$$

where $\widehat{K}$ *is defined as in* (20c). *At any* $\mathbf{x} \in \mathbb{R}^d$, *suppose picking the elements of* $\mathcal{S}$ *randomly according to the probability distribution* (23), *and forming* $\mathbf{H}(\mathbf{x})$ *as in* (19). *We have*

$$\Pr\left(\|\mathbf{H} - \nabla^2 F(\mathbf{x})\| \leq \varepsilon\right) \geq 1 - \delta. \tag{25}$$

Proof Define $\mathbf{B} = \operatorname{diag}\{f_1''(\mathbf{a}_1^T\mathbf{x})/n, \cdots, f_n''(\mathbf{a}_n^T\mathbf{x})/n\} \in \mathbb{R}^{n\times n}$. Let $\mathbf{S} \in \mathbb{R}^{n\times|\mathscr{S}|}$ be the sampling matrix and define $\mathbf{H} \triangleq \mathbf{A}^T\mathbf{S}\mathbf{S}^T\mathbf{B}\mathbf{A}$. Further, let the diagonals of $\mathbf{B}$ be denoted by b_i and define $c \triangleq \sum_{i=1}^n |b_j|\|\mathbf{a}_j\|^2$. Consider s random matrices $\mathbf{H}_j$ such that $\Pr(\mathbf{H}_j = b_i\mathbf{a}_i\mathbf{a}_i^T/p_i) = p_i,\ \forall j = 1, 2, \ldots, |\mathscr{S}|$, where $p_i = |b_i|\|\mathbf{a}_i\|^2/(\sum_{i=1}^n |b_j|\|\mathbf{a}_j\|^2)$. Define

$$\mathbf{X}_j \triangleq \mathbf{H}_j - \mathbf{A}^T\mathbf{B}\mathbf{A}, \quad \mathbf{H} \triangleq \frac{1}{|\mathscr{S}|}\sum_{j=1}^{|\mathscr{S}|}\mathbf{H}_j, \quad \mathbf{X} \triangleq \sum_{j=1}^{|\mathscr{S}|}\mathbf{X}_j = |\mathscr{S}|\left(\mathbf{H} - \mathbf{A}^T\mathbf{B}\mathbf{A}\right).$$

Note that $\mathbb{E}[\mathbf{X}_j] = \sum_{i=1}^n p_i\left(b_i\mathbf{a}_i\mathbf{a}_i^T/p_i - \mathbf{A}^T\mathbf{B}\mathbf{A}\right) = 0$, and

$$\begin{aligned}\mathbb{E}[\mathbf{X}_j^2] &= \mathbb{E}[\mathbf{H}_j - \mathbf{A}^T\mathbf{B}\mathbf{A}] = \mathbb{E}[\mathbf{H}_j^2] + (\mathbf{A}^T\mathbf{B}\mathbf{A})^2 - \mathbb{E}[\mathbf{H}_j]\mathbf{A}^T\mathbf{B}\mathbf{A} - \mathbf{A}^T\mathbf{B}\mathbf{A}\mathbb{E}[\mathbf{H}_j]\\ &= \mathbb{E}[\mathbf{H}_j^2] - (\mathbf{A}^T\mathbf{B}\mathbf{A})^2 \preceq \mathbb{E}[\mathbf{H}_j^2] = \sum_{i=1}^n p_i\left(\frac{b_i}{p_i}\mathbf{a}_i\mathbf{a}_i^T\right)^2 = \sum_{i=1}^n \frac{b_i^2\|\mathbf{a}_i\|^2}{p_i}\mathbf{a}_i\mathbf{a}_i^T\\ &= \sum_{i=1}^n |b_j|\|\mathbf{a}_j\|^2\sum_{i=1}^n |b|_i\mathbf{a}_i\mathbf{a}_i^T = c\sum_{i=1}^n |b|_i\mathbf{a}_i\mathbf{a}_i^T = c\mathbf{A}^T|\mathbf{B}|\mathbf{A}.\end{aligned}$$

