Inexact Newton-type Methods for Optimisation with Nonnegativity Constraints

Oscar Smee [1] Fred Roosta [1 2]

Abstract

We consider solving large scale nonconvex optimisation problems with nonnegativity constraints. Such problems arise frequently in machine learning, such as nonnegative least-squares, nonnegative matrix factorisation, as well as problems with sparsity-inducing regularisation. In such settings, first-order methods, despite their simplicity, can be prohibitively slow on ill-conditioned problems or become trapped near saddle regions, while most second-order alternatives involve nontrivially challenging subproblems. The two-metric projection framework, initially proposed by Bertsekas (1982), alleviates these issues and achieves the best of both worlds by combining projected gradient steps at the boundary of the feasible region with Newton steps in the interior in such a way that feasibility can be maintained by simple projection onto the nonnegative orthant. We develop extensions of the two-metric projection framework, which by inexactly solving the subproblems as well as employing non-positive curvature directions, are suitable for large scale and nonconvex settings. We obtain state-of-the-art convergence rates for various classes of nonconvex problems and demonstrate competitive practical performance on a variety of problems.

1. Introduction

We consider high-dimensional problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \quad \text{subject to} \quad \mathbf{x} \geq \mathbf{0}, \tag{1}$$

where $d \gg 1$ and $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and possibly nonconvex function. Despite the simplicity of its formulation, such problems arise in many applications in science, engineering, and machine learning (ML). Typical examples in ML include nonnegative formulations of least-squares and matrix factorisation (Lee & Seung, 1999; 2000; Gillis, 2020). Additionally, problems involving sparsity inducing regularisation such as ℓ_1 norm, which are typically non-smooth, can be reformulated into a differentiable objective with nonnegativity constraints (Schmidt et al., 2007).

Many methods have been developed to solve (1). First-order methods (Lan, 2020), such as projected gradient descent, can be very simple to implement and as such are popular in ML. However, they come with well-known deficiencies, including relatively-slow convergence on ill-conditioned problems, sensitivity to hyper-parameter settings such as learning rate, and difficulty in escaping flat regions and saddle points. On the other hand, general purpose second-order algorithms, e.g., projected Newton method (Schmidt et al., 2011; Lee et al., 2014) and interior point methods (Nocedal & Wright, 2006), alleviate some of these issues such as susceptibility to ill-conditioning and/or stagnation near flat regions. However, due to not leveraging the simplicity of the constraint, this advantages come at the cost of introducing highly non-trivial and challenging subproblems.

By exploiting the structure of the constraint in (1), Bertsekas (1982) proposed the two-metric projection framework as a natural and simple adaptation of the classical Newton's method for unconstrained problems. By judicious modification of the Hessian matrix, this framework can be effectively seen as projecting Newton's step onto the nonnegative orthant. This allows for the best of both worlds, blending the efficiency of classical Newton's method with the simplicity of projected gradient descent. Indeed, similar to the classical Newton's method, the subproblem amounts to solving a linear system, while like projected gradient-descent, the projection step is straightforward.

Contribution. In this paper, we design, theoretically analyse, and empirically evaluate novel two-metric projection type algorithms (Algorithms 1 and 2) with desirable complexity guarantees for solving large scale and nonconvex optimisation problems with nonnegativity constraints (1). Both Algorithms 1 and 2 are Hessian-free in that the subproblems are solved inexactly using the minimum residual (MINRES) method (Paige & Saunders, 1975) and only re-

[1] School of Mathematics and Physics, University of Queensland, Brisbane, Australia [2] ARC Training Centre for Information Resilience, Brisbane, Australia. Correspondence to: Oscar Smee <o.smee@uq.edu.au>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

quire Hessian-vector product evaluations. To achieve approximate first-order optimality (see Definition 2.1), we leverage the theoretical properties of MINRES, as recently established in (Liu & Roosta, 2022a), e.g., nonnegative curvature detection and monotonicity properties, and we show the following:

(I) Under minimal assumptions, Algorithm 1 achieves global iteration complexity that matches those of first-order alternatives (Theorem 3.3).
(II) Under stronger assumptions, Algorithm 2 enjoys a global iteration complexity guarantee with an improved rate that matches the state of the art for second-order methods (Theorem 3.8).
(III) Both variants obtain competitive oracle complexities, i.e., the total number of gradient and Hessian-vector product evaluations (Corollaries D.2 and D.3).
(IV) Our approach enjoys fast local convergence guarantees (Theorem 3.13 and Corollary 3.14).
(V) Our approach exhibit highly competitive empirical performance on several machine learning problems (Section 4).

To our knowledge, the complexity guarantees outlined in this paper are the first to be established for two-metric projection type algorithms in nonconvex settings.

Notation. Vectors and matrices are denoted, respectively, by bold lowercase and uppercase letters. Denote the nonnegative orthant by $\mathbb{R}^d_+$. The open ball of radius r around $\mathbf{x}$ is denoted by $B(\mathbf{x}, r) \triangleq \{\mathbf{z} \in \mathbb{R}^d \mid \|\mathbf{z} - \mathbf{x}\| < r\}$. The inequalities,"$\geq$" and "$\leq$", are often applied elementwise. Big-$\mathcal{O}$ complexity is denoted by $\mathcal{O}$ with hidden logarithmic factors indicated by $\tilde{\mathcal{O}}$. Denote components of vectors by superscript and iteration counters as subscripts, e.g., $\mathbf{x}_k^i$ is i^{th} component of the k^{th} iterate of $\mathbf{x}$. As a natural extension, a set of indices in the superscript denotes the subvector corresponding to those components, e.g., letting $[d] = \{1, \ldots, d\}$, if $\mathcal{I} \subseteq [d]$ and $\mathbf{v} \in \mathbb{R}^d$ then $\mathbf{v}^{\mathcal{I}} = (\mathbf{v}^i \mid i \in \mathcal{I}) \in \mathbb{R}^{|\mathcal{I}|}$. Let $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ and $\mathbf{H}(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ denote the gradient and Hessian of f, respectively. Denote the δ_k-active and δ_k-inactive sets, respectively, by

$$\mathcal{A}(\mathbf{x}_k, \delta_k) = \{i \in [d] \mid 0 \leq \mathbf{x}_k^i \leq \delta_k\}, \tag{2a}$$

$$\mathcal{I}(\mathbf{x}_k, \delta_k) = \{i \in [d] \mid \mathbf{x}_k^i > \delta_k\}. \tag{2b}$$

When the context is clear, we suppress the dependence on $\mathbf{x}_k$ and δ_k, e.g., $\mathbf{g}_k$ and $\mathbf{H}_k$ for $\mathbf{g}(\mathbf{x}_k)$ and $\mathbf{H}(\mathbf{x}_k)$ and $\mathbf{x}_k^{\mathcal{I}}$ or $\mathbf{x}_k^{\mathcal{I}_k}$ instead of $\mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k, \delta_k)}$. We also denote $\mathbf{H}_k^{\mathcal{I}} = \{(\mathbf{H}_k)_{ij} \mid i, j \in \mathcal{I}(\mathbf{x}_k, \delta_k)\}$.

2. Background and Related Work

We now briefly review related works for solving (1) and some essential background necessary for our presentation.

First-order Methods. The projected gradient method (Lan, 2020) is among the simplest techniques for solving optimisation problems involving convex constraints. Indeed, the projected gradient iteration for minimisation over a convex set Ω is simply given by $\mathbf{x}_{k+1} = \mathcal{P}_\Omega(\mathbf{x}_k - \alpha_k \mathbf{g}_k)$ where $\mathcal{P}_\Omega : \mathbb{R}^d \to \mathbb{R}^d$ is the orthogonal projection onto Ω defined by $\mathcal{P}_\Omega(\mathbf{x}) = \arg\min_{\mathbf{z} \in \Omega} \|\mathbf{z} - \mathbf{x}\|$. When α_k is chosen appropriately, e.g., via line search, the projected gradient method is known to converge under essentially the same conditions and at the same rate as the unconstrained variant (Bertsekas, 1999; Beck, 2017). Many variations of this method have also been considered, e.g., spectral projected gradient (Birgin et al., 2014), proximal gradient (Parikh & Boyd, 2014; Beck, 2017), and accelerated proximal gradient (Nesterov, 2013; Beck & Teboulle, 2009) with its extensions to non-convex settings (Lin et al., 2020; Li et al., 2017).

Of course, the effectiveness of the projected gradient method relies heavily on the computational cost associated with computing $\mathcal{P}_\Omega(\mathbf{x})$. While this can be challenging for general convex sets, in the case of $\Omega = \mathbb{R}^d_+$, it is simply given by $[\mathcal{P}(\mathbf{x})]^i = \mathbf{x}^i$, if $\mathbf{x}^i > 0$, and $[\mathcal{P}(\mathbf{x})]^i = 0$, otherwise. Note that, for notational simplicity, we omit the dependence of $\mathcal{P}$ on Ω in our context. Nonetheless, while the projected gradient method is a simple choice for solving (1), it shares the common drawbacks of first-order methods alluded to earlier, e.g., susceptibility to ill-conditioning.

Second-order Methods. By incorporating Hessian information, second-order methods hold the promise to alleviate many of the well-known deficiencies of first-order alternatives, e.g., they are typically better suited to ill-conditioned problems (Xu et al., 2020b). For constrained problems, generic projected (quasi) Newton methods involve iterations of the form $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ where

$$\mathbf{p}_k = \arg\min_{\mathbf{x} \in \Omega} \langle \mathbf{g}_k, \mathbf{p} \rangle + \langle \mathbf{p}, \mathbf{B}_k \mathbf{p} \rangle / 2, \tag{3}$$

where α_k is an appropriately chosen step-size, e.g., backtracking line search, and $\mathbf{B}_k$ captures some curvature information of f at $\mathbf{x}_k$ (and also potentially the step-length as in the proximal arc search). For $\mathbf{B}_k = \mathbf{I}$ we recover a projected gradient variant, whereas for $\mathbf{B}_k = \mathbf{H}_k$, or some approximation, we obtain projected (or more generally proximal) Newton-type methods (Schmidt et al., 2009; 2011; Becker & Fadili, 2012; Lee et al., 2014; Shi & Liu, 2015). The main drawback of this framework is that the subproblem, (3), may no longer be a simple projection even when Ω is a simple, and one has to resort to an optimisation subroutine to (approximately) solve (3).

An alternative is the interior point framework (Nocedal & Wright, 2006), where the constraints are directly integrated into the objective as "barrier" functions. While the subproblems in this framework amount to solving linear systems, to produce accurate solutions the barrier function must approach the constraint, which can lead to highly ill conditioned subproblems. Some recent works (Bian et al., 2014; Haeser et al., 2017; O'Neill & Wright, 2020) consider interior point methods for (1). In particular, in (O'Neill & Wright, 2020), capped Newton-CG with a preconditioned Hessian is used to optimise a log barrier augmented objective. Due to issues arising from increasingly ill-conditioned subproblems, the practical efficacy of this method seems to be inferior when compared to projection-based methods, including those of first-order (Xie & Wright, 2023).

The issue with the general purpose second-order methods discussed so far is that, unlike projected gradient, they do not leverage the simplicity of the nonnegativity constraints and the corresponding projection. In this light, a naïve adaptation of the projected gradient would imply directly projecting the Newton step on the constraints, e.g., $\mathbf{x}_{k+1} = \mathcal{P}_\Omega(\mathbf{x}_k - \alpha_k \mathbf{H}_k^{-1}\mathbf{g}_k)$. Unfortunately, such a direct adaptation may lead to ascent directions for the objective function at the boundary. To that end, the *two-metric projection (TMP) framework* (Bertsekas, 1982; Gafni & Bertsekas, 1984) offers an ingenious solution. Specifically, at each iteration, the component indices, $[d]$, are divided into the approximately bound, $\mathcal{J}_k^+$, and free sets, $\mathcal{J}_k^-$, given by

$$\mathcal{J}_k^+ = \{i \in [d] \mid \mathbf{x}_k^i \le \delta, \mathbf{g}_k^i > 0\},\ \mathcal{J}_k^- = [d] \setminus \mathcal{J}_k^+. \quad (4)$$

where $\delta > 0$. A matrix, $\mathbf{D}_k$, is then chosen to be "diagonal" with respect to set $\mathcal{J}_k^+$, that is,

$$(\mathbf{D}_k)_{ij} = 0,\ \ i \in \mathcal{J}_k^+,\ \ j \in [d] \setminus \{i\},$$

and the update is simply given by

$$\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k - \alpha_k \mathbf{D}_k \mathbf{g}_k). \quad (5)$$

It has been shown that TMP is asymptotically convergent under certain conditions and reasonable choices of $\mathbf{D}_k$. For example, for strongly convex problems, the non-diagonal portion of $\mathbf{D}_k$ can consist of the inverse of the Hessian submatrix corresponding to the indices in $\mathcal{J}_k^-$. In this case, (5) reduces to a scaled gradient in $\mathcal{J}_k^+$ and a Newton step in $\mathcal{J}_k^-$. Bertsekas (1982) also shows that, under certain conditions, TMP can preserve fast "Newton like" local convergence. Practically, TMP type algorithms has been successfully applied to a range of problems (Gafni & Bertsekas, 1984; Schmidt et al., 2007; Kim et al., 2010; Haber, 2014; Kuang et al., 2015; Cai et al., 2023). In large scale and nonconvex settings, employing the Newton step as part of (5) may be infeasible or even undesirable. Indeed, not only can Hessian storage and inversion costs be prohibitive, the existence of negative curvature can lead to ascent directions.

With a view to eliminate the necessity of forming and inverting the Hessian, Kim et al. (2010) extend TMP to utilise a quasi-Newton update with asymptotic convergence guarantees in the convex setting. Also in this vein, Xie & Wright (2023) considered "projected Newton-CG", which entails a combination of the projected gradient and the inexact Newton steps that preserve the simplicity of projection onto $\mathbb{R}_+^d$. In particular, Newton-CG steps are based on the capped CG procedure of Royer et al. (2018). Unfortunately, the gradient and Newton-CG steps are not taken simultaneously. Instead, the algorithm employs projected gradient steps across all components until optimality is attained in the approximately active set. Only at that point is the Newton-CG step applied in the approximately inactive set. This implies that the algorithm may take projected gradient steps at most iterations, potentially impeding its practical performance.

Hessian-free Inexact Methods. In high-dimensional settings, storing the Hessian matrix may be impractical. Moreover, an approximate direction can often be computed at a fraction of the cost of a full Newton step. In this context, Hessian-free inexact Newton-type algorithms leverage Krylov subspace methods (Saad, 2003), which are particularly well-suited for these scenarios. Krylov subspace solvers can recover a reasonable approximate direction in just a few iterations and only require access to the Hessian-vector product mapping, $\mathbf{v} \mapsto \mathbf{H}(\mathbf{x})\mathbf{v}$. The computational cost of a Hessian-vector product is comparable to that of a gradient evaluation and does not require the explicit formation of $\mathbf{H}$. Indeed, $\mathbf{H}(\mathbf{x})\mathbf{v}$ can be computed by obtaining the gradient of the map $\mathbf{x} \mapsto \langle \mathbf{g}(\mathbf{x}), \mathbf{v} \rangle$ using automatic differentiation, leading to one additional back propagation compared to computing $\mathbf{g}(\mathbf{x})$.

Complexity in Optimisation. Recently, there has been a growing interest in obtaining global worst case *iteration complexity* guarantees for optimisation methods, namely a bound on the number of iterates required for the algorithm to compute an approximate solution. For instance, in unconstrained and nonconvex settings, gradient descent produces an approximate first-order optimal point satisfying $\|\mathbf{g}(\mathbf{x})\| \le \epsilon_g$ in at most $\mathcal{O}(\epsilon_g^{-2})$ iterations for objectives with Lipschitz continuous gradients (Nesterov, 2004). This rate has been shown to be tight (Cartis et al., 2010). Without additional assumptions, similar rates have also been shown for second-order methods (Cartis et al., 2022). However, for objectives with both Lipschitz continuous gradient and Hessian, this rate can be improved to $\mathcal{O}(\epsilon_g^{-3/2})$, which is also shown to be tight over a wide class of second-order algorithms (Cartis et al., 2011b). Second-order methods which achieve this rate include cubic regularised Newton's method and its adaptive variants (Nesterov & Polyak, 2006; Cartis et al., 2011c;a; Xu et al., 2020a), modified trust region based methods (Curtis et al., 2016; 2021; Curtis & Wang, 2023) and line search methods including Newton-CG (Royer et al., 2018) and Newton-MR (Liu & Roosta, 2022b) as well as their inexact variants (Yao et al., 2022; Lim & Roosta, 2023). Many of the above works also provide explicit bounds on the *operational complexity*, that is, a bound on the number of fundamental computational units (e.g. gradient evaluations, Hessian vector products) to obtain an approximate solution.

In the constrained setting, direct comparison between bounds is difficult due to differences in approximate optimality conditions; see discussion in Xie & Wright (2023, Section 3) for the bound constraint case. However, the algorithms in Cartis et al. (2020); Birgin & Martínez (2018) achieve $\mathcal{O}(\epsilon_g^{-3/2})$ for a first-order optimal point with certain types of constraints, which is shown to be tight in Cartis et al. (2020). More specific to the bound constraint case, the Newton-CG log barrier method of O'Neill & Wright (2020) achieves a complexity of $\mathcal{O}(d\epsilon_g^{-1/2} + \epsilon_g^{-3/2})$, while the projected Newton-CG algorithm of Xie & Wright (2023) obtains a rate of $\mathcal{O}(\epsilon_g^{-3/2})$ under a set of approximate optimality conditions similar to this work.

Optimality Conditions. Recall that $\mathbf{x}_*$ satisfies the first-order necessary conditions for (1) if

$$\mathbf{x}_* \geq \mathbf{0}, \quad \text{and} \quad \begin{cases} [\nabla f(\mathbf{x}_*)]^i = 0, & \text{if } \mathbf{x}_*^i > 0, \\ [\nabla f(\mathbf{x}_*)]^i \geq 0, & \text{if } \mathbf{x}_*^i = 0. \end{cases} \tag{6}$$

We seek a point which satisfies these conditions to some "ϵ" tolerance. There are a number of ways to adapt (6) into an approximate condition (Xie & Wright, 2023, Section 3). In this work we adopt Xie & Wright (2023, Definition 1).

Definition 2.1 (ϵ-Optimal Point). A point, $\mathbf{x}$, is called ϵ-approximate first-order optimal (ϵ-FO) if

$$\mathbf{g}^i \geq -\sqrt{\epsilon}, \quad \forall i \in \mathcal{A}(\mathbf{x}, \sqrt{\epsilon}) \tag{7a}$$

$$\|\text{diag}(\mathbf{x}^{\mathcal{A}})\mathbf{g}^{\mathcal{A}}\| \leq \epsilon, \tag{7b}$$

$$\|\mathbf{g}^{\mathcal{I}}\| \leq \epsilon. \tag{7c}$$

We take (7a) and (7b) to be trivially satisfied if $\mathcal{A}(\mathbf{x}, \sqrt{\epsilon}) = \emptyset$ and similar for (7c) if $\mathcal{I}(\mathbf{x}, \sqrt{\epsilon}) = \emptyset$.

This definition has been shown to be asymptotically exact.

Lemma 2.2. *(Xie & Wright, 2023, Lemma 1) Suppose that $\epsilon_k \downarrow 0$ and we have a sequence $\{\mathbf{x}_k\}_{k=1}^{\infty}$ where each $\mathbf{x}_k$ satisfies the corresponding ϵ_k-FO optimality condition. If $\mathbf{x}_k \to \mathbf{x}_*$ then $\mathbf{x}_*$ satisfies (6).*

3. Newton-MR Two-Metric Projection

We now propose and theoretically study our extensions of the TMP framework, which involves simultaneously employing gradient and inexact Newton steps, which are, respectively, restricted to the active and inactive sets.

3.1. MINRES and Its Properties

The inexact Newton step is based on the recently proposed Newton-MR framework (Liu & Roosta, 2022b; Roosta et al., 2022), where instead of CG, subproblems are approximately solved using the minimum residual (MINRES) method (Paige & Saunders, 1975). Recall that the t^{th} iteration of MINRES is formulated as

$$\mathbf{s}^{(t)} = \underset{\mathbf{s} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})}{\arg\min} \|\mathbf{H}\mathbf{s} + \mathbf{g}\|^2. \tag{8}$$

where $\mathcal{K}_t(\mathbf{H}, \mathbf{g}) = \text{Span}\{\mathbf{g}, \mathbf{H}\mathbf{g}, \ldots, \mathbf{H}^{t-1}\mathbf{g}\}$ is the Krylov subspace of degree t generated from $\mathbf{H}$ and $\mathbf{g}$. On each iteration MINRES minimises the squared norm of the residual of the Newton system over the corresponding Krylov subspace. Note that, from an optimisation perspective, the residual itself can be viewed as the *gradient* of the second-order Taylor approximation typically considered by second-order methods (e.g. Newton-CG), that is, $\mathbf{r} \triangleq -\mathbf{H}\mathbf{s} - \mathbf{g} = -\nabla_{\mathbf{s}}(\langle \mathbf{g}, \mathbf{s} \rangle + \frac{1}{2}\langle \mathbf{s}, \mathbf{H}\mathbf{s} \rangle)$. This highlights an advantage of MINRES over CG. Indeed, unlike CG, which aims to minimise the second order Taylor approximation, minimisation of the residual norm remains well defined even if $\mathbf{H}$ is indefinite. For more theoretical and empirical comparisons between CG and MINRES, see Lim et al. (2024).

Recently, Liu & Roosta (2022a) established several properties of MINRES that make it particularly well-suited for nonconvex settings. For example, to assess the availability of a nonpositive curvature (NPC) direction in MINRES, one merely needs to monitor the condition

$$\langle \mathbf{r}^{(t-1)}, \mathbf{H}\mathbf{r}^{(t-1)} \rangle \leq 0, \tag{9}$$

This condition is shown to be both necessary and sufficient for the existence of NPC directions in $\mathcal{K}_t(\mathbf{H}, \mathbf{g})$ (Liu & Roosta, 2022a, Theorem 3.3). In addition, MINRES enjoys a natural termination condition in non-convex settings. More specifically, for any user specified tolerance $\eta > 0$, the termination condition

$$\|\mathbf{H}\mathbf{r}^{(t-1)}\| \leq \eta \|\mathbf{H}\mathbf{s}^{(t-1)}\|, \tag{10}$$

is satisfied at some iteration. Note that the left hand side, $\mathbf{H}\mathbf{r}^{(t-1)}$, is simply the residual of the normal equation $\mathbf{H}^2\mathbf{s} = -\mathbf{H}\mathbf{g}$. Condition (10) is particularly appealing in non-convex settings where we might have $\mathbf{g} \notin \text{Range}(\mathbf{H})$ and therefore $\|\mathbf{r}\| > 0$ for all $\mathbf{s} \in \mathbb{R}^d$. In this case a more typical termination condition $\|\mathbf{r}^{(t-1)}\| \leq \eta$ may never be satisfied for a given $\eta > 0$. By contrast, (10) is applicable in all situations since $\|\mathbf{H}\mathbf{r}^{(t-1)}\|$ is guaranteed to monotonically decrease to zero, while $\|\mathbf{H}\mathbf{s}^{(t-1)}\|$ is monotonically increasing (Liu & Roosta, 2021, Lemma 3.11). Remarkably, both Conditions (9) and (10) can be computed with a scalar update directly from the MINRES iterates without any additional Hessian-vector products; see Lemma A.1.

