OPTIMAL PARAMETER-FREE SECOND-ORDER ACCELERATION WITH INEXACT HESSIANS UNDER GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study second-order convex optimization under generalized Hessian smoothness, where the local Hessian Lipschitz constant may grow linearly with the gradient norm. For inexact Hessians, we develop adaptive algorithms, including an accelerated Monteiro-Svaiter-type method whose complexity matches the lower bound for second-order methods with δ-inexact Hessians in the classical Lipschitz-Hessian setting. Our methods are parameter-free in the sense that they require no problem-specific parameters as input, including the Hessian inexactness level δ and smoothness constants. Instead, their backtracking adapts a single regularization parameter that jointly captures Hessian inexactness and local Hessian smoothness. The resulting rates split into two regimes. Far from the solution, the gradient-dependent smoothness term yields geometric convergence. Near the solution, the rate is governed by the ordinary Hessian smoothness and the inexactness level δ, recovering the classical Lipschitz-Hessian guarantees.

1 INTRODUCTION

We study second-order methods for the problem $\min_{x\in\mathbb{R}^d} f(x)$, where $f \in \mathcal{C}^2$ is a convex function. We assume that a solution $x^* \in \mathbb{R}^d$ exists and we denote $f^* := f(x^*)$. Given an initial point x_0, we define $R := \|x_0 - x^*\|$ as a distance to the solution. The classical theory assumes exact Hessians that are Lipschitz continuous (Nesterov & Polyak, 2006; Nesterov, 2018):

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \le L\|y - x\|, \qquad x, y \in \mathbb{R}^d. \tag{H}$$

In this setting, optimal methods find an ε-solution in $\mathcal{O}\big((LR^3/\varepsilon)^{2/7}\big)$ iterations (Kovalev & Gasnikov, 2022; Carmon et al., 2022), matching the lower bound (Arjevani et al., 2019). We consider two extensions of this theory.

Generalized smoothness. In first-order optimization, (L_0, L_1)-smoothness lets the local Lipschitz-gradient constant grow with the gradient norm (Zhang et al., 2020; Crawshaw et al., 2022). We study its second-order analogue (Xie et al., 2024; Jerad, 2026): for some $M_0 \ge 0$, $M_1 \ge 0$, and $\rho > 0$,

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \le (M_0 + M_1\|\nabla f(x)\|)\,\|y - x\|, \qquad x, y \in \mathbb{R}^d : \ \|y - x\| \le \rho. \tag{LH}$$

It reduces to (H) for $M_1 = 0$ and $\rho = +\infty$, and it covers problems such as Poisson regression and ℓ_p regression with $p \ge 4$, for which (H) fails (see examples in Section 2.1).

Inexact Hessians. Exact Hessians are often too expensive. In practice, methods use approximations obtained by subsampling (Kohler & Lucchi, 2017; Xu et al., 2020), quasi-Newton updates (Kamzolov et al., 2025b), or lazy updates (Doikov et al., 2023). We assume that the approximation $H(x)$ is symmetric and satisfies

$$\|H(x) - \nabla^2 f(x)\| \le \delta. \tag{IH}$$

Parameter-free methods.[1] Both extensions make assumptions more realistic, but they add constants that a method has to know: (LH) involves M_0, M_1, and ρ, and (IH) adds δ. These constants are rarely known. Moreover, under (LH) they are not unique. Poisson regression satisfies (LH) for every $\rho > 0$, with M_0 and M_1 growing in ρ (Example 2.2). A method that takes them as input must fix ρ in advance, whereas the guarantees of a parameter-free method hold for the best ρ.

To the best of our knowledge, even under (H), no parameter-free method attains the optimal complexity with inexact Hessians. The method of Carmon et al. (2022) is optimal and adaptive, but its analysis requires exact Hessians. The optimal method of Chen et al. (2026) requires L, δ, and R

[1] By parameter-free, we mean requiring no problem constants as input and converging for any initialization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of our methods with the fastest known second-order convex methods and the lower bound with inexact Hessians

Method	Assump.	Param. Free	Hessian complexity
Lower bound (Agafonov et al., 2024b)	(H) (IH)	–	$\Omega\big((\frac{LR^3}{\varepsilon})^{2/7} + \sqrt{\frac{\delta R^2}{\varepsilon}}\big)$
Optimal MS Acceleration (Carmon et al., 2022)	(H)	✓	$\mathcal{O}\big((\frac{LR^3}{\varepsilon})^{2/7}\big)$
Optimal inexact Newton (Chen et al., 2026)	(H) (IH)	✗	$\mathcal{O}\big((\frac{LR^3}{\varepsilon})^{2/7} + \sqrt{\frac{\delta R^2}{\varepsilon}}\big)$
Gradient-Regularized Newton (Semenov et al., 2026)	(GH) (IH)[a]	✓	$\widetilde{\mathcal{O}}\big(\sqrt{\frac{M_0 D^3}{\varepsilon}} + \sqrt{M_1} D + \frac{\delta D^2}{\varepsilon}\big)^b$
AR2-LS (Jerad, 2026)	(LH)	✓	$\widetilde{\mathcal{O}}\big(\varepsilon^{-1/2}\big)$
Adaptive regularized Newton (This work, Thm. 3.1)	(LH) (IH)	✓	$\mathcal{O}\big(\sqrt{\frac{M_0 D^3}{\varepsilon}} + \frac{\delta D^2}{\varepsilon}\big) + \widetilde{\mathcal{O}}\big(1 + \sqrt{M_1} D + \frac{D}{\rho}\big)$
Inexact adaptive A-NPE (This work, Thm. E.1)	(LH) (IH)	✓	$\mathcal{O}\big((\frac{M_0 R^3}{\varepsilon})^{1/3} + \sqrt{\frac{\delta R^2}{\varepsilon}}\big) + \widetilde{\mathcal{O}}\big(1 + \sqrt{M_1} R + \frac{R}{\rho}\big)$
Inexact adaptive damped A-NPE (This work, Thm. 4.1)	(LH) (IH)	✓	$\mathcal{O}\big((\frac{M_0 R^3}{\varepsilon})^{2/7} + \sqrt{\frac{\delta R^2}{\varepsilon}}\big)$ $+\widetilde{\mathcal{O}}\big(1 + M_1^{1/3} R^{2/3} + (\frac{R}{\rho})^{2/3}\big)$

$D = \operatorname{diam}\{f \le f(x_0)\} < \infty$, $G = \sup_{f \le f(x_0)} \|\nabla f\|$. $\widetilde{\mathcal{O}}$ hides logarithmic factors. Parameter-free methods require no problem parameters. [a] $H_k \succeq 0$ is also assumed. [b] Combines Corollary 2 and Appendix F.6 of Semenov et al. (2026).

as input. Under generalized smoothness, existing convex guarantees are non-accelerated (Semenov et al., 2026; Jerad, 2026). In view of these results, we ask two questions:

Can a parameter-free method attain the optimal complexity with inexact Hessians?
Can such a method be extended to generalized Hessian smoothness?

Our main result. We answer both questions affirmatively with a single method (Algorithm 3). Using the Monteiro–Svaiter condition (Monteiro & Svaiter, 2013) as an acceptance test, we do not estimate the Hessian inexactness and the local Lipschitz constant of the Hessian separately, but rather adapt them as a single regularization parameter. Under (H) and (IH), the proposed method needs $\mathcal{O}\left((LR^3/\varepsilon)^{2/7} + \sqrt{\delta R^2/\varepsilon}\right)$ Hessian evaluations, up to an additive logarithmic term independent of ε. This matches the lower bound (Agafonov et al., 2024b). Under (LH), the same method converges linearly far from the solution and then at the same optimal rate with $L = M_0$. Table 1 compares our main methods with closest competitors and the lower bound.

1.1 Related Work

Beyond classical smoothness. In first-order optimization, empirical observations on neural networks have motivated (L_0, L_1)-smoothness assumption (Zhang et al., 2020; Crawshaw et al., 2022), which is studied for both convex (Koloskova et al., 2023; Li et al., 2023; Tyurin, 2025; Gorbunov et al., 2025) and nonconvex (Koloskova et al., 2023; Chen et al., 2023; Vankov et al., 2025) objectives. In particular, convex accelerated under (L_0, L_1)-smoothness has recently received considerable attention (Gorbunov et al., 2025; Vankov et al., 2025; Tyurin, 2026), including parameter and line-search free methods (Borodich & Kovalev, 2026). Second-order methods under generalized smoothness have been developed in nonconvex (Xie et al., 2024; Fang et al., 2026) and convex (Semenov et al., 2026; Jerad, 2026) settings, some under the global variant of (LH)

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \le \big(M_0 + M_1 \|\nabla f(x)\|\big)\|y - x\|, \qquad x, y \in \mathbb{R}^d. \tag{GH}$$

Inexact Hessians. Under (H), many methods with inexact Hessians augment cubic regularization with a quadratic term that compensates for the Hessian error (Ghadimi et al., 2017; Agafonov et al., 2024a;b). These methods take both L and δ as input. Kamzolov et al. (2025b) adapt the quadratic coefficient but still require a cubic coefficient $\mathcal{M} \ge 2L$. Antonakopoulos et al. (2022) propose Extra-Newton, a parameter-free but suboptimal method. Semenov et al. (2026) propose an adaptive but non-accelerated method for generalized Hessian smoothness with inexact Hessians. Recently, Chen et al. (2026) proposed an optimal method matching the lower bound (Agafonov et al., 2024b), but their method is non-adaptive.

Adaptive second-order methods. With exact Hessians and under (H), adaptive cubic regularization estimates the cubic coefficient by backtracking (Nesterov & Polyak, 2006; Cartis et al., 2011;

Grapiglia & Nesterov, 2020). Gradient-regularized Newton methods replace the cubic subproblem with a linear system (Mishchenko, 2023; Doikov & Nesterov, 2023), and their adaptive variants tune the regularization coefficient in the same way (Doikov et al., 2024). Carmon et al. (2022) obtain an optimal adaptive method by searching over the regularization of a Newton step using Monteiro–Svaiter acceleration (Monteiro & Svaiter, 2013).

1.2 Contributions

- **Far/near rate separation under** (LH)**.** We analyze an adaptive regularized Newton method and three accelerated variants. Far from the solution, the methods converge linearly at a rate that depends on M_1 and ρ. Near the solution, they recover the standard rates under (H) (see Table 1).
- **Inexact Hessians and adaptivity.** The three methods listed in Table 1 allow inexact Hessians and remain parameter-free. A single regularization parameter in the Newton step accounts for both the Hessian error and the local smoothness, and methods adapt it by backtracking on the Monteiro-Svaiter condition.
- **Optimal adaptive acceleration.** Algorithm 3 combines this backtracking with the damped acceleration scheme of Carmon et al. (2022). Under (LH) and (IH), it converges linearly during the first $\widetilde{\mathcal{O}}\big(1 + M_1^{1/3}R^{2/3} + (R/\rho)^{2/3}\big)$ iterations and then needs $\mathcal{O}\big((M_0R^3/\varepsilon)^{2/7} + \sqrt{\delta R^2/\varepsilon}\big)$ more iterations. Under (H), the first phase reduces to an additive logarithmic term, and the bound matches the lower bound (Agafonov et al., 2024b). The number of gradient evaluations has the same order up to additive terms at most logarithmic in $1/\varepsilon$.
- **Revisited analysis of optimal acceleration.** Our convergence proof is new, and we discuss its differences from the analyses of Carmon et al. (2022) and Chen et al. (2026) in Section 4.3.
- **Recovering local quadratic rate.** For strongly convex functions, our adaptive regularized Newton converges globally linearly and does not accumulate Hessian errors. If the errors vanish (e.g., with lazy Hessian updates), it converges superlinearly. Under a stronger, relative accuracy condition on the Hessian approximation, it also converges locally quadratically (Section 3.2).

2 Preliminaries

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices. For both conditions (GH) and (LH), we define $L(x) := M_0 + M_1\|\nabla f(x)\|$. When using $H(x)$ satisfying (IH), we assume w.l.o.g. that it is symmetric, since replacing it by symmetric matrix $(H(x) + H^T(x))/2$ does not increase δ. We set $\log_+ t := \max\{0, \log t\}$ and $\log_{b,+} t := \max\{0, \log_b t\}$ for $t > 0$ and $b > 1$. Under (GH) we take $\rho = +\infty$, and use $a/(+\infty) = 0$ for finite $a \geq 0$ and $a/0 = +\infty$ for $a > 0$. In particular, $1/\sqrt{M_1} = +\infty$ if $M_1 = 0$.

2.1 Generalized Hessian Smoothness

Relations between assumptions. Functions satisfying (H), such as logistic regression, also satisfy (GH) and (LH). Examples below show that the inclusions (H) $\subset$ (GH) $\subset$ (LH) are strict. In particular, setting $x = x^*$ in (GH) gives $\|\nabla^2 f(y)\| \leq \|\nabla^2 f(x^*)\| + M_0\|y - x^*\|$. Thus, the Hessian norm can grow at most linearly with the distance from x^*, as under (H). However, (GH) is more general than (H), since the coefficient controlling the Hessian variation at x may grow with $\|\nabla f(x)\|$. The local condition (LH) does not imply this global linear bound and therefore covers a broader class of functions.

Example 2.1. The function $f(x) = x^2 + \int_0^x (x-t)\sin(t^2)\mathrm{d}t$ is 1-strongly convex and satisfies (GH) with $M_0 = M_1 = 2$, but it does not satisfy (H).

Example 2.2 (Poisson regression)**.** Let $a_i \in \mathbb{R}^d$ with $[a_i]_1 = 1$ and $b_i \in \mathbb{N}$, $i = 1, \ldots, n$. The Poisson regression loss $f(x) = \sum_{i=1}^n \left(e^{\langle a_i, x\rangle} - b_i \langle a_i, x\rangle\right)$ is convex and satisfies (LH) for every $\rho > 0$ with $M_0 = \|b\|_1 c^3 e^{c\rho}$ and $M_1 = c^3 e^{c\rho}$, where $c = \max_i \|a_i\|$, but it does not satisfy (GH).

Example 2.3 (ℓ_p regression)**.** Let $p > 3$, $A \in \mathbb{R}^{n\times d}$, $A \neq 0$, and $b \in \mathbb{R}^n$. The ℓ_p regression loss $f(x) = \frac{1}{p}\|Ax - b\|_p^p$ is convex and satisfies (LH) for every $\rho > 0$ with

$$M_0 = (p-1)(p-2)\|A\|^3 \left(1 + 2\|b\|_p + \|A\|\rho\right)^{p-3}, \qquad M_1 = 3\sqrt{n}M_0/\sigma_+(A),$$

where $\sigma_+(A)$ is the smallest positive singular value of A, but it does not satisfy (GH).

For $f \in C^3(\mathbb{R}^d)$, condition (LH) implies the pointwise bound (Xie et al., 2024)

$$\|\nabla^3 f(x)\|_{\mathrm{op}} := \sup_{\|u\|=1} \|\nabla^3 f(x)[u]\| \leq M_0 + M_1\|\nabla f(x)\|, \qquad x \in \mathbb{R}^d. \tag{TD}$$

Proposition 2.4. *Suppose that, for some $L_0 > 0$ and $L_1 \geq 0$, $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$. If* (TD) *holds with constants $\widehat{M}_0$ and $\widehat{M}_1$, then, for every $\rho > 0$,* (LH) *holds with*

$$M_0(\rho) = \widehat{M}_0 + \widehat{M}_1 L_0 \rho e^{L_1\rho}, \qquad M_1(\rho) = \widehat{M}_1 e^{L_1\rho}.$$

2.2 Globally Convergent Second-order Methods under Inexact Hessians

With an inexact Hessian, cubic regularization commonly uses the model

$$\omega_x^{\mathcal{M},\widehat{\delta}}(s) := f(x) + \langle \nabla f(x), s\rangle + \tfrac{1}{2}\langle H(x)s, s\rangle + \tfrac{\widehat{\delta}}{2}\|s\|^2 + \tfrac{\mathcal{M}}{6}\|s\|^3, \tag{1}$$

with $s = y - x$. The quadratic term compensates for Hessian error and the cubic term controls the Taylor remainder. Under (H), $\widehat{\delta} \geq \delta$ and $\mathcal{M} \geq L$ ensure model convexity and majorization (Agafonov et al., 2024a, Theorem 1). In Nesterov-type acceleration these coefficients also enter the estimating sequence (Nesterov, 2008; Ghadimi et al., 2017).

For a parameter-free method, a useful observation is that the first-order optimality condition of the cubic model depends on these two regularizers only through their sum:

$$(H(x) + \tau I)s = -\nabla f(x), \qquad \tau = \widehat{\delta} + \tfrac{\mathcal{M}}{2}\|s\|.$$

This suggests treating τ as a singe regularization parameter, rather than adapting $\widehat{\delta}$ and M separately. The difficulty here is that τ depends on unknown step length $\|s\|$. It is a circular dependence, which arises in the accelerated Newton proximal extragradient (A-NPE) method (Monteiro & Svaiter, 2013) and its optimal variants (Kovalev & Gasnikov, 2022; Carmon et al., 2022; Chen et al., 2026). In particular, it can be resolved by searching over the regularization parameter while solving only quadratic subproblems. This perspective motivates our approach: instead of separately adapting two regularization parameters, we search directly for a regularized Newton step that satisfies a verifiable Monteiro-Svaiter condition (Monteiro & Svaiter, 2013).

Definition 2.5. A Monteiro-Svaiter (MS) oracle at x returns $y = x + s$ and $\tau > 0$ such that

$$\|\nabla f(x+s) + \tau s\| \leq \tau\|s\|/2.$$

Under (H), inexact cubic step $y = x + \operatorname{argmin}_s \omega_x^{\mathcal{M},\widehat{\delta}}(s)$ with $\widehat{\delta} \geq 2\delta$ and $\mathcal{M} \geq 2L$ is an instance of MS oracle with $\tau = \widehat{\delta} + \mathcal{M}\|s\|/2$ (Chen et al., 2026, Lemma 3.1). Under exact Hessian the same Lipschitz assumption, the gradient-regularized Newton step (Mishchenko, 2023; Doikov & Nesterov, 2023) is also an MS oracle for $\tau = \sqrt{M\|\nabla f(x)\|} > 0$ and $M \geq L$.

3 Basic Method

Algorithm 1 MSBacktrack(x, τ)

1: **Input:** $x \in \mathbb{R}^d$ and $\tau > 0$.
2: **loop**
3: **if** PD(x, τ) **then**
4: Solve $(H(x) + \tau I)s = -\nabla f(x)$.
5: **if** MS(x, τ, s) **then**
6: **return** $(x + s, \tau)$.
7: Set $\tau := 2\tau$.

Algorithm 2 Adaptive regularized Newton

1: **Input:** $x_0 \in \mathbb{R}^d$ and $\eta_0 > 0$.
2: **for** $k \geq 0$ **do**
3: $(x_{k+1}, \tau_k) :=$ MSBacktrack(x_k, η_k).
4: Set $\eta_{k+1} := \frac{\tau_k}{2}\min\left\{1, \frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|}\right\}$.

We begin by proposing a basic non-accelerated adaptive method under MS oracle. First, we define two tests for our backtracking procedure

$$\mathsf{PD}(x,\tau) : H(x) + \tau I \succ 0, \quad \mathsf{MS}(x,\tau,s) : \|\nabla f(x+s) + \tau s\| \leq \tfrac{\tau}{2}\|s\|. \tag{2}$$

The tests do not require δ, $L(x)$, or ρ. The PD test is checked first, so the linear system is solved only when $H(x)+\tau I$ is positive definite. Algorithm 1 defines MSBacktrack, which ensures that tests (2) hold. The resulting Adaptive regularized Newton methods is listed as Algorithm 2.

At each outer iteration, the method evaluates one $H(x)$ and keeps it fixed throughout backtracking. For example, an SVD-based implementation can reuse the same factorization across all trial values of τ. The gradient $\nabla f(x_k)$ is reused throughout backtracking, and $\nabla f(x_{k+1})$ is reused in the update of η_{k+1} and at the next iteration. The halving in line 4 prevents overregularization, which is important in our convergence analysis (see proof sketch below). Consequently, τ is allowed to both increase and decrease across outer iterations. The gradient-dependent scaling factor is not needed for convex convergence, but allows us to recover local quadratic convergence under exact Hessians.

3.1 Convex analysis

Let $D = \max_{x \in \mathcal{L}(x_0)} \|x - x_0\|$ be the radius of the level set $\mathcal{L}(x_0) = \left\{x \in \mathbb{R}^d : f(x) \leq f(x_0)\right\}$. Since Algorithm 2 is monotone (Lemma C.1), all iterates remain in $\mathcal{L}(x_0)$. Define

$$B := \max\left\{1, 2D/\rho, 6\sqrt{M_1}D\right\}, \qquad \varepsilon_{\mathrm{tr}} := \max\left\{M_0 D^3/B^2, \delta D^2/B\right\}, \tag{3}$$

where $\varepsilon_{\mathrm{tr}}$ marks the switch from the linear rate to the rate of inexact cubic Newton.

Theorem 3.1. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Let $\ell_0 := \max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\text{tr}}}, \frac{\eta_0 D^2}{\sqrt{6}B\varepsilon_{\text{tr}}}\right\}$. For every $\varepsilon > 0$, Algorithm 2 generates an iterate x_N satisfying $f(x_N) - f^* \le \varepsilon$ within the following number of iterations*

$$N = \mathcal{O}\left(1 + \log \ell_0 + B \log_+ \frac{f(x_0)-f^*}{\max\{\varepsilon, \varepsilon_{\text{tr}}\}} + \sqrt{\frac{M_0 D^3}{\varepsilon}} + \frac{\delta D^2}{\varepsilon}\right). \tag{4}$$

Theorem 3.1 bounds the number of outer iterations N, which equals the number of inexact Hessian evaluations. In Appendix C.5, we show that the total number of gradient evaluations is $\mathcal{O}(N + \log_+ \frac{B\|\nabla f(x_0)\|}{D\eta_0})$, matching N up to a constant factor and an additive logarithmic term.

The complexity (4) separates the dependence on M_0, M_1 and δ, and reveals two accuracy regimes. If the target accuracy $\varepsilon \ge \varepsilon_{\text{tr}}$, the method remains in the far regime and converges linearly, requiring $\mathcal{O}(B + \log \ell_0 + B \log_+ \frac{f(x_0)-f^*}{\varepsilon})$ iterations. If $\varepsilon < \varepsilon_{\text{tr}}$, the method first reaches ε_{tr} at a linear rate, after which the complexity is determined by M_0 and δ, matching the rate of inexact cubic Newton under (H). The additional term $\log \ell_0$ is independent of ε and arises from bounding the number of overregularized iterations (see the proof sketch). Let us introduce one of the main lemmas.

Lemma 3.2. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Let $(y, \tau) :=$ MSBACKTRACK(x, η). If at least one trial is rejected (equivalently $\tau > \eta$), then*

$$\|s\| > \min\left\{\frac{\rho}{2}, \frac{1}{6\sqrt{M_1}}\right\} \quad \text{or} \quad \tau \le 8\delta + 24 M_0 \|s\|$$

This shows that for iterations with rejected trials we have two regimes: either regularization matches the scale of inexact cubic Newton step (set $M_0 = L$ under (H) and see example after Definition 2.5) or the step is determined by generalized smoothness constants M_1, ρ and is sufficiently large.

Proof sketch of Theorem 3.1. Denote $\Delta_k = f(x_k) - f(x^*)$, $g_k = \|\nabla f(x_k)\|$, and $s_k = x_{k+1} - x_k$. By MS test we show $\tau_k \|s_k\|/2 \le g_{k+1} \le 3\tau_k \|s_k\|/2$ and using convexity

$$f(x_k) - f(x_{k+1}) \ge 2g_{k+1}^2/(3\tau_k), \qquad g_{k+1} \le 2g_k/\sqrt{3}.$$

Thus, the method is monotone and progress of one step depends on how large the accepted regularization τ_k can be relative to g_{k+1}. We call iteration k controlled if $\tau_k \le \Theta_{\text{c}}(g_{k+1})$, where $\Theta_{\text{c}}(g) := \max\left\{8\delta + \sqrt{48 M_0 g}, 4g/\rho, 12\sqrt{M_1 g}\right\}$. The proof is based on separate counting of controlled and uncontrolled iterations.

First, we show, that if at least one trial was rejected the iteration is controlled. By Lemma 3.2 and MS relation we have $\tau_k \le \Theta_{\text{c}}(g_{k+1})$. Let $U_k := \max\{\tau_{k-1}, \Theta_{\text{c}}(\|\nabla f(x_k)\|)\}$. Then by the definition of U_k we get $U_{k+1} \le (2/\sqrt{3})U_k$ on every controlled iteration.

An uncontrolled iteration has no rejected trials and therefore accepts $\tau_k = \eta_k$. By the update rule for η_k, we obtain $U_{k+1} = \tau_k \le \tau_{k-1}/2 \le U_k/2$. Therefore $\Theta_{\text{c}}(g_m) \le U_0(2/\sqrt{3})^{m_{\text{c}}} 2^{-(m-m_{\text{c}})}$ after m iteration of which m_{c} are controlled. Taking the logarithm yields

$$m \le 2m_{\text{c}} + \log_2(U_0/\Theta_{\text{c}}(g_m)).$$

Thus, to show the convergence, it is enough to count the number of controlled steps. For controlled step, using the boundness of τ_k, convexity and monotonicity we show recurrence

$$\Delta_k - \Delta_{k+1} \gtrsim \frac{\Delta_{k+1}}{B + \sqrt{M_0 D^3/\Delta_{k+1}} + \delta D^2/\Delta_{k+1}}.$$

When B dominates in denominator ($\Delta_{k+1} \ge \varepsilon_{\text{tr}}$) we get linear convergence, otherwise the rate is determined by M_0 and δ terms. To accumulate this progress, we introduce a scalar potential that increases by at least one on every controlled iteration and does not decrease on the others. This gives us a bound

$$m_{\text{c}} \lesssim B \log_+ \Delta_0/\max\{\varepsilon, \varepsilon_{\text{tr}}\} + \sqrt{M_0 D^3/\varepsilon} + \delta D^2/\varepsilon$$

as long as $\Delta_m > \varepsilon$. Bounding the logarithmic term in upper bound on m gives us the rate (4). The complete proof is given in Appendix C.

3.2 Strongly convex analysis

In this section, we denote $\delta_k := \|H(x_k) - \nabla^2 f(x_k)\| \le \delta$ and $r_0 := \min\left\{\rho/2, 1/(6\sqrt{M_1})\right\}$.

Theorem 3.3. *Suppose that f is μ-strongly convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Let*

$$\overline{\tau} := \max\left\{\eta_0, 8\delta + 48 M_0 \sqrt{2(f(x_0) - f^*)/\mu}, 2(f(x_0) - f^*)/r_0^2\right\}.$$

Every completed iteration k of Algorithm 2 satisfies $\tau_k \le \overline{\tau}$ and

$$\tau_k \le 8 \max_{0 \le j \le k} 2^{j-k}\delta_j + \overline{\tau}\left(1 - \frac{2\mu}{3\overline{\tau} + 2\mu}\right)^k, \quad f(x_{k+1}) - f^* \le \left(1 - \frac{2\mu}{3\tau_k + 2\mu}\right)^2 (f(x_k) - f^*).$$

Theorem 3.3 guarantees global linear convergence for any sequence $\delta_k \le \delta$ and shows that Hessian errors do not accumulate. If $\delta_k \to 0$, then $\tau_k \to 0$, so the contraction factor in convergence rate goes to zero and the method converges superlinearly. In particular, for exact Hessians, we obtain global superlinear convergence as established for cubic Newton in Kamzolov et al. (2025a). This theorem also covers lazy Hessian updates (Doikov et al., 2023), where an exact Hessian is reused for a fixed number of iterations. Indeed, for any $m \ge 1$, the choice $H(x_k) = \nabla^2 f(x_{m\lfloor k/m \rfloor})$ satisfies $\delta_k \to 0$.

The rate in Theorem 3.3 bounds the number of outer iterations and, equivalently, Hessian computations. To bound the first-order oracle calls, let N be the first index satisfying $f(x_N) - f^* \le \varepsilon$. By Corollary C.12 it can be bounded as $\mathcal{O}(N + \log_{2,+}\left(\frac{\|\nabla f(x_0)\|}{\eta_0} \max\left\{\frac{M_0}{\mu} + \frac{\delta}{\sqrt{\mu\varepsilon}}, \frac{1}{r_0}\right\}\right))$. For inexact Hessians, this bound guarantees only linear convergence in gradient evaluations, even when $\delta_k \to 0$ gives superlinear convergence in Hessian evaluations. For $\delta = 0$, the additional logarithmic term is independent of ε, and the gradient count is $\mathcal{O}(N+1)$. Thus, for exact Hessians, the method converges superlinearly in terms of both gradient and Hessian evaluations.

Our next goal is to recover the local quadratic convergence typical for second-order methods with exact Hessians. We show that Algorithm 2 preserves local quadratic and global superlinear convergence even with inexact Hessians. Both rates hold in terms of gradient and Hessian evaluations under a stronger accuracy assumption than (IH) (Bellavia et al., 2020; 2022; Chen et al., 2022; Semenov et al., 2026, Examples 7, 8):

$$\delta_k \le \nu \|\nabla f(x_k)\|, \qquad k \ge 0. \tag{5}$$

Theorem 3.4. *Suppose that f is μ-strongly convex, either* (GH) *or* (LH) *holds, and* (5) *holds for some $\nu > 0$. Algorithm 2 converges globally superlinearly. Every completed iteration k satisfies*

$$\tau_k \le \max\left\{\frac{2^{-k}\eta_0}{\|\nabla f(x_0)\|}, \frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\} \|\nabla f(x_k)\|, \; f(x_{k+1}) - f^* \le \left(1 - \frac{2\mu}{3\tau_k + 2\mu}\right)^2 (f(x_k) - f^*).$$

Let $C := \frac{9\left(\|\nabla^2 f(x^*)\| + \mu\right)}{2\mu^2}\left(8\nu + \frac{24M_0}{\mu}\right)^2$. Let for some $K_0 \ge 1$, $f(x_{K_0}) - f^* \le \min\left\{\frac{\mu r_0^2}{8}, \frac{1}{2C}\right\}$. Then, after finitely many additional iterations independent of ε, the method converges quadratically:

$$f(x_{k+1}) - f^* \le C\,(f(x_k) - f^*)^2, \qquad k \ge K \ge K_0.$$

By Corollary C.14, $N \ge 1$ completed iterations require at most N Hessian evaluations and $\mathcal{O}(N + \log_{2,+}\left(\frac{\|\nabla f(x_0)\|}{\eta_0} \max\left\{\frac{M_0}{\mu} + \nu, \frac{1}{r_0}\right\}\right))$ gradient evaluations. The logarithmic term is independent of ε, so the global superlinear iteration complexity also holds for gradient evaluations. Since our method converges globally, the index K_0 is reachable. After iteration K, which is defined in Theorem C.15 and independent of ε, the method requires $\mathcal{O}(\log\log(1/\varepsilon))$ additional Hessian and gradient evaluations suffice to reach $f(x) - f^* \le \varepsilon$.

4 Acceleration under Generalized Smoothness

4.1 Suboptimal Acceleration

Before moving to the optimal acceleration scheme, we briefly discuss generalization of accelerated methods with the suboptimal rate $\mathcal{O}(1/k^3)$ under (H). Inspired by Tyurin (2026), in Appendix D.2 we propose an adaptive Nesterov-type accelerated method with exact Hessians. The main challenge is that under (LH), the regularization parameter $\mathcal{M}_k$ must dominate $L(w_{k+1})$ at the output w_{k+1} of the cubic step (1), even though this point is unknown when $\mathcal{M}_k$ is chosen. We refer to the proof sketch in Appendix D for the key idea behind resolving this difficulty. Furthermore, this acceleration scheme uses the cubic model, which has two parameters: $\mathcal{M}$ and $\widehat{\delta}$. To avoid the difficulty of adapting two parameters simultaneously, in Appendix E we also propose an adaptive A-NPE method with inexact Hessians under (LH), which attains the rate $\mathcal{O}(1/k^3 + \delta/k^2)$ under (H).

4.2 Adaptive Damped A-NPE Method

In this section, we present the adaptive damped A-NPE method, listed as Algorithm 3. The method is parameter-free and attains the optimal convergence rate for convex optimization with inexact Hessians in the classical Lipschitz-Hessian setting. It combines backtracking (Algorithm 1) with acceleration scheme of Carmon et al. (2022). The parameter η_k is the guess regularization parameter for MSBacktrack procedure, which returns $\tau_k \ge \eta_k$. If $\tau_k = \eta_k$ the method performs full step with $A_{k+1} = A_k + a'_k$, otherwise the update of A_{k+1} is damped $A_{k+1} = A_k + (\eta_k/\tau_k)a'_k$. The next guess is set to $\eta_k/2$ when the first trial is accepted and to $2\eta_k$ otherwise.

As in Algorithm 2, each outer iteration uses at most one Hessian, which is reused throughout backtracking. We again maintain one regularization parameter, which adapts to both Hessian inexactness

Algorithm 3 Adaptive damped A-NPE with an inexact Hessian

Input: $x_0 \in \mathbb{R}^d$ and $\eta_0 > 0$.
$(w_1, \eta_1) := \text{MSBACKTRACK}(x_0, \eta_0)$.
Set $A_1 := 1/\eta_1,\ z_1 := x_0 - A_1 \nabla f(w_1)$.
for $k \geq 1$ **do**
  Set $a'_k := \frac{1+\sqrt{1+4\eta_k A_k}}{2\eta_k},\ A'_k := A_k + a'_k,\ v_k := \frac{A_k w_k + a'_k z_k}{A'_k}$.
  $(y_k, \tau_k) := \text{MSBACKTRACK}(v_k, \eta_k)$.
  Set $\gamma_k := \frac{\eta_k}{\tau_k},\ a_{k+1} := \gamma_k a'_k,\ A_{k+1} := A_k + a_{k+1}$.
  Set $w_{k+1} := \frac{(1-\gamma_k)A_k w_k + \gamma_k A'_k y_k}{A_{k+1}},\ z_{k+1} := z_k - a_{k+1}\nabla f(y_k)$.
  if $\tau_k = \eta_k$ **then**
    Set $\eta_{k+1} := \eta_k/2$.
  else
    Set $\eta_{k+1} := 2\eta_k$.

and local smoothness, without requiring them as an input. The intuition behind this acceleration scheme can be found in the proof sketch below. While the outer scheme of Algorithm 3 matches the method of Carmon et al. (2022), our convergence analysis is different. We discuss the key differences below.

For the convergence bounds of Algorithm 3, define

$$B := \max\left\{1, (6\sqrt{M_1}R)^{2/3}, (2R/\rho)^{2/3}\right\}, \quad \varepsilon_{\text{tr}} := \max\left\{48M_0R^3/B^{7/2}, 16\delta R^2/B^2\right\},$$

where $\rho = +\infty$ under (GH). Let $\overline{\tau}_0 := \max\left\{4\delta, 2\sqrt{L(x_0)\|\nabla f(x_0)\|}, 2\|\nabla f(x_0)\|/\rho\right\}$.

Theorem 4.1. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. For any $\varepsilon > 0$, the method generates an iterate w_N satisfying $f(w_N) - f^* \leq \varepsilon$ within*

$$\mathcal{O}\left(B\left(1 + \log_+ \frac{\max\{\eta_0, \overline{\tau}_0\}R^2}{\max\{\varepsilon, \varepsilon_{\text{tr}}\}}\right) + \left(\frac{M_0R^3}{\varepsilon}\right)^{2/7} + \sqrt{\frac{\delta R^2}{\varepsilon}}\right) \tag{6}$$

iterations. Moreover, there exists iterate $K \geq 1$ with $K = \mathcal{O}\left(B\max\left\{1, \log\frac{\max\{\eta_0, \overline{\tau}_0\}R^2}{\varepsilon_{\text{tr}}}\right\}\right)$ such that

$$f(w_k) - f^* = \mathcal{O}\left(\max\{\eta_0, \overline{\tau}_0\}R^2 \exp\left(-\frac{c(k-1)}{B}\right)\right), \qquad 1 \leq k < K, \tag{7}$$

$$f(w_{K+n}) - f^* = \mathcal{O}\left(\frac{M_0R^3}{(B+n)^{7/2}} + \frac{\delta R^2}{(B+n)^2}\right), \qquad n = 0, 1, \ldots. \tag{8}$$

Theorem 4.1 bounds outer iterations, which equals the number of inexact Hessian evaluations. In Appendix F.4, we show that the number of gradient evaluations is of the same order as (6) up to two additive terms $\mathcal{O}(\log_+ \frac{\overline{\tau}_0}{\eta_0})$ and $\widetilde{\mathcal{O}}(B(1 + \log_+ \frac{\max\{\eta_0, \overline{\tau}_0\}R^2}{\varepsilon})^{2/3})$, where $\widetilde{\mathcal{O}}$ hides logarithmic factor independent of ε. By (7) and (8), the method first converges linearly, and then converges in Lipschitz Hessian regime.

Under (H) the complexity is $\mathcal{O}(1 + \log_+ \frac{\max\{\eta_0, \overline{\tau}_0\}}{LR+\delta} + (\frac{LR^3}{\varepsilon})^{2/7} + (\frac{\delta R^2}{\varepsilon})^{1/2})$, matching the optimal method of Chen et al. (2026) and the corresponding lower bound (Agafonov et al., 2024b) up to an additive term independent of ε. Gradient computations have additional term $\mathcal{O}(\log_+ \frac{\overline{\tau}_0}{\eta_0})$. To the best of our knowledge, this is the first second-order method to simultaneously achieve full adaptivity, optimal complexity, and robustness to inexact Hessians. The complete proof is given in Appendix F.

Proof sketch. Let $s_i := y_i - v_i$ and $\Lambda(A) := \max\left\{32\delta, (96M_0R)^{4/7}A^{-3/7}, B^2/A\right\}$. By the algorithm construction, the regularization guess η_i is driven toward $\Lambda(A_i)$. $\Lambda(A)$ serves as an ideal regularization scale, which will allow us to recover the optimal Lipschitz-Hessian rate near the solution and a geometric rate far from it.

We next use the standard Monteiro-Svaiter potential argument. Let $\mathcal{E}_k = A_k(f(w_k) - f^*) + 1/2\|z_k - x^*\|^2$. Then, $3/8\sum_{i=1}^{k-1} A'_i\eta_i\|s_i\|^2 \leq \mathcal{E}_1 \leq R^2/2$. We first count two types of iterations:

$$\mathcal{S}^{>}_{k-1} := \{i \in \{1, \ldots, k-1\} : \tau_i > \eta_i,\ \eta_i \geq \Lambda(A_i)/2\},$$
$$\mathcal{S}^{=}_{k-1} := \{i \in \{1, \ldots, k-1\} : \tau_i = \eta_i,\ \eta_i \leq 2\Lambda(A_i)\}.$$

On $\mathcal{S}^{>}_{k-1}$, at least one trial is rejected and η_i is sufficiently large. Lemma 3.2, together with $A'_i \geq A_i$ lower-bounds the corresponding terms in the potential estimate:

$$\frac{3R^2}{16\sqrt{A_k\Lambda(A_k)}}|\mathcal{S}^{>}_{k-1}| \leq \frac{3}{8}\sum_{i\in\mathcal{S}^{>}_{k-1}} A'_i\eta_i\|s_i\|^2 \leq \frac{R^2}{2} \quad \Rightarrow \quad |\mathcal{S}^{>}_{k-1}| \leq \frac{8}{3}\sqrt{A_k\Lambda(A_k)}.$$

Here, the condition on η_i yields sufficient potential decrease, similarly to how controlled iterations in the basic method yield sufficient function decrease. Note, that substituting $A_k \asymp R^2/\varepsilon$ and definition of $\Lambda(A_k)$ into $\sqrt{A_k\Lambda(A_k)}$ recovers the desired rate (6).

