SIAM J. OPTIM.
Vol. 28, No. 2, pp. 1448–1477

COMPLEXITY ANALYSIS OF SECOND-ORDER LINE-SEARCH ALGORITHMS FOR SMOOTH NONCONVEX OPTIMIZATION*

CLÉMENT W. ROYER† AND STEPHEN J. WRIGHT‡

Abstract. There has been much recent interest in finding unconstrained local minima of smooth functions, due in part to the prevalence of such problems in machine learning and robust statistics. A particular focus is algorithms with good complexity guarantees. Second-order Newton-type methods that make use of regularization and trust regions have been analyzed from such a perspective. More recent proposals, based chiefly on first-order methodology, have also been shown to enjoy optimal iteration complexity rates, while providing additional guarantees on computational cost. In this paper, we present an algorithm with favorable complexity properties that differs in two significant ways from other recently proposed methods. First, it is based on line searches only: Each step involves computation of a search direction, followed by a backtracking line search along that direction. Second, its analysis is rather straightforward, relying for the most part on the standard technique for demonstrating sufficient decrease in the objective from backtracking. In the latter part of the paper, we consider inexact computation of the search directions, using iterative methods in linear algebra: the conjugate gradient and Lanczos methods. We derive modified convergence and complexity results for these more practical methods.

Key words. smooth nonconvex unconstrained optimization, line-search methods, second-order methods, second-order necessary conditions, iteration complexity

AMS subject classifications. 49M05, 49M15, 90C06, 90C60

DOI. 10.1137/17M1134329

1. Introduction.

We consider the unconstrained optimization problem

$$\min f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice Lipschitz continuously differentiable function that is generally nonconvex. Some algorithms for this problem seek points that nearly satisfy the second-order necessary conditions for optimality, which are that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$. These iterative schemes terminate at an iterate x_k for which

$$\|\nabla f(x_k)\| \leq \epsilon_g \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_H, \tag{2}$$

where $\epsilon_g, \epsilon_H \in (0, 1)$ are (typically small) prescribed tolerances. Numerous algorithms have been proposed in recent years for finding points that satisfy (2), each with a complexity guarantee, which is an upper bound on an index k that satisfies (2), in terms of ϵ_g, ϵ_H, and other quantities. We summarize below the main results.

*Received by the editors June 12, 2017; accepted for publication (in revised form) January 30, 2018; published electronically May 8, 2018.

http://www.siam.org/journals/siopt/28-2/M113432.html

Funding: Work of the first author was supported by Subcontract 3F-30222 from Argonne National Laboratory. Work of the second author was supported by NSF Awards IIS-1447449, 1628384, 1634597, and 1740707; AFOSR Award FA9550-13-1-0138; and Subcontract 3F-30222 from Argonne National Laboratory. Part of this work was done while the second author was visiting the Simons Institute for the Theory of Computing, and partially supported by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF Award CCF-1740425.

†Wisconsin Institute of Discovery, University of Wisconsin, Madison, WI 53715 (croyer2@wisc.edu).

‡Computer Sciences Department, University of Wisconsin, Madison, WI 53706 (swright@cs.wisc.edu).

Classical second-order convergent trust-region schemes [10] can be shown to satisfy (2) after at most $\mathcal{O}(\max\{\epsilon_g^{-2}\,\epsilon_H^{-1}, \epsilon_H^{-3}\})$ iterations [9]. Cubic regularization methods in their basic form [6] have better complexity bounds than trust-region schemes, requiring at most $\mathcal{O}(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\})$ iterations. The difference can be explained by the restriction enforced by the trust-region constraint on the norm of the steps. Recent work has shown that it is possible to improve the bound for trust-region algorithms using specific definitions of the trust-region radius [13]. The best-known iteration bound for a second-order algorithm (that is, an algorithm relying on the use of second-order derivatives and Newton-type steps) is $\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$. This bound was established originally (under the form of a global convergence rate) in [17] by considering cubic regularization of Newton's method. The same result is achieved by the adaptive cubic regularization framework under suitable assumptions on the computed step [9]. Recent proposals have shown that the same bound can be attained by algorithms other than cubic regularization. A modified trust-region method [11], a variable-norm trust-region scheme [16], and a quadratic regularization algorithm with cubic descent condition [2] all achieve the same bound.

When $\epsilon_g = \epsilon_H = \epsilon$ for some $\epsilon \in (0, 1)$, all the bounds mentioned above reduce to $\mathcal{O}(\epsilon^{-3})$. It has been established that this order is sharp for the class of second-order methods [9] and it can be proved for a wide range of algorithms that make use of second-order derivative information; see [12]. Setting $\epsilon_H = \epsilon^{1/2}$ and $\epsilon_g = \epsilon$ for some $\epsilon > 0$ yields bounds varying between $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-3/2})$, the latter being again optimal within the class of second-order algorithms [8].

A new trend in complexity analyses has emerged recently that focuses on measuring not just the number of iterations to achieve (2) but also the computational cost of the iterations. Two independent proposals, respectively based on adapting accelerated gradient to the nonconvex setting [4] and approximately solving the cubic subproblem [1], require $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-7/4})$ operations (with high probability, showing only dependency on ϵ) to find a point x_k that satisfies

$$\|\nabla f(x_k)\| \le \epsilon \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) \ge -\sqrt{L_H \epsilon}, \tag{3}$$

with L_H being a Lipschitz constant of the Hessian. The difference factor of $\epsilon^{-1/4}$ by comparison with the complexities of the previous paragraph is due to the cost of computing a negative eigenvalue of $\nabla^2 f(x_k)$ and/or the cost of solving the linear system. A later proposal [3] focuses on solving cubic subproblems via gradient descent, together with an inexact eigenvalue computation: It satisfies (3) in at most $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-2})$ with high probability. Another technique [14] requires only gradient computations, with noise being added to some iterates. It reaches with high probability a point satisfying (3) in at most $\mathcal{O}(\log^4(\frac{1}{\epsilon})\epsilon^{-2})$ iterations. Up to the logarithmic factor, this bound is characteristic of gradient-type methods, but classical work establishes only first-order guarantees [5]. Although this setting is not explicitly addressed in the cited papers, it appears that to reach an iterate satisfying (2) with $\epsilon_g = \epsilon_H = \epsilon$, the methods studied in [1, 4] would require $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-7/2})$ iterations, while the methods described in [3] and [14] could require $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-3})$ and $\mathcal{O}(\log^4(\frac{1}{\epsilon})\epsilon^{-3})$ iterations, respectively. Although these bounds look worse than those of classical nonlinear optimization schemes, they are more informative in that they not only account for the number of outer iterations of the algorithm, but also for the cost of performing each outer iteration (often measured in terms of the number of inner iterations, each of which has similar cost). We note, however, that unlike the classical complexity results, the newer procedures make use of randomization, so the bounds typically hold

only with high probability.

Our goal in this paper is to describe an algorithm that achieves optimal complexity, whether measured by the number of iterations required to satisfy the condition (2) or by an estimate of the number of fundamental operations required (gradient evaluations or Hessian-vector multiplications). Each iteration of our algorithm takes the form of a step calculation followed by a backtracking line search. (To our knowledge, ours is the first line-search algorithm that is endowed with a second-order complexity analysis.) The "reference" version of our algorithm is presented in section 2, along with its complexity analysis. In this version, we assume that two key operations—solution of the linear equations to obtain Newton-like steps and calculation of the most negative eigenvalue of a Hessian—are performed exactly. In section 3, we refine our study by introducing inexactness into these operations and adjusting the complexity bounds appropriately. Finally, we discuss the established results and their practical connections in section 4.

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm, unless otherwise indicated by a subscript. A vector v will be called a *unit vector* if $\|v\| = 1$.

2. A line-search algorithm based on exact step computations. We now describe an algorithm based on exact computation of search directions, in particular, the Newton-like search directions and the eigenvector that corresponds to the most negative eigenvalue of the Hessian.

2.1. Outline. We use a standard line-search framework [18, Chapter 3]. Starting from an initial iterate x_0, we apply an iterative scheme of the form $x_{k+1} = x_k + \alpha_k d_k$, where d_k is a chosen search direction and α_k is a step length computed by a backtracking line-search procedure.

Algorithm 1 defines our method. Each iteration begins by evaluating the gradient, together with the curvature of the function along the gradient direction. This information determines whether the negative gradient direction is a suitable choice for search direction d_k, and if so, what scaling should be applied to it. If not, we compute the minimum eigenvalue of the Hessian. The corresponding eigenvector is used as the search direction whenever the eigenvalue is sufficiently negative. Otherwise, we compute a Newton-like search direction, adding a regularization term if needed to ensure sufficient positive definiteness of the coefficient matrix. There are a total of five possible choices for the search direction d_k (including two different scalings of the negative gradient). Table 1 summarizes the various steps that can be performed and the conditions under which those steps are chosen.

TABLE 1
Steps and associated decrease lemmas for Algorithm 1.

Context			Direction	Decrease
$\|g_k\| = 0$	-	$\lambda_k < -\epsilon_H$	v_k	Lemma 1
	$R_k < -\epsilon_H$		$R_k g_k / \|g_k\|$	Lemma 1
$\|g_k\| > \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$		$-g_k / \|g_k\|^{1/2}$	Lemma 2
$\|g_k\| \le \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$	$\lambda_k < -\epsilon_H$	v_k	Lemma 1
$\|g_k\| \le \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$	$\lambda_k \in [-\epsilon_H, \epsilon_H]$	d_k^r	Lemma 4
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k < -\epsilon_H$	v_k	Lemma 1
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k \in [-\epsilon_H, \epsilon_H]$	d_k^r	Lemma 4
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k > \epsilon_H$	d_k^n	Lemma 3

Once a search direction has been selected, a backtracking line search is applied

with an initial choice of 1. A sufficient condition related to the cube of the step norm must be satisfied; see (7). Such a condition has been instrumental in the complexity analysis of recently proposed Newton-type methods achieving the best known iteration complexity rates [2, 11].

At most one eigenvector computation and one linear system solve are needed per iteration of Algorithm 1, along with a gradient evaluation and the Hessian-vector multiplication required to calculate R_k.

The algorithm contains two tests for termination, with the option of switching to a "Local Phase" instead of terminating at a point that satisfies approximate second-order conditions. The Local Phase aims for rapid local convergence to a point satisfying second-order necessary conditions for a local solution; it is detailed in Algorithm 2. Termination (or switch to the Local Phase) occurs at an iteration k at which an (ϵ_g, ϵ_H)-*approximate second-order critical point* is reached, according to the following definition:

$$\min\{\|g_k\|, \|g_{k+1}\|\} \leq \epsilon_g \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_H, \tag{8}$$

where $g_k = \nabla f(x_k)$, etc. As we see below, the quantity $\min\{\|g_k\|, \|g_{k+1}\|\}$ arises naturally in the decrease formula we establish for the steps computed by Algorithm 1. In fact, for the methods we reviewed in the introduction, one observes that the decrease formulas obtained for their steps either involve only $\|g_k\|$ [1, 3, 4, 14, 17], only $\|g_{k+1}\|$ [2, 11, 16], or the minimum of the two quantities [7]. The latter case appears due to the presence of both gradient-type (see Lemma 2) and Newton-type steps (see Lemmas 3 and 4).

The main convergence results of this section are complexity results on the number of iterations or function evaluations required to satisfy condition (8) *for the first time*. (Algorithm 2 makes provision for reentering the main algorithm if the approximate second-order conditions are violated at any point. This reentry feature is not covered by our complexity analysis.)

2.2. Iteration complexity. We now establish a complexity bound for Algorithm 1 in the form of the maximum number of iterations that may occur before the termination conditions are satisfied for the first time. To this end, we provide guarantees on the decrease that can be obtained for each of the possible choices of search direction.

In the rest of this paper, we make the following assumptions.

Assumption 1. The level set $\mathcal{L}_f(x_0) = \{x | f(x) \leq f(x_0)\}$ is a compact set.

Assumption 2. The function f is twice Lipschitz continuously differentiable on an open neighborhood of $\mathcal{L}_f(x_0)$, and we denote by L_g and L_H the respective Lipschitz constants for ∇f and $\nabla^2 f$ on this set.