A Newton-MR step is computed by running MINRES until (9) is detected, in which case $\mathbf{r}^{(t-1)}$ is returned. Since $\mathbf{r}^{(t-1)}$ is a nonpositive curvature direction, we label this case as a "NPC" step. Otherwise, when the termination condition (10) is satisfied, $\mathbf{s}^{(t-1)}$ is returned. This step serves as an approximate solution to (8) and so we label this case as a "SOL" step. Let $\mathbf{p}$ denote the direction returned by negative curvature detecting MINRES. Liu & Roosta (2022a) shows that $\mathbf{p}$ serves as a direction of first and second-order descent for the function f, namely $\langle \mathbf{p}, \mathbf{g} \rangle < 0$ and $\langle \mathbf{p}, \mathbf{g} \rangle + \langle \mathbf{p}, \mathbf{Hp} \rangle / 2 < 0$ (Liu & Roosta, 2022a, Theorem 3.8), as well as a direction of non-ascent for the norm of its gradient $\|\mathbf{g}\|^2$, that is, $\langle \mathbf{p}, \mathbf{Hg} \rangle < 0$ for a SOL step and $\langle \mathbf{p}, \mathbf{Hg} \rangle = 0$ for a NPC step (Liu & Roosta, 2022a, Lemma 3.1).

We include the full MINRES algorithm (Algorithm 3) as well as some additional properties in Appendix A.

3.2. Global Convergence: Minimal Assumptions

We first present a variant that is globally convergent under minimal assumptions. Algorithm 1 is our simplest variant of the Newton-MR two-metric projection method. Recalling the definition of the δ_k-active and δ_k-inactive sets as in (2)[1], Algorithm 1 combines an active set gradient step (i.e., $\mathbf{p}_k^{\mathcal{A}} = -\mathbf{g}_k^{\mathcal{A}}$) with an inactive set Newton-MR step (i.e., $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$, if $\text{D}_{\text{type}} = \text{SOL}$, and $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$, if $\text{D}_{\text{type}} = \text{NPC}$). In Algorithm 1, the curvature condition (9) is considered with a positive tolerance, $\overline{\varsigma} = (d+1)\varsigma > 0$, i.e., $\langle \mathbf{r}^{(t-1)}, \mathbf{Hr}^{(t-1)} \rangle \leq \overline{\varsigma} \|\mathbf{r}^{(t-1)}\|^2$. Lemma B.1 demonstrates that $\langle \mathbf{r}^{(i)}, \mathbf{Hr}^{(i)} \rangle > \overline{\varsigma} \|\mathbf{r}^{(i)}\|^2$ for all $0 \leq i \leq t-1$ is a certificate that $\mathbf{H}$ is ς-strongly positive definite over $\mathcal{K}_t(\mathbf{H}, \mathbf{g})$.

Once the step direction is computed, the step size is selected with a line search criteria similar to that of Bertsekas (1982). Specifically, letting $\mathbf{x}_k(\alpha) = \mathcal{P}(\mathbf{x}_k + \alpha \mathbf{p}_k)$, we find α that, for some $\rho \in (0, 1/2)$, satisfies

$$\begin{aligned} f(\mathbf{x}_k(\alpha)) - f(\mathbf{x}_k) \leq \rho \left\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha \mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \right\rangle \\ + \alpha \rho \left\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \right\rangle, \end{aligned} \tag{11}$$

Note that the term corresponding to the inexact set in (11) is negative due to the descent properties of $\mathbf{p}_k^{\mathcal{I}}$ discussed earlier. On the other hand, the active set term in (11) is negative due the descent properties of the gradient mapping (Bertsekas, 1999, Proposition 3.3.1). This is crucial for our analysis as it allows us to consider the decrease in the inactive and active sets independently of each other. The two terms are unified since

$$\begin{aligned} \langle \mathbf{g}_k, \mathcal{P}(\mathbf{x}_k + \alpha \mathbf{p}_k) - \mathbf{x}_k \rangle = \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha \mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle \\ + \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle, \end{aligned}$$

so long as α is chosen small enough. This is a direct consequence of $\mathcal{I}(\mathbf{x}_k, \delta_k)$ containing only *strictly* feasible indices.

Algorithm 1 Newton-MR TMP (Minimal Assumptions)

```
Input Initial point $\mathbf{x}_0 \geq \mathbf{0}$, active set tol $\{\delta_k\}$, optimality tol $\{\epsilon_k\}$, MINRES inexactness tol $\eta > 0$, NPC tol $\overline{\varsigma} = (d+1)\varsigma$ for $\varsigma > 0$, Line search parameter $\rho < 1/2$.

for $k = 0, 1, \ldots$ do
    Update sets $\mathcal{A}(\mathbf{x}_k, \delta_k)$ and $\mathcal{I}(\mathbf{x}_k, \delta_k)$ as in (2).
    if (7) is satisfied then
        Terminate.
    end if
    $\mathbf{p}_k : \begin{cases} \mathbf{p}_k^{\mathcal{A}} \leftarrow -\mathbf{g}_k^{\mathcal{A}}, \\ (\mathbf{p}_k^{\mathcal{I}}, \text{D}_{\text{type}}) \leftarrow \text{MINRES}(\mathbf{H}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}}, \eta, \overline{\varsigma}) \end{cases}$
    if $\text{D}_{\text{type}} = \text{SOL}$ then
        $\alpha_k \leftarrow$ Algorithm 5 with $\alpha_0 = 1$ and (11).
    else if $\text{D}_{\text{type}} = \text{NPC}$ then
        $\alpha_k \leftarrow$ Algorithm 6 with $\alpha_0 = 1$ and (11).
    end if
    $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$
end for
```

In Liu & Roosta (2022b), it was shown that when MINRES algorithm the returns an NPC step, the line search for α could run in a forward tracking mode (cf. Algorithm 6). In numerical experiments, it was demonstrated that the forward tracking line search was beneficial because it allowed for very large steps to be taken, particularly in flat regions where progress would otherwise be slow. Our theoretical analysis in Appendix B demonstrates that a forward tracking line search can also be used in Algorithm 1 for NPC type steps.

To analyse the global complexity of Algorithm 1, we only require typical assumptions on Lipschitz continuity of the gradient and lower-boundedness of the objective.

Assumption 3.1. There exists $0 \leq L_g < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^d$, $\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq L_g \|\mathbf{x} - \mathbf{y}\|$.

Assumption 3.2. We have $-\infty < f_* \leq f(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}_+^d$.

With these minimal assumptions we can provide a guarantee of convergence of Algorithm 1 in Theorem 3.3, the proof of which we deferred to Appendix B.

Theorem 3.3 (Global Complexity of Algorithm 1)**.** *Let $\epsilon_g \in (0,1)$ and $\varsigma > 0$. Under Assumptions 3.1 and 3.2, if we choose $\delta_k = \epsilon_k = \epsilon_g^{1/2}$ and $\overline{\varsigma} = (d+1)\varsigma$, Algorithm 1 produces an ϵ_g-FO point in at most $\mathcal{O}(\epsilon_g^{-2})$ iterations.*

Remark 3.4. The "big-$\mathcal{O}$" rate obtained in Theorem 3.3 hides a dependence on the problem constants and algorithm parameters ρ, ς, L_g, η, which are, in particular, independent of d. However, the proof of Theorem 3.3 (and, indeed, Theorem 3.8) implies that the worst case constant hidden by the

[1] Note that (2) differs from (4) as it does not include a gradient positivity condition. This helps with tractability of the global analysis but leads to a relatively smaller inactive set.

big-$\mathcal{O}$ notation could have an unfortunate dependence on the problem constants (e.g., L_g^3). This could suggest poor *practical* performance despite the desirable dependence on ϵ_g. However, as we show numerically in Section 4, such worst case analyses are rarely indicative of typical performance in practice.

3.3. Global Convergence: Improved Rate

It is possible to modify Algorithm 1 to improve upon the convergence rate of Theorem 3.3, albeit under stronger assumptions. This is done in Algorithm 2 where, by appropriate use of curvature information, we can obtain an improved complexity rate. Algorithm 2 shares the same inactive/active sets, line search strategies, and projection based feasibility with Algorithm 1.There are, however, some main differences. A key distinction lies in the certification of *strictly* positive curvature (9) rather than strongly positive curvature, i.e., unlike Algorithm 1 where we set $\bar{\varsigma} > 0$, in Algorithm 2 we set the NPC tolerance to $\bar{\varsigma} = 0$. Another notable difference is the introduction of *Type II* steps. *Type II* steps set the active portion of the step to zero and occur when the active set optimality conditions (7a) and (7b) are satisfied (otherwise *Type I* steps, i.e., steps similar to Algorithm 1, are used) but the inactive set tolerance (7c) is unsatisfied. Because the active set termination conditions are satisfied, removing the active portion of the step is not expected to significantly impede the algorithm's progress. By the same token, we can analyse *Type II* steps using second-order curvature information, similar to the unconstrained Newton-MR algorithm, without having to account for the curvature related to the projected gradient portion of the step. Additionally, to achieve an improved rate over Algorithm 1, MINRES inexactness tolerance must scale with ϵ_k in Algorithm 2.

Algorithm 2 Newton-MR TMP (Improved Rate)

```
Input Initial point $\mathbf{x}_0 \geq \mathbf{0}$, active set tol $\{\delta_k\}$, optimality tol $\{\epsilon_k\}$, MINRES inexactness tol $\eta = \epsilon_k \theta$ and $\theta > 0$, Line search parameter $\rho < 1/2$, NPC tol $\bar{\varsigma} = 0$.
for $k = 0, 1, \dots$ do
    Update sets $\mathcal{A}(\mathbf{x}_k, \delta_k)$ and $\mathcal{I}(\mathbf{x}_k, \delta_k)$ as in (2).
    if $\mathcal{A}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and (not (7a) or not (7b)) then
        Flag = Type I.
    else if $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and not (7c) then
        Flag = Type II.
    else
        Terminate.
    end if
    $\mathbf{p}_k : \begin{cases} \mathbf{p}_k^{\mathcal{A}} \leftarrow \begin{cases} -\mathbf{g}_k^{\mathcal{A}}, & \text{If Flag = Type I,} \\ \mathbf{0}, & \text{If Flag = Type II,} \end{cases} \\ (\mathbf{p}_k^{\mathcal{I}},\ \text{D}_{\text{type}}) \leftarrow \text{MINRES}(\mathbf{H}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}}, \eta, \bar{\varsigma}) \end{cases}$
    if $\text{D}_{\text{type}} = \text{SOL}$ then
        $\alpha_k \leftarrow$ Algorithm 5 with $\alpha_0 = 1$ and (11).
    else if $\text{D}_{\text{type}} = \text{NPC}$ then
        $\alpha_k \leftarrow$ Algorithm 6 with $\alpha_0 = 1$ and (11).
    end if
    $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$
end for
```

For our analysis, we need additional assumptions including the Lipschitz continuity of the Hessian.

Assumption 3.5. There exists $0 \leq L_H < \infty$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^d$, $\|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\| \leq L_H \|\mathbf{x} - \mathbf{y}\|$.

Additionally, we make some regularity assumptions on the output of the MINRES iterations.

Assumption 3.6. There exists a contant $\omega > 0$, independent of $\mathbf{x}$, such that the NPC direction from MINRES, $\mathbf{p} = \mathbf{r}^{(t-1)}$, satifies $\|\mathbf{r}^{(t-1)}\| \geq \omega \|\mathbf{g}\|$.

We note that a lower bound for the relative residual is available directly *prior to termination*. In fact, recall that if an NPC direction is returned, the termination condition (10) must not yet be satisfied. In this case, Assumption 3.1 and Lemma A.1 together imply that $\|\mathbf{r}^{(t-1)}\| \geq \eta \|\mathbf{g}\| / \sqrt{\eta^2 + L_g^2}$. For Algorithm 1, this lower bound is directly utilised to establish convergence with no requirement for Assumption 3.6. However, for Algorithm 2, η depends on ϵ_k, which could lead us to believe that the lower bound on the relative residual prior to termination does too. In particular, at first glance, this might suggest that the smaller the inexactness tolerance η, the more iterations MINRES is expected to perform before NPC detection. We argue that this is not the case. Firstly, an upper bound on the number of MINRES iterations until a NPC direction is encountered is *independent* of the termination criteria η Liu & Roosta (2022b, Corollary 2). In fact, by construction, the MINRES iterates are independent of the termination tolerance η and the magnitude of $\|\mathbf{g}\|$; see discussion and numerical examples around Liu & Roosta (2022b, Assumption 4). Additionally, in the case where $\mathbf{g} \notin \text{Range}(\mathbf{H})$, we always have $\|\mathbf{r}^{(t-1)}\| \geq \|(\mathbf{I} - \mathbf{H}\mathbf{H}^\dagger)\mathbf{g}\|$, which is clearly independent of η. Together, these lines of argumentation constitute our justification for Assumption 3.6.

Recall that Algorithm 1 includes a manual verification of user specified strongly positive curvature over $\mathcal{K}_t(\mathbf{H}, \mathbf{g})$ in $\text{D}_{\text{type}} = \text{SOL}$ case, while Algorithm 2 only certifies strict positive curvature through the NPC condition (9). Liu & Roosta (2022a) demonstrated that as long as the NPC condition (9) has not been detected, we have $\mathbf{T}_t \succ \mathbf{0}$ where $\mathbf{T}_t \in \mathbb{R}^{t \times t}$ is the symmetric tridiagonal matrix obtained in the t^{th} iteration of MINRES (see Appendix A for more details). Our next assumption strengthens this notion.

Assumption 3.7. There exists $\sigma > 0$ such that for any $\mathbf{x}_k$ in the sequence of SOL type iterates returned by Algorithm 2, we have $\mathbf{T}_t \succeq \sigma\mathbf{I}$.

Assumption 3.7 implies that, as long as the NPC condition (9) has not been detected, for any $\mathbf{v} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$ we have $\langle \mathbf{v}, \mathbf{H}\mathbf{v} \rangle \geq \sigma\|\mathbf{v}\|^2$. Assumption 3.7 is satisfied by an objective function whose Hessian contains positive $\mathbf{g}$-relevant eigenvalues (eigenspaces not orthogonal to the gradient) uniformly separated from zero. A simple example is an under-determined least-squares problem.

Together, Assumptions 3.6 and 3.7 allow us to control the curvature of our step, which is necessary to obtain an improved rate over Algorithm 1 using a Lipschitz Hessian upper bound. We now present the convergence result for Algorithm 2. We defer the proof to Appendix C.

Theorem 3.8 (Global Complexity of Algorithm 2)**.** *Let $\epsilon_g \in (0, 1)$. Under Assumptions 3.1, 3.2 and 3.5 to 3.7, if we choose $\delta_k = \epsilon_k = \epsilon_g^{1/2}$, Algorithm 2 produces an ϵ_g-FO point in at most $\mathcal{O}(\epsilon_g^{-3/2})$ iterations.*

Remark 3.9. A direct corollary to Theorems 3.3 and 3.8, under some mild additional assumptions, is a bound on the operational complexity in terms of gradient and Hessian-vector product evaluations. In particular, to produce a ϵ_g-FO point, the operation complexity for Algorithms 1 and 2 is, respectively, $\mathcal{O}(\epsilon_g^{-2})$ and $\tilde{\mathcal{O}}(\epsilon_g^{-3/2})$; see Appendix D.

Remark 3.10. In all our algorithms, each step includes the Newton-MR component. The integration of the gradient and Newton-MR step is feasible in our algorithm due to the properties of the MINRES iterates (Lemmas A.2 and A.3), allowing for a more flexible analysis with only first-order information. In contrast, it appears that second-order information is crucial for achieving descent with the capped-CG procedure, a central aspect of Xie & Wright (2023). This constraint prevents the algorithm from taking a step simultaneously comprised of gradient and Newton-CG components.

3.4. Local Convergence

An advantage of the original TMP method of Bertsekas (1982) is that we get fast local convergence, a property that is shared by many Newton-type methods. We now show that our algorithm, in a slightly modified form, also exhibits this property. The basis for the local convergence is the fact that, under certain conditions, projected gradient algorithms are capable of identifying the true set of active constraints in a *finite* number of iterations. This result was first establish for projected gradient with bound constraints in Bertsekas (1976) but has been extended to a variety of constraints (Burke & Moré, 1988; Burke, 1990; Wright, 1993; Sun et al., 2019). In the case of two-metric projection, once the active set is identified, the combined step reduces to an unconstrained Newton step in the inactive set.

For the analysis, we consider a "local phase" variant of Algorithm 1. Specifically, we maintain flexibility in defining the outer and inner termination conditions and tolerances, eliminate the strongly positive curvature validation, and only perform backtracking line search from $\alpha_0 = 1$ to ensure the step length remains bounded. The pseudo-code for this local phase version is given in Algorithm 4 for completeness. To show that the active set is identified in finite number of iterations, we need non-degeneracy and second-order sufficiency assumptions, which are standard in this context.

Assumption 3.11. A local minima, $\mathbf{x}_*$, is non-degenerate if $[\mathbf{g}(\mathbf{x}_*)]^i > 0, \forall i \in \mathcal{A}(\mathbf{x}_*, 0)$.

Assumption 3.12. A local minima, $\mathbf{x}_*$, satisfies the second-order sufficiency condition if $0 < \langle \mathbf{z}, \mathbf{H}(\mathbf{x}_*)\mathbf{z} \rangle$ for all $\mathbf{z} \neq 0$ such that $\mathbf{z}^i = 0$ if $i \in \mathcal{A}(\mathbf{x}_*, 0)$.

Theorem 3.13 (Active Set Identification)**.** *Let f satisfy Assumption 3.1 and $\mathbf{x}_*$ be a local minima satisfying Assumptions 3.11 and 3.12. Let $\{\mathbf{x}_k\}$ be the sequence of iterates generated by Algorithm 4 with δ chosen according to (44). There exists $\Delta_{actv} > 0$ such that if $\mathbf{x}_{\bar{k}} \in B(\mathbf{x}_*, \Delta_{actv})$, then $\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_k, 0) = \mathcal{A}(\mathbf{x}_*, 0)$ for all $k \geq \bar{k} + 1$.*

We defer the proof to Appendix E. Once the active set is identified, our method reduces to unconstrained Newton-MR on the inactive set. Local convergence is therefore a simple corollary of Theorem 3.13.

Corollary 3.14 (Local Convergence)**.** *For $k \geq \bar{k} + 1$ (cf. Theorem 3.13), the convergence of Algorithm 4 is driven by the local properties of the Newton-MR portion of the step.*

Remark 3.15. The local convergence of Newton-MR is similar to that of other inexact Newton methods. Suppose that we use a relative residual tolerance, $\|\mathbf{r}^{\mathcal{I}}\| \leq \eta\|\mathbf{g}^{\mathcal{I}}\|$, as the criteria for the MINRES termination. Under Assumption 3.12, we know that $\mathbf{H}(\mathbf{x}_*)$ is positive definite on the inactive indices. Therefore, by applying Nocedal & Wright (2006, Theorem 7.1 and 7.2), we obtain a superlinear convergence if we choose $\eta = \mathcal{O}(1)$ and let $\mathbf{x}_k$ be close enough to $\mathbf{x}_*$. If we choose $\eta = \mathcal{O}(\|\mathbf{g}_k\|)$ and the Hessian is Lipschitz then we can improve the rate to quadratic.

Remark 3.16. A central ingredient in the projected Newton-CG of Xie & Wright (2023) is the damping of the Hessian in the form of diagonal perturbation (i.e., $\mathbf{H} + \epsilon\mathbf{I}$) for all Newton-CG steps in the inactive set. While this facilitates an optimal global complexity, an unfortunate consequence, at least in theory, is that the algorithm no longer enjoys a guaranteed fast "Newton-type" local convergence rate. In other words, one can at best show linear rates in local regimes.

4. Numerical Experiments

We now compare the performance of our method for solving (1) with several alternatives using various convex and non-convex examples. Specifically, we consider Algorithm 2 (denoted by **MR**), projected Newton-CG (denoted by **CG**) as in Xie & Wright (2023, Algorithm 1), and projected gradient with line search (denoted by **PG**) (Beck, 2017). For convex problems, we also include **FISTA** with line search (Beck & Teboulle, 2009), while for non-convex settings, we compare against the proximal gradient with momentum and fine-tuned constant step size (denoted by **PGM**) from Lin et al. (2020, Algorithm 4.1). We exclude proximal Newton methods due to the difficulty of solving its subproblems at each iteration. We also do not consider the Newton-CG log barrier method (O'Neill & Wright, 2020) due to poor practical performance observed in Xie & Wright (2023).

For all applicable methods we terminate according to (7) with $\epsilon_g = 10^{-8}$. Instead of the highly implementation dependent "wall-clock" time, here we plot the objective value against the number of *oracle calls*, i.e., the number of equivalent function evaluations. For completeness, however, we also include plots of objective value against wall-clock time in Appendix F.5. The PyTorch (Paszke et al., 2019) implementation for our experiments is available here. All experiments were performed on a GPU cluster. See Appendix F.3 for further experimental details.

4.1. Sparse Regularisation With ℓ_1 Norm

We first consider sparse regression using ℓ_1-regularisation

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1, \tag{12}$$

where f is a smooth function. Although the objective function in (12) is nonsmooth, it can be reformulated into a smooth optimisation problem with nonnegativity constraints; see Appendix F.2 for details. We consider two examples in this context.

Multinomial Regression. In Figures 1 and 2, we consider convex multinomial regression with C classes where f is given by (62). The FISTA method is applied directly to (12). While FISTA clearly outperforms the others, our method is competitive. Further simulations showing fast local convergence of our method on these examples are given in Appendix F.4.

Neural Network. Figure 3 shows the results using a two layer neural network where f is non-convex and defined by (63). Again, PGM is applied directly to (12) and its step size is fine-tuned for best performance. We once again observe superior performance of our method compared with the alternatives.

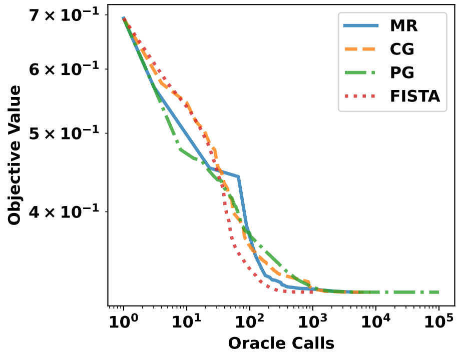

Figure 1. Logistic regression ($C = 2$) on the binarised MNIST dataset (LeCun et al., 1998) ($d = 785$) with $\lambda = 10^{-3}$.

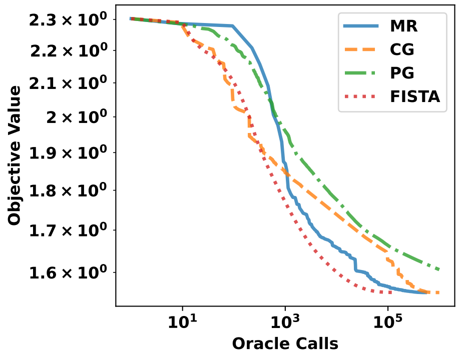

Figure 2. Multinomial regression ($C = 10$) on CIFAR10 dataset (Krizhevsky, 2009) ($d = 27,657$) with $\lambda = 10^{-4}$.

4.2. Nonnegative Matrix Factorisation

Given a nonnegative data matrix $\mathbf{Y} \in \mathbb{R}_{+}^{n\times m}$, nonnegative matrix factorisation (NNMF) aims to produce two low rank, say r, nonnegative matrices $\mathbf{W} \in \mathbb{R}_{+}^{n\times r}$ and $\mathbf{H} \in \mathbb{R}_{+}^{r\times m}$ such that $\mathbf{Y} \approx \mathbf{WH}$. This can be formulated as

$$\min_{\mathbf{W}\geq 0,\ \mathbf{H}\geq 0} D(\mathbf{Y}, \mathbf{WH}) + R_\lambda(\mathbf{W}, \mathbf{H}), \tag{13}$$

where $D(\cdot,\cdot)$ is a 'distance' and $R_\lambda(\cdot,\cdot)$ is a regularisation term. In Figure 4, we consider a text dataset and cosine similarity based distance function, while in Figure 5, we use an image dataset and a Euclidean distance function with a nonconvex regulariser; see Appendix F.3 for details. Clearly, our method outperforms all others across both problems.

