NEWTON METHODS BEYOND HESSIAN LIPSCHITZ CONTINUITY: A NONLINEAR PRECONDITIONING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Newton-type methods are typically analyzed under Lipschitz continuity of the Hessian, an assumption that can fail for objectives with higher-order or polynomial growth. We introduce a class of nonlinearly preconditioned Newton methods that apply Newton's root-finding scheme to a transformed optimality mapping, thereby extending recent nonlinear preconditioning ideas from first-order methods to the second-order setting. The resulting methods are naturally analyzed under Lipschitz continuity of a preconditioned Hessian, a condition that significantly relaxes the classical Hessian Lipschitz continuity assumption. Under this generalized smoothness model, we establish local superlinear and quadratic convergence guarantees, and develop a globalization strategy for the nonregularized method despite the fact that the preconditioned Newton direction need not be a descent direction. We further propose a regularized variant for isotropic preconditioners, which returns an approximate $(\varepsilon_g, \varepsilon_H)$-second-order stationary point in $\mathcal{O}(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\})$ iterations. An adaptive version removes the need to know the smoothness constant and allows inexact subproblem solutions while preserving the same complexity order.

1 INTRODUCTION

We study unconstrained optimization problems of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x), \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function with $\inf f > -\infty$. Gradient descent constitutes an efficient solver for this class of problems, but often requires carefully tuned stepsizes, and may suffer from problems like exploding gradients. In response to this, practitioners developed various adaptive gradient schemes, like AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014), as well as gradient clipping techniques (Zhang et al., 2019). It was recently shown (Oikonomidis et al., 2025b, §1.6) that all of these methods are encapsulated within a single framework of *nonlinearly preconditioned gradient methods*, which, for a stepsize $\gamma > 0$, has iterations

$$x^{k+1} = x^k - \gamma\nabla\phi^*(\nabla f(x^k)). \tag{PG}$$

Here, $\phi : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is referred to as the *reference function* and ϕ^*, its convex conjugate, is the *dual reference function*. Note that $\phi = \frac{1}{2}\|\cdot\|^2$ recovers classical gradient descent. Under a generalized smoothness assumption, (PG) can be analyzed from a majorization-minimization perspective, in the sense that at every iteration minimizes an upper bound of the form

$$f(x) \le f(x^k) + \tfrac{1}{L}\phi(L(x - y^k)) - \tfrac{1}{L}\phi(L(x^k - y^k)) \tag{1}$$

with respect to x, where $y^k = x^k - \frac{1}{L}\nabla\phi^*(\nabla f(x^k))$ and $L = 1/\gamma$. This inequality corresponds to a natural extension of Lipschitz smoothness, called *anisotropic smoothness* (Laude & Patrinos, 2025). Notably, it enables the construction of upper bounds that more closely fit the properties of the objective function f, since ϕ may be chosen to grow faster than a quadratic. Moreover, anisotropic smoothness encompasses (L_0, L_1)-smoothness (Zhang et al., 2019), and was recently shown to hold for key applications such as matrix factorization and phase retrieval (Bodard & Patrinos, 2025).

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.

This work presents a novel class of Newton methods for solving (P) that serves as a second-order counterpart to (PG), and we study its properties under a relaxed Hessian Lipschitz continuity condition. Newton's method performs iterations $x^{k+1} = x^k + d^k$, where the update directions satisfy

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k). \tag{N}$$

Instead, we consider a generalization of this method where

$$\nabla^2 \phi^*(\nabla f(x^k)) \nabla^2 f(x^k) d^k = -\nabla \phi^*(\nabla f(x^k)), \tag{PN}$$

This is nothing else than Newton's root-finding method applied to $\nabla \phi^* \circ \nabla f$ instead of ∇f, and is justified under appropriate assumptions on the reference function ϕ that guarantee the equivalence

$$\forall x^\star \in \mathbb{R}^n : \nabla f(x^\star) = 0 \quad \iff \quad \nabla \phi^*(\nabla f(x^\star)) = 0. \tag{2}$$

Note that Newton's method is *affine invariant*, so linear preconditioning does not modify the update direction. In contrast, *nonlinear preconditioning* (PN) does.

The proposed Newton variants are based on generalized smoothness conditions that more accurately reflect the properties of the objective f. By doing so, we aim to (i) narrow the gap between optimization theory and practice, and, ultimately, (ii) obtain improved practical performance.

Contributions We introduce a class of preconditioned Newton methods, described by (PN), that is naturally analyzed under a Lipschitz continuity assumption on the *preconditioned Hessian* (cf. Assumption 2.2). For appropriate references this *relaxes the usual Hessian Lipschitz continuity* condition. In particular, Assumption 2.2 holds for any 1D polynomial objective, while for higher dimensional problems a mild sufficient condition (cf. (4)) is provided that provably holds for *symmetric matrix factorization and phase retrieval* problems. For the proposed method, we then characterize the region of quadratic convergence in terms of the potentially more favorable Lipschitz constant of the preconditioned Hessian and describe a globalization strategy, which is non-trivial due to the update directions not being descent directions in general.

For isotropic references, we present a regularized variant that attains an $\mathcal{O}(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\})$ iteration complexity under Assumption 2.2. Well-definedness of the subproblems is described in detail. We also present an adaptive variant which does not require explicit knowledge of the smoothness constant and handles inexact subproblem solutions, while maintaining the same complexity.

1.1 Properties of the reference function

Throughout this work, we consider valid the following assumptions on the reference function ϕ.

Assumption 1.1. The following statements hold:

(i) The function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, μ_ϕ-strongly convex and even with $\phi(0) = 0$;

(ii) $\operatorname{int}\operatorname{dom}\phi \neq \emptyset$; $\phi \in \mathcal{C}^2(\operatorname{int}\operatorname{dom}\phi)$; and for any sequence $\{x^k\}_{k\in\mathbb{N}}$ that converges to some boundary point of $\operatorname{int}\operatorname{dom}\phi$, it follows that $\|\nabla\phi(x^k)\| \to \infty$.

Assumption 1.1 ensures in particular that the equivalence (2) holds (cf. Lemma B.1(i)), that $\phi \geq 0$ (cf. Lemma B.1(ii)) and that $\phi^* \in \mathcal{C}^2$ (Rockafellar, 1977, p. 42). Henceforth, we focus on two types of reference functions: (i) *isotropic* functions $\phi(x) = h(\|x\|)$; and (ii) *separable* functions $\phi(x) = \sum_{i=1}^n h(x_i)$ where $h : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ is a scalar kernel. Interesting choices include

$$h_1(x) = \cosh(x) - 1, \quad h_2(x) = \exp(|x|) - |x| - 1, \quad h_3(x) = -|x| - \ln(1 - |x|).$$

Notably, for $\phi(x) = \sum_{i=1}^n h_3(x_i)$ the iterates of (PG) reduce to a variant of Adam without momentum (Oikonomidis et al., 2025b, Example 1.6). We emphasize that h_1, h_2, h_3 grow faster than $\frac{1}{2}x^2$. For such reference functions anisotropic smoothness describes more general, and hence more flexible, upper bounds than Lipschitz smoothness. Note that the definition of Oikonomidis et al. (2025b) involves two constants $L, \bar{L} > 0$. Without loss of generality $\bar{L}$ can be absorbed into ϕ by instead considering $\widetilde{\phi} := \bar{L}\phi$, see also (Oikonomidis et al., 2025c).

Definition 1.2 (Anisotropic smoothness)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-anisotropically smooth relative to ϕ with constant $L > 0$ if, for all $x, \bar{x} \in \mathbb{R}^n$ and with $\bar{y} = \bar{x} - L^{-1}\nabla\phi^*(\nabla f(\bar{x}))$,

$$f(x) \leq f(\bar{x}) + L^{-1}\phi(L(x - \bar{y})) - L^{-1}\phi(L(\bar{x} - \bar{y})). \tag{3}$$

We now briefly describe the gradient and Hessian of ϕ^* in the isotropic and separable settings. For additional details, including the conjugates of h_1, h_2, h_3, we refer to (Oikonomidis et al., 2025b).

Isotropic references If $\phi = h \circ \|\cdot\|$ with h satisfying Assumption 1.1, then $\phi^* = h^* \circ \|\cdot\|$ with $\nabla\phi^*(y) = h^{*\prime}(\|y\|)\,\overline{\operatorname{sgn}}(y)$ for all $y \in \mathbb{R}^n$ (Oikonomidis et al., 2025b, Lemma 1.3), and moreover

$$\left[\nabla^2\phi^*(y)\right]^{-1} = \frac{1}{h^{*\prime\prime}(\|y\|)}\frac{yy^\top}{\|y\|^2} + \frac{\|y\|}{h^{*\prime}(\|y\|)}\left(\mathrm{I} - \frac{yy^\top}{\|y\|^2}\right)$$

for $y \in \mathbb{R}^n \setminus \{0\}$ and $\left[\nabla^2\phi^*(y)\right]^{-1} = {}^1\!/_{h^{*\prime\prime}(\|y\|)}\,\mathrm{I}$ at $y = 0$ (Oikonomidis et al., 2025b, §2.2).

Separable references In this case we have $\phi^*(y) = \sum_{i=1}^n h^*(y_i)$, with gradient $\nabla\phi^*(y) = \big(h^{*\prime}(y_1), \ldots, h^{*\prime}(y_n)\big)$. The Hessian is diagonal with elements $\Big[\left[\nabla^2\phi^*(y)\right]^{-1}\Big]_{i,i} = \frac{1}{h^{*\prime\prime}(y_i)}$.

2 Nonlinearly preconditioned Newton

Not any reference function ϕ that satisfies Assumption 1.1 is desirable in the context of the preconditioned Newton method described by (PN). We discuss two aspects.

Computational cost of the Newton system First, the system (PN) should not be more difficult to solve than the classical Newton system (N). If $\phi = h \circ \|\cdot\|$ is isotropic and $\nabla f(x) \neq 0$, then by Section 1.1 we have that $\nabla^2 f(x)d = -[\nabla^2\phi^*(\nabla f(x))]^{-1}\nabla\phi^*(\nabla f(x)) = -\alpha(x)\nabla f(x)$ where we introduced the shorthand notation $\alpha(x) := \frac{h^{*\prime}(\|\nabla f(x)\|)}{h^{*\prime\prime}(\|\nabla f(x)\|)\|\nabla f(x)\|}$. Thus, for isotropic reference functions (PN) reduces to a conventional Newton method with variable stepsizes $\alpha(x)$. Lemma B.2 provides explicit expressions for $\alpha(x)$ corresponding to h_1, h_2, h_3. For these kernel functions $\alpha(x) \geq 1$, i.e., preconditioned Newton steps are *larger* than vanilla Newton steps. On the other hand, if $\phi(x) = \sum_{i=1}^n h(x_i)$ is separable, then

$$\left[\nabla^2 f(x)d\right]_i = -\left[[\nabla^2\phi^*(\nabla f(x))]^{-1}\nabla\phi^*(\nabla f(x))\right]_i = -\tfrac{h^{*\prime}(\nabla_i f(x))}{h^{*\prime\prime}(\nabla_i f(x))}.$$

In either case, the dominant computational cost remains solving a system with the Hessian $\nabla^2 f$.

Growth of the scalar kernel function Second, the kernel function h should grow *sufficiently fast*.

Example 2.1. Let f be strongly convex with minimizer $x^\star \in \mathbb{R}^n$. Then the preconditioned Newton method (PN) with reference function $\phi(x) = f(x + x^\star) - f(x^\star)$ converges in a *single iteration*. Indeed, we have $\phi^*(y) = f^*(y) - \langle y, x^\star\rangle + f(x^\star)$ by (Rockafellar & Wets, 1998, Eq. 11(3)), and hence $\nabla\phi^*(\nabla f(x^k)) = x^k - x^\star$ and $\nabla^2\phi^*(\nabla f(x)) = \left[\nabla^2 f(x)\right]^{-1}$. It follows that $x^{k+1} = x^\star$.

Of course the 'ideal' reference function from Example 2.1 is impractical, as it requires knowledge of the minimizer $x^\star$, and restricts the objective f. However, it indicates that the growth of ϕ around 0 should ideally *mimic the growth of f around $x^\star$*. In particular, if f is not Lipschitz smooth and grows faster than a quadratic, then so should ϕ. This motivates scalar kernel functions like $h(x) = \cosh(x) - 1 = \sum_{i=1}^\infty \frac{x^{2i}}{(2i)!}$, which can match arbitrarily fast polynomial growth.

2.1 Preconditioned Hessian Lipschitz continuity

Local convergence of Newton's method is often analyzed under Lipschitz continuity of the Hessian $\nabla^2 f$. Instead, we assume Lipschitz continuity of the *preconditioned Hessian* $\mathrm{J}(\nabla\phi^* \circ \nabla f)$, which is the natural assumption for the preconditioned root-finding problem involving $\nabla\phi^* \circ \nabla f$. A similar condition was also used in (Bodard & Patrinos, 2025, Assump. 3.3).

Assumption 2.2. The mapping $H(x) := \nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)$ is L_H-Lipschitz continuous, i.e., $\exists L_H > 0 : \|H(x) - H(y)\| \leq L_H\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$.

Assumption 2.2 subsumes the usual assumption ($\phi = \frac{1}{2}\|\cdot\|^2$), and also holds for costs with stronger growth. Appendix C.1 proves the next example, generalizing (Bodard & Patrinos, 2025, Ex. 3.4).

Example 2.3 ((Bodard & Patrinos, 2025, Example 3.4))**.** Let $f(x) : \mathbb{R} \to \mathbb{R}$ be a 1D polynomial. Then there exists an $L_H > 0$ such that Assumption 2.2 holds relative to $h_1 \circ |\cdot|$ with constant L_H.

Thus, despite a higher-order growth of f, there may exist a *global* constant L_H for which Assumption 2.2 holds. By contrast, a global Hessian Lipschitz constant $L_{f,2}$ does not exist for polynomials of degree at least four. At best, there exists a local constant $L_{f,2}$ which scales with the diameter of a compact set containing the iterates. Since these constants determine both the local and global convergence of Newton-based methods, this motivates analyzing (PN) under Assumption 2.2.

The next result describes a sufficient condition (cf. (4)) for Assumption 2.2, that resembles (L_0, L_1)-smoothness. It was recently introduced by Fang et al. (2026) for analyzing Newton methods.

Theorem 2.4. *Let $f \in \mathcal{C}^3$ be (L_0, L_1)-smooth with constants $L_0, L_1 > 0$ and suppose that there exists a constant $M > 0$ such that for every $x \in \mathbb{R}^n$,*

$$\|\nabla^3 f(x)\| \leq M \left(L_0 + L_1 \|\nabla f(x)\|\right). \tag{4}$$

Then f satisfies Assumption 2.2 with constant $L_H = M + 3L_1$, relative to the isotropic reference function $\phi = h \circ \|x\|$, where $h(z) = \frac{L_0}{L_1}\left[-|z| - \frac{1}{L_1}\ln(1 - L_1|z|)\right]$.

Under (L_0, L_1)-smoothness, also *quasi-self-concordance* (Bach, 2010; Doikov, 2025) implies (4).

Corollary 2.5. *Let $f \in \mathcal{C}^3$ be (L_0, L_1)-smooth with $L_0, L_1 > 0$, and quasi-self-concordant with constant $M > 0$, i.e., for all $x, u, v \in \mathbb{R}^n$, it holds that $|\nabla^3 f(x)[u,u,v]| \leq M\langle u, \nabla^2 f(x) u\rangle \|v\|$. Then Assumption 2.2 holds relative to the reference from Theorem 2.4 with $L_H = M + 3L_1$.*

Example 2.6. Consider the logistic regression loss $f(x) = \sum_{i=1}^m \ln(1 + \exp(-b_i a_i^\top x))$, where $a_i \in \mathbb{R}^n$ denote the feature vectors and $b_i \in \{-1, 1\}$ the binary labels. Since $L_{f,1}$-Lipschitz smoothness implies $(L_{f,1}, L_1)$-smoothness for any $L_1 \geq 0$, and since f is quasi-self-concordant, Corollary 2.5 implies that f satisfies Assumption 2.2.

