

GRADIENT DESCENT FINDS THE CUBIC-REGULARIZED NONCONVEX NEWTON STEP*

YAIR CARMON[†] AND JOHN DUCHI[‡]

Abstract. We consider the minimization of a nonconvex quadratic form regularized by a cubic term, which may exhibit saddle points and a suboptimal local minimum. Nonetheless, we prove that, under mild assumptions, gradient descent approximates the *global minimum* to within ε accuracy in $O(\varepsilon^{-1} \log(1/\varepsilon))$ steps for large ε and $O(\log(1/\varepsilon))$ steps for small ε (compared to a condition number we define), with at most logarithmic dependence on the problem dimension. When we use gradient descent to approximate the cubic-regularized Newton step, our result implies a rate of convergence to second-order stationary points of general smooth nonconvex functions.

Key words. gradient descent, nonconvex quadratics, cubic regularization, global optimization, Newton's method, nonasymptotic rate of convergence, power method, trust region methods

AMS subject classifications. 65K05, 90C06, 90C20, 90C26, 90C30

DOI. 10.1137/17M1113898

1. Introduction. We study the optimization problem

$$(1) \quad \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) \triangleq \frac{1}{2} x^T A x + b^T x + \frac{\rho}{3} \|x\|^3,$$

where the matrix A is symmetric and possibly indefinite. Problem (1) arises in Newton's method with cubic regularization, for (approximately) minimizing a general smooth function g . The method consists of the iterative procedure

$$(2) \quad y_{t+1} = y_t + \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \nabla g(y_t)^T x + \frac{1}{2} x^T \nabla^2 g(y_t) x + \frac{\rho_t}{3} \|x\|^3 \right\},$$

where every iteration requires the solution of a problem of the form (1) and choice of the parameter ρ_t . Griewank [15] first proposed scheme (2) (in a more general setting), and then Nesterov and Polyak [26] and Weiser, Deufhard, and Erdmann [34] independently rediscovered it. Cubic regularization methods, as well as the closely related trust-region methods, are among the most practically successful and theoretically sound approaches to nonconvex optimization [9, 26, 6]. Indeed, Nesterov and Polyak [26] establish that $O(\varepsilon^{-3/2})$ iterations of the form (2) suffice to find an ε -second-order-stationary point of g , meaning a point y_ε such that $\|\nabla g(y_\varepsilon)\| \leq \varepsilon$ and $\lambda_{\min}(\nabla^2 g(y_\varepsilon)) \gtrsim -\sqrt{\varepsilon}$. However, this complexity guarantee does not account for the computational cost of solving subproblems of the form (1).

In this work, we study what is perhaps the simplest algorithm for approximately solving problem (1): gradient descent. Each iteration of gradient descent consists of

*Received by the editors January 27, 2017; accepted for publication (in revised form) April 3, 2019; published electronically September 5, 2019.

<https://doi.org/10.1137/17M1113898>

Funding: This work was partially supported by the SAIL-Toyota Center for AI Research. The first author was partially supported by the Stanford Graduate Fellowship and the Numerical Technologies Fellowship. The second author was partially supported by the National Science Foundation award NSF-CAREER-1553086.

[†]Department of Electrical Engineering, Stanford University, Stanford, CA 94305 (yairc@stanford.edu).

[‡]Departments of Statistics and Electrical Engineering, Stanford University, Stanford, CA 94305 (jduchi@stanford.edu).

the transformation $x \mapsto x - \eta \nabla f(x) = x - \eta(Ax + b + \rho \|x\|x)$ for a step size $\eta \in \mathbb{R}$. Thus, the computational cost of a gradient descent iteration is essentially that of multiplying the matrix A with a vector. Iterative methods requiring only matrix-vector products are called *matrix-free*, and are especially appealing in the setting when d is large and A has structure, such as sparsity (cf. [33]), which enables efficient computation of Ax . Notably, when A is a Hessian as in (2), it is often possible to compute Ax in time linear in d [28, 30], comparable to the time to evaluate a gradient.

We do not claim that gradient descent is the most efficient method for solving problem (1). Indeed, popular matrix-free Krylov subspace solvers [6] provide faster convergence by definition, as the first k iterates of gradient descent lie in the Krylov subspace of order k , $\text{span}\{b, Ab, \dots, A^{k-1}b\}$. Moreover, two-term recursions such as the heavy-ball method [29] and Nesterov's accelerated gradient descent [23] outperform gradient descent in convex problems, with results extending to several nonconvex scenarios [5]. Yet we believe gradient descent—as a workhorse for numerous large-scale problems—is a valuable subject to study, for the following reasons.

1. By proving concrete upper bounds on the number of gradient steps required to achieve an ε -accurate solution to problem (1), we obtain a benchmark for more sophisticated algorithms, such as Krylov subspace methods, and provide dimension-independent guarantees on the number of matrix-vector products such methods require to solve (1) to ε accuracy (see further discussion in section 1.2).
2. Analysis of optimization methods operating on convex quadratic objectives provides important insight about the performance of these methods for general nonlinear objectives close to a local minimum. Analogously, we believe that analyzing gradient descent on the simple structured nonconvex objective (1) will provide useful intuition about the way gradient descent generally navigates saddle points. We show that saddle points may cause gradient descent to stall, but that the overall effect of this stalling on the rate of convergence is bounded, and that the presence of nonconvexity slows convergence by at most a logarithmic factor. We expect a similar qualitative picture to emerge for other nonconvex problems.
3. Unlike more sophisticated methods, gradient descent (with properly chosen step sizes) is often effective in the stochastic setting, where only a noisy estimate of the gradient is available. This effectiveness is well-understood for convex objectives [10, 4], and it extends to several nonconvex problems (notably neural network training), for reasons we do not fully understand [19]. Analyzing gradient descent on the nonstochastic problem (1) is a first step towards understanding stochastic gradient descent methods beyond convex problems, which may prove useful for stochastic variants of the cubic-regularized Newton's method (2) as well a broader theory of nonconvex optimization with stochastic gradient methods.

1.1. Outline of our contribution. We begin our development in section 2 with a number of definitions and results, specifying our assumptions, characterizing the solution to problem (1), and proving that gradient descent converges to the *global minimum* of f . Additionally, we show that gradient descent produces iterates with monotonically increasing norm. This property is essential to our results, and we use it extensively throughout the paper.

In section 3.1 we provide nonasymptotic rates of convergence for gradient descent, which are our main results: gradient descent finds a point x such that $f(x) \leq$

$\inf_{x_* \in \mathbb{R}^d} f(x_*) + \varepsilon$, in a number of steps that scales as $\log \frac{1}{\varepsilon}$ for well-conditioned problems and $\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$ for poorly conditioned problems (for a condition number we define explicitly). Our first convergence guarantee includes the term $\log(1/|v_1^T b|)$, where v_1 is the eigenvector corresponding to the smallest eigenvalue of A . When $v_1^T b = 0$ —as happens in the so-called “hard case” for nonconvex quadratic problems [9]—this term becomes infinite. Nevertheless, by applying gradient descent on a slightly perturbed problem we achieve convergence rates scaling no worse than logarithmically in the problem dimension for any value of $v_1^T b$. (Our results have close connections with the convergence rates of gradient descent on smooth convex functions and of the power method, which we discuss in section 7.)

We illustrate our results with a number of experiments, which we report in section 3.2. We explore the trajectory of gradient descent on nonconvex problem instances, demonstrating its dependence on problem conditioning and the presence of saddle points. We then illustrate our convergence rate guarantees by running gradient descent over an ensemble of random problem instances. This experiment suggests the sharpness of our theoretical analysis.

In section 5 we extend our scope to step sizes chosen by exact line search. If the search is unconstrained, the method may fail to converge to the global minimum, but success is guaranteed for a guarded variation of exact line search. Unfortunately, we have thus far been unable to give rates of convergence for this scheme, though its empirical behavior is at least as strong as standard gradient descent.

As our initial motivation for solving problem (1) is the regularized Newton’s method (2), in section 6 we consider a method for minimizing a general nonconvex function g , which approximates the iterations (2) via gradient descent. In keeping with the theoretical focus of this work, the method is not designed to be efficient in practice, but rather showcases how our analysis applies in the context of subproblem solutions. When g has 2ρ -Lipschitz continuous Hessian, we show that this method finds a point y_ϵ such that $\|\nabla g(y_\epsilon)\| \leq \epsilon$ and $\lambda_{\min}(\nabla^2 g(y_\epsilon)) \geq -\sqrt{\rho\epsilon}$ in ϵ^{-2} gradient and Hessian-vector product evaluations (ignoring constant and logarithmic terms), which is the rate for gradient descent applied directly on g [24, Example 1.2.3]. However, unlike gradient descent, we provide the additional second-order guarantee $\lambda_{\min}(\nabla^2 g(y_\epsilon)) \geq -\sqrt{\rho\epsilon}$, and thus give a first-order method with nonasymptotic convergence guarantees to second-order stationary points at essentially no additional cost over gradient descent. We remark that concurrent works [1, 5] give algorithms attaining such second-order stationary guarantees with an improved first-order complexity scaling as roughly $\epsilon^{-7/4}$.

1.2. Related work. Despite its nonconvexity, problem (1) can be solved to machine precision by means of the iterative solution to linear systems of the form $(A + \lambda I)x = -b$ [6]. However, the cost of this approach generally grows rapidly with the problem dimension d . To address this, several researchers propose matrix-free solvers that allow trading between solution accuracy and computational cost. Griewank [15] and Weiser, Deuffhard, and Erdmann [34] propose variants of the conjugate gradient method, Weiser, Deuffhard, and Erdmann [34] and Cartis, Gould, and Toint [6] propose Krylov subspace solvers based on the Lanczos method, and Bianconcini et al. [3] propose a variant of steepest descent. For generic (i.e., “easy case”) problems and assuming infinite precision arithmetic, Krylov subspace methods solve (1) exactly in d iterations [6], but such guarantees provide limited insight for high-dimensional problems, where the number of iterations is typically $\ll d$. Ideally, a matrix-free solver should provide an ε -accurate solution to (1) in a number of iterations (matrix-vector

products) independent of the problem dimension d , growing instead as the desired tolerance ε decreases, as is the case for first-order methods in convex optimization. The above-mentioned works empirically demonstrate strong performance and scaling to high-dimensional problems, but do not provide such dimension-free convergence guarantees. Our main result shows that gradient descent solves (1) to ε accuracy in $O(\log(d/\varepsilon)/\varepsilon)$ steps, giving a (nearly) dimension-free convergence guarantee. Krylov subspace methods provide solutions at least as accurate as those of gradient descent running the same number of iterations, and therefore our results imply the same convergence guarantee for them as well.

The iterative solvers proposed in [15, 34, 6, 3] approximate subproblem solutions in the cubic regularization scheme (2). It is therefore interesting to understand the total computational cost (in terms of gradient and Hessian-vector product evaluations) of finding an ε -second-order-stationary point for the function g using these approximate solvers. Cartis, Gould, and Toint [7] show that solving the subproblem with a single subspace iteration (known as the Cauchy point) is sufficient for the overall method to converge to an ε -stationary point of g in $O(\varepsilon^{-2})$ outer iterations. However, second-order stationarity is not guaranteed, and the Nesterov–Polyak rate of $O(\varepsilon^{-3/2})$ outer iterations is lost. One naturally asks how many more iterations of the subproblem solver are needed to restore these guarantees. In a follow-up work, Cartis, Gould, and Toint [8] address this question by providing conditions on the quality of subproblem approximations that suffice to guarantee ε -second-order-stationarity after $O(\varepsilon^{-3/2})$ outer iterations. It is unclear how to meet these conditions with a matrix-free method, and in section 6 we show that solving the subproblems with at most $\tilde{O}(\varepsilon^{-1/2})$ gradient descent steps guarantees ε -second-order-stationarity after $O(\varepsilon^{-3/2})$ outer iterations.

Work on the cubic-regularized problem (1) parallels and draws from the literature on the quadratic trust region problem [9, 13, 14, 11], where one replaces the regularizer $(\rho/3)\|x\|^3$ with the constraint $\|x\| \leq R$. Here too, exact solutions are available but scale poorly with dimension, and leading matrix-free solvers include the Steihaug–Toint truncated conjugate gradient method and generalized Lanczos trust region (GLTR), a Lanczos-based subspace method [13]. Tao and An [32] give an analysis of projected gradient descent with a restart scheme that guarantees convergence to the global minimum; however, the number of restarts may be proportional to the problem dimension, suggesting potential difficulties for large-scale problems. Beck and Vaisbourd [2] show convergence to the global minimum for a family of simple first-order methods that includes projected gradient descent. None of these works provides a dimension-free bound on the number of iterations required to solve the subproblem to ε accuracy.

Hazan and Koren [16] address this issue, giving a first-order method that solves the trust-region problem with an accelerated, nearly dimension-free rate. They find an ε -suboptimal point for the trust region problem in $\tilde{O}(1/\sqrt{\varepsilon})$ matrix-vector multiplies by reducing the trust-region problem to a sequence of approximate eigenvector problems. Ho-Nguyen and Kılınç-Karzan [17] provide a different perspective, showing how a single eigenvector calculation can be used to reformulate the nonconvex quadratic trust region problem into a convex QCQP, efficiently solvable with first-order methods.

Concurrent to this work, Agarwal et al. [1] show the same accelerated rate of convergence for the cubic problem (1) via reductions to fast approximate matrix inversion and eigenvector computations. Their rates of convergence are better than those we achieve when ε is large relative to problem conditioning. However, while these works indicate that solving (1) is never harder than approximating the smallest eigenvector

of A , the regime of linear convergence we identify shows that it is sometimes much easier. In work published during the preparation of this paper, Zhang, Shen, and Li [35] demonstrate that Krylov subspace methods indeed achieve (accelerated) linear rates of convergence for trust-region problems, suggesting that such results may be possible for the cubic-regularized problem (1) as well.

Another related line of work is the study of the behavior of gradient descent around saddle points and its ability to escape them [12, 20, 21]. A common theme in these works is an “exponential growth” mechanism that pushes the gradient descent iterates away from critical points with negative curvature. This mechanism plays a prominent role in our analysis as well, highlighting the implications of negative curvature for the dynamics of gradient descent.

2. Preliminaries and basic convergence guarantees. We begin by defining some (mostly standard) notation. Our problem (1) is to solve

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) \triangleq \frac{1}{2} x^T A x + b^T x + \frac{\rho}{3} \|x\|^3,$$

where $\rho > 0$, $b \in \mathbb{R}^d$, and $A \in \mathbb{R}^{d \times d}$ is a symmetric (possibly indefinite) matrix, and $\|\cdot\|$ denotes the Euclidean norm. The eigenvalues of the matrix A are $\lambda^{(1)}(A) \leq \lambda^{(2)}(A) \leq \dots \leq \lambda^{(d)}(A)$, where any of the $\lambda^{(i)}(A)$ may be negative. We define the eigengap of A by $\text{gap} \triangleq \lambda^{(k)}(A) - \lambda^{(1)}(A)$, where k is the first eigenvalue of A strictly larger than $\lambda^{(1)}(A)$. Fix v_1, \dots, v_d to be orthonormal eigenvectors of A such that $Av_i = \lambda^{(i)}(A)v_i$ and $A = \sum_{i=1}^d \lambda^{(i)}(A)v_i v_i^T$. Importantly, throughout the paper we work in the eigenbasis of A , and for any vector $w \in \mathbb{R}^d$ we let

$$(3) \quad w^{(i)} = v_i^T w \text{ denote the } i\text{th coordinate of } w \text{ in the eigenbasis of } A.$$

We let $\|\cdot\|_2$ be the ℓ_2 -operator norm, so $\|A\|_2 = \max_{u: \|u\|=1} \|Au\|$, and define

$$\gamma \triangleq -\lambda^{(1)}(A) \quad \text{and} \quad \beta \triangleq \|A\|_2 = \max\{|\lambda^{(1)}(A)|, |\lambda^{(d)}(A)|\},$$

so that the function f is nonconvex if and only if $\gamma > 0$. Our results also hold when $\beta \geq \|A\|_2$ rather than its exact value. We say a function g is L -smooth on a convex set X if $\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$ for all $x, y \in X$; this is equivalent to $\|\nabla^2 g(x)\|_2 \leq L$ for Lebesgue almost every $x \in X$ and is equivalent to the bound $|g(x) - g(y) - \nabla g(y)^T(x - y)| \leq \frac{L}{2}\|x - y\|^2$ for $x, y \in X$.

2.1. Characterization of f and its global minimizers. Throughout the paper, we let x_* denote a solution to problem (1), i.e., a global minimizer of f , and define the matrix

$$A_* \triangleq A + \rho\|x_*\|I,$$

where I is the $d \times d$ identity matrix. We have the following characterization for x_* .

PROPOSITION 2.1 (cf. [6, Theorem 3.1]). *A solution x_* of problem (1) satisfies*

$$(4) \quad \nabla f(x_*) = A_* x_* + b = 0 \quad \text{and} \quad \rho\|x_*\| \geq \gamma,$$

and x_ is unique whenever $\rho\|x_*\| > \gamma$.*

We may write the gradient and Hessian of f as

$$\nabla f(x) = A_*(x - x_*) - \rho(\|x_*\| - \|x\|)x \quad \text{and} \quad \nabla^2 f(x) = A + \rho\|x\|I + \rho \frac{xx^T}{\|x\|}.$$

The globally minimal value of f admits the expression and bound

$$(5a) \quad f(x_*) = \frac{1}{2}x_*^T A s + b^T x_* + \frac{\rho \|x_*\|^3}{3} = -\frac{1}{2}x_*^T A_* x_* - \frac{\rho \|x_*\|^3}{6} \leq -\frac{\rho \|x_*\|^3}{6},$$

and, using the fact that $x_*^T A_* x_* = -b^T x_* \leq \|b\| \|x_*\|$, we derive the lower bound

$$(5b) \quad f(x_*) \geq -\frac{1}{2}\|b\|\|x_*\| - \frac{\rho \|x_*\|^3}{6}.$$

Algebraic manipulation also shows that

$$(6) \quad f(x) = f(x_*) + \frac{1}{2}(x - x_*)^T A_*(x - x_*) + \frac{\rho}{6}(\|x_*\| - \|x\|)^2(\|x_*\| + 2\|x\|),$$

which makes it clear that x_* is indeed the global minimum, as both of the x -dependent terms are nonnegative and minimized at $x = x_*$, and the minimum is unique whenever $\|x_*\| > \gamma/\rho$, because $A_* \succ 0$ in this case.