5. Conclusions and Future Directions

We developed Newton-MR variants of the two-metric projection framework. By inexactly solving the subproblems using MINRES as well as employing non-positive curvature directions, our proposed variants are suitable for large scale and nonconvex settings. We demonstrated that, under

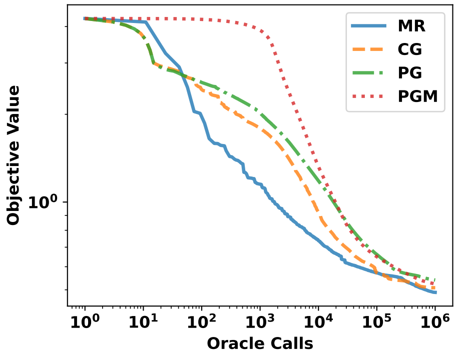

Figure 3. Training a two-layer neural network on the `Fashion MNIST` dataset (Xiao et al., 2017) ($d = 89,610$) with $\lambda = 10^{-3}$.

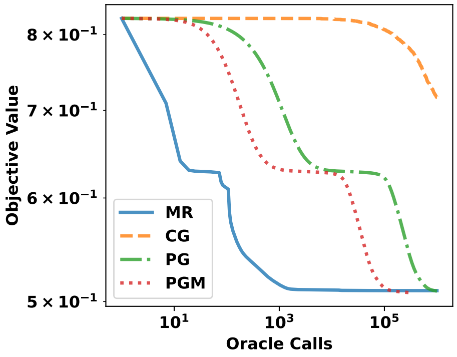

Figure 4. NNMF ($r = 20$) with cosine distance on top 1000 TF-IDF features of the `20 Newsgroup` dataset (Mitchell, 1999) ($d = 385,220$).

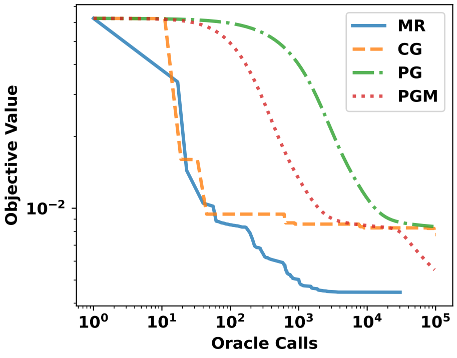

Figure 5. NNMF ($r = 10$) with nonconvex TSCAD regulariser on the `Olivetti faces` dataset (Pedregosa et al., 2011) ($d = 44,960$). We used $a = 3$ and $\lambda = 10^{-4}$ for the TSCAD regulariser.

certain assumptions, the convergence rates of our methods match the state-of-the-art and showcased competitive practical performance across a variety of problems.

Possible avenues for future research include extensions to box constraints, variants with second-order complexity guarantees, and the development of stochastic algorithms.

Acknowledgements

This research was partially supported by the Australian Research Council through an Industrial Transformation Training Centre for Information Resilience (IC200100022).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Beck, A. First-Order Methods in Optimization. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2017. ISBN 9781611974997. URL `https://books.google.com.au/books?id=wrk4DwAAQBAJ`.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009. doi: 10.1137/080716542. URL `https://doi.org/10.1137/080716542`.

Becker, S. and Fadili, J. A quasi-Newton proximal splitting method. Advances in Neural Information Processing Systems, 25, 2012.

Bernstein, D. S. Matrix Mathematics: Theory, Facts, and Formulas (Second Edition). Princeton University Press, 2009. ISBN 9780691140391. URL `http://www.jstor.org/stable/j.ctt7t833`.

Bertsekas, D. P. On the Goldstein-Levitin-Polyak gradient projection method. IEEE Transactions on Automatic Control, 21(2):174–184, 1976. doi: 10.1109/TAC.1976.1101194.

Bertsekas, D. P. Projected Newton methods for optimization problems with simple constraints. SIAM Journal on Control and Optimization, 20(2):221–246, 1982. doi: 10.1137/0320018. URL `https://doi.org/10.1137/0320018`.

Bertsekas, D. P. Constrained Optimization and Lagrange

Multiplier Methods. Athena Scientific, 1996. ISBN 1-886529-04-30.

Bertsekas, D. P. Nonlinear Programming. Athena Scientific, Belmont, Mass., 2nd ed edition, 1999. ISBN 1886529000.

Bian, W., Chen, X., and Ye, Y. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. Mathematical Programming, 149:301–327, 02 2014. doi: 10.1007/s10107-014-0753-5.

Birgin, E. G. and Martínez, J. M. On regularization and active-set methods with complexity for constrained optimization. SIAM Journal on Optimization, 28(2): 1367–1395, 2018. doi: 10.1137/17M1127107. URL https://doi.org/10.1137/17M1127107.

Birgin, E. G., Martínez, J. M., and Raydan, M. Spectral projected gradient methods: Review and perspectives. Journal of Statistical Software, 60(3): 1–21, 2014. doi: 10.18637/jss.v060.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v060i03.

Burke, J. On the identification of active constraints II: The nonconvex case. SIAM Journal on Numerical Analysis, 27(4):1081–1102, 1990. ISSN 00361429. URL http://www.jstor.org/stable/2157700.

Burke, J. V. and Moré, J. J. On the identification of active constraints. SIAM Journal on Numerical Analysis, 25(5): 1197–1211, 1988. doi: 10.1137/0725068. URL https://doi.org/10.1137/0725068.

Cai, J.-F., de Miranda Cardoso, J. V., Palomar, D., and Ying, J. Fast projected Newton-like method for precision matrix estimation under total positivity. In Advances in Neural Information Processing Systems, volume 36, pp. 73348–73370, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e878c8f38381d0964677fb9536c494ee-Paper-Conference.pdf.

Cartis, C., Gould, N. I. M., and Toint, P. L. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM J. on Optimization, 20(6):2833–2852, sep 2010. ISSN 1052-6234. doi: 10.1137/090774100. URL https://doi.org/10.1137/090774100.

Cartis, C., Gould, N., and Toint, P. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. Mathematical programming, 130(2):295–319, 2011a. ISSN 0025-5610. doi: 10.1007/s10107-009-0337-y.

Cartis, C., Gould, N., and Toint, P. Optimal Newton-type methods for nonconvex smooth optimization problems. ERGO Technical Report, 11-009, 2011b.

Cartis, C., Gould, N. I., and Toint, P. L. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. Math. Program., 127(2):245–295, 2011c. ISSN 0025-5610.

Cartis, C., Gould, N. I. M., and Toint, P. L. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. SIAM Journal on Optimization, 30(1):513–541, 2020. doi: 10.1137/17M1144854. URL https://doi.org/10.1137/17M1144854.

Cartis, C., Gould, N. I. M., and Toint, P. L. Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022. doi: 10.1137/1.9781611976991. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611976991.

Clarke, F. H. Optimization and Nonsmooth Analysis. SIAM, 1990.

Curtis, F. E. and Wang, Q. Worst-case complexity of TRACE with inexact subproblem solutions for nonconvex smooth optimization. SIAM Journal on Optimization, 33 (3):2191–2221, 2023. doi: 10.1137/22M1492428. URL https://doi.org/10.1137/22M1492428.

Curtis, F. E., Robinson, D., and Samadi, M. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. Mathematical Programming, 162:1–32, 2016. doi: 10.1007/s10107-016-1026-2.

Curtis, F. E., Robinson, D. P., Royer, C. W., and Wright, S. J. Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization. SIAM Journal on Optimization, 31(1):518–544, 2021. doi: 10.1137/19M130563X. URL https://doi.org/10.1137/19M130563X.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. arXiv preprint arXiv:1702.03118, 2017.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360, 2001. ISSN 01621459. URL http://www.jstor.org/stable/3085904.

Gafni, E. M. and Bertsekas, D. P. Two-metric projection methods for constrained optimization. SIAM Journal on Control and Optimization, 22(6):936–964, 1984. doi: 10.1137/0322061. URL https://doi.org/10.1137/0322061.

Gillis, N. The why and how of nonnegative matrix factorization. arXiv preprint arXiv:1401.5226, 2014.

Gillis, N. Nonnegative Matrix Factorization. SIAM, 2020.

Goodfellow, I., Bengio, Y., and Courville, A. Deep Learning. MIT press, 2016.

Haber, E. Computational Methods in Geophysical Electromagnetics. SIAM, 2014. doi: https://doi.org/10.1137/1.9781611973808.

Haeser, G., Liu, H., and Ye, Y. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. Mathematical Programming, 178, 02 2017. doi: 10.1007/s10107-018-1290-4.

Kim, D., Sra, S., and Dhillon, I. S. Tackling box-constrained optimization via a new projected quasi-newton approach. SIAM Journal on Scientific Computing, 32(6):3548–3563, 2010. doi: 10.1137/08073812X. URL https://doi.org/10.1137/08073812X.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kuang, D., Yun, S., and Park, H. SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. Journal of Global Optimization, 62(3):545–574, July 2015. doi: 10.1007/s10898-014-0247-2. URL https://ideas.repec.org/a/spr/jglopt/v62y2015i3p545-574.html.

Lan, G. First-order and Stochastic Optimization Methods for Machine Learning, volume 1. Springer, 2020.

LeCun, Y., Cortes, C., and Burges, C. The MNIST database of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/.

Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. Nature, 401:788–91, 11 1999. doi: 10.1038/44565.

Lee, D. and Seung, H. S. Algorithms for non-negative matrix factorization. In Leen, T., Dietterich, T., and Tresp, V. (eds.), Advances in Neural Information Processing Systems, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

Lee, J. D., Sun, Y., and Saunders, M. A. Proximal Newton-type methods for minimizing composite functions. SIAM Journal on Optimization, 24(3):1420–1443, 2014. doi: 10.1137/130921428. URL https://doi.org/10.1137/130921428.

Li, Q., Zhou, Y., Liang, Y., and Varshney, P. K. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In International Conference on Machine Learning, pp. 2111–2119. PMLR, 2017.

Lim, A. and Roosta, F. Complexity guarantees for nonconvex Newton-MR under inexact Hessian information. arXiv preprint arXiv:2308.09912, 2023.

Lim, A., Liu, Y., and Roosta, F. Conjugate direction methods under inconsistent systems. arXiv preprint arXiv:2401.11714, 2024.

Lin, Z., Li, H., and Fang, C. Accelerated Optimization for Machine Learning. Springer, 2020.

Liu, Y. and Roosta, F. Convergence of Newton-MR under inexact Hessian information. SIAM Journal on Optimization, 31(1):59–90, 2021. doi: 10.1137/19m1302211. URL https://doi.org/10.1137%2F19m1302211.

Liu, Y. and Roosta, F. MINRES: From negative curvature detection to monotonicity properties. SIAM Journal on Optimization, 32(4):2636–2661, 2022a. doi: 10.1137/21M143666X. URL https://doi.org/10.1137/21M143666X.

Liu, Y. and Roosta, F. A Newton-MR algorithm with complexity guarantees for nonconvex smooth unconstrained optimization. arXiv preprint arXiv:2208.07095, 2022b. doi: 10.48550/ARXIV.2208.07095. URL https://arxiv.org/abs/2208.07095.

Mitchell, T. Twenty Newsgroups. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C5C323.

Nesterov, Y. Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization. Springer US, 2004. ISBN 978-1-4020-7553-7. doi: https://doi.org/10.1007/978-1-4419-8853-9.

Nesterov, Y. Gradient methods for minimizing composite functions. Mathematical Programming, 140(1):125–161, 2013. doi: https://doi.org/10.1007/s10107-012-0629-5.

Nesterov, Y. and Polyak, B. Cubic regularization of Newton method and its global performance. Math. Program., 108: 177–205, 08 2006. doi: 10.1007/s10107-006-0706-8.

Nocedal, J. and Wright, S. J. Numerical Optimization. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, second edition edition, 2006. ISBN 9780387303031.

O'Neill, M. and Wright, S. J. A log-barrier Newton-CG method for bound constrained optimization with complexity guarantees. IMA Journal of Numerical Analysis, 41(1):84–121, 04 2020. ISSN 0272-4979. doi: 10.1093/imanum/drz074. URL https://doi.org/10.1093/imanum/drz074.

Paige, C. C. and Saunders, M. A. Solution of sparse indefinite systems of linear equations. SIAM Journal on Numerical Analysis, 12:617–629, 9 1975. ISSN 0036-1429. doi: 10.1137/0712047.

Parikh, N. and Boyd, S. Proximal algorithms. Found. Trends Optim., 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pearlmutter, B. A. Fast exact multiplication by the Hessian. Neural computation, 6(1):147–160, 1994.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

Roosta, F., Liu, Y., Xu, P., and Mahoney, M. W. Newton-MR: Inexact Newton method with minimum residual sub-problem solver. EURO Journal on Computational Optimization, 10:100035, 2022. ISSN 2192-4406. doi: https://doi.org/10.1016/j.ejco.2022.100035. URL https://www.sciencedirect.com/science/article/pii/S2192440622000119.

Royer, C. W., O'Neill, M., and Wright, S. J. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. arXiv preprint arXiv:1803.02924, 3 2018.

Saad, Y. Iterative Methods for Sparse Linear Systems. SIAM, 2nd edition, 2003. ISBN 9780898715347.

Schmidt, M., Fung, G., and Rosales, R. Fast optimization methods for L1 regularization: A comparative study and two new approaches. In Kok, J. N., Koronacki, J., Mantaras, R. L. d., Matwin, S., Mladenič, D., and Skowron, A. (eds.), Machine Learning: ECML 2007, pp. 286–297, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74958-5.

Schmidt, M., Berg, E., Friedlander, M., and Murphy, K. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In Artificial intelligence and statistics, pp. 456–463. PMLR, 2009.

Schmidt, M., Kim, D., and Sra, S. Projected Newton-type Methods in Machine Learning. In Optimization for Machine Learning. The MIT Press, 09 2011. ISBN 9780262298773. doi: 10.7551/mitpress/8996.003.0013. URL https://doi.org/10.7551/mitpress/8996.003.0013.

Shi, Z. and Liu, R. Large scale optimization with proximal stochastic Newton-type gradient descent. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15, pp. 691–704. Springer, 2015.

Sun, Y., Jeong, H., Nutini, J., and Schmidt, M. Are we there yet? Manifold identification of gradient-related proximal methods. In Chaudhuri, K. and Sugiyama, M. (eds.), Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of Proceedings of Machine Learning Research, pp. 1110–1119. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/sun19a.html.

Wright, S. J. Identifiable surfaces in constrained optimization. SIAM Journal on Control and Optimization, 31(4):1063–17, 07 1993. Copyright - Copyright] © 1993 Society for Industrial and Applied Mathematics; Last updated - 2023-12-04.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Xie, Y. and Wright, S. J. Complexity of a projected Newton-CG method for optimization with bounds. Mathematical Programming, July 2023. doi: https://doi.org/10.1007/s10107-023-02000-z.

Xu, P., Roosta, F., and Mahoney, M. W. Newton-type methods for non-convex optimization under inexact Hessian information. Mathematical Programming, 184(1-2):35–70, 2020a.

Xu, P., Roosta, F., and Mahoney, M. W. Second-order optimization for non-convex machine learning: An empirical study. In Proceedings of the 2020 SIAM International Conference on Data Mining, pp. 199–207. SIAM, 2020b.

Yao, Z., Xu, P., Roosta, F., Wright, S. J., and Mahoney, M. W. Inexact Newton-CG algorithms with complexity guarantees. IMA Journal of Numerical Analysis, 43(3):1855–1897, 08 2022. ISSN 0272-4979. doi: 10.1093/imanum/drac043. URL https://doi.org/10.1093/imanum/drac043.

A. MINRES and Newton-MR

In this section, for completeness, we discuss MINRES (Algorithm 3) and provide some of its fundamental properties. We note that our presentation is essentially that of Liu & Roosta (2022b, Appendix A) as the notation and implementation is well adapted to our setting. Recall that MINRES combines the Lanczos process, a QR decomposition, and an updating formula to iteratively solve a symmetric linear least-squares problem of the form

$$\min_{\mathbf{s}\in\mathbb{R}^d} \|\mathbf{H}\mathbf{s}+\mathbf{g}\|^2.$$

We now discuss each of these aspects in detail.

Lanczos Process. Recall that, starting from $\mathbf{v}_1 = \mathbf{g}/\|\mathbf{g}\|$, after t iterations of the Lanczos process, the Lanczos vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{t+1}\}$, form a basis for the Krylov subspace $\mathcal{K}_{t+1}(\mathbf{H},\mathbf{g})$. Collecting these vectors into an orthogonal matrix

$$\mathbf{V}_{t+1} = [\mathbf{v}_1, \ldots \mathbf{v}_{t+1}] \in \mathbb{R}^{d\times(t+1)},$$

we can write

$$\mathbf{H}\mathbf{V}_t = \mathbf{V}_{t+1}\tilde{\mathbf{T}}_t,$$

where $\tilde{\mathbf{T}}_t \in \mathbb{R}^{(t+1),t}$ is an upper Hessenberg matrix of the form

$$\mathbf{T}_t = \begin{pmatrix} \tilde{\alpha}_1 & \tilde{\beta}_2 & & & \\ \tilde{\beta}_2 & \tilde{\alpha}_2 & \tilde{\beta}_3 & & \\ & \tilde{\beta}_3 & \tilde{\alpha}_3 & \ddots & \\ & & \ddots & \ddots & \tilde{\beta}_t \\ & & & \tilde{\beta}_t & \tilde{\alpha}_t \end{pmatrix}, \ \tilde{\mathbf{T}}_t \triangleq \begin{pmatrix} \mathbf{T}_t \\ \tilde{\beta}_{t+1}\mathbf{e}_t^\intercal \end{pmatrix}.$$

This relation yields the underlying update process of the MINRES iterations for $t \geq 2$ as,

$$\mathbf{H}\mathbf{v}_t = \tilde{\beta}_t\mathbf{v}_{t-1} + \tilde{\alpha}_t\mathbf{v}_t + \tilde{\beta}_{t+1}\mathbf{v}_{t+1}.$$

The Lanczos process terminates when $\tilde{\beta}_{t+1} = 0$. We remark that computing an expansion of the basis requires a single Hessian-vector product, $\mathbf{H}\mathbf{v}_t$. The basis for the Krylov subspace allows us to significantly simplify (8). Indeed, let $\mathbf{s}_t$ be a solution to (8) at iteration t. By $\mathbf{s}_t \in \mathcal{K}_t(\mathbf{H},\mathbf{g})$, we have $\mathbf{s}_t = \mathbf{V}_t\mathbf{y}_t$ for some $\mathbf{y}_t \in \mathbb{R}^t$. Hence, the residual can be written as

$$\mathbf{r}_t = -\mathbf{g} - \mathbf{H}\mathbf{s}_t = -\mathbf{g} - \mathbf{H}\mathbf{V}_t\mathbf{y}_t = -\mathbf{g} - \mathbf{V}_{t+1}\tilde{\mathbf{T}}_t\mathbf{y}_t = -\mathbf{V}_{t+1}(\|\mathbf{g}\|\mathbf{e}_1 + \tilde{\mathbf{T}}_t\mathbf{y}_t).$$

In the final equality, we applied the orthogonality of the basis vectors and $\mathbf{v}_1 = \mathbf{g}/\|\mathbf{g}\|$. Applying this expression to (8) and using the orthogonality of $\mathbf{V}_{t+1}$, we obtain the reduced tridiagonal least-squares problem

$$\min_{\mathbf{y}_t\in\mathbb{R}^t} \left\|\tilde{\beta}_1\mathbf{e}_1 + \tilde{\mathbf{T}}_t\mathbf{y}_t\right\|, \tag{14}$$

where $\tilde{\beta}_1 = \|\mathbf{g}\|$.

QR Factorisation. The next step in the MINRES procedure is to solve (14) by computing the full QR factorisation $\mathbf{Q}_t\tilde{\mathbf{T}}_t = \tilde{\mathbf{R}}_t$ where $\mathbf{Q}_t \in \mathbb{R}^{(t+1)\times(t+1)}$ and $\tilde{\mathbf{R}}_t \in \mathbb{R}^{(t+1)\times t}$. Because $\tilde{\mathbf{T}}_t$ is already close to being upper triangular, we form the QR factorisation using a series of Householder reflections to annihilate the sub-diagonal elements. Each Householder reflection affects only two rows of $\tilde{\mathbf{T}}_t$. We can summarise the effect of two successive Householder reflections for $3 \leq i \leq t-1$ as

$$\begin{aligned}
\begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{i-1} & s_{i-1} \\ 0 & s_{i-1} & -c_{i-1} \end{pmatrix}
\begin{pmatrix} c_{i-2} & s_{i-2} & 0 \\ s_{i-2} & -c_{i-2} & 0 \\ 0 & 0 & 1 \end{pmatrix}
\begin{pmatrix} \gamma_{i-2} & \delta_{i-1} & 0 & 0 \\ \tilde{\beta}_{i-1} & \tilde{\alpha}_{i-1} & \tilde{\beta}_i & 0 \\ 0 & \tilde{\beta}_i & \tilde{\alpha}_i & \tilde{\beta}_{i+1} \end{pmatrix}
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{i-1} & s_{i-1} \\ 0 & s_{i-1} & -c_{i-1} \end{pmatrix}
\begin{pmatrix} \gamma^{[2]}_{i-2} & \delta^{[2]}_{i-1} & \epsilon_i & 0 \\ 0 & \gamma_{i-1} & \delta_i & 0 \\ 0 & \tilde{\beta}_i & \tilde{\alpha}_i & \tilde{\beta}_{i+1} \end{pmatrix} \\
&= \begin{pmatrix} \gamma^{[2]}_{i-2} & \delta^{[2]}_{i-1} & \epsilon_i & 0 \\ 0 & \gamma^{[2]}_{i-1} & \delta^{[2]}_i & \epsilon_{i+1} \\ 0 & 0 & \gamma_i & \delta_{i+1} \end{pmatrix},
\end{aligned}$$

where for $1 \leq j \leq t$ we have

$$c_j = \frac{\gamma_j}{\gamma^{[2]}_j}, \quad s_j = \frac{\tilde{\beta}_{j+1}}{\gamma^{[2]}_j}, \quad \gamma^{[2]}_j = \sqrt{(\gamma_j)^2 + \tilde{\beta}^2_{j+1}} = c_j \gamma_j + s_j \tilde{\beta}_{j+1}.$$

We therefore form $\mathbf{Q}_t$ as a product of the Householder reflection matrices

$$\mathbf{Q}_t = \prod_{i=1}^{t} \mathbf{Q}_{i,i+1}, \quad \mathbf{Q}_{i,i+1} \triangleq \begin{pmatrix} \mathbf{I}_{i-1} & & & \\ & c_t & s_t & \\ & s_t & -c_t & \\ & & & \mathbf{I}_{t-i} \end{pmatrix}.$$

It is also clear that $\tilde{\mathbf{R}}_t$ is given by

$$\mathbf{R}_t \triangleq \begin{pmatrix} \gamma^{[2]}_1 & \delta^{[2]}_2 & \epsilon_3 & & \\ & \gamma^{[2]}_2 & \delta^{[2]}_3 & \ddots & \\ & & \ddots & \ddots & \epsilon_t \\ & & & \gamma^{[2]}_{t-1} & \delta^{[2]}_t \\ & & & & \gamma^{[2]}_t \end{pmatrix}, \quad \tilde{\mathbf{R}}_t = \begin{pmatrix} \mathbf{R}_t \\ \mathbf{0}^\intercal \end{pmatrix}.$$

Applying $\mathbf{Q}_t$ to $\tilde{\beta}_1 \mathbf{e}_1$, we obtain

$$\mathbf{Q}_t \tilde{\beta}_1 \mathbf{e}_1 = \tilde{\beta}_1 \begin{pmatrix} c_1 \\ s_1 c_2 \\ \vdots \\ s_1 s_2 \cdots s_{t-1} c_t \\ s_1 s_2 \cdots s_{t-1} s_t \end{pmatrix} \triangleq \begin{pmatrix} \tau_1 \\ \tau_2 \\ \ldots \tau_t \\ \phi_t \end{pmatrix} \triangleq \begin{pmatrix} \mathbf{t}_t \\ \phi_t \end{pmatrix}.$$

Applying the QR factorisation to solve (14) gives

$$\begin{aligned}
\min_{\mathbf{y}_t} \left\| \tilde{\beta}_1 \mathbf{e}_1 + \tilde{\mathbf{T}}_t \mathbf{y}_t \right\| &= \min_{\mathbf{y}_t} \left\| \mathbf{Q}_t^\intercal (\tilde{\beta}_1 \mathbf{Q}_t \mathbf{e}_1 + \mathbf{Q}_t \tilde{\mathbf{T}}_t \mathbf{y}_t) \right\| \\
&= \min_{\mathbf{y}_t} \left\| \begin{pmatrix} \mathbf{t}_t \\ \phi_t \end{pmatrix} + \begin{pmatrix} \mathbf{R}_t \\ \mathbf{0}^\intercal \end{pmatrix} \mathbf{y}_t \right\|.
\end{aligned}$$

An immediate implication of this result is $\phi_t = \|\mathbf{r}_t\|$.