Condition (4) generalizes *beyond quasi-self-concordance*, as the next results demonstrate. Their proofs in Appendix C use an argument similar to (Bodard & Patrinos, 2025, Theorem 2.9) and conclude (4) from polynomial upper and lower bounds to $\|\nabla^3 f\|$ and $\|\nabla f\|$, respectively.

Example 2.7. Consider the phase retrieval loss $f(x) = \frac{1}{4}\sum_{i=1}^m \left(y_i^2 - (a_i^\top x)^2\right)^2$, and suppose that the measurement vectors $\{a_i\}_{i=1}^m$ span $\mathbb{R}^n$. Then, for any $L_1 > 0$ there exists an $L_0 > 0$ such that (i) f is (L_0, L_1)-smooth; and (ii) there exists an $M > 0$ for which (4) holds with constants L_0, L_1, M.

Example 2.8. Consider the matrix factorization loss $f(X) = \frac{1}{2}\|XX^\top - Y\|_F^2$, with $Y \in \mathbb{R}^{n\times n}$ a symmetric matrix, and $X \in \mathbb{R}^{n\times r}$. Then, for any $L_1 > 0$ there exists an $L_0 > 0$ such that (i) f is (L_0, L_1)-smooth; and (ii) there exists an $M > 0$ for which (4) holds with constants L_0, L_1, M.

2.2 Local convergence and globalization

If initialized sufficiently close to a strong local minimizer, the iterates of the preconditioned Newton method described by (PN) converge at least superlinearly. Note that for isotropic and separable reference functions the Hessian $\nabla^2\phi^*(y)$ is a multiple of identity at $y = 0$.

Theorem 2.9. *Let $x^\star \in \mathbb{R}^n$ be a strong local minimum of $f \in \mathcal{C}^2$, i.e., $\nabla f(x^\star) = 0$ and there exists a $\mu_f > 0$ such that $\nabla^2 f(x^\star) \succeq \mu_f I$. Then the following statements hold:*

(i) *there exists a $\delta > 0$ such that for all $x^0 \in \mathcal{B}(x^\star; \delta)$, the iterates $x^k{}_{k\in\mathbb{N}}$ of* (PN) *remain in $\mathcal{B}(x^\star; \delta)$ and the sequence of iterates converges Q-superlinearly to $x^\star$;*

(ii) *if, additionally, Assumption 2.2 holds with constant $L_H > 0$, if $\nabla^2\phi^*(0) = \mu_\phi^{-1}\,\mathrm{I}$ for some $\mu_\phi > 0$, and if $\|x^0 - x^\star\| < \Delta := \frac{2\mu_f}{3\mu_\phi L_H}$, then the iterates converge Q-quadratically, i.e.,*

$$\forall k \geq 0 : \|x^{k+1} - x^\star\| \leq \frac{\mu_\phi L_H}{2(\mu_f - \mu_\phi L_H \|x^k - x^\star\|)}\|x^k - x^\star\|^2.$$

Theorem 2.9 generalizes (Nesterov, 2018, §1.2.5) from the standard ($\phi = \frac{1}{2}\|\cdot\|^2$) to the preconditioned setting, and replaces the conventional Lipschitz constant $L_{f,2}$ by the product $\mu_\phi L_H$.

Similar to the classical Newton method (N), the preconditioned variant (PN) may diverge if initialized too far from a stationary point. A standard globalization strategy consists of performing an Armijo linesearch under a Lipschitz-smoothness assumption. In our framework, the corresponding

notion would be anisotropic smoothness (cf. Definition 1.2). However, Armijo linesearch does not globalize (PN), as the preconditioned directions *need not be descent directions*, i.e, $\langle \nabla f(x^k), d^k \rangle \leq 0$ may not hold. Instead, Algorithm 1 defines the candidate $x(\tau) := x^k + \tau d^k + (1-\tau)p^k$ as a convex combination of a Newton direction d^k and a negative gradient direction p^k. As $\tau \to 0$ we have $x(\tau) \to x^k + p^k$, yielding sufficient decrease under anisotropic smoothness. A similar idea has been used for globalizing proximal gradient methods (Stella et al., 2017).

Algorithm 1 Globalized preconditioned Newton method

Require: $x^0 \in \mathbb{R}^n$, $\alpha \in (0,1), \sigma \in (0,1)$ and define the stepsize $\gamma = \alpha/L$

1: **for** $k = 0, 1, \ldots$ **do**
2: **If** $\nabla f(x^k) = 0$ **then** Return x^k
3: Let $g^k = \nabla\phi^*(\nabla f(x^k))$; $p^k = -\gamma g^k$; and $H_k = \nabla^2\phi^*(\nabla f(x^k))\nabla^2 f(x^k)$
4: Select $d^k \in \mathbb{R}^n$ such that $H_k d^k = -g^k$ if possible, otherwise set $d^k = -\gamma g^k$
5: $x^{k+1} = x^k + \tau_k d^k + (1-\tau_k)p^k$ where $\tau_k \in \{1/2^i \mid i \in \mathbb{N}\}$ is the largest stepsize for which

$$f(x^{k+1}) - f(x^k) \leq -\gamma\sigma\phi(g^k).$$

If H^k does not have full rank, then there need not exist a d^k satisfying (PN). For simplicity, Algorithm 1 selects a negative preconditioned gradient direction $d^k = -\gamma g^k$ in such cases.

Theorem 2.10 (Global convergence)**.** *Consider the iterates generated by Algorithm 1 and suppose that f is L-anisotropically smooth relative to ϕ, for some $L > 0$. Then, the linesearch procedure is well-defined, i.e., there exists $\bar{\tau}_k > 0$ such that for all $\tau \in [0, \bar{\tau}_k]$*

$$f(x^k + \tau d^k + (1-\tau)p^k) - f(x^k) \leq -\gamma\sigma\phi(\nabla\phi^*(\nabla f(x^k))).$$

Moreover, we have that $\min_{0\leq k\leq K} \phi(\nabla\phi^*(\nabla f(x^k))) \leq \frac{L\left(f(x^0) - \inf f\right)}{\alpha\sigma(K+1)}$.

Although $\phi(\nabla\phi^*(\nabla f(x)))$ is a natural stationarity measure under anisotropic smoothness, it can often be translated to the more standard measure $\|\nabla f(x)\|$. For example, if $\phi(x) = \cosh(\|x\|) - 1$, then $\phi(\nabla\phi^*(\nabla f(x))) = \sqrt{1 + \|\nabla f(x)\|^2} - 1$. Tracing the proof of (Oikonomidis et al., 2025c, Corollary 3.3), this reference function yields an $\mathcal{O}(1/\sqrt{K+1})$ convergence rate in terms of the standard gradient norm. On the other hand, the fast local convergence rate of (PN) is preserved.

Theorem 2.11 (Fast local convergence)**.** *Consider the iterates generated by Algorithm 1 and suppose that $x^k \to x^\star$, with $x^\star$ a strong local minimum of f. Then, eventually, $\tau_k = 1$ is always accepted and the iterates satisfy* (PN)*. In particular, all claims of Theorem 2.9 hold.*

Remark 2.12 (Adaptive estimation of L)**.** Algorithm 1 implicitly assumes knowledge of the constant L of anisotropic smoothness. Yet, this constant can also be estimated *adaptively*, since the convergence proof only requires that $f(x^k - \gamma\nabla\phi^*(\nabla f(x^k))) - f(x^k) \leq -\gamma\phi(\nabla\phi^*(\nabla f(x^k)))$ holds at any iteration. Whenever violated, we can set $L \leftarrow 2L$ and $\gamma \leftarrow \alpha/L$, and restart that iteration. It is easily verified that this can only happen a finite number of times.

3 Regularization

Practical Newton variants typically include *regularization* to deal with singular Hessians. By adding an appropriate nonnegative multiple $\lambda_k \geq 0$ of the identity matrix, the regularized Hessian $\nabla^2 f(x^k) + \lambda_k \mathrm{I}$ becomes positive definite. Regularized Newton variants thus use directions s^k satisfying $\left(\nabla^2 f(x^k) + \lambda_k \mathrm{I}\right) s^k = -\nabla f(x^k)$. *Cubic regularization* methods select $\lambda_k = \sigma\|s^k\|$ for some $\sigma > 0$. This direction s^k can be interpreted as the minimizer of a cubic model of f, hence the name of this type of regularization. Algorithm 2 generalizes this variant to our preconditioning framework and uses directions s^k satisfying

$$\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k = -M_k g^k \tag{5}$$

where $\lambda_k = \sigma\|s^k\|$ is such that $\nabla^2 f(x^k) + \lambda_k M_k \succeq 0$, with $g^k := \nabla\phi^*(\nabla f(x^k))$, and $M_k := [\nabla^2\phi^*(\nabla f(x^k))]^{-1} = \nabla^2\phi(g^k)$. By left-multiplying with M_k^{-1} (5) is equivalent to

$$H(x^k)s^k + \lambda_k s^k = -g(x^k). \tag{6}$$

We stress that, even for isotropic reference functions, regularizing by λM_k with $\lambda \geq 0$ is *not* a rescaling of standard regularized Newton methods, since M_k is not a multiple of the identity matrix. Note also that Algorithm 2 requires $\sigma \geq L_H$ and consequently requires knowledge of the constant L_H from Assumption 2.2. A variant that adaptively estimates σ is presented in Section 3.3.

Algorithm 2 Preconditioned Newton method with regularization

Require: $x^0 \in \mathbb{R}^n$, $\sigma \geq L_H$

1: **for** $k = 0, 1, \ldots$ **do**

2: Let $g^k = \nabla\phi^*(\nabla f(x^k))$; $M_k = \nabla^2\phi(g^k)$ and compute $(s^k, \lambda_k) \in \mathbb{R}^n \times \mathbb{R}$ satisfying

$$M_k^{-1}\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k = -g^k, \quad \lambda_k = \sigma\|s^k\|, \quad \nabla^2 f(x^k) + \lambda_k M_k \succeq 0 \tag{7}$$

3: Set $x^{k+1} = x^k + s^k$.

Remark 3.1 (No cubic model)**.** Although Algorithm 2 generalizes Newton's method with cubic regularization, it cannot, in general, be interpreted as minimizing a cubic model at each iteration. Indeed, define $F_k(s^k) := (\nabla^2 f(x^k) + \sigma\|s^k\|M_k)s^k$. If this is to be the gradient of a (cubic) model, then its Jacobian should be symmetric. However, $J_{F_k}(s^k) = \nabla^2 f(x^k) + \sigma\|s^k\|M_k + \sigma\frac{M_k s^k s^{k\top}}{\|s^k\|}$ is asymmetric in general. A cubic model would exist if $\lambda_k = \sigma\|s^k\|_{M_k}$, but then the subsequent convergence proof breaks down as Lemmas 3.6 and 3.7 would be expressed in different norms.

3.1 Well-definedness of the subproblem

The following result confirms the well-definedness of the subproblem (7), in the sense that a suitable pair (s^k, λ_k) can always be found. The proof is constructive, and follows a variation on standard arguments for cubic regularization methods (Cartis et al., 2022, §8).

Theorem 3.2. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, $M \in \mathbb{R}^{n\times n}$ symmetric positive definite, $g \in \mathbb{R}^n$, and $\sigma > 0$. Denote by $\xi_1 \leq \cdots \leq \xi_n$ the generalized eigenvalues of the pencil (A, M), and define $s(\lambda) := -(A + \lambda M)^\dagger Mg$. Then, there exists a pair $(s^\star, \lambda_\star) \in \mathbb{R}^n \times \mathbb{R}$ satisfying*

$$(A + \lambda_\star M)s^\star = -Mg, \qquad \lambda_\star = \sigma\|s^\star\|, \qquad A + \lambda_\star M \succeq 0.$$

In particular, exactly one of the following statements holds:

(i) *$\lambda_\star > -\xi_1$ and $s^\star = s(\lambda_\star)$;*

(ii) *$\lambda_\star = -\xi_1$, and $s^\star = s(\lambda_\star) + \alpha_\star v_1$, with $\alpha_\star \in \mathbb{R}$ such that $\|s^\star\| = \lambda_\star/\sigma$, and v_1 an eigenvector corresponding to ξ_1.*

While the proof of Theorem 3.2 is constructive, Lanczos–Krylov techniques are often more practical and scale better for solving subproblems of regularized Newton methods. Since $M_k \succ 0$, we can define $B_k := M_k^{-1/2}\nabla^2 f(x^k)M_k^{-1/2}$, $q^k := M_k^{1/2}g^k$ and $d^k = M_k^{1/2}s^k$. Then (7) is equivalent to

$$(B_k + \lambda_k \mathrm{I})\, d^k = -q^k, \quad \lambda_k = \sigma\|M_k^{-1/2}d^k\|, \quad B_k + \lambda_k \mathrm{I} \succeq 0. \tag{8}$$

Thus, these Lanczos–Krylov techniques remain applicable to the transformed subproblem. The main modification is that the scalar equation for λ_k uses a weighted norm, which is easily incorporated. Note that the additional cost stemming from the factors $M_k^{1/2}$ and $M_k^{-1/2}$ remains limited for isotropic reference functions: since M_k is the sum of a symmetric rank-1 term and a multiple of identity, matrix-vector products with these additional factors can be performed using only inner products and vector operations, i.e., without forming dense matrices.

3.2 Convergence analysis for isotropic reference functions

We now analyze the convergence of Algorithm 2 for *isotropic* reference functions.

Assumption 3.3. The reference $\phi = h \circ \|\cdot\|$ is isotropic, with h satisfying Assumption 1.1.

Algorithm 3 Preconditioned Newton method with adaptive regularization

Require: $x^0 \in \mathbb{R}^n$, $\sigma_0 > 0$. Provide parameters $\sigma_{\min} \in (0, \sigma_0], \theta \geq 0$ and $\eta_1, \eta_2, \gamma_1, \gamma_2, \gamma_3$ satisfying $0 < \eta_1 \leq \eta_2 < 1, 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3$.

1: **for** $k = 0, 1, \ldots$ **do**

2: Let $g^k = \nabla\phi^*(\nabla f(x^k))$; $M_k = \nabla^2\phi(g^k)$ and compute $(s^k, \lambda_k, z^k) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ satisfying

$$M_k^{-1}\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k = -g^k + z^k, \quad \lambda_k = \sigma_k \|s^k\|, \quad \nabla^2 f(x^k) + \lambda_k M_k \succeq 0,$$
$$\|z^k\| \leq \frac{1}{2}\theta\|s^k\|^2, \quad \langle M_k z^k, s^k\rangle \leq 0.$$

3: **If** $s^k = 0$ **then** return x^k **else** compute $f(x^k + s^k)$ and define $\rho_k = \frac{f(x^k) - f(x^k + s^k)}{(\mu_\phi/4)\sigma_k\|s^k\|^3}$.

4: **If** $\rho_k \geq \eta_1$ **then** $x^{k+1} = x^k + s^k$ **else** $x^{k+1} = x^k$

5: Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1\sigma_k), \sigma_k] & \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \rho_k < \eta_1. \end{cases} \tag{9}$$

Under Assumption 3.3, the next two lemmas relate the preconditioned gradient $\nabla\phi^* \circ \nabla f$ with the gradient ∇f, and establish that the directions of Algorithm 2 are *descent directions*.