On $\mathcal{S}^{=}_{k-1}$, no trial is rejected, so $\gamma_i = 1$ and $A_{i+1} = A_i + a'_i$. The upper bound on $\eta_i = \tau_i$ prevents overregularization. Using the update in line 7, we have

$$\frac{a'_i}{A_i} \geq \frac{1}{\sqrt{2A_i\Lambda(A_i)}} \;\Rightarrow\; \tfrac{1}{3} \leq \sqrt{A_i\Lambda(A_i)}\log\tfrac{A_{i+1}}{A_i} \leq \textstyle\int_{A_i}^{A_{i+1}} \sqrt{\frac{\Lambda(t)}{t}}dt. \tag{9}$$

Thus, on these iterations the upper bound on η_i guarantees fast growth of A_i. Summing the last inequality over $\mathcal{S}^{=}_{k-1}$ and calculating integral gives the upper bound (6). Note that Λ is chosen to match the progress on $\mathcal{S}^{>}_{k-1}$ and $\mathcal{S}^{=}_{k-1}$.

Iterations that are not in $\mathcal{S}^{>}_{k-1} \cup \mathcal{S}^{=}_{k-1}$ improve regularization guess. If $\tau_i = \eta_i$ and $\eta_i > 2\Lambda(A_i)$, the algorithm halves η_i. Otherwise, $\tau_i > \eta_i$ and $\eta_i < \Lambda(A_i)/2$, and it doubles η_i. The last challenge is to bound the number of such iterations. Similarly to uncontrolled iterations in the basic method, we show

$$k - 1 \leq 2(|\mathcal{S}^{>}_{k-1}| + |\mathcal{S}^{=}_{k-1}|) + \log(A_k\Lambda(A_1)).$$

Combining iteration counts yields

$$k \lesssim B\log_+ \min\{A_k, A_{\mathrm{tr}}\}/A_1 + \sqrt{A_k\Lambda(A_k)},$$

where $A_{\mathrm{tr}} := R^2/(2\varepsilon_{\mathrm{tr}})$. Since $f(w_k) - f^* \leq R^2/(2A_k)$, it suffices to reach $A_\varepsilon := R^2/(2\varepsilon)$. For $\varepsilon \geq \varepsilon_{\mathrm{tr}}$, we have $A_\varepsilon \leq A_{\mathrm{tr}}$ and $\sqrt{A_\varepsilon\Lambda(A_\varepsilon)} = B$, so the complexity depends logarithmically on the target accuracy. For $\varepsilon < \varepsilon_{\mathrm{tr}}$, the logarithmic term is constant, while

$$\sqrt{A_\varepsilon\Lambda(A_\varepsilon)} \simeq \left(M_0R^3/\varepsilon\right)^{2/7} + \sqrt{\delta R^2/\varepsilon}.$$

4.3 Related Proof Techniques

Algorithmically, our method builds upon the optimal MS acceleration of Carmon et al. (2022). Recently, Chen et al. (2026) established the optimal rate for inexact Hessians using the schedule of Adil et al. (2026), but their method is non-adaptive. Below, we discuss the key challenges of extending their analyses to an adaptive method under (LH) and (IH).

Comparison with Carmon et al. (2022). They split iterations into two sets: $\mathcal{S}^{\uparrow} = \{i : \tau_i > \eta_i\}$, on which the potential bound $\sum_{i\in\mathcal{S}^{\uparrow}} A'_i\eta_i\|s_i\|^2 \lesssim R^2$ is used, and $\mathcal{S}^{\downarrow} = \{i : \tau_i = \eta_i\}$, on which the growth of A_i is given by the weight identity $\eta_i(a'_i)^2 = A'_i$. In the latter case, they use the bound of type $\sqrt{A_{i+1}} - \sqrt{A_i} \geq 1/(2\sqrt{\eta_i})$, which is enough for polynomial rate. The first difference is that we use log form in (9) to get geometric far rate.

From this point the proofs follow different ideas. Carmon et al. (2022) transfer the growth of A_i from $\mathcal{S}^{\downarrow}$ to neighboring iterations of $\mathcal{S}^{\uparrow}$. The reason is that on $\mathcal{S}^{\uparrow}$ they have $\|s_i\| \gtrsim \tau_i/L$ because of rejection, which turns the potential bound into $\sum_{i\in\mathcal{S}^{\uparrow}} A'_i\eta_i^3 \lesssim L^2R^2$. Next, since both bounds are on $\mathcal{S}^{\uparrow}$, they use reverse Hölder inequality to remove η_i and get the rate $A_k \gtrsim k^{7/2}/(LR)$.

Because of assumptions (LH) and (IH) we do not have such lower bound on $\|s_i\|$. Instead, Lemma 3.2 gives three regimes for rejected steps: it is proportional to τ_i for $\delta \lesssim \tau_i \lesssim M_0r_0$, it is bounded below only by r_0 for larger τ_i, and it may be arbitrary small for $\tau_i \lesssim \delta$. Thus, we introduce $\Lambda(A)$, whose three terms correspond to these three regimes, and follow the proof of Theorem 3.1 instead. $\Lambda(A_i)$ plays the role of $\Theta_{\mathrm{c}}(g)$, iterations of $\mathcal{S}^{>}_{k-1}$ and $\mathcal{S}^{=}_{k-1}$ (because of additional bound on guess η_i) make progress in the potential or show the growth of A_i. The remaining iterations move η_i toward $\Lambda(A_i)$, so their number is bounded by the multiplier on number of steps with progress and additional logarithm (similarly to uncontrolled iterations in the basic method).

Comparison with Chen et al. (2026). The cubic step is used with explicit regularization ($\tau_i = 2\delta + M_0\|s_i\|$), and the guess is $\eta_i \simeq \delta + (M_0R)^{4/7}A_i^{-3/7}$ (Chen et al., 2026, Lemma 3.1 and Theorem 3.1). Their acceleration scheme keeps η_i fixed until A_i doubles and then resets it. Under (H) this matches our $\Lambda(A)$ up to constants for $A \geq A_{\mathrm{tr}}$. However, the analysis is based on the fact that between two resets the guess is a known constant of order at least δ, while an adaptive guess changes at every iteration and can be arbitrarily small. We resolve this challenge by counting only iterations with $\eta_i \asymp \Lambda(A_i)$ and bounding the remaining as described above.

5 Experiments

We compare our methods with adaptive second-order baselines on logistic and Poisson regression (Figure 1). We use datasets from LIBSVM (Chang & Lin, 2011) `gisette`, `german_numer` and `ijcnn1` for logistic regression, and `abalone` for Poisson regression. Methods start from

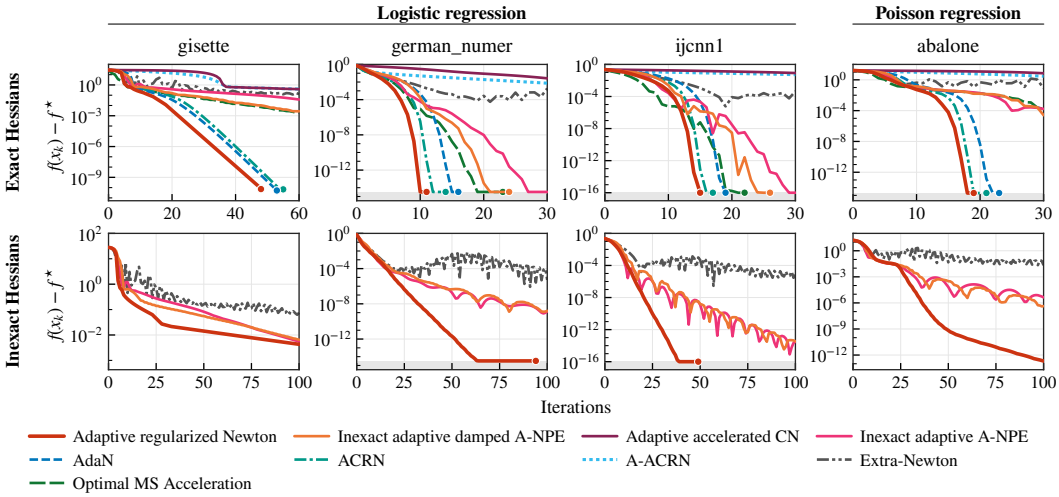

Figure 1: Top row: exact Hessians. Bottom row: Hessians computed on a fixed random 10% of the samples, with exact gradients.

$x_0 = -\mathbf{1}$. We run Algorithms 2, 3, 5 and 6 against AdaN (Mishchenko, 2023), the adaptive cubic regularized Newton (ACRN), its accelerated version (A-ACRN) (Kamzolov et al., 2025b), Extra-Newton (Antonakopoulos et al., 2022), and Optimal MS Acceleration (Carmon et al., 2022)

Parameters. All methods are adaptive, so we do not tune them. For a fair comparison, Algorithms 2, 3 and 6 start backtracking with $\eta_0 = \sqrt{L(x_0)\|\nabla f(x_0)\|}$, Optimal MS Acceleration takes $\lambda_0' = \eta_0$, $\sigma = 1/2$, and AdaN takes $H_0 = L(x_0)/2$, which it doubles before its first trial. For the step-size parameters of Extra-Newton, we take $\gamma = 2\|x_0\|$ as a proxy for the diameter D, following Antonakopoulos et al. (2022, Remark 3.1), and $\beta_0 = L(x_0)\gamma^2\|\nabla f(x_0)\|$. Thus, these six methods perform the same first trial step. ACRN, A-ACRN and Algorithm 5 start from the standard theoretical choice of the cubic regularization parameter, $\mathcal{M}_0 = L(x_0)$.

Inexact Hessians. We consider a centralized distributed setup with $f(x) = \frac{1}{m}\sum_{j=1}^m f_j(x)$, where $f_j(x)$ is the local loss stored on machine j. The server also stores local data and computes the Hessian $H(x) = \nabla^2 f_1(x)$. Under the β-similarity this Hessian satisfies (IH) with $\delta = \beta$ (Zhang & Lin, 2015; Agafonov et al., 2021). To perform a step, the server collects the gradients from the workers and uses the full gradient with its local Hessian. In our experiments, $H(x)$ is computed from a fixed uniformly sampled subset of $\lceil n/10 \rceil$ observations, where n is the number of training samples. We compare only parameter-free methods that allow inexact Hessians.

Results. Under exact Hessians, our Adaptive regularized Newton converges faster on all four problems, with ACRN close behind. Among the accelerated methods, our A-NPE methods (Algorithms 3 and 6) perform best, and the optimal Algorithm 3 usually outperforms the suboptimal Algorithm 6. As in the logistic regression experiments of Carmon et al. (2022), the non-accelerated methods outperform the accelerated ones. However, on toy problems satisfying (LH), optimal acceleration outperforms non-accelerated methods and suboptimal accelerated ones when starting far from the solution (Appendix G.2). With inexact Hessians, our three methods outperform Extra-Newton, and Algorithm 2 is again the best. Experimental details and additional results are provided in Appendix G.

LIMITATIONS

First, our analysis requires exact gradients, and we leave the stochastic setting for future work. Even under (H), the theory of stochastic second-order methods is incomplete: the only methods with the proper dependence on the stochastic gradient variance are suboptimal accelerated methods (Antonakopoulos et al., 2022; Agafonov et al., 2024b). Second, while being parameter-free, our methods still require backtracking. Combining optimal rates with line-search-free adaptivity is an interesting direction for future work. Finally, our theory is limited to convex optimization. Still, convex analysis is informative for nonconvex learning: Nesterov momentum (Nesterov, 1983), developed for convex problems, is important for training deep networks (Sutskever et al., 2013), adaptive gradient methods originate from online convex optimization (Duchi et al., 2011), and learning-rate schedules for large model training closely follow bounds from convex optimization theory (Schaipp et al., 2025).

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, assist with code writing and with parts of the theoretical analysis. We have not used generative AI tools for retrieval and discovery (e.g., finding related work), writing sections of the paper, or generating synthetic datasets. All research ideas, including the proposed methods, as well as the core analysis, are original work of the authors. We have reviewed all AI-assisted work.

Text and code. We used OpenAI ChatGPT, Anthropic Claude, Codex, and Claude Code to edit author-written text and assist with code development.

Mathematical claims. We used Codex to transcribe images of our handwritten proofs into LaTeX. We used ChatGPT and Codex to assist with reviewing proofs throughout the paper. With Codex, we additionally used the Lean 4 skills package [2]. Examples in Section B were proved with assistance of AI. In the convex analysis of the basic method, AI assisted with deriving the bound on the number of controlled iterations in Lemma C.6. For the strongly convex analysis, AI generated an initial draft extending the proof according to a scheme specified by us and based on Kamzolov et al. (2025a). We then iteratively reviewed and refined this analysis. For Algorithm 5, AI generated an initial draft extending the non-adaptive analysis to backtracking strategy specified by us. We subsequently reviewed and refined the analysis through several iterations with AI assistance. For the optimal method, AI assisted with iterative refinement of our initial proof, which was based on ideas from Algorithm 2 and Carmon et al. (2022). The proofs for the suboptimal method, as well as the oracle complexities for all methods, were initially derived by AI from the Hessian convergence rates and our proof for the optimal method.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical claims are proved in the Appendix, and the code for the experiments is available at `https://anonymous.4open.science/r/accelerated_l0_l1-6C02/README.md`.

REFERENCES

Deeksha Adil, Brian Bullins, Arun Jambulapati, and Aaron Sidford. Convex optimization with p-norm oracles. In *37th International Conference on Algorithmic Learning Theory*, 2026. URL https://openreview.net/forum?id=gZax9QXNyR.

Artem Agafonov, Pavel Dvurechensky, Gesualdo Scutari, Alexander Gasnikov, Dmitry Kamzolov, Aleksandr Lukashevich, and Amir Daneshmand. An accelerated second-order method for distributed stochastic optimization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 2407–2413, 2021. ISBN 2576-2370. doi: 10.1109/CDC45484.2021.9683400. URL https://doi.org/10.1109/CDC45484.2021.9683400.

Artem Agafonov, Dmitry Kamzolov, Pavel Dvurechensky, Alexander Gasnikov, and Martin Takáč. Inexact tensor methods and their application to stochastic convex optimization. *Optimization Methods and Software*, 39(1):42–83, 2024a. doi: 10.1080/10556788.2023.2261604. URL https://doi.org/10.1080/10556788.2023.2261604.

Artem Agafonov, Dmitry Kamzolov, Alexander Gasnikov, Ali Kavis, Kimon Antonakopoulos, Volkan Cevher, and Martin Takáč. Advancing the lower bounds: an accelerated, stochastic, second-order method with optimal adaptation to inexactness. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=otU31x3fus.

Kimon Antonakopoulos, Ali Kavis, and Volkan Cevher. Extra-Newton: A first approach to noise-adaptive accelerated second-order methods. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29859–29872. Curran Associates, Inc.,

[2] https://github.com/cameronfreer/lean4-skills

2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c10804702be5a0cca89331315413f1a2-Paper-Conference.pdf.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178:327–360, 2019. ISSN 1436-4646. doi: 10.1007/s10107-018-1293-1. URL https://doi.org/10.1007/s10107-018-1293-1.

S Bellavia, G Gurioli, B Morini, and Ph.L. Toint. Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives. *Journal of Complexity*, 68: 101591, 2022. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2021.101591. URL https://www.sciencedirect.com/science/article/pii/S0885064X21000467.

Stefania Bellavia, Nataša Krejić, and Nataša Krklec Jerinkić. Subsampled inexact Newton methods for minimizing large sums of convex functions. *IMA Journal of Numerical Analysis*, 40(4):2309–2341, 10 2020. ISSN 0272-4979. doi: 10.1093/imanum/drz027. URL https://doi.org/10.1093/imanum/drz027.

Ekaterina Borodich and Dmitry Kovalev. Nesterov finds GRAAL: Optimal and adaptive gradient method for convex optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 95533–95558, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9a419ce12db8db2e35ee68c1f58c1a36-Paper-Conference.pdf.

Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive Monteiro-Svaiter acceleration. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20338–20350. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7ff97417474268e6b5a38bcbfae04944-Paper-Conference.pdf.

Coralia Cartis, Nicholas I M Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127:245–295, 2011. ISSN 1436-4646. doi: 10.1007/s10107-009-0286-5. URL https://doi.org/10.1007/s10107-009-0286-5.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199. URL https://doi.org/10.1145/1961189.1961199.

Lesi Chen, Chengchang Liu, Luo Luo, John Lui, and Jingzhao Zhang. Optimal convex optimization with inexact second-order oracles. *arXiv preprint arXiv:2607.24520*, 2026.

Xi Chen, Bo Jiang, Tianyi Lin, and Shuzhong Zhang. Accelerating adaptive cubic regularization of Newton's method via random sampling. *Journal of Machine Learning Research*, 23(90):1–38, 2022. URL http://jmlr.org/papers/v23/20-910.html.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5396–5427. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ar.html.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9955–9968. Curran Associates, Inc., 2022. doi: 10.52202/068431-0723. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40924475a9bf768bdac3725e67745283-Paper-Conference.pdf.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical Programming*, 2023. ISSN 1436-4646. doi: 10.1007/s10107-023-01943-7. URL https://doi.org/10.1007/s10107-023-01943-7.

Nikita Doikov, El Mahdi Chayti, and Martin Jaggi. Second-order optimization with lazy Hessians. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 8138–8161. PMLR, 9 2023. URL https://proceedings.mlr.press/v202/doikov23a.html.

Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-Universal regularized Newton method. *SIAM Journal on Optimization*, 34:27–56, 2024. doi: 10.1137/22M1519444. URL https://doi.org/10.1137/22M1519444.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Shaoying Fang, Naoki Marumo, and Akiko Takeda. Parameter-free cubic-regularized newton method: Sharp complexity and generalized smoothness. *arXiv preprint arXiv:2607.10741*, 2026.

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

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtarik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97261–97305, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f21a5fedfbcb55406ddaf51cf90d494a-Paper-Conference.pdf.

G. N. Grapiglia and Yu. Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. *SIAM Journal on Optimization*, 30(4):2750–2779, 2020. doi: 10.1137/19M1259432. URL https://doi.org/10.1137/19M1259432.

Sadok Jerad. Fast adaptive tensor methods under local smoothness. *arXiv preprint arXiv:2606.30225*, 2026.

Dmitry Kamzolov, Artem Agafonov, Dmitry Pasechnyuk, Alexander Gasnikov, and Martin Takáč. OPTAMI: Global superlinear convergence of high-order methods. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Cpr6Wv2tfr.

Dmitry Kamzolov, Klea Ziu, Artem Agafonov, and Martin Takáč. Accelerated adaptive cubic regularized quasi-Newton methods. *Journal of Optimization Theory and Applications*, 208:27, 2025b. ISSN 1573-2878. doi: 10.1007/s10957-025-02804-3. URL https://doi.org/10.1007/s10957-025-02804-3.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1895–1904. PMLR, 5 2017. URL https://proceedings.mlr.press/v70/kohler17a.html.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35339–35351. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e56f394bbd4f0ec81393d767caa5a31b-Paper-Conference.pdf.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 40238–40271. Curran Associates, Inc., 2023. doi: 10.52202/075280-1749. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7e8bb8d17bb1cb24dfe972a2f8ff2500-Paper-Conference.pdf.

Konstantin Mishchenko. Regularized Newton method with global $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence. *SIAM Journal on Optimization*, 33:1440–1462, 2023. doi: 10.1137/22M1488752. URL https://doi.org/10.1137/22M1488752.

Renato D C Monteiro and B F Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23:1092–1125, 2013. doi: 10.1137/110833786. URL https://doi.org/10.1137/110833786.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983. (In Russian).

Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112:159–181, 2008. ISSN 1436-4646. doi: 10.1007/s10107-006-0089-x. URL https://doi.org/10.1007/s10107-006-0089-x.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Cham, 2 edition, 2018. ISBN 978-3-319-91577-7. doi: 10.1007/978-3-319-91578-4.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006. doi: 10.1007/s10107-006-0706-8. URL https://doi.org/10.1007/s10107-006-0706-8.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53267–53294. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/schaipp25a.html.

Andrei Semenov, Martin Jaggi, and Nikita Doikov. Gradient-normalized smoothness for optimization with approximate Hessians. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 13974–14029, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1759a83b007f0685c3fbc460fa1b6395-Paper-Conference.pdf.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60493–60514. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tyurin25a.html.

Alexander Tyurin. Near-optimal convergence of accelerated gradient methods under generalized and (L_0, L_1)-smoothness. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=L0T4pmqkMg.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 15953–15979, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/28a9255e2a0e7062e8e0af715f2e5b86-Paper-Conference.pdf.

Chenghan Xie, Chenxi Li, Chuwen Zhang, Qi Deng, Dongdong Ge, and Yinyu Ye. Trust region methods for nonconvex stochastic optimization beyond Lipschitz smoothness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16049–16057, 2024.

Peng Xu, Fred Roosta, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184:35–70, 2020. ISSN 1436-4646. doi: 10.1007/s10107-019-01405-z. URL https://doi.org/10.1007/s10107-019-01405-z.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

Yuchen Zhang and Xiao Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 362–370, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/zhangb15.html.

CONTENTS

A Taylor Remainder Lemmas **16**

B Proofs for the Relations between Assumptions and Examples **16**

C Proof of Algorithm 2 **19**

C.1 Step Estimates 20

C.2 Iteration Count 23

C.3 Proof of Theorem 3.1 24

C.4 Far and Near Convergence Rates 27

C.5 Oracle complexity 29

C.6 Strongly Convex Analysis 30

D Nesterov-type Acceleration **37**

D.1 Proof of Nesterov-type Acceleration under (GH) 38

D.2 Adaptive Method under (LH) with Exact Hessian 45

E Suboptimal A-NPE Method **51**

E.1 Auxiliary Lemmas 53

E.2 Step Estimates 54

E.3 Explicit Choice of Regularization Parameter 56

E.4 Main-Loop Backtracking 57

E.5 Proof of Theorem E.1 58

E.6 Oracle Complexity 61

F Proof of Algorithm 3 **62**

F.1 Initialization and Backtracking 62

F.2 Step Estimates 62

F.3 Proof of Theorem 4.1 64

F.4 Oracle Complexity 69

G Extra Experiments **72**

G.1 Experiment Details 72

G.2 Additional Experiments on (LH) Functions 72

G.3 Sensitivity to Initial Parameters 73

A Taylor Remainder Lemmas

The following two lemmas record the Taylor remainder bounds used throughout our analysis.

Lemma A.1. *Suppose that* (GH) *holds. Then, for all* $x, s \in \mathbb{R}^d$*,*

$$\|\nabla f(x+s) - \nabla f(x) - \nabla^2 f(x)s\| \le \tfrac{L(x)}{2}\|s\|^2,$$
$$\left|f(x+s) - f(x) - \langle \nabla f(x), s\rangle - \tfrac{1}{2}\left\langle \nabla^2 f(x)s, s\right\rangle\right| \le \tfrac{L(x)}{6}\|s\|^3.$$

Lemma A.2. *Suppose that* (LH) *holds. Then, for all* $x, s \in \mathbb{R}^d$ *satisfying* $\|s\| \le \rho$*,*

$$\|\nabla f(x+s) - \nabla f(x) - \nabla^2 f(x)s\| \le \tfrac{L(x)}{2}\|s\|^2,$$
$$\left|f(x+s) - f(x) - \langle \nabla f(x), s\rangle - \tfrac{1}{2}\left\langle \nabla^2 f(x)s, s\right\rangle\right| \le \tfrac{L(x)}{6}\|s\|^3.$$

Proof. We follow Nesterov (2018, Lemma 1.2.4). For $t \in [0,1]$, (GH), or (LH) with $\|ts\| \le \rho$, gives $\|\nabla^2 f(x+ts) - \nabla^2 f(x)\| \le tL(x)\|s\|$. Hence,

$$\|\nabla f(x+s) - \nabla f(x) - \nabla^2 f(x)s\| = \|\int_0^1 \left(\nabla^2 f(x+ts) - \nabla^2 f(x)\right) s\mathrm{d}t\|$$
$$\le \int_0^1 tL(x)\|s\|^2\mathrm{d}t = \frac{L(x)}{2}\|s\|^2.$$

Since $\|ts\| \le \|s\|$, the same bound holds for the step ts:

$$\|\nabla f(x+ts) - \nabla f(x) - t\nabla^2 f(x)s\| \le \frac{L(x)}{2}t^2\|s\|^2, \qquad t \in [0,1].$$

Using $f(x+s) - f(x) = \int_0^1 \langle \nabla f(x+ts), s\rangle\,\mathrm{d}t$ and $\int_0^1 t\mathrm{d}t = \frac{1}{2}$, we obtain

$$\left|f(x+s) - f(x) - \langle \nabla f(x), s\rangle - \frac{1}{2}\left\langle \nabla^2 f(x)s, s\right\rangle\right|$$
$$= \left|\int_0^1 \left\langle \nabla f(x+ts) - \nabla f(x) - t\nabla^2 f(x)s, s\right\rangle \mathrm{d}t\right| \le \int_0^1 \frac{L(x)}{2}t^2\|s\|^3\mathrm{d}t = \frac{L(x)}{6}\|s\|^3.$$

□

B Proofs for the Relations between Assumptions and Examples

Proof of Example 2.1. Direct differentiation gives

$$f'(x) = 2x + \int_0^x \sin(t^2)\mathrm{d}t, \qquad f''(x) = 2 + \sin(x^2), \qquad f^{(3)}(x) = 2x\cos(x^2).$$

Thus, $f''(x) \ge 1$, so f is 1-strongly convex. Moreover, $f'(0) = 0$ and

$$|f'(x)| = \left|\int_0^x f''(t)\mathrm{d}t\right| \ge |x|.$$

If $|y - x| \le 1$, then

$$|f''(y) - f''(x)| = |\sin(y^2) - \sin(x^2)| \le |y^2 - x^2| \le (2|x| + 1)\,|y - x| \le (2 + 2|f'(x)|)\,|y - x|.$$

If $|y - x| > 1$, then

$$|f''(y) - f''(x)| \le 2 \le 2|y - x| \le (2 + 2|f'(x)|)\,|y - x|.$$

Hence, (GH) holds with $M_0 = M_1 = 2$. On the other hand, for $x_n = \sqrt{2\pi n}$,

$$|f^{(3)}(x_n)| = 2\sqrt{2\pi n} \longrightarrow +\infty.$$

Condition (H) would imply $|f^{(3)}(x)| \le L$ by taking the directional limit in (H). Therefore, (H) does not hold. □

Example B.1. The 1-strongly convex function $f(x) = \cosh x - 1$ satisfies (LH) for every $\rho > 0$ with $M_0 = \sinh \rho$ and $M_1 = e^\rho$, but it does not satisfy (GH).

Proof. We have

$$f'(x) = \sinh x, \qquad f''(x) = \cosh x, \qquad f^{(3)}(x) = \sinh x.$$

Since $f''(x) \geq 1$, the function is 1-strongly convex. Let $h = y - x$ and $|h| \leq \rho$. For every $t \in [0, 1]$,

$$\begin{aligned} |\sinh(x + th)| &\leq |\sinh x| \cosh(th) + \cosh x |\sinh(th)| \\ &\leq |\sinh x| \cosh \rho + \cosh x \sinh \rho \leq \sinh \rho + e^\rho |\sinh x|, \end{aligned}$$

where we used $\cosh x \leq 1 + |\sinh x|$. Therefore,

$$|f''(y) - f''(x)| \leq |h| \int_0^1 |\sinh(x + th)| \mathrm{d}t \leq (\sinh \rho + e^\rho |f'(x)|)\, |y - x|.$$

This proves (LH). If (GH) held, then the choice $x = 0$ would give

$$\cosh y - 1 \leq M_0 |y|, \qquad y \in \mathbb{R},$$

which is impossible as $y \to +\infty$. □

Proof of Example 2.2. Let $\mu_i(x) := e^{\langle a_i, x \rangle}$. We have

$$\nabla f(x) = \sum_{i=1}^n (\mu_i(x) - b_i)\, a_i, \qquad \nabla^2 f(x) = \sum_{i=1}^n \mu_i(x) a_i a_i^\top \succeq 0.$$

Since $[a_i]_1 = 1$ and $b_i \geq 0$, the first coordinate of $\nabla f(x)$ equals $\sum_{i=1}^n \mu_i(x) - \|b\|_1$. Hence,

$$\sum_{i=1}^n \mu_i(x) \leq \|b\|_1 + \|\nabla f(x)\|. \tag{10}$$

Let $s = y - x$ and $\|s\| \leq \rho$. Then $|\langle a_i, s \rangle| \leq c\|s\| \leq c\rho$. Using $|e^u - 1| \leq e^{|u|}|u|$ for $u \in \mathbb{R}$, we obtain

$$\begin{aligned} \|\nabla^2 f(y) - \nabla^2 f(x)\| &= \left\| \sum_{i=1}^n \mu_i(x) \left(e^{\langle a_i, s \rangle} - 1 \right) a_i a_i^\top \right\| \leq c^3 e^{c\rho} \|s\| \sum_{i=1}^n \mu_i(x) \\ &\overset{(10)}{\leq} \left(\|b\|_1 c^3 e^{c\rho} + c^3 e^{c\rho} \|\nabla f(x)\| \right) \|y - x\|. \end{aligned}$$

This proves (LH) with the stated constants. Let e_1 be the first coordinate vector. Since $\langle a_i, e_1 \rangle = 1$, we have $\nabla^2 f(x + te_1) = e^t \nabla^2 f(x)$ for $t \in \mathbb{R}$ and

$$\|\nabla^2 f(x)\| \geq \left\langle \nabla^2 f(x) e_1, e_1 \right\rangle = \sum_{i=1}^n \mu_i(x) > 0.$$

If (GH) held, then, for all $t > 0$,

$$\left(e^t - 1\right) \|\nabla^2 f(x)\| = \|\nabla^2 f(x + te_1) - \nabla^2 f(x)\| \leq (M_0 + M_1 \|\nabla f(x)\|)\, t,$$

which is impossible as $t \to +\infty$. □

Proof of Example 2.3. Let $a_i^\top$ be the rows of A, $r := Ax - b$, and $\psi(u) := |u|^{p-2} u$ for $u \in \mathbb{R}$. We have

$$\nabla f(x) = \sum_{i=1}^n \psi(r_i) a_i, \qquad \nabla^2 f(x) = (p - 1) \sum_{i=1}^n |r_i|^{p-2} a_i a_i^\top \succeq 0.$$

First, we bound $\|r\|_p$ by $\|\nabla f(x)\|$. Let P be the orthogonal projector onto the range of $A^\top$. The gradient $\nabla f(x) = A^\top \psi(r)$ belongs to this range, and $\|Az\| \geq \sigma_+(A)\|z\|$ for all z in this range. Therefore, Hölder's inequality and $\|v\| \leq \sqrt{n}\|v\|_\infty \leq \sqrt{n}\|v\|_p$ for $v \in \mathbb{R}^n$ give

$$\|r\|_p^p - \|r\|_p^{p-1}\|b\|_p \leq \sum_{i=1}^n \psi(r_i)\,(r_i + b_i) = \langle \nabla f(x), x\rangle = \langle \nabla f(x), Px\rangle \leq \frac{\|\nabla f(x)\|\|APx\|}{\sigma_+(A)}$$
$$= \frac{\|\nabla f(x)\|\|r+b\|}{\sigma_+(A)} \leq \frac{\sqrt{n}}{\sigma_+(A)}\|\nabla f(x)\|\left(\|r\|_p + \|b\|_p\right).$$

If $\|r\|_p > \max\{1, 2\|b\|_p\}$, then $\|r\|_p - \|b\|_p > \|r\|_p/2$ and $\|r\|_p + \|b\|_p < 3\|r\|_p/2$. In this case, the previous display gives $\|r\|_p^{p-1} < \frac{3\sqrt{n}}{\sigma_+(A)}\|\nabla f(x)\| = \frac{M_1}{M_0}\|\nabla f(x)\|$, and $\|r\|_p > 1$ implies

$$(\|r\|_p + \|A\|\rho)^{p-3} \leq (1 + \|A\|\rho)^{p-3}\|r\|_p^{p-3} \leq (1 + \|A\|\rho)^{p-3}\|r\|_p^{p-1}$$
$$< (1 + 2\|b\|_p + \|A\|\rho)^{p-3}\,\frac{M_1}{M_0}\|\nabla f(x)\|.$$

Otherwise, $\|r\|_p \leq 1 + 2\|b\|_p$ and $(\|r\|_p + \|A\|\rho)^{p-3} \leq (1 + 2\|b\|_p + \|A\|\rho)^{p-3}$. In both cases, the definition of M_0 gives

$$(p-1)(p-2)\|A\|^3\left(\|r\|_p + \|A\|\rho\right)^{p-3} \leq M_0 + M_1\|\nabla f(x)\|. \tag{11}$$

Now let $s = y - x$ and $\|s\| \leq \rho$. Then $Ay - b = r + As$ and $|\,\langle a_i, s\rangle\,| \leq \|A\|\|s\| \leq \|A\|\rho$. By the mean value theorem and $|r_i| \leq \|r\|_p$,

$$\left||r_i + \langle a_i, s\rangle|^{p-2} - |r_i|^{p-2}\right| \leq (p-2)\left(|r_i| + |\,\langle a_i, s\rangle\,|\right)^{p-3}|\,\langle a_i, s\rangle\,|$$
$$\leq (p-2)\left(\|r\|_p + \|A\|\rho\right)^{p-3}\|A\|\|s\|.$$

Since $\|\sum_{i=1}^n w_i a_i a_i^\top\| \leq \max_i |w_i|\|A^\top A\|$ for $w \in \mathbb{R}^n$, we obtain

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| = (p-1)\left\|\sum_{i=1}^n \left(|r_i + \langle a_i, s\rangle|^{p-2} - |r_i|^{p-2}\right) a_i a_i^\top\right\|$$
$$\leq (p-1)(p-2)\|A\|^3\left(\|r\|_p + \|A\|\rho\right)^{p-3}\|s\| \overset{(11)}{\leq} \left(M_0 + M_1\|\nabla f(x)\|\right)\|y - x\|.$$

This proves (LH). Since $A \neq 0$, there is j with $a_j \neq 0$. For $t \geq |b_j|/\|a_j\|^2$,

$$\|\nabla^2 f(ta_j)\|\|a_j\|^2 \geq \left\langle \nabla^2 f(ta_j)a_j, a_j\right\rangle \geq (p-1)\left(t\|a_j\|^2 - |b_j|\right)^{p-2}\|a_j\|^4.$$

If (GH) held, then the choice $x = 0$ and $y = ta_j$ would give

$$\|\nabla^2 f(ta_j)\| \leq \|\nabla^2 f(0)\| + \left(M_0 + M_1\|\nabla f(0)\|\right)t\|a_j\|, \qquad t > 0.$$

Since $p - 2 > 1$, this contradicts the previous display as $t \to +\infty$. □

Example B.2. For $\mu \geq 0$, consider $f(x) = \frac{\mu}{2}\|x\|^2 + \frac{1}{4}\|x\|^4$ for $x \in \mathbb{R}^d$. The function is convex for $\mu = 0$ and μ-strongly convex for $\mu > 0$. For every $\rho > 0$, it satisfies (LH) with $M_0 = 3\rho + 4$ and $M_1 = 2$, but it does not satisfy (GH).

Proof. We have

$$\nabla f(x) = \left(\mu + \|x\|^2\right)x, \qquad \nabla^2 f(x) = \left(\mu + \|x\|^2\right)I + 2xx^\top.$$

Hence, $\nabla^2 f(x) \succeq \mu I$. Let $s = y - x$ and $r = \|s\| \leq \rho$. Using

$$yy^\top - xx^\top = xs^\top + sx^\top + ss^\top,$$

we obtain

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq \left|\|y\|^2 - \|x\|^2\right| + 2\|yy^\top - xx^\top\|$$
$$\leq (2\|x\| + r)\,r + 2\left(2\|x\|r + r^2\right) = (6\|x\| + 3r)\,r.$$

For every $t \geq 0$,

$$4 + 2t^3 - 6t = 2(t-1)^2(t+2) \geq 0.$$

Since $\|\nabla f(x)\| = \left(\mu + \|x\|^2\right)\|x\| \geq \|x\|^3$, we have

$$6\|x\| \leq 4 + 2\|\nabla f(x)\|.$$

Consequently,

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq (3\rho + 4 + 2\|\nabla f(x)\|)\,\|y - x\|,$$

which proves (LH) with the stated constants. At $x = 0$,

$$\|\nabla^2 f(y) - \nabla^2 f(0)\| = 3\|y\|^2, \qquad \nabla f(0) = 0.$$

Thus, (GH) would require $3\|y\|^2 \leq M_0\|y\|$ for all $y \in \mathbb{R}^d$, which is impossible. □

Example B.3. The convex function $f(x) = e^x - x - 1$ satisfies (LH) for every $\rho > 0$ with $M_0 = M_1 = e^\rho$, but it does not satisfy (GH).

Proof. We have

$$f'(x) = e^x - 1, \qquad f''(x) = f^{(3)}(x) = e^x.$$

Let $h = y - x$ and $|h| \leq \rho$. The mean value theorem gives

$$|f''(y) - f''(x)| = e^x|e^h - 1| \leq e^\rho e^x|h| \leq e^\rho\,(1 + |f'(x)|)\,|y - x|.$$

Hence, (LH) holds with $M_0 = M_1 = e^\rho$. If (GH) held, then the choice $x = 0$ would give

$$e^y - 1 \leq M_0|y|, \qquad y \in \mathbb{R},$$

which is impossible as $y \to +\infty$. □

Pointwise consequence of the pairwise conditions. For any $x \in \mathbb{R}^d$, $\|u\| = 1$, and all sufficiently small $|t|$, either pairwise condition gives

$$\frac{\|\nabla^2 f(x + tu) - \nabla^2 f(x)\|}{|t|} \leq M_0 + M_1\|\nabla f(x)\|.$$

Taking $t \to 0$ and then the supremum over $\|u\| = 1$ proves (TD). □

The pointwise first-order condition used in Proposition 2.4 is denoted by

$$\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|, \qquad x \in \mathbb{R}^d. \tag{FO}$$

Proof of Proposition 2.4. Let $s = y - x$, $r = \|s\| \leq \rho$, and $g(t) = \|\nabla f(x + ts)\|$. The function g is absolutely continuous. For almost every $t \in [0, 1]$, condition (FO) gives

$$g'(t) \leq r\|\nabla^2 f(x + ts)\| \leq r\,(L_0 + L_1 g(t))\,.$$

Grönwall's inequality yields

$$g(t) \leq e^{L_1 rt}\,(g(0) + L_0 rt) \leq e^{L_1\rho}\,(\|\nabla f(x)\| + L_0\rho)\,.$$

Using (TD) along the segment from x to y, we obtain

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq r\int_0^1 \left(\widehat{M}_0 + \widehat{M}_1 g(t)\right)\mathrm{d}t \leq \left(\widehat{M}_0 + \widehat{M}_1 L_0\rho e^{L_1\rho} + \widehat{M}_1 e^{L_1\rho}\|\nabla f(x)\|\right) r.$$

This is (LH) with the stated constants. □

C Proof of Algorithm 2

We consider completed iterations before a zero-gradient return.