The global minimizer admits the following equivalent characterization whenever the vector b is not orthogonal to the eigenspace associated with $\lambda^{(1)}(A)$.

PROPOSITION 2.2. *If $b^{(1)} \neq 0$, x_* is the unique solution to the system defined by*

$$\nabla f(s) = 0 \quad \text{and} \quad b^{(1)} s^{(1)} \leq 0.$$

Proof. Let x'_* satisfy $\nabla f(x'_*) = 0$ and $b^{(1)} x'^{(1)}_* \leq 0$. Focusing on the first (eigen)coordinate, we have $0 = [\nabla f(x'_*)]^{(1)} = (-\gamma + \rho \|x'_*\|)x'^{(1)}_* + b^{(1)}$. Therefore, $b^{(1)} \neq 0$ implies both $x'^{(1)}_* \neq 0$ and $-\gamma + \rho \|x'_*\| \neq 0$. This strengthens the inequality $b^{(1)} x'^{(1)}_* \leq 0$ to $b^{(1)} x'^{(1)}_* < 0$. Hence $-\gamma + \rho \|x'_*\| = -b^{(1)} x'^{(1)}_* / [x'^{(1)}_*]^2 > 0$; by Proposition 2.1, if a critical point satisfies $\rho \|x'_*\| > \gamma$, it is the unique global minimum. \square

The norm of x_* plays an important role in our analysis, so we provide a number of bounds on it. First, observe that $\|b\| = \|A_* x_*\| \geq (-\gamma + \rho \|x_*\|)\|x_*\|$. Solving for $\|x_*\|$ gives the upper bound

$$(7a) \quad \|x_*\| \leq \frac{\gamma}{2\rho} + \sqrt{\left(\frac{\gamma}{2\rho}\right)^2 + \frac{\|b\|}{\rho}} \leq \frac{\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{\|b\|}{\rho}} \triangleq R,$$

where we recall that $\beta = \|A\|_2 \geq |\gamma|$. An analogous lower bound on $\|x_*\|$ is available: we have $\|x_*\| \geq \gamma/\rho$, and if $b^{(1)} \neq 0$, then $\|x_*\| = \|A_*^{-1} b\| \geq |b^{(1)}|/(-\gamma + \rho \|x_*\|)$ implies

$$(7b) \quad \|x_*\| \geq \frac{\gamma}{2\rho} + \sqrt{\left(\frac{\gamma}{2\rho}\right)^2 + \frac{|b^{(1)}|}{\rho}} \geq -\frac{\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{|b^{(1)}|}{\rho}} = R - \frac{\beta}{\rho}.$$

We can also prove a different lower bound with the similar form

$$(8) \quad \|x_*\| \geq R_c \triangleq \frac{-b^T A b}{2\rho \|b\|^2} + \sqrt{\left(\frac{b^T A b}{2\rho \|b\|^2}\right)^2 + \frac{\|b\|}{\rho}} \geq -\frac{\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{\|b\|}{\rho}}.$$

The quantity R_c is the *Cauchy radius* [9]—the magnitude of the (global) minimizer of f in the subspace spanned by b : $R_c = \operatorname{argmin}_{\zeta \in \mathbb{R}} f(-\zeta b/\|b\|)$. To see the claimed

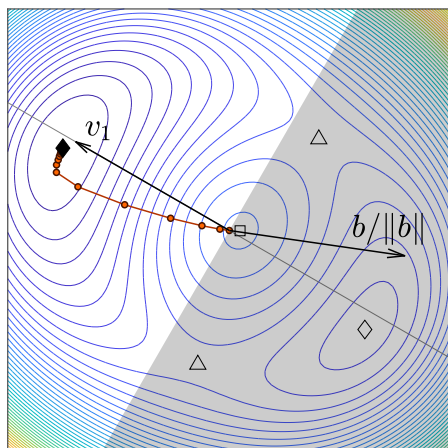


FIG. 1. Contour plot of a two-dimensional instance of (1), featuring a local maximum (\square), saddle points (\triangle), and local minima (\diamond). The line of circles indicates the path of gradient descent initialized at the origin, and the gray shaded area is the half-plane $(v_1^T b)(v_1^T x) = b^{(1)} x^{(1)} > 0$. Note that the global minimum is the only critical point outside this half-plane (Proposition 2.2). The gradient descent iterates have increasing norm (Lemma 2.3), lie outside the half-plane (Lemma 2.4), and converge to x_* (Proposition 2.5).

lower bound (8), set $x_c = -R_c b / \|b\|$ (the *Cauchy point*) and note that $f(x_c) = -(1/2)\|b\|R_c - (\rho/6)R_c^3$. Therefore,

$$0 \leq f(x_c) - f(x_*) \leq \frac{1}{2}\|b\|(\|x_*\| - R_c) + \frac{1}{6}\rho(\|x_*\|^3 - R_c^3),$$

which implies $\|x_*\| \geq R_c$.

For matrices A with distinct eigenvalues, f may have a single suboptimal local minimizer, a single local maximizer, and up to $2(d-1)$ saddle points [15, section 3]; see Figure 1 for an example with $d = 2$.¹

2.2. Properties and convergence of gradient descent. The gradient descent method begins at some initialization $x_0 \in \mathbb{R}^d$ and generates iterates via

$$(9) \quad x_{t+1} = x_t - \eta \nabla f(x_t) = (I - \eta A - \rho \eta \|x_t\| I) x_t - \eta b,$$

where η is a fixed step size. Recalling the definitions (7a) and (8) of R and R_c as well as $\|A\|_2 = \beta$, throughout our analysis we make the following assumptions.

Assumption A. The step size η in (9) satisfies $0 < \eta \leq \frac{1}{4(\beta + \rho R)}$.

Assumption B. The initialization of (9) satisfies $x_0 = -r \frac{b}{\|b\|}$, with $0 \leq r \leq R_c$.

To select a step size η satisfying Assumption A, only a rough upper bound on $\|A\|_2$ is necessary. One way to obtain such a bound (with high probability) is to apply a few power iterations on A . Alternatively, we may perform a line search, as in section 5.

We begin our treatment of the convergence of gradient descent by establishing that $\|x_t\|$ is monotonic and bounded (see Appendix A for a proof).

¹Color figures are available in the online version of this paper.

LEMMA 2.3. *Let Assumptions A and B hold. Then the iterates (9) of gradient descent satisfy $x_t^T \nabla f(x_t) \leq 0$, the norms $\|x_t\|$ are nondecreasing, and $\|x_t\| \leq R$.*

This lemma is the key to our analysis throughout the paper. The next lemma shows that x_t and b have opposite signs at all coordinates in the eigenbasis of A .

LEMMA 2.4. *Let Assumptions A and B hold. For all $t \geq 0$ and $i \in \{1, \dots, d\}$,*

$$x_t^{(i)} b^{(i)} \leq 0, \quad b^{(i)} x_\star^{(i)} \leq 0, \quad \text{and} \quad x_t^{(i)} x_\star^{(i)} \geq 0.$$

Consequently, $x_t^T b \leq 0$ and $x_t^T x_\star \geq 0$ for every t , and $x_\star^T b \leq 0$.

Proof. We first show that $x_t^{(i)} b^{(i)} \leq 0$. Writing the gradient descent recursion in the eigenbasis of A , we have

$$(10) \quad x_t^{(i)} = \left(1 - \eta \lambda^{(i)}(A) - \eta \rho \|x_{t-1}\|\right) x_{t-1}^{(i)} - \eta b^{(i)}.$$

Assumption A and Lemma 2.3 imply $1 - \eta \lambda^{(i)}(A) - \eta \rho \|x_{t-1}\| \geq 1 - \eta(\beta + \rho R) > 0$ for all t, i . Therefore, $x_t^{(i)} b^{(i)} \leq 0$ if $x_0^{(i)} b^{(i)} \leq 0$; the initialization in Assumption B guarantees this. To show $b^{(i)} x_\star^{(i)} \leq 0$, we use the fact that $b = -A_\star x_\star$ to write

$$b^{(i)} x_\star^{(i)} = -\left(\lambda^{(i)}(A) + \rho \|x_\star\|\right) [x_\star^{(i)}]^2 \leq 0$$

as $\lambda^{(i)}(A) + \rho \|x_\star\| \geq 0$ for every i by the condition (4) defining x_\star .

Multiplying $x_t^{(i)} b^{(i)} \leq 0$ and $b^{(i)} x_\star^{(i)} \leq 0$ yields $x_t^{(i)} x_\star^{(i)} [b^{(i)}]^2 \geq 0$. The coordinate-wise update (10) and Assumption B show that $b^{(i)} = 0$ implies $x_t^{(i)} = 0$ for every t , and therefore $x_t^{(i)} x_\star^{(i)} \geq 0$. \square

Lemmas 2.3 and 2.4, and Proposition 2.2 immediately lead to the following guarantee.

PROPOSITION 2.5. *Let Assumptions A and B hold, and assume that $b^{(1)} \neq 0$. Then $x_t \rightarrow x_\star$ and $f(x_t) \downarrow f(x_\star)$ as $t \rightarrow \infty$.*

Proof. By Lemma 2.3, the iterates satisfy $\|x_t\| \leq R$ for all t . Since $\|\nabla^2 f(x)\|_2 \leq \beta + 2\rho\|x\|$, the function f is $\beta + 2\rho R$ -smooth on the set $\{x \in \mathbb{R}^d : \|x\| \leq R\}$ containing all the iterates x_t . Therefore, by the definition of smoothness and the gradient step,

$$f(x_{t+1}) \leq f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\eta^2}{2} (\beta + 2\rho R) \|\nabla f(x_t)\|^2 \leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2,$$

where the final inequality used that $\eta \leq \frac{1}{4(\beta + \rho R)}$ from Assumption A. Consequently, $f(x_t)$ is decreasing and, for every $t > 0$,

$$(11) \quad \frac{\eta}{2} \sum_{\tau=0}^{t-1} \|\nabla f(x_\tau)\|^2 \leq f(x_0) - f(x_t) \leq f(x_0) - f(x_\star).$$

Let x'_\star be any limit point of the sequence x_t (there must be at least one, as the sequence x_t is bounded). Inequality (11) implies $\nabla f(x_t) \rightarrow 0$ and therefore $\nabla f(x'_\star) = 0$ by continuity. By Lemma 2.4, $x_t^{(1)} b^{(1)} \leq 0$ for every t , so $x_\star^{(1)} b^{(1)} \leq 0$. Proposition 2.2 thus implies that x'_\star is the unique global minimizer x_\star . We conclude that x_\star is the only limit point of the sequence x_t . \square

To handle the case when $b^{(1)} = 0$, let $k \geq 1$ be the first index for which $b^{(k)} \neq 0$ (if no such k exists then $b = 0$ and $x_t = 0$ for all t). Consider a modified problem instance, with b , ρ unchanged but A replaced with

$$\tilde{A} = \beta \sum_{i=1}^{k-1} v_i v_i^T + \lambda^{(i)}(A) v_i v_i^T,$$

i.e., we replace the $k-1$ smallest eigenvalues with $\beta \geq \lambda^{(d)}(A)$. Note that gradient descent produces the same iterates on the modified and original problems. Additionally, note that Lemma 2.3 and Proposition 2.5 apply to the modified problem, as the inner product between b and the eigenvector of \tilde{A} corresponding to its smallest eigenvalue is nonzero. Applying these results, we have $\|x_t\| \uparrow \|\hat{x}_\star\|$, where \hat{x}_\star is the unique solution of the modified problem. Finally, we have $\|\hat{x}_\star\| \leq \|x_\star\|$, since $\hat{x}_\star \neq x_\star$ only if $\rho\|\hat{x}_\star\| \leq \gamma$ [6, section 6.1]. Thus, we obtain the following lemma, to which we will refer throughout the paper.

LEMMA 2.6. *Let Assumptions A and B hold. For all $t \geq 0$, the iterates (9) of gradient descent satisfy $x_t^T \nabla f(x_t) \leq 0$, the norms $\|x_t\|$ are nondecreasing and satisfy $\|x_t\| \leq \|x_\star\|$, and f is $(\beta + 2\rho\|x_\star\|)$ -smooth on a ball containing the iterates x_t .*

Figure 1 provides a graphical representation of these results, showing gradient descent's iterates on an instance of problem (1) exhibiting numerous stationary points.

3. Nonasymptotic convergence rates. Proposition 2.5 shows the convergence of the gradient for the cubic-regularized (nonconvex) quadratic problem (1). We now present stronger nonasymptotic guarantees, including a randomized scheme solving (1) in all cases. We follow this with simulations illustrating our theoretical results.

3.1. Theoretical results. Our primary result, Theorem 3.1, gives a convergence rate for gradient descent in the case when $b^{(1)} \neq 0$. (Recall our convention (3): that parenthesized superscripts denote components in the eigenbasis of A .) Further recalling that $\gamma = -\lambda^{(1)}(A)$, $\beta = \|A\|_2$, and gap is the eigengap of A , we define the shorthand

$$\gamma_+ \triangleq \max\{\gamma, 0\} \quad \text{and} \quad \text{gap}' \triangleq \min\{\text{gap}, \rho\|x_\star\|\}.$$

With this notation in hand, we state our result as follows.

THEOREM 3.1. *Let Assumptions A and B hold, $b^{(1)} \neq 0$, and $\varepsilon > 0$. Then $f(x_t) \leq f(x_\star) + \varepsilon$ for all*

$$(12) \quad t \geq T_\varepsilon \triangleq \frac{\tau_{\text{grow}}(b^{(1)}) + \tau_{\text{conv}}(\varepsilon)}{\eta} \begin{cases} \frac{1}{\rho\|x_\star\| - \gamma}, & \frac{1}{\rho\|x_\star\| - \gamma} \leq \frac{10\|x_\star\|^2}{\varepsilon}, \\ \sqrt{\frac{10\|x_\star\|^2}{\varepsilon} \cdot \frac{1}{\text{gap}'}}}, & \frac{1}{\text{gap}'} \leq \frac{10\|x_\star\|^2}{\varepsilon} \leq \frac{1}{\rho\|x_\star\| - \gamma}, \\ \frac{10\|x_\star\|^2}{\varepsilon}, & \text{otherwise,} \end{cases}$$

where

$$\tau_{\text{grow}}(b^{(1)}) = 6 \log \left(1 + \frac{\gamma_+^2}{4\rho|b^{(1)}|} \right) \quad \text{and} \quad \tau_{\text{conv}}(\varepsilon) = 6 \log \left(\frac{(\beta + 2\rho\|x_\star\|)\|x_\star\|^2}{\varepsilon} \right).$$

See section 4.1 for a proof.

Theorem 3.1 shows that the rate of convergence changes from roughly $O(1/\varepsilon)$ to $O(\log(1/\varepsilon))$ as ε decreases, with an intermediate gap-dependent rate of $O(1/\sqrt{\varepsilon})$. The terms τ_{grow} and τ_{conv} correspond to a period (τ_{grow}) in which $\|x_t\|$ grows exponentially in t until reaching the basin of attraction to the global minimum and a period (τ_{conv}) of linear convergence to x_* . Exponential growth occurs only in nonconvex problem instances, as $\tau_{\text{grow}} = 0$ when the problem is convex.

The dependence of our result on $|b^{(1)}|$ (the magnitude of b in the direction of the smallest eigenvector of A) is unavoidable: if $b^{(1)} = 0$, then gradient descent always remains in a subspace orthogonal to the smallest eigenvector of A , while $x_*^{(1)}$ might be nonzero; this is the “hard case” of nonconvex quadratic problems [9, 6]. We use a small random perturbation to guarantee $|b^{(1)}| \neq 0$ except with negligible probability, which yields the following high probability guarantee, whose proof we provide in section 4.2.

THEOREM 3.2. *Let Assumptions A and B hold, let $\varepsilon, \delta > 0$, and let q be uniformly distributed on the unit sphere in \mathbb{R}^d . Let \tilde{x}_t be generated by the gradient descent iteration (9) with $\tilde{b} = b + \sigma q$ replacing b , where*

$$\sigma = \frac{\rho\varepsilon}{\beta + 2\rho\|x_*\|} \cdot \frac{\bar{\sigma}}{12} \text{ with } \bar{\sigma} \leq 1.$$

Then, with probability at least $1 - \delta$, we have $f(\tilde{x}_t) \leq f(x_) + (1 + \bar{\sigma})\varepsilon$ for all*

$$(13) \quad t \geq T_\varepsilon \triangleq \frac{\tilde{\tau}_{\text{grow}}(d, \delta, \bar{\sigma}) + \tilde{\tau}_{\text{conv}}(\varepsilon)}{(1 + \bar{\sigma})^{-1}\eta} \begin{cases} \frac{1}{\rho\|x_*\| - \gamma}, & \frac{1}{\rho\|x_*\| - \gamma} \leq \frac{10\|x_*\|^2}{\varepsilon}, \\ \sqrt{\frac{10\|x_*\|^2}{\varepsilon} \cdot \frac{1}{\text{gap}}}, & \frac{1}{\text{gap}} \leq \frac{10\|x_*\|^2}{\varepsilon} \leq \frac{1-2\bar{\sigma}/3}{\rho\|x_*\| - \gamma}, \\ \frac{10\|x_*\|^2}{\varepsilon}, & \text{otherwise,} \end{cases}$$

where

$$\tilde{\tau}_{\text{grow}}(d, \delta, \bar{\sigma}) \triangleq 6 \log \left(1 + \mathbb{I}_{\{\gamma > 0\}} \frac{3\sqrt{d}}{\bar{\sigma}\delta} \right), \quad \tilde{\tau}_{\text{conv}}(\varepsilon) \triangleq 14 \log \left(\frac{(\beta + 2\rho\|x_*\|)\|x_*\|^2}{\varepsilon} \right).$$

To facilitate later discussion, we define $L_* \triangleq \beta + 2\rho\|x_*\|$; then f is L_* -smooth on the Euclidean ball of radius $\|x_*\|$. The bound (7b) implies $\rho R \leq \beta + \rho\|x_*\|$, and therefore the step size choice $\eta = \frac{1}{4(\beta + \rho R)}$ satisfies $\frac{1}{\eta} \leq 8\beta + 4\rho\|x_*\| \leq 8L_*$. Combining this upper bound with Theorem 3.2, we have the following corollary.