Lemma 3.4. *Suppose that Assumption 3.3 holds. Define* $\xi(y) = \frac{y}{h^{*\prime}(y)}$ *for* $y > 0$*, and* $\xi(0) = h''(0)$*. Then the composite function* $\nu(x) := \xi(\|\nabla f(x)\|)$ *is continuous on* $\mathbb{R}^n$*, and for all* $x \in \mathbb{R}^n$ *satisfies* $\nu(x) \geq \mu_\phi > 0$ *and* $\nabla f(x) = \nu(x)\nabla\phi^*(\nabla f(x))$*.*

Lemma 3.5. *Let* x^k, s^k *be generated by Algorithm 2. If Assumption 3.3 holds, then* s^k *is a descent direction, i.e.,* $\langle \nabla f(x^k), s^k\rangle \leq 0$*.*

Traditional analyses based on $L_{f,2}$-Lipschitz continuous Hessians exploit upper bounds of the form

$$\forall x, s \in \mathbb{R}^n : \left|f(x+s) - f(x) - \nabla f(x)^T s - \tfrac{1}{2}s^T\nabla^2 f(x)s\right| \leq \tfrac{L_{f,2}}{6}\|s\|^3.$$

to establish *sufficient decrease*. However, a similar bound does not hold under Assumption 2.2 for generic references ϕ. Nevertheless, the following result leverages Lemmas 3.4 and 3.5 to establish a similar bound *for the specific directions* s^k *generated by Algorithm 2*.

Lemma 3.6. *Consider* $x^k, s^k \in \mathbb{R}^n$ *generated by Algorithm 2. If Assumptions 2.2 and 3.3 hold, then* $f(x^k + s^k) - f(x^k) \leq -\frac{\mu_\phi L_H}{3}\|s^k\|^3$*.*

Our analysis also requires a lower bound on $\|s^k\|$ in terms of the (preconditioned) gradient norm.

Lemma 3.7. *Under Assumption 2.2, the iterates of Algorithm 2 satisfy* $\|s^k\|^2 \geq \frac{2}{2\sigma + L_H}\|g(x^{k+1})\|$*.*

The combination of Lemmas 3.6 and 3.7 through a telescoping argument yields an iteration complexity for obtaining a *second-order stationary point*. The result below is expressed in terms of the preconditioned gradient $\nabla\phi^* \circ \nabla f$, and the preconditioned Hessian $\lambda_{\min}(H(x)) = \lambda_{\min}(M^{-1}\nabla^2 f(x))$. Note that the asymmetric matrix $H(x^k) = M_k^{-1}\nabla^2 f(x^k)$ is similar to the symmetric matrix $M_k^{-1/2}\nabla^2 f(x^k)M_k^{-1/2}$, and therefore has *real* eigenvalues. Moreover, by positive definiteness of M_k, we have the equivalence $\lambda_{\min}(H(x^k)) \geq 0 \Leftrightarrow \lambda_{\min}(\nabla^2 f(x^k)) \geq 0$, which motivates using $\lambda_{\min}(H(x^k)) \geq 0$ as a second-order stationarity measure. In fact, by specializing to a specific reference function ϕ, these unconventional optimality measures, i.e., the preconditioned gradient norm $\|\nabla\phi^*(\nabla f(x))\|$ and the smallest eigenvalue of the preconditioned Hessian $\lambda_{\min}(H(x))$, can be related to the standard $\|\nabla f(x)\|$ and $\lambda_{\min}(\nabla^2 f(x))$, as shown in the subsequent corollary.

Theorem 3.8. *Consider* $x^k, s^k \in \mathbb{R}^n$ *generated by Algorithm 2. Suppose that Assumptions 2.2 and 3.3 hold and that* $\sigma \geq L_H$*. Then, for any* $\varepsilon_g, \varepsilon_H > 0$*, there exists a* $k^* \leq K$ *such that Algorithm 2 obtains a point* x^{k^*} *for which* $\|\nabla\phi^*(\nabla f(x^{k^*}))\| \leq \varepsilon_g$ *and* $[-\lambda_{\min}(H(x^{k^*}))]_+ \leq \varepsilon_H$*, where*

$$K = 1 + \left\lceil 6\tfrac{(f(x^0) - \inf f)}{\mu_\phi L_H} \max\left\{\left(\tfrac{3\sigma}{2\varepsilon_g}\right)^{3/2}, \left(\tfrac{\sigma}{\varepsilon_H}\right)^3\right\}\right\rceil.$$

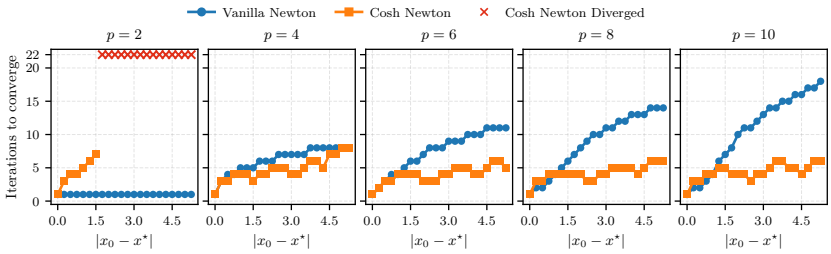

Figure 1: Iterations to minimize $f(x) = x^p/p + x^2/2$.

Corollary 3.9. *Consider $x^k, s^k \in \mathbb{R}^n$ generated by Algorithm 2. Suppose that Assumptions 2.2 and 3.3 hold and that $\sigma \geq L_H$ and $\phi(x) = \mu_\phi h_3(\|x\|)$. Then, for any $\varepsilon_g \in (0, \mu_\phi), \varepsilon_H > 0$, there exists a $k^* \leq K$ such that Algorithm 2 obtains a point x^{k^*} for which $\|\nabla f(x^{k^*})\| \leq \varepsilon_g$ and $[-\lambda_{\min}(\nabla^2 f(x^{k^*}))]_+ \leq \varepsilon_H$, where*

$$K = 1 + \left\lceil 6 \tfrac{(f(x^0) - \inf f)}{\mu_\phi L_H} \max\left\{ \left(\tfrac{3\mu_\phi \sigma}{\varepsilon_g}\right)^{3/2}, \left(\tfrac{4\mu_\phi \sigma}{\varepsilon_H}\right)^3 \right\} \right\rceil .$$

If $\sigma = L_H$, and with $\varepsilon_g = \varepsilon, \varepsilon_H = \sqrt{\mu_\phi L_H \varepsilon}, \varepsilon > 0$, this recovers an $\mathcal{O}(\sqrt{\mu_\phi L_H}\varepsilon^{-3/2})$ complexity in terms of the standard first and second-order optimality measures, similar to Zhou et al. (2025).

3.3 Adaptive Regularization

Finally, Algorithm 3 describes an *adaptive* variant of Algorithm 2. An obvious advantage is that the Lipschitz constant L_H is not explicitly required. Moreover, it supports *inexact* solutions to the subproblems, represented by the vector $z^k \in \mathbb{R}^n$. Clearly, if $z^k = 0$ (e.g. when setting $\theta = 0$), then the obtained subproblem solution is exact and corresponds to that of Algorithm 2. Since $z^k = 0$ can always be chosen, well-definedness immediately follows.

To obtain an analogue to Lemma 3.5, Algorithm 3 requires $\langle M_k z^k, s^k \rangle \leq 0$. This additional condition is compatible with a Lanczos-based Krylov subproblem solver that satisfies the Galerkin condition. Indeed, for a given subspace $\mathcal{K}_k \subseteq \mathbb{R}^n$ and shift $\lambda_k \geq 0$, $s^k \in \mathcal{K}_k$ is computed such that $\langle (\nabla^2 f(x^k) + \lambda_k M_k) s^k + M_k g^k, v \rangle = 0$ for every $v \in \mathcal{K}_k$. Defining $z^k := M_k^{-1}((\nabla^2 f(x^k) + \lambda_k M_k) s^k + M_k g^k)$ we find $M_k z^k \perp \mathcal{K}_k$. Since $s^k \in \mathcal{K}_k$, this actually enforces the stricter condition $\langle M_k z^k, s^k \rangle = 0$.

Theorem 3.10. *Consider $x^k, s^k \in \mathbb{R}^n$ generated by Algorithm 3. Suppose that Assumptions 2.2 and 3.3 hold. Then, for any $\varepsilon_g, \varepsilon_H > 0$, there exists a $k^* \leq K$ such that Algorithm 3 obtains a point x^{k^*} for which $\|\nabla \phi^*(\nabla f(x^{k^*}))\| \leq \varepsilon_g$ and $[-\lambda_{\min}(H(x^{k^*}))]_+ \leq \varepsilon_H$, where $K \leq C_1 \max\left\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\right\} + C_2$ for some constants $C_1, C_2 > 0$ independent of $\varepsilon_g, \varepsilon_H$.*

4 Numerical Results

Higher-order 1D polynomials This experiment illustrates the importance of accurate smoothness conditions. Consider a scalar polynomial $f(x) = \frac{1}{p}x^p + \frac{1}{2}x^2$ where $p \geq 2$ is even. The unique minimizer is clearly $x^\star = 0$. Figure 1 compares vanilla Newton (N) against preconditioned Newton (PN) with $\phi(x) = \cosh(|x|) - 1$, for various initializations x^0. No globalization is applied. Vanilla Newton performs best for $p = 2$, as it converges in a single step. If $p > 2$, then (PN) performs remarkably better. In fact, (N) quickly deteriorates as p or x^0 is increased, whereas (PN) is consistent.

Physics-Informed Neural Network Training Let us consider the reaction-diffusion problem

$$-u''(x) + \beta u(x)^3 = \pi^2 \sin(\pi x) + \beta \sin^3(\pi x), \quad x \in [-1, 1], \quad u(-1) = u(1) = 0, \tag{10}$$

with exact solution $u^\star(x) = \sin(\pi x)$. We focus on the highly nonlinear case where $\beta = 10^4$, and train a $1 - 12 - 12 - 1$ physics-informed neural network (PINN) $N_\theta(x)$ with $\tanh$ activation

Table 1: PINN results over 40 held-out initializations. Success means relative error at most 10^{-4}.

Method	Median error	90th perc. error	Success	Time (s)	Func. evals.
LM ($\mu = 1$)	1.14×10^{-5}	3.55×10^{-1}	30/40	0.066	134
NP-LM ($\mu = 0.1$)	4.62×10^{-6}	$\mathbf{4.58 \times 10^{-5}}$	**37/40**	0.069	168
LM ($\mu = 0.1$)	$\mathbf{3.25 \times 10^{-6}}$	3.05×10^{-1}	33/40	0.068	141.5

Table 2: Transfer to the milder $\beta = 10^3$ problem without retuning. Entries summarize 20 paired initializations. Success means relative error at most 10^{-4}.

Method	Median error	90th perc. error	Success	Time (s)	Func. evals.
LM ($\mu = 1$)	5.45×10^{-5}	1.65×10^{-4}	14/20	0.066	103
NP-LM ($\mu = 0.1$)	$\mathbf{1.10 \times 10^{-5}}$	$\mathbf{1.10 \times 10^{-4}}$	**18/20**	0.068	134.5

function and parameters θ. Boundary conditions are imposed by setting $u_\theta(x) = (1 - x^2)N_\theta(x)$, and we evaluate (10) at $m = 64$ equally spaced interior points. The training objective is the residual loss $f(\theta) = \frac{\|r(\theta)\|^2}{2m}$ where the residual vector $r(\theta) \in \mathbb{R}^m$ measures the violation of (10) by the prediction $u_\theta(x)$ at the m grid points. As a baseline, we consider the Levenberg-Marquardt (LM) method (Rathore et al., 2024), which computes directions d satisfying

$$(J(\theta)^\top J(\theta) + \mu I)d = -\nabla f(\theta), \tag{LM}$$

where $J(\theta)$ is the Jacobian of $r(\theta)/\sqrt{m}$ and $\mu \geq 0$ a tuned constant. Inspired by (5), we compare against a nonlinearly preconditioned variant that regularizes by μM instead of $\mu \mathrm{I}$, and in particular

$$(J(\theta)^\top J(\theta) + \mu M(\theta))d = -M(\theta)\nabla\phi^*(\nabla f(\theta)), \tag{NP-LM}$$

where $M(\theta) = \nabla^2\phi(\nabla\phi^*(\nabla f(\theta)))$. We use $\phi(\theta) = c^2[\exp(\|\theta\|/c) - \|\theta\|/c - 1]$ with $c > 0$.

Both methods use the exact Jacobian $J(\theta)$, solve the resulting system with a Cholesky factorization, globalize with Armijo linesearch, and have the same initializations and budget of 40 iterations. Hyperparameters are tuned on 8 initializations using the final training loss: (LM) tests $\mu \in \{0.1, 0.3, 1, 3, 10\}$; (NP-LM) first selects $c = c_0\|\nabla f(\theta^0)\|$ with $c_0 \in \{0.03, 0.1, 0.3, 1, 3\}$ and then selects μ over the same grid. As such, (LM) selects $\mu = 1$, and (NP-LM) selects $\mu = 0.1$ and $c_0 = 1$. We keep these hyperparameters fixed for 40 new initializations, and report the relative error $\|u_\theta(x) - u^\star(x)\|/\|u^\star(x)\|$ over the grid vector x in Table 1.

(NP-LM) reduces the 90th percentile error, while maintaining a comparable run time. It also increases the number of successful initializations, i.e., such that the error is at most 10^{-4}. To distinguish the nonlinear preconditioning from the different tuning, we also report LM with $\mu = 0.1$. While this improves median error of (LM) slightly, the 90th percentile becomes worse. Finally, Table 2 considers the case $\beta = 10^3$, and assesses performance without retuning the hyperparameters for $\beta = 10^4$. The median error for (NP-LM) is roughly 5 times lower, again at a similar run time.

An additional experiment involving large-scale matrix factorization is reported in Appendix E.

5 Conclusion

This work introduced a novel class of Newton methods, derived by applying a nonlinear preconditioner to the first-order optimality conditions. Our approach is closely related to a generalized Hessian Lipschitz continuity condition, which may better describe an objective in the presence of higher-order polynomial growth. In particular, we provided a mild sufficient condition that holds for symmetric matrix factorization and phase retrieval problems, and related it to quasi-self-concordance. The local and global convergence of our scheme were analyzed under such relaxed smoothness conditions, and practical choices for the preconditioners were discussed. A novel regularization approach based on $\nabla^2\phi$ was then presented, and for isotropic reference functions, we established an $\mathcal{O}(\max\{\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\})$ iteration complexity for a cubic regularization variant. Finally, we validated the proposed ideas on large-scale matrix factorization and PINN training problems.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, creating or editing software code, implementing methods. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, generating synthetic data sets, supporting qualitative and thematic data analysis, interpreting results. designing or providing feedback on research methodology or experiments. Assisting with translation, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for suggesting experimental parameters, editing the research paper to improve readability and checking mathematical proofs. We have reviewed all AI-assisted work. Code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Complete proofs for all theoretical results are provided in the appendix. The manuscript describes all experiments in detail; a software implementation is included in the supplementary material to ensure full reproducibility.

REFERENCES

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384 – 414, 2010.

Alexander Bodard and Panagiotis Patrinos. Escaping saddle points without Lipschitz smoothness: the power of nonlinear preconditioning. In *Advances in Neural Information Processing Systems*, pp. 124173–124208, 2025.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *Advances in Neural Information Processing Systems*, pp. 25311–25374, 2024.

Regina S Burachik, C Yalçın Kaya, and Shoham Sabach. A generalized univariate Newton method motivated by proximal regularization. *Journal of Optimization Theory and Applications*, 155(3): 923–940, 2012.

Regina S Burachik, Bethany I Caldwell, and C Yalçın Kaya. A generalized multivariable Newton method. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):15, 2021.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011a.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011b.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. SIAM, 2022.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pp. 5396–5427, 2023.

Nikita Doikov. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *Mathematical Programming*, pp. 1–39, 2025.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. *Mathematical programming*, 204(1):1–25, 2024.

Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

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.

Jean-Pierre Dussault, Tangi Migot, and Dominique Orban. Scalable adaptive cubic regularization methods. *Mathematical Programming*, 207(1):191–225, 2024.

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

Serge Gratton, Sadok Jerad, and Philippe L Toint. Yet another fast variant of Newton's method for nonconvex optimization. *IMA Journal of Numerical Analysis*, 45(2):971–1008, 2025.

Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takác. A damped Newton method achieves global $\mathcal{O}(1/k^2)$ and local quadratic convergence rate. *Advances in Neural Information Processing Systems*, pp. 25320–25334, 2022.

Chuan He, Zhaosong Lu, and Ting Kei Pong. A Newton-CG based augmented lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. *SIAM Journal on Optimization*, 33(3):1734–1766, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Emanuel Laude and Panagiotis Patrinos. Anisotropic proximal gradient. *Mathematical Programming*, 214(1):801–845, 2025.

Emanuel Laude, Andreas Themelis, and Panagiotis Patrinos. Dualities for non-Euclidean smoothness and strong convexity under the light of generalized conjugacy. *SIAM Journal on Optimization*, 33(4):2721–2749, 2023.

Flavien Léger and Pierre-Cyril Aubin-Frankowski. Gradient descent with a general cost. *arXiv preprint arXiv:2305.04917*, 2023.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

Dong-Hui Li, Masao Fukushima, Liqun Qi, and Nobuo Yamashita. Regularized Newton methods for convex minimization problems with singular solutions. *Computational optimization and applications*, 28(2):131–147, 2004.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and Non-convex Optimization Under Generalized Smoothness. *Advances in Neural Information Processing Systems*, pp. 40238–40271, 2023.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Chris J Maddison, Daniel Paulin, Yee Whye Teh, and Arnaud Doucet. Dual space preconditioning for gradient descent. *SIAM Journal on Optimization*, 31(1):991–1016, 2021.

Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Konstantin Mishchenko. Regularized Newton method with global convergence. *SIAM Journal on Optimization*, 33(3):1440–1462, 2023.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.

Konstantinos Oikonomidis, Emanuel Laude, and Panagiotis Patrinos. Forward-backward splitting under the light of generalized convexity. *arXiv preprint arXiv:2503.18098*, 2025a.

Konstantinos Oikonomidis, Jan Quan, Emanuel Laude, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods under generalized smoothness. In *International Conference on Machine Learning*, pp. 47132–47154, 2025b.

Konstantinos Oikonomidis, Jan Quan, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods: Momentum and stochastic analysis. In *Advances in Neural Information Processing Systems*, pp. 38957–38988, 2025c.

Roman A Polyak. Regularized Newton method for unconstrained convex optimization. *Mathematical programming*, 120(1):125–145, 2009.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *International Conference on Machine Learning*, pp. 42159–42191, 2024.

R Tyrrell Rockafellar. Higher derivatives of conjugate convex functions. *Journal of Applied Analysis*, 1(1):41–43, 1977.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Clément W Royer, Michael O'Neill, and Stephen J Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180(1): 451–488, 2020.

Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *Conference on Decision and Control*, pp. 1939–1944, 2017.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Yuhao Zhou, Jintao Xu, Bingrui Li, Chenglong Bao, Chao Ding, and Jun Zhu. A regularized Newton method for nonconvex optimization with global and local complexity guarantees. In *Advances in Neural Information Processing Systems*, pp. 73447–73502, 2025.

Hong Zhu and Yunhai Xiao. A hybrid inexact regularized Newton and negative curvature method. *Computational Optimization and Applications*, 88(3):849–870, 2024.

A Notation and Related Work

A.1 Notation

By $\langle \cdot, \cdot \rangle$ we denote the Euclidean inner product, and by $\| \cdot \|$ the induced Euclidean norm for vectors as well as the corresponding induced operator norm on (multi)linear maps. For a square matrix A with real spectrum, the largest and smallest eigenvalues are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. Likewise, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ refer to the largest and smallest singular values. We use the notation $\overline{\operatorname{sgn}}(x) = {}^{x}/_{\|x\|}$ for $x \in \mathbb{R}^n \setminus \{0\}$ and 0 otherwise. The class of k times continuously differentiable functions is denoted by $\mathcal{C}^k$. A function $f \in \mathcal{C}^1$ is $L_{f,1}$-Lipschitz smooth or has $L_{f,1}$-Lipschitz continuous gradients if $\|\nabla f(x) - \nabla f(y)\| \leq L_{f,1}\|x - y\|$ for all $x, y \in \mathbb{R}^n$ with $L_{f,1} \geq 0$. A function $f \in \mathcal{C}^2$ has $L_{f,2}$-Lipschitz continuous Hessians if $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_{f,2}\|x - y\|$ for all $x, y \in \mathbb{R}^n$ with $L_{f,2} \geq 0$, and is (L_0, L_1)-smooth if $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$ for all $x \in \mathbb{R}^n$ with $L_0, L_1 \geq 0$. Otherwise, we adopt the notation from Rockafellar & Wets (1998).

A.2 Related work

Nonlinear Preconditioning The nonlinearly preconditioned gradient method (PG) was first analyzed in the convex setting by Maddison et al. (2021), and later in the nonconvex setting Laude & Patrinos (2025); Oikonomidis et al. (2025b) under anisotropic smoothness, a notion introduced in Laude et al. (2023) and closely related to the concept of Φ-convexity Léger & Aubin-Frankowski (2023); Oikonomidis et al. (2025a). The method has also been extended to measure spaces Bonet et al. (2024). Recent works presented variants with momentum and a stochastic analysis Oikonomidis et al. (2025c), and established saddle point avoidance of the scheme Bodard & Patrinos (2025).

Generalized smoothness Gradient-based methods are traditionally analyzed under restrictive Lipschitz smoothness assumptions, see e.g. Nesterov (2018). Bregman relative smoothness is a milder condition that allows for unbounded Hessians Lu et al. (2018). More recently, the (L_0, L_1)-smoothness condition, proposed by Zhang et al. (2019) based on empirical observations in training LSTMs, has gained traction. Further generalizations include α-symmetric smoothness Chen et al. (2023) and ℓ-smoothness Li et al. (2023). Note that anisotropic smoothness subsumes (L_0, L_1)-smoothness and ℓ-smoothness for certain reference functions Bodard & Patrinos (2025). For Newton methods, Lipschitz continuity of the Hessian or self-concordance are often assumed Nesterov & Nemirovskii (1994); Hanzely et al. (2022); Doikov (2025).

Newton's method Following the terminology of Maddison et al. (2021), the proposed method (PN) is derived by 'left'-preconditioning ∇f, cf. (2). A distinct class of 'right'-preconditioned Newton methods has been studied in Burachik et al. (2012; 2021). Various globalization and regularization techniques have been explored for Newton methods. Cubic regularization methods Nesterov & Polyak (2006) attain an $\mathcal{O}\left(\varepsilon^{-3/2}\right)$ complexity under Lipschitz continuity of the Hessian. This is optimal for this class of problems Carmon et al. (2020). Following the works of Cartis et al. (2011a;b), *adaptive* variants that attain the optimal rate without knowledge of the actual Lipschitz constant are a popular subject of study Dussault et al. (2024), and *universal* variants have also received attention Doikov et al. (2024). Quadratic regularization or Levenberg-Marquardt regularization Levenberg (1944); Marquardt (1963) has also been studied for its conceptual simplicity and computational efficiency Li et al. (2004); Polyak (2009). Many recent works analyze its iteration complexity, both in the convex Mishchenko (2023); Doikov & Nesterov (2024) and nonconvex setting Gratton et al. (2025); Royer et al. (2020); He et al. (2023); Zhu & Xiao (2024); Zhou et al. (2025).

B Additional Results

Lemma B.1 (Stationarity)**.** *Suppose that Assumption 1.1 holds. Then,*

(i) $x^\star \in \mathbb{R}^n$ *is a stationary point of* f*, i.e.,* $\nabla f(x^\star) = 0$*, if and only if* $\nabla \phi^*(\nabla f(x^\star)) = 0$

(ii) $\phi(z) \geq 0$ *for all* $z \in \mathbb{R}^n$*.*

Proof. Strong convexity of ϕ implies that ϕ has a unique stationary point $z^\star$ where $\nabla\phi(z^\star) = 0$. Since ϕ is even, $\phi(z^\star) = \phi(-z^\star)$, meaning $-z^\star$ is also a minimizer. By uniqueness, $z^\star = -z^\star$, which implies $z^\star = 0$. Thus, $\nabla\phi(0) = 0$. From the Fenchel-Moreau identity $\nabla\phi(z) = y \iff \nabla\phi^*(y) = z$, it follows that $\nabla\phi^*(y) = 0$ if and only if $y = 0$, which proves the first claim.

The second claim then follows from the fact that $\phi(0) = 0$ and from the fact that 0 is the minimizer of ϕ. □

Lemma B.2. *Let $\phi = h \circ \|\cdot\|$ be isotropic and suppose that $\nabla f(x) \neq 0$. Then, $\alpha(x) := \frac{h^{*\prime}(\|\nabla f(x)\|)}{h^{*\prime\prime}(\|\nabla f(x)\|)\|\nabla f(x)\|}$ can be expressed as follows.*

(i) *If $h(x) = \cosh(x) - 1$, then $\alpha(x) = \operatorname{arcsinh}(\|\nabla f(x)\|)\frac{\sqrt{1+\|\nabla f(x)\|^2}}{\|\nabla f(x)\|}$.*

(ii) *If $h(x) = \exp(|x|) - |x| - 1$, then $\alpha(x) = \ln(1 + \|\nabla f(x)\|)\frac{1+\|\nabla f(x)\|}{\|\nabla f(x)\|}$.*

(iii) *If $h(x) = -|x| - \ln(1 - |x|)$, then $\alpha(x) = 1 + \|\nabla f(x)\|$.*

Moreover, $\alpha(x) \geq 1$ for these three kernel functions.

Proof. The first claim follows immediately by substituting into the definition of α the expressions for $h^{*\prime}$ and $h^{*\prime\prime}$, which can be found in (Oikonomidis et al., 2025b, Table 1).

As for the second claim, standard calculus reveals that for any $y \geq 0$

$$\operatorname{arcsinh}(y)\frac{\sqrt{1+y^2}}{y} \geq 1, \qquad \ln(1+y)\frac{1+y}{y} \geq 1, \qquad 1 + y \geq 1,$$

which proves the claim. □

C Proofs of Section 2

C.1 Proof of Example 2.3

Proof. Straightforward manipulations reveal that $H(x) = f''(x)(1 + (f'(x))^2)^{-1/2}$. It suffices to uniformly upper bound the magnitude of $H'(x)$, which is of the form

$$H'(x) = \frac{f'''(x)}{\sqrt{1 + (f'(x))^2}} - \frac{(f''(x))^2 f'(x)}{(1 + (f'(x))^2)^{3/2}}.$$

Since $H'(x)$ is continuous on $\mathbb{R}$ (because the denominators are always ≥ 1), a sufficient condition for boundedness of $H'(x)$ is that $\lim_{x\to\pm\infty} H'(x) = 0$. Without loss of generality, assume that $f(x) = \sum_i^d \alpha_i x^i$, where $\alpha_d \neq 0$. If $d \geq 3$, then,

$$f'(x) \sim d\alpha_d x^{d-1}, \quad f''(x) \sim d(d-1)\alpha_d x^{d-2}, \quad f'''(x) \sim d(d-1)(d-2)\alpha_d x^{d-3}.$$

We find that $H'(x)$ is the sum of two terms $\mathcal{O}(x^{-2})$, hence $\lim_{x\to\pm\infty} H'(x) = 0$. If $d = 2$, then the first term is 0 (since $f'''(x) = 0$) and the second term is again $\mathcal{O}(x^{-2})$. As before, we conclude $\lim_{x\to\pm\infty} H'(x) = 0$. Finally, for $d \leq 1$ the claim is trivial since $f''(x) = f'''(x) = 0$. This concludes the proof. □

C.2 Proof of Theorem 2.4

We first establish the following auxiliary result.

Lemma C.1. *For any $y \in \mathbb{R}^n \setminus \{0\}$, define $P(y) := \frac{yy^\top}{\|y\|^2}$. Then, for any $w \in \mathbb{R}^n$ we have that*

$$\|DP(y)[w]\| \leq \frac{\|w\|}{\|y\|}.$$

Proof. Define $r := \|y\|$. Since $P(y) = \frac{y}{r}\frac{y}{r}^\top$ we find that

$$DP(y)[w] = D\left(\frac{y}{r}\right)[w]\frac{y}{r}^\top + \frac{y}{r}D\left(\frac{y}{r}\right)[w]^\top = \frac{(I-P(y))w\frac{y}{r}^\top + \frac{y}{r}w^\top(I-P(y))}{r}.$$

Since $I - P(y)$ projects orthogonally to $\frac{y}{r}$, we find that $\|DP(y)[w]\| = \frac{\|(I-P(y))w\|}{r}$. The claim then follows since an orthogonal projector has operator norm 1. □

We now provide the main proof to Theorem 2.4.

Proof. The effective domain of h is the open interval $(-\frac{1}{L_1}, \frac{1}{L_1})$. Note that $h'(z) = \frac{L_0 z}{1-L_1|z|}$ and $h''(z) = \frac{L_0}{(1-L_1|z|)^2}$. It is easily verified that this kernel function h satisfies our standing assumptions. By properties of the convex conjugate, we find $h^{*\prime}(r) = \frac{r}{L_0+L_1|r|}$ and $h^{*\prime\prime}(r) = \frac{L_0}{(L_0+L_1|r|)^2}$. Since the reference function is isotropic, we find for $y \in \mathbb{R}^n \setminus \{0\}$ that

$$\nabla^2\phi^*(y) = \frac{L_0}{(L_0+L_1\|y\|)^2}P(y) + \frac{1}{L_0+L_1\|y\|}\left(\mathrm{I} - P(y)\right), \qquad \text{where} \qquad P(y) := \frac{yy^\top}{\|y\|^2},$$

and else $\nabla^2\phi^*(0) = \frac{1}{L_0}\,\mathrm{I}$, which is the continuous extension at $y = 0$. This Hessian has two eigenvalues: $\lambda_1 := \frac{L_0}{(L_0+L_1\|y\|)^2}$ and $\lambda_2 := \frac{1}{L_0+L_1\|y\|}$. Thus, for all y we obtain the bound $\|\nabla^2\phi^*(y)\| \le \max\{\lambda_1, \lambda_2\} \le \frac{1}{L_0+L_1\|y\|}$.

By regrouping terms, we can write $\nabla^2\phi^*(y) = a(r)\,\mathrm{I} + b(r)P(y)$, where $r = \|y\|$, $a(r) := \frac{1}{L_0+L_1 r}$ and $b(r) := -\frac{L_1 r}{(L_0+L_1 r)^2}$. Clearly, $|a'(r)| = \frac{L_1}{(L_0+L_1 r)^2}$ and $|b'(r)| = |\frac{L_1(L_1 r - L_0)}{(L_0+L_1 r)^3}| \le \frac{L_1}{(L_0+L_1 r)^2}$. The sum and product rule along with Lemma C.1 then yield for any $y \in \mathbb{R}^n \setminus \{0\}$ and $w \in \mathbb{R}^n$ that

$$\|D\nabla^2\phi^*(y)[w]\| \le |a'(r)|\|w\| + |b(r)|\|DP(y)[w]\| + |b'(r)|\|P(y)\|\|w\| \le \frac{3L_1}{(L_0+L_1\|y\|)^2}\|w\|.$$

Together with continuity of $\nabla^2\phi^*$ at the origin, this establishes global Lipschitz continuity of $\nabla^2\phi^*$.