By the continuity of f and its derivatives, Assumption 1 implies that there exist $f_{\text{low}} \in \mathbb{R}$, $U_g > 0$, and $U_H > 0$ such that, for every $x \in \mathcal{L}_f(x_0)$, one has

$$f(x) \geq f_{\text{low}}, \quad \|\nabla f(x)\| \leq U_g, \quad \|\nabla^2 f(x)\| \leq U_H. \tag{9}$$

We point out that the choice $U_H = L_g$ is a valid one for theoretical purposes. However, U_H will serve as an explicit parameter of our inexact method in section 3, so we use separate notation, to allow U_H to be an overestimate of L_g.

An immediate consequence of these assumptions is that, for any x and d such that Assumption 2 is satisfied at x and $x + d$, we have

$$f(x+d) \leq f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \frac{L_H}{6} \|d\|^3. \tag{10}$$

Algorithm 1 Second-order line-search method

Init. Choose $x^0 \in \mathbb{R}^n$, $\theta \in (0,1)$, $\eta > 0$, $\epsilon_g \in (0,1)$, $\epsilon_H \in (0,1)$;
for $k = 0, 1, 2, \ldots$ **do**
  Step 1. (First Order) Set $g_k = \nabla f(x_k)$;
  if $\|g_k\| = 0$ **then**
    Go to Step 2;
  end if
  Compute $R_k = \frac{g_k^\top \nabla^2 f(x_k) g_k}{\|g_k\|^2}$;
  if $R_k < -\epsilon_H$ **then**
    Set $d_k = \frac{R_k}{\|g_k\|} g_k$ and go to Step LS;
  else if $R_k \in [-\epsilon_H, \epsilon_H]$ and $\|g_k\| > \epsilon_g$ **then**
    Set $d_k = -\frac{g_k}{\|g_k\|^{1/2}}$ and go to Step LS;
  else
    Go to Step 2;
  end if
  Step 2. (Second Order) Compute an eigenpair $(v_k, \lambda_k) \in \mathbb{R}^n \times \mathbb{R}$, where $\lambda_k = \lambda_{\min}(\nabla^2 f(x_k))$ and v_k is such that

$$\nabla^2 f(x_k) v_k = \lambda_k v_k, \quad v_k^\top g_k \leq 0, \quad \|v_k\| = [-\lambda_k]_+; \tag{4}$$

  if $\|g_k\| \leq \epsilon_g$ and $\lambda_k \geq -\epsilon_H$ **then**
    Terminate (or go to Local Phase);
  else if $\lambda_k < -\epsilon_H$ **then**
    (Negative Curvature) Set $d_k = v_k$;
  else if $\lambda_k > \epsilon_H$ **then**
    (Newton) Set $d_k = d_k^n$, where

$$\nabla^2 f(x_k) d_k^n = -g_k; \tag{5}$$

  else
    (Regularized Newton) Set $d_k = d_k^r$, where

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d_k^r = -g_k; \tag{6}$$

  end if
  Go to Step LS;
  Step LS. (Line Search) Compute a step length $\alpha_k = \theta^{j_k}$, where j_k is the smallest nonnegative integer such that

$$f(x_k + \alpha_k d_k) < f(x_k) - \frac{\eta}{6} \alpha_k^3 \|d_k\|^3 \tag{7}$$

holds, and set $x_{k+1} = x_k + \alpha_k d_k$.
  if $d_k = d_k^n$ or $d_k = d_k^r$ and $\|\nabla f(x_{k+1})\| \leq \epsilon_g$ **then**
    Terminate (or go to Local Phase);
  end if
end for

Algorithm 2 Local Phase

loop
  Set $g_k = \nabla f(x_k)$;
  if $\|g_k\| > \epsilon_g$ **then**
    Return to Algorithm 1;
  end if
  Compute λ_k and v_k as in (4);
  if $\lambda_k < -\epsilon_H$ **then**
    Return to Algorithm 1;
  else if $\lambda_k \in (-\epsilon_H, 0]$ **then**
    Set $d_k = d_k^r$ from (6);
  else
    Set $d_k = d_k^n$ from (5);
  end if
  Perform backtracking line search as in Step LS of Algorithm 1 to obtain x_{k+1};
  $k \leftarrow k + 1$;
end loop

The following four technical lemmas derive bounds on the decrease obtained from each type of step. The proofs are rather similar to each other, and follow the usual template for backtracking line-search methods.

We begin with negative curvature directions, showing that our choices for initial scaling yield a decrease proportional to the cube of the (negative) curvature in that direction.

LEMMA 1. *Under Assumption* 2*, suppose that the search direction for the* k*th iteration of Algorithm* 1 *is chosen either as* $d_k = \frac{R_k}{\|g_k\|} g_k$ *with* $R_k < -\epsilon_H$ *in Step* 1 *or* $d_k = v_k$ *in Step* 2*. Then the backtracking line search terminates with step length* $\alpha_k = \theta^{j_k}$ *with* $j_k \leq j_e + 1$*, where*

$$j_e := \left\lceil \log_\theta \left(\frac{3}{L_H + \eta} \right) \right\rceil_+, \tag{11}$$

and the decrease in the function value resulting from the chosen step length satisfies

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; c_e \left[\frac{|d_k^\top \nabla^2 f(x_k) d_k|}{\|d_k\|^2} \right]^3 \tag{12}$$

with

$$c_e := \frac{\eta}{6} \min \left\{ 1, \frac{27\theta^3}{(L_H + \eta)^3} \right\}.$$

Proof. For the direction $d_k = R_k g_k / \|g_k\|$, we have

$$d_k^T \nabla^2 f(x_k) d_k = R_k^2 \frac{g_k^T \nabla^2 f(x_k) g_k}{\|g_k\|^2} = R_k^3 = -\|d_k\|^3.$$

For the other choice $d_k = v_k$, we have $d_k^T \nabla^2 f(x_k) d_k = \lambda_k^3 = -\|d_k\|^3$, so that in both cases we have

$$d_k^T \nabla^2 f(x_k) d_k = -\|d_k\|^3 \quad \text{and} \quad \frac{|d_k^T \nabla^2 f(x_k) d_k|}{\|d_k\|^2} = \|d_k\|. \tag{13}$$

Thus, if the unit value $\alpha_k = 1$ is accepted by (7), the result (12) holds trivially. Suppose now that the unit step length is not accepted. Then the choice $\alpha = \theta^j$ does *not* satisfy the decrease condition (7) for some $j \geq 0$. Using (10) and the definition of d_k, we obtain

$$\begin{aligned} -\frac{\eta}{6}\alpha^3\|d_k\|^3 \leq f(x_k+\alpha d_k) - f(x_k) &\leq \alpha g_k^\top d_k + \frac{\alpha^2}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\alpha^3\|d_k\|^3 \\ &\leq \frac{\alpha^2}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\alpha^3\|d_k\|^3 \\ &= -\frac{\alpha^2}{2}\|d_k\|^3 + \frac{L_H}{6}\alpha^3\|d_k\|^3, \end{aligned}$$

where the last line follows from (13). Therefore, we have

$$\alpha = \theta^j \geq \frac{3}{L_H + \eta}, \tag{14}$$

which holds only if $j \leq j_e$ by definition of j_e. Thus, the line search must terminate with (7) being satisfied for some value $j_k \leq j_e + 1$. Because the line search did not stop with step length $\theta^{j_k - 1}$, we must have

$$\theta^{j_k - 1} \geq \frac{3}{L_H + \eta} \Rightarrow \theta^{j_k} \geq \frac{3\theta}{L_H + \eta}.$$

As a result, the decrease satisfied by the step $\alpha_k d_k = \theta^{j_k} d_k$ is such that

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; \frac{\eta}{6}\theta^{3j_k}\|d_k\|^3 \;\geq\; \frac{\eta}{6}\frac{27\theta^3}{(L_H+\eta)^3}\left[\frac{|d_k^\top \nabla^2 f(x_k) d_k|}{\|d_k\|^2}\right]^3.$$

This inequality, together with the analysis for the case in which $\alpha_k = 1$, establishes the desired result. ∎

The second result concerns use of the step $d_k = -g_k/\|g_k\|^{1/2}$ in the case in which the curvature of the function along the gradient direction is small.

LEMMA 2. *Let Assumptions* 1 *and* 2 *hold. Then, if at the kth iteration of Algorithm* 1 *the search direction is* $d_k = -g_k/\|g_k\|^{1/2}$*, the backtracking line search terminates with step length* $\alpha_k = \theta^{j_k}$*, with* $j_k \leq j_g + 1$*, where*

$$j_g := \left\lfloor \log_\theta \left(\min\left\{ \frac{5}{3}, \sqrt{\frac{1}{L_H + \eta}} \right\} \min\left\{ \epsilon_g^{1/2} \epsilon_H^{-1}, 1 \right\} \right) \right\rfloor_+, \tag{15}$$

and the resulting step length α_k *is such that*

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; c_g \min\left\{ \epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2} \right\}, \tag{16}$$

where

$$c_g := \frac{\eta}{6} \min\left\{ 1, \frac{\theta^3}{(L_H+\eta)^{3/2}}, \frac{125\theta^3}{27} \right\}.$$

Proof. Recall that the choice $d_k = -g_k/\|g_k\|^{1/2}$ is adopted only when $\|g_k\| > \epsilon_g$ and $|R_k| \leq \epsilon_H$. If the unit step length $\alpha_k = 1$ is accepted, we have

$$f(x_k) - f(x_k + d_k) \;\geq\; \frac{\eta}{6}\|d_k\|^3 \;=\; \frac{\eta}{6}\|g_k\|^{3/2} \;\geq \frac{\eta}{6}\epsilon_g^{3/2},$$

satisfying (16). Otherwise, it means that there exists $j \geq 0$ for which the decrease condition (7) is not satisfied using the step size θ^j. For such j, we have from (10) that

$$\begin{aligned} -\frac{\eta}{6}\theta^{3j}\|g_k\|^{3/2} &\leq f(x_k - \theta^j\|g_k\|^{-1/2}g_k) - f(x_k) \\ &\leq -\theta^j\|g_k\|^{3/2} + \frac{\theta^{2j}}{2}R_k\|g_k\| + \frac{L_H}{6}\theta^{3j}\|g_k\|^{3/2} \\ &\leq -\theta^j\|g_k\|^{3/2} + \frac{\theta^{2j}}{2}\epsilon_H\|g_k\| + \frac{L_H}{6}\theta^{3j}\|g_k\|^{3/2}, \end{aligned}$$

which leads to

$$0 \leq \left[-\frac{5}{6}\theta^j\|g_k\|^{3/2} + \frac{\theta^{2j}}{2}\epsilon_H\|g_k\|\right] + \left[-\frac{1}{6}\theta^j\|g_k\|^{3/2} + \frac{L_H+\eta}{6}\theta^{3j}\|g_k\|^{3/2}\right]. \tag{17}$$

Therefore, at least one of the two terms between brackets must be nonnegative. If

$$-\frac{5}{6}\theta^j\|g_k\|^{3/2} + \frac{\theta^{2j}}{2}\epsilon_H\|g_k\| \geq 0,$$

we have $\theta^j \geq \frac{5}{3}\|g_k\|^{1/2}\epsilon_H^{-1}$. On the other hand, if

$$-\frac{1}{6}\theta^j\|g_k\|^{3/2} + \frac{L_H+\eta}{6}\theta^{3j}\|g_k\|^{3/2} \geq 0,$$

then $\theta^j \geq \sqrt{\frac{1}{L_H+\eta}}$. Putting the two bounds together, we have that

$$\theta^j \geq \min\left\{\frac{5}{3}\|g_k\|^{1/2}\epsilon_H^{-1}, \sqrt{\frac{1}{L_H+\eta}}\right\} \tag{18a}$$

$$\geq \min\left\{\frac{5}{3}, \sqrt{\frac{1}{L_H+\eta}}\right\}\min\left\{\|g_k\|^{1/2}\epsilon_H^{-1}, 1\right\} \tag{18b}$$

$$\geq \min\left\{\frac{5}{3}, \sqrt{\frac{1}{L_H+\eta}}\right\}\min\left\{\epsilon_g^{1/2}\epsilon_H^{-1}, 1\right\}. \tag{18c}$$