COROLLARY 3.3. *Let the conditions of Theorem 3.2 hold and let $\eta = \frac{1}{4(\beta + \rho R)}$ and $\bar{\sigma} = 1$. Then, with probability at least $1 - \delta$, we have $f(\tilde{x}_t) \leq f(x_*) + \varepsilon$ for all*

$$t \geq \tilde{T}_\varepsilon = O(1) \cdot \min \left\{ \frac{L_*}{\rho\|x_*\| - \gamma}, \frac{L_*\|x_*\|^2}{\varepsilon} \right\} \log \left[\left(1 + \mathbb{I}_{\{\gamma > 0\}} \frac{d}{\delta} \right) \frac{L_*\|x_*\|^2}{\varepsilon} \right].$$

We conclude the presentation of our main results with a few brief remarks.

- (i) Corollary 3.3 highlights parallels between our guarantees and those for gradient descent on smooth convex functions [24]. In our case, $L_*/(\rho\|x_*\| - \gamma) \geq 1$ is a condition number, while L_* and $\|x_*\|$ bound the smoothness of f and iterate radius $\sup_t \|x_t\|$, respectively. We defer further comparison to section 7.
- (ii) We readily obtain relative accuracy guarantees by using the bound (5a); setting $\varepsilon = \rho\|x_*\|^3\varepsilon'/12$, we have $f(\tilde{x}_t) - f(x_*) \leq -\varepsilon'f(x_*) = \varepsilon'(f(0) - f(x_*))$, or $f(\tilde{x}_t) \leq (1 - \varepsilon')f(x_*)$, for any $t \geq \tilde{T}_\varepsilon$, where \tilde{T}_ε is as defined in (13).

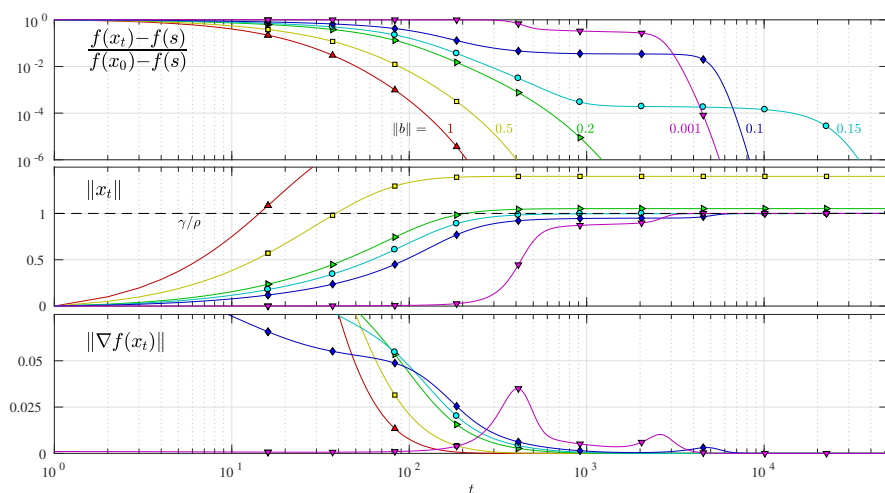


FIG. 2. Trajectories of gradient descent with $\lambda^{(1)}(A) = -\gamma = -0.2$ and $\lambda^{(2)}(A), \dots, \lambda^{(d)}(A)$ equally spaced between -0.18 and $\beta = 1$, and different vectors b proportional to $[0.01, 1, 1, 1, \dots]$ in the eigenbasis of A . The rest of the parameters are $d = 10^3$, $\eta = 0.1$, $\rho = 0.2$, and $x_0 = 0$.

- (iii) Evaluating \tilde{T}_ε for given A , b , and ρ is not straightforward, as $\|x_\star\|$ is generally unknown. Using $\|x_\star\| \leq R$ gives an easily computable upper bound on \tilde{T}_ε , and in section 6 we demonstrate how to apply our results when $\|x_\star\|$ is unknown.

3.2. Illustration of results. We present two experiments that investigate the behavior of gradient descent on problem (1). For the first experiment, we examine the behavior of gradient descent on single problem instances, looking at convergence behavior as we vary the vector b (to effect conditioning of the problem) by scaling its norm $\|b\|$. The selected norm values $\|b\| \in \{1, 0.5, 0.2, 0.15, 0.1, 0.001\}$ correspond to condition numbers $(\beta + \rho\|x_\star\|)/(-\gamma + \rho\|x_\star\|) \in \{7.6, 16, 120, 5.5 \cdot 10^3, 2.9 \cdot 10^4, 3.8 \cdot 10^6\}$; the problem conditioning becomes worse as $\|b\|$ decreases. Figure 2 summarizes our results and describes the settings of the other parameters in the experiment.

The plots show two behaviors of gradient descent. The problem is well-conditioned when $\|b\| \geq 0.2$, and in these cases gradient descent behaves as though the problem is strongly convex, with x_t converging linearly to x_\star . For $\|b\| \leq 0.15$ the problem becomes ill-conditioned and gradient descent stalls around saddle points. Indeed, the third plot of Figure 2 shows that for the ill-conditioned problems we have $\|\nabla f(x_t)\|$ increasing over some iterations, which does not occur in convex quadratic problems. The length of the stall does not depend only on the condition number; for $\|b\| = 10^{-3}$ the stall is shorter than for $\|b\| \in \{0.1, 0.15\}$. Instead, it appears to depend on the norm of the saddle point that causes it, which we observe from the value of $\|x_t\|$ at the time of the stall; we see that the closer the norm is to γ/ρ , the longer the stall takes. This is explained by observing that $\nabla^2 f(x) \succeq (\rho\|x\| - \gamma)I$, which means that every saddle point with norm close to γ/ρ must have only small negative curvature, and therefore is harder to escape (see also Lemma 4.3). Fortunately, as we see in Figure 2, saddle points with large norm have near-optimal objective value—this is the intuition behind our proof of the sublinear convergence rates.

In our second experiment, we test our rate guarantees by considering the performance of gradient descent over an ensemble of random instances. We generate random instances with a fixed value of γ , β , ρ , $\|x_\star\|$, and gap as follows. We set

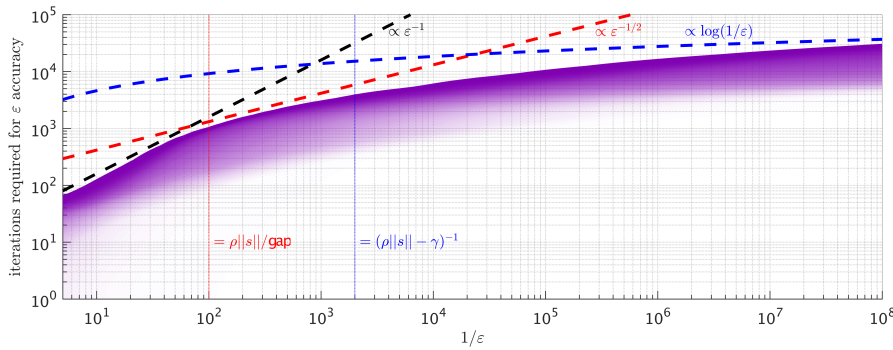


FIG. 3. The shaded curve shows the cumulative distribution function of the number of iterations required to reach relative accuracy ε , computed over 2,500 random problem instances each with $d = 10^4$, $\beta = \rho = 1$, $\gamma = 0.5$, $\text{gap} = 5 \cdot 10^{-3}$ and $\rho\|x_\star\| - \gamma = 5 \cdot 10^{-4}$. We use $x_0 = -R_c b / \|b\|$ and $\eta = 0.25$. The three dashed curves indicate the three convergence regimes Theorem 3.1 identifies.

$A = \text{diag}([-\gamma; -\gamma + \text{gap}; u])$ with u uniformly random in $[-\gamma + \text{gap}, \beta]^{d-2}$. We draw $\tilde{x}_\star = (A + \rho\|x_\star\|)^{-\zeta} \nu$, where $\nu \sim \mathcal{N}(0; I)$ and $\log_2 \zeta$ is uniform on $[-1, 1]$. We then set $x_\star = (\|x_\star\| / \|\tilde{x}_\star\|) \tilde{x}_\star$ and $b = -(A + \rho\|x_\star\|)x_\star$, so that x_\star is the global minimizer of problem instance (A, b, ρ) . The choice of ζ ensures we observe large variety in the values of $\|x_t\|$ at which gradient descent stalls, allowing us to find difficult instances for each value of ε . In Figure 3 we depict the cumulative distribution of the number of iterations required to find an ε -relatively accurate solution versus $1/\varepsilon$. The slopes in the plot agree with our upper bounds, suggesting the sharpness of our theoretical results.

4. Proofs of main results. In this section, we provide proofs of our main results, Theorems 3.1 and 3.2. A number of the steps involve technical lemmas whose proofs we defer to Appendix B. In all lemma statements, we tacitly let Assumptions A and B hold, as in the main theorem statements. Without loss of generality, we assume $\varepsilon \leq \frac{1}{2}\beta\|x_\star\|^2 + \rho\|x_\star\|^3$, as f is $\beta + 2\rho\|x_\star\|$ smooth on the set $\{x : \|x\| \leq \|x_\star\|\}$ and therefore $f(x_0) \leq f(x_\star) + \varepsilon$ for any $\varepsilon \geq \frac{1}{2}\beta\|x_\star\|^2 + \rho\|x_\star\|^3$.

4.1. Proof of Theorem 3.1. We divide the proof of Theorem 3.1 into two main steps: in section 4.1.1 we prove the first case in the bound (12) (linear convergence), and in section 4.1.2 we prove the last two cases in (12) (sublinear convergence).

4.1.1. Linear convergence and exponential growth. We first prove that $f(x_t) \leq f(x_\star) + \varepsilon$ for $t \geq \frac{1}{\eta(\rho\|x_\star\| - \gamma)}(\tau_{\text{grow}}(b^{(1)}) + \tau_{\text{conv}}(\varepsilon))$. We begin with two lemmas that provide regimes in which x_t converges to the solution x_\star linearly.

LEMMA 4.1. *For each $t > 0$, we have*

$$\|x_t - x_\star\|^2 \leq \left(1 - \eta \left[\rho\|x_t\| - \left(\gamma - \frac{\rho\|x_\star\| - \gamma}{2} \right) \right] \right) \|x_{t-1} - x_\star\|^2.$$

See Appendix B.1 for a proof of this lemma.

For nonconvex problem instances (those with $\gamma > 0$), the above recursion is a contraction (implying linear convergence of x_t to x_\star) only when $\rho\|x_t\|$ is larger than $\gamma - \frac{1}{2}(\rho\|x_\star\| - \gamma)$. Using the fact that $\|x_t\|$ is nondecreasing (Lemma 2.6), Lemma 4.1 immediately implies the following result.

LEMMA 4.2. If $\rho\|x_t\| \geq \gamma - \frac{1}{2}(\rho\|x_\star\| - \gamma) + \mu$ for some $t \geq 0$, then, for all $\tau \geq 0$,

$$\|x_{t+\tau} - x_\star\|^2 \leq (1 - \eta\mu)^\tau \|x_t - x_\star\|^2 \leq 2\|x_\star\|^2 e^{-\eta\mu\tau}.$$

Proof. Lemma 4.1 implies that $\|x_{t+\tau} - x_\star\|^2 \leq (1 - \eta\mu)\|x_{t+\tau-1} - x_\star\|^2$ for all $\tau > 1$. Using that $\|x_t - x_\star\|^2 \leq \|x_t\|^2 + \|x_\star\|^2 \leq 2\|x_\star\|^2$ by Lemmas 2.4 and 2.6 and $1 + \alpha \leq e^\alpha$ for all α gives the result. \square

It remains to understand whether the gradient descent iterations satisfy the condition $\rho\|x_t\| \geq \gamma - \frac{1}{2}(\rho\|x_\star\| - \gamma) + \mu$. Fortunately, as long as $\rho\|x_t\|$ is below $\gamma - \nu$, $|x_t^{(1)}|$ grows faster than $(1 + \eta\nu)^t$.

LEMMA 4.3. Let $\nu > 0$. Then $\rho\|x_t\| \geq \gamma - \nu$ for all $t \geq \frac{2}{\eta\nu} \log(1 + \frac{\gamma_+^2}{4\rho|b^{(1)}|})$.

See Appendix B.2 for a proof of this lemma.

We now combine the lemmas to give the linear convergence regime of Theorem 3.1. Applying Lemma 4.3 with $\nu = \frac{1}{3}(\rho\|x_\star\| - \gamma)$ yields $\rho\|x_t\| \geq \gamma - \frac{1}{3}(\rho\|x_\star\| - \gamma)$ for

$$t \geq T_1 \triangleq \frac{6}{\eta(\rho\|x_\star\| - \gamma)} \log\left(1 + \frac{\gamma_+^2}{4\rho|b^{(1)}|}\right) = \frac{1}{\eta(\rho\|x_\star\| - \gamma)} \tau_{\text{grow}}(b^{(1)}).$$

Therefore, by Lemma 4.2 with $\mu = \frac{1}{2}(\rho\|x_\star\| - \gamma) - \nu = \frac{1}{6}(\rho\|x_\star\| - \gamma)$, for any t we have

$$(14) \quad \|x_{T_1+t} - x_\star\|^2 \leq 2\|x_\star\|^2 \exp\left(-\frac{1}{6}\eta(\rho\|x_\star\| - \gamma)t\right).$$

As a consequence, for all $t \geq 0$ we may use the $(\beta + 2\rho\|x_\star\|)$ -smoothness of f and the fact that $\|x_t\| \leq \|x_\star\|$ (by Lemma 2.6) to obtain

$$f(x_t) - f(x_\star) \leq \frac{\beta + 2\rho\|x_\star\|}{2} \|x_t - x_\star\|^2 \leq (\beta + 2\rho\|x_\star\|) \|x_\star\|^2 e^{-\frac{1}{6}\eta(\rho\|x_\star\| - \gamma)(t-T_1)},$$

where we have used that $\nabla f(x_\star) = 0$ and the bound (14). Therefore, if we set

$$T_2 \triangleq \frac{6}{\eta(\rho\|x_\star\| - \gamma)} \log \frac{(\beta + 2\rho\|x_\star\|) \|x_\star\|^2}{\varepsilon} = \frac{1}{\eta(\rho\|x_\star\| - \gamma)} \tau_{\text{conv}}(\varepsilon),$$

then $t \geq T_1 + T_2 = \frac{1}{\eta(\rho\|x_\star\| - \gamma)} (\tau_{\text{grow}}(b^{(1)}) + \tau_{\text{conv}}(\varepsilon))$ implies $f(x_t) - f(x_\star) \leq \varepsilon$.

4.1.2. Sublinear convergence and convergence in subspaces. We now turn to the sublinear convergence regime in Theorem 3.1, which applies when the quantity $\rho\|x_\star\| - \gamma$ is sufficiently small:

$$(15) \quad \rho\|x_\star\| - \gamma \leq \frac{\varepsilon}{10\|x_\star\|^2}.$$

Note that if (15) fails to hold, then (12) is dominated by the $(\rho\|x_\star\| - \gamma)^{-1}$ term. Therefore, to complete the proof of Theorem 3.1 it suffices to show that if (15) holds, then $f(x_t) \leq f(x_\star) + \varepsilon$ whenever

$$(16) \quad t \geq T_\varepsilon^{\text{sub}} \triangleq \frac{\tau_{\text{grow}}(b^{(1)}) + \tau_{\text{conv}}(\varepsilon)}{\eta} \min \left\{ \frac{10\|x_\star\|^2}{\varepsilon}, \sqrt{\frac{10\|x_\star\|^2}{(\min\{\text{gap}, \rho\|x_\star\|\})\varepsilon}} \right\}.$$

Roughly, our proof of the result (16) proceeds as follows: when $\rho\|x_\star\| - \gamma$ is small, the function f is very smooth along eigenvectors with eigenvalues close to $-\gamma = \lambda^{(1)}(A)$. It is therefore sufficient to show convergence in the complementary subspace, which occurs at a linear rate. Appropriately choosing the gap between the eigenvalues in the complementary subspace and $\lambda^{(1)}(A)$ to trade between convergence rate and function smoothness yields the rates (16).

The following analogues of Lemmas 4.1 and 4.2 establish subspace convergence.

LEMMA 4.4. *Let Π be any projection matrix satisfying $\Pi A = A\Pi$ for which $\Pi A_\star \succeq \nu\Pi$ for some $\nu > 0$. For all $t > 0$,*

$$\begin{aligned} \|\Pi A_\star^{1/2}(x_t - x_\star)\|^2 &\leq (1 - \eta\nu) \|\Pi A_\star^{1/2}(x_{t-1} - x_\star)\|^2 \\ &\quad + \sqrt{8}\eta\rho(\|x_\star\| - \|x_{t-1}\|) [\rho(\|x_\star\| - \|x_{t-1}\|) \|x_{t-1}\|^2 + \|(I - \Pi)A_\star\|_2 \|x_\star\|^2]. \end{aligned}$$

See Appendix B.3 for a proof. Letting $\Pi_\nu = \sum_{i: \lambda^{(i)} \geq \nu + \lambda^{(1)}} v_i v_i^T$ be the projection matrix onto the span of eigenvectors of A with eigenvalues at least $\lambda^{(1)}(A) + \nu$, we obtain the following consequence of Lemma 4.4, whose proof we provide in Appendix B.4.