For $g \geq 0$ and $k \geq 0$, recall

$$\begin{aligned} \Theta_{\mathrm{c}}(g) &:= \max\left\{8\delta + \sqrt{48M_0 g}, \frac{4g}{\rho}, 12\sqrt{M_1 g}\right\}, \\ U_k &:= \max\left\{\tau_{k-1}, \Theta_{\mathrm{c}}(\|\nabla f(x_k)\|)\right\}, \qquad \tau_{-1} := 2\eta_0. \end{aligned} \tag{12}$$

C.1 STEP ESTIMATES

Lemma C.1. *Suppose that f is convex. For every accepted step of Algorithm 2, set $s_k := x_{k+1} - x_k$. Then*

$$\frac{\tau_k}{2}\|s_k\| \le \|\nabla f(x_{k+1})\| \le \frac{3\tau_k}{2}\|s_k\|, \tag{13}$$

$$f(x_k) - f(x_{k+1}) \ge \max\left\{\frac{\tau_k}{2}\|s_k\|^2, \frac{2\|\nabla f(x_{k+1})\|^2}{3\tau_k}\right\}, \tag{14}$$

$$\|\nabla f(x_{k+1})\| \le \frac{2}{\sqrt{3}}\|\nabla f(x_k)\|. \tag{15}$$

Proof. By the MS condition in (2)

$$\|\nabla f(x_{k+1})\| \ge \tau_k\|s_k\| - \|\nabla f(x_{k+1}) + \tau_k s_k\| \ge \frac{\tau_k}{2}\|s_k\|,$$
$$\|\nabla f(x_{k+1})\| \le \|\nabla f(x_{k+1}) + \tau_k s_k\| + \tau_k\|s_k\| \le \frac{3\tau_k}{2}\|s_k\|.$$

we prove (13). Squaring the MS condition and expanding the left-hand side gives

$$\|\nabla f(x_{k+1}) + \tau_k s_k\|^2 = \|\nabla f(x_{k+1})\|^2 + 2\tau_k\langle\nabla f(x_{k+1}), s_k\rangle + \tau_k^2\|s_k\|^2 \le \frac{\tau_k^2}{4}\|s_k\|^2. \tag{16}$$

Next, by convexity, (16), and (13),

$$f(x_k) - f(x_{k+1}) \ge -\langle\nabla f(x_{k+1}), s_k\rangle \ge \frac{\|\nabla f(x_{k+1})\|^2}{2\tau_k} + \frac{3\tau_k}{8}\|s_k\|^2$$
$$\ge \frac{1}{2\tau_k}\left(\frac{\tau_k}{2}\|s_k\|\right)^2 + \frac{3\tau_k}{8}\|s_k\|^2 = \frac{\tau_k}{2}\|s_k\|^2.$$

Using instead $\|s_k\| \ge 2\|\nabla f(x_{k+1})\|/(3\tau_k)$ from (13), we obtain

$$f(x_k) - f(x_{k+1}) \ge \frac{\|\nabla f(x_{k+1})\|^2}{2\tau_k} + \frac{3\tau_k}{8}\left(\frac{2\|\nabla f(x_{k+1})\|}{3\tau_k}\right)^2 = \frac{2\|\nabla f(x_{k+1})\|^2}{3\tau_k}.$$

This proves (14). For (15), convexity at both endpoints gives

$$\langle\nabla f(x_k), s_k\rangle \le f(x_{k+1}) - f(x_k) \le \langle\nabla f(x_{k+1}), s_k\rangle.$$

Next, (16) gives

$$\|\nabla f(x_{k+1})\|^2 + \frac{3}{4}\tau_k^2\|s_k\|^2 \le -2\tau_k\langle\nabla f(x_{k+1}), s_k\rangle \le -2\tau_k\langle\nabla f(x_k), s_k\rangle \le 2\tau_k\|\nabla f(x_k)\|\|s_k\|.$$

Rearranging and completing the square gives

$$\|\nabla f(x_{k+1})\|^2 \le 2\tau_k\|\nabla f(x_k)\|\|s_k\| - \frac{3}{4}\tau_k^2\|s_k\|^2$$
$$= \frac{4}{3}\|\nabla f(x_k)\|^2 - \frac{3}{4}\left(\tau_k\|s_k\| - \frac{4}{3}\|\nabla f(x_k)\|\right)^2 \le \frac{4}{3}\|\nabla f(x_k)\|^2.$$

□

Lemma C.2. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Set*

$$\overline{\tau}(x) := \begin{cases} \max\left\{4\delta, 2\sqrt{L(x)\|\nabla f(x)\|}\right\}, & \text{under (GH)}, \\ \max\left\{4\delta, 2\sqrt{L(x)\|\nabla f(x)\|}, \frac{2\|\nabla f(x)\|}{\rho}\right\}, & \text{under (LH)}. \end{cases} \tag{17}$$

Every trial of MSBACKTRACK *with $\tau \ge \overline{\tau}(x)$ satisfies both tests in* (2)*. Consequently, initialized with $\tau = \eta > 0$, the algorithm terminates after at most*

$$1 + \left\lceil \log_2 \max\left\{1, \frac{\overline{\tau}(x)}{\eta}\right\}\right\rceil$$

trials, and its accepted parameter satisfies

$$\eta \le \tau \le \max\{\eta, 2\overline{\tau}(x)\}.$$

Proof. Fix a trial parameter $\tau \geq \overline{\tau}(x)$. By convexity and (IH), $H(x) \succeq -\delta I$. Since $\tau \geq 4\delta$,

$$H(x) + \tau I \succeq (\tau - \delta) I \succeq \frac{3\tau}{4} I \succ 0.$$

Thus $\mathsf{PD}(x, \tau)$ holds, and the linear system $(H(x) + \tau I)s = -\nabla f(x)$ has a unique solution satisfying

$$\|s\| = \|(H(x) + \tau I)^{-1} \nabla f(x)\| \leq \frac{\|\nabla f(x)\|}{\tau - \delta} \leq \frac{2\|\nabla f(x)\|}{\tau}.$$

Under (LH), the definition of $\overline{\tau}(x)$ also gives $\|s\| \leq \rho$. The gradient remainder bound in Lemma A.1 or Lemma A.2, together with (IH) and the linear system, yields

$$\begin{aligned} \|\nabla f(x+s) + \tau s\| &= \|\nabla f(x+s) - \nabla f(x) - H(x)s\| \\ &\leq \frac{L(x)}{2}\|s\|^2 + \delta\|s\| \leq \left(\frac{L(x)\|\nabla f(x)\|}{\tau} + \delta\right)\|s\| \\ &\leq \left(\frac{\tau}{4} + \frac{\tau}{4}\right)\|s\| = \frac{\tau}{2}\|s\|. \end{aligned}$$

Hence $\mathsf{MS}(x, \tau, s)$ also holds. The trial parameters are $2^j\eta$, $j = 0, 1, \ldots$, so the first index with $2^j\eta \geq \overline{\tau}(x)$ gives the stated trial bound. If no trial is rejected, the accepted parameter is $\tau = \eta$. Otherwise, $\tau/2$ was rejected, so $\tau/2 < \overline{\tau}(x)$ and $\tau < 2\overline{\tau}(x)$. □

Lemma C.3. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Set*

$$r_0 := \begin{cases} \frac{1}{6\sqrt{M_1}}, & \textit{under } \text{(GH)}, \\ \min\left\{\frac{\rho}{2}, \frac{1}{6\sqrt{M_1}}\right\}, & \textit{under } \text{(LH)}, \end{cases} \tag{18}$$

where $1/\sqrt{M_1} = +\infty$ when $M_1 = 0$. Suppose that MSBACKTRACK*, initialized at x with $\tau = \eta > 0$, returns (y, τ). Set $s := y - x$. If at least one trial is rejected (equivalently $\tau > \eta$), then*

$$\|s\| > r_0 \quad \textit{or} \quad \tau \leq 8\delta + 24M_0\|s\|. \tag{19}$$

If, in addition, $\tau \geq 16\delta$, then

$$\|s\| \geq \min\left\{r_0, \frac{\tau}{48M_0}\right\}. \tag{20}$$

Proof. We prove (19) by considering three cases.

First, if $\tau < 4\delta$, then $\tau \leq 8\delta + 24M_0\|s\|$, so the second inequality holds.

Second, if $\tau \geq 4\delta$ and $\|s\| > r_0$, the first inequality holds.

It remains to consider the case $\tau \geq 4\delta$ and $\|s\| \leq r_0$. Since $\tau > \eta$, the trial with parameter $\tau/2$ was rejected. Convexity and (IH) give

$$H(x) + \frac{\tau}{2} I \succeq \left(\frac{\tau}{2} - \delta\right) I \succeq \frac{\tau}{4} I \succ 0, \tag{21}$$

so this trial failed the MS test. Its step $s_{1/2}$ and the accepted step s satisfy

$$\begin{aligned} \left(H(x) + \frac{\tau}{2} I\right) s_{1/2} &= -\nabla f(x), \\ (H(x) + \tau I)\, s &= -\nabla f(x). \end{aligned}$$

Subtracting these equations gives

$$s_{1/2} - s = \frac{\tau}{2}\left(H(x) + \frac{\tau}{2} I\right)^{-1} s, \qquad \|s_{1/2} - s\| \leq 2\|s\|, \tag{22}$$

where the last inequality follows from (21). For every $t \in [0, 1]$, (22) gives

$$\begin{aligned} \|x + t s_{1/2} - y\| &= \|t(s_{1/2} - s) - (1-t)s\| \\ &\leq t\|s_{1/2} - s\| + (1-t)\|s\| \leq (1+t)\|s\| \leq 2\|s\|. \end{aligned}$$

Under (LH), (18) gives

$$\|x-y\| = \|s\| \le r_0 \le \frac{\rho}{2}, \qquad \|x + ts_{1/2} - y\| \le 2\|s\| \le 2r_0 \le \rho.$$

Thus, applying (GH) or (LH) at y gives

$$\begin{aligned}\|\nabla^2 f(x+ts_{1/2}) - \nabla^2 f(x)\| &\le \|\nabla^2 f(x+ts_{1/2}) - \nabla^2 f(y)\| + \|\nabla^2 f(x) - \nabla^2 f(y)\| \\ &\le L(y)\left(\|x+ts_{1/2}-y\| + \|x-y\|\right) \le 3L(y)\|s\|. \end{aligned} \tag{23}$$

Since the trial with parameter $\tau/2$ fails the MS test, we have

$$\frac{\tau}{4}\|s_{1/2}\| < \|\nabla f(x+s_{1/2}) + \frac{\tau}{2}s_{1/2}\|. \tag{24}$$

The linear system for $s_{1/2}$ gives

$$\frac{\tau}{2}s_{1/2} = -\nabla f(x) - H(x)s_{1/2}.$$

Therefore,

$$\begin{aligned}\nabla f(x+s_{1/2}) + \frac{\tau}{2}s_{1/2} &= \nabla f(x+s_{1/2}) - \nabla f(x) - H(x)s_{1/2} \\ &= \int_0^1 \left(\nabla^2 f(x+ts_{1/2}) - H(x)\right) s_{1/2}\,dt. \end{aligned} \tag{25}$$

Therefore, by (24), (25), (23), and (IH),

$$\frac{\tau}{4}\|s_{1/2}\| < \|\int_0^1 \left(\nabla^2 f(x+ts_{1/2}) - H(x)\right) s_{1/2}\,dt\| \le \left(\delta + 3L(y)\|s\|\right)\|s_{1/2}\|.$$

Since this trial fails MS, its step $s_{1/2}$ is nonzero. The accepted MS test in (2) gives

$$\|\nabla f(y)\| \le \|\nabla f(y) + \tau s\| + \tau\|s\| \le \frac{3\tau}{2}\|s\|.$$

Dividing by $\|s_{1/2}\|$ and using this bound, we obtain

$$\begin{aligned}\frac{\tau}{4} &< \delta + 3L(y)\|s\| = \delta + 3M_0\|s\| + 3M_1\|\nabla f(y)\|\|s\| \\ &\le \delta + 3M_0\|s\| + \frac{9}{2}M_1\tau\|s\|^2 \le \delta + 3M_0\|s\| + \frac{\tau}{8}. \end{aligned}$$

The last inequality follows from $\|s\| \le r_0$ and (18). Indeed, if $M_1 > 0$, then

$$M_1\|s\|^2 \le M_1 r_0^2 \le M_1\left(\frac{1}{6\sqrt{M_1}}\right)^2 = \frac{1}{36}.$$

If $M_1 = 0$, then $M_1\|s\|^2 = 0$. Thus, in both cases,

$$\frac{9}{2}M_1\tau\|s\|^2 \le \frac{\tau}{8}.$$

Rearranging proves (19). Finally, if $\tau \ge 16\delta$, (19) gives

$$\|s\| \ge \min\left\{r_0, \frac{\tau - 8\delta}{24M_0}\right\} \ge \min\left\{r_0, \frac{\tau}{48M_0}\right\},$$

which proves (20). □

By (12) and (18),

$$\Theta_{\mathrm{c}}(g) = \max\left\{8\delta + \sqrt{48M_0 g}, \frac{2g}{r_0}\right\}, \qquad g \ge 0,$$

where $1/(+\infty) = 0$.

Lemma C.4. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. If iteration k of Algorithm 2 rejects at least one trial, then*

$$\tau_k \le \Theta_c(\|\nabla f(x_{k+1})\|). \tag{26}$$

Proof. Set $s_k := x_{k+1} - x_k$ and apply Lemma C.3 with $x = x_k$, $y = x_{k+1}$, $\eta = \eta_k$, and $\tau = \tau_k$. If $\|s_k\| > r_0$, then (13) and (12) give

$$\tau_k < \frac{2\|\nabla f(x_{k+1})\|}{r_0} \le \Theta_c(\|\nabla f(x_{k+1})\|).$$

If $\|s_k\| \le r_0$, (19) and (13) imply

$$\tau_k \le 8\delta + 24M_0\|s_k\| \le 8\delta + \frac{48M_0\|\nabla f(x_{k+1})\|}{\tau_k},$$
$$\tau_k \le 4\delta + \sqrt{16\delta^2 + 48M_0\|\nabla f(x_{k+1})\|} \le 8\delta + \sqrt{48M_0\|\nabla f(x_{k+1})\|},$$

which proves (26). □

C.2 Iteration Count

We call a completed iteration k of Algorithm 2 controlled if

$$\tau_k \le \Theta_c(\|\nabla f(x_{k+1})\|). \tag{27}$$

By Lemma C.4, every iteration with a rejection is controlled. An iteration accepted at its first trial can belong to either type.

Lemma C.5. *Under the assumptions of Lemma C.4, suppose that m iterations of Algorithm 2 are completed and $f(x_m) > f^*$. If m_c of these iterations are controlled, then*

$$m \le \frac{5}{4}m_c + \log_2 \frac{U_0}{\Theta_c(\|\nabla f(x_m)\|)}. \tag{28}$$

Proof. By (12), $\Theta_c(\|\nabla f(x_k)\|) \le U_k$ for every $k \ge 0$. The function Θ_c is nondecreasing, and (12) gives $\Theta_c(ag) \le a\Theta_c(g)$ for $g \ge 0$ and $a \ge 1$.

If iteration k is controlled, then $\tau_k \le \Theta_c(\|\nabla f(x_{k+1})\|)$ by (27). The definition of U_{k+1} and (15) give

$$U_{k+1} = \Theta_c(\|\nabla f(x_{k+1})\|) \le \frac{2}{\sqrt{3}}\Theta_c(\|\nabla f(x_k)\|) \le \frac{2}{\sqrt{3}}U_k.$$

If iteration k is not controlled, then $\tau_k > \Theta_c(\|\nabla f(x_{k+1})\|)$, so $U_{k+1} = \tau_k$ by definition. By Lemma C.4, this iteration has no rejected trials, so $\tau_k = \eta_k$. For $k \ge 1$, line 4 of Algorithm 2, applied at iteration $k-1$, gives $\eta_k \le \tau_{k-1}/2$. Hence,

$$U_{k+1} = \tau_k = \eta_k \le \frac{\tau_{k-1}}{2} \le \frac{U_k}{2}.$$

For $k = 0$, the same bound holds because $U_1 = \tau_0 = \eta_0 \le U_0/2$. Applying these bounds over the m_c controlled iterations and $m - m_c$ remaining iterations gives

$$\Theta_c(\|\nabla f(x_m)\|) \le U_m \le U_0 \left(\frac{2}{\sqrt{3}}\right)^{m_c} 2^{-(m-m_c)}.$$

Taking logarithms and rearranging, we obtain

$$m \le \left(1 + \log_2 \frac{2}{\sqrt{3}}\right) m_c + \log_2 \frac{U_0}{\Theta_c(\|\nabla f(x_m)\|)}.$$

Since $(2/\sqrt{3})^4 = 16/9 < 2$, this proves (28). □

To count the controlled iterations, assume $D < +\infty$. By (14), every accepted point remains in $\mathcal{L}(x_0)$, which also contains x^*. Hence, by convexity,

$$\begin{aligned} \|x_k - x^*\| &\le \|x_k - x_0\| + \|x_0 - x^*\| \le 2D, \\ f(x_k) - f^* &\le \langle \nabla f(x_k), x_k - x^* \rangle \le 2D\|\nabla f(x_k)\|. \end{aligned} \tag{29}$$

By (3) and (18), $B = \max\{1, D/r_0\}$, where $D/(+\infty) = 0$. For $t > 0$, define

$$\Phi(t) := \max\left\{ \frac{6D}{r_0}, 12\sqrt{6}\sqrt{\frac{M_0 D^3}{t}} + \frac{48\delta D^2}{t} \right\}. \tag{30}$$

The function Φ is positive and nonincreasing, and $t(1+\Phi(t))$ is nondecreasing.

Lemma C.6. *For the prefix and count of Lemma C.5, assume in addition that $0 < D < +\infty$. Then*

$$m_{\mathrm{c}} \le \int_{f(x_m)-f^*}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t}\, dt. \tag{31}$$

Proof. For a controlled iteration, we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\overset{(14)}{\ge} \frac{2\|\nabla f(x_{k+1})\|^2}{3\tau_k} \overset{(27)}{\ge} \frac{2\|\nabla f(x_{k+1})\|^2}{3\Theta_{\mathrm{c}}(\|\nabla f(x_{k+1})\|)} \\ &\overset{(12)}{=} \min\left\{ \frac{2\|\nabla f(x_{k+1})\|^2}{3\left(8\delta + \sqrt{48 M_0 \|\nabla f(x_{k+1})\|}\right)}, \frac{r_0\|\nabla f(x_{k+1})\|}{3} \right\}. \end{aligned}$$

By convexity and Cauchy-Schwarz,

$$f(x_{k+1}) - f^* \le \langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle \le \|\nabla f(x_{k+1})\| \|x_{k+1} - x^*\| \overset{(29)}{\le} 2D\|\nabla f(x_{k+1})\|.$$

Substituting $\|\nabla f(x_{k+1})\| \ge \frac{f(x_{k+1})-f^*}{2D}$ gives

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\ge \min\left\{ \frac{(f(x_{k+1})-f^*)^2}{48\delta D^2 + 12\sqrt{6}\sqrt{M_0 D^3 (f(x_{k+1})-f^*)}}, \frac{r_0(f(x_{k+1})-f^*)}{6D} \right\} \\ &= \frac{f(x_{k+1})-f^*}{\Phi(f(x_{k+1})-f^*)}. \end{aligned}$$

By (14), $f(x_{k+1}) - f^* \le f(x_k) - f^*$. For every $t \in [f(x_{k+1})-f^*, f(x_k)-f^*]$, the monotonicity of $t(1+\Phi(t))$ gives

$$\frac{1+\Phi(t)}{t} \ge \frac{(f(x_{k+1})-f^*)(1+\Phi(f(x_{k+1})-f^*))}{t^2}.$$

Thus, we obtain

$$\begin{aligned} \int_{f(x_{k+1})-f^*}^{f(x_k)-f^*} \frac{1+\Phi(t)}{t}\, dt &\ge (f(x_{k+1})-f^*)(1+\Phi(f(x_{k+1})-f^*)) \int_{f(x_{k+1})-f^*}^{f(x_k)-f^*} \frac{dt}{t^2} \\ &= (f(x_{k+1})-f^*)(1+\Phi(f(x_{k+1})-f^*)) \left(\frac{1}{f(x_{k+1})-f^*} - \frac{1}{f(x_k)-f^*} \right) \\ &= (1+\Phi(f(x_{k+1})-f^*)) \frac{f(x_k)-f(x_{k+1})}{f(x_k)-f^*} \ge \frac{1+\Phi(f(x_{k+1})-f^*)}{\Phi(f(x_{k+1})-f^*)+1} = 1. \end{aligned}$$

Summing over the controlled iterations proves (31). □

C.3 Proof of Theorem 3.1

Theorem C.7. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Define*

$$B := \max\left\{1, \frac{2D}{\rho}, 6\sqrt{M_1 D}\right\}, \qquad \varepsilon_{\mathrm{tr}} := \max\left\{ \frac{M_0 D^3}{B^2}, \frac{\delta D^2}{B} \right\}, \tag{32}$$

where $\rho = +\infty$ under (GH)*. For every $\varepsilon > 0$, Algorithm 2 generates an iterate x_N satisfying $f(x_N) - f^* \le \varepsilon$ within*

$$N \le 2 + \log_2 \max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\mathrm{tr}}}, \frac{\eta_0 D^2}{\sqrt{6}B\varepsilon_{\mathrm{tr}}}\right\} + 100B \log_+ \frac{f(x_0) - f^*}{\max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}} + 80\sqrt{\frac{M_0 D^3}{\varepsilon}} + 64\frac{\delta D^2}{\varepsilon}. \tag{33}$$

iterations.

Proof. Each search terminates by Lemma C.2. Consider a prefix $x_0, \dots, x_m$ with $f(x_m) - f^* > \varepsilon$. By (29), $\|\nabla f(x_m)\| > \varepsilon/(2D)$. Combining (28) and (31) gives

$$\begin{aligned} m &\le \frac{5}{4}\int_{f(x_m)-f^*}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt + \log_2 \frac{U_0}{\Theta_{\mathrm{c}}(\|\nabla f(x_m)\|)} \\ &\le \frac{5}{4}\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt + \log_{2,+} \frac{U_0}{\Theta_{\mathrm{c}}(\varepsilon/(2D))}. \end{aligned} \tag{34}$$

Let N be the first index with $f(x_N) - f^* \le \varepsilon$. If $N \ge 1$, then $f(x_{N-1}) - f^* > \varepsilon$. Therefore,

$$N \le 1 + \frac{5}{4}\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt + \log_{2,+} \frac{U_0}{\Theta_{\mathrm{c}}(\varepsilon/(2D))}. \tag{35}$$

We begin with bounding integral term in (35).

By (32) and (30), $1 + \Phi(t) \le 79B$ for $t \ge \varepsilon_{\mathrm{tr}}$. For $0 < t \le \varepsilon_{\mathrm{tr}}$, we have

$$\max\left\{\sqrt{\frac{M_0 D^3}{t}}, \frac{\delta D^2}{t}\right\} \ge B \ge \max\left\{1, \frac{D}{r_0}\right\}, 1 + \Phi(t) \le (12\sqrt{6}+1)\sqrt{\frac{M_0 D^3}{t}} + 49\frac{\delta D^2}{t}.$$

If $\varepsilon_{\mathrm{tr}} \le \varepsilon < f(x_0) - f^*$, the whole integration interval lies above $\varepsilon_{\mathrm{tr}}$, so

$$\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt \le 79B \int_{\varepsilon}^{f(x_0)-f^*} \frac{dt}{t} = 79B \log \frac{f(x_0) - f^*}{\varepsilon}.$$

If $0 < \varepsilon < \varepsilon_{\mathrm{tr}} \le f(x_0) - f^*$, split the integral as

$$\begin{aligned} &\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt = \int_{\varepsilon}^{\varepsilon_{\mathrm{tr}}} \frac{1+\Phi(t)}{t} dt + \int_{\varepsilon_{\mathrm{tr}}}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt \\ &\le (12\sqrt{6}+1)\sqrt{M_0 D^3}\int_{\varepsilon}^{\varepsilon_{\mathrm{tr}}} \frac{dt}{t^{3/2}} + 49\delta D^2 \int_{\varepsilon}^{\varepsilon_{\mathrm{tr}}} \frac{dt}{t^2} + 79B \int_{\varepsilon_{\mathrm{tr}}}^{f(x_0)-f^*} \frac{dt}{t} \\ &= (24\sqrt{6}+2)\sqrt{M_0 D^3}\left(\frac{1}{\sqrt{\varepsilon}} - \frac{1}{\sqrt{\varepsilon_{\mathrm{tr}}}}\right) + 49\delta D^2\left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon_{\mathrm{tr}}}\right) + 79B \log \frac{f(x_0) - f^*}{\varepsilon_{\mathrm{tr}}} \\ &\le (24\sqrt{6}+2)\sqrt{\frac{M_0 D^3}{\varepsilon}} + 49\frac{\delta D^2}{\varepsilon} + 79B \log \frac{f(x_0) - f^*}{\varepsilon_{\mathrm{tr}}}. \end{aligned}$$

If $0 < \varepsilon < f(x_0) - f^* < \varepsilon_{\mathrm{tr}}$, the whole interval lies below $\varepsilon_{\mathrm{tr}}$, so

$$\begin{aligned} &\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t} dt \le (12\sqrt{6}+1)\sqrt{M_0 D^3}\int_{\varepsilon}^{f(x_0)-f^*} \frac{dt}{t^{3/2}} + 49\delta D^2 \int_{\varepsilon}^{f(x_0)-f^*} \frac{dt}{t^2} \\ &= (24\sqrt{6}+2)\sqrt{M_0 D^3}\left(\frac{1}{\sqrt{\varepsilon}} - \frac{1}{\sqrt{f(x_0) - f^*}}\right) + 49\delta D^2\left(\frac{1}{\varepsilon} - \frac{1}{f(x_0) - f^*}\right) \\ &\le (24\sqrt{6}+2)\sqrt{\frac{M_0 D^3}{\varepsilon}} + 49\frac{\delta D^2}{\varepsilon}. \end{aligned}$$

Here $\log_+ \frac{f(x_0)-f^*}{\varepsilon_{\mathrm{tr}}} = 0$ by definition.

Combining the three cases yields

$$\int_{\varepsilon}^{f(x_0)-f^*} \frac{1+\Phi(t)}{t}dt \le 79B\log_+ \frac{f(x_0)-f^*}{\max\{\varepsilon,\varepsilon_{\rm tr}\}} \quad + (24\sqrt{6}+2)\sqrt{\frac{M_0D^3}{\varepsilon}} + 49\frac{\delta D^2}{\varepsilon}. \tag{36}$$

Now, we need to bound the last term in (35).

For $g_2 \ge g_1 > 0$, (12) gives $\Theta_{\rm c}(g_2) \le (g_2/g_1)\Theta_{\rm c}(g_1)$. Together with monotonicity of $\Theta_{\rm c}$, this yields

$$\log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon/(2D))} \le \log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))} + \log_{2,+}\frac{\varepsilon_{\rm tr}}{\varepsilon}.$$

If $\varepsilon \ge \varepsilon_{\rm tr}$, then $\log_{2,+}(\varepsilon_{\rm tr}/\varepsilon) = 0$. For $0 < \varepsilon < \varepsilon_{\rm tr}$, we have

$$\log_2\frac{\varepsilon_{\rm tr}}{\varepsilon} \le 3\sqrt{\frac{\varepsilon_{\rm tr}}{\varepsilon}} = \frac{3}{B}\sqrt{\frac{M_0D^3}{\varepsilon}}, \qquad \text{if } \varepsilon_{\rm tr} = \frac{M_0D^3}{B^2},$$
$$\log_2\frac{\varepsilon_{\rm tr}}{\varepsilon} \le 2\frac{\varepsilon_{\rm tr}}{\varepsilon} = \frac{2}{B}\frac{\delta D^2}{\varepsilon}, \qquad \text{if } \varepsilon_{\rm tr} = \frac{\delta D^2}{B}.$$

Since $B \ge 1$,

$$\log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon/(2D))} \le \log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))} + 3\sqrt{\frac{M_0D^3}{\varepsilon}} + 2\frac{\delta D^2}{\varepsilon}. \tag{37}$$

Substituting (36) and (37) into (35), and using

$$\frac{5}{4}\cdot 79 \le 100, \qquad \frac{5}{4}(24\sqrt{6}+2)+3 \le 80, \qquad \frac{5}{4}\cdot 49 + 2 \le 64,$$

we obtain

$$N \le 2 + \log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))} + 100B\log_+\frac{f(x_0)-f^*}{\max\{\varepsilon,\varepsilon_{\rm tr}\}} + 80\sqrt{\frac{M_0D^3}{\varepsilon}} + 64\frac{\delta D^2}{\varepsilon}.$$

Recall that $\Theta_{\rm c}$ is nondecreasing and $\Theta_{\rm c}(ag) \le a\Theta_{\rm c}(g)$ for $a \ge 1$. Thus,

$$\Theta_{\rm c}(\|\nabla f(x_0)\|) \le \Theta_{\rm c}\left(\max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\rm tr}}\right\}\frac{\varepsilon_{\rm tr}}{2D}\right) \le \max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\rm tr}}\right\}\Theta_{\rm c}\left(\frac{\varepsilon_{\rm tr}}{2D}\right).$$

If $\varepsilon_{\rm tr} = M_0D^3/B^2$

$$\Theta_{\rm c}\left(\frac{\varepsilon_{\rm tr}}{2D}\right) \ge \sqrt{\frac{24M_0\varepsilon_{\rm tr}}{D}} = \frac{2\sqrt{6}B\varepsilon_{\rm tr}}{D^2}.$$

Otherwise $\varepsilon_{\rm tr} = \delta D^2/B$

$$\Theta_{\rm c}\left(\frac{\varepsilon_{\rm tr}}{2D}\right) \ge 8\delta = \frac{8B\varepsilon_{\rm tr}}{D^2} \ge \frac{2\sqrt{6}B\varepsilon_{\rm tr}}{D^2}.$$

Since $U_0 = \max\{2\eta_0, \Theta_{\rm c}(\|\nabla f(x_0)\|)\}$, these estimates give

$$\begin{aligned}\log_{2,+}\frac{U_0}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))} &= \log_2\max\left\{1, \frac{\Theta_{\rm c}(\|\nabla f(x_0)\|)}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))}, \frac{2\eta_0}{\Theta_{\rm c}(\varepsilon_{\rm tr}/(2D))}\right\} \\ &\le \log_2\max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\rm tr}}, \frac{\eta_0D^2}{\sqrt{6}B\varepsilon_{\rm tr}}\right\}.\end{aligned} \tag{38}$$

Substituting (38) into the bound for N proves (33).

□

C.4 Far and Near Convergence Rates

Corollary C.8. *Under the assumptions of Theorem C.7, suppose that $0 < D < +\infty$. Define*

$$\ell_0 := \max\left\{1, \frac{2D\|\nabla f(x_0)\|}{\varepsilon_{\mathrm{tr}}}, \frac{\eta_0 D^2}{\sqrt{6}B\varepsilon_{\mathrm{tr}}}\right\}.$$

For every $k \geq 0$ with $f(x_k) - f^* > \varepsilon_{\mathrm{tr}}$,

$$f(x_k) - f^* \leq (f(x_0) - f^*) \exp\left(-\frac{(k - \log_2 \ell_0)_+}{100B}\right), \tag{39}$$

where we write $(t)_+ := \max\{0, t\}$. If $f(x_k) - f^* \leq \varepsilon_{\mathrm{tr}}$, the subsequent iterates satisfy

$$f(x_{k+n}) - f^* \leq \frac{7396 M_0 D^3}{\left(B + (n - \log_2 \ell_0)_+\right)^2} + \frac{140\delta D^2}{B + (n - \log_2 \ell_0)_+}, \qquad n = 0, 1, \ldots, \tag{40}$$

where $t_+ := \max\{t, 0\}$.

Proof. If $f(x_k) - f^* > \varepsilon_{\mathrm{tr}}$, (34) gives

$$k \leq \frac{5}{4}\int_{f(x_k)-f^*}^{f(x_0)-f^*} \frac{1 + \Phi(t)}{t} dt + \log_2 \frac{U_0}{\Theta_{\mathrm{c}}(\|\nabla f(x_k)\|)}.$$

For $t \geq \varepsilon_{\mathrm{tr}}$, definitions (32) and (30) give

$$\Phi(t) \leq \max\left\{6B, 12\sqrt{6}\sqrt{\frac{M_0 D^3}{\varepsilon_{\mathrm{tr}}}} + 48\frac{\delta D^2}{\varepsilon_{\mathrm{tr}}}\right\} \leq \max\left\{6B, (12\sqrt{6} + 48)B\right\} = (12\sqrt{6} + 48)B.$$

Therefore, $1 + \Phi(t) \leq 79B$. By (29), monotonicity of Θ_{c}, and (38),

$$\log_2 \frac{U_0}{\Theta_{\mathrm{c}}(\|\nabla f(x_k)\|)} \leq \log_{2,+} \frac{U_0}{\Theta_{\mathrm{c}}(\varepsilon_{\mathrm{tr}}/(2D))} \leq \log_2 \ell_0.$$

Substituting these bounds yields

$$\begin{aligned} k &\leq \frac{5}{4} \cdot 79B \int_{f(x_k)-f^*}^{f(x_0)-f^*} \frac{dt}{t} + \log_2 \ell_0 \\ &= \frac{395}{4} B \log \frac{f(x_0) - f^*}{f(x_k) - f^*} + \log_2 \ell_0 \leq 100B \log \frac{f(x_0) - f^*}{f(x_k) - f^*} + \log_2 \ell_0. \end{aligned}$$

Rearranging and using $f(x_k) \leq f(x_0)$ when $k \leq \log_2 \ell_0$ proves (39).

Now suppose that $f(x_k) - f^* \leq \varepsilon_{\mathrm{tr}}$. Let $n_{\mathrm{tr}} \leq k$ be the first index with $f(x_{n_{\mathrm{tr}}}) - f^* \leq \varepsilon_{\mathrm{tr}}$. For $n \geq 0$ with $f(x_{n_{\mathrm{tr}}+n}) > f^*$, the first inequality in (34), applied from iteration n_{tr}, gives

$$n \leq \frac{5}{4}\int_{f(x_{n_{\mathrm{tr}}+n})-f^*}^{f(x_{n_{\mathrm{tr}}})-f^*} \frac{1 + \Phi(t)}{t} dt + \log_2 \frac{U_{n_{\mathrm{tr}}}}{\Theta_{\mathrm{c}}(\|\nabla f(x_{n_{\mathrm{tr}}+n})\|)}. \tag{41}$$

For $0 < t \leq \varepsilon_{\mathrm{tr}}$, the definitions of $\varepsilon_{\mathrm{tr}}$ and B give

$$\max\left\{\sqrt{\frac{M_0 D^3}{t}}, \frac{\delta D^2}{t}\right\} \geq B = \max\left\{1, \frac{D}{r_0}\right\}.$$

Hence, by the definition of Φ in (30),

$$1 + \Phi(t) \leq 1 + 12\sqrt{6}\sqrt{\frac{M_0 D^3}{t}} + 48\frac{\delta D^2}{t} \leq (12\sqrt{6} + 1)\sqrt{\frac{M_0 D^3}{t}} + 49\frac{\delta D^2}{t}.$$

Since $0 < f(x_{n_{\mathrm{tr}}+n}) - f^* \le f(x_{n_{\mathrm{tr}}}) - f^* \le \varepsilon_{\mathrm{tr}}$,

$$\begin{aligned}\int_{f(x_{n_{\mathrm{tr}}+n})-f^*}^{f(x_{n_{\mathrm{tr}}})-f^*} \frac{1+\Phi(t)}{t}dt &\le (12\sqrt{6}+1)\sqrt{M_0D^3}\int_{f(x_{n_{\mathrm{tr}}+n})-f^*}^{f(x_{n_{\mathrm{tr}}})-f^*}\frac{dt}{t^{3/2}} + 49\delta D^2\int_{f(x_{n_{\mathrm{tr}}+n})-f^*}^{f(x_{n_{\mathrm{tr}}})-f^*}\frac{dt}{t^2}\\ &= (24\sqrt{6}+2)\sqrt{M_0D^3}\left(\frac{1}{\sqrt{f(x_{n_{\mathrm{tr}}+n})-f^*}} - \frac{1}{\sqrt{f(x_{n_{\mathrm{tr}}})-f^*}}\right)\\ &\quad +49\delta D^2\left(\frac{1}{f(x_{n_{\mathrm{tr}}+n})-f^*} - \frac{1}{f(x_{n_{\mathrm{tr}}})-f^*}\right)\\ &\le (24\sqrt{6}+2)\sqrt{\frac{M_0D^3}{f(x_{n_{\mathrm{tr}}+n})-f^*}} + 49\frac{\delta D^2}{f(x_{n_{\mathrm{tr}}+n})-f^*}.\end{aligned}$$

To bound the logarithm in (41), first express $U_{n_{\mathrm{tr}}}$ through Θ_{c}. If there is no controlled iteration before n_{tr}, the proof of Lemma C.5 gives $U_{n_{\mathrm{tr}}} \le U_0$. Otherwise, let $i < n_{\mathrm{tr}}$ be the last controlled iteration. Then

$$U_{n_{\mathrm{tr}}} \le U_{i+1} = \Theta_{\mathrm{c}}(\|\nabla f(x_{i+1})\|).$$

Next, by $f(x_0) - f^* > \varepsilon_{\mathrm{tr}} = \max\left\{M_0D^3/B^2, \delta D^2/B\right\}$

$$\delta \le \frac{B(f(x_0)-f^*)}{D^2}, \qquad M_0 \le \frac{B^2(f(x_0)-f^*)}{D^3}.$$

By the definition of B,

$$\max\left\{\frac{4}{\rho}, 12\sqrt{M_1}\right\} = \frac{2}{D}\max\left\{\frac{2D}{\rho}, 6\sqrt{M_1}D\right\} \le \frac{2B}{D}.$$

If $\|\nabla f(x_{i+1})\| < (f(x_0)-f^*)/D$, then $B \ge 1$ gives $\|\nabla f(x_{i+1})\| < B\frac{f(x_0)-f^*}{D}$.