Since $j > j_g$ contradicts (18c), the line search terminates with (7) being satisfied for some value $j_k \leq j_g + 1$. Since (7) did not hold for $\alpha = \theta^{j_k-1}$, we have from (18b) that

$$\theta^{j_k} \;\geq\; \theta\min\left\{\frac{5}{3}, \sqrt{\frac{1}{L_H+\eta}}\right\}\min\left\{\|g_k\|^{1/2}\epsilon_H^{-1}, 1\right\}.$$

The decrease obtained by the step length $\alpha_k = \theta^{j_k}$ thus satisfies

$$\begin{aligned} f(x_k) - f(x_k+\alpha_k d_k) &\geq \frac{\eta}{6}\theta^{3j_k}\|d_k\|^3 \\ &\geq \frac{\eta}{6}\left[\theta\min\left\{\frac{5}{3}, \sqrt{\frac{1}{L_H+\eta}}\right\}\right]^3 \min\left\{\|g_k\|^{3/2}\epsilon_H^{-3}, 1\right\}\|g_k\|^{3/2} \\ &\geq \frac{\eta}{6}\left[\theta\min\left\{\frac{5}{3}, \sqrt{\frac{1}{L_H+\eta}}\right\}\right]^3 \min\left\{\epsilon_g^3\epsilon_H^{-3}, \epsilon_g^{3/2}\right\}. \end{aligned} \tag{19}$$

Thus (16) is also satisfied in the case of $\alpha_k < 1$, completing the proof. ∎

Lemma 2 describes the reduction that can be achieved along the negative gradient direction when the curvature of the function in this direction is modest. When this curvature is significantly positive (or when this curvature is slightly positive but the gradient is small), we compute the minimum Hessian eigenvalue (Step 2) and consider other options for the search direction.

Our next result concerns the decrease that can be guaranteed by the Newton step, when it is computed.

LEMMA 3. *Let Assumptions* 1 *and* 2 *hold. Suppose that the Newton direction* $d_k = d_k^n$ *is used at the* k*th iteration of Algorithm* 1*. Then the backtracking line search terminates with step length* $\alpha_k = \theta^{j_k}$*, with* $j_k \leq j_n + 1$*, where*

$$j_n := \left\lceil \log_\theta \left(\sqrt{\frac{3}{L_H + \eta}} \frac{\epsilon_H}{\sqrt{U_g}} \right) \right\rceil_+, \tag{20}$$

and we have

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; c_n \min \left\{ \|\nabla f(x_k + \alpha_k d_k)\|^{3/2}, \epsilon_H^3 \right\}, \tag{21}$$

where

$$c_n := \frac{\eta}{6} \min \left\{ \left[\frac{2}{L_H} \right]^{3/2}, \left[\frac{3\theta}{L_H + \eta} \right]^3 \right\}.$$

Proof. Note first that the Newton direction $d_k = d_k^n$ is computed only when $\nabla^2 f(x_k)) \succ \epsilon_H I$, so we have

$$\|d_k\| \leq \|\nabla^2 f(x_k)^{-1}\| \|g_k\| \leq U_g / \epsilon_H. \tag{22}$$

Suppose first that the step length $\alpha_k = 1$ satisfies the decrease condition (7). Then from (5) and (10), we have

$$\begin{aligned} \|\nabla f(x_k + \alpha_k d_k)\| &= \|\nabla f(x_k + d_k) - \nabla f(x_k) + \nabla f(x_k)\| \\ &= \left\|\nabla f(x_k + d_k) - \nabla f(x_k) - \nabla^2 f(x_k) d_k\right\| \leq \frac{L_H}{2} \|d_k\|^2. \end{aligned}$$

We thus have the following bound on the decrease obtained with the unitary Newton step:

$$f(x_k) - f(x_k + d_k) \;\geq\; \frac{\eta}{6} \left[\frac{2}{L_H} \right]^{3/2} \|\nabla f(x_k + d_k)\|^{3/2}. \tag{23}$$

Suppose now that the unit step length does not allow for a sufficient decrease as measured by (7). Then this condition must fail for $\alpha_k = \theta^j$ for some $j \geq 0$. For this

value, we have from (10) that

$$\begin{aligned} -\frac{\eta}{6}\theta^{3j}\|d_k\|^3 &\le f(x_k+\theta^j\,d_k)-f(x_k)\\ &\le \theta^j g_k^T d_k + \frac{\theta^{2j}}{2} d_k^T\nabla^2 f(x_k)d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3\\ &\le \theta^j\left(\frac{\theta^j}{2}-1\right) d_k^T\nabla^2 f(x_k)d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3\\ &\le -\frac{\theta^j}{2} d_k^T\nabla^2 f(x_k)d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3\\ &\le -\frac{\theta^j}{2}\epsilon_H\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3, \end{aligned} \tag{24}$$

where we used $\nabla^2 f(x_k) \succeq \epsilon_H I$ for the final inequality. This relation holds in particular for $j = 0$, in which case it gives

$$-\frac{\eta}{6}\|d_k\|^3 \le -\frac{\epsilon_H}{2}\|d_k\|^2 + \frac{L_H}{6}\|d_k\|^3$$

leading to the following lower bound on the norm of the Newton step:

$$\|d_k\| \;\ge\; \frac{3}{L_H+\eta}\epsilon_H. \tag{25}$$

More generally, for any integer j such that the decrease condition is not satisfied, we have from (24) that

$$\theta^j \;\ge\; \sqrt{\frac{3}{L_H+\eta}}\epsilon_H^{1/2}\|d_k\|^{-1/2}. \tag{26}$$

For any $j > j_n$, the last inequality is violated since

$$\theta^j < \theta^{j_n} \le \sqrt{\frac{3}{L_H+\eta}}\frac{\epsilon_H}{\sqrt{U_g}} = \sqrt{\frac{3}{L_H+\eta}}\epsilon_H^{1/2}\frac{\epsilon_H^{1/2}}{\sqrt{U_g}} \le \sqrt{\frac{3}{L_H+\eta}}\epsilon_H^{1/2}\|d_k\|^{-1/2},$$

where we used (22) for the final inequality. This proves that the condition (7) will be satisfied by some $j_k \le j_n+1$. Since $\alpha = \theta^{j_k-1}$ does not fulfill the decrease requirement, it follows from (26) that

$$\theta^{j_k} \;\ge\; \theta\sqrt{\frac{3}{L_H+\eta}}\epsilon_H^{1/2}\|d_k\|^{-1/2}.$$

By substituting this lower bound into the sufficient decrease condition, and then using (25), we obtain

$$\begin{aligned} f(x_k)-f(x_k+\alpha_k d_k) &= f(x_k)-f(x_k+\theta^{j_k}d_k)\\ &\ge \frac{\eta}{6}\theta^{3\,j_k}\|d_k\|^3\\ &\ge \frac{\eta}{6}\theta^3\left[\frac{3}{L_H+\eta}\right]^{3/2}\epsilon_H^{3/2}\|d_k\|^{-3/2}\|d_k\|^3\\ &\ge \frac{\eta}{6}\theta^3\left[\frac{3}{L_H+\eta}\right]^{3}\epsilon_H^{3}, \end{aligned}$$

where the final inequality is from (25). We obtain the required result by combining this inequality with the bound (23) for the case in which $\alpha_k = 1$. ∎

Our last intermediate result addresses the case of a regularized Newton step.

LEMMA 4. *Let Assumptions* 1 *and* 2 *hold. Suppose that* $d_k = d_k^r$ *at the* k*th iteration of Algorithm* 1. *Then the backtracking line search terminates with step length* $\alpha_k = \theta^{j_k}$, *with* $j_k \leq j_r + 1$, *where*

$$j_r := \left[\log_\theta \left(\frac{6}{L_H + \eta} \frac{\epsilon_H^2}{U_g} \right) \right]_+, \tag{27}$$

and we have

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; c_r \min \left\{ \|\nabla f(x_k + \alpha_k d_k)\|^3 \; \epsilon_H^{-3}, \epsilon_H^3 \right\}, \tag{28}$$

where

$$c_r := \frac{\eta}{6} \min \left\{ \left[\frac{1}{1 + \sqrt{1 + L_H/2}} \right]^3, \left[\frac{6\theta}{L_H + \eta} \right]^3 \right\}.$$

Proof. Note first that the regularized Newton step is taken only when $\nabla^2 f(x_k) \succeq -\epsilon_H I$. Thus the minimum eigenvalue of the coefficient matrix in (6) is $\lambda_k + 2\epsilon_H \geq \epsilon_H$, and we have

$$\|d_k\| \;\leq\; \frac{\|g_k\|}{\lambda_k + 2\epsilon_H} \;\leq\; \frac{\|g_k\|}{\epsilon_H} \;\leq\; \frac{U_g}{\epsilon_H}. \tag{29}$$

Suppose first that the unit step is accepted. Then the gradient norm at the new point satisfies

$$\begin{aligned} \|\nabla f(x_k + d_k)\| &= \|\nabla f(x_k + d_k) - \nabla f(x_k) + \nabla f(x_k)\| \\ &= \left\|\nabla f(x_k + d_k) - \nabla f(x_k) - \nabla^2 f(x_k) d_k - 2\epsilon_H d_k\right\| \\ &\leq \frac{L_H}{2} \|d_k\|^2 + 2\epsilon_H \|d_k\|, \end{aligned}$$

and therefore

$$\frac{L_H}{2} \|d_k\|^2 + 2\epsilon_H \|d_k\| - \|\nabla f(x_k + d_k)\| \;\geq\; 0.$$

By treating the left-hand side as a quadratic in $\|d_k\|$ and applying Lemma 17 with $a = 2$, $b = 2L_H$, and $t = \|\nabla f(x_k + d_k)\|/\epsilon_H^2$, we obtain from this bound that

$$\begin{aligned} \|d_k\| &\geq \frac{-2\epsilon_H + \sqrt{4\epsilon_H^2 + 2L_H \|\nabla f(x_k + d_k)\|}}{L_H} \\ &= \frac{-2 + \sqrt{4 + 2L_H \|\nabla f(x_k + d_k)\|/\epsilon_H^2}}{L_H} \, \epsilon_H \\ &\geq \frac{-2 + \sqrt{4 + 2L_H}}{L_H} \min \left(\|\nabla f(x_k + d_k)\|/\epsilon_H^2, 1 \right) \epsilon_H \\ &= \frac{2L_H}{L_H(2 + \sqrt{4 + 2L_H})} \min \left(\|\nabla f(x_k + d_k)\|/\epsilon_H, \epsilon_H \right) \\ &= \frac{1}{1 + \sqrt{1 + L_H/2}} \min \left(\|\nabla f(x_k + d_k)\|/\epsilon_H, \epsilon_H \right). \end{aligned} \tag{30}$$

Therefore, if the unit step is accepted, we have

$$f(x_k) - f(x_k + d_k)$$
$$\geq \frac{\eta}{6}\|d_k\|^3 \geq \frac{\eta}{6}\left[\frac{1}{1+\sqrt{1+L_H/2)}}\right]^3 \min\left(\|\nabla f(x_k+d_k)\|^3\epsilon_H^{-3}, \epsilon_H^3\right). \tag{31}$$

If the unit step does not yield a sufficient decrease, there must be a value $j \geq 0$ such that (7) is not satisfied for $\alpha = \theta^j$. For such j, and using again (10), we have

$$\begin{aligned}
-\frac{\eta}{6}\theta^{3j}\|d_k\|^3 &\leq f(x_k + \theta^j d_k) - f(x_k) \\
&\leq \theta^j g_k^\top d_k + \frac{\theta^{2j}}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&= \theta^j \left(1 - \frac{\theta^j}{2}\right) g_k^\top d_k - \epsilon_H \theta^{2j}\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&\leq -\epsilon_H \theta^{2j}\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3.
\end{aligned}$$

Thus, for any $j \geq 0$ for which sufficient decrease is not obtained, one has

$$\theta^j \;\geq\; \frac{6}{L_H + \eta}\epsilon_H \|d_k\|^{-1}. \tag{32}$$

Meanwhile, we have from the definition of j_r that

$$\theta^{j_r} \leq \frac{6}{L_H+\eta}\frac{\epsilon_H^2}{U_g} \leq \frac{6}{L_H+\eta}\epsilon_H \frac{\epsilon_H}{U_g} \leq \frac{6}{L_H+\eta}\epsilon_H\|d_k\|^{-1},$$

using the upper bound (29). By comparing this bound with (32), we deduce that the backtracking line-search procedure terminates with $j_k \leq j_r + 1$, where $j_k \geq 1$ by our earlier assumption. Thus, since (32) is satisfied for $j = j_k - 1$, we have

$$\theta^{j_k} \geq \frac{6\theta}{L_H+\eta}\epsilon_H\|d_k\|^{-1},$$

and therefore

$$f(x_k) - f(x_k + \theta^{j_k} d_k) \;\geq\; \frac{\eta}{6}\theta^{3j_k}\|d_k\|^3 \;\geq\; \frac{\eta}{6}\left[\frac{6\theta}{L_H+\eta}\right]^3 \epsilon_H^3.$$

By combining this bound with (31), obtained for the unit-step case, we obtain the result. ∎

By combining the estimates of function decrease proved in the lemmas above, we bound the number of iterations needed by Algorithm 1 to satisfy the approximate second-order optimality conditions (8).