Update. The key to the computational efficiency of MINRES is the existence of vector update formula, which eliminates the requirement to form or store the matrices involved in the Lanczos and QR factorisation processes, i.e., $\mathbf{V}_t$, $\mathbf{Q}_t$, $\tilde{\mathbf{R}}_t$, and $\tilde{\mathbf{T}}_t$. Define $\mathbf{W}_t$ from the upper triangular system $\mathbf{W}_t\mathbf{R}_t = \mathbf{V}_t$ as

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_t \end{pmatrix} = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_t \end{pmatrix} \begin{pmatrix} \gamma_1^{[2]} & \delta_2^{[2]} & \epsilon_3 & & \\ & \gamma_2^{[2]} & \delta_3^{[2]} & \ddots & \\ & & \ddots & \ddots & \epsilon_t \\ & & & \gamma_{t-1}^{[2]} & \delta_t^{[2]} \\ & & & & \gamma_t^{[2]} \end{pmatrix}. \tag{15}$$

By reading off (15), we see that

$$\mathbf{v}_t = \epsilon_t \mathbf{w}_{t-2} + \delta_t^{[2]} \mathbf{w}_{t-1} + \gamma_t^{[2]} \mathbf{w}_t.$$

The computation for the MINRES iterate can now be written as

$$\mathbf{s}_t = \mathbf{V}_t \mathbf{y}_t = \mathbf{W}_t \mathbf{R}_t \mathbf{y}_t = \mathbf{W}_t \mathbf{t}_t = \begin{pmatrix} \mathbf{W}_{t-1} & \mathbf{w}_t \end{pmatrix} \begin{pmatrix} \mathbf{t}_{t-1} \\ \tau_t \end{pmatrix} = \mathbf{s}_{t-1} + \tau_t \mathbf{w}_t,$$

where we set $\mathbf{s}_0 = 0$. With this result in mind, we give the full MINRES method in Algorithm 3. We remark that, in Algorithm 3, we have also included steps for verifying the inexactness condition (10) (Algorithm 3-Line 10) as well as certifying $\langle \mathbf{r}_t, \mathbf{H}\mathbf{r}_t \rangle \geq \vartheta \|\mathbf{r}_t\|^2$ for some user specified $\vartheta \geq 0$ (Algorithm 3-Line 7).

Algorithm 3 MINRES($\mathbf{H}$, $\mathbf{g}$, η, ϑ)

```
1: Input Hessian $\mathbf{H}$, gradient $\mathbf{g}$, inexactness tolerance $\eta > 0$, and NPC tolerance $\vartheta \geq 0$.
2: $\phi_0 = \tilde{\beta}_0 = \|\mathbf{g}\|$, $\mathbf{r}_0 = -\mathbf{g}$, $\mathbf{v}_1 = \mathbf{r}_0/\phi_0$, $\mathbf{v}_0 = \mathbf{s}_0 = \mathbf{w}_0 = \mathbf{w}_{-1} = 0$.
3: $s_0 = 0$, $c_0 = -1$, $\delta_1 = \tau_0 = 0$, $t = 1$.
4: while True do
5:     $\mathbf{q}_t = \mathbf{H}\mathbf{v}_t$, $\tilde{\alpha}_t = \langle \mathbf{v}_t, \mathbf{q}_t \rangle$, $\mathbf{q}_t = \mathbf{q}_t - \tilde{\beta}_t \mathbf{v}_{t-1}$, $\mathbf{q}_t = \mathbf{q}_t - \tilde{\alpha}_t \mathbf{v}_t$, $\tilde{\beta}_{t+1} = \|\mathbf{q}_t\|$.
6:     $\begin{pmatrix} \delta_t^{[2]} & \epsilon_{t+1} \\ \gamma_t & \delta_{t+1} \end{pmatrix} = \begin{pmatrix} c_{t-1} & s_{t-1} \\ s_{t-1} & -c_{t-1} \end{pmatrix} \begin{pmatrix} \delta_t & 0 \\ \tilde{\alpha}_t & \tilde{\beta}_{t+1} \end{pmatrix}$
7:     if $-c_{t-1}\gamma_t \leq \vartheta$ then
8:         return ($\mathbf{r}_{t-1}$, $\text{D}_{\text{type}} = \text{NPC}$).
9:     end if
10:    if $\phi_{t-1}\sqrt{\gamma_t^2 + \delta_{t+1}^2} \leq \eta\sqrt{\phi_0^2 - \phi_{t-1}^2}$ then
11:        return ($\mathbf{s}_{t-1}$, $\text{D}_{\text{type}} = \text{SOL}$).
12:    end if
13:    $\delta_t^{[2]} = \sqrt{\gamma_t^2 + \tilde{\beta}_{t+1}^2}$.
14:    if $\delta_t^{[2]} \neq 0$ then
15:        $c_t = \gamma_t/\gamma_t^{[2]}$, $s_t = \tilde{\beta}_{t+1}/\gamma_t^{[2]}$, $\tau_t = c_t\phi_{t-1}$, $\phi_t = s_t\phi_{t-1}$.
16:        $\mathbf{w}_t = (\mathbf{v}_t - \gamma_t^{[2]}\mathbf{w}_{t-1} - \epsilon_t\mathbf{w}_{t-2})/\gamma_t^{[2]}$, $\mathbf{s}_t = \mathbf{s}_{t-1} + \tau_t\mathbf{w}_t$.
17:        if $\tilde{\beta}_{t+1} \neq 0$ then
18:            $\mathbf{v}_{t+1} = \mathbf{q}_t/\tilde{\beta}_{t+1}$, $\mathbf{r}_t = s_t^2\mathbf{r}_{t-1} - \phi_t c_t \mathbf{v}_{t+1}$.
19:        end if
20:    else
21:        $c_t = 0$, $s_t = 1$, $\tau_t = 0$, $\phi_t = \phi_{t-1}$, $\mathbf{r}_t = \mathbf{r}_{t-1}$, $\mathbf{s}_t = \mathbf{s}_{t-1}$.
22:    end if
23:    $t \leftarrow t + 1$.
24: end while
```

We now collect several properties of the MINRES for reference; see Liu & Roosta (2022a;b) for more details and properties. Firstly, we give some scalar expressions for the quantities of interest in (9) and (10) in the MINRES algorithm

Lemma A.1 (MINRES scalar updates)**.** *We have the following*

$$\|\mathbf{r}^{(t)}\| = \phi_t \tag{16a}$$

$$\langle \mathbf{r}^{(t-1)}, \mathbf{H}\mathbf{r}^{(t-1)}\rangle = -c_{t-1}\gamma_t\|\mathbf{r}^{(t-1)}\|^2, \tag{16b}$$

$$\|\mathbf{H}\mathbf{s}^{(t-1)}\| = \sqrt{\phi_0^2 - \phi_{t-1}^2}, \tag{16c}$$

$$\|\mathbf{H}\mathbf{r}^{(t-1)}\| = \phi_{t-1}\sqrt{\gamma_t^2 + \delta_{t+1}^2}. \tag{16d}$$

Proof. (16a) follows from the construction of the MINRES algorithm. The proof of (16d), (16c) and (16b) is given in Liu & Roosta (2022b, Lemma 11). □

Next we give some helpful properties of the SOL and NPC steps.

Lemma A.2 ($\mathrm{D}_{\text{type}} = \mathrm{SOL}$)**.** *Any iterate of MINRES, $\mathbf{s}^{(t)}$, satisfies*

$$\|\mathbf{H}\mathbf{s}^{(t)}\| \leq \|\mathbf{g}\|, \tag{17}$$

and

$$\langle \mathbf{s}^{(t)}, \mathbf{H}\mathbf{g}\rangle \leq 0. \tag{18}$$

Suppose that negative curvature has not been detected up to iteration t. Then,

$$\langle \mathbf{s}^{(t)}, \mathbf{g}\rangle \leq -\langle \mathbf{s}^{(t)}, \mathbf{H}\mathbf{s}^{(t)}\rangle. \tag{19}$$

Further, consider Assumption 3.1 and suppose there exists some $\varrho > 0$ such that for any $\mathbf{v} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$ we have $\langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle \geq \varrho\|\mathbf{v}\|^2$. Then,

$$C_{\varrho, L_g}\|\mathbf{g}\| \leq \|\mathbf{s}^{(t)}\| \leq \frac{\|\mathbf{g}\|}{\varrho}, \tag{20}$$

where $C_{\varrho,L_g} \triangleq \varrho/L_g^2$.

Proof. The relation (17) follows from Liu & Roosta (2021, Lemma 3.11), while (18) follows from the fact that $\mathbf{0} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$ and $\mathbf{s}^{(t)}$ minimises (8). Also, (19) follows from Liu & Roosta (2022a, Theorem 3.8). For the right-hand-side of (20), we use (17) and the fact that $\mathbf{s}^{(t)} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$ to get

$$\varrho\|\mathbf{s}^{(t)}\|^2 \leq \langle \mathbf{s}^{(t)}, \mathbf{H}\mathbf{s}^{(t)}\rangle \leq \|\mathbf{s}^{(t)}\|\|\mathbf{H}\mathbf{s}^{(t)}\| \leq \|\mathbf{s}^{(t)}\|\|\mathbf{g}\| \implies \|\mathbf{s}^{(t)}\| \leq \|\mathbf{g}\|/\varrho.$$

We show the left-hand-side of (20) using a monotonicity argument. In particular, consider the first iterate $\mathbf{s}^{(1)}$. It is easy to see that the solution to (8) over the Krylov subspace $\mathcal{K}_1(\mathbf{H}, \mathbf{g}) = \mathrm{Span}\{\mathbf{g}\}$ is given by

$$\min_{\mathbf{s}\in\mathcal{K}_1(\mathbf{H},\mathbf{g})} \|\mathbf{H}\mathbf{s} + \mathbf{g}\|^2 = \min_{\beta\in\mathbb{R}} \|\beta\mathbf{H}\mathbf{g} + \mathbf{g}\|^2 \implies \beta = -\frac{\langle \mathbf{g}, \mathbf{H}\mathbf{g}\rangle}{\|\mathbf{H}\mathbf{g}\|^2}.$$

The step is therefore given by

$$\mathbf{s}^{(1)} = -\frac{\langle \mathbf{g}, \mathbf{H}\mathbf{g}\rangle}{\|\mathbf{H}\mathbf{g}\|^2}\mathbf{g}.$$

We can apply $\langle \mathbf{g}, \mathbf{H}\mathbf{g}\rangle \geq \varrho\|\mathbf{g}\|^2$ and $\|\mathbf{H}\mathbf{g}\| \leq L_g\|\mathbf{g}\|$ to obtain

$$\begin{aligned}\|\mathbf{s}^{(1)}\| &= \frac{\langle \mathbf{g}, \mathbf{H}\mathbf{g}\rangle}{\|\mathbf{H}\mathbf{g}\|^2}\|\mathbf{g}\| \\ &\geq \frac{\varrho}{L_g^2}\|\mathbf{g}\|.\end{aligned}$$

The full results follows from the monotonicity of the MINRES iterates (Liu & Roosta, 2022a, Theorem 3.11), that is, as long as negative curvature remains undetected up to iteration $t \geq 1$ we have

$$\|\mathbf{s}^{(t)}\| \geq \|\mathbf{s}^{(1)}\| \geq \frac{\varrho}{L_g^2}\|\mathbf{g}\|.$$

□

Lemma A.3 ($\text{D}_{\text{type}} = \text{NPC}$)**.** *Suppose that the MINRES algorithm returns $D_{type} = NPC$ so that our step is $\mathbf{r}^{(t-1)}$. Then,*

$$\langle \mathbf{r}^{(t-1)}, \mathbf{g}\rangle = -\left\|\mathbf{r}^{(t-1)}\right\|^2. \tag{21}$$

Additionally, the residual norm is upper bounded by the gradient.

$$\|\mathbf{r}^{(t-1)}\| \leq \|\mathbf{g}\|. \tag{22}$$

Proof. The relation (21) follows from the MINRES properties directly (Liu & Roosta, 2022a, Lemma 3.1). We get (22) by noting that

$$\begin{aligned}\|\mathbf{H}\mathbf{s}^{(t-1)}\|^2 &= \|\mathbf{r}^{(t-1)} + \mathbf{g}\|^2 = \|\mathbf{r}^{(t-1)}\|^2 + 2\langle \mathbf{r}^{(t-1)}, \mathbf{g}\rangle + \|\mathbf{g}\|^2 \\ &= \|\mathbf{r}^{(t-1)}\|^2 - 2\|\mathbf{r}^{(t-1)}\|^2 + \|\mathbf{g}\|^2 = \|\mathbf{g}\|^2 - \|\mathbf{r}^{(t-1)}\|^2.\end{aligned}$$

For the third line, we applied (21). The final equality and the nonnegativity of the norm implies the result. □

B. Global Convergence - Minimal Assumptions

In this section, we detail the proof of the global convergence of Algorithm 1, i.e., Theorem 3.3. We first demonstrate that the uniform positive curvature certification of the residuals, $\mathbf{r}^{(i)}$, provides a bound on the curvature of the Hessian over the corresponding Krylov subspace.

Lemma B.1 (Strong Positive Curvature Certification)**.** *By verifying*

$$\langle \mathbf{r}^{(t-1)}, \mathbf{H}\mathbf{r}^{(i)}\rangle > \varsigma\|\mathbf{r}^{(i)}\|^2,$$

for $i = 0, \ldots, t-1$, we obtain

$$\langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle \geq \varsigma/(t+1)\|\mathbf{v}\|^2, \tag{23}$$

for any $\mathbf{v} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$.

Proof. Let $\mathbf{v} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$. We can write (Liu & Roosta, 2022a, Lemma A.1)

$$\mathcal{K}_t(\mathbf{H}, \mathbf{g}) = \text{Span}\left\{\mathbf{r}^{(0)}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(t-1)}\right\},$$

and therefore there exists a set of scalars, $\{\beta_i\}_{i=0}^{t-1}$, such that

$$\mathbf{v} = \sum_{i=0}^{t-1} \beta_i \mathbf{r}^{(i)}.$$

Using this fact and the certificates $\langle \mathbf{r}^{(i)}, \mathbf{H}\mathbf{r}^{(i)}\rangle \geq \varsigma\|\mathbf{r}^{(i)}\|^2$ gathered for $i = 0, \ldots, t-1$, we obtain

$$\begin{aligned}\langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle &= \left\langle \sum_{i=0}^{t-1} \beta_i \mathbf{r}^{(i)}, \sum_{i=0}^{t-1} \beta_i \mathbf{H}\mathbf{r}^{(i)}\right\rangle = \sum_{i=0}^{t-1}\sum_{j=0}^{t-1} \beta_i\beta_j \left\langle \mathbf{r}^{(i)}, \mathbf{H}\mathbf{r}^{(j)}\right\rangle \\ &= \sum_{i=0}^{t-1} \beta_i^2 \left\langle \mathbf{r}^{(i)}, \mathbf{H}\mathbf{r}^{(i)}\right\rangle \geq \sum_{i=0}^{t-1} \beta_i^2 \varsigma\|\mathbf{r}^{(i)}\|^2,\end{aligned} \tag{24}$$

where the second to last equality follows from the $\mathbf{H}$-conjugacy of the residuals (Liu & Roosta, 2022b, Lemma 11). Using Bernstein (2009, Fact 9.7.9), we get

$$\frac{1}{t+1}\left\|\sum_{i=0}^{t-1}\beta_i\mathbf{r}^{(i)}\right\|^2 \leq \sum_{i=0}^{t-1}\beta_i^2\|\mathbf{r}^{(i)}\|^2,$$

which gives the desired result. □

Note that since t appears in the lower bound (23), there is a dependence on the number of MINRES iterations undertaken and hence $\mathbf{x}$. However, t is bounded above by d. For this reason, in the sequel, we choose $\overline{\varsigma} = (d+1)\varsigma$ for some $\varsigma > 0$. Indeed, this choice implies that, under the conditions of Lemma B.1, for any $\mathbf{v} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})$ we have

$$\langle \mathbf{v}, \mathbf{H}\mathbf{v}\rangle \geq \varsigma\|\mathbf{v}\|^2. \tag{25}$$

We now demonstrate that the line search procedure (11) terminates for a small enough step size.

Lemma B.2 (Step-size Lower Bound)**.** *Suppose f satisfies Assumption 3.1. If at iteration k of Algorithm 1, we have $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$, then the largest step size, α_k, that satisfies the line search criteria (11), also satisfies the following lower bound*

$$\alpha_k \geq \min\left\{\frac{2(1-\rho)}{L_g}\min\{1,\varsigma\}, \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|}\right\}. \tag{26}$$

On the other hand, if $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, the bound is given by

$$\alpha_k \geq \frac{2(1-\rho)}{L_g}. \tag{27}$$

Proof. We note that the proof of Lemma C.1 utilises no curvature properties of the residual. With this fact in mind, the proof is entirely the same as Lemma C.1 in Appendix C with ς taking the place of σ. □

The following lemma gives the amount of decrease obtained from the inactive set step whenever the inactive set is nonempty and the inactive set termination condition (7c) is not satisfied.

Lemma B.3 (Sufficient Decrease: Inactive Set Case)**.** *Suppose f satisfies Assumption 3.1. Let $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k\mathbf{p}_k)$ be the update computed at iteration k of Algorithm 1, where α_k satisfies the line search criterion (11). Suppose $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and (7c) is not satisfied. If D$_{type}$ = SOL, then*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\varsigma\min\left\{\frac{2(1-\rho)\min\{1,\varsigma\}C_{\varsigma,L_g}^2}{L_g}\epsilon_k^4, C_{\varsigma,L_g}\delta_k\epsilon_k^2\right\},$$

where C_{ς,L_g} is as in (20). Otherwise, with D$_{type}$ = NPC,

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\min\left\{\frac{2(1-\rho)\eta^2}{L_g(\eta^2+L_g^2)}\epsilon_k^4, \frac{\eta\delta_k}{\sqrt{\eta^2+L_g^2}}\epsilon_k^2\right\}.$$

Proof. Since α_k satisfies the line search condition, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \rho\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha_k\mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle + \alpha_k\rho\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle \leq \alpha_k\rho\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle,$$

where we use the fact that $\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha\mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle \leq 0$. We now consider $\text{D}_{\text{type}} = \text{SOL}$ and $\text{D}_{\text{type}} = \text{NPC}$ cases.

When $\mathrm{D}_{\text{type}} = \text{SOL}$, we have $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$. Using the line search condition, (19), (25), (26), and the left-hand-side inequality in (20) with $\varrho = \varsigma$, we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho \alpha_k \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}} \rangle \\
&\leq -\rho \alpha_k \langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}} \mathbf{s}_k^{\mathcal{I}} \rangle \\
&\leq -\rho \varsigma \alpha_k \|\mathbf{s}_k^{\mathcal{I}}\|^2 \\
&\leq -\rho \varsigma \min \left\{ \frac{2(1-\rho)}{L_g} \min\{1, \varsigma\}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|} \right\} \|\mathbf{s}_k^{\mathcal{I}}\|^2 \\
&\leq -\rho \varsigma \min \left\{ \frac{2(1-\rho)}{L_g} \min\{1, \varsigma\} \|\mathbf{s}_k^{\mathcal{I}}\|^2, \delta_k \|\mathbf{s}_k^{\mathcal{I}}\| \right\} \\
&\leq -\rho \varsigma \min \left\{ \frac{2(1-\rho) \min\{1, \varsigma\} C_{\varsigma, L_g}^2}{L_g} \|\mathbf{g}_k^{\mathcal{I}}\|^2, C_{\varsigma, L_g} \delta_k \|\mathbf{g}_k^{\mathcal{I}}\| \right\} \\
&< -\rho \varsigma \min \left\{ \frac{2(1-\rho) \min\{1, \varsigma\} C_{\varsigma, L_g}^2}{L_g} \epsilon_k^4, C_{\varsigma, L_g} \delta_k \epsilon_k^2 \right\},
\end{aligned}$$

where we applied $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^2$ on the final line.

When $\mathrm{D}_{\text{type}} = \text{NPC}$, we have, $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$. We first note that, since the inexactness condition (10) has not been met, by applying Assumption 3.1 and using the fact that

$$\|\mathbf{H}\mathbf{s}^{(t-1)}\|^2 = \|\mathbf{g}\|^2 - \|\mathbf{r}^{(t-1)}\|^2,$$

we get

$$\|\mathbf{r}^{(t-1)}\| \geq \frac{\eta}{\sqrt{\eta^2 + L_g^2}} \|\mathbf{g}\|.$$

Let $\omega = \eta / \sqrt{\eta^2 + L_g^2}$. Proceeding similarly to the SOL case but using (21), we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho \alpha_k \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{r}_k^{\mathcal{I}} \rangle \\
&\leq -\rho \alpha_k \|\mathbf{r}_k^{\mathcal{I}}\|^2 \\
&\leq -\rho \min \left\{ \frac{2(1-\rho)}{L_g} \|\mathbf{r}_k^{\mathcal{I}}\|^2, \delta_k \|\mathbf{r}_k^{\mathcal{I}}\| \right\} \\
&\leq -\rho \min \left\{ \frac{2(1-\rho)\omega^2}{L_g} \|\mathbf{g}_k^{\mathcal{I}}\|^2, \delta_k \omega \|\mathbf{g}_k^{\mathcal{I}}\| \right\} \\
&< -\rho \min \left\{ \frac{2(1-\rho)\omega^2}{L_g} \epsilon_k^4, \delta_k \omega \epsilon_k^2 \right\},
\end{aligned}$$

again, making use of $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^2$ in the final line. □

The following lemma covers the case when the inactive set termination condition is satisfied, that is, $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$ or (7c) holds. In this case, we expect the inactive set step to be small (cf. (20)) and so we analyse the decrease due to the active set portion of the step, using the fact that at lease one of the active set termination conditions (7a) or (7b) must be unsatisfied.