So we have $\|\mathbb{E}[\mathbf{X}_j^2]\| \le c\|\mathbf{A}^T|\mathbf{B}|\mathbf{A}\|$. Now we can apply the Operator-Bernstein inequality [35, Theorem 1] to get

$$\Pr\left(\|\mathbf{H} - \mathbf{A}^T\mathbf{B}\mathbf{A}\|_2 \ge \varepsilon\right) \le \Pr\left(\|\mathbf{X}\|_2 \ge \varepsilon|\mathscr{S}|\right) \le 2de^{\varepsilon^2|\mathscr{S}|/(4c\|\mathbf{A}^T|\mathbf{B}|\mathbf{A}\|)}.$$

Since $c = \sum_{i=1}^n |b_i|\,\|\mathbf{a}_i\|^2 = \frac{1}{n}\sum_{i=1}^n \left|f_i''\right|\|\mathbf{a}_i\|^2 \le \frac{1}{n}\sum_{i=1}^n K_i = \widehat{K}$ and

$$\left\|\mathbf{A}^T|\mathbf{B}|\mathbf{A}\right\| = \left\|\frac{1}{n}\sum_{i=1}^n \left|f_i''\right|\mathbf{a}_i\mathbf{a}_i^T\right\| \le \frac{1}{n}\sum_{i=1}^n \left\|\left|f_i''\right|\mathbf{a}_i\mathbf{a}_i^T\right\| \le \frac{1}{n}\sum_{i=1}^n K_i = \widehat{K},$$

then we have

$$\Pr\left(\|\mathbf{H} - \mathbf{A}^T\mathbf{B}\mathbf{A}\|_2 \ge \varepsilon\right) \le 2de^{\varepsilon^2|\mathscr{S}|/(4\widehat{K}^2)},$$

which gives the desired result. □

The bound in (24) can be improved by replacing the dimension d with a smaller quantity, known as intrinsic dimension; see Appendix A. As it can be seen from (20b) and (20c), since $\widehat{K} \le K_{\max}$, the sampling complexity given by Lemma 17 always provides a smaller sample-size compared with that prescribed by Lemma 16. Indeed, the advantage of non-uniform sampling is more pronounced in cases where the distribution of K_i's are highly skewed, i.e., a few large ones and many small ones, in which case we can have $\widehat{K} \ll K_{\max}$; see numerical experiments in [61]. Also, from (25), it follows that the approximate matrix $\mathbf{H}$, constructed according to Lemma 17

Table 1 Examples of problems in the form (**P2**) with the corresponding estimates for K_i in (20a)

Problem	Data	$f_i(\mathbf{a}_i^T\mathbf{x})$	$\nabla^2 f_i(\mathbf{a}_i^T\mathbf{x})$	K_i
Robust linear regression	$\mathbf{a}_i \in \mathbb{R}^d$ $b_i \in \mathbb{R}$	$\frac{\left(\mathbf{a}_i^T\mathbf{x}-b_i\right)^2}{1+\left(\mathbf{a}_i^T\mathbf{x}-b_i\right)^2}$	$\left(\frac{2\left(1-3\left(\mathbf{a}_i^T\mathbf{x}\right)^2\right)}{\left(\left(\mathbf{a}_i^T\mathbf{x}\right)^2+1\right)^3}\right)\mathbf{a}_i\mathbf{a}_i^T$	$\frac{\|\mathbf{a}_i\|^2}{6\sqrt{3}}$
Non-linear binary classification	$\mathbf{a}_i \in \mathbb{R}^d$ $b_i \in \{0,1\}$	$\left(\frac{1}{1+\exp\left(-\mathbf{a}_i^T\mathbf{x}\right)}-b_i\right)^2$	$\left(\frac{\exp\left(\mathbf{a}_i^T\mathbf{x}\right)\left(1-\exp\left(\mathbf{a}_i^T\mathbf{x}\right)\right)}{\left(\exp\left(\mathbf{a}_i^T\mathbf{x}\right)+1\right)^3}\right)\mathbf{a}_i\mathbf{a}_i^T$	$2\|\mathbf{a}_i\|^2$

satisfies (3b) with $K_H = \widehat{K} + \varepsilon$, with probability $1-\delta$, which in turn, implies that Condition 1 is ensured, with probability $1-\delta$.