THEOREM 5. *Let Assumptions* 1 *and* 2 *hold. Then Algorithm* 1 *reaches an iterate that satisfies* (8) *in at most*

$$\mathcal{C}\max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_g^{-3/2}, \epsilon_H^{-3}\right\} \tag{33}$$

iterations, where

$$\mathcal{C} := c^{-1}(f(x_0) - f_{\text{low}}), \quad c := \min\{c_g, c_v, c_n, c_r\}. \tag{34}$$

Proof. Suppose l is an iteration at which the conditions for termination are *not* satisfied. We consider in turn the various types of steps that could have been taken at iteration l and obtain a lower bound on the amount of decrease obtained from each. Table 1 is helpful in working through the various cases. We consider two main cases, and several subcases.

Case 1: $\lambda_l < -\epsilon_H$. From Table 1, we see that in this case, the search direction is either a scaling of $-g_k$ or the most-negative-curvature direction v_k. When $R_l < -\epsilon_H$, we have $d_l = \frac{R_l}{\|g_l\|} g_l$, and Lemma 1 indicates the following bound on function decrease:

$$f(x_l) - f(x_{l+1}) \;\geq\; c_e \epsilon_H^3.$$

When $R_l \in [-\epsilon_H, \epsilon_H]$ and $\|g_l\| > \epsilon_g$, we have $d_l = -g_l/\|g_l\|^{1/2}$. Thus, using Lemma 2, we have

$$f(x_l) - f(x_{l+1}) \;\geq\; c_g \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2}\right\}.$$

For the remaining cases in which "$\|g_l\| \leq \epsilon_g$ and $R_l \in [-\epsilon_H, \epsilon_H]$" and "$\|g_l\| > \epsilon_g$ and $R_l > \epsilon_H$," the search direction is necessarily v_l. We have from Lemma 1 that

$$f(x_l) - f(x_{l+1}) \;\geq\; c_e \left[\frac{|d_l^\top \nabla^2 f(x_l) d_l|}{\|d_l\|^2}\right]^3 = c_e |\lambda_l|^3 \geq c_e \epsilon_H^3.$$

Case 2: $\lambda_l \geq -\epsilon_H$, $\|g_l\| > \epsilon_g$, and $\|g_{l+1}\| > \epsilon_g$. In this case, we have three possible choices for the search direction. The first one is $d_l = -g_l/\|g_l\|^{1/2}$, in which case we have from Lemma 2 that

$$f(x_l) - f(x_{l+1}) \;\geq\; c_g \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2}\right\}.$$

The second possible choice is the Newton direction $d_l = d_l^n$. Using Lemma 3, we obtain

$$f(x_l) - f(x_{l+1}) \;\geq\; c_n \min\left\{\epsilon_g^{3/2}, \epsilon_H^3\right\}.$$

The third choice is the regularized Newton direction $d_l = d_l^r$, for which Lemma 4 yields

$$f(x_l) - f(x_{l+1}) \;\geq\; c_r \min\left\{\|g_{l+1}\|^3 \epsilon_H^{-3}, \epsilon_H^3\right\} \;\geq\; c_r \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_H^3\right\}.$$

By putting all these bounds together, we obtain the following lower bound on the decrease in f on iteration l:

$$f(x_l) - f(x_{l+1}) \geq c \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2}, \epsilon_H^3\right\}, \tag{35}$$

where c is defined in (34). Consequently, summing across all iterations up to k yields

$$f(x_0) - f_{\text{low}} \geq \sum_{l=0}^{k-1} f(x_l) - f(x_{l+1}) \geq kc \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2}, \epsilon_H^3\right\},$$

which implies that k is bounded above by (33). Therefore, there must exist a finite index k_ϵ such that (8) is satisfied. For this index, the bound (33) applies, hence the result. ∎

We now look further into the various components of the bound established in Theorem 5.

Dependencies on the tolerances (ϵ_g, ϵ_H). The result (33) makes explicit the variation of the bound with respect to the two tolerances. As this result differs from those in the literature, we follow two usual approaches to ease the comparison with other methods.

Letting $\epsilon_g = \epsilon$ and $\epsilon_H = \sqrt{\epsilon}$ for some $\epsilon \in (0, 1)$ allows us to equate all components of the maximum term in (33); indeed,

$$\epsilon_g^{-3}\epsilon_H^3 = \epsilon_g^{-3/2} = \epsilon_H^{-3} = \epsilon^{-3/2},$$

and therefore our bound is $\mathcal{O}(\epsilon^{-3/2})$. On the other hand, the choice $\epsilon_g = \epsilon_H = \epsilon$, which puts first- and second-order requirement on an equal footing, leads to a bound in $\mathcal{O}(\epsilon^{-3})$. Both match the optimal bounds known for second-order globally convergent methods in terms of iteration count.

Dependencies on problem-algorithmic constants. Although our main goal is to analyze dependencies with respect to the tolerances, our bounds can also reflect dependencies on problem-dependent quantities, namely, the initial function value discrepancy $f(x_0) - f_{\text{low}}$ and the Lipschitz constants L_g and L_H. It can be seen from the lemmas of this subsection that

$$c_e = \mathcal{O}(L_H^{-3}), \quad c_g = \mathcal{O}(L_H^{-3/2}), \quad c_n = \mathcal{O}(L_H^{-3}), \quad c_r = \mathcal{O}(L_H^{-3}).$$

As a result, the iteration complexity of our method is in

$$\mathcal{O}\left((f(x_0) - f_{\text{low}})L_H^3 \max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_g^{-3/2}, \epsilon_H^{-3}\right\}\right).$$

2.3. Evaluation/inner iteration complexity. We now discuss the function evaluation complexity of Algorithm 1, which counts the number of function calls required by the algorithm before its termination conditions are satisfied. We need to refine the iteration complexity analysis of section 2.2 to take into account the function evaluations associated with the backtracking line-search process.

THEOREM 6. *Suppose that Assumptions* 1 *and* 2 *hold. The number of function evaluations required by Algorithm* 1 *prior to reaching a point that satisfies* (8) *is at most*

$$\left[1 + \mathcal{K} + \log_\theta\left(\min\{\epsilon_H^2, \epsilon_g^{1/2}\epsilon_H^{-1}\}\right)\right]\mathcal{C}\max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_g^{-3/2}, \epsilon_H^{-3}\right\}, \tag{36}$$

where

$$\mathcal{K} := \left[\log_\theta\left(\min\left\{\frac{3}{L_H + \eta}, \frac{5}{3}, \frac{1}{(L_H + \eta)^{1/2}}, \sqrt{\frac{3}{(L_H + \eta)U_g}}, \frac{6}{(L_H + \eta)U_g}\right\}\right)\right]_+$$

and $\mathcal{C}$ *is defined as in Theorem* 5.

Proof. Theorem 5 gives a bound on the number of iterations. By Lemmas 1–4, a bound on the corresponding number of function evaluations is

$$(1 + \max\{j_e, j_g, j_n, j_r\})\,\mathcal{C}\max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_g^{-3/2}, \epsilon_H^{-3}\right\}.$$

Using the definitions of j_e, j_g, j_n, and j_r from Lemmas 1, 2, 3, and 4, respectively, and the fact that $\epsilon_g, \epsilon_H \in (0, 1)$ yields the result. ∎

With our specific choices of ϵ_g and ϵ_H mentioned in the previous section, the evaluation complexity bounds are $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-3/2})$ and $\mathcal{O}(\log(\frac{1}{\epsilon})\epsilon^{-3})$, respectively. We can also derive a bound that includes dependencies on problem constants; for instance, the bound corresponding to $\epsilon_g = \epsilon_H = \epsilon$ is

$$\mathcal{O}\left(\log\left(\tfrac{\max\{L_H, U_g, U_H\}}{\epsilon}\right)\,(f(x_0) - f_{\text{low}})L_H^{-3}\epsilon^{-3}\right).$$

2.4. Local convergence. In the previous sections, we have derived global complexity guarantees for Algorithm 1. We now aim to show rapid local convergence for the variant of the algorithm that invokes the Local Phase, Algorithm 2, rather than terminating as soon as the conditions (8) are satisfied. We note that local convergence results like the one we prove here have in the past gone hand in hand with global convergence results in smooth nonconvex optimization (see, for example, [18]). More recently, several works in the optimization literature have established rapid local convergence alongside global complexity guarantees [2, 6, 11].

For this section, we will make the following additional assumption.

Assumption 3. The sequence of iterates generated by Algorithm 1 in conjunction with Algorithm 2 converges to a local minimizer, that is, a point x^* at which $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

Under this assumption, the following result is immediate.

LEMMA 7. *Under Assumptions* 1, 2 *and* 3, *there exists* $k_0 \in \mathbb{N}$ *such that, for every* $k \geq k_0$, *we have for* $\mu := \frac{1}{2}\min(1, \lambda_{\min}(\nabla^2 f(x^*))) > 0$ *that*

$$\mu I \preceq \nabla^2 f(x_k) \preceq U_H I \tag{37}$$

and

$$\|g_k\| < \min\left\{\frac{3\mu^4}{L_H + \eta}, \epsilon_g\right\}. \tag{38}$$

Note that the conditions on k_0 in Lemma 7 are such that the combined strategy of Algorithms 1 and 2 will have entered the Local Phase (Algorithm 2) before iteration k_0, and will stay in this phase at all subsequent iterations.

We now establish a local quadratic convergence result.