Lemma B.4 (Sufficient Decrease: Active Set Case)**.** *Suppose that f satisfies Assumption 3.1. Let $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$ be the update computed at iteration k of Algorithm 1, where α_k satisfies the line search criterion (11). Suppose that at least one of the active set termination conditions, (7a) or (7b), is not satisfied. If $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, then*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min \left\{ \frac{1}{2}, \frac{2(1-\rho)}{L_g} \min \left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2.$$

However, if $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and (7c) is satisfied, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min \left\{ \frac{1}{2}, \min \left\{ \frac{2(1-\rho)}{L_g}, \frac{\delta_k}{\epsilon_k^2} \right\} \min\{1, \varsigma\} \min \left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2.$$

Proof. Since α_k satisfies the line search criterion we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha_k \mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle + \alpha_k \rho \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle \\ &\leq \rho\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} + \alpha_k \mathbf{p}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle, \end{aligned}$$

where we apply $\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle \leq 0$. From here the proof proceeds similarly to Lemma C.3. Indeed, the if (7a) or (7b) are unsatisfied, (36) gives

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{1}{2}, \alpha_k \min\left\{1, \frac{\epsilon_k^2}{2\delta_k^2}\right\}\right\}\epsilon_k^2, \tag{28}$$

and it only remains to apply a bound on α_k. If $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, we use (27) to obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{1}{2}, \frac{2(1-\rho)}{L_g} \min\left\{1, \frac{\epsilon_k^2}{2\delta_k^2}\right\}\right\}\epsilon_k^2.$$

Otherwise, we have $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$. In this case, we must lower bound $\delta_k / \|\mathbf{p}_k^{\mathcal{I}}\|$ in (26). We therefore use (20) with $\varrho = \varsigma$, (22), as well as the fact that (7c) is unsatisfied to obtain

$$\min\{\varsigma, 1\}\|\mathbf{p}_k^{\mathcal{I}}\| \leq \|\mathbf{g}_k\| \leq \epsilon_k^2 \implies \frac{\delta_k \min\{1, \varsigma\}}{\epsilon_k^2} \leq \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|}.$$

We now apply this bound to (26) and combine with (28) to get

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{1}{2}, \min\left\{\frac{2(1-\rho)}{L_g}, \frac{\delta_k}{\epsilon_k^2}\right\}\min\{1, \varsigma\} \min\left\{1, \frac{\epsilon_k^2}{2\delta_k^2}\right\}\right\}\epsilon_k^2.$$

□

Proof of Theorem 3.3. We posit that the algorithm must terminate in at most

$$K = \left\lceil \frac{(f_0 - f_*)\epsilon_g^{-2}}{\min\{c_1, c_2\}} \right\rceil,$$

iterations, where

$$\begin{aligned} c_1 &\triangleq \rho \min\left\{\frac{2\varsigma(1-\rho)\min\{1,\varsigma\}C_{\varsigma,L_g}^2}{L_g}, \varsigma C_{\varsigma,L_g}, \frac{2(1-\rho)\omega^2}{L_g}, \omega\right\}, \quad \text{with} \quad \omega \triangleq \frac{\eta}{\sqrt{\eta^2 + L_g^2}}, \\ c_2 &\triangleq \rho \min\left\{\frac{1}{2}, \frac{1}{2}\min\left\{\frac{2(1-\rho)}{L_g}, 1\right\}\min\{1, \varsigma\}\right\}, \end{aligned}$$

and C_{ς,L_g} is as in (20). Suppose otherwise, that is, the algorithm fails to terminate until at least iteration $K+1$. For iterations $k = 1, \ldots, K$, the termination conditions must be unsatisfied. We divide the iterates up in the following manner

$$\mathcal{K}_1 = \{k \in [K] \mid \mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset, \ \|\mathbf{g}_k^{\mathcal{I}}\| \geq \epsilon_g\},$$

and

$$\mathcal{K}_2 = \{k \in [K] \setminus \mathcal{K}_1 \mid \mathcal{A}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset, \ (\exists i \in \mathcal{A}(\mathbf{x}_k, \epsilon_g^{1/2}), \ \mathbf{g}_k^i < -\sqrt{\epsilon_g} \text{ or } \|\mathrm{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\| \geq \epsilon_g)\}.$$

Since the algorithm has not terminated, $[K] = \mathcal{K}_1 \cup \mathcal{K}_2$. If $k \in \mathcal{K}_1$ we apply Lemma B.3 and combine the SOL and NPC cases with $\epsilon_g < 1$ to obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{2\varsigma(1-\rho)\min\{1,\varsigma\}C_{\varsigma,L_g}^2}{L_g}, \varsigma C_{\varsigma,L_g}, \frac{2(1-\rho)\omega^2}{L_g}, \omega\right\}\epsilon_g^2 = -c_1\epsilon_g^2.$$

If $k \in \mathcal{K}_2$, we instead combine the results of Lemma B.4 to obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{1}{2}, \frac{1}{2}\min\left\{\frac{2(1-\rho)}{L_g}, 1\right\}\min\{1, \varsigma\}\right\}\epsilon_g \leq -c_2\epsilon_g^2.$$

Finally, we obtain

$$\begin{aligned} f_0 - f_* \geq f_0 - f(\mathbf{x}_K) &= \sum_{k=0}^{K-1} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) > |\mathcal{K}_1|c_1\epsilon_g^2 + |\mathcal{K}_2|c_2\epsilon_g^2 \\ &\geq (|\mathcal{K}_1| + |\mathcal{K}_2|)\min\{c_1, c_2\}\epsilon_g^2 = K\min\{c_1, c_2\}\epsilon_g^2, \end{aligned}$$

which contradicts the definition of K. □

C. Global Convergence - Improved Rate

In this section, we provide the proof of Theorem 3.8. Recall that we denote the update to $\mathbf{x}_k$ for some step size, α, by

$$\mathbf{x}_k(\alpha) = \mathcal{P}(\mathbf{x}_k + \alpha\mathbf{p}_k).$$

Recall that Algorithm 2 involves two types of steps: *Type I* and *Type II*. We summarise the step types, the optimality conditions, as well as the corresponding lemmas in Table 1.

Table 1. The step types, the optimality conditions, as well as the corresponding lemmas involved in the proof of Theorem 3.8.

Type	Termination condition	Active Step	Inactive Step	Step size	Sufficient Decrease
I	$\mathcal{A} \neq \emptyset$ and (not (7a) or not (7b))	Gradient	Newton-MR	Lemma C.1	Lemmas C.2 and C.3
II	($\mathcal{A} = \emptyset$ or ((7a) and (7b))) and ($\mathcal{I} \neq \emptyset$ and (not (7c)))	None	Newton-MR	Lemma C.4	Lemmas C.5 and C.6

Our first three lemmas (Lemmas C.1 to C.3) will demonstrate that *Type I* steps produce sufficient decrease in the function value. The analysis of *Type I* steps builds off of Xie & Wright (2023) which demonstrated that projected gradient can achieve good progress (in terms of guaranteed decrease) when the active termination conditions (7a) and (7b) are unsatisfied. However, unlike Xie & Wright (2023), which only uses a first-order step, we also incorporate second-order update in the form of Newton-MR step in the inactive set of indices.

As shown in Lemma C.1, combining the steps in this manner suggests that the lower bound on the step size may depend inversely on the length of the Newton-MR step. This, in turn, could lead to small step sizes, if the Newton-MR step is large. We deal with this issue by splitting our analysis into two cases. The first case (Lemma C.2) deals with large gradients on the inactive set where we expect good progress due to the corresponding large Newton-MR step on the inactive set (cf. (20)). By contrast, the second case (Lemma C.3) deals with small gradients on the inactive set where we can expect to see small inactive set steps (cf. (20)) and therefore lower bounded step sizes. In this way, we trade off the convergence due to the inactive and active sets to always ensure sufficient decrease at the required rate.

Recall that Assumption 3.1 implies that, for any $\mathbf{y}, \mathbf{x} \in \mathbb{R}^d_+$, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L_g}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{29}$$

We now give the proof of Lemmas C.1 to C.3.

Lemma C.1 (*Type I* Step: Step-size Lower Bound)**.** *Assume that f satisfies Assumptions 3.1 and 3.7. Suppose a Type I step is taken at iteration k of Algorithm 2. If $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$,then the largest step size which satisfies the line search criteria (11), α_k, satisfies the following lower bound*

$$\alpha_k \geq \min\left\{\frac{2(1-\rho)}{L_g}\min\{1, \sigma\}, \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|}\right\}. \tag{30}$$

However, if $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, then

$$\alpha_k \geq \frac{2(1-\rho)}{L_g}. \tag{31}$$

Proof. If $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$, suppose

$$\alpha \leq \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|} \leq \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|_\infty},$$

so that for each $i \in \mathcal{I}(\mathbf{x}_k, \delta_k)$ we have $\mathcal{P}(\mathbf{x}_k^i + \alpha \mathbf{p}_k^i) = \mathbf{x}_k^i + \alpha \mathbf{p}_k^i$. The Lipschitz gradient upper bound (29) yields

$$\begin{aligned} f(\mathbf{x}_k(\alpha)) &\leq f(\mathbf{x}_k) + \langle \mathbf{g}_k, \mathcal{P}(\mathbf{x}_k + \alpha \mathbf{p}_k) - \mathbf{x}_k \rangle + \frac{L_g}{2} \| P(\mathbf{x}_k + \alpha \mathbf{p}_k) - \mathbf{x}_k \|^2 \\ &= f(\mathbf{x}_k) + \langle \mathbf{g}_k^{\mathcal{A}}, P(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle + \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle + \frac{L_g}{2} \| \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \|^2 + \frac{L_g \alpha^2}{2} \| \mathbf{p}_k^{\mathcal{I}} \|^2. \end{aligned}$$

It is clear from this bound that the line search will terminate for any α such that

$$\begin{aligned} \langle \mathbf{g}_k^{\mathcal{A}}, P(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle + \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle + \frac{L_g}{2} \| \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \|^2 + \frac{L_g \alpha^2}{2} \| \mathbf{p}_k^{\mathcal{I}} \|^2 \\ - \rho \left(\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle + \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle \right), \end{aligned} \tag{32}$$

is nonpositive. Starting with the active set terms of (32). We use the projection inequality $\|\mathcal{P}(\mathbf{x}) - \mathcal{P}(\mathbf{y})\|^2 \leq \langle \mathbf{x} - \mathbf{y}, \mathcal{P}(\mathbf{x}) - \mathcal{P}(\mathbf{y}) \rangle$ combined with the feasibility of $\mathbf{x}_k^{\mathcal{A}}$ (which implies $\mathcal{P}(\mathbf{x}_k^{\mathcal{A}}) = \mathbf{x}_k^{\mathcal{A}}$) to obtain

$$\begin{aligned} &(1-\rho) \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle + \frac{L_g}{2} \| P(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \|^2 \\ &\leq (1-\rho) \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle - \frac{\alpha L_g}{2} \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle \\ &\leq \left((1-\rho) - \frac{\alpha L_g}{2} \right) \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle. \end{aligned}$$

By $\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle \leq 0$, the active terms of (32) are nonpositive if

$$(1-\rho) - \frac{\alpha L_g}{2} \geq 0 \implies \alpha \leq \frac{2(1-\rho)}{L_g}.$$

If $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, then (31) follows directly from this bound.

Now we consider the inactive terms of (32). If $\text{D}_{\text{type}} = \text{SOL}$, i.e., $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$, we apply (19) and Assumption 3.7 to obtain

$$\begin{aligned} \alpha(1-\rho) \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}} \rangle + \frac{L_g \alpha^2}{2} \| \mathbf{s}_k^{\mathcal{I}} \|^2 &\leq \alpha \left(-(1-\rho) \langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k \mathbf{s}_k^{\mathcal{I}} \rangle + \frac{\alpha L_g}{2} \| \mathbf{s}_k^{\mathcal{I}} \|^2 \right) \\ &\leq \alpha \left(-(1-\rho) \sigma \| \mathbf{s}_k^{\mathcal{I}} \|^2 + \frac{\alpha L_g}{2} \| \mathbf{s}_k^{\mathcal{I}} \|^2 \right) \\ &= \alpha \left(-(1-\rho) \sigma + \frac{\alpha L_g}{2} \right) \| \mathbf{s}_k^{\mathcal{I}} \|^2. \end{aligned}$$

This upper bound will be negative for any step size satisfying

$$\alpha \leq \frac{2\sigma(1-\rho)}{L_g}.$$

If $\text{D}_{\text{type}} = \text{NPC}$, i.e., $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$, we apply (21) to obtain

$$\begin{aligned} \alpha(1-\rho) \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{r}_k^{\mathcal{I}} \rangle + \frac{L_g \alpha^2}{2} \| \mathbf{r}_k^{\mathcal{I}} \|^2 &\leq -\alpha(1-\rho) \| \mathbf{r}_k^{\mathcal{I}} \|^2 + \frac{L_g \alpha}{2} \| \mathbf{r}_k^{\mathcal{I}} \|^2 \\ &= \alpha \left(-(1-\rho) + \frac{L_g \alpha}{2} \right) \| \mathbf{r}_k^{\mathcal{I}} \|^2, \end{aligned}$$

which is negative when

$$\alpha \le \frac{2(1-\rho)}{L_g}.$$

If both the inactive and active terms of (32) are nonpositive then the line search will certainly terminate. Collecting the bounds on the step size, we can see that the largest α_k which satisfies the line search criteria also satisfies the following lower bound

$$\alpha_k \ge \min\left\{\frac{2(1-\rho)}{L_g}\min\{1,\sigma\}, \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|}\right\}.$$

□

Lemma C.2 (*Type I* Step: Inactive Set Decrease)**.** *Assume that f satisfies Assumptions 3.1, 3.6 and 3.7. Suppose that a Type I step is taken at iteration k of Algorithm 2 but both $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^{3/2}$. Let α_k be the largest step size satisfying the line search condition (11) so that $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$. If D_{type} = SOL then*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\sigma \min\left\{\frac{2(1-\rho)\min\{1,\sigma\}C_{\sigma,L_g}^2}{L_g}\epsilon_k^3, C_{\sigma,L_g}\delta_k\epsilon_k^{3/2}\right\}.$$

Otherwise, if D_{type} = NPC,

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\frac{2(1-\rho)\omega^2}{L_g}\epsilon_k^3, \omega\delta_k\epsilon_k^{3/2}\right\}.$$

Proof. Line search criterion and the negativity of $\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha_k \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle$ implies

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \le \rho\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha_k \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}}\rangle + \alpha_k\rho\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle \le \rho\alpha_k\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}}\rangle. \tag{33}$$

We now divide into two cases, depending on the step type selected by MINRES.

If $\text{D}_{\text{type}} = \text{SOL}$, then $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$. Using the line search condition (33), (19), Assumption 3.7, the lower bound on α_k from Lemma C.1, and the left-hand-side inequality of (20) with $\varrho = \sigma$, we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\le \rho\alpha_k\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle \\
&\le -\rho\alpha_k\langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{s}_k^{\mathcal{I}}\rangle \\
&\le -\rho\sigma\alpha_k\|\mathbf{s}_k^{\mathcal{I}}\|^2 \\
&\le -\rho\sigma\min\left\{\frac{2(1-\rho)}{L_g}\min\{1,\sigma\}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|}\right\}\|\mathbf{s}_k^{\mathcal{I}}\|^2 \\
&\le -\rho\sigma\min\left\{\frac{2(1-\rho)}{L_g}\min\{1,\sigma\}\|\mathbf{s}_k^{\mathcal{I}}\|^2, \delta_k\|\mathbf{s}_k^{\mathcal{I}}\|\right\} \\
&\le -\rho\sigma\min\left\{\frac{2(1-\rho)\min\{1,\sigma\}C_{\sigma,L_g}^2}{L_g}\|\mathbf{g}_k^{\mathcal{I}}\|^2, C_{\sigma,L_g}\delta_k\|\mathbf{g}_k^{\mathcal{I}}\|\right\} \\
&< -\rho\sigma\min\left\{\frac{2(1-\rho)\min\{1,\sigma\}C_{\sigma,L_g}^2}{L_g}\epsilon_k^3, C_{\sigma,L_g}\delta_k\epsilon_k^{3/2}\right\},
\end{aligned}$$

where for the last inequality, we used the fact that $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^{3/2}$.

If $\text{D}_{\text{type}} = \text{NPC}$, then $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$. We use (33), but apply (21) and Assumption 3.6 to get

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho\alpha_k \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{r}_k^{\mathcal{I}} \rangle \\ &\leq -\rho\alpha_k \|\mathbf{r}_k^{\mathcal{I}}\|^2 \\ &\leq -\rho \min\left\{ \frac{2(1-\rho)}{L_g} \|\mathbf{r}_k^{\mathcal{I}}\|^2, \delta_k \|\mathbf{r}_k^{\mathcal{I}}\| \right\} \\ &\leq -\rho \min\left\{ \frac{2(1-\rho)\omega^2}{L_g} \|\mathbf{g}_k^{\mathcal{I}}\|^2, \delta_k \omega \|\mathbf{g}_k^{\mathcal{I}}\| \right\} \\ &< -\rho \min\left\{ \frac{2(1-\rho)\omega^2}{L_g} \epsilon_k^3, \delta_k \omega \epsilon_k^{3/2} \right\}, \end{aligned}$$

again, making use of $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^{3/2}$ in the final line. □

Lemma C.3 (*Type I* Step: Sufficient Reduction)**.** *Assume that f satisfies Assumptions 3.1 and 3.7. Suppose that a Type I step is taken on iteration k of Algorithm 2 so that $\mathcal{A}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and either (7a) or (7b) is unsatisfied. Let α_k be the largest step size satisfying the line search condition (11) so that $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$. If $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and $\|\mathbf{g}_k^{\mathcal{I}}\| \leq \epsilon_k^{3/2}$, then*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{ \frac{1}{2}, \min\{1, \sigma\} \min\left\{ \frac{2(1-\rho)}{L_g}, \frac{\delta_k}{\epsilon_k^{3/2}} \right\} \min\left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2.$$

Otherwise, if $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$,

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{ \frac{1}{2}, \frac{2(1-\rho)}{L_g} \min\left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2.$$

Proof. Since α_k satisfies the line search sufficient decrease condition, the negativity of $\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle$, implied by (19) and (21), gives

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho \left(\langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha_k \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle + \alpha_k \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle \right) \\ &\leq \rho \langle \mathbf{g}_k^{\mathcal{A}}, \mathcal{P}(\mathbf{x}_k^{\mathcal{A}} - \alpha \mathbf{g}_k^{\mathcal{A}}) - \mathbf{x}_k^{\mathcal{A}} \rangle \\ &= \rho \sum_{i \in \mathcal{A}(\mathbf{x}_k, \delta_k)} \mathbf{g}_k^i (\mathcal{P}(\mathbf{x}_k^i - \alpha \mathbf{g}_k^i) - \mathbf{x}_k^i). \end{aligned} \tag{34}$$

The analysis proceeds depending on which optimality condition is unsatisfied.

Case 1 (7a): $\mathbf{g}_k^i < -\epsilon_k$ for some $i \in \mathcal{A}(\mathbf{x}_k, \delta_k)$. In this case we can see that

$$\mathbf{g}_k^i (\mathcal{P}(\mathbf{x}_k^i - \alpha_k \mathbf{g}_k^i) - \mathbf{x}_k^i) = -\alpha_k (\mathbf{g}_k^i)^2 < -\alpha_k \epsilon_k^2.$$

We immediately see from the term wise nonpositivity of (34) that

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\alpha_k \epsilon_k^2.$$

Case 2 (7b): Continuing from (34) we obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \rho \sum_{i \in \mathcal{A}(\mathbf{x}_k, \delta_k)} \mathbf{g}_k^i (\mathcal{P}(\mathbf{x}_k^i - \alpha \mathbf{g}_k^i) - \mathbf{x}_k^i) \\ &= \rho \left(\sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha_k \mathbf{g}_k^i \geq \mathbf{x}_k^i}} -\mathbf{g}_k^i \mathbf{x}_k^i + \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha_k \mathbf{g}_k^i < \mathbf{x}_k^i}} -\alpha_k (\mathbf{g}_k^i)^2 \right). \end{aligned} \tag{35}$$

Note each sum in (35) is *term-wise* negative. Since $\|\text{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\| > \epsilon_k^2$, we have

$$\epsilon_k^4 < \|\text{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\|^2 = \left(\sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha_k \mathbf{g}_k^i \geq \mathbf{x}_k^i}} (\mathbf{g}_k^i \mathbf{x}_k^i)^2 + \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha_k \mathbf{g}_k^i < \mathbf{x}_k^i}} (\mathbf{x}_k^i \mathbf{g}_k^i)^2 \right).$$

This implies two possible cases: either

$$\frac{\epsilon_k^4}{2} < \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha \mathbf{g}_k^i \geq \mathbf{x}_k^i}} (\mathbf{x}_k^i \mathbf{g}_k^i)^2 \implies \frac{\epsilon_k^2}{2} < \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha \mathbf{g}_k^i \geq \mathbf{x}_k^i}} \mathbf{x}_k^i \mathbf{g}_k^i,$$

or

$$\frac{\epsilon_k^4}{2} < \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha \mathbf{g}_k^i < \mathbf{x}_k^i}} (\mathbf{x}_k^i \mathbf{g}_k^i)^2 \leq \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha \mathbf{g}_k^i < \mathbf{x}_k^i}} (\delta_k \mathbf{g}_k^i)^2 \implies \frac{\epsilon_k^4}{2\delta_k^2} < \sum_{\substack{i \in \mathcal{A}(\mathbf{x}_k, \delta_k) \\ \alpha \mathbf{g}_k^i < \mathbf{x}_k^i}} (\mathbf{g}_k^i)^2.$$

In either case, the negativity of each term of (35) implies

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min \left\{ \frac{\epsilon_k^2}{2}, \frac{\alpha_k \epsilon_k^4}{2\delta_k^2} \right\}.$$

Combining with Case 1 gives

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &< -\rho \min \left\{ \frac{\epsilon_k^2}{2}, \alpha_k \epsilon_k^2, \frac{\alpha_k \epsilon_k^4}{2\delta_k^2} \right\} \\ &= -\rho \min \left\{ \frac{1}{2}, \alpha_k, \frac{\alpha_k \epsilon_k^2}{2\delta_k^2} \right\} \epsilon_k^2. \end{aligned} \tag{36}$$

If $\mathcal{I}(\mathbf{x}_k, \delta_k) = \emptyset$, we apply (31) to obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min \left\{ \frac{1}{2}, \frac{2(1-\rho)}{L_g} \min \left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2.$$

On the other hand, if $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$, the lower bound for α_k in (30) depends inversely on the inactive portion of the step $\|\mathbf{p}_k^{\mathcal{I}}\|$. The step size can therefore become small if $\|\mathbf{p}_k^{\mathcal{I}}\|$ is too large. To avoid this, we will make use of the fact that the gradient is bounded. In particular, by combining the right inequality of (20) and (22), we obtain

$$\min\{1, \sigma\} \|\mathbf{p}_k^{\mathcal{I}}\| \leq \|\mathbf{g}_k^{\mathcal{I}}\| \leq \epsilon_k^{3/2},$$

which implies

$$\frac{\delta_k \min\{1, \sigma\}}{\epsilon_k^{3/2}} \leq \frac{\delta_k}{\|\mathbf{p}_k^{\mathcal{I}}\|}.$$

Imposing this on the step size lower bound (30) gives

$$\alpha_k \geq \min\{1, \sigma\} \min \left\{ \frac{2(1-\rho)}{L_g}, \frac{\delta_k}{\epsilon_k^{3/2}} \right\}.$$