Now, let $H(z) := \nabla^2\phi^*(\nabla f(z))\nabla^2 f(z)$. Then for any point z such that $\nabla f(z) \neq 0$ and any direction v the chain and product rule yields

$$DH(z)[v] = D\nabla^2\phi^*(\nabla f(z))[\nabla^2 f(z)v]\nabla^2 f(z) + \nabla^2\phi^*(\nabla f(z))D\nabla^2 f(z)[v].$$

Using (L_0, L_1)-smoothness and the bound on $\|D\nabla^2\phi^*(\cdot)[\cdot]\|$, the first term can be bounded by

$$\|D\nabla^2\phi^*(\nabla f(z))[\nabla^2 f(z)v]\nabla^2 f(z)\| \le \frac{3L_1}{(L_0+L_1\|\nabla f(z)\|)^2}\|\nabla^2 f(z)v\|\|\nabla^2 f(z)\| \le 3L_1\|v\|.$$

For the second term, (4) and the bound on $\|\nabla^2\phi^*(\cdot)\|$ yield $\|\nabla^2\phi^*(\nabla f(z))D\nabla^2 f(z)[v]\| \le M\|v\|$. Therefore we conclude that, $\|DH(z)[v]\| \le (M+3L_1)\|v\|$.

If H were continuously differentiable everywhere with this bound holding for any z, then the claim would follow by integrating along the line segment between x and y. However, at points z where $\nabla f(z) = 0$, this need not be the case. To address this, fix $x, y \in \mathbb{R}^n$, set $d := y - x$, and define

$$\gamma(t) := x + td, \qquad \Psi(t) := H(\gamma(t)).$$

Since $\nabla^2\phi^*$ is globally Lipschitz continuous and $f \in C^3$, the path Ψ is absolutely continuous. Whenever $\nabla f(\gamma(t)) \neq 0$, the preceding calculation gives

$$\|\Psi'(t)\| \le (M+3L_1)\|d\|.$$

On the set $E := \{t \in [0,1] : \nabla f(\gamma(t)) = 0\}$, one has $(\nabla f \circ \gamma)'(t) = 0$ for almost every $t \in E$. Since $\nabla^2\phi^*$ is Lipschitz, it follows that $\left(\nabla^2\phi^* \circ \nabla f \circ \gamma\right)'(t) = 0$ for almost every $t \in E$. Therefore, for almost every $t \in E$, $\Psi'(t) = \nabla^2\phi^*(0)\,D\nabla^2 f(\gamma(t))[d]$. Using $\nabla^2\phi^*(0) = L_0^{-1}I$ and (4), we obtain

$$\|\Psi'(t)\| \le \frac{1}{L_0}M\big(L_0 + L_1\|\nabla f(\gamma(t))\|\big)\|d\| = M\|d\|.$$

Hence $\|\Psi'(t)\| \le (M+3L_1)\|d\|$ for almost every $t \in [0,1]$. The fundamental theorem of calculus for absolutely continuous functions now yields

$$\|H(y) - H(x)\| \le \int_0^1 \|\Psi'(t)\|\,dt \le (M+3L_1)\|y-x\|.$$

This proves the claim. □

C.3 Proof of Corollary 2.5

Proof. By symmetry of $D\nabla^2 f(x)[v]$ we obtain

$$\|D\nabla^2 f(x)[v]\| = \sup_{\|u\|=1} \left|u^\top D\nabla^2 f(x)[v]u\right| = \sup_{\|u\|=1} |\nabla^3 f(x)[u,u,v]| \leq M\|\nabla^2 f(x)\|\,\|v\|.$$

Then (4) follows by (L_0, L_1)-smoothness and the claim follows from Theorem 2.4. □

C.4 Proof of Example 2.7

The first claim is proven in (Bodard & Patrinos, 2025, Theorem 2.12).

For the second claim, note that the objective f is a fourth-order polynomial in the variables x. As such, it is easily verified that $\|\nabla^3 f(x)\| \leq C_1\|x\|$, for some constant $C_1 > 0$. By (Bodard & Patrinos, 2025, Lemma C.1) there also exists a constant $C_2 > 0$ such that $\|\nabla f(x)\| \geq C_2\|x\|^3 - \sum_{j=1}^m \|y_j^2 a_j a_j^\top\|\|x\|$. Therefore,

$$\limsup_{\|x\|\to\infty} \frac{\|\nabla^3 f(x)\|}{\|\nabla f(x)\|} \leq \limsup_{\|x\|\to\infty} \frac{C_1\|x\|}{C_2\|x\|^3 - \sum_{j=1}^m \|y_j^2 a_j a_j^\top\|\|x\|} = 0.$$

By nonnegativity of norms, we conclude that $\lim_{\|x\|\to\infty} \frac{\|\nabla^3 f(x)\|}{\|\nabla f(x)\|} = 0$. We now mirror the arguments of (Bodard & Patrinos, 2025, Theorem 2.9). The obtained limit implies that for any $\widetilde{L_1} > 0$ there exists a $\delta > 0$ such that

$$\|x\| > \delta \Rightarrow \|\nabla^3 f(x)\| \leq \widetilde{L_1}\|\nabla f(x)\|.$$

Moreover, by continuity of $\nabla^3 f$, the norm $\|\nabla^3 f(x)\|$ remains bounded on the compact set $\Omega := \{x \mid \|x\| \leq \delta\}$. Thus, it follows that

$$\|\nabla^3 f(x)\| \leq \widetilde{L_0} + \widetilde{L_1}\|\nabla f(x)\| \leq \max\left\{\frac{\widetilde{L_0}}{L_0}, \frac{\widetilde{L_1}}{L_1}\right\}(L_0 + L_1\|\nabla f(x)\|)$$

where $\widetilde{L_0} = \max_{x\in\Omega}\|\nabla^3 f(x)\|$. Thus, (4) holds with $M = \max\left\{\frac{\widetilde{L_0}}{L_0}, \frac{\widetilde{L_1}}{L_1}\right\}$.

C.5 Proof of Example 2.8

We can trace the arguments from Example 2.7 and instead use the gradient norm lower bound from (Bodard & Patrinos, 2025, Lemma C.2).

C.6 Proof of Theorem 2.9

Proof. Let $g(x) = \nabla\phi^*(\nabla f(x))$ with Jacobian matrix $H(x) = \nabla^2\phi^*(\nabla f(x))\nabla^2 f(x)$. The iteration (PN) can then be written as $x^{k+1} = x^k - H(x^k)^{-1}g(x^k)$.

"Theorem 2.9(i)": At $x^\star$, we have $H(x^\star) = \nabla^2\phi^*(\nabla f(x^\star))\nabla^2 f(x^\star) = \nabla^2\phi^*(0)\nabla^2 f(x^\star)$. Since ϕ is strongly convex, we have $\nabla^2\phi(0) \succ 0$. By properties of the convex conjugate, $\nabla^2\phi^*(0) = (\nabla^2\phi(0))^{-1} \succ 0$. As $\nabla^2 f(x^\star) \succeq \mu_f I$, we observe that $H(x^\star)$ is the product of two positive definite matrices and is thus invertible. By continuity of $H(x)$, there exists a $\delta > 0$ and $\mu > 0$ such that for all $x \in \mathcal{B}(x^\star; \delta)$, the preconditioned Hessian $H(x)$ is invertible and $\|H(x)^{-1}\| \leq 1/\mu$.

Taylor's theorem with integral remainder states that for $g(x)$ we have

$$g(y) = g(x) + H(x)(y - x) + R(x, y),$$

with the residual term given by $R(x,y) = \int_0^1 [H(x+t(y-x)) - H(x)]\,(y-x)\mathrm{d}t$. Using the iteration map $x^{k+1} = x^k - H(x^k)^{-1}g(x^k)$, we can write the error $x^{k+1} - x^\star$ as

$$\begin{aligned} x^{k+1} - x^\star &= x^k - H(x^k)^{-1}g(x^k) - x^\star \\ &= (x^k - x^\star) - H(x^k)^{-1}g(x^k) \\ &= H(x^k)^{-1}\left(H(x^k)(x^k - x^\star) - g(x^k)\right) \\ &= H(x^k)^{-1}\left(H(x^k)(x^k - x^\star) - (g(x^k) - g(x^\star))\right) \\ &= H(x^k)^{-1}\left(g(x^\star) - g(x^k) - H(x^k)(x^\star - x^k)\right) \\ &= H(x^k)^{-1}R(x^k, x^\star). \end{aligned}$$

Here the fourth equality uses $g(x^\star) = 0$ (cf. Lemma B.1). Note that for all $x \in \mathbb{R}^n$ we have

$$\begin{aligned} \|R(x,x^\star)\| &\le \int_0^1 \|[H(x+t(x^\star - x)) - H(x)](x^\star - x)\|\mathrm{d}t \\ &\le \|x^\star - x\| \int_0^1 \|H(x+t(x^\star - x)) - H(x)\|\mathrm{d}t \end{aligned}$$

and by continuity of H it moreover holds that $\lim_{x\to x^\star}\int_0^1 \|H(x+t(x^\star - x)) - H(x)\|\mathrm{d}t = 0$.

Thus, we can shrink $\delta > 0$ (if necessary) such that for all $x \in \mathcal{B}(x^\star;\delta)$ it holds that

$$\int_0^1 \|H(x+t(x^\star - x)) - H(x)\|\mathrm{d}t \le \frac{\mu}{2},$$

which implies $\|R(x,x^\star)\| \le \frac{\mu}{2}\|x - x^\star\|$. Using this bound together with the fact that $\|H^{-1}(x^k)\| \le 1/\mu$ for all $x^k \in \mathcal{B}(x^\star;\delta)$, we obtain

$$\|x^{k+1} - x^\star\| \le \|H^{-1}(x^k)\|\|R(x^k,x^\star)\| \le \frac{1}{\mu}\left(\frac{\mu}{2}\|x^k - x^\star\|\right) = \frac{1}{2}\|x^k - x^\star\|,$$

which shows by induction that the iterates remain in $\mathcal{B}(x^\star;\delta)$ and converge (at least) Q-linearly to $x^\star$. The Q-superlinear convergence follows from the fact that

$$\lim_{k\to\infty}\frac{\|x^{k+1} - x^\star\|}{\|x^k - x^\star\|} \le \lim_{k\to\infty}\|H^{-1}(x^k)\|\int_0^1 \|H(x^k + t(x^\star - x^k)) - H(x^k)\|\mathrm{d}t = 0.$$

Indeed, as $k \to \infty$, we have $x^k \to x^\star$. The first term satisfies $\|H^{-1}(x^k)\| \le 1/\mu$, and the second term (the integral) goes to 0 by continuity of H.

"Theorem 2.9(ii)": Since H is L_H-Lipschitz continuous, we have

$$H(x^k) = H(x^\star) + E(x^k), \qquad E(x^k) := H(x^k) - H(x^\star), \qquad \text{where } \|E(x^k)\| \le L_H\|x^k - x^\star\|.$$

We remark that $\nabla^2\phi^*(0) = \mu_\phi^{-1}\,\mathrm{I}$ is a multiple of identity, and hence $H(x^\star)$ is symmetric positive definite and satisfies $\lambda_{\min}(H(x^\star)) \ge \frac{\mu_f}{\mu_\phi} > 0$. For any $v \in \mathbb{R}^n$ with $\|v\| = 1$, we have

$$\|H(x^k)v\| = \|(H(x^\star) + E(x^k))v\| \ge \|H(x^\star)v\| - \|E(x^k)v\| \ge \tfrac{\mu_f}{\mu_\phi} - L_H\|x^k - x^\star\|.$$

Taking the infimum over all unit vectors yields

$$\sigma_{\min}(H(x^k)) = \inf_{\|v\|=1}\|H(x^k)v\| \ge \frac{\mu_f}{\mu_\phi} - L_H\|x^k - x^\star\|.$$

Consequently, if $\|x^k - x^\star\| < \frac{\mu_f}{\mu_\phi L_H}$, then $H(x^k)$ is invertible and

$$\|H(x^k)^{-1}\| = \sigma_{\min}(H(x^k))^{-1} \le \left(\frac{\mu_f}{\mu_\phi} - L_H\|x^k - x^\star\|\right)^{-1} = \frac{\mu_\phi}{\mu_f - \mu_\phi L_H\|x^k - x^\star\|}.$$

Assumption 2.2 also implies that $\|H(x+t(x^\star - x)) - H(x)\| \le L_H\|t(x^\star - x)\| = L_H t\|x - x^\star\|$. Plugging this into the residual bound yields

$$\|R(x^k,x^\star)\| \le \|x^k - x^\star\|\int_0^1 (L_H t\|x^k - x^\star\|)\mathrm{d}t = L_H\|x^k - x^\star\|^2\int_0^1 t\mathrm{d}t = \frac{L_H}{2}\|x^k - x^\star\|^2.$$

The desired quadratic rate now follows:

$$\|x^{k+1} - x^\star\| \leq \|H^{-1}(x^k)\|\|R(x^k, x^\star)\| \leq \frac{\mu_\phi L_H}{2(\mu_f - \mu_\phi L_H \|x^k - x^\star\|)}\|x^k - x^\star\|^2.$$

Remark also that $\|x^k - x^\star\| \leq \Delta = \frac{2\mu_f}{3\mu_\phi L_H}$ implies

$$\|x^{k+1} - x^\star\| \leq \frac{\mu_\phi L_H}{2(\mu_f - \mu_\phi L_H \Delta)}\|x^k - x^\star\|^2 = \frac{1}{\Delta}\|x^k - x^\star\|^2 \leq \Delta.$$

□

C.7 Proof of Theorem 2.10

Proof. We first prove the well-definedness of the linesearch procedure. Then, we establish the claimed convergence rate.

Well-definedness of the linesearch procedure Let us fix an arbitrary iterate $k \in \mathbb{N}$. We omit sub- and superscripts for notational simplicity. Suppose, by contradiction, that for all $\varepsilon > 0$ there exists $\tau_\varepsilon \in [0, \varepsilon]$ such that the point $x_\varepsilon := x + \tau_\varepsilon d + (1 - \tau_\varepsilon)p$ satisfies

$$f(x_\varepsilon) - f(x) > -\gamma\sigma\phi(\nabla\phi^*(\nabla f(x))).$$

By taking the limit $\varepsilon \searrow 0$ it follows by (strict) continuity of f that

$$f(x + p) - f(x) \geq -\gamma\sigma\phi(\nabla\phi^*(\nabla f(x))). \tag{11}$$

On the other hand, the anisotropic smoothness inequality (3) evaluated at the points $x + p$ and x yields

$$f(x + p) - f(x) \leq L^{-1}\left[\phi((1 - \alpha)\nabla\phi^*(\nabla f(x))) - \phi(\nabla\phi^*(\nabla f(x)))\right].$$

By convexity of ϕ we have that for any $\theta \in [0, 1]$

$$\phi(\theta x) = \phi((1 - \theta)0 + \theta x) \leq (1 - \theta)\phi(0) + \theta\phi(x) = \theta\phi(x). \tag{12}$$

Note that $\alpha \in (0, 1)$ and $\gamma = \alpha/L$, such that we obtain

$$f(x + p) - f(x) \leq -\gamma\phi(\nabla\phi^*(\nabla f(x))).$$

Since $\sigma \in (0, 1)$, the combination of (11) and (12) yields a contradiction, i.e., $f(x + p) - f(x) > f(x + p) - f(x)$, which proves the well-definedness of the linesearch procedure.