LEMMA 4.5. *Let $t \geq 0$, let $\nu \geq 0$, and define $\bar{\nu} = \max\{\nu, \text{gap}\}$. If $\rho\|x_\star\| \leq \gamma + \sqrt{\nu\bar{\nu}}$ and $\rho\|x_t\| \geq \gamma - \frac{1}{3}\sqrt{\nu\bar{\nu}}$, then, for any $\tau \geq 0$,*

$$\begin{aligned} \|\Pi_\nu A_\star^{1/2}(x_{t+\tau} - x_\star)\|^2 &\leq (1 - \eta\bar{\nu})^\tau \|\Pi_\nu A_\star^{1/2}(x_t - x_\star)\|^2 + 13\|x_\star\|^2\nu \\ &\leq 2(\beta + \rho\|x_\star\|) \|x_\star\|^2 e^{-\eta\bar{\nu}\tau} + 13\|x_\star\|^2\nu. \end{aligned}$$

We use these lemmas to prove the desired bound (16) by appropriate separation of the eigenspaces over which we guarantee convergence. To that end, we define

$$(17) \quad \nu \triangleq \frac{\varepsilon}{10\|x_\star\|^2}, \quad \bar{\nu} \triangleq \max\{\nu, \text{gap}\}, \quad \text{and} \quad \bar{\nu}' \triangleq \max\{\nu, (\min\{\text{gap}, \rho\|x_\star\|\}) \leq \bar{\nu}\},$$

and note that the definition of **gap** immediately implies $\Pi_\nu = \Pi_{\bar{\nu}}$. The growth guaranteed by Lemma 4.3 shows that $\rho\|x_t\| \geq \gamma - \frac{1}{3}\sqrt{\nu\bar{\nu}'}$ for every

$$t \geq T_1^{\text{sub}} \triangleq \frac{6}{\eta\sqrt{\nu\bar{\nu}'}} \log \left(1 + \frac{\gamma_+^2}{4\rho|b^{(1)}|} \right) = \frac{1}{\eta\sqrt{\nu\bar{\nu}'}} \tau_{\text{grow}}(b^{(1)}).$$

Additionally, for $t \geq T_1^{\text{sub}}$ we have $\rho\|x_t\| \geq \gamma - \frac{1}{3}\sqrt{\nu\bar{\nu}'} \geq \gamma - \frac{1}{3}\sqrt{\nu\bar{\nu}}$ because $\bar{\nu} \geq \bar{\nu}'$. Thus, using that $\nu, \bar{\nu}' \leq \bar{\nu}$ and that $(\beta + 2\rho\|x_\star\| \|x_\star\|^2)/\varepsilon \geq 2$ as in the beginning of section 4, we may define

$$T_2^{\text{sub}} \triangleq \frac{1}{\eta\bar{\nu}} \log \frac{2(\beta + \rho\|x_\star\|)}{\nu} \leq \frac{1}{\eta\sqrt{\nu\bar{\nu}'}} \log \left(\left[\frac{(\beta + 2\rho\|x_\star\|) \|x_\star\|^2}{\varepsilon} \right]^6 \right) = \frac{\tau_{\text{conv}}(\varepsilon)}{\eta\sqrt{\nu\bar{\nu}'}}.$$

Thus, $2(\beta + \rho\|x_\star\|) \|x_\star\|^2 e^{-\eta\bar{\nu}t} \leq \|x_\star\|^2\nu$ for every $t \geq T_2^{\text{sub}}$, and by Lemma 4.5 we have

$$(18) \quad \|\Pi_\nu A_\star^{1/2}(x_t - x_\star)\|^2 \leq \|x_\star\|^2\nu + 13\|x_\star\|^2\nu = 14\|x_\star\|^2\nu$$

for every $t \geq T^{\text{sub}} = T_1^{\text{sub}} + T_2^{\text{sub}}$.

We now translate the guarantee (18) on the distance from x_t to x_\star in the subspace of “large” eigenvectors of A to a guarantee on the solution quality $f(x_t)$. Using

expression (6) for $f(x)$, the orthogonality of $I - \Pi_\nu$ and Π_ν , and $\|x_t\| \leq \|x_\star\|$, we have

$$\begin{aligned} f(x_t) &\leq f(x_\star) + \frac{1}{2} \|(I - \Pi_\nu)A_\star^{\frac{1}{2}}(x_t - x_\star)\|^2 \\ &\quad + \frac{1}{2} \|\Pi_\nu A_\star^{\frac{1}{2}}(x_t - x_\star)\|^2 + \frac{\rho\|x_\star\|}{2} (\|x_\star\| - \|x_t\|)^2. \end{aligned}$$

Now we note that

$$(19) \quad \|(I - \Pi_\nu)A_\star\|_2 = \max_{i: \lambda^{(i)} < \lambda^{(1)} + \nu} |\lambda^{(i)} + \rho\|x_\star\|| \leq -\gamma + \nu + \rho\|x_\star\| \leq 2\nu,$$

where we have used our assumption (15) that $\rho\|x_\star\| - \gamma \leq \frac{\varepsilon}{10\|x_\star\|^2} = \nu$. Using this gives

$$(20) \quad f(x_t) \leq f(x_\star) + \nu\|x_t - x_\star\|^2 + 7\|x_\star\|^2\nu + \frac{\rho\|x_\star\|}{2} (\|x_\star\| - \|x_t\|)^2,$$

where we use inequality (18). Because $\rho\|x_t\| \geq \gamma - \frac{1}{3}\sqrt{\nu\bar{\nu}'}$ for $t \geq T_1^{\text{sub}}$, we obtain

$$0 \leq \rho(\|x_\star\| - \|x_t\|) \leq \rho\|x_\star\| - \gamma - (\rho\|x_t\| - \gamma) \leq \frac{4}{3}\sqrt{\nu\bar{\nu}'}. \quad (21)$$

The above inequality provides an upper bound on $(\|x_\star\| - \|x_t\|)^2$. Alternatively, we may bound $(\|x_\star\| - \|x_t\|)^2 \leq \|x_\star\|^2$ using $\|x_t\| \leq \|x_\star\|$ (Lemma 2.6). Therefore,

$$(21) \quad \frac{\rho\|x_\star\|}{2} (\|x_\star\| - \|x_t\|)^2 \leq \|x_\star\|^2 \min \left\{ \frac{\rho\|x_\star\|}{2}, \frac{16\bar{\nu}'\nu}{18\rho\|x_\star\|} \right\} \leq \|x_\star\|^2\nu,$$

where the final inequality follows as $\bar{\nu}' \leq \max\{\nu, \rho\|x_\star\|\}$. Substituting the bound (21) into (20) with $\|x_\star - x_t\|^2 \leq 2\|x_\star\|^2$ (by Lemma 2.4), we find

$$f(x_t) \leq f(x_\star) + 9\|x_\star\|^2\nu \leq f(x_t) + \varepsilon,$$

where we substitute $\nu = \frac{\varepsilon}{10\|x_\star\|^2}$. Summarizing, if $\rho\|x_\star\| - \gamma \leq \nu = \frac{\varepsilon}{10\|x_\star\|^2}$, the point x_t is ε -suboptimal for problem (1) whenever

$$t \geq \frac{\tau_{\text{grow}}(b^{(1)}) + \tau_{\text{conv}}(\varepsilon)}{\eta\sqrt{\nu\bar{\nu}'}} \geq T_1^{\text{sub}} + T_2^{\text{sub}},$$

where

$$\sqrt{\nu\bar{\nu}'} = \max \left\{ \frac{\varepsilon}{10\|x_\star\|^2}, \sqrt{\frac{\varepsilon}{10\|x_\star\|^2} \min\{\text{gap}, \rho\|x_\star\|\}} \right\}.$$

4.2. Proof of Theorem 3.2. Theorem 3.2 follows from three basic observations about the effect of adding a small uniform perturbation to b , which we summarize in the following lemma (see section B.5 for a proof).

LEMMA 4.6. *Set $\tilde{b} = b + \sigma q$, where q is uniform on the unit sphere in \mathbb{R}^d and $\sigma > 0$. Let $\tilde{f}(x) = \frac{1}{2}x^T A x + \tilde{b}^T x + \frac{1}{3}\rho\|x\|^3$ and let \tilde{x}_\star be a global minimizer of \tilde{f} . Then, for any $\delta > 0$,*

- (i) *for $d > 2$, $\mathbb{P}(|\tilde{b}^{(1)}| \leq \sqrt{\pi}\sigma\delta/\sqrt{2d}) \leq \delta$;*
- (ii) *$|f(x) - \tilde{f}(x)| \leq \sigma\|x\|$ for all $x \in \mathbb{R}^d$;*
- (iii) *$\|x_\star\|^2 - \|\tilde{x}_\star\|^2 \leq 2\sigma/\rho$.*

With Lemma 4.6 in hand, our proof can be split into three parts: in the first two, we provide bounds on the iteration complexity of each of the modes of convergence that Theorem 3.1 exhibits in the perturbed problem with vector \tilde{b} ; the final part shows that the quality of the (approximate) solutions \tilde{x}_t and \tilde{x}_\star is not much worse than x_\star .

Let \tilde{f} , \tilde{b} , and \tilde{x}_\star be as defined in Lemma 4.6. By Theorem 3.1, we know that $\tilde{f}(\tilde{x}_t) \leq \tilde{f}(\tilde{x}_\star) + \varepsilon$ for all

$$(22a) \quad t \geq \frac{6}{\eta} \left(\log \left(1 + \frac{\gamma_+^2/4}{\rho|\tilde{b}^{(1)}|} \right) + \log \frac{(\beta + 2\rho\|\tilde{x}_\star\|)\|\tilde{x}_\star\|^2}{\varepsilon} \right) \min \left\{ \frac{1}{\rho\|\tilde{x}_\star\| - \gamma}, \frac{10\|\tilde{x}_\star\|^2}{\varepsilon} \right\},$$

and that if $\rho\|\tilde{x}_\star\| - \gamma \leq \frac{\varepsilon}{10\|\tilde{x}_\star\|}$, then $\tilde{f}(\tilde{x}_t) \leq \tilde{f}(\tilde{x}_\star) + \varepsilon$ for all

$$(22b) \quad t \geq \frac{6}{\eta} \left(\log \left(1 + \frac{\gamma_+^2}{4\rho|\tilde{b}^{(1)}|} \right) + \log \frac{(\beta + 2\rho\|\tilde{x}_\star\|)\|\tilde{x}_\star\|^2}{\varepsilon} \right) \sqrt{\frac{10\|\tilde{x}_\star\|^2}{\varepsilon} \frac{1}{\min\{\text{gap}, \rho\|\tilde{x}_\star\|\}}},$$

We now turn to bounding expressions (22a) and (22b) appropriately: section 4.2.1 deals with the occurrences of $\|\tilde{x}_\star\|$ outside the logarithm, and section 4.2.2 bounds the terms $\tilde{b}^{(1)}$ and $\|\tilde{x}_\star\|$ appearing inside the logarithm.

4.2.1. Part 1: Upper bounding terms outside the log. Recalling that $\sigma = \frac{\rho\bar{\sigma}\varepsilon}{12(\beta+2\rho\|x_\star\|)}$ and $\varepsilon \leq (\frac{1}{2}\beta + \rho\|x_\star\|)\|x_\star\|^2$, we have $\sigma \leq \frac{\rho}{24}\bar{\sigma}\|x_\star\|^2$. Thus, part (iii) of Lemma 4.6 gives

$$\|x_\star\|^2 - \|\tilde{x}_\star\|^2 \leq 2\sigma/\rho \leq \bar{\sigma}\|x_\star\|^2/12, \quad \text{so} \quad \|\tilde{x}_\star\|^2 \in (1 \pm \bar{\sigma}/12)\|x_\star\|^2.$$

Consequently, using $\bar{\sigma} \leq 1$ we have

$$(23) \quad \left| \|x_\star\| - \|\tilde{x}_\star\| \right| \leq \frac{2\sigma}{\rho(\|x_\star\| + \|\tilde{x}_\star\|)} \leq \frac{2\bar{\sigma}\varepsilon}{12(1 + \sqrt{11/12})\|x_\star\|(\beta + 2\rho\|x_\star\|)} \leq \frac{\bar{\sigma}\varepsilon}{20\rho\|x_\star\|^2}.$$

Now, suppose that $\frac{\varepsilon}{10\|x_\star\|^2} \leq \rho\|x_\star\| - \gamma$. Substituting this into the bound (23) yields

$\|x_\star\| - \|\tilde{x}_\star\| \leq \frac{\bar{\sigma}}{2\rho}(\rho\|x_\star\| - \gamma)$, and rearranging, we obtain

$$\rho\|\tilde{x}_\star\| - \gamma \geq (1 - 0.5\bar{\sigma})(\rho\|x_\star\| - \gamma) \geq \frac{\rho\|x_\star\| - \gamma}{1 + \bar{\sigma}}$$

because $\bar{\sigma} \leq 1$. We combine the preceding bounds to obtain

$$(24a) \quad \min \left\{ \frac{1}{\rho\|\tilde{x}_\star\| - \gamma}, \frac{10\|\tilde{x}_\star\|^2}{\varepsilon} \right\} \leq (1 + \bar{\sigma}) \min \left\{ \frac{1}{\rho\|x_\star\| - \gamma}, \frac{10\|x_\star\|^2}{\varepsilon} \right\}$$

and

$$(24b) \quad \sqrt{\frac{10\|\tilde{x}_\star\|^2}{\varepsilon} \cdot \frac{1}{\min\{\text{gap}, \rho\|\tilde{x}_\star\|\}}} \leq (1 + \bar{\sigma}) \sqrt{\frac{10\|x_\star\|^2}{\varepsilon} \cdot \frac{1}{\min\{\text{gap}, \rho\|x_\star\|\}}},$$

where we have used $\|\tilde{x}_\star\| \leq (1 + \bar{\sigma})\|x_\star\|$ and $\|\tilde{x}_\star\| \geq \sqrt{1 - \bar{\sigma}/12}\|x_\star\| \geq \|x_\star\|/(1 + \bar{\sigma})$.

The bound (23) also implies $\rho\|\tilde{x}_\star\| - \gamma \leq \rho\|x_\star\| - \gamma + \frac{\bar{\sigma}\varepsilon}{20\|x_\star\|^2}$. When $\rho\|x_\star\| - \gamma \leq (1 - 2\bar{\sigma}/3)\frac{\varepsilon}{10\|x_\star\|^2}$, we thus have

$$\rho\|\tilde{x}_\star\| - \gamma \leq \frac{\varepsilon}{10\|x_\star\|^2} \left(1 - \frac{2}{3}\bar{\sigma} + \frac{1}{2}\bar{\sigma} \right) \leq \frac{\varepsilon(1 + \bar{\sigma}/12)(1 - \bar{\sigma}/6)}{10\|\tilde{x}_\star\|^2} \leq \frac{\varepsilon}{10\|\tilde{x}_\star\|^2},$$

where we have used $\|\tilde{x}_\star\|^2 \leq (1 + \bar{\sigma}/12)\|x_\star\|^2$. Therefore, $\tilde{f}(\tilde{x}_t) \leq \tilde{f}(\tilde{x}_\star) + \varepsilon$ whenever the conditions $\rho\|x_\star\| - \gamma \leq (1 - 2\bar{\sigma}/3)\frac{\varepsilon}{10\|x_\star\|^2}$ and (22b) hold.

4.2.2. Part 2: Upper bounding terms inside the log. Fix a confidence level $\delta \in (0, 1)$. By Lemma 4.6(i), $1/|\tilde{b}^{(1)}| \leq \sqrt{2d}/(\sqrt{\pi}\sigma\delta) \leq \sqrt{d}/(\sigma\delta)$ with probability at least $1 - \delta$, so

$$\begin{aligned} 6 \log \left(1 + \frac{\gamma_+^2}{4\rho|\tilde{b}^{(1)}|} \right) &\leq 6 \log \left(1 + \frac{\gamma_+^2 \sqrt{d}}{4\rho\sigma\delta} \right) \\ &\stackrel{(*)}{\leq} 6 \log \left(1 + \mathbb{I}_{\{\gamma > 0\}} \frac{3\sqrt{d}}{\bar{\sigma}\delta} \right) + 6 \log \frac{(\beta + 2\rho\|x_\star\|)\|x_\star\|^2}{\varepsilon} \\ &= \tilde{\tau}_{\text{grow}}(d, \delta, \bar{\sigma}) + \frac{6}{14} \tilde{\tau}_{\text{conv}}(\varepsilon), \end{aligned}$$

where inequality $(*)$ uses that $\rho\|x_\star\| \geq \gamma_+$ and $\varepsilon \leq (\beta + \frac{1}{2}\rho\|x_\star\|)\|x_\star\|^2$. Using $\|\tilde{x}_\star\| \leq \sqrt{1 + \bar{\sigma}/12}\|x_\star\|$ yields the upper bound

$$6 \log \frac{(\beta + 2\rho\|\tilde{x}_\star\|)\|\tilde{x}_\star\|^2}{\varepsilon} \leq 6 \log \frac{(\beta + 2\rho\|x_\star\|)\|x_\star\|^2}{\varepsilon} + 9 \log(1 + \bar{\sigma}/12) \leq \frac{8}{14} \tilde{\tau}_{\text{conv}}(\varepsilon),$$

where the second inequality follows as $9 \log(1 + \bar{\sigma}/12) < 2 \log 2 \leq 2 \log \frac{(\beta + 2\rho\|x_\star\|)\|x_\star\|^2}{\varepsilon}$.

Substituting the above bounds and the upper bounds (24a) and (24b) into expressions (22a) and (22b), we see that the iteration bounds claimed in Theorem 3.2 hold. To complete the proof we need only bound the quality of the solution \tilde{x}_t .

4.2.3. Part 3: Bounding solution quality. We recall that $\sigma = \frac{\rho\bar{\sigma}\varepsilon}{12(\beta + 2\rho\|x_\star\|)} \leq \frac{\bar{\sigma}\varepsilon}{24\|x_\star\|}$ and $\|\tilde{x}_\star\| \leq \sqrt{1 + \bar{\sigma}/12}\|x_\star\| \leq \sqrt{2}\|x_\star\|$, so $\sigma \leq \frac{\bar{\sigma}\varepsilon}{\|x_\star\| + \|\tilde{x}_\star\|}$. Thus, whenever $\tilde{f}(\tilde{x}_t) \leq \tilde{f}(\tilde{x}_\star) + \varepsilon$,

$$\begin{aligned} f(\tilde{x}_t) &\stackrel{(a)}{\leq} \tilde{f}(\tilde{x}_t) + \sigma\|\tilde{x}_t\| \leq \tilde{f}(\tilde{x}_\star) + \varepsilon + \sigma\|\tilde{x}_t\| \stackrel{(b)}{\leq} \tilde{f}(\tilde{x}_\star) + \varepsilon + \sigma\|\tilde{x}_\star\| \\ &\stackrel{(c)}{\leq} \tilde{f}(x_\star) + \varepsilon + \sigma\|\tilde{x}_\star\| \stackrel{(d)}{\leq} f(x_\star) + \sigma(\|\tilde{x}_\star\| + \|x_\star\|) + \varepsilon \leq f(x_\star) + (1 + \bar{\sigma})\varepsilon, \end{aligned}$$

where transitions (a) and (d) follow from part (ii) of Lemma 4.6, transition (b) follows from $\|\tilde{x}_t\| \leq \|\tilde{x}_\star\|$ (Lemma 2.6), and transition (c) follows from $\tilde{f}(\tilde{x}_\star) = \min_{z \in \mathbb{R}^d} \tilde{f}(z)$.