Otherwise, since iteration i is controlled, by (14), (27), and the definition of Θ_{c}

$$\begin{aligned}\|\nabla f(x_{i+1})\| &\le \frac{3}{2}(f(x_0)-f^*)\max\left\{\frac{8\delta}{\|\nabla f(x_{i+1})\|} + \sqrt{\frac{48M_0}{\|\nabla f(x_{i+1})\|}}, \frac{4}{\rho}, 12\sqrt{M_1}\right\}\\ &\le \frac{3(8+\sqrt{48})B(f(x_0)-f^*)}{2D} \le \frac{23B(f(x_0)-f^*)}{D}.\end{aligned}$$

Thus,

$$\|\nabla f(x_{i+1})\| \le 23B\frac{f(x_0)-f^*}{D} \le 23B\|\nabla f(x_0)\|.$$

Substituting into Θ_{c} and using its monotonicity and $\Theta_{\mathrm{c}}(ag) \le a\Theta_{\mathrm{c}}(g)$ for $a \ge 1$, we obtain

$$U_{n_{\mathrm{tr}}} \le \Theta_{\mathrm{c}}(\|\nabla f(x_{i+1})\|) \le \Theta_{\mathrm{c}}(23B\|\nabla f(x_0)\|) \le 23B\Theta_{\mathrm{c}}(\|\nabla f(x_0)\|) \le 23BU_0.$$

Using this bound, $\|\nabla f(x_{n_{\mathrm{tr}}+n})\| \ge (f(x_{n_{\mathrm{tr}}+n})-f^*)/(2D)$, and (37) with (38), we obtain

$$\begin{aligned}\log_2\frac{U_{n_{\mathrm{tr}}}}{\Theta_{\mathrm{c}}(\|\nabla f(x_{n_{\mathrm{tr}}+n})\|)} &\le \log_2(23B) + \log_{2,+}\frac{U_0}{\Theta_{\mathrm{c}}((f(x_{n_{\mathrm{tr}}+n})-f^*)/(2D))}\\ &\le \log_2\ell_0 + \log_2(23B) + 3\sqrt{\frac{M_0D^3}{f(x_{n_{\mathrm{tr}}+n})-f^*}} + 2\frac{\delta D^2}{f(x_{n_{\mathrm{tr}}+n})-f^*}.\end{aligned}$$

Substitution into (41) yields

$$n \le \log_2\ell_0 + \log_2(23B) + 80\sqrt{\frac{M_0D^3}{f(x_{n_{\mathrm{tr}}+n})-f^*}} + 64\frac{\delta D^2}{f(x_{n_{\mathrm{tr}}+n})-f^*}. \tag{42}$$

The definition of $\varepsilon_{\mathrm{tr}}$ and $f(x_{n_{\mathrm{tr}}+n}) - f^* \le \varepsilon_{\mathrm{tr}}$ imply

$$B \le \max\left\{\sqrt{\frac{M_0D^3}{f(x_{n_{\mathrm{tr}}+n})-f^*}}, \frac{\delta D^2}{f(x_{n_{\mathrm{tr}}+n})-f^*}\right\}, \qquad \log_2(23B) \le 5B.$$

Inequality (42) therefore gives

$$B + (n - \log_2 \ell_0)_+ \le 6B + 80\sqrt{\frac{M_0 D^3}{f(x_{n_{\rm tr}+n}) - f^*}} + 64\frac{\delta D^2}{f(x_{n_{\rm tr}+n}) - f^*}$$
$$\le 86\sqrt{\frac{M_0 D^3}{f(x_{n_{\rm tr}+n}) - f^*}} + 70\frac{\delta D^2}{f(x_{n_{\rm tr}+n}) - f^*}.$$

Multiplying by $f(x_{n_{\rm tr}+n}) - f^*$ and using $2ab \le a^2 + b^2$, we obtain

$$(B + (n - \log_2 \ell_0)_+)\,(f(x_{n_{\rm tr}+n}) - f^*) \le 86\sqrt{M_0 D^3\,(f(x_{n_{\rm tr}+n}) - f^*)} + 70\delta D^2$$
$$\le \frac{(B + (n - \log_2 \ell_0)_+)\,(f(x_{n_{\rm tr}+n}) - f^*)}{2} + \frac{3698 M_0 D^3}{B + (n - \log_2 \ell_0)_+} + 70\delta D^2,$$
$$f(x_{n_{\rm tr}+n}) - f^* \le \frac{7396 M_0 D^3}{(B + (n - \log_2 \ell_0)_+)^2} + \frac{140\delta D^2}{B + (n - \log_2 \ell_0)_+}.$$

Since $f(x_{k+n}) \le f(x_{n_{\rm tr}+n})$, this proves (40). □

C.5 Oracle complexity

Corollary C.9. *Let $N := \min\{k \ge 0 : f(x_k) - f^* \le \varepsilon\}$ and $g_0 = \|\nabla f(x_0)\|$. The total number of rejected trials of Algorithm 2 before obtaining x_N, including failed* PD *tests, is at most*

$$\frac{5}{4}(N-1) + 4 + \log_{2,+}\frac{Bg_0}{D\eta_0} + \log_{2,+}\frac{\varepsilon_{\rm tr}}{\varepsilon}. \tag{43}$$

Thus, the total number of gradient evaluations and rejected trials is bounded by

$$\mathcal{O}\left(1 + \log \ell_0 + \log_+\frac{Bg_0}{D\eta_0} + B\log_+\frac{f(x_0) - f^*}{\max\{\varepsilon, \varepsilon_{\rm tr}\}} + \sqrt{\frac{M_0 D^3}{\varepsilon}} + \frac{\delta D^2}{\varepsilon}\right). \tag{44}$$

Proof. Let b_j be the number of rejected trials on iteration j. Then, $b_0 = \log_2 \frac{\tau_0}{\eta_0}$ and $\tau_j = 2^{b_j}\eta_j$. Let $g_j = \|\nabla f(x_j)\|$ Therefore, for $j \ge 1$ by line 4 of Algorithm 2

$$b_j = \log_2\frac{\tau_j}{\eta_j} = \log_2\frac{2\tau_j}{\tau_{j-1}\min\{1, g_j/g_{j-1}\}} = 1 + \log_2\frac{\tau_j}{\tau_{j-1}} + \log_2\max\left\{1, \frac{g_{j-1}}{g_j}\right\}.$$

Next, let us sum the number of rejected trials to obtain x_N

$$\sum_{j=0}^{N-1} b_j = \log_2\frac{\tau_0}{\eta_0} + (N-1) + \log_2\frac{\tau_{N-1}}{\tau_0} + \sum_{j=1}^{N-1}\log_{2,+}\frac{g_{j-1}}{g_j}.$$

Let us bound the last sum

$$\sum_{j=1}^{N-1}\log_{2,+}\frac{g_{j-1}}{g_j} = \sum_{j=1}^{N-1}\left(\log_{2,+}\frac{g_j}{g_{j-1}} + \log_2\frac{g_{j-1}}{g_j}\right) = \sum_{j=1}^{N-1}\log_{2,+}\frac{g_j}{g_{j-1}} + \log_2\frac{g_0}{g_{N-1}}$$

Therefore,

$$\sum_{j=0}^{N-1} b_j = N - 1 + \log_2\frac{g_0\tau_{N-1}}{\eta_0 g_{N-1}} + \sum_{j=1}^{N-1}\log_{2,+}\frac{g_j}{g_{j-1}}. \tag{45}$$

By (15), we have $g_j \le (2/\sqrt{3})g_{j-1}$. Therefore,

$$\sum_{j=1}^{N-1}\log_{2,+}\frac{g_j}{g_{j-1}} \le (N-1)\log_{2,+}\frac{2}{\sqrt{3}} \le \frac{N-1}{4}. \tag{46}$$

Now, it remains to bound $\log_{2,+}(\tau_{N-1}/g_{N-1})$. By Lemma C.2, every accepted τ satisfies

$$\tau_j \le \max\left\{\eta_j, 8\delta, 4\sqrt{L(x_j)g_j}, \frac{4g_j}{\rho}\right\}.$$

Therefore,

$$\frac{\tau_j}{g_j} \le \max\left\{\frac{\eta_j}{g_j}, \frac{8\delta}{g_j}, 4\sqrt{\frac{M_0}{g_j} + M_1}, \frac{4}{\rho}\right\}. \tag{47}$$

By the definitions (32) of B and $\varepsilon_{\rm tr}$

$$\delta \le \frac{B\varepsilon_{\rm tr}}{D^2}, \qquad M_0 \le \frac{B^2\varepsilon_{\rm tr}}{D^3}, \qquad M_1 \le \frac{B^2}{36D^2}, \qquad \frac{1}{\rho} \le \frac{B}{2D}.$$

Next, by definition, N is the first iteration, such that $f(x_N) - f^* \le \varepsilon$. Then for $j < N$ we have $g_j \ge \varepsilon/(2D)$. Therefore,

$$\frac{8\delta}{g_j} \le \frac{16\delta D}{\varepsilon} \le \frac{16B}{D}\frac{\varepsilon_{\rm tr}}{\varepsilon}, \qquad \frac{4}{\rho} \le \frac{2B}{D} \le \frac{16B}{D}$$

$$4\sqrt{\frac{M_0}{g_j} + M_1} \le 4\sqrt{\frac{2M_0 D}{\varepsilon} + M_1} \le \frac{4B}{D}\sqrt{2\frac{\varepsilon_{\rm tr}}{\varepsilon} + \frac{1}{36}} \le \frac{16B}{D}\max\left\{1, \frac{\varepsilon_{\rm tr}}{\varepsilon}\right\},$$

Moreover, for $j \ge 1$, the update for η_j gives

$$\frac{\eta_j}{g_j} = \frac{\tau_{j-1}}{2g_{j-1}}\min\left\{1, \frac{g_{j-1}}{g_j}\right\} \le \frac{\tau_{j-1}}{2g_{j-1}}.$$

Substituting these estimates into (47) gives for $1 \le j < N$

$$\frac{\tau_0}{g_0} \le \max\left\{\frac{\eta_0}{g_0}, \frac{16B}{D}\max\left\{1, \frac{\varepsilon_{\rm tr}}{\varepsilon}\right\}\right\}, \quad \frac{\tau_j}{g_j} \le \max\left\{\frac{\tau_{j-1}}{2g_{j-1}}, \frac{16B}{D}\max\left\{1, \frac{\varepsilon_{\rm tr}}{\varepsilon}\right\}\right\}.$$

Induction therefore gives

$$\frac{\tau_j}{g_j} \le \max\left\{\frac{\eta_0}{g_0}, \frac{16B}{D}\max\left\{1, \frac{\varepsilon_{\rm tr}}{\varepsilon}\right\}\right\}, \qquad 0 \le j < N. \tag{48}$$

Finally, by combining (45), (46), and (48), we obtain

$$\begin{aligned}\sum_{j=0}^{N-1} b_j &\le \frac{5}{4}(N-1) + \log_2 \max\left\{1, \frac{16Bg_0}{D\eta_0}\max\left\{1, \frac{\varepsilon_{\rm tr}}{\varepsilon}\right\}\right\} \\ &\le \frac{5}{4}(N-1) + 4 + \log_{2,+}\frac{Bg_0}{D\eta_0} + \log_{2,+}\frac{\varepsilon_{\rm tr}}{\varepsilon}.\end{aligned}$$

This proves (43).

Each trial requires at most one new gradient evaluation. Hence, the total number of gradient evaluations is at most $1 + N + \sum_{j=0}^{N-1} b_j$. Next, by the definition of $\varepsilon_{\rm tr}$, we have $\max\left\{\sqrt{\frac{M_0 D^3}{\varepsilon_{\rm tr}}}, \frac{\delta D^2}{\varepsilon_{\rm tr}}\right\} = B \ge 1$. Using $\log_2 t \le \min\left\{2\sqrt{t}, t\right\}$ for $t \ge 1$, we obtain $\log_{2,+}\frac{\varepsilon_{\rm tr}}{\varepsilon} \le 2\sqrt{\frac{M_0 D^3}{\varepsilon}} + \frac{\delta D^2}{\varepsilon}$. Combining these bounds with (4) proves (44).

□

C.6 Strongly Convex Analysis

Let f be strongly convex μ-strongly convex for $\mu > 0$:

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \qquad x, y \in \mathbb{R}^d. \tag{49}$$

This implies

$$\frac{\mu}{2}\|x - x^*\|^2 \le f(x) - f^* \le \frac{\|\nabla f(x)\|^2}{2\mu}. \tag{50}$$

In particular, strong convexity with monotonicity (Lemma C.1) imply

$$\|x_k - x^*\| \le \sqrt{\frac{2(f(x_k) - f^*)}{\mu}} \le \sqrt{\frac{2(f(x_0) - f^*)}{\mu}}. \tag{51}$$

We also denote $\delta_k := \|\nabla^2 f(x_k) - H(x_k)\| \le \delta$ for all $h \ge 0$.

C.6.1 One Step Estimates

Lemma C.10. *Under* (49)*, every completed iteration of Algorithm 2 satisfies*

$$f(x_{k+1}) - f^* \le \left(\frac{3\tau_k}{3\tau_k + 2\mu}\right)^2 (f(x_k) - f^*), \tag{52}$$

$$\|s_k\| \le \frac{\|\nabla f(x_k)\|}{\mu + \tau_k/2} \le \frac{\|\nabla f(x_k)\|}{\mu}. \tag{53}$$

Proof. By (16),

$$-\langle \nabla f(x_{k+1}), s_k\rangle \ge \frac{\|\nabla f(x_{k+1})\|^2}{2\tau_k} + \frac{3\tau_k}{8}\|s_k\|^2.$$

Apply (49) at x_{k+1}, then use (13) and (50):

$$f(x_k) - f(x_{k+1}) \ge -\langle \nabla f(x_{k+1}), s_k\rangle + \frac{\mu}{2}\|s_k\|^2 \ge \frac{\|\nabla f(x_{k+1})\|^2}{2\tau_k} + \left(\frac{3\tau_k}{8} + \frac{\mu}{2}\right)\|s_k\|^2$$
$$\ge \left(\frac{1}{2\tau_k} + \frac{4}{9\tau_k^2}\left(\frac{3\tau_k}{8} + \frac{\mu}{2}\right)\right)\|\nabla f(x_{k+1})\|^2 = \left(\frac{2}{3\tau_k} + \frac{2\mu}{9\tau_k^2}\right)\|\nabla f(x_{k+1})\|^2$$
$$\ge \left(\frac{4\mu}{3\tau_k} + \frac{4\mu^2}{9\tau_k^2}\right)(f(x_{k+1}) - f^*).$$

The last inequality uses (50) at x_{k+1}:

$$\|\nabla f(x_{k+1})\|^2 \ge 2\mu(f(x_{k+1}) - f^*).$$

Adding $f(x_{k+1}) - f^*$ to the bound for $f(x_k) - f(x_{k+1})$ and completing the square gives

$$f(x_k) - f^* \ge \left(1 + \frac{4\mu}{3\tau_k} + \frac{4\mu^2}{9\tau_k^2}\right)(f(x_{k+1}) - f^*)$$
$$= \left(1 + \frac{2\mu}{3\tau_k}\right)^2 (f(x_{k+1}) - f^*),$$

which is (52). For (53), recall that $s_k = x_{k+1} - x_k$. Strong convexity gives

$$\mu\|s_k\|^2 \le \langle \nabla f(x_{k+1}) - \nabla f(x_k), s_k\rangle.$$

The accepted MS test (2) gives

$$\|\nabla f(x_{k+1}) + \tau_k s_k\| \le \frac{\tau_k}{2}\|s_k\|.$$

Hence, by Cauchy-Schwarz,

$$\langle \nabla f(x_{k+1}), s_k\rangle = \langle \nabla f(x_{k+1}) + \tau_k s_k, s_k\rangle - \tau_k\|s_k\|^2$$
$$\le \|\nabla f(x_{k+1}) + \tau_k s_k\|\|s_k\| - \tau_k\|s_k\|^2 \le \frac{\tau_k}{2}\|s_k\|^2 - \tau_k\|s_k\|^2 = -\frac{\tau_k}{2}\|s_k\|^2.$$

Combining the two bounds and applying Cauchy-Schwarz once more gives

$$\mu\|s_k\|^2 \le -\frac{\tau_k}{2}\|s_k\|^2 - \langle \nabla f(x_k), s_k\rangle,$$
$$\left(\mu + \frac{\tau_k}{2}\right)\|s_k\|^2 \le -\langle \nabla f(x_k), s_k\rangle \le \|\nabla f(x_k)\|\|s_k\|.$$

Dividing by $\|s_k\| > 0$ and using $\tau_k > 0$ gives

$$\|s_k\| \le \frac{\|\nabla f(x_k)\|}{\mu + \tau_k/2} \le \frac{\|\nabla f(x_k)\|}{\mu}.$$

□

C.6.2 Global Linear Convergence with Inexact Hessians

We recall definition of r_0 from (18):

$$r_0 = \begin{cases} \frac{1}{6\sqrt{M_1}}, & \text{under (GH),} \\ \min\left\{\frac{\rho}{2}, \frac{1}{6\sqrt{M_1}}\right\}, & \text{under (LH).} \end{cases}$$

Theorem C.11. *Suppose that* (49) *and* (IH) *hold, and either* (GH) *or* (LH) *holds. Define*

$$\overline{\tau} := \max\left\{\eta_0,\ 8\delta + 48M_0\sqrt{\frac{2(f(x_0)-f^*)}{\mu}}, \frac{2(f(x_0)-f^*)}{r_0^2}\right\}. \tag{54}$$

Every completed iteration satisfies $\tau_k \le \overline{\tau}$, and

$$f(x_k) - f^* \le \left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^{2k}(f(x_0)-f^*), \qquad k \ge 0. \tag{55}$$

Proof. All searches terminate by Lemma C.2. We prove $\tau_k \le \overline{\tau}$ by induction.

Suppose that on iteration k at least one trial is rejected, so $\tau_k > \eta_k$. If $\|s_k\| \le r_0$, then (19) and (51) give

$$\begin{aligned}\tau_k &\le 8\delta_k + 24M_0\|s_k\| \le 8\delta + 24M_0\left(\|x_k - x^*\| + \|x_{k+1} - x^*\|\right) \\ &\le 8\delta + 48M_0\sqrt{\frac{2(f(x_0)-f^*)}{\mu}} \le \overline{\tau}.\end{aligned}$$

If $\|s_k\| > r_0$, then (14) gives

$$\frac{\tau_k}{2}r_0^2 < \frac{\tau_k}{2}\|s_k\|^2 \le f(x_k) - f(x_{k+1}) \le f(x_0) - f^*,$$

so again $\tau_k \le \overline{\tau}$.

Now, let's assume that in iteration k there is no rejection. For $k = 0$, the accepted value is $\tau_0 = \eta_0 \le \overline{\tau}$. For $k \ge 1$, line 4 gives

$$\tau_k = \eta_k \le \frac{\tau_{k-1}}{2} \le \overline{\tau}.$$

Thus the induction closes. Since $0 < \tau_k \le \overline{\tau}$,

$$\left(\frac{3\tau_k}{3\tau_k+2\mu}\right)^2 = \left(1 - \frac{2\mu}{3\tau_k+2\mu}\right)^2 \le \left(1 - \frac{2\mu}{3\overline{\tau}+2\mu}\right)^2 = \left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^2.$$

Substitution into (52) and iteration prove (55). □

Corollary C.12. *Under the assumptions of Theorem C.11, let $0 < \varepsilon < f(x_0) - f^*$ and $N := \min\{k \ge 0 : f(x_k) - f^* \le \varepsilon\}$. The total number of gradient evaluations before obtaining x_N is at most*

$$1 + \frac{9}{4}N + \log_{2,+}\left(\frac{\|\nabla f(x_0)\|}{\eta_0}\max\left\{\frac{24M_0}{\mu} + \frac{8\delta}{\sqrt{2\mu\varepsilon}}, \frac{2}{r_0}\right\}\right). \tag{56}$$

Consequently, it is bounded by

$$1 + \frac{9}{4}\left\lceil\frac{\log((f(x_0)-f^*)/\varepsilon)}{2\log\left(1+\frac{2\mu}{3\overline{\tau}}\right)}\right\rceil + \log_{2,+}\left(\frac{\|\nabla f(x_0)\|}{\eta_0}\max\left\{\frac{24M_0}{\mu} + \frac{8\delta}{\sqrt{2\mu\varepsilon}}, \frac{2}{r_0}\right\}\right).$$

For $\delta = 0$, the logarithmic term in (56) *is independent of ε, and the bound holds for any $N \ge 1$ completed iterations.*

Proof. Let $g_k := \|\nabla f(x_k)\|$ and let b_k count the rejected trials on iteration k. By (45) and (46),

$$\sum_{k=0}^{N-1} b_k \le \frac{5}{4}(N-1) + \log_2\frac{g_0\tau_{N-1}}{\eta_0 g_{N-1}}.$$

We prove by induction that

$$\frac{\tau_k}{g_k} \le \max\left\{\frac{2^{-k}\eta_0}{g_0}, \frac{24M_0}{\mu} + \frac{8\delta}{g_k}, \frac{2}{r_0}\right\}. \tag{57}$$

Suppose first that at least one trial is rejected. If $\|s_k\| \le r_0$, then (19) and (53) give

$$\tau_k \le 8\delta + 24M_0\|s_k\| \le 8\delta + \frac{24M_0}{\mu} g_k.$$

If $\|s_k\| > r_0$, then (53) gives

$$\tau_k \le \frac{2g_k}{\|s_k\|} < \frac{2g_k}{r_0}.$$

Both estimates imply (57) for every $k \ge 0$ with a rejection.

For $k = 0$, the estimates above apply if a trial is rejected. Otherwise, $\tau_0 = \eta_0$, and

$$\frac{\tau_0}{g_0} = \frac{\eta_0}{g_0} \le \max\left\{\frac{\eta_0}{g_0}, \frac{24M_0}{\mu} + \frac{8\delta}{g_0}, \frac{2}{r_0}\right\}.$$

Thus (57) holds for $k = 0$. Now let $k \ge 1$ and assume it holds for $k - 1$. If a trial is rejected, the estimates above apply again. If there is no rejection, line 4 gives

$$\begin{aligned}\tau_k = \frac{\tau_{k-1}}{2}\min\left\{1, \frac{g_k}{g_{k-1}}\right\} &\le \max\left\{2^{-k}\eta_0\frac{g_k}{g_0}, 4\delta + \frac{12M_0}{\mu}g_k, \frac{g_k}{r_0}\right\} \\ &\le \max\left\{2^{-k}\eta_0\frac{g_k}{g_0}, 8\delta + \frac{24M_0}{\mu}g_k, \frac{2g_k}{r_0}\right\}.\end{aligned}$$

This proves (57) by induction. By the definition of N, we have $f(x_{N-1}) - f^* > \varepsilon$. By (50), $g_{N-1} > \sqrt{2\mu\varepsilon}$. Substituting this into (57), we obtain

$$\begin{aligned}\frac{g_0\tau_{N-1}}{\eta_0 g_{N-1}} &\le \max\left\{2^{-(N-1)}, \frac{g_0}{\eta_0}\left(\frac{24M_0}{\mu} + \frac{8\delta}{g_{N-1}}\right), \frac{2g_0}{\eta_0 r_0}\right\} \\ &\le \max\left\{1, \frac{g_0}{\eta_0}\left(\frac{24M_0}{\mu} + \frac{8\delta}{\sqrt{2\mu\varepsilon}}\right), \frac{2g_0}{\eta_0 r_0}\right\}.\end{aligned}$$

Each trial passing PD requires at most one new gradient evaluation. The accepted gradient is reused at the next iteration. Hence the total number of gradient evaluations is at most

$$1 + N + \sum_{k=0}^{N-1} b_k,$$

which, together with the preceding bounds, proves (56).

□

C.6.3 Proof of Theorem 3.3

We refine the global bound from Theorem C.11 to show superlinear convergence by retaining the individual Hessian errors δ_k instead of replacing them by δ.

Theorem C.13. *Under the assumptions of Theorem C.11, let $\overline{\tau}$ be defined by* (54) *and set*

$$\overline{\tau}_k := \min\left\{\overline{\tau}, 8\max_{0\le j\le k} 2^{j-k}\delta_j + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau} + 2\mu}\right)^k\right\}, \qquad \zeta_k := \left(\frac{3\overline{\tau}_k}{3\overline{\tau}_k + 2\mu}\right)^2.$$

Every completed iteration satisfies

$$\begin{aligned}\tau_k &\le \overline{\tau}_k, \\ f(x_{k+1}) - f^* &\le \zeta_k(f(x_k) - f^*), \qquad 0 < \zeta_k < 1.\end{aligned} \tag{58}$$

Proof. For every $i \geq 1$, by (13) and (53),

$$\|\nabla f(x_i)\| \leq \frac{3\tau_{i-1}}{2}\|s_{i-1}\| \leq \frac{3\tau_{i-1}}{\tau_{i-1}+2\mu}\|\nabla f(x_{i-1})\|.$$

Dividing by $\|\nabla f(x_{i-1})\| > 0$ and substituting into the update in line 4 gives

$$\frac{\eta_i}{\tau_{i-1}} = \frac{1}{2}\min\left\{1, \frac{\|\nabla f(x_i)\|}{\|\nabla f(x_{i-1})\|}\right\} \leq \min\left\{\frac{1}{2}, \frac{3\tau_{i-1}}{2\tau_{i-1}+4\mu}\right\} \leq \min\left\{\frac{1}{2}, \frac{3\overline{\tau}}{2\overline{\tau}+4\mu}\right\} \leq \frac{3\overline{\tau}}{3\overline{\tau}+2\mu}, \tag{59}$$

where we used $\tau_{i-1} \leq \overline{\tau}$ from Theorem C.11. Fix a completed iteration k.

If no trial was rejected during iterations $0, \ldots, k$, then $\tau_i = \eta_i$ for $i = 0, \ldots, k$. By (59),

$$\tau_k = \eta_0 \prod_{i=1}^{k} \frac{\eta_i}{\tau_{i-1}} \leq \eta_0 \left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^k \leq \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^k.$$

Otherwise, let j be the last iteration among $0, \ldots, k$ on which a trial is rejected. By (50) and (55),

$$\|s_j\| \leq \|x_j - x^*\| + \|x_{j+1} - x^*\|$$
$$\leq \sqrt{\frac{2(f(x_j)-f^*)}{\mu}} + \sqrt{\frac{2(f(x_{j+1})-f^*)}{\mu}} \leq 2\sqrt{\frac{2(f(x_0)-f^*)}{\mu}}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^j.$$

Since a trial is rejected on iteration j, the two cases in the proof of Theorem C.11 give the first bound below. Substituting the step estimate above and (55), we obtain

$$\tau_j \leq \max\left\{8\delta_j + 24M_0\|s_j\|, \frac{2(f(x_j)-f^*)}{r_0^2}\right\}$$
$$\leq \max\left\{8\delta_j + 48M_0\sqrt{\frac{2(f(x_0)-f^*)}{\mu}}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^j, \frac{2(f(x_0)-f^*)}{r_0^2}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^{2j}\right\}.$$

By (54) we obtain

$$\tau_j \leq 8\delta_j + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^j. \tag{60}$$

By the choice of j, no trial is rejected on iterations $j+1, \ldots, k$, so $\tau_i = \eta_i$ for $i = j+1, \ldots, k$. Applying (59), we obtain

$$\prod_{i=j+1}^{k}\frac{\eta_i}{\tau_{i-1}} \leq \prod_{i=j+1}^{k}\frac{1}{2} = 2^{j-k}, \quad \prod_{i=j+1}^{k}\frac{\eta_i}{\tau_{i-1}} \leq \prod_{i=j+1}^{k}\frac{3\overline{\tau}}{3\overline{\tau}+2\mu} = \left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^{k-j}.$$

Substituting (60), we obtain

$$\tau_k = \tau_j \prod_{i=j+1}^{k}\frac{\eta_i}{\tau_{i-1}} \leq \left(8\delta_j + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^j\right)\prod_{i=j+1}^{k}\frac{\eta_i}{\tau_{i-1}} \leq 8\delta_j 2^{j-k} + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^j\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^{k-j}$$
$$= 8\delta_j 2^{j-k} + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^k \leq 8\max_{0\leq i\leq k} 2^{i-k}\delta_i + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^k.$$

Thus, in both cases,

$$\tau_k \leq 8\max_{0\leq j\leq k} 2^{j-k}\delta_j + \overline{\tau}\left(\frac{3\overline{\tau}}{3\overline{\tau}+2\mu}\right)^k. \tag{61}$$

Combining (61) with $\tau_k \leq \overline{\tau}$ gives $\tau_k \leq \overline{\tau}_k$.

Since $0 < \tau_k \leq \overline{\tau}_k$,

$$\left(\frac{3\tau_k}{3\tau_k+2\mu}\right)^2 = \left(1 - \frac{2\mu}{3\tau_k+2\mu}\right)^2 \leq \left(1 - \frac{2\mu}{3\overline{\tau}_k+2\mu}\right)^2 = \zeta_k.$$

Substitution into (52) proves (58). □

C.6.4 Global Superlinear Convergence under Relative Hessian Accuracy

We consider Hessian approximations satisfying

$$\delta_k = \|H(x_k) - \nabla^2 f(x_k)\| \le \nu \|\nabla f(x_k)\|, \qquad k \ge 0, \tag{62}$$

for some $\nu > 0$.

Corollary C.14. *Under the assumptions of Theorem C.11, suppose additionally that* (62) *holds. The method either returns a minimizer or converges superlinearly in function value:*

$$\frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} \le \left(\frac{3\tau_k}{3\tau_k + 2\mu}\right)^2 \longrightarrow 0.$$

The first $N \ge 1$ completed iterations require at most N Hessian evaluations and

$$1 + \frac{9}{4}N + \log_{2,+}\left(\frac{\|\nabla f(x_0)\|}{\eta_0} \max\left\{\frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\}\right) \tag{63}$$

gradient evaluations.

Proof. Let $g_k := \|\nabla f(x_k)\|$. In the proof of (57), retain the current error δ_k on iterations with a rejection and use $\delta_k/g_k \le \nu$ from (62). The same bounds for the two cases give

$$\frac{\tau_0}{g_0} \le \max\left\{\frac{\eta_0}{g_0}, \frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\}, \qquad \frac{\tau_k}{g_k} \le \max\left\{\frac{\tau_{k-1}}{2g_{k-1}}, \frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\}, \quad k \ge 1.$$

Iterating this inequality gives

$$\frac{\tau_k}{g_k} \le \max\left\{\frac{2^{-k}\eta_0}{g_0}, \frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\}. \tag{64}$$

By (14) and (55), for $k \ge 1$,

$$g_k^2 \le \frac{3\tau_{k-1}}{2}\left(f(x_{k-1}) - f(x_k)\right) \le \frac{3\overline{\tau}}{2}\left(f(x_{k-1}) - f^*\right) \le \frac{3\overline{\tau}(f(x_0) - f^*)}{2}\left(\frac{3\overline{\tau}}{3\overline{\tau} + 2\mu}\right)^{2(k-1)}.$$

Substituting into (64) gives

$$\tau_k \le \sqrt{\frac{3\overline{\tau}(f(x_0) - f^*)}{2}}\left(\frac{3\overline{\tau}}{3\overline{\tau} + 2\mu}\right)^{k-1} \max\left\{\frac{2^{-k}\eta_0}{g_0}, \frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\} \longrightarrow 0.$$

Thus, (52) proves superlinear convergence.

Let b_k count the rejected trials on iteration k. By (45) and (46),

$$\sum_{k=0}^{N-1} b_k \le \frac{5}{4}(N-1) + \log_2 \frac{g_0\tau_{N-1}}{\eta_0 g_{N-1}}.$$

By (64),

$$\frac{g_0\tau_{N-1}}{\eta_0 g_{N-1}} \le \max\left\{2^{-(N-1)}, \frac{g_0}{\eta_0}\left(\frac{24M_0}{\mu} + 8\nu\right), \frac{2g_0}{\eta_0 r_0}\right\} \le \max\left\{1, \frac{g_0}{\eta_0}\left(\frac{24M_0}{\mu} + 8\nu\right), \frac{2g_0}{\eta_0 r_0}\right\}.$$

Hence

$$\sum_{k=0}^{N-1} b_k \le \frac{5}{4}(N-1) + \log_{2,+}\left(\frac{g_0}{\eta_0}\max\left\{\frac{24M_0}{\mu} + 8\nu, \frac{2}{r_0}\right\}\right).$$

The gradient count is at most $1 + N + \sum_{k=0}^{N-1} b_k$, which proves (63). □

C.6.5 Local Quadratic Convergence

Theorem C.15. *Under the assumptions of Corollary C.14, set*

$$C := \frac{9\left(\|\nabla^2 f(x^*)\| + \mu\right)}{2\mu^2}\left(8\nu + \frac{24M_0}{\mu}\right)^2.$$

Unless a minimizer is returned earlier, let $K_0 \geq 1$ satisfy

$$f(x_{K_0}) - f^* \leq \min\left\{\frac{\mu r_0^2}{8}, \frac{1}{2C}\right\} \tag{65}$$

and define

$$K := K_0 + \left\lceil \log_{2,+} \frac{\tau_{K_0-1}}{2\left(8\nu + \frac{24M_0}{\mu}\right)\sqrt{2\mu(f(x_{K_0-1}) - f^*)}} \right\rceil. \tag{66}$$

Then, for every $k \geq K$,

$$f(x_{k+1}) - f^* \leq C\left(f(x_k) - f^*\right)^2. \tag{67}$$

The gradient norms and the distances to x^* also converge quadratically. After a finite initial cost independent of ε, attaining $f(x) - f^* \leq \varepsilon$ requires $\mathcal{O}(\log\log(1/\varepsilon))$ additional iterations, Hessian evaluations, and gradient evaluations.

Proof. By Theorem C.11, an index K_0 satisfying (65) exists unless a minimizer is returned earlier. Since $\|\nabla^2 f(x^*)\| \geq \mu$, the definition of C gives

$$C \geq \frac{9}{\mu}\left(\frac{24M_0}{\mu}\right)^2 \geq \frac{4M_0^2}{\mu^3}, \qquad \frac{1}{2C} \leq \frac{\mu^3}{8M_0^2}.$$

Together with monotonicity and (50), condition (65) gives, for $k \geq K_0$,

$$\|x_k - x^*\| \leq \sqrt{\frac{2(f(x_k) - f^*)}{\mu}} \leq \frac{1}{2}\min\left\{r_0, \frac{\mu}{M_0}\right\},$$

$$\|s_k\| \leq \|x_k - x^*\| + \|x_{k+1} - x^*\| \leq 2\sqrt{\frac{2(f(x_{K_0}) - f^*)}{\mu}} \leq r_0.$$

Let $g_k := \|\nabla f(x_k)\|$. Since $\|s_k\| \leq r_0$, the proof of (64) applies without the term $2/r_0$. Starting at K_0, it gives, for $j \geq 0$,

$$\frac{\tau_{K_0+j}}{g_{K_0+j}} \leq \max\left\{\frac{2^{-j-1}\tau_{K_0-1}}{g_{K_0-1}}, 8\nu + \frac{24M_0}{\mu}\right\} \leq \max\left\{\frac{2^{-j-1}\tau_{K_0-1}}{\sqrt{2\mu(f(x_{K_0-1}) - f^*)}}, 8\nu + \frac{24M_0}{\mu}\right\}.$$

By (66), for every $k \geq K$,

$$\tau_k \leq \left(8\nu + \frac{24M_0}{\mu}\right) g_k. \tag{68}$$

Combining (13) and (53), we obtain

$$g_{k+1} \leq \frac{3\tau_k}{2}\|s_k\| \leq \frac{3\tau_k}{2\mu} g_k \leq \frac{3}{2\mu}\left(8\nu + \frac{24M_0}{\mu}\right) g_k^2. \tag{69}$$

Thus, the gradient norms converge quadratically. Next, we relate the gradient norm to the function gap. For $\|z - x^*\| \leq \min\{r_0, \mu/M_0\}$, either (GH) or (LH) gives

$$\|\nabla^2 f(z)\| \leq \|\nabla^2 f(x^*)\| + M_0\|z - x^*\| \leq \|\nabla^2 f(x^*)\| + \mu,$$

since $\nabla f(x^*) = 0$ and $r_0 \leq \rho/2$ under (LH). Thus, f has $(\|\nabla^2 f(x^*)\| + \mu)$-Lipschitz gradient on this ball. Hence, for $k \geq K_0$,

$$g_k \leq \left(\|\nabla^2 f(x^*)\| + \mu\right)\|x_k - x^*\|,$$

$$\left\|x_k - \frac{\nabla f(x_k)}{\|\nabla^2 f(x^*)\| + \mu} - x^*\right\| \leq \|x_k - x^*\| + \frac{g_k}{\|\nabla^2 f(x^*)\| + \mu} \leq 2\|x_k - x^*\| \leq \min\left\{r_0, \frac{\mu}{M_0}\right\}.$$

By Lipschitznes of the gradient, for any y in the ball,

$$f(y) \le f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{\|\nabla^2 f(x^*)\| + \mu}{2} \|y - x_k\|^2.$$

Substituting the point above, we obtain

$$f^* \le f\left(x_k - \frac{\nabla f(x_k)}{\|\nabla^2 f(x^*)\| + \mu}\right) \le f(x_k) - \frac{g_k^2}{\|\nabla^2 f(x^*)\| + \mu} + \frac{\left(\|\nabla^2 f(x^*)\| + \mu\right) g_k^2}{2\left(\|\nabla^2 f(x^*)\| + \mu\right)^2}$$
$$= f(x_k) - \frac{g_k^2}{2\left(\|\nabla^2 f(x^*)\| + \mu\right)}.$$

Rearranging gives

$$g_k^2 \le 2\left(\|\nabla^2 f(x^*)\| + \mu\right)(f(x_k) - f^*).$$

Substituting this and (68) into (52), for $k \ge K$ we get

$$f(x_{k+1}) - f^* \le \left(\frac{3\tau_k}{3\tau_k + 2\mu}\right)^2 (f(x_k) - f^*)$$
$$\le \frac{9}{4\mu^2}\left(8\nu + \frac{24M_0}{\mu}\right)^2 g_k^2 (f(x_k) - f^*) \le C(f(x_k) - f^*)^2.$$

This proves (67). For the distances, strong convexity and (69) give

$$\|x_{k+1} - x^*\| \le \frac{g_{k+1}}{\mu} \le \frac{3\left(\|\nabla^2 f(x^*)\| + \mu\right)^2}{2\mu^2}\left(8\nu + \frac{24M_0}{\mu}\right)\|x_k - x^*\|^2.$$

By (65), $C(f(x_K) - f^*) \le 1/2$. Iterating (67) gives

$$f(x_{K+n}) - f^* \le \frac{1}{C}\left(C(f(x_K) - f^*)\right)^{2^n} \le \frac{2^{-2^n}}{C}, \qquad n \ge 0.$$

Thus, $f(x_{K+n}) - f^* \le \varepsilon$ after $n = \mathcal{O}(\log\log(1/\varepsilon))$ additional iterations. By Corollary C.14, this requires $\mathcal{O}(K + \log\log(1/\varepsilon))$ Hessian and gradient evaluations in total. □

D Nesterov-type Acceleration

We present an accelerated cubic method for the global condition (GH). Fix a positive radius estimate $\bar{R} \ge R$ and an initial scale $\Gamma_0 > 0$ satisfying

$$\Gamma_0 \bar{R}^3 \ge 3\left(f(x_0) - f^*\right). \tag{70}$$

We discuss how to choose parameters satisfying (70) in Section D.2.

We present the resulting method as Algorithm 4. The acceleration mechanism aggregates the linear models $\ell_w(x) := f(w) + \langle \nabla f(w), x - w \rangle$. At each iteration, α_k is defined by the scalar equation in line 5 of Algorithm 4. This definition is well posed because the mapping $\alpha \mapsto \alpha^3/(1-\alpha)$ is continuous and strictly increasing from 0 to $+\infty$ on $(0, 1)$. Hence, for $\Gamma_k > 0$ and $\mathcal{M}_k > 0$, the equation has a unique solution, which can be found by bisection.

Theorem D.1. *Assume that f is convex,* (GH) *and* (70) *hold. For every $\varepsilon > 0$, Algorithm 4 returns an iterate w_N satisfying $f(w_N) - f^* \le \varepsilon$ for*

$$N = \mathcal{O}\left(\left(\frac{M_0 \bar{R}^3}{\varepsilon}\right)^{1/3} + \left(1 + \sqrt{M_1}\bar{R}\right)\log\max\left\{1, \min\left\{\frac{M_1^{3/2}\bar{R}^3\Gamma_0}{M_0}, \frac{\Gamma_0 \bar{R}^3}{\varepsilon}\right\}\right\}\right). \tag{74}$$

Proof sketch. We use the standard accelerated-cubic estimating sequence technique Nesterov (2008). The main obstruction is that under (GH) the regularization coefficient $\mathcal{M}_k$ must dominate the generalized smoothness parameter $L(w_{k+1})$ at the *output* w_{k+1} of the cubic step (73), although this output is not known when the coefficient is selected. Inspired by Tyurin (2026), we provide an explicit way of choosing regularization parameter $\mathcal{M}_k$ (71), such that this condition is satisfied. The complete proof is given in Appendix D.1.