Convergence rate The linesearch condition guarantees that for all $k \geq 0$

$$f(x^{k+1}) - f(x^k) \leq -\gamma\sigma\phi(\nabla\phi^*(\nabla f(x^k))).$$

A standard telescoping argument yields

$$\sum_{k=0}^{K} \phi(\nabla\phi^*(\nabla f(x^k))) \leq \frac{f(x^0) - f(x^K)}{\gamma\sigma} \leq \frac{f(x^0) - \inf f}{\gamma\sigma}.$$

Lower bounding $\sum_{k=0}^{K} \phi(\nabla\phi^*(\nabla f(x^k))) \geq (K + 1)\min_{0\leq k\leq K} \phi(\nabla\phi^*(\nabla f(x^k)))$ and substituting $\gamma = \alpha/L$ establishes the claim. □

C.8 Proof of Theorem 2.11

Proof. By Theorem 2.9 the update directions d^k are eventually Q-superlinear, meaning that

$$\lim_{k\to\infty} \frac{\|x^k + d^k - x^\star\|}{\|x^k - x^\star\|} = 0.$$

Since $\nabla^2 f(x^\star) \succ 0$, there exist constants $c, C > 0$ and a neighborhood of $x^\star$ such that

$$c\|z - x^\star\|^2 \leq f(z) - f(x^\star) \leq C\|z - x^\star\|^2$$

for all z sufficiently close to $x^\star$. Hence, it follows that

$$0 \leq \varepsilon_k := \frac{f(x^k + d^k) - f(x^\star)}{f(x^k) - f(x^\star)} \leq \frac{C}{c} \left(\frac{\|x^k + d^k - x^\star\|}{\|x^k - x^\star\|} \right)^2 \to 0 \qquad \text{as } k \to \infty.$$

Since $x^\star$ is a strong local minimum, and since $x^k + p^k \to x^\star$, eventually we must have $f(x^k + p^k) \geq f(x^\star)$. From the anisotropic descent inequality we obtain that

$$f(x^k) - f(x^\star) \geq f(x^k) - f(x^k + p^k) \geq \gamma\phi(\nabla\phi^*(\nabla f(x^k))).$$

Consequently,

$$\begin{aligned} f(x^k + d^k) - f(x^k) &= -(1 - \varepsilon_k)(f(x^k) - f(x^\star)) \\ &\leq -(1 - \varepsilon_k)\gamma\phi(\nabla\phi^*(\nabla f(x^k))) \\ &\leq -\gamma\sigma\phi(\nabla\phi^*(\nabla f(x^k))) \end{aligned}$$

where the last step follows from $\varepsilon_k \to 0$ and $\sigma < 1$ so that eventually $1 - \varepsilon_k \geq \sigma$. We conclude that eventually (for large enough k) the linesearch condition holds and unit stepsize $\tau_k = 1$ is always accepted. Algorithm 1 then reduces to $x^{k+1} = x^k + d^k$, where d^k satisfies (PN). □

D Proofs of Section 3

D.1 Proof of Theorem 3.2

We aim to find a pair $(s^\star, \lambda_\star) \in \mathbb{R}^n \times \mathbb{R}$ satisfying

$$(A + \lambda_\star M)s^\star = -Mg, \tag{13a}$$

$$\lambda_\star = \sigma\|s^\star\|, \tag{13b}$$

$$A + \lambda_\star M \succeq 0. \tag{13c}$$

Denote by $\xi_1 \leq \cdots \leq \xi_n$ the generalized eigenvalues of the pencil (A, M), and by v_i for $1 \leq i \leq n$ the corresponding generalized eigenvectors. The diagonal matrix Ξ has the generalized eigenvalues as diagonal elements, V is the matrix whose columns are the generalized eigenvectors, i.e.,

$$AV = MV\Xi, \qquad \text{where} \qquad V^\top MV = \mathrm{I}. \tag{14}$$

We first present two auxiliary lemmas.

Lemma D.1. *V is invertible with $V^{-1} = V^\top M$.*

Proof. Starting from $V^\top MV = \mathrm{I}$ and applying the determinant yields

$$\det(V^\top MV) = \det(\mathrm{I}) = 1.$$

From the multiplicative property of determinants it follows that

$$\det(V^\top MV) = \det(V^\top)\det(M)\det(V) = \det(V)^2\det(M).$$

Therefore,

$$|\det(V)| = \frac{1}{\sqrt{\det(M)}} \neq 0,$$

where the inequality follows from $\det M > 0$ (by positive definiteness). We conclude that V is invertible. The expression for V^{-1} follows by right-multiplying $V^\top MV = \mathrm{I}$ by V^{-1} (or noting that the inverse is unique). □

Lemma D.2. *$A + \lambda M$ is positive semidefinite if and only if $\lambda \geq -\xi_1$.*

Proof. Positive semidefiniteness of $A + \lambda M$ means that

$$\langle u, (A + \lambda M)u \rangle \geq 0$$

for all non-zero $u \in \mathbb{R}^n$. Since V is invertible by Lemma D.1, the mapping $u = Vy$ is a bijection. Substituting this into the quadratic form:

$$\begin{aligned}\langle u, (A + \lambda M)u\rangle &= \langle Vy, (A + \lambda M)Vy\rangle \\ &= y^\top (V^\top AV + \lambda V^\top MV)y \\ &= y^\top (\Xi + \lambda \mathrm{I})y \\ &= \sum_{i=1}^{n} (\xi_i + \lambda) y_i^2.\end{aligned}$$

This sum is non-negative for all $y \neq 0$ if and only if $\xi_i + \lambda \geq 0$ for all i. Since the eigenvalues are ordered $\xi_1 \leq \cdots \leq \xi_n$, this is equivalent to $\lambda \geq -\xi_1$. □

Define $s_V(\lambda) = V^\top M s(\lambda)$, such that by Lemma D.1 $s(\lambda) = V s_V(\lambda)$. Then, by (14)

$$(A + \lambda M)s(\lambda) = (A + \lambda M)V s_V(\lambda) = MV(\Xi + \lambda \mathrm{I})s_V(\lambda) = -Mg$$

Since M is nonsingular, this yields $V(\Xi + \lambda \mathrm{I})s_V(\lambda) = -g$. If $\lambda \neq -\xi_i$ for $i \in \{1, \ldots, n\}$, then

$$s(\lambda) = V s_V(\lambda) = -V(\Xi + \lambda \mathrm{I})^{-1} V^{-1} g = -V(\Xi + \lambda \mathrm{I})^{-1} V^\top Mg.$$

We are now ready to prove Theorem 3.2.

Proof. By Lemma D.2, we have that $A + \lambda M \succeq 0$ if and only if $\lambda \geq -\xi_1$. Since $\lambda = \sigma\|s\| \geq 0$, the admissible domain is $\lambda \in [\lambda_s, +\infty)$, where $\lambda_s := \max\{-\xi_1, 0\}$. Thus any value of $\lambda \in [\lambda_s, +\infty)$ satisfies (13c).

For all $\lambda > \lambda_s$, the matrix $A + \lambda M$ is positive definite, so based on (13a) we have that $s(\lambda) = -(A + \lambda M)^{-1} Mg$. Define

$$\pi(\lambda) := \|s(\lambda)\|^2, \qquad \Psi(\lambda) := \pi(\lambda) - \frac{\lambda^2}{\sigma^2}.$$

Using the generalized eigenvalue decomposition of the pencil (A, M), with generalized eigenvalues $\xi_1 \leq \cdots \leq \xi_n$ and corresponding eigenvectors $v_1, \ldots, v_n$, we obtain the expression

$$s(\lambda) = -\sum_{i=1}^{n} \frac{g_{V\,i}}{\xi_i + \lambda} v_i, \qquad g_{V\,i} := v_i^\top Mg. \tag{15}$$

Any pair $(s(\lambda), \lambda)$ satisfies (13a), and it remains to enforce (13b) by finding a λ for which $\Psi(\lambda) = 0$. Define $c(\lambda) \in \mathbb{R}^n$ as the vector with components $c_i(\lambda) := \frac{g_{V\,i}}{\xi_i + \lambda}$ for $1 \leq i \leq n$, and $G \in \mathbb{R}^{n \times n}$ as the Gram matrix with entries $G_{ij} = v_i^\top v_j$ for $1 \leq i, j \leq n$. Note that the Gram matrix G is positive definite by linear independence of the vectors v_i, which follows by invertibility of V (cf. Lemma D.1). Then we have that $\pi(\lambda) = \|s(\lambda)\|^2 = c(\lambda)^\top G c(\lambda)$. It follows that $\pi(\lambda)$ and therefore also $\Psi(\lambda)$ are continuous on $(\lambda_s, +\infty)$. Moreover, by positive definiteness of G and since $c(\lambda) \to 0$ as $\lambda \to +\infty$ we have

$$0 \leq \lim_{\lambda \to +\infty} \lambda_{\min}(G)\|c(\lambda)\|^2 \leq \lim_{\lambda \to +\infty} \pi(\lambda) \leq \lim_{\lambda \to +\infty} \lambda_{\max}(G)\|c(\lambda)\|^2 = 0.$$

We conclude that $\lim_{\lambda \to +\infty} \pi(\lambda) = 0$ and hence $\lim_{\lambda \to +\infty} \Psi(\lambda) = \lim_{\lambda \to +\infty} -\frac{\lambda^2}{\sigma^2} = -\infty$. We now distinguish three cases:

Case 1: $\xi_1 \geq 0$. If $g = 0$ and $\xi_1 \geq 0$, then $A \succeq 0$ and the pair $s^\star = 0, \lambda_\star = 0$ satisfies (13a) - (13c). Suppose therefore that $g \neq 0$. Clearly, $\xi_1 \geq 0$ yields $\lambda_s = 0$. The limit of Ψ as $\lambda \searrow \lambda_s$

$$\lim_{\lambda \searrow \lambda_s} \Psi(\lambda) = \lim_{\lambda \searrow 0} \pi(\lambda) \geq \lim_{\lambda \searrow 0} \lambda_{\min}(G)\|c(\lambda)\|^2 > 0$$

may be finite or infinite, but is never equal to zero. Indeed, $\lim_{\lambda \searrow 0} \|c(\lambda)\| = 0$ if and only if $g = 0$, which is excluded by assumption. Thus, by the intermediate value theorem there exists a $\lambda_\star \in (\lambda_s, +\infty)$ for which $\Psi(\lambda_\star) = 0$. This means that $s^\star = s(\lambda_\star)$ as in (15) is well-defined, and that the pair $(s^\star, \lambda_\star)$ satisfies (13a) - (13c).

Case 2: $\xi_1 < 0$ and there exists at least one i with $\xi_i = \xi_1$ and $g_{V\,i} \neq 0$. Hence $\lambda_s = -\xi_1 > 0$. It follows that

$$\lim_{\lambda \searrow \lambda_s} \pi(\lambda) \geq \lim_{\lambda \searrow \lambda_s} \lambda_{\min}(G)\|c(\lambda)\|^2 = +\infty.$$

and hence that $\lim_{\lambda \searrow \lambda_s} \Psi(\lambda) = +\infty$. The reasoning from Case 1 applies again.

Case 3: $\xi_1 < 0$ and $g_{V\,i} = 0$ for all $i \in \mathcal{I} := \{1 \leq i \leq n \mid \xi_i = \xi_1\}$. We have that $\lambda_s = -\xi_1 > 0$, and the expression (15) for $s(\lambda)$ can be written as

$$s(\lambda) = -\sum_{\substack{i=1 \\ i \notin \mathcal{I}}}^{n} \frac{g_{V\,i}}{\xi_i + \lambda} v_i$$

In this case the limit

$$\lim_{\lambda \searrow \lambda_s} \pi(\lambda) = \lim_{\lambda \searrow \lambda_s} \|s(\lambda)\|^2 < +\infty.$$

This implies that $\Psi(\lambda_s) := \lim_{\lambda \searrow \lambda_s} \Psi(\lambda)$ is finite. If $\Psi(\lambda_s) > 0$, then the intermediate value theorem again guarantees the existence of a $\lambda_\star \in (\lambda_s, +\infty)$ for which $\Psi(\lambda_\star) = 0$ and for which $(s(\lambda_\star), \lambda_\star)$ satisfies (13a) - (13c).

Otherwise $\Psi(\lambda_s) \leq 0$. This is the *hard case*. The vector $s_s := \lim_{\lambda \searrow \lambda_s} s(\lambda)$ is well-defined and satisfies $(A + \lambda_s M)s_s = -Mg$. Since $\Psi(\lambda_s) \leq 0$ we find $\|s_s\| \leq \frac{\lambda_s}{\sigma}$, whereas equality is required. The vector $s(\lambda_s) = -(A + \lambda_s M)^\dagger Mg$ is the minimal Euclidean norm solution to the same linear system, hence $\|s(\lambda_s)\| \leq \|s_s\| \leq \frac{\lambda_s}{\sigma}$. Because $\xi_1 = -\lambda_s$ is a generalized eigenvalue of the pencil (A, M) corresponding to v_1, i.e., $(A + \lambda_s M)v_1 = 0$, we have for any $\alpha \in \mathbb{R}$ that

$$(A + \lambda_s M)(s(\lambda_s) + \alpha v_1) = -Mg.$$

Note that $\alpha \mapsto \|s(\lambda_s) + \alpha v_1\|^2$ is continuous, quadratic and unbounded above. One can therefore always select an $\alpha_s \in \mathbb{R}$ for which $\|s(\lambda_s) + \alpha_s v_1\|^2 = \frac{\lambda_s^2}{\sigma^2}$, and it then follows that the pair $(s(\lambda_s) + \alpha_s v_1, \lambda_s)$ satisfies (13a) - (13c). □

D.2 Proof of Lemma 3.4

Proof. For isotropic references $\phi = h \circ \|\cdot\|$, the domain ϕ is radial, and $\operatorname{int} \operatorname{dom} \phi$ is an open ball around 0 with some radius $\rho \in (0, \infty]$. Let us fix a unit vector $e \in \mathbb{R}^n$, satisfying $\|e\| = 1$. Then, the restriction $t \mapsto \phi(te) = h(|t|) = h(t)$ is $\mathcal{C}^2$ near 0 by Assumption 1.1, hence $h \in \mathcal{C}^2([0, \rho))$. Using $\nabla\phi(x) = h'(\|x\|)\,\overline{\operatorname{sgn}}(x)$ and μ_ϕ-strong convexity of ϕ at points re and se with $0 \leq s < r < \rho$, we obtain

$$(h'(r) - h'(s))\,(r - s) \geq \mu_\phi (r - s)^2$$

By selecting $s = 0$ and using $h'(0) = 0$, we conclude that $h'(r) \geq \mu_\phi r$ for $r \in [0, \rho)$.