As concrete examples of the problems in the form (**P2**) where Lemma 17 can be readily used, Table 1 gives estimates for K_i in (20a) for robust linear regression with smooth non-convex bi-weight loss, [3], as well as non-convex binary-classification using logistic regression with least squares loss, [61].

3.2 Probabilistic convergence analysis

Now, we are in the position to give iteration complexity for Algorithms 1 and 2 where the inexact Hessian matrix $\mathbf{H}_t$ is constructed according to Lemmas 16 or 17. Since the approximation is a probabilistic construction, in order to guarantee success, we need to ensure that we require a small failure probability across all iterations. In particular, in order to get an overall and accumulative success probability of $1-\delta$ for the entire T iterations, the per-iteration failure probability is set as $(1-\sqrt[T]{(1-\delta)}) \in \mathcal{O}(\delta/T)$. This failure probability appears only in the "log factor" for sample size in all of our results, and so it is not the dominating cost. Hence, requiring that all T iterations are successful for a large T, only necessitates a small (logarithmic) increase in the sample size. For example, for $T \in \mathcal{O}(\max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\})$, as in Theorem 2, we can set the per-iteration failure probability to $\delta \min\{\varepsilon_g^2, \varepsilon_H^3\}$, and ensure that when Algorithm 2 terminates, all Hessian approximations have been, accumulatively, successful with probability of $1-\delta$.

Using these results, we can have the following probabilistic, but optimal, guarantee on the worst-case iteration complexity of Algorithm 1 for solving finite-sum problem (**P1**) (or (**P2**)) and in the case where the inexact Hessian is formed by sub-sampling. Their proofs follow very similar line of reasoning as that used for obtaining the results of Sect. 2, and hence are omitted.

Theorem 4 (Optimal complexity of Algorithm 1 for finite-sum problem) *Consider any* $0 < \varepsilon_g, \varepsilon_H, \delta < 1$. *Let* ε *be as in* (11) *and set* $\delta_0 = \delta \min\{\varepsilon_g^2 \varepsilon_H, \varepsilon_H^3\}$. *Furthermore, for such* (ε, δ_0), *let the sample-size* $|\mathcal{S}|$ *be as in* (21) *(or* (24)*) and form the sub-sampled matrix* $\mathbf{H}$ *as in* (19). *For Problem* (**P1**) *(or* (**P2**)*), under Assumption 1 and Condition 2, Algorithm 1 terminates in at most* $T \in \mathcal{O}(\max\{\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\})$ *iterations, upon which, with probability* $1-\delta$, *we have that* $\|\nabla F(\mathbf{x})\| \leq \varepsilon_g$, *and* $\lambda_{\min}(\nabla^2 F(\mathbf{x})) \geq -(\varepsilon + \varepsilon_H)$.

Similarly, in the setting of optimization problems (**P1**) and (**P2**), with appropriate sub-sampling of the Hessian as in Lemmas 16 and 17 , we can also obtain probabilistic worst-case iteration complexities for Algorithm 2 as in the deterministic case. Again, the proofs are similar to those in Sect. 2, and hence are omitted.