Algorithm 4 Accelerated cubic regularization under global generalized Hessian smoothness

1: **Input:** $x_0 \in \mathbb{R}^d$, $M_0 > 0$, $M_1 \geq 0$, $\bar{R} \geq R$ with $\bar{R} > 0$, and $\Gamma_0 > 0$ satisfying $\Gamma_0 \bar{R}^3 \geq 3\,(f(x_0) - f^*)$
2: Set $A_0 := 1$, $w_0 := x_0$, and $z_0 := x_0$, $\psi_0(x) := f(x_0) + \frac{\Gamma_0}{3}\|x - x_0\|^3$.
3: **for** $k \geq 0$ **do**
4: Set

$$\mathcal{M}_k := \max\left\{2M_0, 12\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_k\right\}. \tag{71}$$

5: Find the unique $\alpha_k \in (0,1)$ satisfying $\frac{\alpha_k^3}{1-\alpha_k} = \frac{\Gamma_k}{72\mathcal{M}_k}$.
6: Set

$$A_{k+1} := \tfrac{A_k}{1-\alpha_k}, \quad a_{k+1} := \alpha_k A_{k+1}, \quad \Gamma_{k+1} := (1-\alpha_k)\Gamma_k. \tag{72}$$

7: Set $v_k := (1-\alpha_k)w_k + \alpha_k z_k$.
8: Compute

$$s_k := \operatorname*{argmin}_{s\in\mathbb{R}^d}\left\{\langle\nabla f(v_k), s\rangle + \tfrac{1}{2}\left\langle\nabla^2 f(v_k)s, s\right\rangle + \tfrac{15\mathcal{M}_k}{6}\|s\|^3\right\}, \quad w_{k+1} := v_k + s_k. \tag{73}$$

9: Set $\psi_{k+1}(x) := \psi_k(x) + a_{k+1}\ell_{w_{k+1}}(x)$ and compute $z_{k+1} := \operatorname{argmin}_{x\in\mathbb{R}^d}\psi_{k+1}(x)$.

Let us consider one step of Algorithm 4. For some $\mathcal{M} > 0$, let $A_+(\mathcal{M})$, $w(\mathcal{M})$, and $\psi_+(\cdot;\mathcal{M})$ denote the quantities produced by one iteration, and set $\psi_+^*(\mathcal{M}) = \min_{x\in\mathbb{R}^d}\psi_+(x;\mathcal{M})$. Assume first that $\mathcal{M} \geq L(w(\mathcal{M}))$. Then the standard accelerated-cubic estimating-sequence argument gives

$$A_+(\mathcal{M})f(w(\mathcal{M})) + \tfrac{A_+(\mathcal{M})}{9}\tfrac{\|\nabla f(w(\mathcal{M}))\|^{3/2}}{\sqrt{\mathcal{M}}} \leq \psi_+^*(\mathcal{M}) \leq A_+(\mathcal{M})f^* + \tfrac{2}{3}\Gamma_0\bar{R}^3. \tag{75}$$

Thus, conditioned on proper choice of $\mathcal{M} \geq L(w(\mathcal{M}))$, the usual estimating-sequence invariant closes. It remains to choose $\mathcal{M}$ before $w(\mathcal{M})$ is known. Consider the virtual family of steps and set $\mathcal{U}_k = \{\mathcal{M} > 0 : \mathcal{M} < L(w(\mathcal{M}))\}$. If this set is nonempty, let $\mathcal{M}_k^\dagger = \sup\mathcal{U}_k$. Since all sufficiently large $\mathcal{M}$ are admissible, $\mathcal{M}_k^\dagger < +\infty$, and continuity gives $\mathcal{M}_k^\dagger = L(w(\mathcal{M}_k^\dagger))$. Let $g^\dagger = \nabla f(w(\mathcal{M}_k^\dagger))$. Combining the two sides of (75) at this boundary yields

$$\tfrac{\|g^\dagger\|^{3/2}}{\sqrt{\mathcal{M}_k^\dagger}} \leq 6\Gamma_k\bar{R}^3, \qquad \mathcal{M}_k^\dagger \leq \max\left\{2M_0, 12\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_k\right\} = \mathcal{M}_k.$$

Hence the explicit $\mathcal{M}_k$ used by the algorithm is admissible. This is why we retain the positive remainder in (75). Although it can be omitted in the standard convergence argument, here it controls the gradient at the unknown output and closes the explicit choice of $\mathcal{M}$. For the actual step, admissibility gives

$$A_{k+1}(f(w_{k+1}) - f^*) \leq \psi_{k+1}^* - A_{k+1}f^* \leq \tfrac{2}{3}\Gamma_0\bar{R}^3.$$

Then, recurrence for Γ_k then gives the rate (74).

D.1 Proof of Nesterov-type Acceleration under (GH)

D.1.1 Auxiliary Lemmas

Lemma D.2 (Nesterov (2008, Lemma 2))**.** *For any $a \geq 0$, $g \in \mathbb{R}^d$, and $\beta > 0$,*

$$\min_{h\in\mathbb{R}^d}\left\{\frac{\beta}{3!}\|h\|^3 + a\,\langle g, h\rangle\right\} = -\frac{2\sqrt{2}}{3}\frac{a^{3/2}\|g\|^{3/2}}{\sqrt{\beta}}. \tag{76}$$

Lemma D.3 (Agafonov et al. (2024a, Lemma 7))**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be convex, $x_0 \in \mathbb{R}^d$, and $\theta > 0$. Define*

$$\varphi(x) := h(x) + \frac{\theta}{3}\|x - x_0\|^3, \qquad \bar{x} := \operatorname*{argmin}_{x\in\mathbb{R}^d}\varphi(x).$$

Then, for every $u \in \mathbb{R}^d$,

$$\varphi(\bar{x} + u) \geq \varphi(\bar{x}) + \frac{\theta}{6}\|u\|^3.$$

Lemma D.4. *The weight updates in Algorithm 4 admit the following equivalent representation:*

$$a_{k+1} := A_{k+1} - A_k,\ \ \Gamma_k = \frac{\Gamma_0}{A_k},\ \ v_k = \frac{A_k w_k + a_{k+1} z_k}{A_{k+1}},\ \ a_{k+1}^3 = \frac{\Gamma_0 A_{k+1}^2}{72\mathcal{M}_k}. \tag{77}$$

Proof. Since $A_0 = 1$, the updates of A_k and Γ_k (72) give

$$A_{k+1}\Gamma_{k+1} = \frac{A_k}{1-\alpha_k}(1-\alpha_k)\Gamma_k = A_k\Gamma_k = \Gamma_0,$$

and hence $\Gamma_k = \Gamma_0/A_k$. Moreover,

$$A_{k+1} - A_k = A_{k+1} - (1-\alpha_k)A_{k+1} = \alpha_k A_{k+1} = a_{k+1},$$
$$A_{k+1}v_k = (1-\alpha_k)A_{k+1}w_k + \alpha_k A_{k+1} z_k = A_k w_k + a_{k+1} z_k.$$

Finally,

$$\frac{\alpha_k^3}{1-\alpha_k} = \frac{\Gamma_k}{72\mathcal{M}_k} \quad\Longleftrightarrow\quad \alpha_k^3 A_{k+1}^3 = \frac{\Gamma_k A_k A_{k+1}^2}{72\mathcal{M}_k} \quad\Longleftrightarrow\quad a_{k+1}^3 = \frac{\Gamma_0 A_{k+1}^2}{72\mathcal{M}_k}.$$

□

D.1.2 Cubic Step Progress

Lemma D.5. *Fix $v \in \mathbb{R}^d$ and $\mathcal{M} > 0$, and define*

$$s := \operatorname*{argmin}_{h\in\mathbb{R}^d} \left\{ \langle \nabla f(v), h\rangle + \frac{1}{2}\langle \nabla^2 f(v)h, h\rangle + \frac{15\mathcal{M}}{6}\|h\|^3 \right\}. \tag{78}$$

Set $w := v + s$. Suppose that either (GH) *holds, or* (LH) *holds and $\|s\| \le \rho$. If*

$$\mathcal{M} \ge L(w) = M_0 + M_1\|\nabla f(w)\|,$$

then

$$\|\nabla f(w)\| \le 9\mathcal{M}\|s\|^2, \tag{79}$$
$$\langle \nabla f(w), v - w\rangle \ge 6\mathcal{M}\|s\|^3, \tag{80}$$
$$\langle \nabla f(w), v - w\rangle \ge \frac{2}{9}\frac{\|\nabla f(w)\|^{3/2}}{\sqrt{\mathcal{M}}}. \tag{81}$$

Proof. By optimality condition for the cubic subproblem (78)

$$\nabla f(v) + \nabla^2 f(v)s + \frac{15\mathcal{M}}{2}\|s\|s = 0. \tag{82}$$

Under (GH), Lemma A.1 applies at w with $v - w = -s$. Under (LH), Lemma A.2 applies in the same way, since $\|s\| \le \rho$. Together with the corresponding smoothness condition, this gives

$$\|\nabla f(v) - \nabla f(w) + \nabla^2 f(w)s\| \le \frac{L(w)}{2}\|s\|^2 \le \frac{\mathcal{M}}{2}\|s\|^2,$$
$$\|\nabla^2 f(v) - \nabla^2 f(w)\| \le L(w)\|s\| \le \mathcal{M}\|s\|.$$

By the triangle inequality

$$\|\nabla f(v) - \nabla f(w) + \nabla^2 f(v)s\|$$
$$\le \|\nabla f(v) - \nabla f(w) + \nabla^2 f(w)s\| + \|\nabla^2 f(v) - \nabla^2 f(w)\|\|s\| \le \frac{3\mathcal{M}}{2}\|s\|^2.$$

Combining this estimate with (82)

$$\|\nabla f(w)\| = \|\nabla f(w) - \big(\nabla f(v) + \nabla^2 f(v)s + \tfrac{15\mathcal{M}}{2}\|s\|s\big)\|$$
$$\le \|\nabla f(v) - \nabla f(w) + \nabla^2 f(v)s\| + \frac{15\mathcal{M}}{2}\|s\|^2 \le 9\mathcal{M}\|s\|^2,$$

which proves (79).

Next,

$$\langle \nabla f(w), -s\rangle \overset{(82)}{=} \left\langle \nabla f(v) - \nabla f(w) + \nabla^2 f(v)s, s\right\rangle + \frac{15\mathcal{M}}{2}\|s\|^3,$$
$$\langle \nabla f(w), -s\rangle \geq -\frac{3\mathcal{M}}{2}\|s\|^3 + \frac{15\mathcal{M}}{2}\|s\|^3 = 6\mathcal{M}\|s\|^3,$$

which proves (80). Finally, (79) implies

$$\|s\| \geq \sqrt{\frac{\|\nabla f(w)\|}{9\mathcal{M}}}.$$

Substituting this bound into (80) proves (81). □

D.1.3 Estimating Sequence One Step Analysis

Lemma D.6. *Suppose that, at the beginning of iteration k,*

$$\psi_k^* := \min_{x\in\mathbb{R}^d} \psi_k(x) \geq A_k f(w_k). \tag{83}$$

Then

$$\psi_{k+1}^* \geq A_{k+1} f(w_{k+1}) + A_{k+1}\langle \nabla f(w_{k+1}), v_k - w_{k+1}\rangle - \frac{A_{k+1}}{9}\frac{\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\mathcal{M}_k}}. \tag{84}$$

If, in addition, the hypotheses of Lemma D.5 hold for $v = v_k$ and $\mathcal{M} = \mathcal{M}_k$, then

$$\psi_{k+1}^* \geq A_{k+1} f(w_{k+1}) + \frac{A_{k+1}}{9}\frac{\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\mathcal{M}_k}}. \tag{85}$$

Consequently,

$$\psi_{k+1}^* \geq A_{k+1} f(w_{k+1}). \tag{86}$$

Proof. By the definition of ψ_k,

$$\psi_k(x) = f(x_0) + \sum_{i=0}^{k-1} a_{i+1}\ell_{w_{i+1}}(x) + \frac{\Gamma_0}{3}\|x - x_0\|^3.$$

The first two terms are affine and hence convex. Therefore, Lemma D.3 with $\theta = \Gamma_0$ gives for any $h \in \mathbb{R}^d$

$$\psi_k(z_k + h) \geq \psi_k^* + \frac{\Gamma_0}{6}\|h\|^3. \tag{87}$$

Next, using assumption of this Lemma (83) and Lemma D.2,

$$\begin{aligned}
\psi_{k+1}^* &= \min_{x\in\mathbb{R}^d} \psi_{k+1}(x) = \min_{h\in\mathbb{R}^d} \left\{\psi_k(z_k + h) + a_{k+1}\ell_{w_{k+1}}(z_k + h)\right\} \\
&= a_{k+1} f(w_{k+1}) + \min_{h\in\mathbb{R}^d} \left\{\psi_k(z_k + h) + a_{k+1}\langle \nabla f(w_{k+1}), z_k + h - w_{k+1}\rangle\right\} \\
&\overset{(87)}{\geq} \psi_k^* + a_{k+1} f(w_{k+1}) + a_{k+1}\langle \nabla f(w_{k+1}), z_k - w_{k+1}\rangle \\
&\quad + \min_{h\in\mathbb{R}^d} \left\{\frac{\Gamma_0}{6}\|h\|^3 + a_{k+1}\langle \nabla f(w_{k+1}), h\rangle\right\} \\
&\overset{(83)}{\geq} A_k f(w_k) + a_{k+1} f(w_{k+1}) + a_{k+1}\langle \nabla f(w_{k+1}), z_k - w_{k+1}\rangle \\
&\quad + \min_{h\in\mathbb{R}^d} \left\{\frac{\Gamma_0}{6}\|h\|^3 + a_{k+1}\langle \nabla f(w_{k+1}), h\rangle\right\} \\
&\overset{(76)}{=} A_k f(w_k) + a_{k+1} f(w_{k+1}) + a_{k+1}\langle \nabla f(w_{k+1}), z_k - w_{k+1}\rangle - \frac{2\sqrt{2}}{3}\frac{a_{k+1}^{3/2}\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\Gamma_0}}.
\end{aligned} \tag{88}$$

Relations (77) imply

$$A_{k+1}v_k = A_k w_k + a_{k+1}z_k, \qquad A_{k+1}w_{k+1} = A_k w_{k+1} + a_{k+1}w_{k+1},$$
$$a_{k+1}(z_k - w_{k+1}) = A_{k+1}(v_k - w_{k+1}) - A_k(w_k - w_{k+1}). \tag{89}$$

Consequently,

$$\begin{aligned} &A_k f(w_k) + a_{k+1} f(w_{k+1}) + a_{k+1}\langle \nabla f(w_{k+1}), z_k - w_{k+1}\rangle \\ &\overset{(89)}{=} A_{k+1} f(w_{k+1}) + A_{k+1}\langle \nabla f(w_{k+1}), v_k - w_{k+1}\rangle \\ &\quad + A_k\left(f(w_k) - f(w_{k+1}) - \langle \nabla f(w_{k+1}), w_k - w_{k+1}\rangle\right) \\ &\geq A_{k+1} f(w_{k+1}) + A_{k+1}\langle \nabla f(w_{k+1}), v_k - w_{k+1}\rangle, \end{aligned} \tag{90}$$

where the inequality follows from convexity of f. By (77),

$$\frac{2\sqrt{2}}{3}\frac{a_{k+1}^{3/2}}{\sqrt{\Gamma_0}} = \frac{A_{k+1}}{9\sqrt{\mathcal{M}_k}}.$$

Substituting this identity and (90) into (88) proves (84). Under the hypotheses of Lemma D.5, substituting (81) into (84) gives

$$\psi_{k+1}^* \geq A_{k+1} f(w_{k+1}) + \left(\frac{2A_{k+1}}{9\sqrt{\mathcal{M}_k}} - \frac{A_{k+1}}{9\sqrt{\mathcal{M}_k}}\right)\|\nabla f(w_{k+1})\|^{3/2},$$

which proves (85). Dropping the nonnegative second term proves (86). □

Lemma D.7. *For every $k \geq 0$,*

$$\psi_k(x^*) \leq A_k f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}, \tag{91}$$

where $\Delta_0 := f(x_0) - f^*$. Moreover, under (70)*,*

$$\Delta_0 + \frac{\Gamma_0 R^3}{3} \leq \frac{2}{3}\Gamma_0 \bar{R}^3. \tag{92}$$

Proof. By convexity, $\ell_w(x^*) \leq f^*$ for every w. Since $A_k = 1 + \sum_{i=0}^{k-1} a_{i+1}$,

$$\psi_k(x^*) = f(x_0) + \frac{\Gamma_0}{3}R^3 + \sum_{i=0}^{k-1} a_{i+1}\ell_{w_{i+1}}(x^*)$$
$$\leq f(x_0) + \frac{\Gamma_0}{3}R^3 + (A_k - 1)f^* = A_k f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}.$$

Finally, (70) gives $\Delta_0 \leq \Gamma_0 \bar{R}^3/3$, and $R \leq \bar{R}$ gives $\Gamma_0 R^3/3 \leq \Gamma_0 \bar{R}^3/3$, proving the statement. □

D.1.4 Explicit Choice of Regularization Parameter

Fix an iteration k and suppose that (83) holds. For $\mathcal{M} > 0$, let $\alpha(\mathcal{M}) \in (0,1)$ solve

$$\frac{\alpha(\mathcal{M})^3}{1 - \alpha(\mathcal{M})} = \frac{\Gamma_k}{72\mathcal{M}}. \tag{93}$$

Define

$$A_+(\mathcal{M}) := \frac{A_k}{1 - \alpha(\mathcal{M})}, \qquad a(\mathcal{M}) := \alpha(\mathcal{M})A_+(\mathcal{M}),$$
$$v(\mathcal{M}) := (1 - \alpha(\mathcal{M}))\, w_k + \alpha(\mathcal{M}) z_k, \tag{94}$$

and let $w(\mathcal{M})$ be the output of the cubic subproblem (78) with base point $v(\mathcal{M})$ and regularization parameter $\mathcal{M}$. Finally, set

$$\psi_+(x;\mathcal{M}) := \psi_k(x) + a(\mathcal{M})\ell_{w(\mathcal{M})}(x), \qquad \psi_+^*(\mathcal{M}) := \min_{x\in\mathbb{R}^d} \psi_+(x;\mathcal{M}).$$

We call $\mathcal{M}$ admissible if $\mathcal{M} \geq L(w(\mathcal{M}))$. The proof of Lemma D.6 directly applies to these quantities, since the weight identities in Lemma D.4 remain valid. Thus, every admissible $\mathcal{M}$ satisfies

$$\psi_+^*(\mathcal{M}) \geq A_+(\mathcal{M}) f(w(\mathcal{M})) + \frac{A_+(\mathcal{M})}{9}\frac{\|\nabla f(w(\mathcal{M}))\|^{3/2}}{\sqrt{\mathcal{M}}}. \tag{95}$$

It remains to verify admissibility for the explicit choice in Algorithm 4.

Lemma D.8. *The mappings $\mathcal{M} \mapsto \|w(\mathcal{M}) - v(\mathcal{M})\|$ and $\mathcal{M} \mapsto \mathcal{M} - L(w(\mathcal{M}))$ are continuous on $(0, +\infty)$, and*

$$\|w(\mathcal{M}) - v(\mathcal{M})\| \to 0, \qquad \mathcal{M} - L(w(\mathcal{M})) \to +\infty \quad \text{as } \mathcal{M} \to +\infty. \tag{96}$$

Proof. The mapping $\alpha \mapsto \alpha^3/(1-\alpha)$ is continuous and strictly increasing on $(0, 1)$. Hence $\alpha(\mathcal{M})$ and $v(\mathcal{M})$ are continuous.

Let $s(\mathcal{M}) := w(\mathcal{M}) - v(\mathcal{M})$. For any sequence $\mathcal{M}_j \to \mathcal{M} > 0$, the optimality condition of the cubic subproblem and $\nabla^2 f(v(\mathcal{M}_j)) \succeq 0$ give

$$\|s(\mathcal{M}_j)\|^2 \leq \frac{2\|\nabla f(v(\mathcal{M}_j))\|}{15\mathcal{M}_j}. \tag{97}$$

By (94) $\alpha(\mathcal{M}_j) \to \alpha(\mathcal{M})$ as $\mathcal{M}_j \to \mathcal{M}$, therefore $v(\mathcal{M}_j) \to v(\mathcal{M})$. Since f is twice continuously differentiable $\nabla f(v(\mathcal{M}_j)) \to \nabla f(v(\mathcal{M}))$. Thus $\{\nabla f(v(\mathcal{M}_j))\}$ is bounded and $\{s(\mathcal{M}_j)\}$ is bounded as well.

Next, we show that $s(\mathcal{M}_j) \to s(\mathcal{M})$. By Bolzano-Weierstrass theorem bounded sequence $\{s(\mathcal{M}_j)\}$ has a convergent subsequence $\{s(\mathcal{M}_{j_l})\} \to \hat{s}$. Taking the limit in the optimality condition of cubic subproblem shows that $\hat{s}$ satisfies this optimality condition corresponding to $\mathcal{M}$. Since cubic subproblem is strictly convex, it has unique minimizer and $\hat{s} = s(\mathcal{M})$. The same argument applies to any convergent subsequence of $\{s(\mathcal{M}_j)\}$. Thus, $s(\mathcal{M}_j) \to s(\mathcal{M})$ since $s(\mathcal{M}_j)$ is bounded sequence with only one accumulation point. Therefore $w(\mathcal{M}_j) \to w(\mathcal{M})$ and thus $w(\mathcal{M})$ is continuous. Since $L(\cdot)$ is continuous, the mappings $\|s(\mathcal{M})\|$ and $\mathcal{M} - L(w(\mathcal{M}))$ are continuous as well.

The estimate (97) gives $s(\mathcal{M}) \to 0$ as $\mathcal{M} \to +\infty$, while $\alpha(\mathcal{M}) \to 0$ by (93) and $v(\mathcal{M}) \to w_k$ by (94). Hence $w(\mathcal{M}) \to w_k$, so $L(w(\mathcal{M}))$ remains bounded. This gives us (96). □

Lemma D.9. *Suppose that* (GH) *and* (83) *hold. Then the parameter*

$$\mathcal{M}_k = \max\left\{2M_0, 12\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_k\right\}$$

used in Algorithm 4 satisfies

$$\mathcal{M}_k \geq L(w_{k+1}) = M_0 + M_1\|\nabla f(w_{k+1})\|. \tag{98}$$

Consequently, estimates (85) *and* (86) *hold for the actual iteration.*

Proof. Define

$$\mathcal{U}_k := \{\mathcal{M} > 0 : \mathcal{M} < L(w(\mathcal{M}))\}.$$

If $\mathcal{U}_k = \emptyset$, the statement is immediate. Otherwise, let $\mathcal{M}_k^\dagger := \sup \mathcal{U}_k$. Lemma D.8 shows that $\mathcal{M}_k^\dagger < +\infty$. By continuity,

$$\mathcal{M}_k^\dagger = L(w(\mathcal{M}_k^\dagger)), \tag{99}$$

and every $\mathcal{M} \geq \mathcal{M}_k^\dagger$ is admissible.

Set

$$w^\dagger := w(\mathcal{M}_k^\dagger), \qquad g^\dagger := \nabla f(w^\dagger), \qquad A_+^\dagger := A_+(\mathcal{M}_k^\dagger).$$

Applying (95) at the admissible boundary and evaluating $\psi_+(\cdot; \mathcal{M}_k^\dagger)$ at x^*, we obtain

$$A_+^\dagger f(w^\dagger) + \frac{A_+^\dagger}{9}\frac{\|g^\dagger\|^{3/2}}{\sqrt{\mathcal{M}_k^\dagger}} \leq \psi_+^*(\mathcal{M}_k^\dagger) \leq \psi_+(x^*; \mathcal{M}_k^\dagger) \leq A_+^\dagger f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}.$$

Since $f(w^\dagger) \geq f^*$, $A_+^\dagger \geq A_k$, and $\Gamma_k = \Gamma_0/A_k$, equation (92) gives

$$\frac{\|g^\dagger\|^{3/2}}{\sqrt{\mathcal{M}_k^\dagger}} \leq 6\Gamma_k\bar{R}^3. \tag{100}$$

For every $g \in \mathbb{R}^d$,

$$M_0 + M_1\|g\| \le \max\left\{2M_0, 2\sqrt{2}M_1^{3/2}\frac{\|g\|^{3/2}}{\sqrt{M_0 + M_1\|g\|}}\right\}. \tag{101}$$

Indeed, the first term applies when $M_1\|g\| \le M_0$. Otherwise, $M_0 + M_1\|g\| \le 2M_1\|g\|$ and

$$\frac{\|g\|^{3/2}}{\sqrt{M_0 + M_1\|g\|}} \ge \frac{\|g\|}{\sqrt{2M_1}}.$$

Applying this bound with $g = g^\dagger$ and using (99) and (100), we obtain

$$\mathcal{M}_k^\dagger \le \max\left\{2M_0, 2\sqrt{2}M_1^{3/2}\frac{\|g^\dagger\|^{3/2}}{\sqrt{\mathcal{M}_k^\dagger}}\right\} \le \max\left\{2M_0, 12\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_k\right\} = \mathcal{M}_k.$$

Therefore $\mathcal{M}_k$ is admissible and (98) holds. The last statement follows from Lemma D.6. □

D.1.5 Estimating Sequence Bounds

Lemma D.10. *Suppose that* (GH) *holds. For every $k \ge 0$, the parameter $\mathcal{M}_k$ in Algorithm 4 satisfies* (98)*, and*

$$\psi_k^* \ge A_k f(w_k). \tag{102}$$

Moreover,

$$\psi_{k+1}^* \ge A_{k+1}f(w_{k+1}) + \frac{A_{k+1}}{9}\frac{\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\mathcal{M}_k}}. \tag{103}$$

Proof. For $k = 0$,

$$\psi_0^* = \min_{x\in\mathbb{R}^d}\left\{f(x_0) + \frac{\Gamma_0}{3}\|x - x_0\|^3\right\} = f(x_0) = A_0 f(w_0).$$

Suppose that (102) holds at iteration k. Theorem D.9 gives (98). Hence, Lemma D.6 gives (103) and

$$\psi_{k+1}^* \ge A_{k+1}f(w_{k+1}),$$

which closes the induction. □

Lemma D.11. *Suppose that* (GH) *holds. For every $k \ge 0$,*

$$f(w_k) - f^* \le \frac{\Delta_0 + \Gamma_0 R^3/3}{A_k} \le \frac{2}{3}\Gamma_k\bar{R}^3. \tag{104}$$

Moreover, every cubic output satisfies

$$f(w_{k+1}) - f^* + \frac{1}{9}\frac{\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\mathcal{M}_k}} \le \frac{\Delta_0 + \Gamma_0 R^3/3}{A_{k+1}} \le \frac{2}{3}\Gamma_{k+1}\bar{R}^3. \tag{105}$$

Proof. Combining (102) and (91),

$$A_k f(w_k) \overset{(102)}{\le} \psi_k^* \le \psi_k(x^*) \overset{(91)}{\le} A_k f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}.$$

Therefore,

$$f(w_k) - f^* \le \frac{\Delta_0 + \Gamma_0 R^3/3}{A_k} \overset{(92)}{\le} \frac{2\Gamma_0\bar{R}^3}{3A_k} \overset{(77)}{=} \frac{2}{3}\Gamma_k\bar{R}^3.$$

Similarly,

$$A_{k+1}f(w_{k+1}) + \frac{A_{k+1}}{9}\frac{\|\nabla f(w_{k+1})\|^{3/2}}{\sqrt{\mathcal{M}_k}} \overset{(103)}{\le} \psi_{k+1}^* \le \psi_{k+1}(x^*) \overset{(91)}{\le} A_{k+1}f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}.$$

Dividing by A_{k+1} and using (92) and (77) proves (105). □

D.1.6 CONVERGENCE RATES

Since $\alpha_k \in (0,1)$ and $\Gamma_0 > 0$,

$$0 < \Gamma_{k+1} = (1-\alpha_k)\Gamma_k < \Gamma_k.$$

Thus, $\{\Gamma_k\}_{k\geq 0}$ is strictly decreasing. Define the transition level

$$\Gamma_{\mathrm{tr}} := \begin{cases} \dfrac{M_0}{6\sqrt{2}M_1^{3/2}\bar{R}^3}, & M_1 > 0, \\ +\infty, & M_1 = 0. \end{cases} \tag{106}$$

Lemma D.12 (Far regime)**.** *Suppose that $M_1 > 0$ and $\Gamma_k > \Gamma_{\mathrm{tr}}$. Then*

$$\alpha_k \geq \frac{1}{12\sqrt[6]{2}\left(1+\sqrt{M_1}\bar{R}\right)}, \tag{107}$$

and

$$\Gamma_{k+1} \leq \exp\left(-\frac{1}{12\sqrt[6]{2}\left(1+\sqrt{M_1}\bar{R}\right)}\right)\Gamma_k. \tag{108}$$

Proof. In this regime,

$$\mathcal{M}_k = 12\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_k,$$

and hence

$$\frac{\alpha_k^3}{1-\alpha_k} = \frac{1}{864\sqrt{2}M_1^{3/2}\bar{R}^3}.$$

If $\alpha_k > 1/2$, then (107) holds immediately. If $\alpha_k \leq 1/2$, then

$$\alpha_k^3 = \tfrac{1-\alpha_k}{864\sqrt{2}M_1^{3/2}\bar{R}^3} \geq \tfrac{1}{1728\sqrt{2}M_1^{3/2}\bar{R}^3}, \quad \alpha_k \geq \tfrac{1}{12\sqrt[6]{2}\sqrt{M_1}\bar{R}}.$$

Combining the two cases proves (107). Finally, $1-\alpha_k \leq \exp(-\alpha_k)$ proves (108). □

Lemma D.13 (Near regime)**.** *Suppose that* (GH) *holds and $\Gamma_K \leq \Gamma_{\mathrm{tr}}$. Then $\mathcal{M}_j = 2M_0$ for every $j \geq K$, and*

$$A_{K+n}^{1/3} \geq A_K^{1/3} + \frac{n}{3}\left(\frac{\Gamma_0}{144M_0}\right)^{1/3}, \qquad n \geq 1. \tag{109}$$

Consequently,

$$\Gamma_{K+n} \leq \frac{3888M_0}{n^3}, \qquad n \geq 1, \tag{110}$$

and

$$f(w_{K+n}) - f^* \leq \frac{2592M_0\bar{R}^3}{n^3}, \qquad n \geq 1. \tag{111}$$

Proof. The monotonicity of Γ_k and (106) give $\mathcal{M}_j = 2M_0$ for every $j \geq K$. Therefore, (77) gives

$$a_{j+1}^3 = \frac{\Gamma_0 A_{j+1}^2}{144M_0}.$$

Using $A_{j+1} - A_j = a_{j+1}$,

$$A_{j+1}^{1/3} - A_j^{1/3} = \frac{a_{j+1}}{A_{j+1}^{2/3} + A_{j+1}^{1/3}A_j^{1/3} + A_j^{2/3}} \geq \frac{a_{j+1}}{3A_{j+1}^{2/3}} = \frac{1}{3}\left(\frac{\Gamma_0}{144M_0}\right)^{1/3}.$$

Summing over $j = K, \ldots, K+n-1$ proves (109). Hence,

$$A_{K+n} \geq \frac{n^3\Gamma_0}{3888M_0}, \quad \Gamma_{K+n} \overset{(77)}{=} \frac{\Gamma_0}{A_{K+n}} \leq \frac{3888M_0}{n^3},$$

which proves (110). Combining (104) and (110) proves (111). □

Theorem D.14. *Suppose that f is convex,* (GH) *and* (70) *hold. For every $\varepsilon > 0$, Algorithm 4 returns an iterate w_N satisfying*

$$f(w_N) - f^* \le \varepsilon$$

after at most

$$N \le \left\lceil 12\sqrt[6]{2}\left(1+\sqrt{M_1}\bar{R}\right)\log\max\left\{1,\min\left\{\frac{6\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_0}{M_0},\frac{2\Gamma_0\bar{R}^3}{3\varepsilon}\right\}\right\}\right\rceil + \left\lceil\left(\frac{2592M_0\bar{R}^3}{\varepsilon}\right)^{1/3}\right\rceil \tag{112}$$

iterations.

Proof. Define

$$\Gamma_\varepsilon := \frac{3\varepsilon}{2\bar{R}^3}.$$

By (104), if $\Gamma_k \le \Gamma_\varepsilon$, then

$$f(w_k) - f^* \le \frac{2}{3}\Gamma_k\bar{R}^3 \le \frac{2}{3}\Gamma_\varepsilon\bar{R}^3 = \varepsilon.$$

Suppose first that $M_1 > 0$, and set

$$N_{\text{far}} := \left\lceil 12\sqrt[6]{2}\left(1+\sqrt{M_1}\bar{R}\right)\log\max\left\{1,\frac{\Gamma_0}{\max\{\Gamma_{\text{tr}},\Gamma_\varepsilon\}}\right\}\right\rceil.$$

Suppose that $\Gamma_k > \max\{\Gamma_{\text{tr}},\Gamma_\varepsilon\}$ for every $k = 0,\dots,N_{\text{far}}-1$. Then (108) applies at each of these iterations, and

$$\Gamma_{N_{\text{far}}} \le \exp\left(-\frac{N_{\text{far}}}{12\sqrt[6]{2}\left(1+\sqrt{M_1}\bar{R}\right)}\right)\Gamma_0 \le \max\{\Gamma_{\text{tr}},\Gamma_\varepsilon\}.$$

Therefore, after at most N_{far} iterations, either $\Gamma_k \le \Gamma_\varepsilon$, or $\Gamma_k \le \Gamma_{\text{tr}}$ and the near regime begins. Moreover,

$$\frac{\Gamma_0}{\Gamma_{\text{tr}}} = \frac{6\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_0}{M_0}, \qquad \frac{\Gamma_0}{\Gamma_\varepsilon} = \frac{2\Gamma_0\bar{R}^3}{3\varepsilon},$$
$$\frac{\Gamma_0}{\max\{\Gamma_{\text{tr}},\Gamma_\varepsilon\}} = \min\left\{\frac{6\sqrt{2}M_1^{3/2}\bar{R}^3\Gamma_0}{M_0},\frac{2\Gamma_0\bar{R}^3}{3\varepsilon}\right\}.$$

If the near regime is needed, Lemma D.13 shows that

$$N_{\text{near}} := \left\lceil\left(\frac{2592M_0\bar{R}^3}{\varepsilon}\right)^{1/3}\right\rceil$$

additional iterations suffice. Hence, $N \le N_{\text{far}} + N_{\text{near}}$, which proves (112).

If $M_1 = 0$, then $\Gamma_{\text{tr}} = +\infty$ and the near regime starts at $K = 0$. The first term in (112) is zero, while Lemma D.13 gives the second term. □

D.2 Adaptive Method under (LH) with Exact Hessian

Algorithm 4 uses the problem parameters in its regularization coefficient and requires an initial scale satisfying (70). We now give one implementation that does not use M_0, M_1, R, ρ, or f^*.

Initialization step. Our first goal is to construct a computable initial scale Γ_0 satisfying

$$\Gamma_0 R^3 \ge 3\left(f(x_0) - f^*\right). \tag{113}$$

without knowing R and f^*. We construct them by solving the cubic subproblem (115) with backtracking regularization parameter $\tilde{\mathcal{M}}$ until condition (116) holds. The whole warm-up backtracking procedure is listed in lines 4-7 of Algorithm 5. The following lemma shows that the backtracking procedure terminates after a finite number of iterations and provides an exact way of satisfying initial condition (113).

Lemma D.15. *Suppose that either* (GH) *or* (LH) *holds. If $\nabla f(x_0) = 0$, Algorithm 5 returns the optimal point x_0. Otherwise, the initialization backtracking (lines 4-7 of Algorithm 5) satisfies* (116) *after at most $\left\lceil \log_2 \max\left\{1, \frac{T_0}{\tilde{\mathcal{M}}_0}\right\}\right\rceil$ rejected trials, where*

$$T_0 = \begin{cases} M_0 + 2M_1\|\nabla f(x_0)\|, & \text{under (GH)}, \\ \max\left\{M_0 + 2M_1\|\nabla f(x_0)\|, \frac{2\|\nabla f(x_0)\|}{15\rho^2}\right\}, & \text{under (LH)}. \end{cases} \tag{114}$$

For the accepted step, $\Gamma_0 := \frac{3\|\nabla f(x_0)\|^3}{(f(x_0)-f(\tilde{x}))^2}$ is well defined and satisfies initial condition (113).

Algorithm 5 Fully adaptive accelerated cubic regularization

Input: $x_0 \in \mathbb{R}^d$ and $\tilde{\mathcal{M}}_0 > 0$.
If $\nabla f(x_0) = 0$, return x_0.
Set $\tilde{\mathcal{M}} := \tilde{\mathcal{M}}_0$.
Compute

$$\tilde{s} := \operatorname*{argmin}_{s\in\mathbb{R}^d}\left\{\langle \nabla f(x_0), s\rangle + \tfrac{1}{2}\left\langle \nabla^2 f(x_0)s, s\right\rangle + \tfrac{15\tilde{\mathcal{M}}}{6}\|s\|^3\right\}, \quad \tilde{x} := x_0 + \tilde{s}. \tag{115}$$

while condition

$$\langle \nabla f(\tilde{x}), -\tilde{s}\rangle \geq \tfrac{2}{9\sqrt{\tilde{\mathcal{M}}}}\|\nabla f(\tilde{x})\|^{3/2}. \tag{116}$$

does not hold **do**
  Set $\tilde{\mathcal{M}} := 2\tilde{\mathcal{M}}$.
  Recompute $\tilde{s}$ and $\tilde{x}$ according to (115).
Set $\Gamma_0 := \frac{3\|\nabla f(x_0)\|^3}{(f(x_0)-f(\tilde{x}))^2}$, $A_0 := 1$, $w_0 := x_0$, $z_0 := x_0$, and $\psi_0(x) := f(x_0) + \frac{\Gamma_0}{3}\|x - x_0\|^3$.
Set $\mathcal{M}_0 := \Gamma_0$.
for $k \geq 0$ **do**
  Find the unique $\alpha_k \in (0,1)$ satisfying $\frac{\alpha_k^3}{1-\alpha_k} = \frac{\Gamma_k}{72\mathcal{M}_k}$.
  Set A_{k+1}, a_{k+1}, and Γ_{k+1} according to (72).
  Set $v_k := (1-\alpha_k)w_k + \alpha_k z_k$.
  Compute s_k and w_{k+1} according to (73) with regularization coefficient $\mathcal{M}_k$.
  while condition

$$\langle \nabla f(w_{k+1}), v_k - w_{k+1}\rangle \geq \tfrac{2}{9\sqrt{\mathcal{M}_k}}\|\nabla f(w_{k+1})\|^{3/2}. \tag{117}$$

  does not hold **do**
    Set $\mathcal{M}_k := 2\mathcal{M}_k$.
    Recompute α_k, A_{k+1}, a_{k+1}, Γ_{k+1}, and v_k as above.
    Recompute w_{k+1} according to (73) with regularization coefficient $\mathcal{M}_k$.
  Set $\psi_{k+1}(x) := \psi_k(x) + a_{k+1}\ell_{w_{k+1}}(x)$ and compute $z_{k+1} := \operatorname{argmin}_{x\in\mathbb{R}^d}\psi_{k+1}(x)$.
  Set $\mathcal{M}_{k+1} := \frac{\Gamma_{k+1}}{\Gamma_k}\mathcal{M}_k$.