Consequently, on $[0, \rho)$ the derivative h' is strictly increasing and has range $[0, +\infty)$: if $\rho = +\infty$ the latter follows from $h'(r) \geq \mu_\phi r$, and otherwise Assumption 1.1 gives $h'(r) = \|\nabla\phi(re)\| \to +\infty$ as $r \uparrow \rho$. Thus, for every $y \geq 0$, there exists a unique $r \in [0, \rho)$ such that $h'(r) = y$. This r is the unique maximizer in the definition of the convex conjugate $h^*(y) := \sup_s \{ys - h(s)\}$. Indeed, by μ_ϕ-strong convexity of ϕ at points re and se we obtain for every $s \geq 0$ in the effective domain of h that

$$h(s) \geq h(r) + h'(r)(s - r) + \frac{\mu_\phi}{2}(s - r)^2 = h(r) + y(s - r) + \frac{\mu_\phi}{2}(s - r)^2$$

and because $\mu_\phi > 0$ this yields $yr - h(r) > ys - h(s)$ for $r \neq s \geq 0$. On the other hand, if $s < 0$, set $t := -s$. By evenness of h, we have $ys - h(s) = -yt - h(t) \leq yt - h(t) \leq yr - h(r)$. If $t \neq r$, the second inequality is strict; if $t = r > 0$ then $y = h'(r) > 0$ so the first inequality is strict. Hence, r is the unique maximizer in the definition of h^* corresponding to $y \geq 0$. Thus, by properties of the convex conjugate, we find that

$$h^{*\prime}(y) = r = (h')^{-1}(y).$$

Since h' is continuous and strictly increasing, also its inverse $h^{*\prime}$ is continuous and strictly increasing. For $y > 0$, this yields $r = (h^*)'(y) > 0$ and

$$\xi(y) = \frac{y}{h^{*\prime}(y)} = \frac{h'(r)}{r} \geq \mu_\phi,$$

along with continuity of ξ on $(0, +\infty)$. It remains to verify the case $y = 0$. We have that

$$\lim_{y \searrow 0} \xi(y) = \lim_{y \searrow 0} \frac{y}{h^{*\prime}(y)} = \lim_{r \searrow 0} \frac{h'(r)}{r} = h''(0) =: \xi(0).$$

Thus, ξ is continuous on $[0, +\infty)$ and the uniform lower bound follows by strong convexity: $h''(0) \geq \mu_\phi$. The continuity and lower boundedness of ν then follow trivially, and the final claim follows from the fact that $\nu(x) \geq \mu_\phi > 0$ and

$$\nabla \phi^*(\nabla f(x)) = h^{*\prime}(\|\nabla f(x)\|) \, \overline{\operatorname{sgn}}(\nabla f(x)) = \begin{cases} \frac{1}{\nu(x)} \nabla f(x) & \text{if } \|\nabla f(x)\| \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

□

D.3 Proof of Lemma 3.5

Proof. Combining (5) and $M_k = \left[\nabla^2 \phi^*(\nabla f(x^k))\right]^{-1}$ yields

$$\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k = -M_k g(x^k) = -\frac{1}{\nu(x^k) h^{*\prime\prime}(\|\nabla f(x^k)\|)} \nabla f(x^k).$$

By positive semidefiniteness of $\nabla^2 f(x^k) + \lambda_k M_k$, by the uniform lower bound $\nu(x^k) \geq \mu_\phi > 0$ (Lemma 3.4) and by $h^{*\prime\prime}(\|\nabla f(x^k)\|) = \frac{1}{h''(h^{*\prime}(\|\nabla f(x^k)\|))} > 0$, the update directions s^k satisfy

$$\langle \nabla f(x^k), s^k \rangle = -\underbrace{\nu(x^k) h^{*\prime\prime}(\|\nabla f(x^k)\|)}_{>0} \underbrace{\langle s^k, \left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k \rangle}_{\geq 0} \leq 0. \tag{16}$$

This proves the claim. □

D.4 Proof of Lemma 3.6

Proof. Let ν be defined as in Lemma 3.4, and apply the fundamental theorem of calculus to f, i.e.,

$$f(x^k + s^k) - f(x^k) = \int_0^1 \langle \nabla f(x^k + ts^k), s^k \rangle \mathrm{d}t = \int_0^1 \nu(x^k + ts^k) \langle g(x^k + ts^k), s^k \rangle \mathrm{d}t.$$

We upper bound $\langle g(x^k + ts^k), s^k \rangle$. The fundamental theorem of calculus applied to g yields

$$g(x^k + ts^k) = g(x^k) + \int_0^t H(x^k + \tau s^k) s^k \mathrm{d}\tau = g(x^k) + tH(x^k)s^k + r(ts^k),$$

where the residual term equals

$$r(ts^k) := \int_0^t \left(H(x^k + \tau s^k) - H(x^k)\right) s^k \mathrm{d}\tau. \tag{17}$$

Since $H(x^k)s^k = -g(x^k) - \lambda_k s^k$ by (6) it follows that

$$g(x^k + ts^k) = (1 - t)g(x^k) - t\lambda_k s^k + r(ts^k).$$

Taking the inner product with s^k yields

$$\begin{aligned} \langle g(x^k + ts^k), s^k \rangle &= (1 - t)\langle g(x^k), s^k \rangle - t\lambda_k \|s^k\|^2 + \langle r(ts^k), s^k \rangle \\ &\leq -t\lambda_k \|s^k\|^2 + \langle r(ts^k), s^k \rangle, \end{aligned}$$

where the inequality follows from the fact that $1 - t \geq 0$ and $\langle g(x^k), s^k \rangle = \frac{1}{\nu(x^k)} \langle \nabla f(x^k), s^k \rangle \leq 0$ (cf. Lemmas 3.4 and 3.5). Under Assumption 2.2 we can upper bound

$$\|r(ts^k)\| \leq \int_0^t \left\| \left(H(x^k + \tau s^k) - H(x^k)\right) s^k \right\| \mathrm{d}\tau \leq \int_0^t L_H \tau \|s^k\|^2 \mathrm{d}\tau = \frac{L_H t^2}{2} \|s^k\|^2. \tag{18}$$

From Cauchy-Schwarz and because $\sigma \geq L_H$ we have that

$$\begin{aligned}\langle g(x^k + ts^k), s^k\rangle &\leq -t\lambda_k\|s^k\|^2 + \langle r(ts^k), s^k\rangle \\ &\leq -t\lambda_k\|s^k\|^2 + \|r(ts^k)\|\|s^k\| \\ &\leq -t\sigma\|s^k\|^3 + \frac{L_H t^2}{2}\|s^k\|^3 \\ &\leq -\left(tL_H - \frac{L_H t^2}{2}\right)\|s^k\|^3.\end{aligned}$$

Thus, using $\nu(z) \geq \mu_\phi$ for all $z \in \mathbb{R}^n$ (Lemma 3.4) we obtain

$$\begin{aligned}f(x^k + s^k) - f(x^k) &= \int_0^1 \nu(x^k + ts^k)\langle g(x^k + ts^k), s^k\rangle \mathrm{d}t \\ &\leq -\mu_\phi \int_0^1 \left(tL_H - \frac{L_H t^2}{2}\right)\|s^k\|^3 \mathrm{d}t \\ &\leq -\mu_\phi L_H\|s^k\|^3 \int_0^1 \left(t - \frac{t^2}{2}\right)\mathrm{d}t \\ &= -\mu_\phi L_H\|s^k\|^3\left(\frac{1}{2} - \frac{1}{6}\right) = -\frac{\mu_\phi L_H}{3}\|s^k\|^3.\end{aligned}$$

□

D.5 Proof of Lemma 3.7

Proof. By the fundamental theorem of calculus we have $g(x^k + s^k) = g(x^k) + H(x^k)s^k + r(s^k)$, where r is given by (17). Combined with (6) and $\lambda_k = \sigma\|s^k\|$ this yields

$$\begin{aligned}g(x^{k+1}) = g(x^k + s^k) &= g(x^k) + H(x^k)s^k + r(s^k) \\ &= -\lambda_k s^k + r(s^k) = -\sigma\|s^k\|s^k + r(s^k).\end{aligned}$$

The claim then follows from the triangle inequality and (18):

$$\|g(x^{k+1})\| \leq \sigma\|s^k\|^2 + \frac{L_H}{2}\|s^k\|^2 = \frac{2\sigma + L_H}{2}\|s^k\|^2.$$

□

D.6 Proof of Theorem 3.8

Proof. Lemma 3.7 and $\sigma \geq L_H$ ensure for $k \geq 1$ that

$$\|s^{k-1}\|^3 \geq \left(\frac{2}{3\sigma}\right)^{3/2}\|g(x^k)\|^{3/2}.$$

Moreover, Algorithm 2 ensures that $\nabla^2 f(x^k) + \lambda_k M_k \succeq 0$, for $k \geq 0$ and therefore

$$M_k^{-1/2}\nabla^2 f(x^k)M_k^{-1/2} + \lambda_k \mathrm{I} \succeq 0.$$

This matrix is similar to $H(x^k) + \lambda_k \mathrm{I}$, and hence we conclude that $\lambda_{\min}(H(x^k)) \geq -\lambda_k = -\sigma\|s^k\|$.

If $k^* = 1$, then the claimed bound $k^* \leq K$ holds trivially. Thus, let $K \geq 2$ be arbitrary. Summing the sufficient decrease condition from Lemma 3.6 yields

$$\frac{\mu_\phi L_H}{3}\sum_{k=0}^{K-1}\|s^k\|^3 \leq \sum_{k=0}^{K-1}(f(x^k) - f(x^{k+1})) \leq f(x^0) - \inf f.$$

Let $q_k := \max\{\|s^{k-1}\|, \|s^k\|\}$ for $k \geq 1$. Then we obtain

$$\sum_{k=1}^{K-1} q_k^3 \leq \sum_{k=1}^{K-1}\left(\|s^{k-1}\|^3 + \|s^k\|^3\right) \leq 2\sum_{k=0}^{K-1}\|s^k\|^3 \leq 6\frac{f(x^0) - \inf f}{\mu_\phi L_H}.$$

Thus, there exists an index $k^* \in \{1, \dots, K-1\}$ for which

$$q_{k^*}^3 \leq 6\frac{f(x^0) - \inf f}{\mu_\phi L_H (K-1)}.$$

Using $q_k^3 \geq \left(\frac{2}{3\sigma}\right)^{3/2} \|g(x^k)\|^{3/2}$ for $k \geq 1$ and $q_k^3 \geq \left(-\frac{\lambda_{\min}(H(x^k))}{\sigma}\right)^3$ for $k \geq 0$ this means that

$$\|g(x^{k^*})\| \leq \frac{3\sigma}{2}\left(6\frac{f(x^0) - \inf f}{\mu_\phi L_H (K-1)}\right)^{2/3}, \qquad \lambda_{\min}(H(x^{k^*})) \geq -\sigma\left(6\frac{f(x^0) - \inf f}{\mu_\phi L_H (K-1)}\right)^{1/3}.$$

Clearly, both $\|g(x^{k^*})\| \leq \varepsilon_g$ and $\lambda_{\min}(H(x^{k^*})) \geq -\varepsilon_H$ hold if

$$6\frac{f(x^0) - \inf f}{\mu_\phi L_H (K-1)} \leq \left(\frac{2}{3}\frac{\varepsilon_g}{\sigma}\right)^{3/2} \qquad 6\frac{f(x^0) - \inf f}{\mu_\phi L_H (K-1)} \leq \left(\frac{\varepsilon_H}{\sigma}\right)^3$$

which is the case if

$$K \geq 1 + 6\frac{f(x^0) - \inf f}{\mu_\phi L_H} \max\left\{\left(\frac{3\sigma}{2\varepsilon_g}\right)^{3/2}, \left(\frac{\sigma}{\varepsilon_H}\right)^3\right\}.$$

The claim follows. □

D.7 Proof of Corollary 3.9

Proof. Invoking Theorem 3.8 with tolerances $\widetilde{\varepsilon}_g := \frac{\varepsilon_g}{2\mu_\phi} \in (0, 1/2)$ and $\widetilde{\varepsilon}_H := \frac{\varepsilon_H}{4\mu_\phi}$, we are guaranteed to find a point x^{k^*} satisfying $\|\nabla\phi^*(\nabla f(x^{k^*}))\| \leq \widetilde{\varepsilon}_g < \frac{1}{2}$ and $\lambda_{\min}(H(x^{k^*})) \geq -\widetilde{\varepsilon}_H$ after at most

$$K = 1 + \left\lceil 6\frac{(f(x^0) - \inf f)}{\mu_\phi L_H} \max\left\{\left(\frac{3\sigma}{2\widetilde{\varepsilon}_g}\right)^{3/2}, \left(\frac{\sigma}{\widetilde{\varepsilon}_H}\right)^3\right\}\right\rceil. \tag{19}$$

iterations. Exploiting the specific choice of reference function, we find by properties of the convex conjugate that $\nabla\phi^*(y) = \frac{y}{\mu_\phi + \|y\|}$ for any $y \in \mathbb{R}^n$. By substituting $y = \nabla f(x^{k^*})$ and remarking that $\|\nabla\phi^*(\nabla f(x^{k^*}))\| < 1$, we find that

$$\|\nabla f(x^{k^*})\| = \frac{\mu_\phi \|\nabla\phi^*(\nabla f(x^{k^*}))\|}{1 - \|\nabla\phi^*(\nabla f(x^{k^*}))\|} \leq 2\mu_\phi \|\nabla\phi^*(\nabla f(x^{k^*}))\| \leq 2\mu_\phi \widetilde{\varepsilon}_g = \varepsilon_g.$$

On the other hand, by similarity of $H(x^k)$ to $M_k^{-1/2}\nabla^2 f(x^k) M_k^{-1/2}$ for $k \geq 0$ we have

$$\lambda_{\min}(H(x^k)) \leq \begin{cases} \lambda_{\max}(M_k^{-1})\lambda_{\min}(\nabla^2 f(x^k)) & \text{if } \lambda_{\min}(\nabla^2 f(x^k)) \geq 0 \\ \lambda_{\min}(M_k^{-1})\lambda_{\min}(\nabla^2 f(x^k)) & \text{if } \lambda_{\min}(\nabla^2 f(x^k)) < 0. \end{cases}$$

Hence,

$$[-\lambda_{\min}(\nabla^2 f(x^{k^*}))]_+ \leq \frac{1}{\lambda_{\min}(M_{k^*}^{-1})}\left[-\lambda_{\min}(H(x^{k^*}))\right]_+ \leq \frac{1}{\lambda_{\min}(M_{k^*}^{-1})}\widetilde{\varepsilon}_H$$

For isotropic reference functions, $\nabla^2\phi^*(y)$ has two distinct eigenvalues $h^{*\prime\prime}(\|y\|)$ and $\frac{h^{*\prime}(\|y\|)}{\|y\|}$ at $y \neq 0$, which for the specific reference function in this theorem reduce to

$$\lambda_1 = \frac{\mu_\phi}{(\mu_\phi + \|y\|)^2}, \qquad \lambda_2 = \frac{1}{\mu_\phi + \|y\|}.$$

At $y = 0$, we have that $\nabla^2\phi^*(0) = \mu_\phi^{-1}\,\mathrm{I}$. Thus, at $y = \nabla f(x^{k^*})$ we find from $\|\nabla f(x^{k^*})\| \leq \varepsilon_g \leq \mu_\phi$ that $\lambda_{\min}(M_k^{-1}) \geq \frac{1}{4}\mu_\phi^{-1}$, yielding

$$[-\lambda_{\min}(\nabla^2 f(x^k))]_+ \leq 4\mu_\phi \widetilde{\varepsilon}_H = \varepsilon_H.$$

The claim then follows by substituting $\widetilde{\varepsilon}_g = \frac{\varepsilon_g}{2\mu_\phi}$ and $\widetilde{\varepsilon}_H = \frac{\varepsilon_H}{4\mu_\phi}$ into (19). □

D.8 Proof of Theorem 3.10

The following lemma serves as a counterpart to Lemma 3.5, and guarantees that s^k is a descent direction.

Lemma D.3. *Let x^k, s^k be generated by Algorithm 3. If Assumption 3.3 holds, then s^k is a descent direction, i.e., $\langle \nabla f(x^k), s^k \rangle \leq 0$.*

Proof. Combining $M_k^{-1}\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k = -g^k + z^k$ and $M_k = \left[\nabla^2 \phi^*(\nabla f(x^k))\right]^{-1}$ yields

$$\left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k - M_k z^k = -M_k g(x^k) = -\frac{1}{\nu(x^k) h^{*\prime\prime}(\|\nabla f(x^k)\|)} \nabla f(x^k).$$

By positive semidefiniteness of $\nabla^2 f(x^k) + \lambda_k M_k$, by the uniform lower bound $\nu(x^k) \geq \mu_\phi > 0$ (Lemma 3.4) and by $h^{*\prime\prime}(\|\nabla f(x^k)\|) = \frac{1}{h^{\prime\prime}(h^{*\prime}(\|\nabla f(x^k)\|))} > 0$, the update directions s^k satisfy

$$\langle \nabla f(x^k), s^k \rangle = -\underbrace{\nu(x^k) h^{*\prime\prime}(\|\nabla f(x^k)\|)}_{>0} \left(\underbrace{\langle s^k, \left(\nabla^2 f(x^k) + \lambda_k M_k\right) s^k \rangle}_{\geq 0} - \underbrace{\langle M_k z^k, s^k \rangle}_{\leq 0} \right) \leq 0. \tag{20}$$

This proves the claim. □

The following lemma serves as a counterpart to Lemma 3.6 and ensures sufficient decrease for sufficiently large values of σ_k.

Lemma D.4. *Suppose that Assumptions 2.2 and 3.3 hold. If $\sigma_k \geq L_H + \theta$, then*

$$f(x^k + s^k) - f(x^k) \leq -\frac{\mu_\phi \sigma_k}{4} \|s^k\|^3.$$

Proof. The proof is similar to that of Lemma 3.6.