Theorem 5 (Complexity of Algorithm 2 for finite-sum problem) *Consider any* $0 < \varepsilon_g, \varepsilon_H, \delta < 1$. *Let* ε *be as in* (16) *and set* $\delta_0 = \delta \min\{\varepsilon_g^2, \varepsilon_H^3\}$. *Furthermore, for such* (ε, δ_0), *let the sample-size* $|\mathcal{S}|$ *be as in* (21) *(or* (24)*) and form the sub-sampled matrix* $\mathbf{H}$ *as in* (19). *For Problem* (**P1**) *(or* (**P2**)*), under Assumption 1 and Condition 3, Algorithm 2 terminates in at most* $T \in \mathcal{O}(\max\{\varepsilon_g^{-2}, \varepsilon_H^{-3}\})$ *iterations, upon which, with probability* $1 - \delta$, *we have that* $\|\nabla F(\mathbf{x})\| \leq \varepsilon_g$, *and* $\lambda_{\min}(\nabla^2 F(\mathbf{x})) \geq -(\varepsilon + \varepsilon_H)$.

Theorem 6 (Optimal complexity of Algorithm 2 for finite-sum problem) *Consider any* $0 < \varepsilon_g, \varepsilon_H, \delta < 1$. *Let* ε *be as in Theorem 3 and set* $\delta_0 = \delta \min\{\varepsilon_g^{3/2}, \varepsilon_H^3\}$. *Furthermore, for such* (ε, δ_0), *let the sample-size* $|\mathcal{S}|$ *be as in* (21) *(or* (24)*) and form the sub-sampled matrix* $\mathbf{H}$ *as in* (19). *For Problem* (**P1**) *(or* (**P2**)*), under Assumption 1, Conditions 3 and 4 , Algorithm 2 terminates in at most* $T \in \mathcal{O}(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\})$ *iterations, upon which, with probability* $1 - \delta$, *we have that* $\|\nabla F(\mathbf{x})\| \leq \varepsilon_g$, *and* $\lambda_{\min}(\nabla^2 F(\mathbf{x})) \geq -(\varepsilon + \varepsilon_H)$.

As it can be seen, the main difference between Theorems 5 and 6 is in the solution to the sub-problem (12). More specifically, if in addition to Condition 3, Condition 4 is also satisfied, then Theorem 6 gives *optimal* worst-case iteration complexity.

4 Conclusion

We considered non-convex optimization settings and developed efficient variants of the trust region and adaptive cubic regularization methods in which both the sub-problems as well as the the curvature information are suitably approximated. For all of our proposed variants, we obtained iteration complexities to achieve approximate second order criticality, which are shown to be the same (up to some constant) as that of the exact variants.

As compared with previous works, our proposed Hessian approximation condition offers a range of theoretical and practical advantages. As a concrete example, we considered the large-scale finite-sum optimization problem and proposed uniform and non-uniform sub-sampling strategies as ways to efficiently construct the desired approximate Hessian. We then, probabilistically, established optimal iteration complexity for variants of trust region and adaptive cubic regularization methods in which the Hessian is appropriately sub-sampled.

In this paper, we focused on approximating the Hessian under the exact gradient information. Arguably, the bottleneck of the computations in such second-order methods involves the computations with the Hessian, e.g., matrix-vector products in the (approximate) solution of the sub-problem. In fact, the cost of the exact gradient computation is typically amortized by that of the operations with the Hessian. In spite of this, approximating the gradient in a computationally feasible way and with minimum

assumptions could improve upon the efficiency of the methods proposed here. However, care has to be taken as cheaper iterations with inaccurate gradients could in fact result in more iterations overall. This could have the adverse effect of slowing down the algorithm's convergence. As a result, approximating the gradient has to be done with care to avoid such pitfalls.

Finally, we mention that our focus here has been solely on developing the theoretical foundations of such randomized algorithms. Extensive empirical evaluations of these algorithms on various machine learning applications are given in the [61].

Acknowledgements Fred Roosta gratefully acknowledges the generous support by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). He was also partially supported by the Australian Research Council through a Discovery Early Career Researcher Award (DE180100923). Michael W. Mahoney would like to acknowledge ARO, DARPA, and NSF for providing partial support of this work.