5. Convergence of a line search method. The maximum step size allowed by Assumption A may be too conservative (as is common with gradient descent). With that in mind, in this section we briefly analyze line search schemes of the form

$$(25) \quad x_{t+1} = x_t - \eta_t \nabla f(x_t), \quad \text{where } \eta_t = \underset{\eta \in C_t}{\operatorname{argmin}} f(x_t - \eta \nabla f(x_t))$$

and C_t is a (possibly time-varying) interval of allowed step sizes. For problem (1), η_t is computable for any interval C_t , as the critical points of the function $h(\eta) = f(x_t - \eta \nabla f(x_t))$ are roots of a quartic polynomial with coefficients determined by $\|x\|$, $\|g\|$, $g^T A g$, and $x^T g$, so η_t must be a root or an edge of the interval C_t .

The unconstrained choice $C_t = \mathbb{R}$ yields the *steepest descent* method [27]. For steepest descent it is possible that $\eta_t < 0$ and that convergence to a suboptimal local minimum of f occurs. Consequently, we propose choosing the updates (25) using the

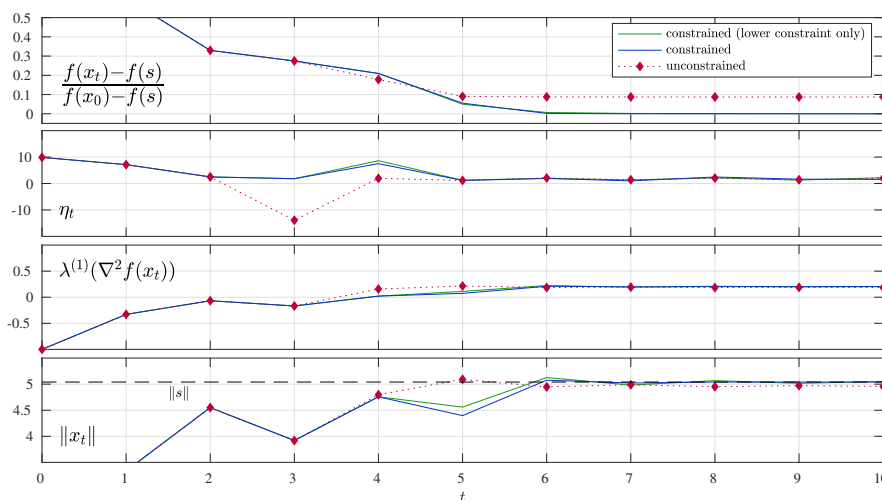


FIG. 4. Steepest descent variants applied on $A = \text{diag}([-1; -0.8; -0.5])$, $b = [0.04; 0.15; 0.3]$, and $\rho = 0.2$. The dotted red, solid green, and solid blue curves correspond to $C_t = \mathbb{R}$, $C_t = [0, \infty)$, and C_t given by (26), respectively.

interval

$$(26) \quad C_t = \left[0, \left[\frac{\nabla f(x_t)^T A \nabla f(x_t)}{\|\nabla f(x_t)\|^2} + \rho \|x_t\| \right]_+^{-1} \right].$$

Scheme (26) converges to the global minimum of f (see Appendix C for proof).

PROPOSITION 5.1. *Let x_t be the iterates of gradient descent with step sizes selected by the constrained minimization (26). Let Assumption B hold and assume $b^{(1)} \neq 0$. Then x_\star is the unique global minimizer of f and $\lim_{t \rightarrow \infty} x_t = x_\star$.*

In Figure 4, we display the quantities $f(x_t)$, η_t , $\lambda^{(1)}(\nabla^2 f(x_t))$, and $\|x_t\|$ for the above line-search variants on a three-dimensional problem instance. The step sizes differ at iteration $t = 3$, where the unconstrained gradient step makes almost 50% more progress than steps restricted to be positive. However, it then converges to a suboptimal local minimum (note $\lambda^{(1)}(\nabla^2 f(x_t)) > 0$) approximately 9% worse than the global minimum achieved by the guarded sequence (26). The step sizes these methods choose are significantly larger than the η Assumption A allows, which is approximately 0.12. Figure 4 reveals a difference between fixed-step-size gradient descent and the line-search schemes—the norm $\|x_t\|$ of the line-search-based iterates is nonmonotonic and overshoots $\|x_\star\|$. Our convergence rate proofs hinge on Lemma 2.6, i.e., that $\|x_t\|$ is increasing, so the extension of our rates to line-search schemes is not straightforward.

We believe that the rate guarantees of Theorem 3.1 also apply to the step size choice (26). To lend credence to this hypothesis, we repeat the ensemble experiment detailed in section 3.2 (Figure 3), where we use the step size given by (26) instead of the fixed step size. Figure 5 shows that the rates we prove in section 3 seem to accurately describe the behavior of guarded steepest descent as well, with constant factors.

We remark that we introduce the upper constraint (26) only because we require it in the proof of Proposition 5.1. Empirically, a scheme with the simpler constraint

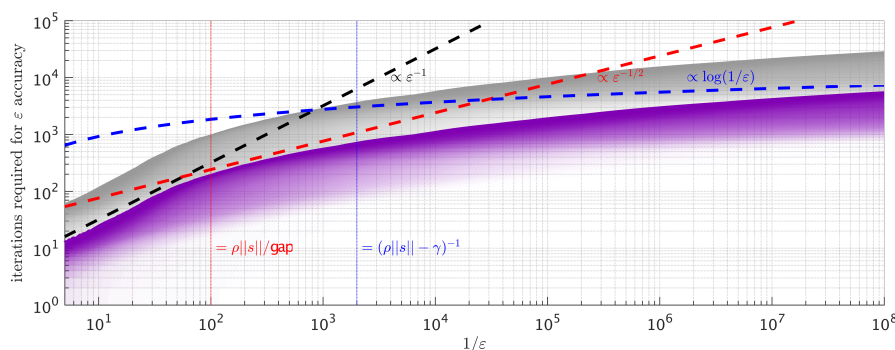


FIG. 5. Reproduction of the ensemble experiment reported in section 3.2 (Figure 3), with the scheme (26) used instead of fixed-step-size gradient descent. The cumulative distribution function for fixed step size $\eta = 0.25$ is shown in light gray for comparison.

$C_t = [0, \infty)$ appears to converge to the global minimum as well, though we remain unable to prove this. While such a step size can differ from the choice in (26) (see time $t = 4$ in Figure 4), the variants seem equally practicable. Indeed, we performed the ensemble experiment (Figures 3 and 5) with $C_t = [0, \infty)$ and the results are indistinguishable.

6. Application: A Hessian-free majorization method. In this section we use our main results to analyze a simple optimization scheme that approximates the cubic-regularized Newton steps with gradient descent. We expect more elaborate schemes to be more efficient in practice, as the current procedure is simplified and uses gradient descent rather than Krylov subspace methods (see the introduction). We believe that our analysis extends to such practical schemes beyond the scope of this paper.

We consider functions g satisfying the following.

Assumption A. The function g satisfies $\inf g \geq \underline{g} > -\infty$, is β -smooth, and has 2ρ -Lipschitz Hessian, i.e., $\|\nabla^2 g(y) - \nabla^2 g(y')\| \leq 2\rho\|y - y'\|$ for every $y, y' \in \mathbb{R}^d$.

The first two parts of Assumption A (boundedness and smoothness) are standard. The third implies that g is upper bounded by its cubic-regularized quadratic approximation [26, Lemma 1]: for $y, \Delta \in \mathbb{R}^d$ one has

$$(27) \quad g(y + \Delta) \leq g(y) + \nabla g(y)^T \Delta + \frac{1}{2} \Delta^T \nabla^2 g(y) \Delta + \frac{\rho}{3} \|\Delta\|^3.$$

For simplicity we assume that the constants β and ρ are known. From a theoretical perspective this is a benign assumption, as we may estimate these constants without significantly affecting the complexity bounds [25]. In practice, however, careful adaptive estimation of ρ is crucial for good performance; this is a primary strength of the adaptive regularization algorithm using cubics (ARC) method [6].

Following [26, 12], our goal is to find an ϵ -second-order stationary point y_ϵ :

$$(28) \quad \|\nabla g(y_\epsilon)\| \leq \epsilon \quad \text{and} \quad \nabla^2 g(y_\epsilon) \succeq -\sqrt{\rho\epsilon}I.$$

Intuitively, ϵ -second-order stationary points provide a finer approximation to local minima than ϵ -stationary points (with only $\|\nabla g(y)\| \leq \epsilon$). Throughout this section, we use ϵ to denote approximate stationarity in g , and continue to use ε to denote approximate optimality for subproblems of the form (1).

Algorithm 1 A second-order majorization method.

```

1: function SOLVE-PROBLEM( $y_0, g, \beta, \rho, \epsilon, \delta$ )
2:   Set  $K_{\text{prog}} = 1/324$  and  $\eta = 1/(10\beta)$ 
3:   for  $k = 1, 2, \dots$  do  $\triangleright$  guaranteed to terminate in at most  $O(\epsilon^{-3/2})$  iterations
4:      $\Delta_k \leftarrow \text{SOLVE-SUBPROBLEM}(\nabla^2 g(y_{k-1}), \nabla g(y_{k-1}), \rho, \eta, \sqrt{\frac{\epsilon}{9\rho}}, \frac{1}{2}, \frac{\delta}{2k^2})$ 
5:     if  $g(y_{k-1} + \Delta_k) \leq g(y_k) - K_{\text{prog}} \epsilon^{3/2} \rho^{-1/2}$  then
6:        $y_k \leftarrow y_{k-1} + \Delta_k$ 
7:     else
8:        $\Delta_k \leftarrow \text{SOLVE-FINAL-SUBPROBLEM}(\nabla^2 g(y_{k-1}), \nabla g(y_{k-1}), \rho, \eta, \frac{\epsilon}{2})$ 
9:     return  $y_{k-1} + \Delta_k$ 

```

Algorithm 2 A Hessian-free subproblem solver.

```

1: function SOLVE-SUBPROBLEM( $A, b, \rho, \eta, r, \epsilon', \delta$ )
2:   Set  $f(x) = (1/2)x^T A x + b^T x + (\rho/3)\|x\|^3$  and  $x_0 = \text{CAUCHY-POINT}(A, b, \rho)$ 
3:   if  $f(x_0) \leq -(1 - \epsilon')\rho r^3/6$  then return  $x_0$ 
4:   Set
      
$$T = \frac{240}{\eta \rho r \epsilon'} \left[ 6 \log \left( 1 + \frac{\sqrt{\delta}}{\delta} \right) + 32 \log \left( \frac{6}{\eta \rho r \epsilon'} \right) \right]$$

5:   Set  $\sigma = \frac{\rho^2 r^3 \epsilon'}{144(\beta + 2\rho r)}$ , draw  $q$  uniformly from the unit sphere, set  $\tilde{b} = b + \sigma q$ 
6:   Set  $\tilde{f}(x) = (1/2)x^T A x + \tilde{b}^T x + (\rho/3)\|x\|^3$  and  $\tilde{x}_0 = \text{CAUCHY-POINT}(A, \tilde{b}, \rho)$ 
7:   for  $t = 1, 2, \dots, T$  do
8:      $\tilde{x}_t \leftarrow \tilde{x}_{t-1} - \eta \nabla \tilde{f}(\tilde{x}_{t-1})$ 
9:     if  $f(\tilde{x}_t) \leq -(1 - \epsilon')\rho r^3/6$  then return  $\tilde{x}_t$ 
10:  return  $\tilde{x}_t$ 

1: function SOLVE-FINAL-SUBPROBLEM( $A, b, \rho, \eta, \epsilon_g$ )
2:   Set  $f(x) = (1/2)x^T A x + b^T x + (\rho/3)\|x\|^3$  and  $\Delta = \text{CAUCHY-POINT}(A, b, \rho)$ 
3:   while  $\|\nabla f(\Delta)\| > \epsilon_g$  do  $\Delta \leftarrow \Delta - \eta \nabla f(\Delta)$ 
4:   return  $\Delta$ 

1: function CAUCHY-POINT( $A, b, \rho$ )
2:   return  $-R_c b / \|b\|$ , where  $R_c = \frac{-b^T A b}{2\rho \|b\|^2} + \sqrt{\left( \frac{b^T A b}{2\rho \|b\|^2} \right)^2 + \frac{\|b\|}{\rho}}$ 

```

We outline a majorization-minimization strategy [9, 27] for optimization of g in Algorithm 1. At each iteration, the method approximately minimizes a local model of g , halting once progress in decreasing g falls below a certain threshold. In Algorithm 2, we describe a simple Hessian-free subproblem solver that uses gradient descent with a small perturbation to the linear term and *fixed step size* (as in Theorem 3.2); we write the method in terms of an input matrix $A = \nabla^2 g(y)$, noting that it requires only matrix-vector products Av implementable by a first-order oracle for g .

The method SOLVE-SUBPROBLEM takes as inputs a problem instance (A, b, ρ) , confidence level δ , relative accuracy ϵ' , and a threshold for the magnitude of the global minimizer x_* , which we denote by r . As an immediate consequence of Theorem 3.2, as long as $\|x_*\| \geq r$ the method is guaranteed to terminate before reaching line 10, and if the gradient is sufficiently large, termination occurs before entering the loop. We formalize this in the following lemma, whose proof we provide in Appendix D.1.

LEMMA 6.2. Let $A \in \mathbb{R}^{d \times d}$ satisfy $\|A\|_2 \leq \beta$, $b \in \mathbb{R}^d$, $\rho > 0$, $r > 0$, $\epsilon' \in (0, 1)$, $\delta \in (0, 1)$, and $\eta \leq 1/(8\beta + 4\rho)$. With probability at least $1 - \delta$, if

$$\|x_\star\| \geq r \text{ or } \|b\| \geq \max\{\sqrt{\beta\rho}r^{3/2}, \rho r^2\},$$

then $x = \text{SOLVE-SUBPROBLEM}(A, b, \rho, \eta, r, \epsilon', \delta)$ satisfies $f(x) \leq (1 - \epsilon')\rho r^3/6$.

Let Δ_k^\star be the global minimizer (in Δ) of the model (27) at $y = y_k$, the k th iterate of Algorithm 1. Lemma 6.2 guarantees that, with high probability, if SOLVE-SUBPROBLEM fails to meet the progress condition in line 5 at iteration k , then $\|\Delta_k^\star\| \leq \sqrt{\epsilon/(9\rho)}$, and therefore $\lambda^{(1)}(\nabla^2 g(y_k)) \geq -\rho\|\Delta_k^\star\| \geq -\sqrt{\rho\epsilon}$. It is possible, nonetheless, that $\|\nabla g(y_k)\| > \epsilon$; to address this, we correctively apply gradient descent on the final subproblem ($\text{SOLVE-FINAL-SUBPROBLEM}$).

Building on an argument from Nesterov and Polyak [26, Lemma 5], we obtain the following guarantee for Algorithm 1, whose proof we provide in Appendix D.2.

PROPOSITION 6.3. Let g satisfy Assumption A, let $y_0 \in \mathbb{R}^d$ be arbitrary, and let $\delta \in (0, 1]$ and $\epsilon \leq \min\{\beta^2/\rho, \rho^{1/3}(g(y_0) - \underline{g})^{2/3}\}$. With probability at least $1 - \delta$, Algorithm 1 finds an ϵ -second-order stationary point (28) in at most

$$(29) \quad O(1) \cdot \frac{\beta(g(y_0) - \underline{g})}{\epsilon^2} \log \left(\frac{d}{\delta} \cdot \frac{\beta(g(y_0) - \underline{g})}{\epsilon^2} \right)$$

Hessian-vector product evaluations, and at most

$$O(1) \cdot \frac{\sqrt{\rho}(g(y_0) - \underline{g})}{\epsilon^{3/2}}$$

calls to SOLVE-SUBPROBLEM and gradient evaluations.

In Proposition 6.3, the assumption $\epsilon \leq \beta^2/\rho$ gives no loss of generality, as otherwise the Hessian guarantee (28) is trivial, and we may obtain the gradient guarantee by simply running gradient descent on g for $2\beta(g(y_0) - \underline{g})\epsilon^{-2}$ iterations. Similarly, if $\epsilon > \rho^{1/3}(g(y_0) - \underline{g})^{2/3}$, then we require at most $1 + 1/K_{\text{prog}} = 325$ calls to SOLVE-SUBPROBLEM , and the proof of Proposition 6.3 reveals that the overall first-order complexity scales as $\epsilon^{-1/2}$ instead of ϵ^{-2} .

There are other Hessian-free methods that provide the guarantee (28), and recent schemes using acceleration techniques [1, 5] provide it in roughly $\epsilon^{-7/4} \log \frac{d}{\delta}$ first-order operations, which is better than Algorithm 1. Nevertheless, this section illustrates how gradient descent on the structured problem (1) can be straightforwardly leveraged to optimize general smooth nonconvex functions.

7. Discussion. Our results have a number of connections to rates of convergence in classical (smooth) convex optimization and the power method for symmetric eigenvector computation; here, we explore these in more detail.

7.1. Comparison with convex optimization. For L -smooth α -strongly convex functions, gradient descent finds an ϵ -suboptimal point within

$$O(1) \cdot \min \left\{ \frac{L}{\alpha} \log \frac{LD^2}{\epsilon}, \frac{LD^2}{\epsilon} \right\}$$

iterations [24], where D is any constant $D \geq \|x_0 - x^\star\|$ and x^\star a global minimizer. For our (possibly nonconvex) problem (1), Corollary 3.3 guarantees that gradient descent

finds an ε -suboptimal point (with probability at least $1 - \delta$) within

$$O(1) \cdot \min \left\{ \frac{L_\star}{\rho \|x_\star\| - \gamma}, \frac{L_\star \|x_\star\|^2}{\varepsilon} \right\} \left[\log \frac{L_\star \|x_\star\|^2}{\varepsilon} + \log \left(1 + \mathbb{I}_{\{\gamma > 0\}} \frac{d}{\delta} \right) \right]$$

iterations, where $L_\star = \beta + 2\rho \|x_\star\|$. The parallels are immediate: by Lemma 2.6, L_\star and $\|x_\star\|$ are precise analogues of L and D in the convex setting. Moreover, the quantity $\rho \|x_\star\| - \gamma$ plays the role of the strong convexity parameter α , but it is well-defined even when f is not convex. When $\lambda^{(1)}(A) = -\gamma \geq 0$, f is $-\gamma$ -strongly convex, and because $\rho \|x_\star\| - \gamma > -\gamma$, our analysis for the cubic problem (1) guarantees better conditioning than the generic convex result. The difference between $\rho \|x_\star\| - \gamma$ and $-\gamma$ becomes significant when b is sufficiently large, as we observe from the bounds (7b) and (8). Even in the nonconvex case when $\gamma > 0$, gradient descent still exhibits linear convergence whenever high accuracy is desired, that is, when $\varepsilon / \|x_\star\|^2 \leq \rho \|x_\star\| - \gamma$.