We apply the fundamental theorem of calculus to the objective f, and denote by ν the function defined in Lemma 3.4, i.e.,

$$f(x^k + s^k) - f(x^k) = \int_0^1 \langle \nabla f(x^k + t s^k), s^k \rangle \mathrm{d}t = \int_0^1 \nu(x^k + t s^k) \langle g(x^k + t s^k), s^k \rangle \mathrm{d}t.$$

We upper bound $\langle g(x^k + t s^k), s^k \rangle$. The fundamental theorem of calculus applied to g yields

$$g(x^k + t s^k) = g(x^k) + \int_0^t H(x^k + \tau s^k) s^k \mathrm{d}\tau = g(x^k) + t H(x^k) s^k + r(t s^k),$$

where the residual term equals

$$r(t s^k) := \int_0^t \left(H(x^k + \tau s^k) - H(x^k)\right) s^k \mathrm{d}\tau.$$

Since $H(x^k) s^k = -g(x^k) - \lambda_k s^k + z^k$ it follows that

$$g(x^k + t s^k) = (1 - t) g(x^k) - t \lambda_k s^k + r(t s^k) + t z^k.$$

Taking the inner product with s^k yields

$$\begin{aligned} \langle g(x^k + t s^k), s^k \rangle &= (1 - t) \langle g(x^k), s^k \rangle - t \lambda_k \|s^k\|^2 + \langle r(t s^k), s^k \rangle + t \langle z^k, s^k \rangle \\ &\leq -t \lambda_k \|s^k\|^2 + \langle r(t s^k), s^k \rangle + t \langle z^k, s^k \rangle, \end{aligned}$$

where the inequality follows from the fact that $1 - t \geq 0$ and $\langle g(x^k), s^k \rangle = \frac{1}{\nu(x^k)} \langle \nabla f(x^k), s^k \rangle \leq 0$ (cf. Lemmas 3.4 and D.3). Under Assumption 2.2 we can upper bound

$$\|r(t s^k)\| \leq \int_0^t \left\| \left(H(x^k + \tau s^k) - H(x^k)\right) s^k \mathrm{d}\tau \right\| \leq \int_0^t L_H \tau \|s^k\|^2 \mathrm{d}\tau = \frac{L_H t^2}{2} \|s^k\|^2.$$

In combination with $\|z^k\| \leq \frac{1}{2}\theta\|s^k\|^2$ and $\sigma_k \geq \theta + L_H \geq \theta + tL_H$ for $t \in [0,1]$ this yields

$$\begin{aligned}\langle g(x^k + ts^k), s^k\rangle &\leq -t\lambda_k\|s^k\|^2 + \langle r(ts^k), s^k\rangle + t\langle z^k, s^k\rangle \\ &\leq -t\lambda_k\|s^k\|^2 + \|r(ts^k)\|\|s^k\| + t\|z^k\|\|s^k\| \\ &\leq -t\sigma_k\|s^k\|^3 + \frac{L_H t^2}{2}\|s^k\|^3 + \frac{t\theta}{2}\|s^k\|^3 \\ &\leq -t\left(\sigma_k - \frac{1}{2}(tL_H + \theta)\right)\|s^k\|^3 \\ &\leq -t\frac{\sigma_k}{2}\|s^k\|^3.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}f(x^k + s^k) - f(x^k) &= \int_0^1 \nu(x^k + ts^k)\langle g(x^k + ts^k), s^k\rangle \mathrm{d}t \leq -\mu_\phi \int_0^1 t\frac{\sigma_k}{2}\|s^k\|^3 \mathrm{d}t \\ &\leq -\mu_\phi\frac{\sigma_k}{2}\|s^k\|^3\int_0^1 t\mathrm{d}t = -\mu_\phi\frac{\sigma_k}{4}\|s^k\|^3.\end{aligned}$$

□

Under the conditions of Lemma D.4 we have $\rho_k \geq 1$. From the update rule (9) we can deduce that σ_k remains bounded.

Lemma D.5. *Suppose that Assumptions 2.2 and 3.3 hold. Then for all iterations k of Algorithm 3,*

$$\sigma_k \leq \sigma_{\max} := \max(\sigma_0, \gamma_3(L_H + \theta)) \tag{21}$$

Proof. If $\sigma_k \geq L_H + \theta$, then by Lemma D.4 $\rho_k \geq 1$ and σ_k is not further increased. Thus the mechanism of Algorithm 3 implies the claimed bound (σ_{k-1} is increased by at most a factor γ_3, which requires $\rho_{k-1} < 1$). □

As before, we lower bound $\|s^k\|$ in terms of $\|g(x^{k+1})\|$.

Lemma D.6. *Suppose that Assumptions 2.2 and 3.3 hold. If iteration k of Algorithm 3 is successful, then*

$$\|s^k\|^2 \geq \frac{2}{2\sigma_{\max} + L_H + \theta}\|g(x^{k+1})\|.$$

Proof. The proof is similar to that of Lemma 3.7. By the fundamental theorem of calculus we have $g(x^k+s^k) = g(x^k)+H(x^k)s^k+r(s^k)$, where r is given by (17). Combined with $H(x^k)s^k+\lambda_k s^k = -g(x^k) + z^k$ and $\lambda_k = \sigma_k\|s^k\|$ this yields

$$\begin{aligned}g(x^{k+1}) = g(x^k + s^k) &= g(x^k) + H(x^k)s^k + r(s^k) \\ &= -\lambda_k s^k + r(s^k) + z^k = -\sigma_k\|s^k\|s^k + r(s^k) + z^k.\end{aligned}$$

By the triangle inequality and $\|z^k\| \leq \frac{\theta}{2}\|s^k\|^2$ we obtain

$$\|g(x^{k+1})\| \leq \sigma_k\|s^k\|^2 + \frac{L_H}{2}\|s^k\|^2 + \frac{\theta}{2}\|s^k\|^2 = \frac{2\sigma_k + L_H + \theta}{2}\|s^k\|^2.$$

The claim follows by Lemma D.5. □

Denote by $\mathcal{S}_k \subseteq \{i \in \mathbb{N} \mid 0 \leq i \leq k\}$ the set of indices corresponding to successful iterations, i.e., for which $\rho_i \geq \eta_1$. By construction, Algorithm 3 increases σ_k at unsuccessful iterations, and Lemma D.4 guarantees that for sufficiently large σ_k we obtain $\rho_k \geq 1 > \eta_1$. Thus, a successful iteration is eventually obtained. The proof of the following lemma follows a strategy similar to that of Theorem 3.8.

Lemma D.7. *Consider $x^k, s^k \in \mathbb{R}^n$ generated by Algorithm 3, and suppose that Assumptions 2.2 and 3.3 hold. Then, for any $\varepsilon_g, \varepsilon_H > 0$, Algorithm 3 obtains a point x^{k^*} such that $\|\nabla\phi^*(\nabla f(x^{k^*}))\| \le \varepsilon_g$ and $[-\lambda_{\min}(H(x^{k^*}))]_+ \le \varepsilon_H$ after at most $N_{\mathcal{S}}$ successful iterations where*

$$N_{\mathcal{S}} \le 2 + \left\lceil 8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}} \max\left\{ \left(\frac{2\sigma_{\max} + L_H + \theta}{2\varepsilon_g}\right)^{3/2}, \left(\frac{\sigma_{\max}}{\varepsilon_H}\right)^3 \right\} \right\rceil.$$

Proof. If a point x^{k^*} satisfying the requirements is found after $N_{\mathcal{S}} < 2$ successful iterations, then the claim follows trivially. Thus, without loss of generality, let $N_{\mathcal{S}} \ge 2$ – recall that Algorithm 3 will always produce a successful iteration after finitely many iterations. Throughout this proof, denote the indices corresponding to the first $N_{\mathcal{S}}$ successful iterations by $\mathcal{S} = \{i_1, \dots, i_{N_{\mathcal{S}}}\}$.

For successful iterates $i_\ell \in \mathcal{S}$, we have $\rho_{i_\ell} \ge \eta_1$ and hence by definition of ρ_{i_ℓ}

$$f(x^{i_\ell}) - f(x^{i_\ell+1}) \ge \eta_1 \frac{\mu_\phi\sigma_{\min}}{4}\|s^{i_\ell}\|^3.$$

Summing over the successful iterations, and leveraging that an unsuccessful iteration leaves the iterate unchanged, i.e., $x^{i_\ell} = x^{i_{\ell-1}+1}$ for $2 \le \ell \le N_{\mathcal{S}}$, we find

$$\frac{\eta_1\mu_\phi\sigma_{\min}}{4}\sum_{i_\ell\in\mathcal{S}}\|s^{i_\ell}\|^3 \le \sum_{i_\ell\in\mathcal{S}}(f(x^{i_\ell}) - f(x^{i_\ell+1})) = f(x^0) - f(x^{i_{N_{\mathcal{S}}}+1}) \le f(x^0) - \inf f.$$

Let $q_\ell := \max\{\|s^{i_{\ell-1}}\|, \|s^{i_\ell}\|\}$ for $\ell \ge 2$. Then we obtain

$$\sum_{\substack{i_\ell\in\mathcal{S}\\ \ell\ge 2}} q_\ell^3 \le \sum_{\substack{i_\ell\in\mathcal{S}\\ \ell\ge 2}} \left(\|s^{i_{\ell-1}}\|^3 + \|s^{i_\ell}\|^3\right) \le 2\sum_{i_\ell\in\mathcal{S}}\|s^{i_\ell}\|^3 \le 8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}}.$$

Thus, there exists an index $\ell^* \ge 2$ for which

$$q_{\ell^*}^3 \le 8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}(N_{\mathcal{S}} - 1)}.$$

Lemma D.6 ensures for $\ell \ge 2$ that

$$\|s^{i_{\ell-1}}\|^3 \ge \left(\frac{2}{2\sigma_{\max} + L_H + \theta}\right)^{3/2}\|g(x^{i_{\ell-1}+1})\|^{3/2} = \left(\frac{2}{2\sigma_{\max} + L_H + \theta}\right)^{3/2}\|g(x^{i_\ell})\|^{3/2}.$$

Moreover, Algorithm 3 ensures that $\nabla^2 f(x^{i_\ell}) + \lambda_{i_\ell} M_{i_\ell} \succeq 0$, for $\ell \ge 1$ and therefore

$$M_{i_\ell}^{-1/2}\nabla^2 f(x^{i_\ell})M_{i_\ell}^{-1/2} + \lambda_{i_\ell}\,\mathrm{I} \succeq 0.$$

Since this matrix is similar to $H(x^{i_\ell}) + \lambda_{i_\ell}\,\mathrm{I}$, we conclude that

$$\lambda_{\min}(H(x^{i_\ell})) \ge -\lambda_{i_\ell} = -\sigma_{i_\ell}\|s^{i_\ell}\| \ge -\sigma_{\max}\|s^{i_\ell}\|.$$

Let $k^* = i_{\ell^*}$, then $q_{\ell^*}^3 \ge \left(\frac{2}{2\sigma_{\max}+L_H+\theta}\right)^{3/2}\|g(x^{i_{\ell^*}})\|^{3/2}$ and $q_{\ell^*}^3 \ge \left(-\frac{\lambda_{\min}(H(x^{i_{\ell^*}}))}{\sigma_{\max}}\right)^3$ yields

$$\|g(x^{k^*})\| \le \frac{2\sigma_{\max} + L_H + \theta}{2}\left(8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}(N_{\mathcal{S}} - 1)}\right)^{2/3},$$

$$\lambda_{\min}(H(x^{k^*})) \ge -\sigma_{\max}\left(8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}(N_{\mathcal{S}} - 1)}\right)^{1/3}.$$

Clearly, both $\|g(x^{k^*})\| \le \varepsilon_g$ and $\lambda_{\min}(H(x^{k^*})) \ge -\varepsilon_H$ hold if

$$8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}(N_{\mathcal{S}} - 1)} \le \left(\frac{2\varepsilon_g}{2\sigma_{\max} + L_H + \theta}\right)^{3/2} \qquad 8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}(N_{\mathcal{S}} - 1)} \le \left(\frac{\varepsilon_H}{\sigma_{\max}}\right)^3$$

which is the case if

$$N_{\mathcal{S}} = 1 + \left\lceil 8\frac{f(x^0) - \inf f}{\eta_1\mu_\phi\sigma_{\min}} \max\left\{ \left(\frac{2\sigma_{\max} + L_H + \theta}{2\varepsilon_g}\right)^{3/2}, \left(\frac{\sigma_{\max}}{\varepsilon_H}\right)^3 \right\} \right\rceil.$$

The claimed bound holds. □

The update rule (9) and the upper bound $\sigma_k \leq \sigma_{\max}$ from Lemma D.5 imply an upper bound of the total number of iterations in terms of the number of successful iterations.

Lemma D.8 ((Cartis et al., 2022, Lemma 2.4.1))**.** *Suppose that σ_k is updated according to* (9)*, and that $\sigma_k \leq \sigma_{\max}$ for some $\sigma_{\max} > 0$. Then,*

$$k \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0}\right). \tag{22}$$

We can now prove Theorem 3.10.

Proof. This is a direct combination of Lemmas D.7 and D.8. □

E ADDITIONAL EXPERIMENTS

Large-scale matrix factorization We consider symmetric matrix factorization problems $\text{minimize}_{X \in \mathbb{R}^{n \times r}} f(X) := \frac{1}{4}\|XX^\top - A\|_F^2$ where $A \in \mathbb{R}^{n \times n}$ is large, sparse and symmetric. If $r \ll n$, then $f, \nabla f$, and Hessian-vector products can be evaluated without forming dense $n \times n$ or $nr \times nr$ matrices. We use $r = 5$ and $n = 30\ 000$, corresponding to $150\ 000$ variables. The matrix A is randomly generated from a weighted stochastic block model, and has around 3 million non-zero elements. We compare standard adaptive regularization with cubics (ARC) (Cartis et al., 2011b) against Algorithm 3 with the reference function $\phi(X) = c[-\|X\|_F - \ln(1 - \|X\|_F)]$, where $c = 10^{-2}\|\nabla f(X^0)\|_F$. Algorithm 3 uses $\theta = 1$ and $\sigma_0 = 10^{-4}$, ARC uses $\sigma_0 = 10$. These hyperparameters were obtained after a grid search, and then fixed; the subsequent results are for different random realizations. Both methods use $\eta_1 = 0.1, \eta_2 = 0.9$, multiply σ_k by a factor of 0.5 (4) after a very successful (unsuccessful) iteration. A run is called successful when $\|\nabla f(X^k)\|_F \leq 10^{-7}\|\nabla f(X^0)\|_F$ and $f(X^k) - f^\star \leq 10^{-10}(f(X^0) - f^\star)$, with $f^\star = \inf f$ computed from the largest eigenvalues of A. Both methods use the same Lanczos-Krylov based method for their subproblems, as standard in the literature (Cartis et al., 2022, §10). In the case of Algorithm 3 this method is applied to the transformed system (8).

Table 3: Large sparse factorization problems ($n = 30{,}000$, $r = 5$). Performance is reported as median [first quartile, third quartile] over 20 held-out instances.

Method	Iterations	HVPs	Time (s)	Success
ARC ($\sigma_0 = 10^1$)	22.0 [22.0, 23.0]	880 [857, 906]	20.32 [19.93, 20.74]	20/20
Algorithm 3 ($\sigma_0 = 10^{-4}$)	**17.0 [17.0, 18.0]**	**720 [706, 742]**	**16.52 [15.94, 17.01]**	20/20
ARC (matched $\sigma_0 = 10^{-4}$)	32.0 [32.0, 33.0]	1462 [1434, 1496]	35.58 [34.85, 36.47]	20/20

Table 3 summarizes the results for 20 random realizations of this symmetric factorization problem (distinct from the hyperparameter selection). Both methods solve all problems successfully. We observe that Algorithm 3 reduces the median run time by 18.7%, and similar observations hold for the number of iterations and Hessian-vector products (HVPs). Finally, as a sensitivity ablation, ARC with the same numerical initial value $\sigma_0 = 10^{-4}$ is also included. This illustrates that the two methods can have substantially different useful ranges for the regularization parameter.