THEOREM 8. *Suppose that Assumptions* 1, 2, *and* 3 *are satisfied, and let* μ *and* k_0 *be as defined in Lemma* 7. *Then, for every* $k \geq k_0$, *the method always takes the Newton direction with a unit step length and we have*

$$\|g_{k+1}\| \leq \frac{L_H}{2\mu^2}\|g_k\|^2 \leq \frac{3}{8}\|g_k\|. \tag{39}$$

Proof. Let $k \geq k_0$, so that we are in the Local Phase (Algorithm 2) at iteration k. By Lemma 7, the Hessian at $\nabla^2 f(x_k)$ is positive definite, with smallest eigenvalue bounded below by $\mu > 0$. Thus Algorithm 2 computes the Newton direction $d_k = d_k^n$ and we have

$$\|d_k\| \leq \|g_k\|/\mu, \quad g_k^\top d_k \leq -\mu\|g_k\|^2.$$

We thus have

$$\begin{aligned} f(x_k + d_k) - f(x_k) &\leq g_k^\top d_k + \frac{1}{2}d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\|d_k\|^3 \\ &= \frac{1}{2}g_k^\top d_k + \frac{L_H}{6}\|d_k\|^3 \leq -\frac{\mu}{2}\|g_k\|^2 + \frac{L_H}{6}\|d_k\|^3. \end{aligned}$$

Thus if the sufficient decrease condition $f(x_k+d_k)-f(x_k) \le -\frac{\eta}{6}\|d_k\|^3$ is *not* satisfied for the unit step, we must have

$$\frac{L_H+\eta}{6}\|d_k\|^3 \ge \frac{\mu}{2}\|g_k\|^2,$$

which by the bound $\|d_k\| \le \|g_k\|/\mu$ can be true only if

$$\frac{L_H+\eta}{6}\frac{\|g_k\|^3}{\mu^3} \ge \frac{\mu}{2}\|g_k\|^2 \quad \Leftrightarrow \quad \|g_k\| \ge \frac{3\mu^4}{L_H+\eta},$$

which contradicts (38). Thus the unit Newton step is taken and we have

$$\begin{aligned}\|g_{k+1}\| = \|\nabla f(x_k+d_k)\| &= \left\|\nabla f(x_k+d_k) - \nabla f(x_k) - \nabla^2 f(x_k)d_k\right\| \\ &\le \frac{L_H}{2}\|d_k\|^2 \\ &\le \frac{L_H}{2\mu^2}\|g_k\|^2 \\ &< \frac{L_H}{2\mu^2}\frac{3\mu^4}{L_H+\eta}\|g_k\| \\ &\le \frac{3}{2}\mu^2\|g_k\| \le \frac{3}{8}\|g_k\|,\end{aligned}$$

completing the proof. ∎

3. A variant with inexact directions. In section 2, we have assumed that certain linear-algebra operations in Algorithm 1—the linear system solves of (5) and (6) and the eigenvalue/eigenvector computation of (4)—are performed exactly. In a large-scale setting, the cost of these operations can be prohibitive, so iterative techniques that perform these operations *inexactly* are of interest. In this section, we describe inexact methods for these key operations and examine their consequences for the complexity analysis.

3.1. Inexact eigenvector calculation: Randomized Lanczos method. The problem of finding the minimum eigenvalue of the matrix in (4) and its associated eigenvector can be reformulated as one of finding the maximum eigenvalue and eigenvector of a positive semidefinite matrix. The Lanczos algorithm with a random starting vector is an appealing option for the latter problem, yielding an ϵ-approximate eigenvector in $\mathcal{O}(\log(n/\delta)\epsilon^{-1/2})$ iterations, with probability at least $1-\delta$ [15]. This fact has been used in several methods that achieve fast convergence rates [1, 3, 4]. In order to apply this method to a matrix that is not positive definite, one must make use of a bound on the Hessian norm. For sake of completeness, we spell out the procedure in the following lemma.

LEMMA 9. *Let H be a symmetric matrix satisfying $\|H\| \le M$ for some $M > 0$. Suppose that the Lanczos procedure is applied to find the largest eigenvalue of $MI - H$ starting at a random vector uniformly distributed over the unit sphere. Then, for any $\varepsilon > 0$ and $\delta \in (0,1)$, there is a probability of at least $1-\delta$ that the procedure outputs a unit vector v such that*

$$v^\top H v \le \lambda_{\min}(H) + \varepsilon \tag{40}$$

in at most

$$\min\left\{n, \frac{\ln(n/\delta^2)}{2\sqrt{2}}\sqrt{\frac{M}{\varepsilon}}\right\} \tag{41}$$

iterations.

After at most n iterations, the procedure obtains a unit vector v such that $v^\top Hv = \lambda_{\min}(H)$ *with probability* 1.

Proof. By definition, the matrix $H' = MI - H$ is a symmetric positive semidefinite matrix with its spectrum lying in $[0, 2M]$. Applying the Lanczos procedure to this matrix from a starting point drawn randomly from the unit sphere yields a unit vector v such that

$$v^\top H'v \geq \left(1 - \frac{\varepsilon}{2M}\right) \lambda_{\max}(H') \geq \left(1 - \frac{\varepsilon}{2M}\right) (M - \lambda_{\min}(H)) \tag{42}$$

in no more than $\min\{n, \ln(n/\delta^2)/4\sqrt{\varepsilon/(2M)}\}$ iterations with probability at least $1-\delta$. (This result is from [15, Theorem 4.2] extended by a continuity argument from the positive definite case to the positive semidefinite case; see [15, Remark 7.5].) Moreover, using (42), we have

$$\begin{aligned} v^\top Hv &= -v^\top H'v + M \\ &\leq -\left(1 - \frac{\varepsilon}{2M}\right)(M - \lambda_{\min}(H)) + M \\ &= -M + \lambda_{\min}(H) + \frac{\varepsilon}{2} - \frac{\varepsilon}{2M}\lambda_{\min}(H) + M \\ &= \lambda_{\min}(H) + \frac{\varepsilon}{2} - \frac{\varepsilon}{2M}\lambda_{\min}(H) \\ &\leq \lambda_{\min}(H) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2M}M \\ &= \lambda_{\min}(H) + \varepsilon, \end{aligned}$$

as required. ∎

Lemma 9 admits the following variant for the case in which we fix the number of Lanczos iterations.

LEMMA 10. *Let H be a symmetric matrix with* $\|H\| \leq M$. *Suppose that q iterations of the Lanczos procedure are applied to find the largest eigenvalue of* $MI - H$ *starting at a random vector uniformly distributed over the unit sphere. Then, for any* $\varepsilon > 0$, *the procedure outputs a unit vector v such that* $v^\top Hv \leq \lambda_{\min}(H) + \varepsilon$ *with probability at least*

$$1 - \delta \;=\; 1 - \sqrt{n}\exp\left[-\sqrt{2}q\sqrt{\frac{\varepsilon}{M}}\right]. \tag{43}$$

We point out that the choice $\delta = 0$ (or, equivalently, $q = n$) is possible, that is, after n iterations, the Lanczos procedure started with a random vector uniformly generated over the unit sphere returns an approximate eigenvector with probability one [15, Theorem 4.2(a)].

3.2. Inexact Newton and regularized Newton directions: Conjugate gradient method. Here we describe the use of the conjugate gradient (CG) algorithm to solve the symmetric positive definite linear systems (5) or (6)—the Newton and regularized Newton equations, respectively. The CG method is the most popular iterative method for positive definite linear systems due to its rich convergence theory and strong practical performance. It has also been popular in the context of nonconvex smooth minimization; see [19]. It requires only matrix-vector operations

involving the coefficient matrix (often these can be found or approximated without explicit knowledge of the matrix) together with some vector operations. It does not require knowledge or estimation of the extreme eigenvalues of the matrix.

We apply CG to a system $Hd = -g$ where there are positive quantities m and M such that $mI \preceq H \preceq MI$, so that the condition number κ of H is bounded above by M/m. Standard convergence theory indicates that CG outputs a vector d such that $\|Hy + g\| \leq \zeta\|g\|$ (for $\zeta \in (0,1)$) in

$$\mathcal{O}\left(\min\left\{n, \kappa^{1/2}\log(\kappa/\zeta)\right\}\right)$$

iterations, with κ being the condition number of H (we obtain the result as a corollary of Lemma 11 below). We use a different stopping criterion, namely,

$$\|Hd + g\| \leq \frac{1}{2}\zeta \min\{\|g\|m\|d\|\} \tag{44}$$

for some $\zeta \in (0,1)$. This criterion is stronger than the one typically used in truncated Newton–Krylov methods in that we require the residual norm to be bounded by a multiple of the norm of the approximate direction, as well as being bounded by a specified fraction of the initial residual norm. The extra criterion resembles the so-called *s-condition* arising in cubic regularization techniques, where the approximate minimizer s_k of the cubic model m_k is required to satisfy

$$\|\nabla m_k(s_k)\| \leq \mathcal{O}\left(\|s_k\|^2\right). \tag{45}$$

This property provides a lower bound on $\|s_k\|$, which is instrumental in obtaining the optimal complexity order of $\mathcal{O}(\epsilon_g^{-3/2})$ for first-order convergence [7]. Our condition replaces $\|s_k\|^2$ by $m\|d_k\|$, but serves a similar purpose.

The next lemma establishes a bound on the number of CG iterations needed to reach the desired accuracy.

LEMMA 11. *Let $Hd = -g$ be a linear system with H symmetric and $mI \preceq H \preceq MI$, where $m \in (0,1)$, $M > 0$, and $\|g\| > 0$. Then the conjugate gradient algorithm computes a vector d such that* (44) *holds for some $\zeta \in (0,1)$ in at most*

$$\min\left\{n, \tfrac{1}{2}\sqrt{\kappa}\ln\left(4\kappa^{3/2}/\zeta\right)\right\} \tag{46}$$

iterations, where $\kappa = M/m$.

Proof. Let $d^{(q)}$ be the iterate obtained at the qth iteration of the conjugate gradient method applied to $Hd = -g$, with $d^{(0)} = 0$. The classical bound on the behavior of the conjugate gradient residual [18, section 5.1] yields

$$\left\|d^{(q)} + H^{-1}g\right\|_H \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^q \|H^{-1}g\|_H, \tag{47}$$

where $\|x\|_H = \sqrt{x^\top Hx}$. From this definition and the bounds on the spectrum of H, we have

$$\begin{aligned}\|d^{(q)} + H^{-1}g\|_H^2 &= (d^{(q)} + H^{-1}g)^T H(d^{(q)} + H^{-1}g)\\ &= (Hd^{(q)} + g)^T H^{-1}(Hd^{(q)} + g) \geq \frac{1}{M}\|Hd^{(q)} + g\|^2,\end{aligned}$$

as well as

$$\|H^{-1}g\|_H^2 = g^T H^{-1} g \leq \frac{1}{m}\|g\|^2.$$

By substituting these bounds into (47), we obtain the following relation:

$$\left\|Hd^{(q)} + g\right\| \leq 2\kappa^{1/2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^q \|g\|. \tag{48}$$

Thus, as long as our stopping criterion is not satisfied, we have

$$\frac{1}{2}\zeta \min\left\{\|g\|, m\|d^{(q)}\|\right\} \leq 2\kappa^{1/2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^q \|g\|. \tag{49}$$

Furthermore, defining $r^{(q)} = Hd^{(q)} + g$, we have

$$\|d^{(q)}\| = \|H^{-1}(-g + r^{(q)})\| \geq \frac{\|g - r^{(q)}\|}{M} \geq \frac{\sqrt{\|g\|^2 - 2g^T r^{(q)} + \|r^{(q)}\|^2}}{M} \geq \frac{\|g\|}{M}$$

for all $q \geq 1$, where we used the facts that $r^{(0)} = g$ and that in CG the residuals are orthogonal: $(r^{(i)})^T r^{(j)} = 0$ for $i \neq j$. Using this bound within (49), we obtain

$$\frac{\zeta}{2}\min\{1, m/M\}\|g\| \leq 2\kappa^{1/2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^q \|g\| \quad \Leftrightarrow \quad \frac{\zeta}{4\kappa^{3/2}} \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^q.$$

By taking logarithms on both sides, we arrive at

$$q \leq \frac{\ln(\zeta/(4\kappa^{3/2}))}{\ln\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)} = \frac{\ln(4\kappa^{3/2}/\zeta)}{\ln\left(1 + \frac{2}{\sqrt{\kappa}-1}\right)} \leq \frac{1}{2}\sqrt{\kappa}\ln\left(\frac{4\kappa^{3/2}}{\zeta}\right),$$

where the bound $\ln(1 + \frac{1}{t}) \geq \frac{1}{t+1/2}$ was used to obtain the last inequality. ∎

3.3. Complexity analysis based on inexact computations. We present a variant of our main algorithm, specified as Algorithm 3, in which computation of approximate eigenvectors and linear system solves are performed inexactly by the means described above. Algorithm 3 requires two parameters not used in Algorithm 1: the upper bound U_H on the Hessian norms, defined in (9), and a probability threshold δ. As we expect only to recover inexact global complexity guarantees, the method does not exploit a local phase.

When Algorithm 3 terminates, condition (8) must hold. At termination, we have $\min(\|g_k\|, \|g_{k+1}\|) \leq \epsilon_g$ and $\lambda_k^i \geq -\frac{1}{2}\epsilon_H$. With high probability, λ_k^i is within $\frac{1}{2}\epsilon_H$ of $\lambda_{\min}(\nabla^2 f(x_k))$, so we must have $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_H$, thus satisfying (8).