For the statement below, define

$$B := \begin{cases} 12\sqrt{2}M_1^{3/2}R^3, & \text{under (GH)}, \\ \max\left\{12\sqrt{2}M_1^{3/2}R^3, \frac{2R^3}{9\rho^3}\right\}, & \text{under (LH)}. \end{cases} \tag{118}$$

Theorem D.16. *Suppose that f is convex and either* (GH) *or* (LH) *holds. All backtracking loops in Algorithm 5 terminate. For every $\varepsilon > 0$, under* (GH)*, the method returns an iterate w_N satisfying $f(w_N) - f^* \leq \varepsilon$ for*

$$N = \mathcal{O}\left(\left(\tfrac{M_0R^3}{\varepsilon}\right)^{1/3}\right) + \widetilde{\mathcal{O}}\left(1 + \sqrt{M_1}R\right). \tag{119}$$

Under (LH)*, the corresponding bound is*

$$N = \mathcal{O}\left(\left(\tfrac{M_0R^3}{\varepsilon}\right)^{1/3}\right) + \widetilde{\mathcal{O}}\left(1 + \sqrt{M_1}R + \tfrac{R}{\rho}\right). \tag{120}$$

The number of rejected initialization trials is bounded as in Lemma D.15. Before returning w_N, the main backtracking loops reject at most $\left\lceil \log_2 \max\left\{1, B, \frac{4M_0R^3}{3\varepsilon}\right\}\right\rceil$ trials in total, where B is defined in (118)*.*

D.2.1 Proof of Lemma D.15

Proof. Suppose that $\nabla f(x_0) \neq 0$ and fix a trial value $\tilde{\mathcal{M}} > 0$. The optimality condition for (115) is

$$\nabla f(x_0) + \nabla^2 f(x_0)\tilde{s} + \frac{15\tilde{\mathcal{M}}}{2}\|\tilde{s}\|\tilde{s} = 0. \tag{121}$$

Since f is convex, $\nabla^2 f(x_0) \succeq 0$. Taking the inner product of (121) with $\tilde{s}$, we obtain

$$\frac{15\tilde{\mathcal{M}}}{2}\|\tilde{s}\|^3 \leq -\langle \nabla f(x_0), \tilde{s}\rangle \leq \|\nabla f(x_0)\|\|\tilde{s}\|,$$

and hence

$$\|\tilde{s}\|^2 \leq \frac{2\|\nabla f(x_0)\|}{15\tilde{\mathcal{M}}}. \tag{122}$$

Let $\tilde{\mathcal{M}} \geq T_0$, defined in (114). Under (LH), equations (114) and (122) give

$$\|\tilde{s}\|^2 \leq \frac{2\|\nabla f(x_0)\|}{15\tilde{\mathcal{M}}} \leq \rho^2.$$

Therefore, the Taylor remainder at x_0 can be applied under both assumptions. Using (121), we get

$$\begin{aligned}\|\nabla f(\tilde{x})\| &\leq \|\nabla f(\tilde{x}) - \nabla f(x_0) - \nabla^2 f(x_0)\tilde{s}\| + \frac{15\tilde{\mathcal{M}}}{2}\|\tilde{s}\|^2 \\ &\leq \frac{M_0 + M_1\|\nabla f(x_0)\| + 15\tilde{\mathcal{M}}}{2}\|\tilde{s}\|^2 \\ &\leq \left(1 + \frac{M_0 + M_1\|\nabla f(x_0)\|}{15\tilde{\mathcal{M}}}\right)\|\nabla f(x_0)\| \leq \frac{16}{15}\|\nabla f(x_0)\|.\end{aligned}$$

Consequently,

$$M_0 + M_1\|\nabla f(\tilde{x})\| \leq M_0 + 2M_1\|\nabla f(x_0)\| \leq \tilde{\mathcal{M}}.$$

Since $\|\tilde{s}\| \leq \rho$ under (LH), the hypotheses of Lemma D.5 hold under both assumptions. Therefore, (81) gives

$$\langle \nabla f(\tilde{x}), -\tilde{s}\rangle \geq \frac{2}{9\sqrt{\tilde{\mathcal{M}}}}\|\nabla f(\tilde{x})\|^{3/2},$$

which is condition (116). Hence, every trial with $\tilde{\mathcal{M}} \geq T_0$ is accepted.

Starting from the input value $\tilde{\mathcal{M}}_0$, the j-th trial uses $2^j\tilde{\mathcal{M}}_0$. For

$$j = \left\lceil \log_2 \max\left\{1, \frac{T_0}{\tilde{\mathcal{M}}_0}\right\}\right\rceil,$$

we have $2^j\tilde{\mathcal{M}}_0 \geq T_0$. Therefore, the backtracking rejects at most j trials.

It remains to verify (113). If $\nabla f(\tilde{x}) \neq 0$, condition (116) and convexity give

$$f(x_0) - f(\tilde{x}) \geq \langle \nabla f(\tilde{x}), x_0 - \tilde{x}\rangle > 0.$$

If $\nabla f(\tilde{x}) = 0$, then $\tilde{x}$ is optimal. Since $\nabla f(x_0) \neq 0$, in this case we also have $f(x_0) - f(\tilde{x}) > 0$. Moreover, by convexity,

$$0 < f(x_0) - f(\tilde{x}) \leq f(x_0) - f^* \leq \langle \nabla f(x_0), x_0 - x^*\rangle \leq \|\nabla f(x_0)\|R.$$

Thus,

$$0 < \frac{f(x_0) - f(\tilde{x})}{\|\nabla f(x_0)\|} \leq R.$$

Using this estimate one more time, we obtain

$$\begin{aligned}f(x_0) - f^* \leq \|\nabla f(x_0)\|R &= (f(x_0) - f(\tilde{x}))\,\frac{\|\nabla f(x_0)\|R}{f(x_0) - f(\tilde{x})} \\ &\leq (f(x_0) - f(\tilde{x}))\left(\frac{\|\nabla f(x_0)\|R}{f(x_0) - f(\tilde{x})}\right)^3 = \frac{\Gamma_0 R^3}{3}.\end{aligned}$$

This proves (113). □

D.2.2 Proof of Theorem D.16

Fix an iteration k. For a trial coefficient $\mathcal{M} > 0$, Algorithm 5 computes α_k, A_{k+1}, a_{k+1}, v_k, and w_{k+1} in the same way as Algorithm 4 with $\mathcal{M}_k = \mathcal{M}$. Hence, under the induction assumption (83), Lemma D.6 gives (84) for every trial. If the trial passes (117), which is (81) for $v = v_k$ and $\mathcal{M} = \mathcal{M}_k$, then the proof of Lemma D.6 gives (85) and (86).

We now prove that (117) holds for every sufficiently large trial coefficient. Let B be defined by (118).

Lemma D.17. *Suppose that either* (GH) *or* (LH) *holds and that* (83) *is satisfied. Every trial with* $\mathcal{M} \geq \max\{2M_0, B\Gamma_k\}$ *satisfies* (117).

Proof. By (113), condition (70) holds with $\bar{R} = R$. Under (GH), we set $\rho = +\infty$. The trial with coefficient $\mathcal{M}$ coincides with the virtual step $w(\mathcal{M})$ defined after (94). Set $r(\mathcal{M}) := \|w(\mathcal{M}) - v(\mathcal{M})\|$ and

$$\mathcal{U}_k := \{\mathcal{M} > 0 : r(\mathcal{M}) > \rho \text{ or } \mathcal{M} < L(w(\mathcal{M}))\}.$$

For $\mathcal{M} \notin \mathcal{U}_k$, the hypotheses of Lemma D.5 hold for $v = v(\mathcal{M})$, and (81) gives (117). Thus, it suffices to show that every $\mathcal{M} \geq \max\{2M_0, B\Gamma_k\}$ satisfies $\mathcal{M} \notin \mathcal{U}_k$.

If $\mathcal{U}_k = \emptyset$, there is nothing to prove. Otherwise, Lemma D.8 shows that $\mathcal{U}_k$ is open and bounded above. Let $\mathcal{M}_k^\dagger := \sup \mathcal{U}_k < +\infty$. Then $\mathcal{M} \notin \mathcal{U}_k$ for every $\mathcal{M} \geq \mathcal{M}_k^\dagger$. By continuity,

$$r^\dagger \leq \rho, \qquad \mathcal{M}_k^\dagger \geq L(w^\dagger), \qquad r^\dagger = \rho \text{ or } \mathcal{M}_k^\dagger = L(w^\dagger), \tag{123}$$

where a dagger denotes evaluation at $\mathcal{M}_k^\dagger$ and $g^\dagger := \nabla f(w^\dagger)$.

By (123), the hypotheses of Lemma D.5 hold at $\mathcal{M}_k^\dagger$. Hence, (79), (80), and (81) give

$$\langle g^\dagger, v^\dagger - w^\dagger \rangle - \frac{\|g^\dagger\|^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}} \geq 6\mathcal{M}_k^\dagger (r^\dagger)^3 - \frac{\left(9\mathcal{M}_k^\dagger (r^\dagger)^2\right)^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}} = 3\mathcal{M}_k^\dagger (r^\dagger)^3,$$
$$\langle g^\dagger, v^\dagger - w^\dagger \rangle - \frac{\|g^\dagger\|^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}} \geq \frac{\|g^\dagger\|^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}}.$$

Lemma D.6 applies to the virtual step, as explained before (95). Therefore, (84) and evaluation of $\psi_+(\cdot; \mathcal{M}_k^\dagger)$ at x^* give

$$A_+^\dagger f(w^\dagger) + A_+^\dagger \max\left\{3\mathcal{M}_k^\dagger (r^\dagger)^3, \frac{\|g^\dagger\|^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}}\right\} \leq \psi_+^*(\mathcal{M}_k^\dagger) \leq \psi_+(x^*; \mathcal{M}_k^\dagger) \leq A_+^\dagger f^* + \Delta_0 + \frac{\Gamma_0 R^3}{3}.$$

Since $f(w^\dagger) \geq f^*$, $A_+^\dagger \geq A_k$, and $\Gamma_k = \Gamma_0 / A_k$, equation (92) with $\bar{R} = R$ gives

$$\max\left\{3\mathcal{M}_k^\dagger (r^\dagger)^3, \frac{\|g^\dagger\|^{3/2}}{9\sqrt{\mathcal{M}_k^\dagger}}\right\} \leq \frac{2}{3}\Gamma_k R^3. \tag{124}$$

If $r^\dagger = \rho$, then (124) gives

$$\mathcal{M}_k^\dagger \leq \frac{2R^3}{9\rho^3}\Gamma_k \leq B\Gamma_k.$$

If $\mathcal{M}_k^\dagger = L(w^\dagger)$, then (101) and (124) give

$$\mathcal{M}_k^\dagger \leq \max\left\{2M_0, 2\sqrt{2}M_1^{3/2}\frac{\|g^\dagger\|^{3/2}}{\sqrt{\mathcal{M}_k^\dagger}}\right\} \leq \max\left\{2M_0, 12\sqrt{2}M_1^{3/2}R^3\Gamma_k\right\} \leq \max\{2M_0, B\Gamma_k\}.$$

In both cases, every $\mathcal{M} \geq \max\{2M_0, B\Gamma_k\}$ satisfies $\mathcal{M} \geq \mathcal{M}_k^\dagger$, and hence $\mathcal{M} \notin \mathcal{U}_k$. □

Lemma D.18. *Suppose that either* (GH) *or* (LH) *holds. Every main backtracking loop in Algorithm 5 terminates. For every accepted iteration $k \geq 0$,*

$$\mathcal{M}_k \leq \max\{4M_0, \max\{1, 2B\}\,\Gamma_k\}\,. \tag{125}$$

For every $K \geq 1$, the total number of rejected main trials during iterations $k = 0, \ldots, K-1$ is at most

$$\left\lceil \log_2 \max\left\{1, B, \frac{2M_0}{\Gamma_{K-1}}\right\}\right\rceil. \tag{126}$$

Proof. At $k = 0$, (83) holds with equality. Suppose that it holds at the beginning of iteration k. Repeated doubling sends $\mathcal{M}_k$ to $+\infty$. Hence, Lemma D.17 shows that condition (117) is satisfied after finitely many trials. For the accepted trial, (86) gives (83) at iteration $k+1$. Thus, every backtracking loop terminates by induction.

It remains to prove (125). At $k = 0$, if the first trial is accepted, then

$$\mathcal{M}_0 = \Gamma_0 \leq \max\{1, 2B\}\,\Gamma_0.$$

If at least one trial is rejected, the coefficient immediately preceding the accepted one is $\mathcal{M}_0/2$. Lemma D.17 gives $\frac{\mathcal{M}_0}{2} < \max\{2M_0, B\Gamma_0\}$. Therefore, $\mathcal{M}_0 < \max\{4M_0, 2B\Gamma_0\}$, and (125) follows.

Suppose that (125) holds at iteration $k-1$. If the first trial is accepted at iteration k, then

$$\mathcal{M}_k = \frac{\Gamma_k}{\Gamma_{k-1}}\mathcal{M}_{k-1} \leq \frac{\Gamma_k}{\Gamma_{k-1}} \max\{4M_0, \max\{1, 2B\}\,\Gamma_{k-1}\} \leq \max\{4M_0, \max\{1, 2B\}\,\Gamma_k\}\,,$$

where we used $\Gamma_k \leq \Gamma_{k-1}$. If at least one trial is rejected, the same argument as for $k = 0$ proves the bound. This closes the induction.

Finally, the initialization and the update of $\mathcal{M}_k$ in Algorithm 5 imply that the ratio $\mathcal{M}_k/\Gamma_k$ is initially equal to one, remains unchanged between accepted iterations, and doubles after every rejected trial. Since Γ_k is decreasing, for every $k = 0, \ldots, K-1$,

$$\max\left\{B, \frac{2M_0}{\Gamma_k}\right\} \leq \max\left\{B, \frac{2M_0}{\Gamma_{K-1}}\right\}.$$

By Lemma D.17, every trial satisfying

$$\frac{\mathcal{M}_k}{\Gamma_k} \geq \max\left\{B, \frac{2M_0}{\Gamma_{K-1}}\right\}$$

is accepted. After J rejected main trials, the current ratio is 2^J. This proves (126). □

Lemma D.19. *Suppose that f is convex and either* (GH) *or* (LH) *holds. For every accepted iterate of Algorithm 5,*

$$f(w_k) - f^* \leq \frac{2}{3}\Gamma_k R^3. \tag{127}$$

Proof. The induction in the proof of Lemma D.18 shows that (83) holds for every accepted iteration. The argument in the proof of Lemma D.11, applied with $\bar{R} = R$, proves (127). □

For the scalar analysis, set

$$\widehat{B} := \max\{1, 2B\}\,, \qquad \Gamma_{\mathrm{tr}}^{\mathrm{ad}} := \frac{4M_0}{\widehat{B}}. \tag{128}$$

The following two lemmas separate the convergence of the fully adaptive method into the far and near regimes. This separation is used only in the analysis and does not require the algorithm to know the transition level.

Lemma D.20 (Far regime)**.** *Suppose that f is convex, either* (GH) *or* (LH) *holds, and $\Gamma_k > \Gamma_{\mathrm{tr}}^{\mathrm{ad}}$. Then*

$$\alpha_k \geq \frac{1}{2 + (144\widehat{B})^{1/3}}, \tag{129}$$

and

$$\Gamma_{k+1} \leq \exp\left(-\frac{1}{2 + (144\widehat{B})^{1/3}}\right)\Gamma_k. \tag{130}$$

Proof. By (128), $\widehat{B}\Gamma_k > 4M_0$. Therefore, equation (125) gives

$$\mathcal{M}_k \leq \widehat{B}\Gamma_k, \qquad \frac{\alpha_k^3}{1-\alpha_k} = \frac{\Gamma_k}{72\mathcal{M}_k} \geq \frac{1}{72\widehat{B}}.$$

If $\alpha_k \leq 1/2$, then

$$\alpha_k^3 \geq \frac{1}{144\widehat{B}}.$$

If $\alpha_k > 1/2$, the bound $\alpha_k > 1/2$ holds directly. Hence,

$$\alpha_k \geq \min\left\{\frac{1}{2}, \frac{1}{(144\widehat{B})^{1/3}}\right\} \geq \frac{1}{2 + (144\widehat{B})^{1/3}}.$$

This proves (129). Equation (130) follows from $\Gamma_{k+1} = (1-\alpha_k)\Gamma_k$ and $1 - t \leq \exp(-t)$. □

Lemma D.21 (Near regime)**.** *Suppose that f is convex, either* (GH) *or* (LH) *holds, and $\Gamma_K \leq \Gamma_{\mathrm{tr}}^{\mathrm{ad}}$. Then, for every $j \geq K$,*

$$\mathcal{M}_j \leq 4M_0. \tag{131}$$

Moreover,

$$A_{K+n}^{1/3} \geq A_K^{1/3} + \frac{n}{3}\left(\frac{\Gamma_0}{288M_0}\right)^{1/3}, \qquad n \geq 1, \tag{132}$$

and

$$\Gamma_{K+n} \leq \frac{7776M_0}{n^3}, \tag{133}$$

$$f(w_{K+n}) - f^* \leq \frac{5184M_0R^3}{n^3}, \qquad n \geq 1. \tag{134}$$

Proof. The sequence $\{\Gamma_k\}_{k\geq 0}$ is decreasing. Thus, $\widehat{B}\Gamma_j \leq 4M_0$ for every $j \geq K$, and equation (125) proves (131). By (77),

$$a_{j+1}^3 = \frac{\Gamma_0 A_{j+1}^2}{72\mathcal{M}_j} \geq \frac{\Gamma_0 A_{j+1}^2}{288M_0}.$$

Therefore,

$$\begin{aligned} A_{j+1}^{1/3} - A_j^{1/3} &= \frac{a_{j+1}}{A_{j+1}^{2/3} + A_{j+1}^{1/3}A_j^{1/3} + A_j^{2/3}} \\ &\geq \frac{a_{j+1}}{3A_{j+1}^{2/3}} \geq \frac{1}{3}\left(\frac{\Gamma_0}{288M_0}\right)^{1/3}. \end{aligned}$$

Summing this inequality proves (132). Hence,

$$A_{K+n} \geq \frac{n^3\Gamma_0}{7776M_0}, \qquad \Gamma_{K+n} = \frac{\Gamma_0}{A_{K+n}} \leq \frac{7776M_0}{n^3}.$$

This proves (133). Combining it with (127) proves (134). □

Proof of Theorem D.16. If $\nabla f(x_0) = 0$, then x_0 is optimal by convexity. Otherwise, Lemma D.15 shows that the initialization backtracking terminates, gives (113), and proves the stated bound on the number of rejected initialization trials.

Lemma D.19 gives (127). It remains to consider $0 < \varepsilon < f(w_0) - f^*$; otherwise the convergence claim and the rejected-trial bound are immediate. In this case, condition (113) implies $\Gamma_0 > 3\varepsilon/R^3$, so the first index below is positive. Let N be the first index satisfying

$$\Gamma_N \le \frac{3\varepsilon}{2R^3}.$$

Then (127) gives $f(w_N) - f^* \le \varepsilon$.

As long as $\Gamma_k > \max\left\{\Gamma_{\mathrm{tr}}^{\mathrm{ad}}, 3\varepsilon/(2R^3)\right\}$, Lemma D.20 gives geometric contraction. Therefore, after at most

$$N_{\mathrm{far}} := \left\lceil \left(2 + (144\widehat{B})^{1/3}\right) \log \max\left\{1, \frac{\Gamma_0}{\max\left\{\Gamma_{\mathrm{tr}}^{\mathrm{ad}}, 3\varepsilon/(2R^3)\right\}}\right\} \right\rceil$$

iterations, either $\Gamma_k \le 3\varepsilon/(2R^3)$, or the near regime begins. In the second case, Lemma D.21 shows that

$$N_{\mathrm{near}} := \left\lceil \left(\frac{5184 M_0 R^3}{\varepsilon}\right)^{1/3} \right\rceil$$

additional iterations suffice. Consequently, $N \le N_{\mathrm{far}} + N_{\mathrm{near}}$. By (128),

$$\frac{\Gamma_0}{\max\left\{\Gamma_{\mathrm{tr}}^{\mathrm{ad}}, 3\varepsilon/(2R^3)\right\}} = \min\left\{\frac{\widehat{B}\Gamma_0}{4M_0}, \frac{2\Gamma_0 R^3}{3\varepsilon}\right\}.$$

Thus,

$$N \le \left\lceil \left(2 + (144\widehat{B})^{1/3}\right) \log \max\left\{1, \min\left\{\frac{\widehat{B}\Gamma_0}{4M_0}, \frac{2\Gamma_0 R^3}{3\varepsilon}\right\}\right\} \right\rceil + \left\lceil \left(\frac{5184 M_0 R^3}{\varepsilon}\right)^{1/3} \right\rceil. \tag{135}$$

Under (GH), definition (118) gives

$$\widehat{B}^{1/3} = \mathcal{O}\left(1 + \sqrt{M_1} R\right).$$

Under (LH), it gives

$$\widehat{B}^{1/3} = \mathcal{O}\left(1 + \sqrt{M_1} R + \frac{R}{\rho}\right).$$

Substituting these estimates into (135) proves (119) and (120).

The minimality of N gives

$$\Gamma_{N-1} > \frac{3\varepsilon}{2R^3}, \qquad \frac{2M_0}{\Gamma_{N-1}} < \frac{4M_0 R^3}{3\varepsilon}.$$

Therefore, (126) proves that the main backtracking loops reject at most $\left\lceil \log_2 \max\left\{1, B, \frac{4M_0R^3}{3\varepsilon}\right\}\right\rceil$ trials. □

E SUBOPTIMAL A-NPE METHOD

In this section, we propose suboptimal adaptive A-NPE method, which matches the convergence rate of Nesterov-type acceleration in exact Hessian case. For (H) and (IH) it achieves convergence rate $\mathcal{O}\left(\frac{LR^3}{N^3} + \frac{\delta R^2}{N^2}\right)$.

We repeat the two backtracking tests for convenience:

$$\mathsf{PD}(x,\tau) : H(x) + \tau I \succ 0, \quad \mathsf{MS}(x,\tau,s) : \|\nabla f(x+s) + \tau s\| \le \frac{\tau}{2}\|s\|. \tag{136}$$

For $w, v \in \mathbb{R}^d$ and $\tau > 0$, we first check $\mathsf{PD}(w, \tau)$. If it holds, we compute the trial step

$$(H(w) + \tau I)s = -\nabla f(v), \qquad y = v + s, \tag{137}$$

and accept it when $\mathsf{MS}(v, \tau, s)$ holds.

The initialization backtracking keeps x_0, $\nabla f(x_0)$, and $H(x_0)$ fixed. In the main loop, $H(w_k)$ remains fixed during backtracking, whereas v_k and $\nabla f(v_k)$ are recomputed after each rejection.

Algorithm 6 Adaptive A-NPE with an inexact Hessian

Input: $x_0 \in \mathbb{R}^d$ and $\tau_0 > 0$.
$(w_1, \tau_1) := \mathrm{MSBACKTRACK}(x_0, \tau_0)$.
Set

$$A_1 := \tfrac{1}{\tau_1}, \qquad z_1 := x_0 - A_1 \nabla f(w_1). \tag{138}$$

for $k \geq 1$ **do**
  Query $H(w_k)$ once.
  Set

$$\alpha_k := \tfrac{2}{1+\sqrt{1+4A_k\tau_k}}, \qquad A_{k+1} := \tfrac{A_k}{1-\alpha_k}, \qquad a_{k+1} := \alpha_k A_{k+1}. \tag{139}$$

  Set $v_k := (1 - \alpha_k)w_k + \alpha_k z_k$.
  if $\mathsf{PD}(w_k, \tau_k)$ **then**
    Compute s_k and w_{k+1} according to (137) with $w = w_k$, $v = v_k$, and $\tau = \tau_k$.
  while not $\mathsf{PD}(w_k, \tau_k)$ **or not** $\mathsf{MS}(v_k, \tau_k, s_k)$ **do**
    Set $\tau_k := 2\tau_k$.
    Recompute α_k, A_{k+1}, and a_{k+1} according to (139).
    Recompute $v_k := (1 - \alpha_k)w_k + \alpha_k z_k$.
    if $\mathsf{PD}(w_k, \tau_k)$ **then**
      Recompute s_k and w_{k+1} according to (137).
  Set

$$z_{k+1} := z_k - a_{k+1}\nabla f(w_{k+1}), \qquad \tau_{k+1} := (1 - \alpha_k)\tau_k.$$

For the convergence bounds of Algorithm 6, we use the convention $\rho = +\infty$ under (GH) and define

$$B := \max\left\{1, 8\sqrt{M_1}R, \tfrac{8R}{\rho}\right\}. \tag{140}$$

If $R = 0$, then x_0 is already a minimizer. For $R > 0$, set

$$\varepsilon_{\mathrm{tr}} := \max\left\{\tfrac{20M_0R^3}{B^3}, \tfrac{4\delta R^2}{B^2}\right\}.$$

Theorem E.1. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Assume that $R > 0$. For every $\varepsilon > 0$, the method generates an iterate w_N satisfying $f(w_N) - f^* \leq \varepsilon$ within*

$$N = \mathcal{O}\left(B\left(1 + \log_+ \tfrac{R^2}{2A_1 \max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}}\right) + \left(\tfrac{M_0R^3}{\varepsilon}\right)^{1/3} + \sqrt{\tfrac{\delta R^2}{\varepsilon}}\right)$$

iterations. The number of rejections is at most

$$\left\lceil \log_2 \max\left\{1, \tfrac{4\delta}{\tau_0}, \tfrac{2\sqrt{L(x_0)\|\nabla f(x_0)\|}}{\tau_0}, \tfrac{2\|\nabla f(x_0)\|}{\rho\tau_0}\right\}\right\rceil + \left\lceil \log_2 \max\left\{B^2, \tfrac{4\delta R^2}{\varepsilon}, \left(\tfrac{20M_0R^3}{\varepsilon}\right)^{2/3}\right\}\right\rceil.$$

Moreover, there exists an integer $K \geq 1$ with $K = \mathcal{O}\left(B \max\left\{1, \log \tfrac{R^2}{2A_1\varepsilon_{\mathrm{tr}}}\right\}\right)$ such that

$$f(w_k) - f^* = \mathcal{O}\left(\tfrac{R^2}{A_1}\exp\left(-\tfrac{c(k-1)}{B}\right)\right), \qquad 1 \leq k \leq K,$$
$$f(w_{K+n}) - f^* = \mathcal{O}\left(\tfrac{M_0R^3}{(B+n)^3} + \tfrac{\delta R^2}{(B+n)^2}\right), \qquad n = 0, 1, \ldots.$$

Here $c > 0$ and the constants in $\mathcal{O}(\cdot)$ are universal.

E.1 Auxiliary Lemmas

Lemma E.2. *For every main-loop trial of Algorithm 6, the parameters defined in* (139) *satisfy*

$$A_{k+1} = A_k + a_{k+1}, \quad \frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{\tau_k}, \quad v_k = \frac{A_k w_k + a_{k+1} z_k}{A_{k+1}}, \quad A_{k+1}\tau_k = \frac{1}{\alpha_k^2}.$$

Proof. By (139), we have

$$\alpha_k = \frac{2}{1 + \sqrt{1 + 4A_k\tau_k}} = \frac{\sqrt{1 + 4A_k\tau_k} - 1}{2A_k\tau_k}.$$

Thus α_k is the positive root of the quadratic equation

$$A_k\tau_k\alpha_k^2 + \alpha_k = 1, \qquad \frac{\alpha_k^2}{1 - \alpha_k} = \frac{1}{A_k\tau_k}. \tag{141}$$

Using (139) and (141) gives

$$A_{k+1} - a_{k+1} = (1 - \alpha_k)A_{k+1} = A_k, \qquad \frac{a_{k+1}^2}{A_{k+1}} = \alpha_k^2 A_{k+1} = \frac{\alpha_k^2 A_k}{1 - \alpha_k} = \frac{1}{\tau_k}.$$

The definition of v_k and the identities above imply

$$v_k = (1 - \alpha_k)w_k + \alpha_k z_k = \frac{A_k}{A_{k+1}} w_k + \frac{a_{k+1}}{A_{k+1}} z_k = \frac{A_k w_k + a_{k+1} z_k}{A_{k+1}},$$
$$A_{k+1}\tau_k = \frac{A_{k+1}^2}{a_{k+1}^2} = \frac{1}{\alpha_k^2}.$$

□

Lemma E.3. *Suppose that either* (GH) *or* (LH) *holds. Fix $w, v, s \in \mathbb{R}^d$, set $y := v + s$, and let $H(w)$ satisfy* (IH) *at w. Under* (LH)*, assume additionally that*

$$\|v - w\| + \|s\| \le \rho.$$

Then

$$\|\nabla f(y) - \nabla f(v) - H(w)s\| \le \left(\delta + L(y)\left(\|v - w\| + \frac{3}{2}\|s\|\right)\right)\|s\|. \tag{142}$$

Proof. By Lemmas A.1 A.2 and assumption of the lemma

$$\|\nabla f(y) - \nabla f(v) - \nabla^2 f(y)s\| = \|\nabla f(y - s) - \nabla f(y) + \nabla^2 f(y)s\| \le \frac{L(y)}{2}\|s\|^2.$$

Also,

$$\|y - w\| = \|v + s - w\| \le \|v - w\| + \|s\|,$$

which is at most ρ under (LH). Applying (GH) or (LH) with base point y, together with (IH) at w, gives

$$\begin{aligned}\|\nabla^2 f(y) - H(w)\| &\le \|\nabla^2 f(y) - \nabla^2 f(w)\| + \|\nabla^2 f(w) - H(w)\| \\ &\le L(y)\|y - w\| + \delta \le \delta + L(y)\left(\|v - w\| + \|s\|\right).\end{aligned}$$

Adding and subtracting $\nabla^2 f(y)s$, we obtain

$$\begin{aligned}&\|\nabla f(y) - \nabla f(v) - H(w)s\| \le \|\nabla f(y) - \nabla f(v) - \nabla^2 f(y)s\| + \|\nabla^2 f(y) - H(w)\|\|s\| \\ &\le \frac{L(y)}{2}\|s\|^2 + \left(\delta + L(y)\left(\|v - w\| + \|s\|\right)\right)\|s\| = \left(\delta + L(y)\left(\|v - w\| + \frac{3}{2}\|s\|\right)\right)\|s\|.\end{aligned}$$

This proves (142). □

For a trial step (137), the linear system gives

$$\nabla f(y) + \tau s = \nabla f(y) - \nabla f(v) - H(w)s.$$

E.2 Step Estimates

Lemma E.4. *Let f be convex and differentiable. Fix $w, z \in \mathbb{R}^d$ and $A \geq 0$. Choose $a > 0$ and $\tau > 0$ such that*

$$A_+ := A + a, \qquad \frac{a^2}{A_+} = \frac{1}{\tau}, \qquad v := \frac{A}{A_+} w + \frac{a}{A_+} z.$$

Let $\mathsf{MS}(v, \tau, s)$ in (136) hold, i.e.

$$\|\nabla f(v + s) + \tau s\| \leq \frac{\tau}{2} \|s\|, \tag{143}$$

and define

$$y := v + s, \qquad z_+ := z - a \nabla f(y).$$

Then

$$A_+(f(y) - f^*) + \frac{1}{2}\|z_+ - x^*\|^2 \leq A(f(w) - f^*) + \frac{1}{2}\|z - x^*\|^2 - \frac{3}{8} A_+ \tau \|s\|^2. \tag{144}$$

Proof. Set $g := \nabla f(y)$. Squaring (143) gives

$$\|g\|^2 + 2\tau \langle g, s \rangle + \tau^2 \|s\|^2 = \|g + \tau s\|^2 \leq \frac{\tau^2}{4} \|s\|^2.$$

Rearranging and dividing by $2\tau > 0$, we obtain

$$\langle g, s \rangle \leq -\frac{1}{2\tau} \|g\|^2 - \frac{3\tau}{8} \|s\|^2. \tag{145}$$

Convexity at y gives

$$f(w) \geq f(y) + \langle g, w - y \rangle, \qquad f^* \geq f(y) + \langle g, x^* - y \rangle.$$

Multiply these inequalities by A and a, respectively, and add them:

$$A f(w) + a f^* \geq A_+ f(y) + \langle g, Aw + a x^* - A_+ y \rangle.$$

The identities $A_+ v = Aw + az$ and $y = v + s$ imply

$$Aw + a x^* - A_+ y = Aw + a x^* - (Aw + az) - A_+ s = a(x^* - z) - A_+ s.$$

Substituting this identity and using $A_+ = A + a$ yields

$$A_+(f(y) - f^*) \leq A(f(w) - f^*) + a \langle g, z - x^* \rangle + A_+ \langle g, s \rangle. \tag{146}$$

The update $z_+ = z - ag$ gives

$$\frac{1}{2}\|z_+ - x^*\|^2 = \frac{1}{2}\|z - x^* - ag\|^2 = \frac{1}{2}\|z - x^*\|^2 - a \langle g, z - x^* \rangle + \frac{a^2}{2}\|g\|^2. \tag{147}$$

Adding (146) and (147) gives

$$A_+(f(y) - f^*) + \frac{1}{2}\|z_+ - x^*\|^2 \leq A(f(w) - f^*) + \frac{1}{2}\|z - x^*\|^2 + A_+ \langle g, s \rangle + \frac{a^2}{2}\|g\|^2.$$

Finally, (145) and $a^2 = A_+/\tau$ give

$$A_+ \langle g, s \rangle + \frac{a^2}{2}\|g\|^2 \leq \left(\frac{a^2}{2} - \frac{A_+}{2\tau} \right) \|g\|^2 - \frac{3}{8} A_+ \tau \|s\|^2 = -\frac{3}{8} A_+ \tau \|s\|^2.$$

Substitution proves (144). □

Lemma E.2 shows that the weights used in the main loop of Algorithm 6 satisfy the relations assumed in Lemma E.4.

To bound the potential after initialization step in Algorithm 6, we apply Lemma E.4 with $A = 0$ and $w = z = x_0$. For the accepted step, take $a = A_+ = A_1 = 1/\tau_1$ and $\tau = \tau_1$. These choices

satisfy the weight condition, and the updates in (138) give $y = w_1$ and $z_+ = z_1$. Since $A_1\tau_1 = 1$, the lemma yields

$$A_1(f(w_1) - f^*) + \frac{1}{2}\|z_1 - x^*\|^2 + \frac{3}{8}\|w_1 - x_0\|^2 \le \frac{R^2}{2}. \tag{148}$$

For the accepted iterates, write

$$\mathcal{E}_k := A_k(f(w_k) - f^*) + \frac{1}{2}\|z_k - x^*\|^2.$$

For any accepted iterate w_k, summing (144) over $i = 1, \dots, k-1$ and using the initialization bound gives

$$\mathcal{E}_k + \frac{3}{8}\sum_{i=1}^{k-1} A_{i+1}\tau_i\|s_i\|^2 \le \mathcal{E}_1 \le \frac{R^2}{2}, \tag{149}$$

where $s_i = w_{i+1} - v_i$ and τ_i is accepted on iteration i. Since all terms on the left-hand side of (149) are nonnegative,

$$A_k(f(w_k) - f^*) \le \mathcal{E}_k \le \frac{R^2}{2}, \qquad \sum_{i=1}^{k-1} A_{i+1}\tau_i\|s_i\|^2 \le \frac{4}{3}R^2.$$

Consider any trial satisfying the conditions of Lemma E.4 with $A = A_k$, $w = w_k$, and $z = z_k$, and write $\alpha = a/A_+$. Let $y := v + s$ be the trial point. For an accepted main-loop step of Algorithm 6, $y = v_k + s_k = w_{k+1}$. The left-hand side of (144) is nonnegative. Hence, by (149),

$$\frac{3}{8}A_+\tau\|s\|^2 \le \mathcal{E}_k \le \frac{R^2}{2}.$$

The MS test (143) also gives

$$\|\nabla f(y)\| \le \|\nabla f(y) + \tau s\| + \tau\|s\| \le \frac{\tau}{2}\|s\| + \tau\|s\| = \frac{3\tau}{2}\|s\|.$$

Using $A_+\tau = A_+^2/a^2 = 1/\alpha^2$, we obtain

$$\|s\| \le \frac{2}{\sqrt{3}}\alpha R, \qquad \|\nabla f(y)\| \le \frac{3}{2}\tau\|s\|. \tag{150}$$

Lemma E.5. *For every $k \ge 1$, the accepted iterates of Algorithm 6 satisfy*

$$\|z_k - x^*\| \le R, \tag{151}$$

$$\|w_k - x^*\| \le \left(1 + \frac{2}{\sqrt{3}}\right)R, \qquad \|z_k - w_k\| \le \left(2 + \frac{2}{\sqrt{3}}\right)R. \tag{152}$$

Proof. By (149),

$$\frac{1}{2}\|z_k - x^*\|^2 \le \mathcal{E}_k \le \frac{R^2}{2},$$

which gives (151). We prove the bound on $\|w_k - x^*\|$ by induction. The initialization bound (148) gives

$$\|w_1 - x_0\| \le \frac{2}{\sqrt{3}}R,$$

$$\|w_1 - x^*\| \le \|w_1 - x_0\| + \|x_0 - x^*\| \le \left(1 + \frac{2}{\sqrt{3}}\right)R.$$

Suppose the bound holds for w_k. For the next accepted step, $w_{k+1} = (1 - \alpha_k)w_k + \alpha_k z_k + s_k$. Using (150) and (151), we obtain

$$\|w_{k+1} - x^*\| \le (1 - \alpha_k)\|w_k - x^*\| + \alpha_k\|z_k - x^*\| + \|s_k\|$$
$$\le (1 - \alpha_k)\left(1 + \frac{2}{\sqrt{3}}\right)R + \alpha_k R + \frac{2}{\sqrt{3}}\alpha_k R = \left(1 + \frac{2}{\sqrt{3}}\right)R.$$

This completes the induction. The bound on $\|z_k - w_k\|$ follows from

$$\|z_k - w_k\| \le \|z_k - x^*\| + \|w_k - x^*\| \le \left(2 + \frac{2}{\sqrt{3}}\right)R.$$

□

Fix already accepted w_k, z_k. These points remain fixed during backtracking. For $v = (1-\alpha)w_k + \alpha z_k$ with any $\alpha \in (0,1)$, inequality (152) gives

$$\|v - w_k\| = \alpha\|z_k - w_k\| \le \left(2 + \frac{2}{\sqrt{3}}\right)\alpha R.$$

This bound holds for every trial, whether or not the MS test passes.

E.3 Explicit Choice of Regularization Parameter

With B defined in (140), set

$$Q(A) := \max\left\{B^2, 8\delta A, (40M_0RA)^{2/3}\right\}, \qquad A > 0. \tag{153}$$

Lemma E.6. *Suppose that f is convex and either* (GH) *or* (LH) *holds. Fix an accepted state A_k, w_k, z_k of Algorithm 6 and a matrix $H(w_k)$ satisfying* (IH) *at w_k. Every main-loop trial with*

$$\tau \ge \frac{Q(A_k)}{A_k} \tag{154}$$

satisfies $\mathsf{PD}(w_k, \tau)$ and $\mathsf{MS}(v, \tau, s)$ in (136)*, where the trial weights are defined by* (139) *with $\tau_k = \tau$, $v = (1-\alpha_k)w_k + \alpha_k z_k$, and s is given by* (137) *with $w = w_k$.*

Proof. Keep $A_k, w_k, z_k, H(w_k)$ fixed and vary $q \ge Q(A_k)$ continuously. For each q, use the weights in (139) with $\tau_k = \tau(q)$, where

$$\tau(q) := \frac{q}{A_k}, \quad \alpha(q) := \frac{2}{1+\sqrt{1+4q}}, \quad v(q) := (1-\alpha(q))w_k + \alpha(q)z_k.$$

Convexity and (IH) give $H(w_k) \succeq -\delta I$. Since $q \ge Q(A_k) \ge \max\{1, 8\delta A_k\}$, we have

$$H(w_k) + \tau(q)I \succeq (\tau(q) - \delta)I \succeq \frac{7}{8}\tau(q)I \succ 0.$$

Thus, $\mathsf{PD}(w_k, \tau(q))$ holds throughout this range, and we can define

$$\begin{aligned} s(q) &:= -\left(H(w_k) + \tau(q)I\right)^{-1}\nabla f(v(q)), \qquad & y(q) &:= v(q) + s(q), \\ d(q) &:= \|v(q) - w_k\| = \alpha(q)\|z_k - w_k\|, \qquad & r(q) &:= \|s(q)\|. \end{aligned}$$

All these quantities are continuous because $H(w_k)$ is fixed and the inverse exists. As $q \to +\infty$,

$$\tau(q) \to +\infty, \;\; \alpha(q) \to 0, \;\; v(q) \to w_k, \;\; d(q) \to 0, \;\; r(q) \le \frac{\|\nabla f(v(q))\|}{\tau(q) - \delta} \to 0, \;\; y(q) \to w_k.$$

By Lemma E.3 and (137), the following conditions are sufficient for the test (136):

$$d(q) + r(q) \le \rho \;\text{ only under (LH)}, \quad \delta + L(y(q))\left(d(q) + \frac{3}{2}r(q)\right) \le \frac{\tau(q)}{2}. \tag{155}$$

Indeed, if (155) holds, then Lemma E.3 and (137) give

$$\begin{aligned} \|\nabla f(y(q)) + \tau(q)s(q)\| &= \|\nabla f(y(q)) - \nabla f(v(q)) - H(w_k)s(q)\| \\ &\le \left(\delta + L(y(q))\left(d(q) + \frac{3}{2}r(q)\right)\right) r(q) \le \frac{\tau(q)}{2}\|s(q)\|. \end{aligned}$$

Since $L(y(q)) \to L(w_k)$,

$$\frac{\delta + L(y(q))\left(d(q) + \frac{3}{2}r(q)\right)}{\tau(q)} \to 0.$$

Hence, all sufficiently large q satisfy (155) strictly.