When $\gamma > 0$, our guarantee becomes probabilistic and contains a $\log(d/\delta)$ term. Such a term does not appear in results on convex optimization [24], and it is fundamentally related to the presence of saddle points in the objective [31].

7.2. Comparison with the power method. The power method for finding the smallest eigenvector of A is the recursion $x_{t+1} = (I - (1/\beta)A)x_t / \|(I - (1/\beta)A)x_t\|$, where x_0 is uniform on the unit sphere [18, 22]. This method guarantees that, with probability at least $1 - \delta$, $x_t^T A x_t \leq -\gamma + \varepsilon$ for all

$$t \geq O(1) \cdot \min \left\{ \frac{\beta}{\varepsilon} \log \left(\frac{d}{\delta} \right), \frac{\beta}{\text{gap}} \log \left(\frac{\beta}{\varepsilon} \cdot \frac{d}{\delta} \right) \right\}.$$

When $b = 0$ and $\lambda^{(1)}(A) = -\gamma < 0$, any global minimizer of problem (1) is an eigenvector of A with eigenvalue $-\gamma$ and $\rho \|x_\star\| = \gamma$, so it is natural to compare gradient descent and the power method. For simplicity, let us assume that $\rho = \gamma$ so that $\|x_\star\| = 1$, and both methods converge to unit eigenvectors. Under these assumptions $f(x) = \frac{1}{2}x^T A x + \frac{\gamma}{3}\|x\|^3$ and $f(x_\star) = -\gamma/6$, so $f(x) \leq f(x_\star) + \varepsilon'$ implies

$$\frac{x^T A x}{\|x\|^2} \leq -\frac{\gamma}{3} \left[\frac{1}{\|x\|^2} + 2\|x\| \right] + \frac{2\varepsilon'}{\|x\|^2} \leq -\gamma + \frac{2\varepsilon'}{\|x\|^2}.$$

Consider gradient descent applied to f with a random perturbation as described in Theorem 3.2, with $\bar{\sigma} = 1$. Inspecting the proofs of our theorems (section 4), we see that Lemmas 4.3 and 4.6 imply that, with probability at least $1 - \delta$, we have $\|\tilde{x}_t\| \geq 1/2$ for every $t \geq O(1) \log(\frac{L_\star \|x_\star\|^2}{\varepsilon} \cdot \frac{d}{\delta})$. As in Corollary 3.3, setting $\eta = \frac{1}{4(\beta + \rho R)} = \frac{1}{8\beta}$ guarantees that, with probability at least $1 - \delta$, $\tilde{x}_t^T A \tilde{x}_t / \|\tilde{x}_t\|^2 \leq -\gamma + \varepsilon$ for all

$$t \geq O(1) \cdot \min \left\{ \frac{\beta}{\varepsilon} \log \left(\frac{\beta}{\varepsilon} \cdot \frac{d}{\delta} \right), \frac{\beta}{\sqrt{\varepsilon} \min\{\text{gap}, \gamma\}} \log \left(\frac{\beta}{\varepsilon} \cdot \frac{d}{\delta} \right) \right\}.$$

Comparing the rates of convergence, we see that both exhibit the $\log(d/\delta)$ hallmark of nonconvexity and gap-free and gap-dependent convergence regimes. Of course, the power method also finds eigenvectors when $\gamma < 0$, while the unique solution to problem (1) when $b = 0$ and $\gamma < 0$ is simply $x_\star = 0$. In the gap-dependent regime, however, the power method enjoys linear convergence when $\varepsilon < \text{gap}$, while our bounds have a $1/\sqrt{\varepsilon}$ factor. Although this may be due to looseness in our analysis, we suspect it is real and related to the fact that gradient descent needs to “grow”

the iterates to have norm $\|x_t\| \approx \gamma_+/\rho$, while the power method iterates always have unit norm. If one is only interested in finding eigenvectors of A , there is probably no reason to prefer the cubic-regularized objective to the power method.

Appendix A. Proof of Lemma 2.3. Before proving Lemma 2.3, we state and prove two technical lemmas (see section A.1 for the proof conditional on these lemmas). For the first lemma, let $\kappa \in \mathbb{R}^d$ satisfy $\kappa^{(1)} \leq \kappa^{(2)} \leq \dots \leq \kappa^{(d)}$, let ν_t be a nonnegative and nondecreasing sequence, $0 \leq \nu_1 \leq \nu_2 \leq \dots$, and consider the process

$$(30) \quad z_t^{(i)} = (1 - \kappa^{(i)} - \nu_{t-1})z_{t-1}^{(i)} + 1.$$

Additionally, assume $1 - \kappa^{(i)} - \nu_{t-1} \geq 0$ for all i and t .

LEMMA A.1. *Let $z_0^{(i)} = c_0 \geq 0$ for every $i \in [d]$. Then, for every $t \in \mathbb{N}$ and $j \in [d]$, the following hold.*

- (i) *If $z_t^{(j)} \leq z_{t-1}^{(j)}$, then we also have $z_{t'}^{(j)} \leq z_{t'-1}^{(j)}$ for every $t' > t$.*
- (ii) *If $z_t^{(j)} \geq z_{t-1}^{(j)}$, then $z_t^{(j)}/z_{t+1}^{(j)} \geq z_t^{(i)}/z_{t+1}^{(i)}$ for every $i \leq j$.*
- (iii) *If $z_{t+1}^{(i)} \leq z_t^{(i)}$, then $z_{t+1}^{(j)} \leq z_t^{(j)}$ for every $j \geq i$.*

Proof. For shorthand, we define $\delta_t^{(i)} \triangleq \kappa^{(i)} + \nu_t$.

We first establish part (i) of the lemma. By (30), we have

$$z_{t+1}^{(j)} - z_t^{(j)} = (1 - \delta_{t-1}^{(j)})(z_t^{(j)} - z_{t-1}^{(j)}) - (\delta_t^{(j)} - \delta_{t-1}^{(j)})z_t^{(j)}.$$

By our assumptions that $z_0^{(j)} \geq 0$ and that $1 - \delta_t^{(j)} \geq 0$ for every t we immediately have that $z_t^{(j)} \geq 0$, and therefore also that $(\delta_t^{(j)} - \delta_{t-1}^{(j)})z_t^{(j)} = (\nu_t - \nu_{t-1})z_t^{(j)} \geq 0$. We therefore conclude that

$$z_{t+1}^{(j)} - z_t^{(j)} \leq (1 - \delta_{t-1}^{(j)})(z_t^{(j)} - z_{t-1}^{(j)}) \leq 0,$$

and induction gives part (i).

To establish part (ii) of the lemma, first note that by the contrapositive of part (i), $z_t^{(j)} \geq z_{t-1}^{(j)}$ for some t implies $z_{t'}^{(j)} \geq z_{t'-1}^{(j)}$ for any $t' \leq t$. We prove by induction that

$$(31) \quad z_{t'}^{(i)} - z_{t'}^{(j)} \leq (\kappa^{(j)} - \kappa^{(i)})z_{t'}^{(i)}z_{t'}^{(j)}$$

for any $i \leq j$ and $t' \leq t$. The basis of the induction is immediate from the assumption $z_0^{(i)} = z_0^{(j)} \geq 0$. Assuming the property holds through time $t' - 1$ for $t' \leq t$, we obtain

$$\begin{aligned} \frac{z_{t'}^{(i)} - z_{t'}^{(j)}}{z_{t'}^{(i)}z_{t'}^{(j)}} &= \frac{(1 - \delta_{t'-1}^{(i)})(z_{t'-1}^{(i)} - z_{t'-1}^{(j)}) + (\delta_{t'-1}^{(j)} - \delta_{t'-1}^{(i)})z_{t'-1}^{(j)}}{z_{t'}^{(i)}z_{t'}^{(j)}} \\ &\leq \frac{(1 - \delta_{t'-1}^{(i)})(\kappa^{(j)} - \kappa^{(i)})z_{t'-1}^{(i)}z_{t'-1}^{(j)}}{z_{t'}^{(i)}z_{t'}^{(j)}} = (\kappa^{(j)} - \kappa^{(i)})\frac{z_{t'-1}^{(j)}}{z_{t'}^{(j)}} \leq \kappa^{(j)} - \kappa^{(i)}, \end{aligned}$$

where the first inequality uses inequality (31) (assumed by induction) and the second uses $z_{t'-1}^{(j)} \leq z_{t'}^{(j)}$ for any $t' \leq t$, as argued above. With the bound

$$z_t^{(i)} - z_t^{(j)} \leq (\kappa^{(j)} - \kappa^{(i)})z_t^{(i)}z_t^{(j)}$$

in place, we may finish the proof of part (ii) by noting that

$$\frac{z_t^{(j)}}{z_{t+1}^{(j)}} - \frac{z_t^{(i)}}{z_{t+1}^{(i)}} = \frac{z_{t+1}^{(i)}z_t^{(j)} - z_{t+1}^{(j)}z_t^{(i)}}{z_{t+1}^{(j)}z_{t+1}^{(i)}} = \frac{(\kappa^{(j)} - \kappa^{(i)})z_t^{(i)}z_t^{(j)} - (z_t^{(i)} - z_t^{(j)})}{z_{t+1}^{(j)}z_{t+1}^{(i)}} \geq 0.$$

Lastly, we prove part (iii). If $z_t^{(j)} \leq z_{t-1}^{(j)}$, then we have $z_{t+1}^{(j)} \leq z_t^{(j)}$ by part (i). Otherwise we have $z_t^{(j)} \geq z_{t-1}^{(j)}$, and so $z_t^{(j)}/z_{t+1}^{(j)} \geq z_t^{(i)}/z_{t+1}^{(i)}$ by part (ii). As $z_{t+1}^{(i)} \leq z_t^{(i)}$, this implies $z_t^{(j)}/z_{t+1}^{(j)} \geq z_t^{(i)}/z_{t+1}^{(i)} \geq 1$ and therefore $z_{t+1}^{(j)} \leq z_t^{(j)}$, as required. \square

Our second technical lemma provides a lower bound on certain inner products in the gradient descent iterations. In the lemma, we recall the definition (7a) of R .

LEMMA A.2. *Assume that $\|x_\tau\|$ is nondecreasing in τ for $\tau \leq t$, that $\|x_t\| \leq R$, and that $x_t^T \nabla f(x_t) \leq 0$. Then $x_t^T A \nabla f(x_t) \geq \beta x_t^T \nabla f(x_t)$.*

Proof. If we define $z_t^{(i)} = x_t^{(i)} / (-\eta b^{(i)})$, then evidently

$$z_{t+1}^{(i)} = (1 - \underbrace{\eta \lambda^{(i)}(A)}_{\triangleq \kappa^{(i)}} - \underbrace{\eta \rho \|x_t\|}_{\triangleq \nu_t}) z_t^{(i)} + 1.$$

We verify that $z_t^{(i)}$ satisfies the conditions of Lemma A.1.

- (i) By definition, $\kappa^{(i)}$ are increasing in i , and $\nu_0 \leq \nu_1 \leq \dots \leq \nu_t$ by our assumption that $\|x_\tau\|$ is nondecreasing for $\tau \leq t$.
- (ii) As $\eta \leq 1/(\beta + \rho R)$ for $\tau \leq t$, we have that $\kappa^{(i)} + \nu_\tau \leq 1$ for $\tau \leq t$ and $i \in [d]$.
- (iii) As $x_0 = -rb/\|b\|$, $z_0^{(i)} = r/(\eta\|b\|) \geq 0$ for every i .

We may therefore apply part (iii) of Lemma A.1 to conclude that $z_t^{(i)} - z_{t+1}^{(i)} \geq 0$ implies $z_t^{(j)} - z_{t+1}^{(j)} \geq 0$ for every $j \geq i$. Since $z_t^{(i)} \geq 0$ for every i ,

$$\text{sign} \left(x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) \right) = \text{sign} \left(z_t^{(i)} \left(z_t^{(i)} - z_{t+1}^{(i)} \right) \right) = \text{sign} \left(z_t^{(i)} - z_{t+1}^{(i)} \right),$$

and there must thus exist some $i^* \in [d]$ such that $x_t^{(i)}(x_t^{(i)} - x_{t+1}^{(i)}) \leq 0$ for every $i \leq i^*$ and $x_t^{(i)}(x_t^{(i)} - x_{t+1}^{(i)}) \geq 0$ for every $i > i^*$. We thus have (by expanding in the eigenbasis of A) that

$$\begin{aligned} x_t^T A \nabla f(x_t) &= \frac{1}{\eta} \sum_{i=1}^{i^*} \lambda^{(i)}(A) x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) + \frac{1}{\eta} \sum_{i=i^*+1}^d \lambda^{(i)}(A) x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) \\ &\geq \lambda^{(i^*)}(A) \frac{1}{\eta} \sum_{i=1}^{i^*} x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) + \lambda^{(i^*+1)}(A) \frac{1}{\eta} \sum_{i=i^*+1}^d x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) \\ &\geq \lambda^{(i^*)}(A) \frac{1}{\eta} \sum_{i=1}^d x_t^{(i)} \left(x_t^{(i)} - x_{t+1}^{(i)} \right) = \lambda^{(i^*)}(A) x_t^T \nabla f(x_t) \geq \beta x_t^T \nabla f(x_t), \end{aligned}$$

where the first two inequalities use the fact the $\lambda^{(i)}$ is nondecreasing with i , and the last inequality uses our assumption that $x_t^T \nabla f(x_t) \leq 0$ along with $\lambda^{(d)}(A) \leq \beta$. \square

A.1. Proof of Lemma 2.3. By definition of the gradient descent iteration, we have

$$(32) \quad \|x_{t+1}\|^2 = \|x_t\|^2 - 2\eta x_t^T \nabla f(x_t) + \eta^2 \|\nabla f(x_t)\|^2,$$

and therefore if we can show that $x_t^T \nabla f(x_t) \leq 0$ for all t , the lemma holds. We give a proof by induction. The basis of the induction $x_0^T \nabla f(x_0) \leq 0$ is immediate as $r \mapsto f(-rb/\|b\|)$ is decreasing until $r = R_c$ (recall the definition (8)), and $x_0^T \nabla f(x_0) = 0$ for $r \in \{0, R_c\}$. Our induction assumption is that $x_{t'-1}^T \nabla f(x_{t'-1}) \leq 0$ (and hence also $\|x_{t'}\| \geq \|x_{t'-1}\|$) for $t' \leq t$ and we wish to show that $x_t^T \nabla f(x_t) \leq 0$. Note that

$$x^T \nabla f(x) = x^T A x + \rho \|x\|^3 + b^T x \geq \rho \|x\|^3 - \gamma \|x\|^2 - \|b\| \|x\|$$

and therefore $x^T \nabla f(x) > 0$ for every $\|x\| > R_{\text{low}} \triangleq \frac{\gamma}{2\rho} + \sqrt{(\frac{\gamma}{2\rho})^2 + \frac{\|b\|}{\rho}}$. Therefore, our induction assumption also implies $\|x_{t'-1}\| \leq R_{\text{low}} \leq R$ for every $t' \leq t$.

Using that $\nabla^2 f$ is 2ρ -Lipschitz, a Taylor expansion immediately implies [26, Lemma 1] that, for all vectors Δ , we have

$$(33) \quad \|\nabla f(x + \Delta) - (\nabla f(x) + \nabla^2 f(x)\Delta)\| \leq \rho \|\Delta\|^2.$$

Thus, if we define $\Delta_t \triangleq \frac{1}{\eta^2} [\nabla f(x_t) - (\nabla f(x_{t-1}) - \eta \nabla^2 f(x_{t-1}) \nabla f(x_{t-1}))]$, we have $\|\Delta_t\| \leq \rho \|\nabla f(x_{t-1})\|^2$, and using the iteration $x_t = x_{t-1} - \eta \nabla f(x_{t-1})$ yields

$$(34) \quad \begin{aligned} x_t^T \nabla f(x_t) &= x_{t-1}^T \nabla f(x_{t-1}) - \eta \|\nabla f(x_{t-1})\|^2 - \underbrace{\eta x_{t-1}^T \nabla^2 f(x_{t-1}) \nabla f(x_{t-1})}_{\triangleq \mathcal{T}_1} \\ &\quad + \underbrace{\eta^2 \nabla f(x_{t-1})^T \nabla^2 f(x_{t-1}) \nabla f(x_{t-1})}_{\triangleq \mathcal{T}_2} + \underbrace{\eta^2 x_t^T \Delta_t}_{\triangleq \mathcal{T}_3}. \end{aligned}$$

We bound each of the terms \mathcal{T}_i in turn. We have that

$$\begin{aligned} \mathcal{T}_1 &= x_{t-1}^T \nabla^2 f(x_{t-1}) \nabla f(x_{t-1}) = x_{t-1}^T A \nabla f(x_{t-1}) + 2\rho \|x_{t-1}\| x_{t-1}^T \nabla f(x_{t-1}) \\ &\geq (\beta + 2\rho \|x_{t-1}\|) x_{t-1}^T \nabla f(x_{t-1}) \geq (\beta + 2\rho R) x_{t-1}^T \nabla f(x_{t-1}), \end{aligned}$$

where both inequalities follow from the induction assumption; the first is Lemma A.2 and the second is due to $\|x_{t-1}\| \leq R$ and $x_{t-1}^T \nabla f(x_{t-1}) \leq 0$.