Algorithm 3 Inexact second-order line-search method

Init. Choose $x^0 \in \mathbb{R}^n$, $\theta \in (0,1)$, $\zeta, \delta \in [0,1)$, $\eta > 0$, $\epsilon_g \in (0,1)$, $\epsilon_H \in (0,1)$, $U_H > 0$ satisfying (9);
for $k = 0, 1, 2, \ldots$ **do**
  Step 1. (First Order) Set $g_k = \nabla f(x_k)$;
  if $\|g_k\| = 0$ **then**
    Go to Step 2;
  end if
  Compute $R_k = \frac{g_k^\top \nabla^2 f(x_k) g_k}{\|g_k\|^2}$;
  if $R_k < -\epsilon_H$ **then**
    Set $d_k = \frac{R_k}{\|g_k\|} g_k$ and go to Step LS;
  else if $R_k \in [-\epsilon_H, \epsilon_H]$ and $\|g_k\| > \epsilon_g$ **then**
    Set $d_k = -\frac{g_k}{\|g_k\|^{1/2}}$ and go to Step LS;
  else
    Go to Step 2;
  end if
  Step 2. (Inexact Second Order) Compute an inexact eigenvector v_k^i such that (with probability $1 - \delta$)

$$[v_k^i]^\top g_k \le 0, \quad \|v_k^i\| = |\lambda_k^i|, \quad \lambda_k^i \le \lambda_{\min}\left(\nabla^2 f(x_k)\right) + \tfrac{1}{2}\epsilon_H, \tag{50}$$

where $\lambda_k^i = [v_k^i]^\top \nabla^2 f(x_k) v_k^i / \|v_k^i\|^2$;
  if $\|g_k\| \le \epsilon_g$ and $\lambda_k^i \ge -\frac{1}{2}\epsilon_H$ **then**
    Terminate;
  else if $\lambda_k^i < -\frac{1}{2}\epsilon_H$ **then**
    (Negative Curvature) Set $d_k = v_k^i$;
  else if $\lambda_k^i > \frac{3}{2}\epsilon_H$ **then**
    (Inexact Newton) Use conjugate gradient to calculate $d_k = d_k^{in}$, where

$$\|\nabla^2 f(x_k) d_k^{in} + g_k\| \le \tfrac{1}{2}\zeta \min\left\{\|g_k\|, \epsilon_H \|d_k^{in}\|\right\}; \tag{51}$$

  else
    (Inexact Regularized Newton) Use conjugate gradient to calculate $d_k = d_k^{ir}$, where

$$\|(\nabla^2 f(x_k) + 2\epsilon_H I) d_k^{ir} + g_k\| \le \tfrac{1}{2}\zeta \min\left\{\|g_k\|, \epsilon_H \|d_k^{ir}\|\right\}; \tag{52}$$

  end if
  Go to Step LS;
  Step LS. (Line Search) Compute a step length $\alpha_k = \theta^{j_k}$, where j_k is the smallest nonnegative integer such that

$$f(x_k + \alpha_k d_k) < f(x_k) - \tfrac{1}{6}\eta \alpha_k^3 \|d_k\|^3 \tag{53}$$

holds, and set $x_{k+1} = x_k + \alpha_k d_k$;
  if $d_k = d_k^{in}$ or $d_k = d_k^{ir}$ and $\|\nabla f(x_{k+1})\| \le \epsilon_g$ **then**
    Terminate;
  end if
end for

TABLE 2
Steps and associated decrease lemmas for Algorithm 3.

Context			Direction	Decrease
$\|g_k\| = 0$	-	$\lambda_k^i < -\frac{\epsilon_H}{2}$	v_k^i	Lemma 1
	$R_k < -\epsilon_H$		$R_k g_k/\|g_k\|$	Lemma 1
$\|g_k\| > \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$		$-g_k/\|g_k\|^{1/2}$	Lemma 2
$\|g_k\| \le \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$	$\lambda_k^i < -\frac{1}{2}\epsilon_H$	v_k^i	Lemma 1
$\|g_k\| \le \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$	$\lambda_k^i \in [-\frac{1}{2}\epsilon_H, \frac{3}{2}\epsilon_H]$	d_k^{ir}	Lemma 13
$\|g_k\| \le \epsilon_g$	$R_k \in [-\epsilon_H, \epsilon_H]$	$\lambda_k^i > \frac{3}{2}\epsilon_H$	d_k^{in}	Lemma 12
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k^i < -\frac{1}{2}\epsilon_H$	v_k^i	Lemma 1
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k^i \in [-\frac{1}{2}\epsilon_H, \frac{3}{2}\epsilon_H]$	d_k^{ir}	Lemma 13
$\|g_k\| > \epsilon_g$	$R_k > \epsilon_H$	$\lambda_k^i > \frac{3}{2}\epsilon_H$	d_k^{in}	Lemma 12

Table 2 shows a summary of the possible choices for the search direction. It shows the same number of cases as Table 1, with the context now determined by the eigenvalue estimate λ_k^i, with one exception. There is an extra row for the case in which $\|g_k\| \le \epsilon_g$, $R_k \in [-\epsilon_H, \epsilon_H]$, $\lambda_k^i > \frac{3}{2}\epsilon_H$, because of possible (but low-probability) failure of the randomized Lanczos process to detect the smallest eigenvalue of $\nabla^2 f(x_k)$ to the required accuracy. Table 2 mentions two additional lemmas, which respectively replace Lemmas 3 and 4 in order to take inexactness into account. We state and prove these results next.

LEMMA 12. *Let Assumptions* 1 *and* 2 *hold. Suppose that an inexact Newton direction $d_k = d_k^{in}$ is computed at the kth iteration of Algorithm* 3. *Then, with probability at least $1-\delta$, the backtracking line search terminates with step length $\alpha_k = \theta^{j_k}$, with $j_k \le j_{inr} + 1$, where*

$$j_{inr} := \left\lfloor \frac{1}{2} \log_\theta \left(\frac{3}{L_H + \eta} \frac{(1-\zeta)\epsilon_H^2}{U_g \sqrt{1+\zeta^2/4}} \right) \right\rfloor_+, \tag{54}$$

and we have

$$f(x_k) - f(x_k + \alpha_k d_k) \;\ge\; c_{in} \min\left\{ \|\nabla f(x_k + \alpha_k d_k)\|^3 \epsilon_H^{-3}, \epsilon_H^3 \right\}, \tag{55}$$

where

$$c_{in} := \frac{\eta}{6} \min\left\{ \left[\frac{4}{\zeta + \sqrt{\zeta^2 + 8L_H}} \right]^3, \left[\frac{3\theta^2(1-\zeta)}{L_H + \eta} \right]^3 \right\}.$$

Proof. We observe first that when the Newton step is computed in Algorithm 3 we have from (40) that

$$\frac{3\epsilon_H}{2} < \lambda_k^i \le \lambda_{\min}\left(\nabla^2 f(x_k)\right) + \frac{\epsilon_H}{2} \;\Rightarrow\; \lambda_{\min}\left(\nabla^2 f(x_k)\right) \ge \epsilon_H,$$

with probability $1-\delta$. Suppose first that the step length $\alpha_k = 1$ satisfies the decrease condition (53). Then, defining

$$r_k := \nabla^2 f(x_k) d_k + g_k \tag{56}$$

and using the inexactness criterion for the inexact Newton step d_k, we find that the gradient at the next point $x_k + d_k$ satisfies

$$\begin{aligned}\|\nabla f(x_k + d_k)\| &= \|\nabla f(x_k + d_k) - \nabla f(x_k) + \nabla f(x_k)\| \\ &= \left\|\nabla f(x_k + d_k) - \nabla f(x_k) - \nabla^2 f(x_k) d_k + r_k\right\| \\ &\leq \frac{L_H}{2}\|d_k\|^2 + \|r_k\| \leq \frac{L_H}{2}\|d_k\|^2 + \frac{\zeta}{2}\epsilon_H \|d_k\|.\end{aligned}$$

We obtain a lower bound on $\|d_k\|$ by taking the root of the above quadratic and applying Lemma 17 with $a = \zeta\epsilon_H/2$, $b = 2L_H\epsilon_H^2$, and $t = \|\nabla f(x_k + d_k)\|/\epsilon_H^2$ to obtain

$$\begin{aligned}\|d_k\| &\geq \frac{-\frac{\zeta}{2}\epsilon_H + \sqrt{\frac{\zeta^2}{4}\epsilon_H^2 + 2L_H\|\nabla f(x_k + d_k)\|}}{L_H} \\ &\geq \frac{-\frac{\zeta}{2}\epsilon_H + \sqrt{\frac{\zeta^2}{4}\epsilon_H^2 + 2L_H\epsilon_H^2}}{L_H} \min\left(\|\nabla f(x_k + d_k)\|/\epsilon_H^2, 1\right) \\ &= \frac{-\zeta + \sqrt{\zeta^2 + 8L_H}}{2L_H} \min\left(\|\nabla f(x_k + d_k)\|/\epsilon_H, \epsilon_H\right) \\ &= \frac{4}{\zeta + \sqrt{\zeta^2 + 8L_H}} \min\left(\|\nabla f(x_k + d_k)\|/\epsilon_H, \epsilon_H\right). \qquad (57)\end{aligned}$$

Therefore, taking the inexact Newton step with a unit step length guarantees

$$f(x_k) - f(x_k + d_k) \geq \frac{\eta}{6}\|d_k\|^3 \geq \frac{\eta}{6}\left[\frac{4}{\zeta + \sqrt{\zeta^2 + 8L_H}}\right]^3 \min\left(\|\nabla f(x_k + d_k)\|^3 \epsilon_H^{-3}, \epsilon_H^3\right),$$

so the inequality (55) is satisfied in the case of a unit step $\alpha_k = 1$.

To complete the proof, consider the case in which the unit step length does not lead to sufficient decrease. In that case, for any value $j \geq 0$ such that (53) is not satisfied, we have

$$\begin{aligned}-\frac{\eta}{6}\theta^{3j}\|d_k\|^3 &\leq f(x_k + \theta^j d_k) - f(x_k) \\ &\leq \theta^j g_k^\top d_k + \frac{\theta^{2j}}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\ &\leq \theta^j(-\nabla^2 f(x_k)d_k + r_k)^\top d_k + \frac{\theta^{2j}}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\ &= -\theta^j\left(1 - \frac{\theta^j}{2}\right) d_k^\top \nabla^2 f(x_k) d_k + \theta^j d_k^\top r_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\ &\leq -\frac{\theta^j}{2}\epsilon_H\|d_k\|^2 + \theta^j\|d_k\|\|r_k\| + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\ &\leq -\frac{\theta^j}{2}(1-\zeta)\epsilon_H\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3.\end{aligned}$$

Thus, for any $j \geq 0$ for which sufficient decrease is not obtained, we have

$$\theta^{2j} \geq \frac{3}{L_H + \eta}(1-\zeta)\epsilon_H\|d_k\|^{-1}. \qquad (58)$$

In particular, since (58) holds for $j = 0$, we have

$$\|d_k\| \;\geq\; \frac{3}{L_H + \eta}(1-\zeta)\epsilon_H. \tag{59}$$

By the definitions of d_k and of r_k in (56), we also have the upper bound on its norm

$$\begin{aligned}\|d_k\| = \left\|\nabla^2 f(x_k)^{-1}\left(g_k - r_k\right)\right\| \leq \|\nabla^2 f(x_k)^{-1}\|\,\|g_k - r_k\| &\leq \frac{1}{\epsilon_H}\sqrt{\|g_k\|^2 + \|r_k\|^2} \\ &\leq \frac{\sqrt{1+\zeta^2/4}}{\epsilon_H}\|g_k\| \\ &\leq \frac{\sqrt{1+\zeta^2/4}}{\epsilon_H}U_g,\end{aligned}$$

using again the fact that g_k and r_k are orthogonal (by the properties of the CG algorithm), as well as the criterion (51) and the bound (9).