The decrease is therefore given by

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &< -\rho \min \left\{ \frac{1}{2}, \alpha_k, \frac{\alpha_k \epsilon_k^2}{2\delta_k^2} \right\} \epsilon_k^2 \\ &\leq -\rho \min \left\{ \frac{1}{2}, \alpha_k \min \left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2 \\ &\leq -\rho \min \left\{ \frac{1}{2}, \min\{1, \sigma\} \min \left\{ \frac{2(1-\rho)}{L_g}, \frac{\delta_k}{\epsilon_k^{3/2}} \right\} \min \left\{ 1, \frac{\epsilon_k^2}{2\delta_k^2} \right\} \right\} \epsilon_k^2. \end{aligned}$$

□

The next three lemmas (Lemmas C.4 to C.6) demonstrate the sufficient decrease of *Type II* steps. Recall that a *Type II* steps occurs once active set optimality is reached. *Type II* steps are taken until the inactive set optimality (7c) is satisfied (termination) or a new index falls into the active set and disrupts active set optimality, in which case we resume *Type I* steps. The *Type II* step consists of only a Newton-MR step in the inactive indices (no step is taken in the active indices). Indeed, a *Type II* direction can be written (with possible reordering of indices) as

$$\mathbf{x}_k(\alpha) - \mathbf{x}_k = \begin{pmatrix} 0 \\ \mathcal{P}(\mathbf{x}_k^{\mathcal{I}} + \alpha \mathbf{p}_k^{\mathcal{I}}) - \mathbf{x}_k^{\mathcal{I}} \end{pmatrix}.$$

Eliminating the active portion of the step allows us to leverage a "second-order analysis" of the inactive indices without having to account for the curvature of the projected gradient portion of the step. Indeed, the analysis of the algorithm reverts to essentially that of unconstrained Newton-MR (Liu & Roosta, 2022b), with some minor modifications to account for the projection. Specifically, with possible reordering of the indices, we partition the Hessian into four blocks as

$$\mathbf{H}_k = \begin{pmatrix} \mathbf{H}_k^{\mathcal{A}} & \mathbf{H}_k^{O} \\ \mathbf{H}_k^{O} & \mathbf{H}_k^{\mathcal{I}} \end{pmatrix},$$

where $\mathbf{H}_k^{\mathcal{A}}$ and $\mathbf{H}_k^{\mathcal{I}}$ are the sub matrices corresponding to the active and inactive indices respectively and $\mathbf{H}_k^{O}$ is the remaining off diagonal blocks of the Hessian. Under the Lipschitz Hessian condition (Assumption 3.5) and using $\alpha \leq \delta_k / \|\mathbf{p}_k^{\mathcal{I}}\|$ so that $\mathcal{P}(\mathbf{x}_k^{\mathcal{I}} + \alpha \mathbf{p}_k^{\mathcal{I}}) = \mathbf{x}_k^{\mathcal{I}} + \alpha \mathbf{p}_k^{\mathcal{I}}$, we can write

$$\begin{aligned} f(\mathbf{x}_k(\alpha)) &\leq f(\mathbf{x}_k) + \left\langle \begin{pmatrix} \mathbf{g}_k^{\mathcal{A}} \\ \mathbf{g}_k^{\mathcal{I}} \end{pmatrix}, \begin{pmatrix} 0 \\ \alpha \mathbf{p}_k^{\mathcal{I}} \end{pmatrix} \right\rangle + \frac{1}{2} \left\langle \begin{pmatrix} 0 \\ \alpha \mathbf{p}_k^{\mathcal{I}} \end{pmatrix}, \begin{pmatrix} \mathbf{H}_k^{\mathcal{A}} & \mathbf{H}_k^{O} \\ \mathbf{H}_k^{O} & \mathbf{H}_k^{\mathcal{I}} \end{pmatrix} \begin{pmatrix} 0 \\ \alpha \mathbf{p}_k^{\mathcal{I}} \end{pmatrix} \right\rangle + \frac{\alpha^3 L_H}{6} \left\| \begin{pmatrix} 0 \\ \mathbf{p}_k^{\mathcal{I}} \end{pmatrix} \right\|^3 \\ &= f(\mathbf{x}_k) + \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle + \frac{\alpha^2}{2} \langle \mathbf{p}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}} \mathbf{p}_k^{\mathcal{I}} \rangle + \frac{\alpha^3 L_H}{6} \|\mathbf{p}_k^{\mathcal{I}}\|^3. \end{aligned} \tag{37}$$

Our first lemma uses the expansion in (37) to show that the largest step size satisfying the line search criterion is lower bounded.

Lemma C.4 (*Type II* Step: Step-size Lower Bound)**.** *Assume that f satisfies Assumption 3.5. If Algorithm 2 selects a Type II step at iteration k and MINRES returns D_{type} = NPC, then for the largest step size, α_k, satisfying the line search criterion* (11)*, we must have*

$$\alpha_k \geq \min \left\{ \sqrt{\frac{6(1-\rho)}{L_H \|\mathbf{r}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{r}_k^{\mathcal{I}}\|} \right\}. \tag{38}$$

Otherwise, if D_{type} = SOL and Assumption 3.7 holds, then

$$\alpha_k \geq \min \left\{ 1, \sqrt{\frac{3\sigma(1-2\rho)}{L_H \|\mathbf{s}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|} \right\}. \tag{39}$$

Proof. We have already seen that, if $\alpha \leq \delta_k / \|\mathbf{p}_k^{\mathcal{I}}\|$, (37) holds. From (11), the line search is satisfied for any α such that

$$f(\mathbf{x}_k(\alpha)) - f(\mathbf{x}_k) - \rho \alpha \langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{p}_k^{\mathcal{I}} \rangle \leq 0.$$

We now consider $\mathrm{D}_{\text{type}} = \text{SOL}$ and $\mathrm{D}_{\text{type}} = \text{NPC}$ cases. Let $\mathrm{D}_{\text{type}} = \text{SOL}$ so that $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$. Applying (37), $\alpha \leq 1$, the

MINRES curvature condition (19) and Assumption 3.7 we have

$$
\begin{aligned}
f(\mathbf{x}_k(\alpha)) - f(\mathbf{x}_k) - \rho\alpha\langle \mathbf{g}_k, \mathbf{s}_k^{\mathcal{I}}\rangle &\leq \alpha\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha^2}{2}\langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 - \rho\alpha\langle \mathbf{g}_k, \mathbf{s}_k^{\mathcal{I}}\rangle \\
&\leq \alpha(1-\rho)\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha}{2}\langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 \\
&= \alpha\left(\frac{1}{2} - \rho\right)\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha}{2}(\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle + \langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{s}_k^{\mathcal{I}}\rangle) + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 \\
&\leq \alpha\left(\frac{1}{2} - \rho\right)\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 \\
&\leq -\alpha\left(\frac{1}{2} - \rho\right)\langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{s}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 \\
&\leq -\alpha\left(\frac{1}{2} - \rho\right)\sigma\|\mathbf{s}_k^{\mathcal{I}}\|^2 + \frac{\alpha^3 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|^3 \\
&= \alpha\left(-\left(\frac{1}{2} - \rho\right)\sigma + \frac{\alpha^2 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\|\right)\|\mathbf{s}_k^{\mathcal{I}}\|^2.
\end{aligned}
$$

It can be seen that this upper bound is nonpositive if

$$
-\left(\frac{1}{2} - \rho\right)\sigma + \frac{\alpha^2 L_H}{6}\|\mathbf{s}_k^{\mathcal{I}}\| \leq 0 \implies \alpha \leq \sqrt{\frac{3\sigma(1-2\rho)}{L_H\|\mathbf{s}_k^{\mathcal{I}}\|}}.
$$

Collecting the bounds on α, the largest step size that satisfies the line search condition can be lower bounded as

$$
\alpha_k \geq \min\left\{1, \sqrt{\frac{3\sigma(1-2\rho)}{L_H\|\mathbf{s}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|}\right\}.
$$

Now let $\mathrm{D}_{\text{type}} = \text{NPC}$ so that $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$. Applying the negative curvature of $\mathbf{r}_k^{\mathcal{I}}$, (21) and (37)

$$
\begin{aligned}
f(\mathbf{x}_k(\alpha)) - f(\mathbf{x}_k) - \rho\alpha\langle \mathbf{g}_k, \mathbf{r}_k^{\mathcal{I}}\rangle &\leq \alpha\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{r}_k^{\mathcal{I}}\rangle + \frac{\alpha^2}{2}\langle \mathbf{r}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}}\mathbf{r}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{r}_k^{\mathcal{I}}\|^3 - \rho\alpha\langle \mathbf{g}_k, \mathbf{r}_k^{\mathcal{I}}\rangle \\
&\leq \alpha(1-\rho)\langle \mathbf{g}_k^{\mathcal{I}}, \mathbf{r}_k^{\mathcal{I}}\rangle + \frac{\alpha^3 L_H}{6}\|\mathbf{r}_k^{\mathcal{I}}\|^3 \\
&\leq -\alpha(1-\rho)\|\mathbf{r}_k^{\mathcal{I}}\|^2 + \frac{\alpha^3 L_H}{6}\|\mathbf{r}_k^{\mathcal{I}}\|^3 \\
&= \alpha\left(-(1-\rho) + \frac{\alpha^2 L_H}{6}\|\mathbf{r}_k^{\mathcal{I}}\|\right)\|\mathbf{r}_k^{\mathcal{I}}\|^2.
\end{aligned}
$$

This upper bound is nonpositive if

$$
-(1-\rho) + \frac{\alpha^2 L_H}{6}\|\mathbf{r}_k^{\mathcal{I}}\| \leq 0 \implies \alpha \leq \sqrt{\frac{6(1-\rho)}{L_H\|\mathbf{r}_k^{\mathcal{I}}\|}}.
$$

Therefore the largest step size that satisfies the line search condition, in the NPC case, is lower bounded as

$$
\alpha_k \geq \min\left\{\sqrt{\frac{6(1-\rho)}{L_H\|\mathbf{r}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{r}_k^{\mathcal{I}}\|}\right\}.
$$

□

From Lemma C.4 we can see that, for a judicious choice of δ_k, the step size is inversely scaling with the step length, except for the $\alpha_k = 1$ in $\mathrm{D}_{\text{type}} = \text{SOL}$ case. This inverse scaling is key to obtaining an improved rate. We therefore deal with the $\alpha_k = 1$ case separately. Indeed, in Lemma C.5 we show that if $\alpha_k = 1$ with $\mathrm{D}_{\text{type}} = \text{SOL}$ the step length must be lower bounded by norm of the gradient of the next iterate (over the same inactive set). This lemma is similar to the result in Liu & Roosta (2022b, Lemma 7), we include it for completeness.

Lemma C.5. *Suppose Algorithm 2 selects a Type II step at iteration k with $D_{type} = SOL$ and $\alpha_k = 1$, that is, an update of the form*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \begin{pmatrix} 0 \\ \mathbf{s}_k^{\mathcal{I}} \end{pmatrix},$$

with possible reordering. Under Assumptions 3.1, 3.5 and 3.7, we have

$$\|\mathbf{s}_k^{\mathcal{I}_k}\| \geq c_0 \min\left\{ \left\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\right\| / \epsilon_k, \epsilon_k \right\},$$

where

$$c_0 \triangleq \frac{2\sigma}{\theta L_g + \sqrt{\theta^2 L_g^2 + 2 L_H \sigma^2}}.$$

Proof. Since $(\mathbf{r}_k^{\mathcal{I}})^{(t-1)} = -\mathbf{H}_k^{\mathcal{I}} (\mathbf{s}_k^{\mathcal{I}})^{(t-1)} - \mathbf{g}_k^{\mathcal{I}} \in \mathcal{K}_t(\mathbf{H}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}})$ and NPC has not been detected, Assumption 3.7 implies

$$\sigma \|\mathbf{r}_k^{\mathcal{I}}\|^2 \leq \langle \mathbf{r}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}} \mathbf{r}_k^{\mathcal{I}} \rangle \leq \|\mathbf{r}_k^{\mathcal{I}}\| \|\mathbf{H}_k^{\mathcal{I}} \mathbf{r}_k^{\mathcal{I}}\| \implies \|\mathbf{r}_k^{\mathcal{I}}\| \leq \frac{\|\mathbf{H}_k^{\mathcal{I}} \mathbf{r}_k^{\mathcal{I}}\|}{\sigma}. \tag{40}$$

For clarity, in the sequel we make the dependence of inactive set on the iteration explicit. Consider

$$\mathbf{g}_{k+1}^{\mathcal{I}_k} = \left(\frac{\partial f(\mathbf{x}_{k+1})}{\partial \mathbf{x}^i} \mid i \in \mathcal{I}(\mathbf{x}_k, \delta_k) \right),$$

that is, the indices of the gradient evaluated at $\mathbf{x}_{k+1}$ corresponding to the inactive set at $\mathbf{x}_k$. This portion of the *next* gradient "lives" in the same subset of the indices as $\mathbf{g}_k^{\mathcal{I}_k}$. The mean value theorem therefore implies that

$$\mathbf{g}_{k+1}^{\mathcal{I}_k} - \mathbf{g}_k^{\mathcal{I}_k} - \mathbf{H}_k^{\mathcal{I}_k} \mathbf{s}_k^{\mathcal{I}_k} = \int_0^1 \left(\mathbf{H}\left(\mathbf{x}_k + t \begin{pmatrix} 0 \\ \mathbf{s}_k^{\mathcal{I}} \end{pmatrix} \right)^{\mathcal{I}_k} - \mathbf{H}_k^{\mathcal{I}_k} \right) \mathbf{s}_k^{\mathcal{I}_k} \, dt.$$

Assumption 3.5 implies

$$\left\| \mathbf{g}_{k+1}^{\mathcal{I}_k} - \mathbf{g}_k^{\mathcal{I}_k} - \mathbf{H}_k^{\mathcal{I}_k} \mathbf{s}_k^{\mathcal{I}_k} \right\| \leq \frac{L_H}{2} \|\mathbf{s}_k^{\mathcal{I}_k}\|.$$

Using this bound, (10), and (40), we obtain

$$\begin{aligned}
\left\| \mathbf{g}_{k+1}^{\mathcal{I}_k} \right\| &= \left\| \mathbf{g}_{k+1}^{\mathcal{I}_k} - \mathbf{g}_k^{\mathcal{I}_k} - \mathbf{H}_k^{\mathcal{I}_k} \mathbf{s}_k^{\mathcal{I}_k} - \mathbf{r}_k^{\mathcal{I}_k} \right\| \\
&\leq \left\| \mathbf{g}_{k+1}^{\mathcal{I}_k} - \mathbf{g}_k^{\mathcal{I}_k} - \mathbf{H}_k^{\mathcal{I}_k} \mathbf{s}_k^{\mathcal{I}_k} \right\| + \|\mathbf{r}_k^{\mathcal{I}_k}\| \\
&\leq \frac{L_H}{2} \|\mathbf{s}_k^{\mathcal{I}_k}\|^2 + \frac{\|\mathbf{H}_k^{\mathcal{I}_k} \mathbf{r}_k^{\mathcal{I}_k}\|}{\sigma} \\
&\leq \frac{L_H}{2} \|\mathbf{s}_k^{\mathcal{I}_k}\|^2 + \frac{\theta \epsilon_k \|\mathbf{H}_k^{\mathcal{I}_k} \mathbf{s}_k^{\mathcal{I}_k}\|}{\sigma} \\
&\leq \frac{L_H}{2} \|\mathbf{s}_k^{\mathcal{I}_k}\|^2 + \frac{\theta \epsilon_k L_g \|\mathbf{s}_k^{\mathcal{I}_k}\|}{\sigma},
\end{aligned}$$

where the second to last line follows from the MINRES termination condition in Algorithm 2 and the last line follows from Assumption 3.1. Rearranging this expression, we obtain a quadratic inequality in $\|\mathbf{s}_k^{\mathcal{I}_k}\|$ as

$$0 \leq L_H \sigma \|\mathbf{s}_k^{\mathcal{I}_k}\|^2 + 2\theta \epsilon_k L_g \|\mathbf{s}_k^{\mathcal{I}_k}\| - 2\sigma \left\| \mathbf{g}_{k+1}^{\mathcal{I}_k} \right\|.$$

We can bound $\|\mathbf{s}_k^{\mathcal{I}_k}\|$ by the positive root of this quadratic as

$$\begin{aligned}\left\|\mathbf{s}_k^{\mathcal{I}_k}\right\| &\geq \frac{-2\theta\epsilon_k L_g + \sqrt{4\theta^2\epsilon_k^2 L_g^2 + 8L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|}}{2L_H\sigma} \\ &= \left(\frac{-\theta L_g + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}}{L_H\sigma}\right)\epsilon_k \\ &= \left(\frac{\theta^2 L_g^2 - \left(\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2\right)}{L_H\sigma\left(-\eta L_g - \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}\right)}\right)\epsilon_k \\ &= \left(\frac{2\sigma\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}{L_g\theta + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}}\right)\epsilon_k.\end{aligned}$$

We now consider two cases. If $\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2 > 1$

$$\begin{aligned}\frac{2\sigma\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}{\theta L_g + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}} &= \frac{2\sigma}{\theta L_g\epsilon_k^2/\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\| + \sqrt{\theta^2 L_g^2\epsilon_k^4/\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|^2 + 2L_H\sigma^2\epsilon_k^2/\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|}} \\ &\geq \frac{2\sigma}{\theta L_g + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2}}.\end{aligned}$$

On the other hand, if $\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2 \leq 1$

$$L_g\theta + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2} \leq L_g\theta + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2}.$$

Together, these cases imply that

$$\begin{aligned}\left\|\mathbf{s}_k^{\mathcal{I}_k}\right\| &= \left(\frac{2\sigma\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}{L_g\theta + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|/\epsilon_k^2}}\right)\epsilon_k \\ &\geq \frac{2\sigma}{L_g\theta + \sqrt{\theta^2 L_g^2 + 2L_H\sigma^2}}\min\left\{\left\|\mathbf{g}_{k+1}^{\mathcal{I}_k}\right\|/\epsilon_k^2, 1\right\}\epsilon_k.\end{aligned}$$

□

We now demonstrate the sufficient decrease of the *Type II* step.

Lemma C.6 (*Type II* Step: Sufficient Decrease)**.** *Assume that f satisfies Assumptions 3.1 and 3.5. Suppose that a Type II step is taken on iteration k of Algorithm 2 (i.e., $\mathcal{I}(\mathbf{x}_k, \delta_k) \neq \emptyset$ and $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^2$). Let $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k\mathbf{p}_k)$ where α_k is the largest step size satisfying the termination condition (11) (cf. Lemma C.4). Suppose that MINRES returns D$_{type}$ $=$ SOL and Assumption 3.7 is satisfied. Then, if $\|\mathbf{g}_{k+1}^{\mathcal{I}}\| > 0$, we have*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\sigma\min\left\{\sqrt{\frac{3\sigma(1-2\rho)}{L_H}}C_{\sigma,L_g}^{3/2}\epsilon_k^3, C_{\sigma,L_g}\delta_k\epsilon_k^2, \frac{c_0^2\left\|\mathbf{g}_{k+1}^{\mathcal{I}}\right\|^2}{2\epsilon_k^2}, \frac{c_0^2\epsilon_k^2}{2}\right\}.$$

where c_0 is defined in Lemma C.5. Note that if $\|\mathbf{g}_{k+1}^{\mathcal{I}}\| = 0$ strict inequality must be replaced with "$\leq$". On the other hand, if D$_{type}$ $=$ NPC and Assumption 3.6 is satisfied, then

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho\min\left\{\sqrt{\frac{6(1-\rho)}{L_H}}\omega^{3/2}\epsilon_k^3, \omega\delta_k\epsilon_k^2\right\}.$$

Proof. If $\mathrm{D}_{\text{type}} = \text{SOL}$, $\mathbf{p}_k^{\mathcal{I}} = \mathbf{s}_k^{\mathcal{I}}$. Combining the line search sufficient decrease (11), the descent condition for the SOL step (19) and Assumption 3.7, we obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \alpha_k \rho \langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}} \rangle \\ &\leq -\alpha_k \rho \langle \mathbf{s}_k^{\mathcal{I}}, \mathbf{H}_k^{\mathcal{I}} \mathbf{s}_k^{\mathcal{I}} \rangle \\ &\leq -\alpha_k \rho \sigma \|\mathbf{s}_k^{\mathcal{I}}\|^2. \end{aligned}$$

Since $\alpha_k \leq 1$, if the step size returned by the line search satisfies $\alpha_k < 1$, then we must have

$$\min \left\{ \sqrt{\frac{3\sigma(1-2\rho)}{L_H \|\mathbf{s}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|} \right\} \leq \alpha_k,$$

as otherwise (39) would imply $\alpha_k \geq 1$. Therefore, by applying (20) with $\varrho = \sigma$, we obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -\rho\sigma \min \left\{ \sqrt{\frac{3\sigma(1-2\rho)}{L_H \|\mathbf{s}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{s}_k^{\mathcal{I}}\|} \right\} \|\mathbf{s}_k^{\mathcal{I}}\|^2 \\ &\leq -\rho\sigma \min \left\{ \sqrt{\frac{3\sigma(1-2\rho)}{L_H}} \|\mathbf{s}_k^{\mathcal{I}}\|^{3/2}, \delta_k \|\mathbf{s}_k^{\mathcal{I}}\| \right\} \\ &\leq -\rho\sigma \min \left\{ \sqrt{\frac{3\sigma(1-2\rho)}{L_H}} C_{\sigma,L_g}^{3/2} \|\mathbf{g}_k^{\mathcal{I}}\|^{3/2}, C_{\sigma,L_g} \delta_k \|\mathbf{g}_k^{\mathcal{I}}\| \right\} \\ &< -\rho\sigma \min \left\{ \sqrt{\frac{3\sigma(1-2\rho)}{L_H}} C_{\sigma,L_g}^{3/2} \epsilon_k^3, C_{\sigma,L_g} \delta_k \epsilon_k^2 \right\}, \end{aligned}$$

on the last line we use the fact that by assumption, $\|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_k^2$. If the step size $\alpha_k = 1$ is selected by the line search, we can use Lemma C.5 to obtain

$$\|\mathbf{s}_k^{\mathcal{I}}\| \geq c_0 \min \left\{ \left\|\mathbf{g}_{k+1}^{\mathcal{I}}\right\| / \epsilon_k, \epsilon_k \right\},$$

which implies

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -\rho\sigma \|\mathbf{s}_k^{\mathcal{I}}\|^2 \\ &< -\frac{\rho\sigma c_0^2}{2} \min \left\{ \left\|\mathbf{g}_{k+1}^{\mathcal{I}}\right\|^2 / \epsilon_k^2, \epsilon_k^2 \right\}. \end{aligned}$$

If $\|\mathbf{g}_{k+1}^{\mathcal{I}}\| = 0$, the strict inequality must be replaced with "$\leq$". Combining the bounds we obtain the result.

If $\mathrm{D}_{\text{type}} = \text{NPC}$, $\mathbf{p}_k^{\mathcal{I}} = \mathbf{r}_k^{\mathcal{I}}$. The line search condition (11), the step size lower bound (38), (21) and Assumption 3.6 imply

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \alpha_k \rho \langle \mathbf{r}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}} \rangle \\ &\leq -\alpha_k \rho \|\mathbf{r}_k^{\mathcal{I}}\|^2 \\ &\leq -\rho \min \left\{ \sqrt{\frac{6(1-\rho)}{L_H \|\mathbf{r}_k^{\mathcal{I}}\|}}, \frac{\delta_k}{\|\mathbf{r}_k^{\mathcal{I}}\|} \right\} \|\mathbf{r}_k^{\mathcal{I}}\|^2 \\ &\leq -\rho \min \left\{ \sqrt{\frac{6(1-\rho)}{L_H}} \|\mathbf{r}_k^{\mathcal{I}}\|^{3/2}, \delta_k \|\mathbf{r}_k^{\mathcal{I}}\| \right\} \\ &\leq -\rho \min \left\{ \sqrt{\frac{6(1-\rho)}{L_H}} \omega^{3/2} \|\mathbf{g}_k^{\mathcal{I}}\|^{3/2}, \delta_k \omega \|\mathbf{g}_k^{\mathcal{I}}\| \right\} \\ &< -\rho \min \left\{ \sqrt{\frac{6(1-\rho)}{L_H}} \omega^{3/2} \epsilon_k^3, \omega \delta_k \epsilon_k^2 \right\}, \end{aligned}$$

where the final inequality follows from the non-termination condition. □

We are finally ready to prove Theorem 3.8.