Treating the second-order term \mathcal{T}_2 , we obtain that

$$\mathcal{T}_2 \leq \|\nabla^2 f(x_{t-1})\|_2 \|\nabla f(x_{t-1})\|^2 \leq (\beta + 2\rho R) \|\nabla f(x_{t-1})\|^2,$$

and, by the Lipschitz bound (33), the remainder term \mathcal{T}_3 satisfies

$$\begin{aligned} \mathcal{T}_3 &= x_t^T \Delta_t \leq \|x_t\| \|r\| \leq \rho \|x_t\| \|\nabla f(x_{t-1})\|^2 \leq \rho \|x_{t-1} - \eta \nabla f(x_{t-1})\| \|\nabla f(x_{t-1})\|^2 \\ &\leq \rho \|x_{t-1}\| \|\nabla f(x_{t-1})\|^2 + \rho \eta \|\nabla f(x_{t-1})\|^3. \end{aligned}$$

Using that $\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(x_*)\| \leq (\beta + 2R) \|x - x_*\| \leq R(\beta + 2\rho R)$ for $\|x\| \leq R$ and that $\eta \leq 1/(2(\beta + 2\rho R))$, our inductive assumption that $\|x_{t-1}\| \leq R$ thus guarantees that $\mathcal{T}_3 \leq 2\rho R \|\nabla f(x_{t-1})\|^2$. Combining our bounds on the terms \mathcal{T}_i in (34), we have that

$$x_t^T \nabla f(x_t) \leq (1 - \eta(\beta + 2\rho R)) x_{t-1}^T \nabla f(x_{t-1}) - (\eta - \eta^2(\beta + 4\rho R)) \|\nabla f(x_{t-1})\|^2.$$

Using $\eta \leq 1/(\beta + 4\rho R)$ shows that $x_t^T \nabla f(x_t) \leq 0$, completing our induction. By the expansion (32), we have $\|x_t\| \leq \|x_{t+1}\|$ as desired, and that $x_t^T \nabla f(x_t) \leq 0$ for all t guarantees that $\|x_t\| \leq R_{\text{low}} \leq R$.

Appendix B. Proofs of technical results from section 4. As in the statement of our major theorems and as we note in the beginning of section 4, we tacitly assume Assumptions A and B hold throughout this section.

B.1. Proof of Lemma 4.1. Expanding $x_t = x_{t-1} - \eta \nabla f(x_{t-1})$, we have

$$(35) \quad \|x_t - x_*\|^2 = \|x_{t-1} - x_*\|^2 - 2\eta (x_{t-1} - x_*)^T \nabla f(x_{t-1}) + \eta^2 \|\nabla f(x_{t-1})\|^2.$$

Using the equality $\nabla f(x) = A_\star(x - x_\star) - \rho(\|x_\star\| - \|x\|)x$, we rewrite the cross-term $(x_{t-1} - x_\star)^T \nabla f(x_{t-1})$ as

$$\begin{aligned} & (x_{t-1} - x_\star)^T A_\star(x_{t-1} - x_\star) + \rho(\|x_{t-1}\| - \|x_\star\|)(\|x_{t-1}\|^2 - x_\star^T x_{t-1}) \\ &= (x_{t-1} - x_\star)^T \left(A_\star + \frac{\rho}{2}(\|x_{t-1}\| - \|x_\star\|)I \right) (x_{t-1} - x_\star) \\ &+ \frac{\rho}{2}(\|x_\star\| - \|x_{t-1}\|)^2(\|x_{t-1}\| + \|x_\star\|). \end{aligned} \quad (36)$$

Moving to the second-order term $\|\nabla f(x_{t-1})\|^2$ from the expansion (35), we find

$$\begin{aligned} \|\nabla f(x_{t-1})\|^2 &= \|A_\star(x_{t-1} - x_\star) + \rho(\|x_{t-1}\| - \|x_\star\|)x_{t-1}\|^2 \\ &\leq 2(x_{t-1} - x_\star)^T A_\star^2(x_{t-1} - x_\star) + 2\rho^2(\|x_{t-1}\| - \|x_\star\|)^2\|x_{t-1}\|^2. \end{aligned}$$

Combining this inequality with the cross-term calculation (36) and the squared distance (35) we obtain

$$\begin{aligned} \|x_t - x_\star\|^2 &\leq (x_{t-1} - x_\star)^T (I - 2\eta A_\star(I - \eta A_\star) - \eta\rho(\|x_{t-1}\| - \|x_\star\|)I) (x_{t-1} - x_\star) \\ &- \eta\rho(\|x_\star\| - \|x_{t-1}\|)^2(\|x_{t-1}\|(1 - 2\eta\rho\|x_{t-1}\|) + \|x_\star\|). \end{aligned}$$

Using $\eta \leq \frac{1}{4(\beta + \rho R)} \leq \frac{1}{4\|A_\star\|_2}$ yields $2\eta A_\star(I - \eta A_\star) \succeq \frac{3}{2}\eta A_\star \succeq \frac{3}{2}\eta(-\gamma + \rho\|x_\star\|)I$, so

$$\begin{aligned} \|x_t - x_\star\|^2 &\leq \left(1 - \frac{\eta}{2}[-3\gamma + \rho(\|x_\star\| + 2\|x_{t-1}\|)]\right)\|x_{t-1} - x_\star\|^2 \\ &- \eta\rho(\|x_\star\| - \|x_{t-1}\|)^2\|x_\star\|. \end{aligned}$$

B.2. Proof of Lemma 4.3. The claim is trivial when $\gamma \leq 0$ as it clearly implies $\rho\|x_t\| \geq \gamma$, so we assume $\gamma_+ = \gamma > 0$. Using the statement in Proposition 2.5 that gradient descent is convergent, we may define $t^\star = \max\{t : \rho\|x_t\| \leq \gamma - \nu\}$. Then, for every $t \leq t^\star$, the gradient descent iteration (9) satisfies

$$\begin{aligned} \frac{x_t^{(1)}}{-\eta b^{(1)}} &= (1 + \eta\gamma - \eta\rho\|x_{t-1}\|) \frac{x_{t-1}^{(1)}}{-\eta b^{(1)}} + 1 \\ &\geq (1 + \eta\nu) \frac{x_{t-1}^{(1)}}{-\eta b^{(1)}} + 1 \geq \dots \geq \frac{1}{\eta\nu} \left((1 + \eta\nu)^t - 1 \right). \end{aligned}$$

Multiplying both sides of the equality by $\eta|b^{(1)}|$ and using that $x_t^{(1)}b^{(1)} \leq 0$, we have

$$\frac{\gamma - \nu}{\rho} \geq \|x_{t^\star}\| \geq |x_{t^\star}^{(1)}| \geq \frac{|b^{(1)}|}{\nu} \left((1 + \eta\nu)^{t^\star} - 1 \right).$$

Consequently,

$$t^\star \leq \frac{\log\left(1 + \frac{(\gamma - \nu)\nu}{\rho|b^{(1)}|}\right)}{\log(1 + \eta\nu)} \leq \frac{2}{\eta\nu} \log\left(1 + \frac{\gamma_+^2}{4\rho|b^{(1)}|}\right),$$

where we have used $\eta\nu \leq \eta\gamma \leq \gamma/\beta \leq 1$, whence $\log(1 + \eta\nu) \geq \frac{\eta\nu}{2}$, and $\gamma\nu - \nu^2 \leq \sup_{x \geq 0} \{x(\gamma - x)\} \leq \frac{\gamma_+^2}{4}$.

B.3. Proof of Lemma 4.4. For typographical convenience, we prove the result with $t + 1$ replacing t . Using the commutativity of Π and A , we have $\Pi A_\star = A_\star \Pi$, so

$$(37) \quad \begin{aligned} \|\Pi A_\star^{1/2} (x_{t+1} - x_\star)\|^2 &= \|\Pi A_\star^{1/2} (x_t - x_\star)\|^2 \\ &\quad - 2\eta (x_t - x_\star)^T A_\star \Pi \nabla f(x_t) + \eta^2 \|\Pi A_\star^{1/2} \nabla f(x_t)\|^2. \end{aligned}$$

We substitute $\nabla f(x) = A_\star(x - x_\star) - \rho(\|x_\star\| - \|x\|)x$ in the cross term to obtain

$$\begin{aligned} (x_t - x_\star)^T \Pi A_\star \nabla f(x_t) &= (x_t - x_\star)^T \Pi A_\star^2 \Pi (x_t - x_\star) - \rho(\|x_\star\| - \|x_t\|) x_t^T \Pi A_\star (x_t - x_\star). \end{aligned}$$

Substituting $A_\star(x - x_\star) = \nabla f(x) + \rho(\|x_\star\| - \|x\|)x$ in the last term yields

$$(38) \quad x_t^T \Pi A_\star (x_t - x_\star) = x_t^T \Pi \nabla f(x_t) + \rho(\|x_\star\| - \|x_t\|) \|\Pi x_t\|^2.$$

Invoking Lemma 2.6 and the fact that $x_t^T \nabla f(x_t) \leq 0$, we get

$$\begin{aligned} x_t^T \Pi \nabla f(x_t) &= x_t^T \nabla f(x_t) - x_t^T (I - \Pi) \nabla f(x_t) \\ &\leq -x_t^T (I - \Pi) \nabla f(x_t) \\ &= -x_t^T (I - \Pi) A_\star (x_t - x_\star) + \rho(\|x_\star\| - \|x_t\|) \|(I - \Pi) x_t\|^2 \\ &\leq \|(I - \Pi) A_\star\|_2 \|x_t\| \|x_t - x_\star\| + \rho(\|x_\star\| - \|x_t\|) \|(I - \Pi) x_t\|^2 \\ &\leq \sqrt{2} \|(I - \Pi) A_\star\|_2 \|x_\star\|^2 + \rho(\|x_\star\| - \|x_t\|) \|(I - \Pi) x_t\|^2, \end{aligned}$$

where in the last line we used $x_t^T x_\star \geq 0$ (by Lemma 2.4). Combining this with the cross terms (38), we find that

$$(39a) \quad x_t^T \Pi A_\star (x_t - x_\star) \leq \sqrt{2} \|(I - \Pi) A_\star\|_2 \|x_\star\|^2 + \rho(\|x_\star\| - \|x_t\|) \|x_t\|^2.$$

Moving on to the second-order term in the expansion (37), we have

$$(39b) \quad \begin{aligned} \|\Pi A_\star^{1/2} \nabla f(x_t)\|^2 &= \|\Pi A_\star^{3/2} (x_t - x_\star) + \rho(\|x_t\| - \|x_\star\|) A_\star^{1/2} \Pi x_t\|^2 \\ &\leq 2\|\Pi A_\star^{3/2} (x_t - x_\star)\|^2 + 2\rho^2 \|\Pi A_\star\|_2 (\|x_t\| - \|x_\star\|)^2 \|x_t\|^2. \end{aligned}$$

Substituting the bounds (39a) and (39b) into the expansion (37), we have

$$\begin{aligned} \|\Pi A_\star^{1/2} (x_{t+1} - x_\star)\|^2 &\leq (x_t - x_\star)^T (I - 2\eta \Pi A_\star (I - \eta \Pi A_\star)) \Pi A_\star (x_t - x_\star) \\ &\quad + 2\eta \rho (\|x_\star\| - \|x_t\|) [\sqrt{2} \|(I - \Pi) A_\star\|_2 \|x_\star\|^2 \\ &\quad + (1 + \eta \|\Pi A_\star\|_2) \rho (\|x_t\| - \|x_\star\|) \|x_t\|^2]. \end{aligned}$$

Using $\eta \leq 1/(4(\beta + \rho R))$, which guarantees $0 \preceq \eta \Pi A_\star \preceq I/4 \prec I/2$, together with the assumption that $\Pi A_\star \succeq \nu \Pi$ gives

$$0 \preceq I - 2\eta \Pi A_\star (I - \eta \Pi A_\star) \preceq (1 - \eta \nu) I$$

and therefore

$$\begin{aligned} \|\Pi A_\star^{1/2} (x_{t+1} - x_\star)\|^2 &\leq (1 - \eta \nu) \|\Pi A_\star^{1/2} (x_t - x_\star)\|^2 \\ &\quad + \sqrt{8} \eta \rho (\|x_\star\| - \|x_t\|) [\rho (\|x_\star\| - \|x_t\|) \|x_t\|^2 + \|(I - \Pi) A_\star\|_2 \|x_\star\|^2]. \end{aligned}$$

B.4. Proof of Lemma 4.5. The conditions of the lemma imply that, for $\tau \geq 0$,

$$\rho(\|x_\star\| - \|x_{t+\tau}\|) \leq 4\sqrt{\nu\bar{\nu}}/3$$

and also that $\|(I - \Pi_\nu)A_\star\|_2 \leq 2\nu \leq 2\sqrt{\nu\bar{\nu}}$ (see (19)), and $\Pi_\nu A_\star \succeq \bar{\nu}I$. Substituting these bounds into Lemma 4.4 along with $\|x_{t-1}\| \leq \|x_\star\|$ (Lemma 2.6), we get

$$\|\Pi_\nu A_\star^{1/2}(x_{t+\tau} - x_\star)\|^2 \leq (1 - \eta\bar{\nu}) \|\Pi_\nu A_\star^{1/2}(x_{t+\tau-1} - x_\star)\|^2 + 13\eta\nu\bar{\nu}\|x_\star\|^2.$$

Iterating this τ times gives

$$\begin{aligned} \|\Pi_\nu A_\star^{1/2}(x_{t+\tau} - x_\star)\|^2 &\leq (1 - \eta\bar{\nu})^\tau \|\Pi_\nu A_\star^{1/2}(x_t - x_\star)\|^2 + 13\nu\|x_\star\|^2(1 - (1 - \eta\bar{\nu})^\tau) \\ &\leq 2(\beta + \rho\|x_\star\|)\|x_\star\|^2 e^{-\eta\bar{\nu}\tau} + 13\|x_\star\|^2\nu, \end{aligned}$$

where the last transition uses that

$$\|\Pi_\nu A_\star^{1/2}(x_t - x_\star)\|^2 \leq \|A_\star\|_2\|x_t - x_\star\|^2 \leq (\beta + \rho\|x_\star\|)2\|x_\star\|^2.$$

B.5. Proof of Lemma 4.6. To establish part (i) of the lemma, note that marginally $[q^{(1)}]^2 \sim \text{Beta}(\frac{1}{2}, \frac{d-1}{2})$ and that $q^{(1)}$ is symmetrically distributed. Therefore, for $d > 2$ the density of $\tilde{b}^{(1)} = b^{(1)} + \sigma q^{(1)}$ is maximal at $b^{(1)}$ and is monotonically decreasing in the distance from $b^{(1)}$. Therefore, we have

$$\mathbb{P}\left(|\tilde{b}^{(1)}| \leq \sigma\sqrt{\pi}\delta/\sqrt{2d}\right) \leq \mathbb{P}\left(|q^{(1)}| \leq \sqrt{\pi}\delta/\sqrt{2d}\right) \leq \delta,$$

where the bound $p_1(u) \leq \sqrt{d/(2\pi u)}$ on the density p_1 of $q^{(1)}$ yields the last inequality.

Part (iii) of the lemma is immediate, as

$$|f(x) - \tilde{f}(x)| = |(b - \tilde{b})^T x| \leq \sigma\|q\|\|x\| = \sigma\|x\|.$$

Part (ii) of the lemma follows by viewing $\|x_\star\|^2$ as a function of b and noting that $b \mapsto \|x_\star\|^2$ is $2/\rho$ -Lipschitz continuous. To see this claim, we use the inverse function theorem. First, note that $\|x_\star\|^2$ is a well-defined function of b , because x_\star is not unique only when $\|x_\star\| = \gamma/\rho$ (see Proposition 2.1). Next, from the relation $b = -A_\star x_\star$ we see that the inverse mapping $x_\star \mapsto b$ is a smooth function, with Jacobian

$$\frac{\partial b}{\partial x_\star} = -A_\star - \rho \frac{x_\star x_\star^T}{\|x_\star\|} = -\nabla^2 f(x_\star).$$

Let us now evaluate $\partial\|x_\star\|^2/\partial b$ when the mapping $x_\star \mapsto b(x_\star) = -(A + \rho\|x_\star\|I)x_\star$ is invertible (i.e., in the case when $\|x_\star\| > \gamma/\rho$); the inverse function theorem yields

$$\frac{\partial\|x_\star\|^2}{\partial b} = \frac{\partial(x_\star^T x_\star)}{\partial b} = 2 \frac{\partial x_\star}{\partial b} x_\star = -2(\nabla^2 f(x_\star))^{-1} x_\star.$$

The mapping $x_\star \mapsto (\nabla^2 f(x_\star))^\dagger x_\star$ is continuous in x_\star even when $A_\star \succeq 0$ is singular, and therefore the preceding expression is valid (as the natural limit) when $\|x_\star\| \rightarrow \gamma/\rho$. Moreover, since $\nabla^2 f(x_\star) \succeq \rho ss^T/\|x_\star\|$, we have

$$\left\| \frac{\partial\|x_\star\|^2}{\partial b} \right\| = 2\|(\nabla^2 f(x_\star))^\dagger x_\star\| \leq 2\|(\rho x_\star x_\star^T/\|x_\star\|)^\dagger x_\star\| = \frac{2}{\rho}.$$

We thus conclude that $b \mapsto \|x_\star\|^2$ is a $2/\rho$ -Lipschitz continuous function of b , and therefore $|\|x_\star\|^2 - \|\tilde{x}_\star\|^2| \leq (2/\rho)\|b - \tilde{b}\| = 2\sigma/\rho$.

Appendix C. Proof of Proposition 5.1. We begin with a lemma implicitly assuming the conditions of Proposition 5.1.

LEMMA C.1. *For all t we have $\|x_t\| \leq 2R$, with R given by (7a).*

Proof. Note that R minimizes the polynomial $-\|b\|r - \beta r^2/2 + \rho r^3/3$ as it solves $-\|b\| - \beta R + \rho R^2 = 0$. This implies that for every $\|x\| > 2R$ we have

$$f(x) \geq -\|b\|\|x\| - \frac{\beta}{2}\|x\|^2 + \frac{\rho}{3}\|x\|^3 > 2R \left(-\|b\| - \beta R + \frac{4\rho}{3}R^2 \right) = \frac{2\rho}{3}R^3 \geq 0,$$

where the first inequality follows because $b^T x \geq -\|b\|\|x\|$ and $\beta \geq \|A\|_2$, the second follows because $-\|b\|\|x\| - \beta\|x\|^2/2 + \rho\|x\|^3/3$ is increasing in $\|x\|$ for $\|x\| \geq R$, and in the last inequality we substituted $\|b\| = \rho R^2 - \beta R$. By Assumption B, $f(x_0) \leq 0$, and the definition (26) of the step size η_t guarantees that $f(x_t)$ is nonincreasing. Thus, $f(x_t) \leq 0$ for all t , so $\|x_t\| \leq 2R$. \square

As in our proof of Lemma 2.4, we focus on the on the first coordinate of the iteration (25) (i.e., $x_{t+1} = x_t - \eta_t \nabla f(x_t)$) in the eigenbasis of A , writing

$$x_{t+1}^{(1)} = (1 - \eta_t [-\gamma + \rho\|x_t\|]) x_t^{(1)} - \eta_t b^{(1)}.$$

By the constraints in the definition (26) of the step size η_t , we have

$$1 - \eta_t(-\gamma + \rho\|x_t\|) \geq 1 - \eta_t \left[\frac{\nabla f(x_t)^T A \nabla f(x_t)}{\|\nabla f(x_t)\|^2} + \rho\|x_t\| \right]_+ \geq 0.$$

By Assumption B, $b^{(1)}x_0^{(1)} \leq 0$, so $b^{(1)}x_t^{(1)} \leq 0$ for every t . Since $u^T A u / \|u\|^2 \leq \|A\|_2 \leq \beta$ for all u and $\|x_t\| \leq 2R$ for every t , the step size $\eta_{\text{feas}} \triangleq 1/(\beta + 4\rho R)$ is always feasible, and we have $f(x_{t+1}) \leq f(x_t - \eta_{\text{feas}} \nabla f(x_t))$. Moreover, since f is $\beta + 4\rho R$ -smooth on the set $\mathbb{B}_{2R} = \{x \in \mathbb{R}^d : \|x\| \leq 2R\}$, and as $x_t \in \mathbb{B}_{2R}$ for all t by Lemma C.1, we have $f(x_{t+1}) \leq f(x_t) - \frac{\eta_{\text{feas}}}{2} \|\nabla f(x_t)\|^2$, which implies $\nabla f(x_t) \rightarrow 0$. Having established $b^{(1)}x_t^{(1)} \leq 0$ for every t and $\nabla f(x_t) \rightarrow 0$ as $t \rightarrow \infty$, the remainder of the proof is identical to that of Proposition 2.5.