Appendix A: Intrinsic dimension and improving the sampling complexity (24)

We can still improve the sampling complexity (24) by considering the intrinsic dimension of the matrix $\mathbf{A}^T|\mathbf{B}|\mathbf{A}$. Recall that for a SPSD matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, the intrinsic dimension is defined as $t(A) = \mathrm{tr}(\mathbf{A})/\|\mathbf{A}\|$, where $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$. The intrinsic dimension can be regarded as a measure for the number of dimensions where $\mathbf{A}$ has a significant spectrum. It is easy to see that $1 \le t(\mathbf{A}) \le \mathrm{rank}(\mathbf{A}) \le d$; see [59] for more details. Now let $t = \mathrm{tr}(\mathbf{A}^T|\mathbf{B}|\mathbf{A})/\|\mathbf{A}^T|\mathbf{B}|\mathbf{A}\|$ be the intrinsic dimension of the SPSD matrix $\mathbf{A}^T|\mathbf{B}|\mathbf{A}$. We have the following improved sampling complexity result:

Lemma 18 (Complexity of non-uniform sampling: intrinsic dimension) *The result of Lemma 17 holds with* (24) *replaced with*

$$|\mathscr{S}| \ge \frac{16\widehat{K}^2}{3\varepsilon^2}\log\frac{8t}{\delta}, \tag{26}$$

where $t = tr(\mathbf{A}^T|\mathbf{B}|\mathbf{A})/\|\mathbf{A}^T|\mathbf{B}|\mathbf{A}\| \le d$ *is the intrinsic dimension of the matrix* $\mathbf{A}^T|\mathbf{B}|\mathbf{A}$.

Proof It is easy to see that $\mathrm{Var}(\mathbf{X}) = \mathbb{E}(\mathbf{X}^2) = \sum_{j=1}^{|\mathscr{S}|}\mathbb{E}(\mathbf{X}_j^2) \preceq |\mathscr{S}|c\mathbf{A}^T|\mathbf{B}|\mathbf{A}$, where $\mathbf{X}$ and c are given in the proof of Lemma 17. For $\mathbf{H}_j = \frac{b_i}{p_i}\mathbf{a}_i\mathbf{a}_i^T$, we have

$$\begin{aligned}\lambda_{\max}(\mathbf{X}_j) \le \|\mathbf{X}_j\| = \|\frac{b_i}{p_i}\mathbf{a}_i\mathbf{a}_i^T - \mathbf{A}^T\mathbf{B}\mathbf{A}\| &= \left\|\left(\frac{1-p_i}{p_i}\right)b_i\mathbf{a}_i\mathbf{a}_i^T - \sum_{j\neq i}b_j\mathbf{a}_j\mathbf{a}_j^T\right\| \\ &\le \left(\frac{1-p_i}{p_i}\right)|b_i|\|\mathbf{a}_i\|^2 + \sum_{j\neq i}|b_j|\|\mathbf{a}_j\|^2\end{aligned}$$

$$= (1 - p_i) \sum_{i=1}^{n} |b_j| \|\mathbf{a}_j\|^2 + \sum_{j \neq i} |b_j| \|\mathbf{a}_j\|^2$$

$$= 2 \sum_{j \neq i} |b_j| \|\mathbf{a}_j\|^2 \leq 2 \sum_{i=1}^{n} |b_j| \|\mathbf{a}_j\|^2 = 2c.$$

Hence, if $\varepsilon |\mathcal{S}| \geq \sqrt{|\mathcal{S}| c \|\mathbf{A}^T |\mathbf{B}| \mathbf{A}\|} + 2c/3$, we can apply Matrix Bernstein using the intrinsic dimension [59, Theorem 7.7.1] to get for $\varepsilon \leq 1/2$

$$\Pr\left(\lambda_{\max}(\mathbf{X}) \geq \varepsilon |\mathcal{S}|\right) \leq 4t \exp\left\{\frac{-\varepsilon^2 |\mathcal{S}|}{2c \|\mathbf{A}^T |\mathbf{B}| \mathbf{A}\| + 4c\varepsilon/3}\right\} \leq 4t \exp\left\{\frac{-3\varepsilon^2 |\mathcal{S}|}{16c^2}\right\}.$$