Proof of Theorem 3.8. Let $f^0 = f(\mathbf{x}_0)$. We posit that the algorithm terminates in

$$K \triangleq \left\lceil \frac{2(f^0 - f_*)\epsilon_g^{3/2}}{\min\{c_{(1,1)}, c_{(1,2,)}, c_{(2,2)}\}} + 1 \right\rceil,$$

iterations where $c_{(1,1)}$, $c_{(1,2,)}$ and $c_{(2,2)}$ are constants that will be defined later. Suppose, to the contrary, that the termination conditions is unsatisfied until at least iteration $K+1$. Then for iterations $k = 0, \dots, K$ at least one of the termination (7a) to (7c) conditions must be unsatisfied. We divide the *Type I* iterations

$$\mathcal{K}_1 = \{k \in [K] \mid \mathcal{A}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset, \quad \text{and} \quad (\exists i \in \mathcal{A}(\mathbf{x}_k, \epsilon_g^{1/2}), \quad \mathbf{g}_k^i < -\sqrt{\epsilon_g}, \quad \text{or} \quad \|\text{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\| \geq \epsilon_g)\},$$

into two sets

$$\begin{aligned}
\mathcal{K}_{1,1} &= \mathcal{K}_1 \cap \{k \in [K] \mid \|\mathbf{g}_k^{\mathcal{I}}\| > \epsilon_g^{3/4}, \quad \text{and} \quad \mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset\},\\
\mathcal{K}_{1,2} &= \mathcal{K}_1 \cap \{k \in [K] \mid \|\mathbf{g}_k^{\mathcal{I}}\| \leq \epsilon_g^{3/4}, \quad \text{or} \quad \mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) = \emptyset\}.
\end{aligned}$$

For the *Type II* iterations $\mathcal{K}_2 = [K] \setminus \mathcal{K}_1$ we have $\mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset$ and $\|\mathbf{g}_k^{\mathcal{I}_k}\| > \epsilon_g$. We divide them as follows

$$\begin{aligned}
\mathcal{K}_{2,1} &= \mathcal{K}_2 \cap \{k \in [K] \mid \mathcal{I}(\mathbf{x}_{k+1}, \epsilon_g^{1/2}) \neq \emptyset, \quad \text{and} \quad \|\mathbf{g}_{k+1}^{\mathcal{I}_{k+1}}\| > \epsilon_g\},\\
\mathcal{K}_{2,2} &= \mathcal{K}_2 \cap \{k \in [K] \mid \mathcal{I}(\mathbf{x}_{k+1}, \epsilon_g^{1/2}) = \emptyset, \quad \text{or} \quad \|\mathbf{g}_{k+1}^{\mathcal{I}_{k+1}}\| \leq \epsilon_g\}.
\end{aligned}$$

We now restate the results obtained for per-iteration decrease.

***Type I* step**. For $k \in \mathcal{K}_{1,1}$, Lemma C.2 applies and by combining the NPC and SOL cases and using $\epsilon_g < 1$ we obtain,

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\min\left\{c_{(1,1)}^a \epsilon_g^{3/2}, c_{(1,1)}^b \epsilon_g^{5/4}\right\} \leq -\min\left\{c_{(1,1)}^a, c_{(1,1)}^b\right\}\epsilon_g^{3/2} = -c_{(1,1)}\epsilon_g^{3/2}, \tag{41}$$

where

$$\begin{aligned}
c_{(1,1)}^a &\triangleq \rho \min\left\{\frac{2(1-\rho)\omega^2}{L_g}, \frac{2(1-\rho)\min\{1,\sigma\}\sigma C_{\sigma,L_g}^2}{L_g}\right\},\\
c_{(1,1)}^b &\triangleq \rho \min\left\{\omega, \sigma C_{\sigma,L_g}\right\}, \quad \text{and} \quad c_{(1,1)} \triangleq \min\{c_{(1,1)}^a, c_{(1,1)}^b\}.
\end{aligned}$$

For $k \in \mathcal{K}_{1,2}$, Lemma C.3 applies. Indeed, for $\mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) \neq \emptyset$ we obtain a decrease

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &< -\frac{\rho}{2}\min\left\{1, \min\{1,\sigma\}\min\left\{\frac{2(1-\rho)}{L_g}, \frac{1}{\epsilon_g^{1/4}}\right\}\right\}\epsilon_g\\
&\leq -\frac{\rho}{2}\min\left\{1, \min\{1,\sigma\}\min\left\{\frac{2(1-\rho)}{L_g}, 1\right\}\right\}\epsilon_g^{3/2},
\end{aligned}$$

where on the second line we used $\epsilon_g < 1$. The decrease in the case where $\mathcal{I}(\mathbf{x}_k, \epsilon_g^{1/2}) = \emptyset$ is given by

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\frac{\rho}{2}\min\left\{1, \frac{2(1-\rho)}{L_g}\right\}\epsilon_g^{3/2}.$$

Combining these results we obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -c_{(1,2)}\epsilon_g^{3/2}, \tag{42}$$

where

$$c_{(1,2)} \triangleq \frac{\rho}{2}\min\left\{1, \min\{1,\sigma\}\min\left\{\frac{2(1-\rho)}{L_g}, 1\right\}\right\}.$$

***Type II* Step**: For $k \in \mathcal{K}_{2,1}$, we can apply $\|\mathbf{g}_{k+1}^{\mathcal{I}_{k+1}}\| > \epsilon_g$ to further refine the bound for the SOL case. Note that because $\delta_k = \delta_{k+1} = \epsilon_g^{1/2}$ and a *Type II* step is taken $\mathcal{I}\left(\mathbf{x}_{k+1}, \epsilon_g^{1/2}\right) \subseteq \mathcal{I}\left(\mathbf{x}_k, \epsilon_g^{1/2}\right)$. Indeed, if $i \in \mathcal{A}\left(\mathbf{x}_k, \epsilon_g^{1/2}\right)$ we have $\mathbf{p}_k^i = 0$ and hence $\mathbf{x}_{k+1}^i = \mathbf{x}_k^i \leq \epsilon_g^{1/2}$ and $\mathcal{A}\left(\mathbf{x}_k, \epsilon_g^{1/2}\right) \subseteq \mathcal{A}\left(\mathbf{x}_{k+1}, \epsilon_g^{1/2}\right)$. Together these results imply that

$$\epsilon_g < \|\mathbf{g}_{k+1}^{\mathcal{I}_{k+1}}\| \leq \|\mathbf{g}_{k+1}^{\mathcal{I}_k}\|.$$

With $\mathrm{D_{type}} = \mathrm{SOL}$ and using $\epsilon_g < 1$, Lemma C.6 implies

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &< -\rho\sigma \min\left\{\sqrt{\frac{3\sigma(1-2\rho)C_{\sigma,L_g}^3}{L_H}}\epsilon_g^{3/2}, C_{\sigma,L_g}\epsilon_g^{3/2}, \frac{c_0^2}{2}, \frac{c_0^2\epsilon_g}{2}, \right\} \\ &\leq -\rho\sigma \min\left\{\sqrt{\frac{3\sigma(1-2\rho)C_{\sigma,L_g}^3}{L_H}}, C_{\sigma,L_g}, \frac{c_0^2}{2}\right\}\epsilon_g^{3/2}. \end{aligned}$$

With $\mathrm{D_{type}} = \mathrm{NPC}$, this becomes

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -\rho \min\left\{\sqrt{\frac{6(1-\rho)}{L_H}}\omega^{3/2}, \omega\right\}\epsilon_g^{3/2},$$

and so by combining the $\mathrm{D_{type}} = \mathrm{NPC}$ and $\mathrm{D_{type}} = \mathrm{SOL}$ cases, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < -c_{(2,2)}\epsilon_g^{3/2}, \tag{43}$$

where

$$c_{(2,2)} \triangleq \rho \min\left\{\sqrt{\frac{3\sigma^3(1-2\rho)C_{\sigma,L_g}^3}{L_H}}, \sigma C_{\sigma,L_g}, \frac{\sigma c_0^2}{2}, \sqrt{\frac{6(1-\rho)\omega^3}{L_H}}, \omega\right\}.$$

For $k \in \mathcal{K}_{2,2}$, the lower bound for the next gradient norm is no longer available. However, due to Lemma C.1, we have at least

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq 0.$$

Additionally, due to the non-termination of the algorithm, $k \in \mathcal{K}_{2,2}$ implies $k+1 \in \mathcal{K}_1$ unless $k = K$, in which case $K+1$ could be the iteration the algorithm terminates. We can therefore write

$$|\mathcal{K}_{(2,1)}| \leq |\mathcal{K}_1| + 1.$$

We now bound the total decrease in terms of the number of iterations that must have occurred using (41) to (43)

$$\begin{aligned} f^0 - f^* &\geq f^0 - f(\mathbf{x}_{K+1}) \\ &= \sum_{k=0}^{K} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \\ &= \sum_{k \in \mathcal{K}_{(1,1)}} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) + \sum_{k \in \mathcal{K}_{(1,2)}} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \\ &+ \sum_{k \in \mathcal{K}_{(2,1)}} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) + \sum_{k \in \mathcal{K}_{(2,2)}} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \\ &> \sum_{k \in \mathcal{K}_{(1,1)}} c_{(1,1)}\epsilon_g^{3/2} + \sum_{k \in \mathcal{K}_{(1,2)}} + c_{(1,2)}\epsilon_g^{3/2} + \sum_{k \in \mathcal{K}_{(2,2)}} c_{(2,2)}\epsilon_g^{3/2} \\ &= |\mathcal{K}_{(1,1)}| c_{(1,1)}\epsilon_g^{3/2} + |\mathcal{K}_{(1,2)}| c_{(1,2)}\epsilon_g^{3/2} + |\mathcal{K}_{(2,2)}| c_{(2,2)}\epsilon_g^{3/2}. \end{aligned}$$

Since each term is positive, we get

$$|\mathcal{K}_{(1,1)}| < \frac{(f^0 - f^*)\epsilon_g^{-3/2}}{c_{(1,1)}},\ |\mathcal{K}_{(1,2)}| < \frac{(f^0 - f^*)\epsilon_g^{-3/2}}{c_{(1,2)}},\ |\mathcal{K}_{(2,2)}| < \frac{(f^0 - f^*)\epsilon_g^{-3/2}}{c_{(2,2)}}.$$

Hence, if we add up the total number of iterations that must have been taken

$$\begin{aligned} K &= |\mathcal{K}_{(1,1)}| + |\mathcal{K}_{(1,2)}| + |\mathcal{K}_{(2,1)}| + |\mathcal{K}_{(2,2)}| \\ &\leq 2(|\mathcal{K}_{(1,1)}| + |\mathcal{K}_{(1,2)}|) + |\mathcal{K}_{(2,2)}| + 1 \\ &< \frac{2(f^0 - f^*)\epsilon_g^{3/2}}{c_{(1,1)}} + \frac{2(f^0 - f^*)\epsilon_g^{3/2}}{c_{(1,2)}} + \frac{(f^0 - f^*)\epsilon_g^{3/2}}{c_{(2,2)}} + 1 \\ &\leq \left\lceil \frac{2(f^0 - f^*)\epsilon_g^{3/2}}{\min\{c_{(1,1)}, c_{(1,2,)}, c_{(2,2)}\}} + 1 \right\rceil \\ &= K, \end{aligned}$$

we arrive at a contradiction. □

D. Operational Complexity

The results in this section are corollaries of Theorem 3.3 and Theorem 3.8 and the MINRES iteration bounds in Liu & Roosta (2022b). The following definitions are included from (Liu & Roosta, 2022b) for completeness.

Let $\Psi(\mathbf{H}, \mathbf{g})$ denote the set of $\mathbf{g}$-relevant eigenvalues[2], that is, the eigenvalues whose eigenspace is *not* orthogonal to $\mathbf{g}$. Denote $\psi = |\Psi(\mathbf{H}, \mathbf{g})|$ and let ψ_-, ψ_0 and ψ_+ be the number of negative, zero and positive $\mathbf{g}$-relevant eigenvalues so that $\psi = \psi_- + \psi_0 + \psi_+$. We impose the following order on the eigenvalues

$$\lambda_1 > \lambda_2 > \ldots > \lambda_{\psi_+} > 0 > \lambda_{\psi_+ + \psi_0 + 1} > \ldots > \lambda_{\psi}.$$

Denote by $\mathbf{U}_i$ the matrix with columns which form an orthonormal basis of the i^{th} eigenspace with the convention that the leading column is the only column onto which the gradient has nonzero projection. For $1 \leq i \leq \psi_+$ and $\psi_+ + \psi_0 + 1 \leq j \leq \psi$, define the following matrices

$$\mathbf{U}_{i+} = [\mathbf{U}_1 \ldots \mathbf{U}_i],\ \mathbf{U}_{j-} = [\mathbf{U}_j, \ldots, \mathbf{U}_\psi].$$

The columns of $\mathbf{U}_{i+}$ represent the eigenspaces of the i most positive $\mathbf{g}$-relevant eigenvalues, while $\mathbf{U}_{j-}$ represents the eigenspaces corresponding to the j most negative $\mathbf{g}$-relevant eigenvalues. As a special case, let $\mathbf{U}_+ = \mathbf{U}_{\psi+}$ and $\mathbf{U}_- = \mathbf{U}_{(\psi_+ + \psi_0 + 1)-}$. Finally, let

$$\mathbf{U} = [\mathbf{U}_+, \mathbf{U}_-].$$

We now state a key assumption for the result.

Assumption D.1. (Liu & Roosta, 2022b, Assumption 5) There exists $\tau > 0$ and $L_g^2/(L_g^2 + \eta^2) < \nu \leq 1$ such that for any $\mathbf{x} \in \mathbb{R}^d_+$ with $\mathbf{g} \notin \text{Null}(\mathbf{H})$ at least one of the following statements (i)-(iii) must hold

(i) If $\psi_+ \geq 1$ and $\psi_i \geq 1$ then there exists $1 \leq i \leq \psi_+$ and $\psi_+ + \psi_0 + 1 \leq j \leq \psi$ such that

$$\begin{gathered} \min\{\lambda_i, -\lambda_j\} \geq \tau, \\ \|(\mathbf{U}_{i+}\mathbf{U}_{i+}^\intercal + \mathbf{U}_{j-}\mathbf{U}_{j-}^\intercal)\mathbf{g}\|^2 \geq \nu\|\mathbf{U}\mathbf{U}^\intercal\mathbf{g}\|^2. \end{gathered}$$

(ii) If $\psi_+ \geq 1$ then there exists $1 \leq i \leq \psi_+$ such that

$$\begin{gathered} \lambda_i \geq \tau, \\ \|\mathbf{U}_{i+}\mathbf{U}_{i+}^\intercal\mathbf{g}\|^2 \geq \nu\|\mathbf{U}\mathbf{U}^\intercal\mathbf{g}\|^2. \end{gathered}$$

[2] Eigenvalues outside of $\Psi(\mathbf{H}, \mathbf{g})$ are essentially “invisible” to the Krylov subspace built out of products of $\mathbf{H}$ and $\mathbf{g}$.

(iii) If $\psi_- \geq 1$ then there exists some $\psi_+ + \psi_0 + 1 \leq j \leq \psi$ such that

$$\begin{gathered} -\lambda_j \geq \tau, \\ \|\mathbf{U}_{j-}\mathbf{U}_{j-}^{\intercal}\mathbf{g}\|^2 \geq \nu \|\mathbf{U}\mathbf{U}^{\intercal}\mathbf{g}\|^2. \end{gathered}$$

Recall that $\mathbf{U}\mathbf{U}^{\intercal}$, $\mathbf{U}_{j-}\mathbf{U}_{j-}^{\intercal}$ and $\mathbf{U}_{i+}\mathbf{U}_{i+}^{\intercal}$ represent a projection onto corresponding eigenspaces. Each part of Assumption D.1 has a natural interpretation as a requirement that there is at least one *large enough* magnitude $\mathbf{g}$-relevant eigenvalue for which the projection of the gradient onto the corresponding eigenspaces is not *too small*. This is a significant relaxation of a more conventional uniform bound on the magnitude of the eigenvalues. For example, a uniform bound on the smallest magnitude eigenvalues

$$\min\{\lambda_{\psi_+}, -\lambda_{\psi_+ + \psi_0 + +1}\} \geq \tau,$$

immediately implies Assumption D.1-(i) with $i = \psi_+$, $j = \psi_+ + \psi_0 + 1$ and $\nu = 1$. See Liu & Roosta (2022b, Assumption 5) for further discussion of this assumption.

Assumption D.1 allows us to bound the number of Hessian vector products that are required for MINRES to satisfy (10). Indeed, assuming $\mathbf{g} \notin \text{Null}(\mathbf{H})$, we appeal to Liu & Roosta (2022b, Eqn. (20)) to bound the number of iterations until the MINRES termination tolerance (10) is satisfied as

$$T_{\text{SOL}} = \min\left\{ \left\lceil \frac{\sqrt{L_g/\mu}}{4} \log\left(4 \middle/ \left(\frac{\eta^2}{L_g^2 + \eta^2} - (1-\nu)\right)\right) + 1 \right\rceil, g \right\},$$

where g denotes the grade of $\mathbf{g}$ with respect to $\mathbf{H}$ (Liu & Roosta, 2022a, Definition 1.3). We note that T_{SOL} has a logarithmic dependence on the inexactness rolernance, η.

On the other hand if $\psi_- \geq 1$ and Assumption D.1-(iii) holds, we appeal to Liu & Roosta (2022b, Eqn. (19)) to bound the iterations required to obtain a NPC direction as

$$T_{\text{NPC}} = \min\left\{ \max\left\{ \left\lceil \left(\frac{\sqrt{2(L_g + \mu)/\mu}}{4} \right) \log\left(\frac{2(L_g + \mu)(1-\nu)}{\mu\nu} \right) + 1 \right\rceil, 1 \right\}, g \right\}.$$

When $\nu = 1$, it is clear from the statement of Assumption D.1-(iii) that all $\mathbf{g}$-relevant eigenvalues are negative, which implies that negative curvature is detected at the very first iteration, i.e., $T_{\text{NPC}} = 1$. If we adopt the convention that $T_{\text{NPC}} = \infty$ when $\psi_- = 0$ or Assumption D.1-(iii) is unsatisfied we bound the number of MINRES iterations as $T = \min\{T_{\text{NPC}}, T_{\text{SOL}}\}$. If $\mathbf{g} \in \text{Null}(\mathbf{H})$ then $\mathbf{g}$ is declared a zero curvature direction at the very first iteration. We now prove the operational complexity results.

Corollary D.2 (First Order Operational Complexity Algorithm 1)**.** *Under the conditions of Theorem 3.3 and Assumption D.1, the total number of gradient evaluations and Hessian vector products in Algorithm 1 to obtain an ϵ_g-FO point is $\mathcal{O}(\epsilon_g^{-2})$, for d sufficiently large.*

Proof. Due to Theorem 3.3, the total number of outer iterations is $\mathcal{O}(\epsilon_g^{-2})$. To obtain the operational result we simply need to count the total number of gradient evaluations and Hessian vector products *per iteration*. The work required for each step of Algorithm 1 is equivalent to the number of MINRES iterations (i.e. Hessian vector product) plus a single gradient evaluation. In the case of Algorithm 1 the termination tolerance η has no dependence on ϵ_g. Considering the discussion above, for sufficiently large d, we bound the number of Hessian vector products as $\mathcal{O}(1)$. The conclusion follows from the fact that $\mathcal{O}(\epsilon_g^{-2})(1 + \mathcal{O}(1)) \in \mathcal{O}(\epsilon_g^{-2})$. □

Corollary D.3 (First Order Operational Complexity Algorithm 2)**.** *Under the conditions of Theorem 3.8 and Assumption D.1, the total number of gradient evaluations and Hessian vector products in Algorithm 2 to obtain an ϵ_g-FO point is $\tilde{\mathcal{O}}(\epsilon_g^{-3/2})$, for d sufficiently large.*

Proof. The result is similar to Corollary D.2. We utilise Theorem 3.8 to bound the total number of outer iterations as $\mathcal{O}(\epsilon_g^{-3/2})$. For Algorithm 2, the MINRES termination tolerance is $\eta = \theta\sqrt{\epsilon_g}$, so we bound the total number of Hessian vector products as $\tilde{\mathcal{O}}(1)$ for d large. The conclusion follows. □

Algorithm 4 Newton-MR TMP (Local Phase Version)

```
for $k = 0, 1, \ldots$ do
    Update sets $\mathcal{A}(\mathbf{x}_k, \delta_k)$ and $\mathcal{I}(\mathbf{x}_k, \delta_k)$ as in (2).
    if Some termination condition is satisfied then
        Terminate.
    end if
    $\mathbf{p}_k : \begin{cases} \mathbf{p}_k^{\mathcal{A}} \leftarrow -\mathbf{g}_k^{\mathcal{A}}, \\ (\mathbf{p}_k^{\mathcal{I}},\ \text{D}_{\text{type}}) \leftarrow \text{MINRES}(\mathbf{H}_k^{\mathcal{I}}, \mathbf{g}_k^{\mathcal{I}}, \eta, 0) \end{cases}$    # See Remark 3.15 regarding the choices for $\eta$.
    $\alpha_k \leftarrow$ Algorithm 5 with $\alpha_0 = 1$ and (11).
    $\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$
end for
```

E. Local Convergence

In this section we provide the detailed proof for Theorem 3.13. Our proof follows a similar line of reasoning as that in Bertsekas (1982, Proposition 3) but with several modifications and alterations specific to our setting and methodology. We assume in this section that $\mathcal{I}(\mathbf{x}_*, 0) \neq \emptyset$ as otherwise the analysis boils down to convergence of projected gradient to a trivial solution $\mathbf{x}_* = 0$. Our main aim is to show that after a finite number of iterations, the iterates eventually end up in the following subspace

$$X_* = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x}^i = 0,\ i \in \mathcal{A}(\mathbf{x}_*, 0)\}.$$

We start with a lemma to show that, by choosing our inexactness tolerance $\delta_k = \delta$, with

$$0 < \delta < \frac{1}{2} \min_{i \in \mathcal{I}(\mathbf{x}_*, 0)} \mathbf{x}_*^i, \tag{44}$$

where $\mathbf{x}_*$ is some local minima, we can properly "separate" the true active and inactive set if $\mathbf{x}_k$ is close enough to $\mathbf{x}_*$. That is, we apply the correct update to the true active and inactive indices.