Let $\mathcal{U}_k$ be the set of $q \ge Q(A_k)$ for which at least one condition in (155) fails. If $\mathcal{U}_k = \emptyset$, the statement follows. Otherwise, set $q^\dagger := \sup \mathcal{U}_k$. Since the conditions in (155) hold for all sufficiently large q, the set $\mathcal{U}_k$ is bounded above. Hence, $q^\dagger < +\infty$. By continuity, $q^\dagger$ lies on the boundary of the

admissible set. Hence, the conditions in (155) hold at $q^\dagger$, and at least one is an equality. A dagger below denotes evaluation at $q^\dagger$.

The boundary trial therefore satisfies the test (155). Moreover, (149) gives $\mathcal{E}_k \le \frac{R^2}{2}$. Consequently, the bounds (150) apply to this trial. Together with (152), they give

$$r^\dagger \le \frac{2}{\sqrt{3}}\alpha^\dagger R, \quad \|\nabla f(y^\dagger)\| \le \frac{3}{2}\tau^\dagger r^\dagger \le \sqrt{3}\tau^\dagger\alpha^\dagger R, \quad d^\dagger = \alpha^\dagger\|z_k - w_k\| \le \left(2 + \frac{2}{\sqrt{3}}\right)\alpha^\dagger R. \tag{156}$$

In particular,

$$d^\dagger + r^\dagger \le d^\dagger + \frac{3}{2}r^\dagger \le \left(2 + \frac{5}{\sqrt{3}}\right)\alpha^\dagger R \le 5\alpha^\dagger R. \tag{157}$$

The definition of $\alpha(q)$ gives

$$\alpha^\dagger = \frac{2}{1 + \sqrt{1 + 4q^\dagger}} \le \frac{1}{\sqrt{q^\dagger}}.$$

By (153), we have $q^\dagger \ge Q(A_k) \ge B^2$. Under (LH), definition (140) also gives $B \ge \frac{8R}{\rho}$. Together with (157), this gives

$$d^\dagger + r^\dagger \le \frac{5R}{\sqrt{q^\dagger}} \le \frac{5R}{B} \le \frac{5}{8}\rho < \rho.$$

Using $L(y^\dagger) = M_0 + M_1\|\nabla f(y^\dagger)\|$, equations (156) and (157) give

$$\begin{aligned}\frac{\delta + L(y^\dagger)\left(d^\dagger + \frac{3}{2}r^\dagger\right)}{\tau^\dagger} &\le \frac{\delta}{\tau^\dagger} + \frac{5M_0R\alpha^\dagger}{\tau^\dagger} + 5\sqrt{3}M_1R^2(\alpha^\dagger)^2 \\ &\le \frac{\delta A_k}{q^\dagger} + \frac{5M_0RA_k}{(q^\dagger)^{3/2}} + \frac{5\sqrt{3}M_1R^2}{q^\dagger} \le \frac{1}{8} + \frac{1}{8} + \frac{5\sqrt{3}}{64} < \frac{1}{2},\end{aligned}$$

where we used $q^\dagger \ge Q(A_k)$, (153), and (140). Thus, all conditions in (155) are strict at $q^\dagger$, contradicting its boundary property. Therefore, $\mathcal{U}_k = \emptyset$, and every trial satisfying (154) passes both tests. □

E.4 Main-Loop Backtracking

Lemma E.7. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Every main backtracking loop of Algorithm 6 terminates. For every $k \ge 1$, let τ_k denote the parameter accepted at iteration k. Then*

$$A_k\tau_k \le 2Q(A_k). \tag{158}$$

For every $K \ge 1$, let J_K be the total number of rejected main trials during iterations $1, \dots, K$, excluding initialization. Then

$$J_K \le \lceil \log_2 Q(A_K) \rceil. \tag{159}$$

Proof. We prove by induction on k that iteration k is reached after finitely many trials. The case $k = 1$ follows from Lemma C.2 with $(w_1, \tau_1) = \text{MSBacktrack}(x_0, \tau_0)$ and $\overline{\tau}_0 := \overline{\tau}(x_0)$. Suppose this holds for some $k \ge 1$. Then (149) and (152) hold. Since A_k is fixed, repeated doubling reaches $\tau \ge \frac{Q(A_k)}{A_k}$, and Lemma E.6 guarantees acceptance. Thus, iteration $k + 1$ is reached after finitely many trials, completing the induction.

We next track the product of the current weight and trial parameter. By (138), before the first main trial this product is

$$A_1\tau_1 = \frac{1}{\tau_1}\tau_1 = 1.$$

After accepting iteration k, the first trial at iteration $k+1$ uses $\tau = (1-\alpha_k)\tau_k$. By (139), its product is

$$A_{k+1}(1 - \alpha_k)\tau_k = \frac{A_k}{1 - \alpha_k}(1 - \alpha_k)\tau_k = A_k\tau_k.$$

Every rejected trial doubles τ_k, while A_k remains unchanged. Therefore,

$$A_k\tau_k = 2^{J_k}, \qquad k \ge 1. \tag{160}$$

Fix $K \geq 1$. The weights increase by (139), and Q is nondecreasing by (153). Hence, $Q(A_i) \leq Q(A_K)$ for every $i \leq K$. If $J_K = 0$, then (160) and $Q(A_K) \geq B^2 \geq 1$ give both (158) and (159).

Suppose that $J_K \geq 1$, and let $i \leq K$ be the iteration at which the last rejection occurs. For this trial, the product of A_i and the trial parameter is 2^{J_K-1}. By Lemma E.6, the trial would be accepted if this product were at least $Q(A_i)$. Therefore,

$$2^{J_K-1} < Q(A_i) \leq Q(A_K).$$

Multiplying by 2 and using (160), we obtain

$$A_K \tau_K = 2^{J_K} < 2Q(A_K),$$

which proves (158). Taking logarithms in $2^{J_K-1} < Q(A_K)$ gives $J_K < 1 + \log_2 Q(A_K)$. Since J_K is an integer,

$$J_K \leq \lceil \log_2 Q(A_K) \rceil,$$

which proves (159). □

E.5 Proof of Theorem E.1

For the next lemma, recall the definitions in (140) and (153):

$$B = \begin{cases} \max\left\{1, 8\sqrt{M_1}R\right\}, & \text{under (GH)}, \\ \max\left\{1, 8\sqrt{M_1}R, \frac{8R}{\rho}\right\}, & \text{under (LH)}, \end{cases}$$

$$Q(A) = \max\left\{B^2, 8\delta A, (40M_0RA)^{2/3}\right\}, \qquad A > 0.$$

We also recall the error threshold

$$\varepsilon_{\mathrm{tr}} = \max\left\{\frac{20M_0R^3}{B^3}, \frac{4\delta R^2}{B^2}\right\}. \tag{161}$$

For $R > 0$, set

$$A_{\mathrm{tr}} := \min\left\{\frac{B^3}{40M_0R}, \frac{B^2}{8\delta}\right\} = \frac{R^2}{2\varepsilon_{\mathrm{tr}}}. \tag{162}$$

Here $\frac{B^2}{8\delta} = +\infty$ when $\delta = 0$. In particular,

$$\sqrt{Q(A)} = \max\left\{B, (40M_0RA)^{1/3}, \sqrt{8\delta A}\right\}, \qquad A > 0.$$

The threshold A_{tr} in (162) is the largest weight for which $Q(A) = B^2$.

Lemma E.8. *Under the assumptions of Lemma E.7, suppose that $R > 0$. For the accepted steps of Algorithm 6, the following bounds hold for every $k \geq 1$. If $A_k \leq A_{\mathrm{tr}}$, then*

$$A_{k+1} \geq \exp\left(\frac{1}{3B}\right) A_k. \tag{163}$$

If $A_k \geq A_{\mathrm{tr}}$, then

$$\sqrt{Q(A_{k+1})} - \sqrt{Q(A_k)} \geq \frac{1}{9}. \tag{164}$$

In particular, for any $K \geq 1$ with $A_K \geq A_{\mathrm{tr}}$ and every integer $n \geq 0$,

$$\sqrt{Q(A_{K+n})} \geq \sqrt{Q(A_K)} + \frac{n}{9} \geq B + \frac{n}{9}. \tag{165}$$

Proof. By (158), the accepted parameter satisfies $A_k\tau_k \leq 2Q(A_k)$. Since $\sqrt{Q(A_k)} \geq B \geq 1$, the definition of α_k in (139) gives

$$\frac{1}{\alpha_k} = \frac{1 + \sqrt{1 + 4A_k\tau_k}}{2} \leq \frac{1 + \sqrt{1 + 8Q(A_k)}}{2} \leq 3\sqrt{Q(A_k)}.$$

Hence,

$$\alpha_k \geq \frac{1}{3\sqrt{Q(A_k)}}, \qquad k \geq 1. \tag{166}$$

If $A_k \le A_{\mathrm{tr}}$, $\max\left\{(40M_0RA_k)^{1/3}, \sqrt{8\delta A_k}\right\} \le B$. Thus, by definition of $Q(A_k)$, we have $B = \sqrt{Q(A_k)}$. By (166), $\alpha_k \ge \frac{1}{3B}$. Using (139) and $1 - t \le \exp(-t)$ for $0 \le t < 1$, we obtain

$$A_{k+1} = \frac{A_k}{1-\alpha_k} \ge \exp(\alpha_k) A_k \ge \exp\left(\frac{1}{3B}\right) A_k,$$

which proves (163).

Since $A \ge A_{\mathrm{tr}}$, $B \le \max\left\{(40M_0RA)^{1/3}, \sqrt{8\delta A}\right\} = \sqrt{Q(A)}$. Suppose that $A_k \ge A_{\mathrm{tr}}$. Since $A_{k+1} > A_k \ge A_{\mathrm{tr}}$, the identity holds for $A = A_k$ and $A = A_{k+1}$. Since $0 < 1 - \alpha_k < 1$, substitution of (139) gives

$$\begin{aligned}\sqrt{Q(A_{k+1})} &= \max\left\{\frac{(40M_0RA_k)^{1/3}}{(1-\alpha_k)^{1/3}}, \frac{\sqrt{8\delta A_k}}{(1-\alpha_k)^{1/2}}\right\} \\ &\ge \frac{\max\left\{(40M_0RA_k)^{1/3}, \sqrt{8\delta A_k}\right\}}{(1-\alpha_k)^{1/3}} = \frac{\sqrt{Q(A_k)}}{(1-\alpha_k)^{1/3}}.\end{aligned}$$

The inequality $(1-t)^{-1/3} \ge 1 + \frac{t}{3}$ for $0 \le t < 1$ and (166) now imply

$$\sqrt{Q(A_{k+1})} - \sqrt{Q(A_k)} \ge \sqrt{Q(A_k)}\left((1-\alpha_k)^{-1/3} - 1\right) \ge \frac{\alpha_k\sqrt{Q(A_k)}}{3} \ge \frac{1}{9}.$$

This proves (164). Finally, if $A_K \ge A_{\mathrm{tr}}$, then $A_j \ge A_{\mathrm{tr}}$ for every $j \ge K$. Summing (164) gives

$$\sqrt{Q(A_{K+n})} = \sqrt{Q(A_K)} + \sum_{j=K}^{K+n-1}\left(\sqrt{Q(A_{j+1})} - \sqrt{Q(A_j)}\right) \ge \sqrt{Q(A_K)} + \frac{n}{9} \ge B + \frac{n}{9},$$

which proves (165). □

We restate Theorem E.1 for convenience.

Theorem E.9. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Assume that $R > 0$. All backtracking loops in Algorithm 6 terminate. For every $\varepsilon > 0$, the method generates an iterate w_N satisfying $f(w_N) - f^* \le \varepsilon$ within*

$$N = \mathcal{O}\left(B\left(1 + \log_+ \frac{R^2}{2A_1\max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}}\right) + \left(\frac{M_0R^3}{\varepsilon}\right)^{1/3} + \sqrt{\frac{\delta R^2}{\varepsilon}}\right) \tag{167}$$

iterations. Under (GH)*, initialization rejects at most*

$$\left\lceil \log_2 \max\left\{1, \frac{4\delta}{\tau_0}, \frac{2\sqrt{L(x_0)\|\nabla f(x_0)\|}}{\tau_0}\right\}\right\rceil \tag{168}$$

trials. Under (LH)*, the maximum additionally includes $\frac{2\|\nabla f(x_0)\|}{\rho\tau_0}$. The total number of rejected main-loop trials before producing w_N is at most*

$$\left\lceil \log_2 \max\left\{B^2, \frac{4\delta R^2}{\varepsilon}, \left(\frac{20M_0R^3}{\varepsilon}\right)^{2/3}\right\}\right\rceil. \tag{169}$$

Moreover, there exists an integer $K \ge 1$ with

$$K \le 1 + \left\lceil 3B\log_+ \frac{R^2}{2A_1\varepsilon_{\mathrm{tr}}}\right\rceil. \tag{170}$$

For this K,

$$f(w_k) - f^* \le \frac{R^2}{2A_1}\exp\left(-\frac{k-1}{3B}\right), \qquad 1 \le k \le K, \tag{171}$$

$$f(w_{K+n}) - f^* \le \max\left\{\frac{20M_0R^3}{(B+n/9)^3}, \frac{4\delta R^2}{(B+n/9)^2}\right\}, \qquad n = 0, 1, \ldots. \tag{172}$$

Proof. Lemmas C.2 and E.7 show that all backtracking loops terminate. By (149),

$$f(w_k) - f^* \le \frac{\mathcal{E}_k}{A_k} \le \frac{R^2}{2A_k}, \qquad k \ge 1. \tag{173}$$

Set $K := \min\{k \ge 1 : A_k \ge A_{\mathrm{tr}}\}$, with $K = +\infty$ if this set is empty. If $A_1 \ge A_{\mathrm{tr}}$, then $K = 1$, and (170) holds. Otherwise, set

$$m := \left\lceil 3B \log \frac{A_{\mathrm{tr}}}{A_1} \right\rceil.$$

If $A_j < A_{\mathrm{tr}}$ for $j = 1, \dots, m$, then (163) yields

$$A_{m+1} \ge A_1 \exp\left(\frac{m}{3B}\right) \ge A_1 \exp\left(\log \frac{A_{\mathrm{tr}}}{A_1}\right) = A_{\mathrm{tr}}.$$

The threshold is therefore reached by index $m+1$. Since $A_{\mathrm{tr}} = \frac{R^2}{2\varepsilon_{\mathrm{tr}}}$, this proves (170).

Fix $1 \le k \le K$. By the definition of K, we have $A_j < A_{\mathrm{tr}}$ for $1 \le j < k$. Applying (163) successively and using (173) gives

$$A_k \ge A_1 \exp\left(\frac{k-1}{3B}\right),$$
$$f(w_k) - f^* \le \frac{R^2}{2A_k} \le \frac{R^2}{2A_1} \exp\left(-\frac{k-1}{3B}\right).$$

This proves (171).

For every $A \ge A_{\mathrm{tr}}$,

$$\sqrt{Q(A)} = \max\left\{(40M_0RA)^{1/3}, \sqrt{8\delta A}\right\}.$$

Hence, $Q(A)^{3/2} \ge 40M_0RA$ and $Q(A) \ge 8\delta A$, so

$$\frac{20M_0R^3}{Q(A)^{3/2}} \le \frac{R^2}{2A}, \qquad \frac{4\delta R^2}{Q(A)} \le \frac{R^2}{2A}.$$

At least one of these inequalities is an equality, because $\sqrt{Q(A)}$ equals at least one of the two terms defining it. Therefore,

$$\frac{R^2}{2A} = \max\left\{\frac{20M_0R^3}{Q(A)^{3/2}}, \frac{4\delta R^2}{Q(A)}\right\}, \qquad A \ge A_{\mathrm{tr}}. \tag{174}$$

For $A = A_{K+n}$, combine (173), (174), and (165) to obtain

$$f(w_{K+n}) - f^* \le \frac{R^2}{2A_{K+n}} = \max\left\{\frac{20M_0R^3}{Q(A_{K+n})^{3/2}}, \frac{4\delta R^2}{Q(A_{K+n})}\right\}$$
$$\le \max\left\{\frac{20M_0R^3}{(B+n/9)^3}, \frac{4\delta R^2}{(B+n/9)^2}\right\}.$$

This proves (172).

It remains to bound the iteration and backtracking counts. Fix $\varepsilon > 0$. Set

$$A_\varepsilon := \frac{R^2}{2\varepsilon}, \qquad N := \min\{k \ge 1 : A_k \ge A_\varepsilon\}.$$

By (165) and (174), $\frac{R^2}{2A_{K+n}} \to 0$ as $n \to +\infty$. Thus, N is finite, and (173) gives $f(w_N) - f^* \le \varepsilon$.

First, suppose that $\varepsilon \ge \varepsilon_{\mathrm{tr}}$. Then $A_\varepsilon \le A_{\mathrm{tr}}$, so every iteration with $A_k < A_\varepsilon$ satisfies (163). The same argument used to prove (170), with A_ε in place of A_{tr}, gives

$$N \le 1 + \left\lceil 3B \log_+ \frac{A_\varepsilon}{A_1} \right\rceil = 1 + \left\lceil 3B \log_+ \frac{R^2}{2A_1\varepsilon} \right\rceil.$$

Now suppose that $0 < \varepsilon < \varepsilon_{\mathrm{tr}}$. Set

$$n := \left\lceil 9 \max\left\{ \left(\frac{20M_0R^3}{\varepsilon}\right)^{1/3}, 2\sqrt{\frac{\delta R^2}{\varepsilon}} \right\} \right\rceil.$$

Equations (174) and (165) give

$$\frac{R^2}{2A_{K+n}} \le \max\left\{ \frac{20M_0R^3}{(B+n/9)^3}, \frac{4\delta R^2}{(B+n/9)^2} \right\} \le \varepsilon.$$

Hence $A_{K+n} \ge A_\varepsilon$, and $N \le K + n$. Using (170), we obtain

$$N \le 1 + \left\lceil 3B \log_+ \frac{A_{\mathrm{tr}}}{A_1} \right\rceil + \left\lceil 9 \max\left\{ \left(\frac{20M_0R^3}{\varepsilon}\right)^{1/3}, 2\sqrt{\frac{\delta R^2}{\varepsilon}} \right\} \right\rceil.$$

Combining the two cases and using (161) and $B \ge 1$ proves (167).

We now bound the number of rejected trials before producing w_N. If $N = 1$, no main trial is needed. If $N \ge 2$, the main iterations needed to compute w_N are $1, \dots, N-1$. The minimality of N gives $A_{N-1} < A_\varepsilon$. Since Q is nondecreasing, Lemma E.7 yields

$$J_{N-1} \le \lceil \log_2 Q(A_{N-1}) \rceil \le \lceil \log_2 Q(A_\varepsilon) \rceil.$$

Substituting $A_\varepsilon = \frac{R^2}{2\varepsilon}$ into (153) gives

$$Q(A_\varepsilon) = \max\left\{ B^2, \frac{4\delta R^2}{\varepsilon}, \left(\frac{20M_0R^3}{\varepsilon}\right)^{2/3} \right\}.$$

This proves the rejection bound (169). By Lemma C.2, initialization rejects at most

$$\left\lceil \log_2 \max\left\{ 1, \frac{\overline{\tau}_0}{\tau_0} \right\} \right\rceil$$

trials, where $\overline{\tau}_0 = \overline{\tau}(x_0)$ is defined in (17). Substituting (17) gives the initialization bound (168) and its counterpart under (LH).

□

E.6 Oracle Complexity

Corollary E.10. *Under the assumptions of Theorem E.9, fix $\varepsilon > 0$ and let N be the first index with $A_N \ge R^2/(2\varepsilon)$. Let $\overline{\tau}_0 := \overline{\tau}(x_0)$, where $\overline{\tau}(x)$ is defined in* (17)*. Then Algorithm 6 computes w_N using at most N Hessian evaluations and at most*

$$2N + 3 + \log_{2,+} \frac{\overline{\tau}_0}{\tau_0} + 2 \log_2 \max\left\{ B^2, \frac{4\delta R^2}{\varepsilon}, \left(\frac{20M_0R^3}{\varepsilon}\right)^{2/3} \right\}$$

gradient evaluations.

Proof. Initialization uses one Hessian, the gradient at x_0, and at most one gradient per trial. By Lemma C.2, it has at most $2 + \log_{2,+} \frac{\overline{\tau}_0}{\tau_0}$ trials. Each of the $N-1$ main iterations uses one Hessian $H(w_k)$ and at most two gradients per trial, $\nabla f(v_k)$ and $\nabla f(v_k + s_k)$, since v_k changes after each rejection. These iterations have $N-1$ accepted trials, and (169) bounds the number of rejected ones. The updates reuse the gradient at the accepted point. Hence, the number of Hessian evaluations is at most $1 + (N-1) = N$, and the number of gradient evaluations is at most

$$1 + \left(2 + \log_{2,+} \frac{\overline{\tau}_0}{\tau_0}\right) + 2\left(N - 1 + \left\lceil \log_2 \max\left\{ B^2, \frac{4\delta R^2}{\varepsilon}, \left(\frac{20M_0R^3}{\varepsilon}\right)^{2/3} \right\} \right\rceil\right)$$
$$\le 2N + 3 + \log_{2,+} \frac{\overline{\tau}_0}{\tau_0} + 2 \log_2 \max\left\{ B^2, \frac{4\delta R^2}{\varepsilon}, \left(\frac{20M_0R^3}{\varepsilon}\right)^{2/3} \right\}.$$

□

Algorithm 7 Adaptive damped A-NPE with an inexact Hessian

1: **Input:** $x_0 \in \mathbb{R}^d$ and $\eta_0 > 0$.
2: $(w_1, \eta_1) := \text{MSBACKTRACK}(x_0, \eta_0)$.
3: Set
$$A_1 := \frac{1}{\eta_1}, \qquad z_1 := x_0 - A_1 \nabla f(w_1). \tag{175}$$
4: **for** $k \geq 1$ **do**
5: Set
$$a'_k := \frac{1 + \sqrt{1 + 4\eta_k A_k}}{2\eta_k}, \qquad A'_k := A_k + a'_k, \qquad v_k := \frac{A_k w_k + a'_k z_k}{A'_k}. \tag{176}$$
6: $(y_k, \tau_k) := \text{MSBACKTRACK}(v_k, \eta_k)$.
7: Set
$$\gamma_k := \frac{\eta_k}{\tau_k}, \qquad a_{k+1} := \gamma_k a'_k, \qquad A_{k+1} := A_k + a_{k+1}. \tag{177}$$
8: Set
$$w_{k+1} := \frac{(1-\gamma_k) A_k w_k + \gamma_k A'_k y_k}{A_{k+1}}, \qquad z_{k+1} := z_k - a_{k+1} \nabla f(y_k). \tag{178}$$
9: **if** $\tau_k = \eta_k$ **then**
10: Set $\eta_{k+1} := \frac{\eta_k}{2}$.
11: **else**
12: Set $\eta_{k+1} := 2\eta_k$.

F Proof of Algorithm 3

We restate Algorithm 3.

F.1 Initialization and Backtracking

Lemma C.2 guarantees termination of all calls to MSBACKTRACK in Algorithm 7.

Lemma F.1. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Initialization in Algorithm 7 (line 2) uses at most*

$$1 + \left\lceil \log_2 \max \left\{ 1, \frac{\overline{\tau}_0}{\eta_0} \right\} \right\rceil$$

trials, where $\overline{\tau}_0 = \overline{\tau}(x_0)$ from (17)*, and accepted parameter satisfies*

$$\eta_0 \leq \eta_1 \leq \max\{\eta_0, 2\overline{\tau}_0\}.$$

The algorithm either returns a minimizer or produces an initial state A_1, w_1, z_1, η_1 satisfying $A_1 \eta_1 = 1$ and

$$A_1(f(w_1) - f^*) + \frac{1}{2}\|z_1 - x^*\|^2 + \frac{3}{8}\|w_1 - x_0\|^2 \leq \frac{R^2}{2}. \tag{179}$$

Proof. If $\nabla f(x_0) = 0$, the algorithm returns x_0 without any trials. Otherwise, Lemma C.2 with $x = x_0$, $\eta = \eta_0$, and $\overline{\tau}(x_0) = \overline{\tau}_0$ proves the bound. For the initialization (175), Lemma E.4 applies with $A = 0$, $w = z = x_0$, $a = A_+ = A_1 = 1/\eta_1$, and $\tau = \eta_1$. The same updates give $y = w_1$ and $z_+ = z_1$. Since $A_1 \tau = 1$, the lemma yields (179). □

F.2 Step Estimates

Lemma F.2. *Let f be convex. Every completed main iteration of Algorithm 7 satisfies*

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k - \frac{3}{8} A'_k \eta_k \|s_k\|^2, \tag{180}$$

where $\mathcal{E}_k := A_k(f(w_k) - f^*) + \frac{1}{2}\|z_k - x^*\|^2$.

Proof. By (176) and (177), $(a'_k)^2 = A'_k/\eta_k$ and $\gamma_k\tau_k = \eta_k$, so

$$\frac{a_{k+1}^2}{\gamma_k A'_k} = \frac{1}{\tau_k}, \qquad v_k = \frac{\gamma_k A_k w_k + a_{k+1} z_k}{\gamma_k A'_k}.$$

Apply Lemma E.4 with $A = \gamma_k A_k$, $a = a_{k+1}$, $A_+ = \gamma_k A'_k$, $w = w_k$, $z = z_k$, and $\tau = \tau_k$. By (178), $z_+ = z_{k+1}$, so it gives

$$\gamma_k A'_k(f(y_k)-f^*)+\frac{1}{2}\|z_{k+1}-x^*\|^2 \le \gamma_k A_k(f(w_k)-f^*)+\frac{1}{2}\|z_k-x^*\|^2-\frac{3}{8}A'_k\eta_k\|s_k\|^2. \quad (181)$$

Since $0 < \gamma_k \le 1$ and

$$\frac{(1-\gamma_k)A_k}{A_{k+1}} + \frac{\gamma_k A'_k}{A_{k+1}} = \frac{A_k + \gamma_k a'_k}{A_{k+1}} = 1,$$

convexity of f and (178) give

$$A_{k+1}(f(w_{k+1}) - f^*) \le (1-\gamma_k)A_k(f(w_k) - f^*) + \gamma_k A'_k(f(y_k) - f^*). \quad (182)$$

Using (182) and then (181), we obtain

$$\begin{aligned}
\mathcal{E}_{k+1} &= A_{k+1}(f(w_{k+1}) - f^*) + \frac{1}{2}\|z_{k+1} - x^*\|^2 \\
&\le (1-\gamma_k)A_k(f(w_k) - f^*) + \gamma_k A'_k(f(y_k) - f^*) + \frac{1}{2}\|z_{k+1} - x^*\|^2 \\
&\le (1-\gamma_k)A_k(f(w_k) - f^*) + \gamma_k A_k(f(w_k) - f^*) + \frac{1}{2}\|z_k - x^*\|^2 - \frac{3}{8}A'_k\eta_k\|s_k\|^2 \\
&= \mathcal{E}_k - \frac{3}{8}A'_k\eta_k\|s_k\|^2.
\end{aligned}$$

□

Summing (180) over $i = 1, \dots, k-1$ and using (179), we obtain

$$\mathcal{E}_k + \frac{3}{8}\sum_{i=1}^{k-1} A'_i\eta_i\|s_i\|^2 \le \mathcal{E}_1 \le \frac{R^2}{2}. \quad (183)$$

Since all terms on the left-hand side of (183) are nonnegative,

$$f(w_k) - f^* \le \frac{R^2}{2A_k}, \qquad \sum_{i=1}^{k-1} A'_i\eta_i\|s_i\|^2 \le \frac{4}{3}R^2.$$

For every completed main iteration, (180) and (183) give

$$\frac{3}{8}A'_k\eta_k\|s_k\|^2 \le \mathcal{E}_k - \mathcal{E}_{k+1} \le \mathcal{E}_k \le \frac{R^2}{2}.$$

The MS test (143) also gives

$$\|\nabla f(y_k)\| \le \|\nabla f(y_k) + \tau_k s_k\| + \tau_k\|s_k\| \le \frac{3}{2}\tau_k\|s_k\|.$$

Using $(a'_k)^2 = A'_k/\eta_k$, we obtain

$$\|s_k\| \le \frac{2R}{\sqrt{3A'_k\eta_k}} = \frac{2a'_k}{\sqrt{3}A'_k}R, \qquad \|\nabla f(y_k)\| \le \frac{3}{2}\tau_k\|s_k\|. \quad (184)$$

Lemma F.3. *For every $k \ge 1$, the points w_k, z_k, v_k generated by Algorithm 7 satisfy*

$$\|z_k - x^*\| \le R, \quad (185)$$

$$\|w_k - x^*\| \le \left(1 + \frac{2}{\sqrt{3}}\right)R, \qquad \|v_k - x^*\| \le \left(1 + \frac{2}{\sqrt{3}}\right)R. \quad (186)$$

Proof. By (183),

$$\frac{1}{2}\|z_k - x^*\|^2 \le \mathcal{E}_k \le \frac{R^2}{2},$$

which gives (185). We prove the bound on $\|w_k - x^*\|$ by induction. The initialization bound (179) gives

$$\|w_1 - x^*\| \le \|w_1 - x_0\| + \|x_0 - x^*\| \le \left(1 + \frac{2}{\sqrt{3}}\right) R.$$

Suppose the bound holds for w_k. Substituting $y_k = v_k + s_k$ and $A'_k v_k = A_k w_k + a'_k z_k$ into (178) yields

$$w_{k+1} = \frac{(1-\gamma_k)A_k w_k + \gamma_k A'_k (v_k + s_k)}{A_{k+1}} = \frac{A_k}{A_{k+1}} w_k + \frac{a_{k+1}}{A_{k+1}} z_k + \frac{\gamma_k A'_k}{A_{k+1}} s_k.$$

Using (184), (185), and $\gamma_k a'_k = a_{k+1}$, we obtain

$$\begin{aligned}\|w_{k+1} - x^*\| &\le \frac{A_k}{A_{k+1}}\|w_k - x^*\| + \frac{a_{k+1}}{A_{k+1}}\|z_k - x^*\| + \frac{\gamma_k A'_k}{A_{k+1}}\|s_k\| \\ &\le \frac{A_k}{A_{k+1}}\left(1 + \frac{2}{\sqrt{3}}\right) R + \frac{a_{k+1}}{A_{k+1}} R + \frac{2a_{k+1}}{\sqrt{3}A_{k+1}} R = \left(1 + \frac{2}{\sqrt{3}}\right) R.\end{aligned}$$

This completes the induction. Finally,

$$\|v_k - x^*\| \le \frac{A_k}{A'_k}\|w_k - x^*\| + \frac{a'_k}{A'_k}\|z_k - x^*\| \le \frac{A_k}{A'_k}\left(1 + \frac{2}{\sqrt{3}}\right) R + \frac{a'_k}{A'_k} R \le \left(1 + \frac{2}{\sqrt{3}}\right) R,$$

which proves (186). □

For iterations of Algorithm 7 with $\gamma_k < 1$, apply Lemma C.3 with $x = v_k$, $y = y_k$, $\eta = \eta_k$, and $\tau = \tau_k$. If $\tau_k \ge 16\delta$, (20) bounds the step length from below. As in Carmon et al. (2022, Appendix A), we combine this bound with (183) to control the number of these iterations.

F.3 Proof of Theorem 4.1

Define

$$\begin{aligned} B &:= \begin{cases} \max\left\{1, \left(6\sqrt{M_1}R\right)^{2/3}\right\}, & \text{under (GH)}, \\ \max\left\{1, \left(6\sqrt{M_1}R\right)^{2/3}, \left(\frac{2R}{\rho}\right)^{2/3}\right\}, & \text{under (LH)}, \end{cases} \\ \varepsilon_{\mathrm{tr}} &:= \max\left\{\frac{48M_0R^3}{B^{7/2}}, \frac{16\delta R^2}{B^2}\right\}. \end{aligned} \tag{187}$$

Equivalently, with r_0 defined in (18),$B = \max\left\{1, \left(\frac{R}{r_0}\right)^{2/3}\right\}$, where $R/(+\infty) = 0$. For $A > 0$, define

$$\Lambda(A) := \max\left\{32\delta, (96M_0R)^{4/7}A^{-3/7}, \frac{B^2}{A}\right\}. \tag{188}$$

The function Λ is nonincreasing, and $A\Lambda(A)$ is nondecreasing with $A\Lambda(A) \ge B^2 \ge 1$.

Lemma F.4. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Assume that $R > 0$ and set*

$$A_{\mathrm{tr}} := \min\left\{\frac{B^{7/2}}{96M_0R}, \frac{B^2}{32\delta}\right\}, \tag{189}$$

where $B^2/(32\delta) = +\infty$ when $\delta = 0$. The weights A_k, $k \ge 1$, generated by Algorithm 7 satisfy

$$k - 1 \le \frac{16}{3}B + 2\log_2 B + \left(6B + \frac{3}{2}\right)\log_+ \frac{\min\{A_k, A_{\mathrm{tr}}\}}{A_1} + 35\left(\sqrt{A_k\Lambda(A_k)} - B\right). \tag{190}$$

Proof. Fix $k \geq 1$ and define

$$\mathcal{S}_{k-1}^{=} := \{i \in \{1, \dots, k-1\} : \tau_i = \eta_i,\ \eta_i \leq 2\Lambda(A_i)\},$$
$$\mathcal{S}_{k-1}^{>} := \{i \in \{1, \dots, k-1\} : \tau_i > \eta_i,\ \eta_i \geq \Lambda(A_i)/2\}.$$

We first bound $|\mathcal{S}_{k-1}^{=}|$ in terms of A_k. On these iterations, the bound $\eta_i \leq 2\Lambda(A_i)$ gives an increase of A_i. For $i \in \mathcal{S}_{k-1}^{=}$, we have $\gamma_i = 1$, $A_{i+1} = A_i'$, and $(a_i')^2 = A_i'/\eta_i$. Hence,

$$\frac{a_i'}{A_i} = \sqrt{\frac{A_i'}{\eta_i A_i^2}} \geq \frac{1}{\sqrt{\eta_i A_i}} \geq \frac{1}{\sqrt{2A_i\Lambda(A_i)}}.$$

Using $\log(1+t) \geq t/(1+t)$ for $t \geq 0$ gives

$$\log \frac{A_{i+1}}{A_i} = \log\left(1 + \frac{a_i'}{A_i}\right) \geq \log\left(1 + \frac{1}{\sqrt{2A_i\Lambda(A_i)}}\right) \geq \frac{1}{\sqrt{2A_i\Lambda(A_i)}+1} \geq \frac{1}{3\sqrt{A_i\Lambda(A_i)}}.$$

Since $\sqrt{t\Lambda(t)}$ is nondecreasing, each $i \in \mathcal{S}_{k-1}^{=}$ therefore satisfies

$$\int_{A_i}^{A_{i+1}} \sqrt{\frac{\Lambda(t)}{t}}dt = \int_{A_i}^{A_{i+1}} \frac{\sqrt{t\Lambda(t)}}{t}dt \geq \sqrt{A_i\Lambda(A_i)}\log\frac{A_{i+1}}{A_i} \geq \frac{1}{3}.$$

Summing over $\mathcal{S}_{k-1}^{=}$ and using the nonnegativity of the integrand, we obtain

$$|\mathcal{S}_{k-1}^{=}| = \sum_{i\in\mathcal{S}_{k-1}^{=}} 1 \leq 3\sum_{i\in\mathcal{S}_{k-1}^{=}} \int_{A_i}^{A_{i+1}} \sqrt{\frac{\Lambda(t)}{t}}dt \leq 3\sum_{i=1}^{k-1}\int_{A_i}^{A_{i+1}} \sqrt{\frac{\Lambda(t)}{t}}dt = 3\int_{A_1}^{A_k}\sqrt{\frac{\Lambda(t)}{t}}dt. \tag{191}$$

We next bound $|\mathcal{S}_{k-1}^{>}|$ using (183). We first bound the term $\frac{3}{8}A_i'\eta_i\|s_i\|^2$ from below. For $i \in \mathcal{S}_{k-1}^{>}$, we have $\tau_i > \eta_i \geq \frac{\Lambda(A_i)}{2} \geq 16\delta$. Thus, by Lemma C.3, (20) gives

$$\|s_i\| \geq \min\left\{r_0, \frac{\tau_i}{48M_0}\right\} \geq \min\left\{r_0, \frac{\Lambda(A_i)}{96M_0}\right\}.$$

Since $A_i' \geq A_i$ and $\eta_i \geq \Lambda(A_i)/2$, it follows that

$$\begin{aligned}\frac{3}{8}A_i'\eta_i\|s_i\|^2 &\geq \frac{3}{16}A_i\Lambda(A_i)\min\left\{r_0^2, \frac{\Lambda(A_i)^2}{(96M_0)^2}\right\} \\ &= \frac{3}{16\sqrt{A_i\Lambda(A_i)}}\min\left\{(A_i\Lambda(A_i))^{3/2}\, r_0^2, \frac{A_i^{3/2}\Lambda(A_i)^{7/2}}{(96M_0)^2}\right\}.\end{aligned} \tag{192}$$

We now use the definition of Λ (188) to bound both terms in the last minimum. For $r_0 < +\infty$, the definition of B and (188) give

$$(A_i\Lambda(A_i))^{3/2}\, r_0^2 \geq B^3 r_0^2 \geq \left(\frac{R}{r_0}\right)^2 r_0^2 = R^2.$$

Also, (188) gives

$$\frac{A_i^{3/2}\Lambda(A_i)^{7/2}}{(96M_0)^2} \geq \frac{A_i^{3/2}\left((96M_0R)^{4/7}A_i^{-3/7}\right)^{7/2}}{(96M_0)^2} = R^2.$$

Substituting these bounds in (192) and using that $A_i \leq A_k$ and $A\Lambda(A)$ is nondecreasing, we obtain

$$\frac{3}{8}A_i'\eta_i\|s_i\|^2 \geq \frac{3R^2}{16\sqrt{A_i\Lambda(A_i)}} \geq \frac{3R^2}{16\sqrt{A_k\Lambda(A_k)}}.$$

Summing over $i \in \mathcal{S}_{k-1}^{>}$ and using (183), we obtain

$$\frac{3R^2}{16\sqrt{A_k\Lambda(A_k)}}|\mathcal{S}_{k-1}^{>}| = \sum_{i\in\mathcal{S}_{k-1}^{>}} \frac{3R^2}{16\sqrt{A_k\Lambda(A_k)}} \leq \frac{3}{8}\sum_{i\in\mathcal{S}_{k-1}^{>}} A_i'\eta_i\|s_i\|^2 \leq \frac{3}{8}\sum_{i=1}^{k-1} A_i'\eta_i\|s_i\|^2 \leq \frac{R^2}{2}.$$

Therefore,

$$|\mathcal{S}^{>}_{k-1}| \le \frac{8}{3}\sqrt{A_k\Lambda(A_k)}. \tag{193}$$

It remains to bound the number $(k-1) - |\mathcal{S}^{=}_{k-1}| - |\mathcal{S}^{>}_{k-1}|$ of iterations outside $\mathcal{S}^{=}_{k-1} \cup \mathcal{S}^{>}_{k-1}$. For these iterations, there are two cases:

- First, if $\tau_i = \eta_i$, then $\eta_i > 2\Lambda(A_i)$, $\eta_{i+1} = \frac{\eta_i}{2}$.
- Second, if $\tau_i > \eta_i$, then $\eta_i < \frac{\Lambda(A_i)}{2}$, $\eta_{i+1} = 2\eta_i$.