Meanwhile, for any $j > j_{inr}$, we have

$$\begin{aligned}\theta^{2j} < \theta^{2j_{inr}} \leq \frac{3}{L_H+\eta}(1-\zeta)\frac{\epsilon_H^2}{U_g\sqrt{1+\zeta^2/4}} &\leq \frac{3}{L_H+\eta}(1-\zeta)\epsilon_H\frac{\epsilon_H}{U_g\sqrt{1+\zeta^2/4}} \\ &\leq \frac{3}{L_H+\eta}(1-\zeta)\epsilon_H\|d_k\|^{-1}.\end{aligned}$$

As a result, (58) is violated for $j > j_{inr}$, which means that the line search must terminate with a step length $\alpha_k = \theta^{j_k}$ satisfying (53), with $1 \leq j_k \leq j_{inr}+1$. Since the index $j = j_k - 1 \geq 0$ satisfies (58), we have

$$\theta^{j_k} \geq \sqrt{\frac{3\theta^2}{L_H+\eta}(1-\zeta)}\epsilon_H^{1/2}\|d_k\|^{-1/2}, \tag{60}$$

and from the sufficient decrease condition we have

$$\begin{aligned}f(x_k) - f(x_k + \theta^{j_k}d_k) \geq \frac{\eta}{6}\theta^{3j_k}\|d_k\|^3 &\geq \frac{\eta}{6}\left[\frac{3\theta^2}{L_H+\eta}(1-\zeta)\right]^{3/2}\epsilon_H^{3/2}\|d_k\|^{3/2} \\ &\geq \frac{\eta}{6}\left[\frac{3\theta^2}{L_H+\eta}(1-\zeta)\right]^{3}\epsilon_H^{3},\end{aligned}$$

where the second inequality follows from (60) and the third inequality follows from (59) (using the fact that $\theta \in (0,1)$). Hence, the claim (55) is satisfied in the case of nonunit step length α_k too, and the proof is complete. ∎

LEMMA 13. *Let Assumptions* 1 *and* 2 *hold. Suppose that an inexact regularized Newton direction $d_k = d_k^{ir}$ is computed at the kth iteration of Algorithm* 3*. Then, with probability at least $1-\delta$, the backtracking line search terminates with step length $\alpha_k = \theta^{j_k}$, with $j_k \leq j_{inr}+1$, where j_{inr} is defined as in* (54)*, and we have*

$$f(x_k) - f(x_k + \alpha_k d_k) \;\geq\; c_{ir}\min\left\{\|\nabla f(x_k+\alpha_k d_k)\|^3\epsilon_H^{-3}, \epsilon_H^3\right\}, \tag{61}$$

where

$$c_{ir} := \frac{\eta}{6}\min\left\{\left[\frac{4}{4+\zeta+\sqrt{(4+\zeta)^2+8L_H}}\right]^3, \left[\frac{3\theta^2(1-\zeta)}{L_H+\eta}\right]^3\right\}.$$

Proof. The inexact regularized Newton step is computed only when $-\frac{1}{2}\epsilon_H \leq \lambda_k^i \leq \frac{3}{2}\epsilon_H$, so from (40) with $\varepsilon = \frac{1}{2}\epsilon_H$, we have

$$\lambda_{\min}(\nabla^2 f(x_k)) + 2\epsilon_H \geq \lambda_k^i - \tfrac{1}{2}\epsilon_H + 2\epsilon_H \geq \epsilon_H, \tag{62}$$

with probability at least $1-\delta$. Suppose first that the step length $\alpha_k = 1$ satisfies the decrease condition (53). Then, defining $r_k = (\nabla^2 f(x_k) + 2\epsilon_H I)d_k + g_k$, we have that

$$\begin{aligned}
\|\nabla f(x_k + d_k)\| &= \|\nabla f(x_k + d_k) - \nabla f(x_k) + \nabla f(x_k)\| \\
&= \left\|\nabla f(x_k + d_k) - \nabla f(x_k) - \nabla^2 f(x_k)d_k - 2\epsilon_H d_k + r_k\right\| \\
&\leq \frac{L_H}{2}\|d_k\|^2 + 2\epsilon_H\|d_k\| + \|r_k\| \\
&\leq \frac{L_H}{2}\|d_k\|^2 + \frac{4+\zeta}{2}\epsilon_H\|d_k\|.
\end{aligned}$$

Reasoning as in (57), with $\frac{4+\zeta}{2}$ replacing $\frac{\zeta}{2}$, we obtain the following lower bound on $\|d_k\|$:

$$\|d_k\| \;\geq\; \frac{4}{4+\zeta+\sqrt{(4+\zeta)^2+8L_H}} \min\left(\|\nabla f(x_k+d_k)\|\epsilon_H^{-1}, \epsilon_H\right). \tag{63}$$

Therefore, taking the unit regularized Newton step guarantees that

$$\begin{aligned}
f(x_k) - f(x_k + d_k) &\geq \frac{\eta}{6}\|d_k\|^3 \\
&\geq \frac{\eta}{6}\left[\frac{4}{4+\zeta+\sqrt{(4+\zeta)^2+8L_H}}\right]^3 \min\left(\|\nabla f(x_k+d_k)\|^3\epsilon_H^{-3}, \epsilon_H^3\right),
\end{aligned}$$

so the result of the theorem holds in the case in which the unit step satisfies the sufficient decrease condition.

To complete the proof, we consider the case in which $\alpha_k < 1$. In that case, for any value $j \geq 0$ such that (53) is not satisfied, we have from the definition of r_k, the bound on $\|r_k\|$ in the definition of d_k^{ir}, and (62) that

$$\begin{aligned}
-\frac{\eta}{6}\theta^{3j}\|d_k\|^3 &\leq f(x_k + \theta^j d_k) - f(x_k) \\
&\leq \theta^j g_k^\top d_k + \frac{\theta^{2j}}{2} d_k^\top \nabla^2 f(x_k) d_k + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&= -\theta^j \left[\nabla^2 f(x_k)d_k + 2\epsilon_H d_k - r_k\right]^\top d_k + \frac{\theta^{2j}}{2} d_k^\top \nabla^2 f(x_k) d_k \\
&\qquad + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&= -\theta^j\left(1 - \frac{\theta^j}{2}\right) d_k^\top [\nabla^2 f(x_k) + 2\epsilon_H I] d_k - \theta^{2j}\epsilon_H\|d_k\|^2 + \theta^j r_k^\top d_k \\
&\qquad + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&\leq -\frac{\theta^j}{2}\epsilon_H\|d_k\|^2 - \theta^{2j}\epsilon_H\|d_k\|^2 + \theta^j\|r_k\|\|d_k\| + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&\leq -\frac{\theta^j}{2}\epsilon_H\|d_k\|^2 + \theta^j\frac{\zeta}{2}\epsilon_H\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3 \\
&= -\frac{\theta^j}{2}(1-\zeta)\epsilon_H\|d_k\|^2 + \frac{L_H}{6}\theta^{3j}\|d_k\|^3.
\end{aligned}$$

Thus, for any $j \geq 0$ for which sufficient decrease is not obtained, one has

$$\theta^{2j} \geq \frac{3}{L_H + \eta}(1-\zeta)\epsilon_H \|d_k\|^{-1}. \tag{64}$$

In particular, setting $j = 0$ in this expression, we obtain

$$\|d_k\| \geq \frac{3}{L_H + \eta}(1-\zeta)\epsilon_H. \tag{65}$$

The right-hand side of (64) is bounded below, since

$$\begin{aligned} \|d_k\| = \left\|[\nabla^2 f(x_k) + 2\epsilon_H I]^{-1}(-g_k + r_k)\right\| &\leq \|[\nabla^2 f(x_k) + 2\epsilon_H]^{-1}\| \, \|-g_k + r_k\| \\ &\leq \frac{1}{\epsilon_H}\sqrt{\|g_k\|^2 + \|r_k\|^2} \\ &\leq \frac{\sqrt{1+\zeta^2/4}}{\epsilon_H}\|g_k\| \\ &\leq \frac{\sqrt{1+\zeta^2/4}}{\epsilon_H}U_g, \end{aligned} \tag{66}$$

where we used again the orthogonality of g_k and r_k (from the properties of conjugate gradient) as well as the condition (52). For any $j > j_{inr}$, we have

$$\theta^{2j} < \theta^{2j_{inr}} \leq \frac{3(1-\zeta)}{L_H + \eta}\frac{\epsilon_H^2}{U_g\sqrt{1+\zeta^2/4}} \leq \frac{3(1-\zeta)}{L_H + \eta}\epsilon_H \|d_k\|^{-1},$$

where the last inequality follows from (66). Therefore, (64) is violated for $j > j_{inr}$, which means that the line search must terminate with a step length $\alpha_k = \theta^{j_k}$, with $1 \leq j_k \leq j_{inr} + 1$. The previous index $j = j_k - 1$ satisfies (64), so we have

$$\theta^{2j_k} \geq \frac{3\theta^2}{L_H + \eta}(1-\zeta)\epsilon_H \|d_k\|^{-1},$$

so that

$$\begin{aligned} f(x_k) - f(x_k + \theta^{j_k} d_k) &\geq \frac{\eta}{6}\theta^{3j_k}\|d_k\|^3 \\ &\geq \frac{\eta}{6}\left[\frac{3\theta^2}{L_H + \eta}(1-\zeta)\right]^{3/2}\epsilon_H^{3/2}\|d_k\|^{3/2} \\ &\geq \frac{\eta}{6}\left[\frac{3\theta^2}{L_H + \eta}(1-\zeta)\right]^{3}\epsilon_H^{3}, \end{aligned}$$

where the final inequality follows from (65), using the fact that $\theta \in (0,1)$. Thus, condition(61) also holds in the case in which $\alpha_k < 1$, and the proof is complete. ∎

THEOREM 14. *Let Assumptions* 1 *and* 2 *hold. Then, Algorithm* 3 *returns a point* x_k *satisfying* (2) *in at most*

$$\hat{K} := \hat{\mathcal{C}} \max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_g^{-3/2}, \epsilon_H^{-3}\right\} \tag{67}$$

iterations, where

$$\hat{\mathcal{C}} := \frac{f(x_0) - f_{\text{low}}}{\hat{c}}, \quad \hat{c} := \min\left\{\frac{c_e}{8}, c_g, c_{in}, c_{ir}\right\},$$

with probability at least $1 - \hat{K}\delta$. *The constants* c_e, c_g, c_{in}, *and* c_{ir} *are defined in Lemmas* 1, 2, 12, *and* 13, *respectively.*

Proof. For any iteration l such that x_l does not satisfy (8), we must have that either $\min(\|g_l\|, \|g_{l+1}\|) > \epsilon_g$ or $\lambda_{\min}(\nabla^2 f(x_l)) < -\epsilon_H$, where the latter implies that $\lambda_l^i < -\frac{1}{2}\epsilon_H$. Thus, similarly to the proof of Theorem 5, we can consider the following two cases.

Case 1: $\boldsymbol{\lambda_l^i < -\frac{1}{2}\epsilon_H}$**.** From Table 2, we see that the same three choices for d_l as in the exact version are possible. If $d_l = \frac{R_l}{\|g_l\|} g_l$, we have exactly as in Lemma 1 that

$$f(x_l) - f(x_{l+1}) \;\geq\; c_e \epsilon_H^3.$$

When $d_l = -g_l/\|g_l\|^{1/2}$, we have from Lemma 2 that

$$f(x_l) - f(x_{l+1}) \;\geq\; c_g \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_g^{3/2}\right\}. \tag{68}$$

The remaining case corresponds to the choice $d_l = v_l^i$. Since

$$\frac{d_l^\top \nabla^2 f(x_l) d_l}{\|d_l\|^2} = \lambda_l^i < -\frac{\epsilon_H}{2},$$

with probability at least $1 - \delta$ in that case, one can use the result of Lemma 1 to deduce that

$$f(x_l) - f(x_{l+1}) \geq c_e |\lambda_l^i|^3 \geq \frac{c_e}{8} \epsilon_H^3,$$

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

Case 2: $\boldsymbol{\lambda_l^i \geq -\frac{1}{2}\epsilon_H}$**,** $\boldsymbol{\|g_l\| > \epsilon_g}$**, and** $\boldsymbol{\|g_{l+1}\| > \epsilon_g}$**.** In this situation, we have three possible choices of search direction d_l. If $d_l = -g_l/\|g_l\|^{1/2}$, we have again from Lemma 2 that (68) holds. If the inexact Newton direction is taken, we obtain by Lemma 12 that

$$f(x_k) - f(x_{k+1}) \;\geq\; c_{in} \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_H^3\right\}.$$

Finally, if the search direction is the inexact regularized Newton direction, that is, $d_l = d_l^{ir}$, we have from Lemma 13 that

$$f(x_k) - f(x_{k+1}) \;\geq\; c_{ir} \min\left\{\epsilon_g^3 \epsilon_H^{-3}, \epsilon_H^3\right\}.$$

By putting all these bounds together, as in the proof of Theorem 5, we obtain that the number of iterations before reaching a point satisfying (8) is bounded above by $\hat{K}$ defined in the statement of the theorem.