Applying the same bound for $\mathbf{Y}_j = -\mathbf{X}_j$ and $\mathbf{Y} = \sum_{j=1}^{s} \mathbf{Y}_j$, followed by the union bound, we get the desired result. □

Appendix B: Computation of Approximate Negative Curvature Direction

Throughout our analysis, we assume that, if a sufficiently negative curvature exists, i.e., $\lambda_{\min}(\mathbf{H}) \leq -\varepsilon_H$ for some $\varepsilon_H \in (0, 1)$, we can approximately compute the corresponding negative curvature direction vector $\mathbf{u}$, i.e., $\langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle \leq -\nu \varepsilon_H \|\mathbf{u}\|^2$, for some $\nu \in (0, 1)$. We note that this can be done efficiently by applying a variety of methods such as Lanczos [38] or shift-and-invert [25] on the SPSD matrix $\tilde{\mathbf{H}} = K_H - \mathbf{H}$. These methods only employ matrix vector products and, hence, are suitable for large scale problems. More specifically, with any $\kappa \in (0, 1)$, these methods using $\mathcal{O}(\log(d/\delta)\sqrt{K_H/\kappa})$ matrix-vector products and with probability $1 - \delta$, yield a vector $\mathbf{u}$ satisfying $K_H \|\mathbf{u}\|^2 - \langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle = \langle \mathbf{u}, \tilde{\mathbf{H}}\mathbf{u} \rangle \geq \kappa \lambda_{\min}(\tilde{\mathbf{H}}) \|\mathbf{u}\|^2 = \kappa(K_H - \lambda_{\min}(\mathbf{H})) \|\mathbf{u}\|^2$. Rearranging, we obtain $\langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle \leq (1 - \kappa) K_H \|\mathbf{u}\|^2 + \kappa \lambda_{\min}(\mathbf{H}) \|\mathbf{u}\|^2$. Setting $1 > \nu = 2\kappa \geq (2K_H)/(2K_H + \varepsilon_H)$, gives $\langle \mathbf{u}, \mathbf{H}\mathbf{u} \rangle \leq -\nu \varepsilon_H \|\mathbf{u}\|^2$.

References

1. Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., Ma, T.: Finding approximate local minima faster than gradient descent (2016). ArXiv preprint arXiv:1611.01146
2. Bandeira, A.S., Scheinberg, K., Vicente, L.N.: Convergence of trust-region methods based on probabilistic models. SIAM J. Optim. **24**(3), 1238–1264 (2014)
3. Beaton, A.E., Tukey, J.W.: The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics **16**(2), 147–185 (1974)
4. Bianconcini, T., Liuzzi, G., Morini, B., Sciandrone, M.: On the use of iterative methods in cubic regularization for unconstrained optimization. Comput. Optim. Appl. **60**(1), 35–57 (2015)
5. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust region method for nonconvex optimization (2018). ArXiv preprint arXiv:1609.07428v3
6. Bollapragada, R., Byrd, R., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization (2016). ArXiv preprint arXiv:1609.08502
7. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning (2016). ArXiv preprint arXiv:1606.04838