Appendix D. Proofs from section 6.

D.1. Proof of Lemma 6.2. For x_0 defined in line 2 of Algorithm 2, we have $f(x_0) = -(1/2)R_c\|b\| - (\rho/6)R_c^3$, where R_c is the Cauchy radius (8). Therefore, a sufficient condition for $f(x_0) \leq -(1 - \varepsilon')\rho r^3/6$ is $R_c\|b\| \geq \rho r^3/3$. We have

$$R_c \geq \frac{-\beta}{2\rho} + \sqrt{\left(\frac{\beta}{2\rho}\right)^2 + \frac{\|b\|}{\rho}} \geq \min \left\{ \frac{2\|b\|}{3\beta}, \sqrt{\frac{\|b\|}{3\rho}} \right\} \geq \frac{1}{3} \min \left\{ \frac{\|b\|}{\beta}, \sqrt{\frac{\|b\|}{\rho}} \right\},$$

where the second inequality follows from $\sqrt{1 + \alpha} \geq 1 + (\min\{\alpha/3, \sqrt{\alpha/3}\})$ for every $\alpha \geq 0$. Thus, Algorithm 2 returns x_0 whenever $\|b\| \min\{\|b\|/\beta, \sqrt{\|b\|/\rho}\} \geq \rho r^3$, which is equivalent to the second part of the “or” condition in the lemma.

Now, suppose that the algorithm does not return x_0 , i.e., $f(x_0) > -(1 - \varepsilon')\rho r^3/6$. Since $f(x_0) < -(\rho/6)R_c^3$, this implies that $R_c < r$. Since $\rho R_c \geq \rho R - \beta$, with R defined in (7a), we have $\rho r > \rho R - \beta$. Therefore, $\eta \leq 1/(8\beta + 4\rho r) \leq 1/(4\beta + 4\rho R)$,

and so the step size η required in the lemma statement satisfies Assumption A. Since we choose \tilde{x}_0 in accordance with Assumption B, we may invoke Theorem 3.2 with $\varepsilon = \rho\|x_\star\|^3\varepsilon'/12$.

Our setting $\sigma = \frac{\rho^2 r^3 \varepsilon'}{144(\beta+2\rho r)} = \frac{\rho r^3 \varepsilon}{12\|x_\star\|^3(\beta+2\rho r)}$ implies

$$\bar{\sigma} = \frac{12\sigma(\beta+2\rho\|x_\star\|)}{\rho\varepsilon} = \frac{\beta+2\rho\|x_\star\|}{\beta+2\rho r} \cdot \frac{r^3}{\|x_\star\|^3}.$$

Therefore, assuming $r \leq \|x_\star\|$, we have that $(r/\|x_\star\|)^3 \leq \bar{\sigma} \leq 1$. On substituting these upper and lower bounds, Theorem 3.2 shows that, with probability at least $1 - \delta$, $f(\tilde{x}_t) \leq f(x_\star) + (1 + \bar{\sigma})\varepsilon \leq f(x_\star) + 2\varepsilon$ for all

$$t \geq \frac{20\|x_\star\|^2}{\eta\varepsilon} \left(6\log \left(1 + \frac{3\sqrt{d}}{\delta} \cdot \frac{\|x_\star\|^3}{r^3} \right) + 14\log \left(\frac{(\beta+2\rho\|x_\star\|)\|x_\star\|^2}{\varepsilon} \right) \right) \triangleq \tilde{T}_\varepsilon^{\text{sub}}.$$

Using $1/\eta \leq 2(\beta+2\rho\|x_\star\|)$ and plugging in $\varepsilon = \rho\|x_\star\|^3\varepsilon'/12$, we see that

$$\begin{aligned} \tilde{T}_\varepsilon^{\text{sub}} &\leq \frac{240}{\eta\|x_\star\|\varepsilon'} \left(6\log \left(1 + \frac{\sqrt{d}}{\delta} \right) + 6\log \left(\left[\frac{1}{\eta\rho r} \right]^3 \right) + 14\log \left(\frac{6}{\eta\rho\|x_\star\|\varepsilon'} \right) \right) \\ &\leq \frac{240}{\eta r \varepsilon'} \left(6\log \left(1 + \frac{\sqrt{d}}{\delta} \right) + 32\log \left(\frac{6}{\eta\rho r \varepsilon'} \right) \right), \end{aligned}$$

where we have used $r \leq \|x_\star\|$ and $\varepsilon' < 1$. Therefore, T defined in line 4 is larger than $\tilde{T}_\varepsilon^{\text{sub}}$, so with probability at least $1 - \delta$ there exists $t \leq T$ for which $f(\tilde{x}_t) \leq f(x_\star) + \rho\|x_\star\|^3\varepsilon'/6$. Recalling that $f(x_\star) \leq -\rho\|x_\star\|^3/6 \leq -\rho r^3/6$ by the bound (5a) completes the proof.

D.2. Proof of Proposition 6.3. We always call SOLVE-SUBPROBLEM with $\varepsilon' = 1/2$ and $r = \sqrt{\varepsilon/(9\rho)}$. As $\varepsilon \leq \beta^2/\rho$ we have that $\eta = 1/(10\beta) \leq 1/(8\beta + 4\rho r)$. Since $\|\nabla^2 g(x)\|_2 \leq \beta$ by Assumption A, we conclude that Lemma 6.2 applies to each call of SOLVE-SUBPROBLEM. Note that, by construction of Algorithm 1, every call to SOLVE-SUBPROBLEM—except the last one—reduces the value of g by at least $K_{\text{prog}}\varepsilon^{3/2}\rho^{-1/2}$ (line 5). Therefore, by a standard progress argument, the algorithm calls SOLVE-SUBPROBLEM at most

$$(40) \quad K_{\max} = 1 + \left\lceil \frac{\sqrt{\rho}(g(y_0) - g)}{K_{\text{prog}}\varepsilon^{3/2}} \right\rceil \leq O(1) \cdot \frac{\sqrt{\rho}(g(y_0) - g)}{\varepsilon^{3/2}}$$

times, where we have used $\varepsilon \leq \rho^{1/3}(g(y_0) - g)^{2/3}$. Letting \mathcal{E} be the event that the conclusions of Lemma 6.2 hold at each call to SOLVE-SUBPROBLEM, a union bound and our choice $\delta' = \delta/(2k^2)$ at outer iteration k guarantee that

$$\mathbb{P}(\mathcal{E}) \geq 1 - \sum_{k=1}^{\infty} \frac{\delta}{2k^2} \geq 1 - \delta.$$

We perform our subsequent analysis deterministically, conditional on the event \mathcal{E} .

Let f_k be the cubic-regularized quadratic model at iteration k . We call the iteration *successful* (line 5) whenever SOLVE-SUBPROBLEM finds a point Δ_k such that

$$f(\Delta_k) \leq -(1 - \varepsilon')\rho r^3/6 = -\frac{1}{2} \left(\frac{\varepsilon}{9\rho} \right)^{3/2} \frac{\rho}{6} = -K_{\text{prog}}\varepsilon^{3/2}\rho^{-1/2}.$$

The bound (27) shows that $g(y_{k-1} + \Delta_k) \leq g(y_{k-1}) - K_{\text{prog}} \epsilon^{3/2} \rho^{-1/2}$ at each successful iteration, so the last iteration of Algorithm 1 is the only unsuccessful one.

Let K be the index of the final iteration with model f_K , $\Delta_K^* = \text{argmin } f_K$, $A_K = \nabla^2 g(y_{K-1})$, and let $b_K = \nabla g(y_{K-1})$. Since the final iteration is unsuccessful, Lemma 6.2 implies $\|\Delta_K^*\| \leq \sqrt{\epsilon/(9\rho)}$. Let Δ_K be the point produced by the call to SOLVE-FINAL-SUBPROBLEM, and let $y_{\text{out}} = y_K + \Delta_K$ denote the output of Algorithm 1. Note that SOLVE-FINAL-SUBPROBLEM guarantees that $\|\nabla f_K(\Delta_K)\| \leq \epsilon/2$ (we show in the end of this proof that the while loop in line 3 terminates after a finite number of iterations). Moreover, by the same argument as we use in the proof of Lemma 6.2, η satisfies Assumption A. Since Assumption B is also satisfied, we have by Lemma 2.6 that $\|\Delta_K\| \leq \|\Delta_K^*\|$. Therefore, by Assumption A we have that

$$\nabla^2 g(y_{\text{out}}) \succeq A_K - 2\rho \|\Delta_K\| I \succeq -\sqrt{\rho\epsilon} I,$$

where we have used $A_K \succeq -\rho \|\Delta_K^*\| I$ and $\rho \|\Delta_K\| \leq \rho \|\Delta_K^*\| \leq \sqrt{\rho\epsilon}/3$. That is, the output y_{out} satisfies the second condition in (28).

It remains to show that $\nabla g(y_{\text{out}})$ is small. Using $\nabla f_K(\Delta_K) = b_K + A\Delta_K + \rho \|\Delta_K\| \Delta_K$ we have

$$(41a) \quad \|b_K + A\Delta_K\| \leq \|\nabla f_K(\Delta_K)\| + \rho \|\Delta_K\|^2.$$

Recalling that $\nabla^2 g$ is 2ρ -Lipschitz continuous (Assumption A) we have [26, Lemma 1]

$$(41b) \quad \|\nabla g(y_{\text{out}}) - (b_K + A\Delta_K)\| \leq \rho \|\Delta_K\|^2.$$

Combining the norm bounds (41a) and (41b) and using $\|\nabla f_K(\Delta_K)\| \leq \epsilon/2$ and $\rho \|\Delta_K\|^2 \leq \rho \|\Delta_K^*\|^2 \leq \epsilon/9$ yields

$$\|\nabla g(y_{\text{out}})\| \leq \|\nabla f_K(\Delta_K)\| + 2\rho \|\Delta_K\|^2 \leq \epsilon,$$

which completes the proof of the ϵ -second-order stationarity (28) of y_{out} .

We now bound the total number of gradient descent iterations Algorithm 1 uses. Noting that $d/\delta' \leq 2K_{\text{max}}^2 d/\delta > 1$ and that $1/(\eta\rho r) > \beta/\sqrt{\rho\epsilon} \geq 1$, we see that a call to SOLVE-SUBPROBLEM performs at most $O(1)\beta\rho^{-1/2}\epsilon^{-1/2}(\log(d/\delta) + \log K_{\text{max}} + \log(\beta/\sqrt{\rho\epsilon}))$ iterations. Substituting the bound (40) on K_{max} , the number of iterations has the further upper bound

$$O(1) \cdot \frac{\beta}{\sqrt{\rho\epsilon}} \log \left(\frac{d}{\delta} \cdot \frac{\beta g(y_0) - \underline{g}}{\epsilon^2} \right).$$

Multiplying this bound by the upper bound on K_{max} shows that the total number of steps in all calls SOLVE-SUBPROBLEM is bounded by (29).

Finally, standard analysis [24, Example 1.2.3] of gradient descent on smooth functions shows that SOLVE-FINAL-SUBPROBLEM, which we call exactly once, terminates after at most $2 \frac{f(x_0) - f(\Delta_K^*)}{\eta(\epsilon/2)^2} \leq 80\beta(g(y_0) - \underline{g})\epsilon^{-2}$ iterations, as f_K is $\beta + 2\rho R$ -smooth and $\eta \leq \frac{1}{\beta + 2\rho R}$.

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 Forty-Ninth Annual ACM Symposium on the Theory of Computing, Association for Computing Machinery, New York, NY, 2017.

- [2] A. BECK AND Y. VAISBOURD, *Globally solving the trust region subproblem using simple first-order methods*, SIAM J. Optim., 28 (2018), pp. 1951–1967.
- [3] T. BIANCONCINI, G. LIUZZI, B. MORINI, AND M. SCIANDRONE, *On the use of iterative methods in cubic regularization for unconstrained optimization*, Comput. Optim. Appl., 60 (2015), pp. 35–57.
- [4] L. BOTTOU, F. CURTIS, AND J. NOCEDAL, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311.
- [5] Y. CARMON, J. C. DUCHI, O. HINDER, AND A. SIDFORD, *Accelerated methods for non-convex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772, <https://doi.org/10.1137/17M1114296>.
- [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. 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. GOULD, AND P. L. TOINT, *Complexity bounds for second-order optimality in unconstrained optimization*, J. Complexity, 28 (2012), pp. 93–108.
- [9] A. R. CONN, N. I. M. GOULD, AND P. L. TOINT, *Trust Region Methods*, MPS-SIAM Ser. Optim., SIAM, Philadelphia, PA, 2000.
- [10] J. C. DUCHI, *Introductory lectures on stochastic convex optimization*, in The Mathematics of Data, IAS/Park City Math. Ser. 25, American Mathematical Society, Providence, RI, 2018, pp. 99–186.
- [11] J. B. ERWAY AND P. E. GILL, *A subspace minimization method for the trust-region step*, SIAM J. Optim., 20 (2010), pp. 1439–1461.
- [12] R. GE, F. HUANG, C. JIN, AND Y. YUAN, *Escaping from saddle points: Online stochastic gradient for tensor decomposition*, in Proceedings of the Twenty Eighth Annual Conference on Computational Learning Theory, Proc. Mach. Learn. Res. 40, 2015, pp. 797–842; available at <http://proceedings.mlr.press/v40/>.
- [13] N. I. M. GOULD, S. LUCIDI, M. ROMA, AND P. L. TOINT, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
- [14] N. I. M. GOULD, D. P. ROBINSON, AND H. S. THORNE, *On solving trust-region and other regularised subproblems in optimization*, Math. Program. Comput., 2 (2010), pp. 21–57.
- [15] A. GRIEWANK, *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.
- [16] E. HAZAN AND T. KOREN, *A linear-time algorithm for trust region problems*, Math. Program., 158 (2016), pp. 363–381.
- [17] N. HO-NGUYEN AND F. KILINÇ-KARZAN, *A Second-Order Cone Based Approach for Solving the Trust-Region Subproblem and Its Variants*, preprint, <https://arxiv.org/abs/1603.03366>, 2016.
- [18] 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.
- [19] Y. LECUN, Y. BENGIO, AND G. HINTON, *Deep learning*, Nature, 521 (2015), pp. 436–444.
- [20] J. D. LEE, M. SIMCHOWITZ, M. I. JORDAN, AND B. RECHT, *Gradient descent converges to minimizers*, in Proceedings of the Twenty Ninth Annual Conference on Computational Learning Theory, Proc. Mach. Learn. Res. 49, 2016, pp. 1246–1257; available at <http://proceedings.mlr.press/v49/>.
- [21] K. Y. LEVY, *The Power of Normalization: Faster Evasion of Saddle Points*, preprint, <https://arxiv.org/abs/1611.04831>, 2016.
- [22] C. MUSCO AND C. MUSCO, *Randomized block Krylov methods for stronger and faster approximate singular value decomposition*, in Adv. Neural Inf. Process. Syst. 28, Curran Associates, Red Hook, NY, 2015, pp. 1396–1404.
- [23] Y. NESTEROV, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$* , Sov. Math. Dokl., 27 (1983), pp. 372–376.
- [24] Y. NESTEROV, *Introductory Lectures on Convex Optimization*, Kluwer Academic, Norwell, MA, 2004.
- [25] Y. NESTEROV, *Gradient Methods for Minimizing Composite Objective Function*, Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.
- [26] Y. NESTEROV AND B. POLYAK, *Cubic regularization of Newton method and its global*

- performance*, Math. Program., 108 (2006), pp. 177–205.
- [27] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, 2nd ed., Springer Ser. Oper. Res. Financ. Eng., Springer, New York, NY, 2006.
 - [28] B. A. PEARLMUTTER, *Fast exact multiplication by the Hessian*, Neural Comput., 6 (1994), pp. 147–160.
 - [29] B. T. POLYAK, *Some methods of speeding up the convergence of iteration methods*, U.S.S.R. Comput. Math. and Math. Phys., 4(5) (1964), pp. 1–17.
 - [30] N. N. SCHRAUDOLPH, *Fast curvature matrix-vector products for second-order gradient descent*, Neural Comput., 14 (2002), pp. 1723–1738.
 - [31] M. SIMCHOWITZ, A. E. ALOUI, AND B. RECHT, *Tight query complexity lower bounds for PCA via finite sample deformed Wigner law*, in Proceedings of the Fiftieth Annual ACM SIGACT Symposium on the Theory of Computing, Association for Computing Machinery, New York, NY, 2018, pp. 1249–1259, <https://doi.org/10.1145/3188745.3188796>.
 - [32] P. D. TAO AND L. T. H. AN, *A D.C. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.
 - [33] L. N. TREFETHEN AND D. BAU, III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
 - [34] M. WEISER, P. DEUFLHARD, AND B. ERDMANN, *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*, Optim. Methods Softw., 22 (2007), pp. 413–431.
 - [35] L.-H. ZHANG, C. SHEN, AND R.-C. LI, *On the generalized Lanczos trust-region method*, SIAM J. Optim., 27 (2017), pp. 2110–2142.