Recalling that for each of these iterations there is a probability δ that the randomized Lanczos iteration in (50) will fail, we bound the probability of failure during the course of the algorithm by $\hat{K}\delta$. ∎

Note that if δ is chosen large enough such that $1 - \hat{K}\delta < 0$, Theorem 14 is not informative. The same remark holds for the corollary below, which makes use of the results from sections 3.1 and 3.2 to obtain a bound on the total number of Hessian-vector multiplications and gradient evaluations needed by the procedure (assuming that these operations cost roughly the same).

COROLLARY 15. *Suppose the assumptions of Theorem* 14 *hold and let* $\delta \in (0,1)$ *be given. Then the total number of gradient evaluations and Hessian-vector multiplications required by Algorithm* 3 *to reach an iterate satisfying* (8) *is bounded by*

$$\left[2 + \min\left\{n, \tfrac{1}{\sqrt{2}}(U_H+2)^{1/2}\epsilon_H^{-1/2}\ln\left(\tfrac{4(U_H+2)^{3/2}\epsilon_H^{-3/2}}{\zeta}\right)\right\} + \min\left\{n, (U_H+2)^{1/2}\epsilon_H^{-1/2}\tfrac{\ln(n/\delta^2)}{2}\right\}\right] \times \hat{K}, \tag{69}$$

with probability $1 - \hat{K}\delta$.

Proof. The proof follows directly from Lemmas 9 and 11, setting $M = U_H + 2$ and $\varepsilon = \epsilon_H/2$, noting that for both Newton and regularized Newton steps the condition number of the respective coefficient matrices can be bounded by $(U_H+2)/\epsilon_H$. ∎

As in section 2.2, we can particularize this result to a specific choice of tolerances.

COROLLARY 16. *Suppose that the assumptions of Theorem* 14 *hold and let* $\delta \in (0,1)$ *be given. Define* $\epsilon_g = \epsilon$ *and* $\epsilon_H = \sqrt{\epsilon}$ *for some* $\epsilon \in (0,1)$. *Then the number of gradient evaluations and Hessian-vector products needed in Algorithm* 3 *to satisfy* (8) *is bounded by*

$$\left[2 + \min\left\{n, \tfrac{1}{\sqrt{2}}(U_H+2)^{1/2}\epsilon^{-1/4}\ln\left(\tfrac{4(U_H+2)^{3/2}\epsilon^{-3/4}}{\zeta}\right)\right\} + \min\left\{n, (U_H+2)^{1/2}\epsilon^{-1/4}\tfrac{\ln(n/\delta^2)}{2}\right\}\right] \times \hat{C}\epsilon^{-3/2}, \tag{70}$$

where $\hat{C}$ *is defined as in Theorem* 14, *with probability at least* $1 - \hat{C}\epsilon^{-3/2}\delta$.

This result is meaningful when $\delta \ll \epsilon^{3/2}$. In terms of the complexity bound, such a choice is not prohibitively small because δ enters into the bound (70) only inside a log term.

We can obtain a bound for the case in which $\delta = 0$ (that is, almost certainty), at the cost of taking n Lanczos iterations whenever the smallest eigenvalue is needed (see Lemma 9). In this case, the bound (70) either becomes $\mathcal{O}((n + \ln(\epsilon^{-1}))\epsilon^{-7/4})$ or $\mathcal{O}(n\epsilon^{-3/2})$, depending on which term dominates in the quantity corresponding to conjugate gradient iterations.

For very large n and $\delta > 0$, we can consider that the term involving ϵ is smaller than n in both minimum expressions in Corollary 16. In this case, the bound is

$$\mathcal{O}\left(\ln\left(\frac{1}{\min\{\epsilon, \delta/\sqrt{n}\}}\right)\epsilon^{-7/4}\right).$$

This complexity matches other recent findings [1, 4].

In terms of dependencies with respect to problem constants, we can reproduce the analysis from section 2.2, replacing c_n and c_r by $c_{in} = \mathcal{O}(L_H^{-3})$ and $c_{ir} = \mathcal{O}(L_H^{-3})$, respectively. For instance, the bound from the previous paragraph is in

$$\mathcal{O}\left((f(x_0) - f_{\text{low}})U_H^{1/2}L_H^3\ln\left(\frac{U_H}{\min\{\epsilon, \delta/\sqrt{n}\}}\right)\epsilon^{-7/4}\right).$$

We point out that the dependency on L_H of our bound is worse than those of [1, 4] due to the lack of explicit use of this constant within our algorithm. Still, we believe

our dependency to match that of other Newton-type methods (although those are not emphasized in the related literature) and we consider such schemes as being more amenable to highly nonlinear settings where estimating such a constant would likely be impractical.

As a final note, we observe that one could also include the number of line-search iterations into our complexity bound. However, this cost is essentially logarithmic in $1/\epsilon$, therefore it is dominated by the cost of the linear algebra techniques.

4. Discussion. Among the many algorithmic frameworks that have been proposed for smooth nonconvex optimization with second-order complexity guarantees, it can be difficult to determine the algorithmic features that affect the complexity analysis or to understand how the guarantees provided by different algorithms relate to one another. We have presented a second-order complexity analysis of a framework that is based exclusively on line searches along certain directions. It does not require solution of cubic-regularized or trust-region subproblems, or minimization of convexified functions—operations that are needed by other approaches. Our search directions are of several types—gradient, negative-curvature, Newton, and regularized Newton—and we presented a variant of our method that allows inexact direction computation using iterative methods. We believe that ours is the first approach of line-search type to achieve known optimal complexity, among methods that identify points that satisfy approximate second-order necessary conditions.

In addition to the results of this paper, we observe that it is possible to modify our algorithms to attain points that satisfy termination conditions of the form (2) (rather than (8)) by continuing to iterate in the situation in which $\|g_{k+1}\| \leq \epsilon_g$ but $\lambda_{min}(\nabla^2 f(x_{k+1})) < -\epsilon_H$. Step $k+1$ then yields a decrease that is a multiple of ϵ_H^3 (per Lemma 1), so the overall complexity estimates are preserved, even if step k in this situation fails to produce a significant decrease in f.

In designing the framework of Algorithms 1 and 3, we have made some choices to give preference to one direction choice over another and we have also incorporated several types of steps. Given the recent literature in this area, our proposed scheme is actually one particular instance of a broader class of methods with similar complexity guarantees but possibly diverse practical performance. An implementation of our approach would raise several delicate issues, for example, issues associated with failure of the randomized Lanczos procedure for obtaining an estimate of the smallest eigenvalue. An incorrect estimate here could lead to the conjugate gradient method subsequently being applied to an indefinite matrix; a robust implementation would need to detect and recover from such an occurrence. Additionally, the choice of suitable values for the bound on the Hessian norm is likely to be of critical importance. Addressing these concerns in the aim of developing a practical algorithm with good complexity guarantees is the subject of ongoing research.

Appendix A. Technical result. We prove a technical result that is used in several proofs, including that of Lemma 4.

LEMMA 17. *For positive scalars a and b, and $t \geq 0$, we have*

$$-a + \sqrt{a^2 + bt} \geq (-a + \sqrt{a^2 + b}) \min(t, 1).$$

Proof. For the case in which $t \geq 1$, we have

$$-a + \sqrt{a^2 + bt} \geq -a + \sqrt{a^2 + b} = (-a + \sqrt{a^2 + b}) \min(t, 1),$$

so the result holds in this case. For $t \in (0,1)$, we need to show that

$$-a + \sqrt{a^2 + bt} \geq (-a + \sqrt{a^2 + b})t.$$

This claim follows from the chain of equivalences

$$\begin{aligned}
& -a + \sqrt{a^2 + bt} \geq (-a + \sqrt{a^2 + b})t \\
\Leftrightarrow \quad & \sqrt{a^2 + bt} \geq (-a + \sqrt{a^2 + b})t + a \\
\Leftrightarrow \quad & a^2 + bt \geq (-a + \sqrt{a^2 + b})^2 t^2 + 2a(-a + \sqrt{a^2 + b})t + a^2 \\
\Leftrightarrow \quad & (b + 2a^2 - 2a\sqrt{a^2 + b})t \geq (-a + \sqrt{a^2 + b})^2 t^2 \\
\Leftrightarrow \quad & (-a + \sqrt{a^2 + b})^2 t \geq (-a + \sqrt{a^2 + b})^2 t^2 \\
\Leftrightarrow \quad & 1 \geq t,
\end{aligned}$$

thereby completing the proof. ∎

Acknowledgments. We are grateful to the anonymous referees and associate editor of the original version of the paper, whose constructive comments led to numerous improvements.

REFERENCES

[1] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, *Finding approximate local minima faster than gradient descent*, in Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Association for Computing Machinery (ACM), New York, 2017, pp. 1195–1199.

[2] E. G. Birgin and J. M. Martínez, *The use of quadratic regularization with a cubic descent condition for unconstrained optimization*, SIAM J. Optim., 27 (2017), pp. 1049–1074.

[3] Y. Carmon and J. C. Duchi, *Gradient Descent Efficiently Finds the Cubic-Regularized Non-Convex Newton Step*, preprint, arXiv:1612.00547v2, 2017.

[4] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Accelerated Methods for Non-Convex Optimization*, preprint, arXiv:1611.00756v2, 2017.

[5] C. Cartis, N. I. M. Gould, and P. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems*, SIAM J. Optim., 20 (2010), pp. 2833–2852.

[6] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* I*: Motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245–295.

[7] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* II*: Worst-case function- and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295–319.

[8] C. Cartis, N. I. M. Gould, and P. L. Toint, *Optimal Newton-Type Methods for Nonconvex Optimization*, Technical report naXys-17-2011, Department of Mathematics, University of Namur, Belgium, 2011.

[9] C. Cartis, N. I. M. Gould, and P. L. Toint, *Complexity bounds for second-order optimality in unconstrained optimization*, J. Complexity, 28 (2012), pp. 93–108.

[10] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim., SIAM, Philadelphia, 2000.

[11] F. E. Curtis, D. P. Robinson, and M. Samadi, *A trust region algorithm with a worst-case iteration complexity of* $O(\epsilon^{-3/2})$ *for nonconvex optimization*, Math. Program., 162 (2017), pp. 1–32.

[12] G. N. Grapiglia, J. Yuan, and Y.-X. Yuan, *Nonlinear stepsize control algorithms: Complexity bounds for first- and second-order optimality*, J. Optim. Theory Appl., 171 (2016), pp. 980–997.

[13] S. Gratton, C. W. Royer, and L. N. Vicente, *A Decoupled First/Second-Order Steps Technique for Nonconvex Nonlinear Unconstrained Optimization with Improved Complexity Bounds*, Technical report 17-21, Department of Mathematics, University of Coimbra, Portugal, 2017.

[14] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, in Proceedings of the 34th International Conference on Machine Learning, Proc. Mach. Learn. Res. 70, PMLR, 2017, pp. 1724–1732.

[15] J. Kuczyński and H. Woźniakowski, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122.

[16] J. M. Martínez and M. Raydan, *Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization*, J. Global Optim., 68 (2017), pp. 367–385.

[17] Y. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2006.

[19] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.