8. Carmon, Y., Duchi, J.C.: Gradient descent efficiently finds the cubic-regularized non-convex Newton step (2016). ArXiv preprint arXiv:1612.00547
9. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM J. Optim. **20**(6), 2833–2852 (2010)
10. Cartis, C., Gould, N.I., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math. Program. **127**(2), 245–295 (2011)
11. Cartis, C., Gould, N.I., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. Math. Program. **130**(2), 295–319 (2011)
12. Cartis, C., Gould, N.I., Toint, P.L.: Optimal Newton-type methods for nonconvex smooth optimization problems. Tech. rep., ERGO technical report 11-009, School of Mathematics, University of Edinburgh (2011)
13. Cartis, C., Gould, N.I., Toint, P.L.: Complexity bounds for second-order optimality in unconstrained optimization. J. Complex. **28**(1), 93–108 (2012)
14. Cartis, C., Gould, N.I., Toint, P.L.: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. Optim. Methods Softw. **27**(2), 197–219 (2012)
15. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Math. Program. **169**(2), 337–375 (2018)
16. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models (2015). ArXiv preprint arXiv:1504.04231
17. Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, Philadelphia (2000)
18. Conn, A.R., Scheinberg, K., Vicente, L.N.: Global convergence of general derivative-free trust-region algorithms to first-and second-order critical points. SIAM J. Optim. **20**(1), 387–415 (2009)
19. Curtis, F.E., Robinson, D.P., Samadi, M.: An inexact regularized Newton framework with a worst-case iteration complexity of $\mathcal{O}(\varepsilon^{-3/2})$ for nonconvex optimization (2017). ArXiv preprint arXiv:1708.00475
20. Curtis, F.E., Robinson, D.P., Samadi, M.: A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\varepsilon^{-3/2})$ for nonconvex optimization. Math. Program. **162**(1–2), 1–32 (2017)
21. Dennis, J.E., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**(126), 549–560 (1974)
22. Drineas, P., Mahoney, M.W.: RandNLA: randomized numerical linear algebra. Commun. ACM **59**(6), 80–90 (2016)
23. Erway, J.B., Gill, P.E., Griffin, J.D.: Iterative methods for finding a trust-region step. SIAM J. Optim. **20**(2), 1110–1131 (2009)
24. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer Series in Statistics, vol. 1. Springer, Berlin (2001)
25. Garber, D., Hazan, E., Jin, C., Musco, C., Netrapalli, P., Sidford, A., et al.: Faster eigenvector computation via shift-and-invert preconditioning. In: International Conference on Machine Learning, pp. 2626–2634 (2016)
26. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points-online stochastic gradient for tensor decomposition. In: COLT, pp. 797–842 (2015)
27. Ghadimi, S., Liu, H., Zhang, T.: Second-order methods with cubic regularization under inexact information (2017). ArXiv preprint arXiv:1710.05782
28. Gould, N.I., Lucidi, S., Roma, M., Toint, P.L.: Solving the trust-region subproblem using the Lanczos method. SIAM J. Optim. **9**(2), 504–525 (1999)
29. Gould, N.I., Robinson, D.P., Thorne, H.S.: On solving trust-region and other regularised subproblems in optimization. Math. Program. Comput. **2**(1), 21–57 (2010)
30. Grapiglia, G.N., Yuan, J., Yuan, Y.: On the worst-case complexity of nonlinear stepsize control algorithms for convex unconstrained optimization. Optim. Methods Softw. **31**(3), 591–604 (2016)
31. Gratton, S., Mouffe, M., Toint, P.L., Weber-Mendonça, M.: A recursive-trust-region method for bound-constrained nonlinear optimization. IMA J. Numer. Anal. **28**(4), 827–861 (2008)
32. Gratton, S., Royer, C.W., Vicente, L.N., Zhang, Z.: Complexity and global rates of trust-region methods based on probabilistic models. IMA J. Numer. Anal. **38**(3), 1579–1597 (2018)
33. Gratton, S., Sartenaer, A., Toint, P.L.: Recursive trust-region methods for multiscale nonlinear optimization. SIAM J. Optim. **19**(1), 414–444 (2008)