To measure the change relative to $\Lambda(A_i)$, set $V_i := \left|\log \frac{\eta_i}{\Lambda(A_i)}\right|$. In the first case, $\eta_{i+1} > \Lambda(A_i)$, so

$$\left|\log \frac{\eta_{i+1}}{\Lambda(A_i)}\right| = \log \frac{\eta_i}{2\Lambda(A_i)} = \log \frac{\eta_i}{\Lambda(A_i)} - \log 2 = V_i - \log 2.$$

In the second case, $\eta_{i+1} < \Lambda(A_i)$, so

$$\left|\log \frac{\eta_{i+1}}{\Lambda(A_i)}\right| = -\log \frac{2\eta_i}{\Lambda(A_i)} = -\log \frac{\eta_i}{\Lambda(A_i)} - \log 2 = V_i - \log 2.$$

For $i \in \mathcal{S}^{=}_{k-1} \cup \mathcal{S}^{>}_{k-1}$, the update of η_i gives $\left|\log \frac{\eta_{i+1}}{\Lambda(A_i)}\right| \le V_i + \log 2$.

Since $\Lambda(A_{i+1}) \le \Lambda(A_i)$,

$$V_{i+1} = \left|\log \frac{\eta_{i+1}}{\Lambda(A_i)} + \log \frac{\Lambda(A_i)}{\Lambda(A_{i+1})}\right| \le \left|\log \frac{\eta_{i+1}}{\Lambda(A_i)}\right| + \log \frac{\Lambda(A_i)}{\Lambda(A_{i+1})}.$$

Summing over $i = 1, \dots, k-1$ and telescoping the logarithms gives

$$\begin{aligned} V_k - V_1 &\le \left(|\mathcal{S}^{=}_{k-1}| + |\mathcal{S}^{>}_{k-1}|\right)\log 2 - \left((k-1) - |\mathcal{S}^{=}_{k-1}| - |\mathcal{S}^{>}_{k-1}|\right)\log 2 + \sum_{i=1}^{k-1} \log \frac{\Lambda(A_i)}{\Lambda(A_{i+1})} \\ &= \left(2\left(|\mathcal{S}^{=}_{k-1}| + |\mathcal{S}^{>}_{k-1}|\right) - (k-1)\right)\log 2 + \log \frac{\Lambda(A_1)}{\Lambda(A_k)}. \end{aligned}$$

Initialization gives $\eta_1 = 1/A_1$. Since $A_1\Lambda(A_1) \ge 1$,

$$V_1 = \left|\log \frac{1}{A_1\Lambda(A_1)}\right| = \log\left(A_1\Lambda(A_1)\right).$$

Also, monotonicity of $A\Lambda(A)$ implies

$$\log \frac{\Lambda(A_1)}{\Lambda(A_k)} = \log \frac{A_k}{A_1} + \log \frac{A_1\Lambda(A_1)}{A_k\Lambda(A_k)} \le \log \frac{A_k}{A_1}.$$

Using $V_k \ge 0$, combining the logarithms, and dividing by $\log 2$, we obtain

$$k - 1 \le 2\left(|\mathcal{S}^{=}_{k-1}| + |\mathcal{S}^{>}_{k-1}|\right) + \log_2\left(A_k\Lambda(A_1)\right). \tag{194}$$

Substituting (191) and (193) into (194) gives

$$k - 1 \le 6\int_{A_1}^{A_k} \sqrt{\frac{\Lambda(t)}{t}}\,dt + \frac{16}{3}\sqrt{A_k\Lambda(A_k)} + \log_2\left(A_k\Lambda(A_1)\right). \tag{195}$$

We now estimate the integral and logarithmic term in (195). By (188) and (189),

$$\sqrt{A\Lambda(A)} = \begin{cases} B, & 0 < A \le A_{\mathrm{tr}}, \\ \max\left\{(96M_0RA)^{2/7}, \sqrt{32\delta A}\right\}, & A > A_{\mathrm{tr}}. \end{cases}$$

Consider an interval $[u, v] \subset [A_{\mathrm{tr}}, +\infty)$ on which the same term attains the maximum. If $\sqrt{t\Lambda(t)} = (96M_0Rt)^{2/7}$ on this interval, then

$$\int_u^v \sqrt{\frac{\Lambda(t)}{t}}dt = (96M_0R)^{2/7}\int_u^v t^{-5/7}dt = \frac{7}{2}\left(\sqrt{v\Lambda(v)} - \sqrt{u\Lambda(u)}\right), \quad \log\frac{v}{u} = \frac{7}{2}\log\frac{\sqrt{v\Lambda(v)}}{\sqrt{u\Lambda(u)}}.$$

If $\sqrt{t\Lambda(t)} = \sqrt{32\delta t}$ on this interval, then

$$\int_u^v \sqrt{\frac{\Lambda(t)}{t}}dt = \sqrt{32\delta}\int_u^v t^{-1/2}dt = 2\left(\sqrt{v\Lambda(v)} - \sqrt{u\Lambda(u)}\right), \quad \log\frac{v}{u} = 2\log\frac{\sqrt{v\Lambda(v)}}{\sqrt{u\Lambda(u)}}.$$

Both coefficients are at most $7/2$. Summing over the intervals above A_{tr}, the intermediate endpoint terms cancel in both bounds. Below A_{tr}, we have $\sqrt{t\Lambda(t)} = B$ and integrate B/t. Thus, for every $A \geq A_1$,

$$\int_{A_1}^A \sqrt{\frac{\Lambda(t)}{t}}dt \leq B\log_+\frac{\min\{A, A_{\mathrm{tr}}\}}{A_1} + \frac{7}{2}\left(\sqrt{A\Lambda(A)} - B\right), \tag{196}$$

$$\log\frac{A}{A_1} \leq \log_+\frac{\min\{A, A_{\mathrm{tr}}\}}{A_1} + \frac{7}{2}\log\frac{\sqrt{A\Lambda(A)}}{B}. \tag{197}$$

These bounds also hold when $A_1 > A_{\mathrm{tr}}$, since $\sqrt{A_1\Lambda(A_1)} \geq B$. Using $\log_2 t \leq \frac{3}{2}\log t$ for $t \geq 1$, (197), and the monotonicity of $A\Lambda(A)$, we obtain

$$\begin{aligned}\log_2\left(A_k\Lambda(A_1)\right) = \log_2\frac{A_k}{A_1} + \log_2\left(A_1\Lambda(A_1)\right) &\leq \frac{3}{2}\log\frac{A_k}{A_1} + 2\log_2 B + 3\log\frac{\sqrt{A_k\Lambda(A_k)}}{B} \\ &\leq 2\log_2 B + \frac{3}{2}\log_+\frac{\min\{A_k, A_{\mathrm{tr}}\}}{A_1} + \frac{33}{4}\log\frac{\sqrt{A_k\Lambda(A_k)}}{B}. \end{aligned} \tag{198}$$

Substituting (196) and (198) into (195), we obtain

$$\begin{aligned}k - 1 \leq \frac{16}{3}B + 2\log_2 B + \left(6B + \frac{3}{2}\right)\log_+\frac{\min\{A_k, A_{\mathrm{tr}}\}}{A_1} \\ + \frac{79}{3}\left(\sqrt{A_k\Lambda(A_k)} - B\right) + \frac{33}{4}\log\frac{\sqrt{A_k\Lambda(A_k)}}{B}.\end{aligned}$$

Finally, since $\sqrt{A_k\Lambda(A_k)} \geq B \geq 1$,

$$\log\frac{\sqrt{A_k\Lambda(A_k)}}{B} \leq \frac{\sqrt{A_k\Lambda(A_k)} - B}{B}, \quad \frac{79}{3} + \frac{33}{4B} \leq \frac{415}{12} < 35.$$

Substituting these inequalities proves (190). □

We restate Theorem 4.1 with explicit constants.

Theorem F.5. *Suppose that f is convex,* (IH) *holds, and either* (GH) *or* (LH) *holds. Let $\overline{\tau}_0 := \overline{\tau}(x_0)$, where $\overline{\tau}(x)$ is defined in* (17)*. For every $\varepsilon > 0$, the method generates an iterate w_N satisfying $f(w_N) - f^* \leq \varepsilon$ within*

$$\begin{aligned}N \leq 2 + \frac{16}{3}B + 2\log_2 B + \left(6B + \frac{3}{2}\right)\log_+\frac{\max\{\eta_0, \overline{\tau}_0\}R^2}{\max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}} \\ + 35\left(\max\left\{B, \left(\frac{48M_0R^3}{\varepsilon}\right)^{2/7}, 4\sqrt{\frac{\delta R^2}{\varepsilon}}\right\} - B\right)\end{aligned} \tag{199}$$

outer iterations, including initialization, each using at most one Hessian evaluation. Moreover, set

$$K := 2 + \left\lfloor \frac{16}{3}B + 2\log_2 B + \left(6B + \frac{3}{2}\right)\log_+\frac{\max\{\eta_0, \overline{\tau}_0\}R^2}{\varepsilon_{\mathrm{tr}}} \right\rfloor. \tag{200}$$

For this K,

$$f(w_k) - f^* \le \max\{\eta_0, \overline{\tau}_0\} R^2 \exp\left(-\frac{\max\left\{0, k-1-\frac{16}{3}B - 2\log_2 B\right\}}{6B+\frac{3}{2}}\right), \qquad 1 \le k < K, \tag{201}$$

$$f(w_{K+n}) - f^* \le \max\left\{\frac{48M_0R^3}{(B+n/35)^{7/2}}, \frac{16\delta R^2}{(B+n/35)^2}\right\}, \qquad n = 0, 1, \dots. \tag{202}$$

Proof. Lemmas C.2 and F.1 show that all backtracking loops terminate. If initialization returns a minimizer, all claims are immediate. By (183),

$$f(w_k) - f^* \le \frac{\mathcal{E}_k}{A_k} \le \frac{R^2}{2A_k}. \tag{203}$$

Since $a_{k+1} > 0$, the A_k strictly increase.

Fix $\varepsilon > 0$ and set $A_\varepsilon := R^2/(2\varepsilon)$. If $A_1 \ge A_\varepsilon$, initialization suffices by (203). Otherwise, by (190), unless a minimizer is returned earlier, there exists a first index N with $A_N \ge A_\varepsilon$. Then $N \ge 2$, $A_{N-1} < A_\varepsilon \le A_N$, and (203) gives $f(w_N) - f^* \le \varepsilon$. Applying (190) at $N-1$ and using monotonicity of its right-hand side, we obtain

$$N \le 2 + \frac{16}{3}B + 2\log_2 B + \left(6B+\frac{3}{2}\right)\log_+ \frac{\min\{A_\varepsilon, A_{\mathrm{tr}}\}}{A_1} + 35\left(\sqrt{A_\varepsilon \Lambda(A_\varepsilon)} - B\right).$$

By (187) and (189), $A_{\mathrm{tr}} = R^2/(2\varepsilon_{\mathrm{tr}})$. Hence,

$$\log_+ \frac{\min\{A_\varepsilon, A_{\mathrm{tr}}\}}{A_1} = \log_+ \frac{R^2}{2A_1 \max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}},$$

$$\sqrt{A_\varepsilon \Lambda(A_\varepsilon)} = \max\left\{B, \left(\frac{48M_0R^3}{\varepsilon}\right)^{2/7}, 4\sqrt{\frac{\delta R^2}{\varepsilon}}\right\}.$$

Next, by Lemma F.1 and (175) we have

$$\frac{1}{A_1} = \eta_1 \le \max\{\eta_0, 2\overline{\tau}_0\} \le 2\max\{\eta_0, \overline{\tau}_0\}. \tag{204}$$

Hence,

$$\frac{R^2}{2A_1} \le \max\{\eta_0, \overline{\tau}_0\} R^2, \qquad 2\max\{\eta_0, \overline{\tau}_0\} A_{\mathrm{tr}} = \frac{\max\{\eta_0, \overline{\tau}_0\} R^2}{\varepsilon_{\mathrm{tr}}}, \tag{205}$$

and

$$\log_+ \frac{R^2}{2A_1 \max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}} \le \log_+ \frac{\max\{\eta_0, \overline{\tau}_0\} R^2}{\max\{\varepsilon, \varepsilon_{\mathrm{tr}}\}}.$$

Substituting these estimates proves (199). If a zero gradient is queried before reaching A_ε, let N count the outer iterations up to that return. The last completed weight still satisfies $A_{N-1} < A_\varepsilon$, so the same estimate applies, and the returned point is a minimizer. There is one Hessian evaluation during initialization and at most one on each main iteration, since all inner trials reuse that Hessian.

We now prove (201) and (202) for iterates produced before a zero-gradient return. The right-hand side of (190) is nonincreasing in A_1, and $1/A_1 \le 2\max\{\eta_0, \overline{\tau}_0\}$ by (204). Hence, (190) remains valid with $1/A_1$ replaced by $2\max\{\eta_0, \overline{\tau}_0\}$. First, let $1 \le k < K$. If $A_k \le A_{\mathrm{tr}}$, then $\sqrt{A_k\Lambda(A_k)} = B$, and (190) gives

$$k - 1 \le \frac{16}{3}B + 2\log_2 B + \left(6B+\frac{3}{2}\right)\log\left(2\max\{\eta_0, \overline{\tau}_0\} A_k\right).$$

If $A_k > A_{\mathrm{tr}}$, then (200), (205), and $2\max\{\eta_0, \overline{\tau}_0\} A_k \ge 1$ give

$$\begin{aligned} k - 1 \le K - 2 &\le \frac{16}{3}B + 2\log_2 B + \left(6B+\frac{3}{2}\right)\log_+\left(2\max\{\eta_0, \overline{\tau}_0\} A_{\mathrm{tr}}\right) \\ &\le \frac{16}{3}B + 2\log_2 B + \left(6B+\frac{3}{2}\right)\log\left(2\max\{\eta_0, \overline{\tau}_0\} A_k\right). \end{aligned}$$

Thus, in both cases, since $\log\left(2\max\left\{\eta_0,\overline{\tau}_0\right\}A_k\right)\geq 0$,

$$\log\left(2\max\left\{\eta_0,\overline{\tau}_0\right\}A_k\right)\geq\frac{\max\left\{0,k-1-\frac{16}{3}B-2\log_2 B\right\}}{6B+\frac{3}{2}}.$$

Combining this bound with (203) proves (201).

Next, let $k=K+n$ with $n\geq 0$. By (200) and (205),

$$\frac{16}{3}B+2\log_2 B+\left(6B+\frac{3}{2}\right)\log_+\left(2\max\left\{\eta_0,\overline{\tau}_0\right\}A_{\mathrm{tr}}\right)+n<k-1.$$

Since $\min\left\{A_k,A_{\mathrm{tr}}\right\}\leq A_{\mathrm{tr}}$, (190) also gives

$$k-1\leq\frac{16}{3}B+2\log_2 B+\left(6B+\frac{3}{2}\right)\log_+\left(2\max\left\{\eta_0,\overline{\tau}_0\right\}A_{\mathrm{tr}}\right)+35\left(\sqrt{A_k\Lambda(A_k)}-B\right).$$

Combining these inequalities yields

$$\sqrt{A_k\Lambda(A_k)}>B+\frac{n}{35}\geq B.$$

Therefore, by (188),

$$\max\left\{(96M_0RA_k)^{2/7},\sqrt{32\delta A_k}\right\}>B+\frac{n}{35},\quad \frac{1}{A_k}\leq\max\left\{\frac{96M_0R}{(B+n/35)^{7/2}},\frac{32\delta}{(B+n/35)^2}\right\}.$$

Multiplying the last inequality by $R^2/2$ and using (203) proves (202). After a zero-gradient return, the output sequence consists of minimizers, so both rate bounds continue to hold. □

F.4 Oracle Complexity

In this section, we bound the numbers of Hessian and gradient evaluations of Algorithm 7. Let

$$\overline{g}:=\max_{\|x-x^*\|\leq(1+2/\sqrt{3})R}\|\nabla f(x)\|. \tag{206}$$

By (186), $\|\nabla f(v_i)\|\leq\overline{g}$ for any v_i. Recall that r_0 is defined in (18).

Corollary F.6. *Under the assumptions of Theorem F.5, fix $\varepsilon>0$. Let N be the number of outer iterations, including initialization, until $A_N\geq R^2/(2\varepsilon)$ or an earlier zero-gradient return. Then Algorithm 7 uses at most N Hessian evaluations and at most*

$$\begin{aligned}&7N+\log_{2,+}\frac{\overline{\tau}_0}{\eta_0}+30\sqrt{\frac{2\delta R^2}{\varepsilon}}+120\left(\frac{2M_0R^3}{\varepsilon}\right)^{2/7}\\&+8B\log_{2,+}\frac{\delta+\overline{g}/r_0}{\delta+M_0r_0}\left(1+\log_+\frac{4\max\left\{\eta_0,\overline{\tau}_0\right\}R^2}{\varepsilon}\right)^{2/3}\end{aligned} \tag{207}$$

gradient evaluations.

We prove Corollary F.6 after three auxiliary lemmas. Lemma F.9 bounds the total number of rejected trials. Its proof uses the bounds on the sums of $\eta_i^{-1/2}$ and $(\|s_i\|/\eta_i)^{2/7}$ from Lemma F.7 and the bound on the number of rejected iterations with $\|s_i\|>r_0$ from Lemma F.8. By (183) and $A_i\leq A_i'$,

$$\sum_{i=1}^{m}A_i\eta_i\|s_i\|^2\leq\frac{4}{3}R^2,\qquad m\geq 1. \tag{208}$$

Lemma F.7. *Under the assumptions of Theorem F.5, every completed main iteration of Algorithm 7 satisfies*

$$\eta_iA_i\geq\frac{1}{2},\qquad A_{i+1}\leq(2+\sqrt{3})A_i<4A_i. \tag{209}$$

For every $m\geq 1$ and $A=A_{m+1}$,

$$\sum_{i=1}^{m}\eta_i^{-1/2}\leq 5\sqrt{A},\qquad\sum_{i=1}^{m}\left(\frac{\|s_i\|}{\eta_i}\right)^{2/7}\leq 20(RA)^{2/7}, \tag{210}$$

$$\sum_{i=1}^{m}\frac{1}{\sqrt{\eta_iA_i}}\leq 8\log 6+4\log\frac{A}{A_1}. \tag{211}$$

Proof. Initialization gives $\eta_1 A_1 = 1$. If $\tau_i > \eta_i$, then $\eta_{i+1} = 2\eta_i$ and $A_{i+1} > A_i$. If $\tau_i = \eta_i$, then $A_{i+1} = A_i'$, $a_i' \geq 1/\eta_i$, and

$$\eta_{i+1} A_{i+1} = \frac{1}{2}\eta_i A_i' = \frac{1}{2}(\eta_i a_i')^2 \geq \frac{1}{2}.$$

Induction proves the first bound in (209). Since $\eta_i A_i \geq 1/2$ and $0 < \gamma_i \leq 1$,

$$\frac{A_{i+1}}{A_i} \leq \frac{A_i'}{A_i} = 1 + \frac{1 + \sqrt{1 + 4\eta_i A_i}}{2\eta_i A_i} \leq 2 + \sqrt{3} < 4.$$

On every iteration with $\tau_i = \eta_i$, we have $a_i' = \sqrt{A_{i+1}/\eta_i}$ and

$$\sqrt{A_{i+1}} - \sqrt{A_i} = \frac{a_i'}{\sqrt{A_{i+1}} + \sqrt{A_i}} = \frac{\sqrt{A_{i+1}}}{\sqrt{\eta_i}\left(\sqrt{A_{i+1}} + \sqrt{A_i}\right)} \geq \frac{1}{2\sqrt{\eta_i}}.$$

All weight increments are nonnegative, so

$$\sum_{\substack{1 \leq i \leq m \\ \tau_i = \eta_i}} \eta_i^{-1/2} \leq 2(\sqrt{A} - \sqrt{A_1}).$$

The update of η_i gives $\eta_{i+1}^{-1/2} = \sqrt{2}\eta_i^{-1/2}$ when $\tau_i = \eta_i$, and $\eta_{i+1}^{-1/2} = \eta_i^{-1/2}/\sqrt{2}$ otherwise. Telescoping and $\eta_1^{-1/2} = \sqrt{A_1}$ yield

$$(1 - 1/\sqrt{2}) \sum_{\substack{1 \leq i \leq m \\ \tau_i > \eta_i}} \eta_i^{-1/2} = (\sqrt{2} - 1) \sum_{\substack{1 \leq i \leq m \\ \tau_i = \eta_i}} \eta_i^{-1/2} + \eta_1^{-1/2} - \eta_{m+1}^{-1/2},$$

$$\sum_{i=1}^{m} \eta_i^{-1/2} \leq (1 + \sqrt{2}) \sum_{\substack{1 \leq i \leq m \\ \tau_i = \eta_i}} \eta_i^{-1/2} + (2 + \sqrt{2})\sqrt{A_1} \leq 2(1 + \sqrt{2})\sqrt{A} - \sqrt{2}\sqrt{A_1} \leq 5\sqrt{A}.$$

Set $S_0 = \sqrt{A_1}$ and $S_i = S_0 + \sum_{j=1}^{i} \eta_j^{-1/2}$. Applying the preceding bound to each shorter prefix gives

$$S_m \leq 6\sqrt{A}, \qquad S_{i-1} \leq 6\sqrt{A_i}, \qquad S_i \leq (6 + \sqrt{2})\sqrt{A_i} \leq 8\sqrt{A_i}. \tag{212}$$

Here $\eta_i^{-1/2} \leq \sqrt{2A_i}$ follows from (209). Since $\eta_i^{-1/2} = S_i - S_{i-1}$, integration of the decreasing function $1/t$ yields

$$\sum_{i=1}^{m} \frac{1}{\sqrt{\eta_i A_i}} \leq 8 \sum_{i=1}^{m} \frac{S_i - S_{i-1}}{S_i} \leq 8 \int_{S_0}^{S_m} \frac{dt}{t} \leq 8 \log 6 + 4 \log \frac{A}{A_1}.$$

This proves (211).

For the second bound in (210), Hölder's inequality gives

$$\sum_{i=1}^{m} \left(\frac{\|s_i\|}{\eta_i}\right)^{2/7} = \sum_{i=1}^{m} \left(A_i \eta_i \|s_i\|^2\right)^{1/7} \left(\frac{1}{\eta_i^{1/2} A_i^{1/6}}\right)^{6/7} \leq \left(\sum_{i=1}^{m} A_i \eta_i \|s_i\|^2\right)^{1/7} \left(\sum_{i=1}^{m} \frac{1}{\eta_i^{1/2} A_i^{1/6}}\right)^{6/7},$$

$$\sum_{i=1}^{m} \frac{1}{\eta_i^{1/2} A_i^{1/6}} \leq 2 \sum_{i=1}^{m} \frac{S_i - S_{i-1}}{S_i^{1/3}} \leq 2 \int_{S_0}^{S_m} t^{-1/3} dt \leq 3 S_m^{2/3} \leq 3 \cdot 6^{2/3} A^{1/3}.$$

The last display uses (212). Substituting these bounds and (208) gives

$$\sum_{i=1}^{m} \left(\frac{\|s_i\|}{\eta_i}\right)^{2/7} \leq (4R^2/3)^{1/7} (3 \cdot 6^{2/3} A^{1/3})^{6/7} \leq 20 (RA)^{2/7}.$$

□

Lemma F.8. *Under the assumptions of Theorem F.5, for $m \geq 1$ and $A = A_{m+1}$, define*

$$\mathcal{I}_m := \{i \in \{1, \dots, m\} : \tau_i > \eta_i,\ \|s_i\| > r_0\}.$$

Then

$$|\mathcal{I}_m| \leq 8B \left(1 + \log \frac{A}{A_1}\right)^{2/3}. \tag{213}$$

If $r_0 = +\infty$, then $\mathcal{I}_m$ is empty.

Proof. Let $r_0 < +\infty$. By (187), $(R/r_0)^{2/3} \leq B$. Hölder's inequality, (208), and (211) give

$$|\mathcal{I}_m| \leq r_0^{-2/3} \sum_{i \in \mathcal{I}_m} \|s_i\|^{2/3} = r_0^{-2/3} \sum_{i \in \mathcal{I}_m} \left(A_i \eta_i \|s_i\|^2\right)^{1/3} \left(\frac{1}{\sqrt{\eta_i A_i}}\right)^{2/3}$$

$$\leq r_0^{-2/3} \left(\sum_{i=1}^m A_i \eta_i \|s_i\|^2\right)^{1/3} \left(\sum_{i=1}^m \frac{1}{\sqrt{\eta_i A_i}}\right)^{2/3} \leq (4/3)^{1/3} B \left(8 \log 6 + 4 \log \frac{A}{A_1}\right)^{2/3}.$$

Since $(4/3)^{1/3}(8 \log 6)^{2/3} < 8$ and $8 \log 6 > 4$, this proves (213). □

Lemma F.9. *Under the assumptions of Theorem F.5, let $b_i := \log_2(\tau_i/\eta_i)$ be the number of rejected trials in main iteration i, including failed* PD *tests. For every $m \geq 1$ and $A = A_{m+1}$,*

$$\sum_{i=1}^m b_i \leq 5m + 30\sqrt{\delta A} + 120(M_0 R A)^{2/7} + 8B \log_{2,+} \frac{\delta + \overline{g}/r_0}{\delta + M_0 r_0} \left(1 + \log \frac{A}{A_1}\right)^{2/3}, \tag{214}$$

where the last term is absent if $r_0 = +\infty$.

Proof. If $\tau_i = \eta_i$, then $b_i = 0$. Let $\tau_i > \eta_i$. If $\|s_i\| \leq r_0$, then Lemma C.3 gives $\tau_i \leq 8\delta + 24 M_0 \|s_i\|$. If $\|s_i\| > r_0$, then (IH), convexity, and (206) give

$$(\tau_i - \delta)\|s_i\|^2 \leq \langle (H(v_i) + \tau_i I) s_i, s_i \rangle = -\langle \nabla f(v_i), s_i \rangle \leq \overline{g} \|s_i\|,$$

$$\tau_i \leq \delta + \frac{\overline{g}}{r_0} \leq \max\left\{1, \frac{\delta + \overline{g}/r_0}{\delta + M_0 r_0}\right\} (\delta + M_0 \|s_i\|).$$

Since $8\delta + 24 M_0 \|s_i\| \leq 32 \eta_i \left(1 + \frac{\delta}{\eta_i}\right)\left(1 + \frac{M_0 \|s_i\|}{\eta_i}\right)$, in both cases

$$b_i \leq 5 + \log_2 \left(1 + \frac{\delta}{\eta_i}\right) + \log_2 \left(1 + \frac{M_0 \|s_i\|}{\eta_i}\right) + \mathbf{1}_{\{i \in \mathcal{I}_m\}} \log_{2,+} \frac{\delta + \overline{g}/r_0}{\delta + M_0 r_0}.$$

Since $\log(1 + x) = \int_0^x \frac{dt}{1+t} \leq \frac{x^p}{p}$ for $x \geq 0$ and $p \in (0, 1]$, we have $\log_2(1 + x) \leq 6x^p$ for $p \in \left\{\frac{1}{2}, \frac{2}{7}\right\}$. Hence,

$$\sum_{i=1}^m b_i \leq 5m + 6\sqrt{\delta} \sum_{i=1}^m \eta_i^{-1/2} + 6 M_0^{2/7} \sum_{i=1}^m \left(\frac{\|s_i\|}{\eta_i}\right)^{2/7} + |\mathcal{I}_m| \log_{2,+} \frac{\delta + \overline{g}/r_0}{\delta + M_0 r_0}.$$

Substituting (210) and (213) proves (214). □

Proof of Corollary F.6. By Lemma F.1, initialization rejects at most $1 + \log_{2,+} \frac{\overline{\tau}_0}{\eta_0}$ trials. If $A_1 \geq R^2/(2\varepsilon)$, then $N = 1$ and there are no main iterations. Otherwise, the minimality of N and (209) give $A_N < 4A_{N-1} < 2R^2/\varepsilon$. Together with (204), this gives

$$\sqrt{\delta A_N} \leq \sqrt{\frac{2\delta R^2}{\varepsilon}}, \quad (M_0 R A_N)^{2/7} \leq \left(\frac{2 M_0 R^3}{\varepsilon}\right)^{2/7}, \quad \frac{A_N}{A_1} \leq \frac{4 \max\{\eta_0, \overline{\tau}_0\} R^2}{\varepsilon}.$$

Substituting these bounds into (214) with $m = N - 1$ and adding the initialization rejections, we obtain that the total number of rejected trials is at most the right-hand side of (207) with $7N$ replaced by $5N$. Each outer iteration makes one call MSBACKTRACK(x, η), which uses $H(x)$, $\nabla f(x)$, and one gradient for each trial with a positive definite matrix. The updates reuse the gradient at the accepted point. Hence, the method uses at most N Hessian evaluations, and the number of gradient evaluations is at most $2N$ plus the number of rejected trials. This proves (207).

□

G Extra Experiments

G.1 Experiment Details

Table 2 gives the initial parameters for experiments on Figure 1. ACRN and A-ACRN use $\delta_0 = 0$ and adapt $\mathcal{M}$ by doubling at each rejected trial Kamzolov et al. (2025b).

Table 2: Initial parameters of the runs in Figure 1, from $x_0 = -\mathbf{1}$.

	`gisette`	`german_numer`	`ijcnn1`	`abalone`
$L(x_0)$	0.1	0.1	0.1	789.5
$\|\nabla f(x_0)\|$	0.4233	0.3268	0.07210	19.48
$\|x_0\|$	70.71	4.899	4.690	3
η_0 (Algorithms 2, 3, 6)	0.2057	0.1808	0.08491	124.0
H_0 (AdaN)	0.05	0.05	0.05	394.8
γ (Extra-Newton)	141.4	9.798	9.381	6
β_0 (Extra-Newton)	846.6	3.138	0.6345	$5.537 \cdot 10^5$
$\mathcal{M}_0$ (ACRN, A-ACRN, Algorithm 5)	0.1	0.1	0.1	789.5

G.2 Additional Experiments on (LH) Functions

We also run all the methods with exact Hessians on three one-dimensional functions from Gorbunov et al. (2025) and Tyurin (2026) (Figure 2): x^4, x^6 and $e^x + e^{1-x}$. Up to scaling and a shift, x^4 and $e^x + e^{1-x}$ are Examples B.2 and B.1, which satisfy (LH) but not (GH). We start x^4 and x^6 from $x_0 \in \{1, 20, 100\}$ and $e^x + e^{1-x}$ from $x_0 \in \{-1, -24, -48\}$, compute derivatives in closed form, and set all parameters from $L(x_0) = M_0 + M_1|f'(x_0)|$ by the rules of Section 5, where (TD) holds with $(M_0, M_1) = (16, 2)$, $(48, 12)$ and $(0, 1)$, respectively. Every run has 100 iterations.

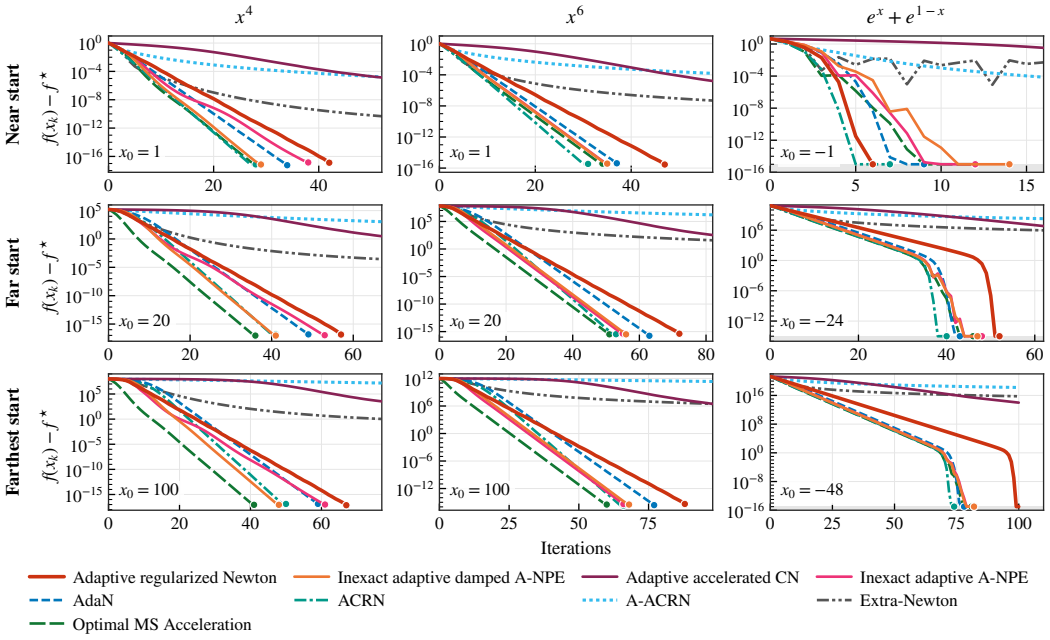

Figure 2: One-dimensional functions with exact Hessians, from the nearest start (top row) to the farthest (bottom row). Each panel ends ten iterations after Algorithm 2 stops.

Results. Six methods converge from every start: Algorithms 2, 3 and 6, AdaN, ACRN and Optimal MS Acceleration (Carmon et al., 2022). On x^4 and x^6 they converge linearly, as expected at a degenerate minimizer. On $e^x + e^{1-x}$ they converge in a few iterations once close to x^*, after a far phase whose length grows linearly with $|x_0 - x^*|$. Unlike in Figure 1, acceleration pays off here:

from far starts, the accelerated Monteiro–Svaiter-type methods, Algorithms 3 and 6 and Optimal MS Acceleration, converge faster than the non-accelerated Algorithm 2.

G.3 Sensitivity to Initial Parameters

We start from $x_0 = -\mathbf{1}$, as in Figure 1, and run Algorithm 2, AdaN and ACRN with exact Hessians. Their initial parameters η_0, H_0 and M_0 take the same value c from the grid $\{10^{-10}, 10^{-9}, \ldots, 10^{10}\}$. Figure 3 shows the number of iterations to reach $f(x_k) - f^* \le 10^{-10}$ for every c, and Figure 4 shows the convergence for five values of c.

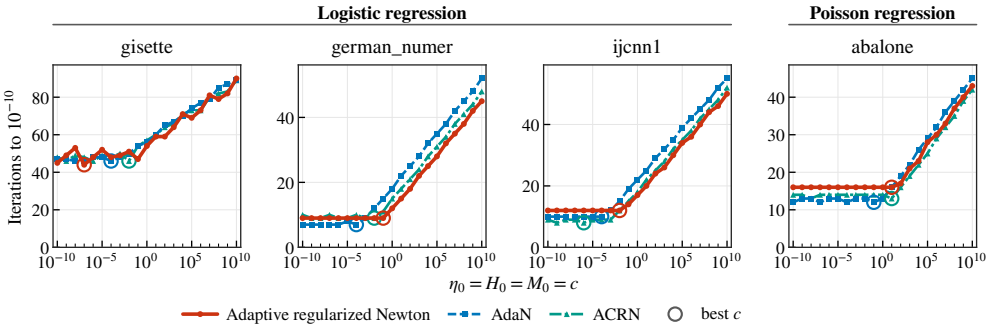

Figure 3: Iterations to reach $f(x_k) - f^* \le 10^{-10}$ for $\eta_0 = H_0 = M_0 = c$. Circles mark the best c of each method.

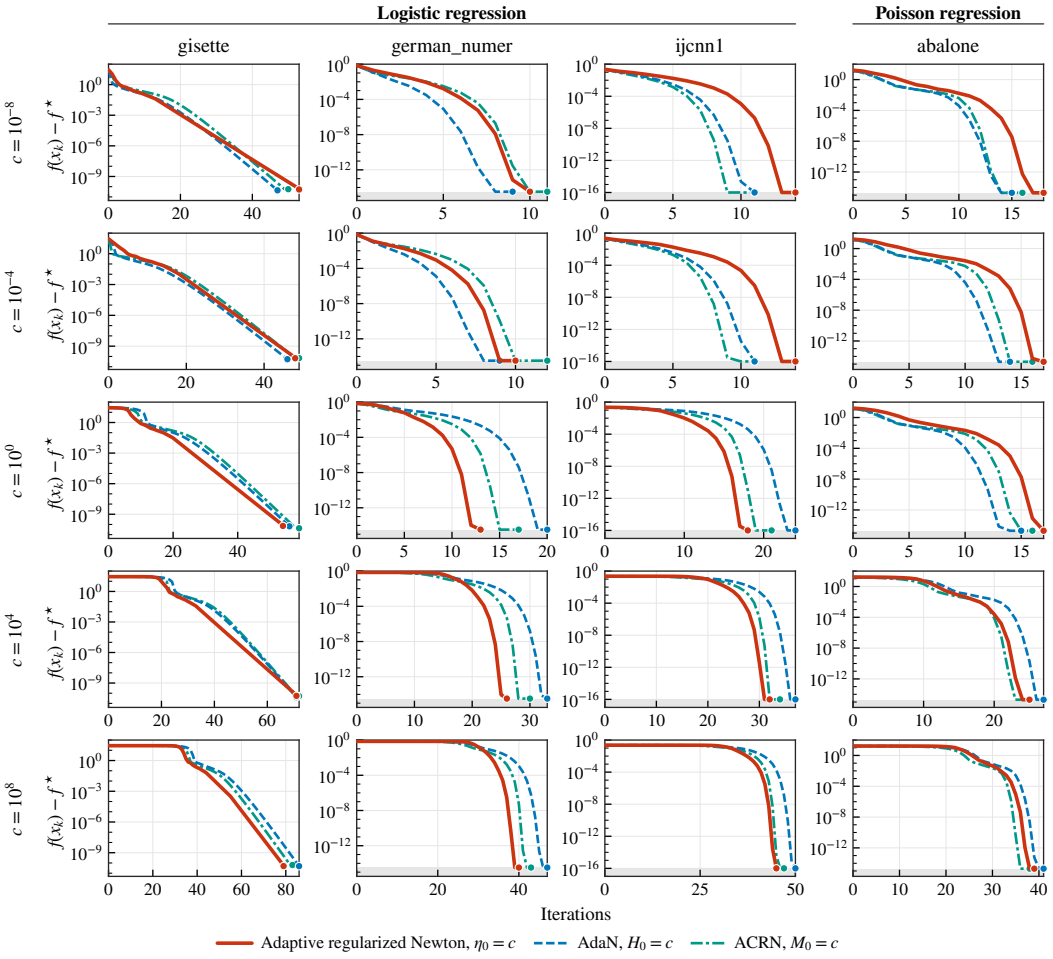

Figure 4: Convergence of Algorithm 2, AdaN and ACRN for $\eta_0 = H_0 = M_0 = c$, one row per value of c.