34. Griewank, A.: The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12. Department of Applied Mathematics and Theoretical Physics, University of Cambridge (1981)
35. Gross, D., Nesme, V.: Note on sampling without replacing from a finite collection of matrices (2010). ArXiv preprint arXiv:1001.2738
36. Hazan, E., Koren, T.: A linear-time algorithm for trust region problems. Math. Program. **158**(1–2), 363–381 (2016)
37. Kohler, J.M., Lucchi, A.: Sub-sampled cubic regularization for non-convex optimization (2017). ArXiv preprint arXiv:1705.05933
38. Kuczyński, J., Woźniakowski, H.: Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. SIAM J. Matrix Anal. Appl. **13**(4), 1094–1122 (1992)
39. Larson, J., Billups, S.C.: Stochastic derivative-free optimization using a trust region framework. Comput. Optim. Appl. **64**(3), 619–645 (2016)
40. Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent only converges to minimizers. In: Conference on Learning Theory, pp. 1246–1257 (2016)
41. Lenders, F., Kirches, C., Potschka, A.: trlib: a vector-free implementation of the GLTR method for iterative solution of the trust region problem (2016). ArXiv preprint arXiv:1611.04718
42. Mahoney, M.W.: Randomized algorithms for matrices and data. Found. Trends® Mach. Learn. **3**(2), 123–224 (2011)
43. Moré, J.J., Sorensen, D.C.: Computing a trust region step. SIAM J. Sci. Stat. Comput. **4**(3), 553–572 (1983)
44. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1), 159–181 (2008)
45. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
46. Nocedal, J., Wright, S.: Numerical Optimization. Springer, Berlin (2006)
47. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Data completion and stochastic algorithms for PDE inversion problems with many measurements. Electron. Trans. Numer. Anal. **42**, 177–196 (2014)
48. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Stochastic algorithms for inverse problems involving PDEs and many measurements. SIAM J. Sci. Comput. **36**(5), S3–S22 (2014)
49. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods. Math. Program. **174**(1), 293–326 (2019)
50. Roosta-Khorasani, F., Székely, G.J., Ascher, U.: Assessing stochastic algorithms for large scale nonlinear least squares problems using extremal probabilities of linear combinations of gamma random variables. SIAM/ASA J. Uncertain. Quantif. **3**(1), 61–90 (2015)
51. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
52. Shashaani, S., Hashemi, F., Pasupathy, R.: ASTRO-DF: a class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization (2016). ArXiv preprint arXiv:1610.06506
53. Sorensen, D.: Minimization of a large-scale quadratic functionsubject to a spherical constraint. SIAM J. Optim. **7**(1), 141–161 (1997)
54. Sorensen, D.C.: Newton's method with a model trust region modification. SIAM J. Numer. Anal. **19**(2), 409–426 (1982)
55. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. MIT Press, Cambridge (2012)
56. Steihaug, T.: The conjugate gradient method and trust regions in large scale optimization. SIAM J. Numer. Anal. **20**(3), 626–637 (1983)
57. Tripuraneni, N., Stern, M., Jin, C., Regier, J., Jordan, M.I.: Stochastic cubic regularization for fast nonconvex optimization (2017). ArXiv preprint arXiv:1711.02838
58. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**(4), 389–434 (2012)
59. Tropp, J.A.: An introduction to matrix concentration inequalities (2015). ArXiv preprint arXiv:1501.01571
60. Woodruff, D.P.: Sketching as a tool for numerical linear algebra (2014). ArXiv preprint arXiv:1411.4357
61. Xu, P., Roosta-Khorasani, F., Mahoney, M.W.: Second-order optimization for non-convex machine learning: an empirical study (2017). ArXiv preprint arXiv:1708.07827

62. Xu, P., Yang, J., Roosta-Khorasani, F., Ré, C., Mahoney, M.W.: Sub-sampled Newton methods with non-uniform sampling. In: Advances in Neural Information Processing Systems, pp. 3000–3008 (2016)