Lemma E.1. *Let $\mathbf{x}_*$ be a local minima of (1) and $\mathbf{x}_k$ be an iterate of Algorithm 4 with δ chosen according to (44). There exists Δ_{sep} such that if $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{sep})$, then $\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_*, 0)$.*

Proof. Define

$$\Delta_{\text{sep}} \triangleq \min\left\{\frac{1}{2}\left(\min_{i \in \mathcal{I}(\mathbf{x}_*, 0)} \mathbf{x}_*^i - \delta\right), \delta\right\} > 0.$$

We first we prove $\mathcal{I}(\mathbf{x}_k, \delta_k) \supseteq \mathcal{I}(\mathbf{x}_*, 0)$. For any $i \in \mathcal{I}(\mathbf{x}_*, 0)$ and $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{sep}})$ we have

$$\begin{aligned} \mathbf{x}_*^i - \mathbf{x}_k^i < \Delta_{\text{sep}} &\le \frac{\mathbf{x}_*^i - \delta}{2} \\ \implies \frac{\mathbf{x}_*^i}{2} - \mathbf{x}_k^i &\le -\frac{\delta}{2} \\ \implies \frac{\mathbf{x}_*^i}{2} + \frac{\delta}{2} &\le \mathbf{x}_k^i \\ \implies \frac{3\delta}{2} &\le \mathbf{x}_k^i \\ \implies \delta &< \mathbf{x}_k^i, \end{aligned}$$

where the second to last line follows from (44).

Next we show that $\mathcal{I}(\mathbf{x}_k, \delta) \subseteq \mathcal{I}(\mathbf{x}_*, 0)$. In particular, we prove the contrapositive $i \in \mathcal{A}(\mathbf{x}_*, 0) \implies i \in \mathcal{A}(\mathbf{x}_k, \delta)$. For $i \in \mathcal{A}(\mathbf{x}_*, 0)$ we know that $\mathbf{x}_*^i = 0$ and so for $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{sep}})$ we have

$$\mathbf{x}_k^i = \mathbf{x}_k^i - \mathbf{x}_* < \Delta_{\text{sep}} \le \delta.$$

That is, $i \in \mathcal{A}(\mathbf{x}_k, \delta)$. □

With this result in hand, we know that we apply the "correct" update to $\mathbf{x}_k$. That is, true active indices receive a projected gradient update, while indices in the true inactive set receive a Newton-type step.

Recall the the second-order sufficient condition in Assumption 3.12. This condition is equivalent to

$$\langle \mathbf{z}, \nabla^2 f(\mathbf{x}_*)\mathbf{z}\rangle > 0, \ \mathbf{z} \in X_*.$$

By the continuity of the Hessian, we are free to choose $\Delta_{\text{cvx}} > 0$ such that for any $\mathbf{x} \in B(\mathbf{x}_*, \Delta_{\text{cvx}})$ the Hessian remains strongly positive definite on the subspace X_*. In other words, the constant, μ, satisfying

$$\mu \triangleq \min_{\mathbf{z}\in X_*, \ \mathbf{x}\in B(\mathbf{x}_*, \Delta_{\text{cvx}})} \frac{\langle \mathbf{z}, \nabla^2 f(\mathbf{x})\mathbf{z}\rangle}{\|\mathbf{z}\|^2} > 0, \tag{45}$$

is well defined. In the current notation, even if $\mathcal{I}(\mathbf{x}_k, \delta) = \mathcal{I}(\mathbf{x}_*, 0)$ we have $\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \notin X_*$ as $\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \in \mathbb{R}^{|\mathcal{I}(\mathbf{x}_k,\delta)|}$ is a subvector. Therefore, as a notational convenience, in this section we take subvectors and submatrices corresponding to a certain subset of indices, e.g., $\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}$, to be padded with zeros in the removed indices. Note that this implies that $\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}, \mathbf{p}_k^{\mathcal{A}(\mathbf{x}_k,\delta)} \in \mathbb{R}^d$ and $\mathbf{p}_k = \mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} + \mathbf{p}_k^{\mathcal{A}(\mathbf{x}_k,\delta)}$ but leaves the mechanics of Algorithm 4 unchanged. Now if $\mathcal{I}(\mathbf{x}_k, \delta) = \mathcal{I}(\mathbf{x}_*, 0)$ we have

$$\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} = \mathbf{p}_k^{\mathcal{I}(\mathbf{x}_*,0)} \in X_*.$$

Indeed, it is easy to see that for any $t = 0, \ldots, g$

$$\mathcal{K}_t(\mathbf{H}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}, \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}) \subseteq X_*.$$

In addition, if $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{cvx}})$ then, by $\mathbf{H}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} = \mathbf{H}_k^{\mathcal{I}(\mathbf{x}_*,0)}$, μ plays the role of Krylov subspace regularity constant, σ, (cf. Assumption 3.7) on $\mathcal{K}_t(\mathbf{H}_k^{\mathcal{I}(\mathbf{x}_*,0)}, \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_*,0)})$. Indeed, together, these results imply that, for any $t = 0, \ldots, g$, we have

$$\mathbf{s} \in \mathcal{K}_t(\mathbf{H}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}, \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}) \implies \langle \mathbf{s}, \mathbf{H}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\mathbf{s}\rangle \geq \mu\|\mathbf{s}\|^2. \tag{46}$$

From (46) it is clear that we can do away with the $\text{D}_{\text{type}} = \text{NPC}$ case and Assumption 3.7. With this in mind, we now demonstrate that the step size produced by the line search in Algorithm 4 is bounded.

Lemma E.2. *Assume that f satisfies Assumption 3.1 and $\mathbf{x}_*$ is a local minima of (1) satisfying Assumption 3.12. Then if $\mathbf{x}_k \in B(\mathbf{x}_*, \min\{\Delta_{cvx}, \Delta_{sep}\})$ the step size produced by the line search in Algorithm 4 satisfies $\alpha_k \in [\bar{\alpha}, 1]$ where*

$$\bar{\alpha} \triangleq \min\left\{1, \frac{2(1-\rho)\mu}{L_g}, \frac{\delta}{\|\mathbf{p}_k^{\mathcal{I}}\|}\right\}. \tag{47}$$

Proof. $\mathbf{x}_k \in B(\mathbf{x}_*, \min\{\Delta_{\text{cvx}}, \Delta_{\text{sep}}\})$ implies that $\mathcal{I}(\mathbf{x}_k, \delta) = \mathcal{I}(\mathbf{x}_*, 0) \neq \emptyset$ and (46) holds. It follows that MINRES always selects $\text{D}_{\text{type}} = \text{SOL}$ step.

The result follows from the step size selection procedure in Algorithm 4 and the analysis in the SOL case of Lemma C.1.

□

Building on Lemma E.1, our next result, Lemma E.3, will show that, close enough to $\mathbf{x}_*$, the active set update will be *large enough* and the inactive set update *small enough* that the zero bound constraints at $\mathbf{x}_{k+1}$ coincide with the zero bound constraints at $\mathbf{x}_*$, i.e., $\mathcal{A}(\mathbf{x}_{k+1}, 0) = \mathcal{A}(\mathbf{x}^*, 0)$. The intuition for this result is that the gradient and hence the step (cf. (20)) in the inactive indices should be going to zero as $\mathbf{x}_k$ approaches $\mathbf{x}_*$. By contrast, in the active set, a non-degeneracy condition (Assumption 3.11) ensures there is positive gradient in the active indices arbitrarily close to the boundary. When $\mathcal{A}(\mathbf{x}_{k+1}, 0) = \mathcal{A}(\mathbf{x}^*, 0)$, the fixed active set and small inactive step can also be used to ensure that our iterates do not drift too far from the starting point. This is the second part of Lemma E.3.

Lemma E.3. *Suppose that f satisfies Assumption 3.1. Let $\mathbf{x}_*$ be a local minima satisfying Assumptions 3.11 and 3.12. If δ is chosen according to (44), then the following two results hold:*

1. There exists $\Delta_{bnd} > 0$ such that $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{bnd})$ implies $\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_{k+1}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$.

2. Given a $\Delta > 0$, we can choose $\Delta_{cls} \in (0, \Delta_{bnd})$ such that $\|\mathbf{x}_k - \mathbf{x}_*\| < \Delta_{cls}$ implies that $\|\mathbf{x}_{k+1} - \mathbf{x}_*\| < \Delta$.

Proof. We stipulate that $\Delta_{\text{bnd}} \leq \min\{\Delta_{\text{sep}}, \Delta_{\text{cvx}}\}$. Note that, in this case, $\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_*, 0)$ by Lemma E.1 and $\alpha_k \in [\bar{\alpha}, 1]$ where

$$\bar{\alpha} \triangleq \min\left\{1, \frac{2(1-\rho)\mu}{L_g}, \frac{\delta}{\|\mathbf{p}_k^{\mathcal{I}}\|}\right\}. \tag{48}$$

by Lemma E.2. It is also clear that MINRES selects $\text{D}_{\text{type}} = \text{SOL}$. We first show that the step size, α_k, can be uniformly lower bounded for $\mathbf{x}_k$ close enough to $\mathbf{x}_*$. We do this by showing that the step $\|\mathbf{p}_k^{\mathcal{I}}\|$ can be upper bounded. Specifically, due to (46) the step, $\mathbf{p}_k^{\mathcal{I}}$, is upper bounded by the gradient magnitude (cf. (20) with $\varrho = \mu$)

$$\|\mathbf{p}_k^{\mathcal{I}}\| \leq \|\mathbf{g}_k^{\mathcal{I}}\|/\mu. \tag{49}$$

Next we use the continuity of $\nabla f(\mathbf{x})$ and the fact that $\mathbf{g}_*^{\mathcal{I}(\mathbf{x}_*,0)} = 0$ to choose $\Delta_0 \leq \min\{\Delta_{\text{sep}}, \Delta_{\text{cvx}}\}$ such that $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_0)$ implies

$$\|\mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\| = \|\mathbf{g}_k^{\mathcal{I}(\mathbf{x}_*,0)}\| \leq \frac{\mu\delta}{2},$$

where in the first equality we used Lemma E.1. This implies

$$\|\mathbf{p}_k^{\mathcal{I}}\| \leq \delta/2, \tag{50}$$

and hence by (48)

$$\bar{\alpha} \triangleq \min\left\{1, \frac{2(1-\rho)\mu}{L_g}\right\}.$$

We now show that $\mathcal{A}(\mathbf{x}^*, 0) \subseteq \mathcal{A}(\mathbf{x}_{k+1}, 0)$. Let $i \in \mathcal{A}(\mathbf{x}^*, 0)$. Define $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}_*$. By Assumption 3.11 and the continuity of $\nabla f(\mathbf{x})$, there exists Δ_1 such that, for $\|\mathbf{e}_k\| \leq \Delta_1$,

$$(\mathbf{g}(\mathbf{x}_k))^j = (\mathbf{g}(\mathbf{x}_* + \mathbf{e}_k))^j > \frac{\gamma}{2}, \ \forall j \in \mathcal{A}(\mathbf{x}^*, 0).$$

Consider $\Delta_2 = \min\{\Delta_0, \Delta_1, \overline{\alpha}\gamma/2\}$. If $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_2)$, we have

$$\mathbf{x}_k^i = \mathbf{x}_k^i - \mathbf{x}_*^i < \Delta_2 \leq \frac{\overline{\alpha}\gamma}{2}.$$

Using this bound and the lower bound for the gradient and step size we compute the update as

$$\mathbf{x}_k^i - \alpha_k \mathbf{g}_k^i \leq \mathbf{x}_k^i - \frac{\overline{\alpha}\gamma}{2} \leq 0 \implies \mathbf{x}_{k+1}^i = \mathcal{P}(\mathbf{x}_k^i + \alpha_k \mathbf{p}_k^i) = 0,$$

which implies $i \in \mathcal{A}(\mathbf{x}_{k+1}, 0)$.

Next, we show $\mathcal{A}(\mathbf{x}^*, 0) \supseteq \mathcal{A}(\mathbf{x}_{k+1}, 0)$. In particular, we prove the contrapositive statement $i \in \mathcal{I}(\mathbf{x}_*, 0) \implies i \in \mathcal{I}(\mathbf{x}_{k+1}, 0)$. Suppose $i \in \mathcal{I}(\mathbf{x}_*, 0)$ the result will follow by showing that the $\mathbf{x}_{k+1}$ remains bounded away from zero. Let $\Delta_{\text{bnd}} = \min\{\Delta_0, \Delta_1, \Delta_2, \Delta_{\text{sep}}, \Delta_{\text{cvx}}\}$. Having $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{bnd}})$ implies $\mathcal{I}(\mathbf{x}_*, 0) = \mathcal{I}(\mathbf{x}_k, \delta)$. Additionally, the bound (50) applies and so $\alpha \in [\overline{\alpha}, 1]$ implies

$$\alpha|\mathbf{p}_k^i| \leq \|\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\| \leq \delta/2,$$

which yields

$$\mathbf{x}_k^i + \alpha \mathbf{p}_k^i \geq \mathbf{x}_k^i - \alpha |\mathbf{p}_k^i| \geq \mathbf{x}_k^i - \frac{\delta}{2} > \frac{\delta}{2}, \tag{51}$$

where the final inequality follows from $i \in \mathcal{I}(\mathbf{x}_*, 0) = \mathcal{I}(\mathbf{x}_k, \delta) \implies \mathbf{x}_k^i > \delta$. Finally we compute the step as

$$\mathbf{x}_{k+1} = \mathcal{P}(\mathbf{x}_k^i + \alpha \mathbf{p}_k^i) = \mathbf{x}_k^i + \alpha \mathbf{p}_k^i > 0,$$

which is the result.

Now for the second part of the result. Fix $\Delta > 0$. From the first part of the result we know that for $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{bnd}})$ we have $\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_{k+1}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$, which implies $\mathbf{x}_{k+1}^{\mathcal{A}(\mathbf{x}_k,\delta)} = \mathbf{x}_{k+1}^{\mathcal{A}(\mathbf{x}_{k+1},0)} = 0$ and $\mathbf{x}_*^{\mathcal{A}(\mathbf{x}_k,\delta)} = \mathbf{x}_*^{\mathcal{A}(\mathbf{x}_*,0)} = 0$. Applying these equalities we obtain

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_*\| &= \left\| \mathbf{x}_{k+1}^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| \\
&= \left\| [\mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)]^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| \\
&= \left\| [\mathbf{x}_k + \alpha_k \mathbf{p}_k]^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| \\
&\leq \left\| \mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| + \alpha_k \left\| \mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| \\
&\leq \|\mathbf{x}_k - \mathbf{x}_*\| + \left\| \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| / \mu,
\end{aligned}$$

where we drop the projection on line three due to $[\mathbf{x}_k + \alpha_k \mathbf{p}_k]^{\mathcal{I}(\mathbf{x}_k,\delta)} > \delta/2$ when $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{bnd}})$ (cf. (51)). Again, $\mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} = \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_*,0)}$ so, by the continuity of $\nabla f(\mathbf{x})$, we are free to choose Δ_3 so that $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_3)$ implies

$$\left\| \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| = \left\| \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_*,0)} \right\| < \frac{\mu \Delta}{2},$$

Finally, we can choose $\Delta_{\text{cls}} = \min\{\Delta_{\text{bnd}}, \Delta_3, \Delta/2\}$ so that, if $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{cls}})$, we have

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_*\| &\leq \|\mathbf{x}_k - \mathbf{x}_*\| + \left\| \mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} \right\| / \mu \\
&< \frac{\Delta}{2} + \frac{\Delta}{2} \\
&= \Delta.
\end{aligned}$$

□

The second part of Lemma E.3 can be used with the choice $\Delta = \Delta_{\text{bnd}}$ to obtain

$$\|\mathbf{x}_k - \mathbf{x}_*\| < \Delta_{\text{cls}} \implies \|\mathbf{x}_{k+1} - \mathbf{x}_*\| < \Delta_{\text{bnd}}.$$

In this case, we can guarantee, due to $\Delta_{\text{cls}} \leq \Delta_{\text{bnd}}$ and the first part of Lemma E.3 applied to $\mathbf{x}_k$, that

$$\mathcal{A}(\mathbf{x}_k, \delta) = \mathcal{A}(\mathbf{x}_{k+1}, 0) = \mathcal{A}(\mathbf{x}_*, 0),$$

and from $\mathbf{x}_{k+1} \in B(\mathbf{x}_*, \Delta_{\text{bnd}})$ and the first part of Lemma E.3 again

$$\mathcal{A}(\mathbf{x}_{k+1}, \delta) = \mathcal{A}(\mathbf{x}_{k+2}, 0) = \mathcal{A}(\mathbf{x}_*, 0).$$

Together, these results show that $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{cls}})$ implies $\mathbf{x}_{k+1}, \mathbf{x}_{k+2} \in X_*$. This means that the iterates of our algorithm essentially "looks" like unconstrained minimisation in this subspace. Unfortunately, with the results we have so far, we cannot guarantee that the iterates continue to stay close enough to the minima beyond iteration $k+2$. Lemma E.4 will overcome this problem by using the second-order sufficient condition and adapting an unconstrained optimisation result (Bertsekas, 1996, Proposition 1.12). The main idea is that the "strict convexity" on X_* induced by Assumption 3.12 (cf. (45)) implies that there exists a small "basin" (restricted to X_*) that the iterates will not leave once they enter. We can then use Lemma E.3 to show that our iterates eventually enter X_* and the corresponding basin.

Lemma E.4. *Let f satisfy Assumption 3.1 and $\mathbf{x}_*$ be a local minima satisfying Assumptions 3.11 and 3.12. Let δ be chosen according to (44). If there is an iterate, $\mathbf{x}_k$, of Algorithm 4 such that $\mathcal{A}(\mathbf{x}_k,\delta) = \mathcal{A}(\mathbf{x}_k,0) = \mathcal{A}(\mathbf{x}_*,0)$ and $\mathcal{A}(\mathbf{x}_{k+1},0) = \mathcal{A}(\mathbf{x}_*,0)$ (i.e. $\mathbf{x}_k, \mathbf{x}_{k+1} \in X_*$) then there exists a neighbourhood (restricted to X_*) of $\mathbf{x}_*$, $\mathcal{N}(\mathbf{x}_*)$, such that if $\mathbf{x}_k \in \mathcal{N}(\mathbf{x}_*)$ and then $\mathbf{x}_{k+1} \in \mathcal{N}(\mathbf{x}_*)$. Additionally, $\mathcal{N}(\mathbf{x}_*)$ is independent of the iterates and can be chosen arbitrarily small, i.e., for any $\Delta > 0$ we have $\mathcal{N}(\mathbf{x}_*) \subset B(\mathbf{x}_*,\Delta)$.*

Proof. We fix $\Delta \leq \Delta_{\text{cvx}}$ and define

$$\mathcal{N}(\mathbf{x}_*) = \left\{ \mathbf{x} \in B(\mathbf{x}_*,\Delta) \cap X_* \mid f(\mathbf{x}) \leq f(\mathbf{x}_*) + \frac{\mu}{2}\left(\frac{\Delta}{1+L_g/\mu}\right)^2 \right\}.$$

We will show that this set is the desired neighbourhood on X_* in the sense there exists an open ball in the relative interior of X_*. The mean value theorem implies that there is a constant $t \in (0,1)$ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{1}{2}\langle \mathbf{y}-\mathbf{x}, \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})\rangle. \tag{52}$$

We obtain, by Assumption 3.1,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L_g}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

Let $\mathbf{x} = \mathbf{x}_*$ and $\mathbf{y} \in B(\mathbf{x}_*,\Delta) \cap X_*$. The fact that $\mathbf{y}$ and $\mathbf{x}_*$ only have nonzero components in $\mathcal{I}(\mathbf{x}_*,0)$, while the optimality condition (7c) implies $\nabla f(\mathbf{x}_*)$ only has zero components in $\mathcal{I}(\mathbf{x}_*,0)$, allows us to write

$$f(\mathbf{y}) \leq f(\mathbf{x}_*) + \frac{L_g}{2}\|\mathbf{y}-\mathbf{x}_*\|^2,$$

so by choosing $\mathbf{y}$ close enough to $\mathbf{x}_*$ we have $\mathbf{y} \in \mathcal{N}(\mathbf{x}_*)$. This implies that $\mathcal{N}(\mathbf{x}_*)$ is a neighbourhood of $\mathbf{x}_*$, in X_*.

Let $\mathbf{x} = \mathbf{x}_*$ and $\mathbf{y} = \mathbf{x}_k$ for $\mathbf{x}_k \in B(\mathbf{x}_*,\Delta) \cap X_*$. Then, $\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*) \in B(\mathbf{x}_*,\Delta) \cap X_*$ for any $t \in [0,1]$. Hence, (45) applied to (52) yields

$$\frac{\mu}{2}\|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_*). \tag{53}$$

Next, we seek to bound the distance between subsequent errors. Since $\mathcal{A}(\mathbf{x}_k,\delta) = \mathcal{A}(\mathbf{x}_{k+1},0) = \mathcal{A}(\mathbf{x}_*,0)$, we have

$$\mathbf{x}_{k+1}^{\mathcal{A}(\mathbf{x}_k,\delta)} = \mathbf{x}_{k+1}^{\mathcal{A}(\mathbf{x}_{k+1},0)} = 0,$$

and

$$\mathbf{x}_{*}^{\mathcal{A}(\mathbf{x}_k,\delta)} = \mathbf{x}_{*}^{\mathcal{A}(\mathbf{x}_*,0)} = 0.$$

We compute

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_*\| &= \|\mathbf{x}_{k+1}^{\mathcal{I}(\mathbf{x}_{k+1},0)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_*,0)}\| \\
&= \|\mathbf{x}_{k+1}^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\| \\
&= \|[\mathcal{P}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)]^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\| \\
&= \|\mathbf{x}_{k}^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)} + \alpha_k \mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\| \\
&\leq \|\mathbf{x}_{k}^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\| + \alpha_k\|\mathbf{p}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\|,
\end{aligned} \tag{54}$$

where the fourth line follows from $\mathcal{I}(\mathbf{x}_k,\delta) = \mathcal{I}(\mathbf{x}_{k+1},0)$ and $i \in \mathcal{I}(\mathbf{x}_{k+1},0)$ implying that $0 < \mathbf{x}_{k+1}^i = \mathcal{P}(\mathbf{x}_k^i + \alpha_k \mathbf{p}_k^i) \implies \mathcal{P}(\mathbf{x}_k^i + \alpha_k \mathbf{p}_k^i) = \mathbf{x}_k^i + \alpha_k \mathbf{p}_k^i$. Since $\mathcal{I}(\mathbf{x}_k,\delta) = \mathcal{I}(\mathbf{x}_*,0)$ and $\mathbf{x}_k \in B(\mathbf{x}_*,\Delta_{\text{cvx}})$, we know (49) holds. We can refine (49) by combining Assumption 3.1, $\mathcal{I}(\mathbf{x}_k,\delta) = \mathcal{I}(\mathbf{x}_k,0) = \mathcal{I}(\mathbf{x}_*,0)$ and (7c) to obtain

$$\left\|\mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,\delta)}\right\| = \left\|\mathbf{g}_k^{\mathcal{I}(\mathbf{x}_k,0)} - \mathbf{g}_*^{\mathcal{I}(\mathbf{x}_*,0)}\right\| \leq \|\mathbf{g}_k - \mathbf{g}_*\| \leq L_g\|\mathbf{x}_k - \mathbf{x}_*\| = L_g\left\|\mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\right\|.$$

Combining this bound, $\alpha_k \leq 1$, (49) and (54) we have

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_*\| &\leq \|\mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\| + \frac{L_g}{\mu}\left\|\mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\right\| \\ &= \left(1 + \frac{L_g}{\mu}\right)\|\mathbf{x}_k^{\mathcal{I}(\mathbf{x}_k,\delta)} - \mathbf{x}_*^{\mathcal{I}(\mathbf{x}_k,\delta)}\| \\ &= \left(1 + \frac{L_g}{\mu}\right)\|\mathbf{x}_k - \mathbf{x}_*\|. \end{aligned} \tag{55}$$

We now show that this is enough to guarantee that $\mathbf{x}_{k+1} \in \mathcal{N}(\mathbf{x}_*)$. In particular, if $\mathbf{x}_k \in \mathcal{N}(\mathbf{x}_*)$ then $\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{cvx})$, so by combining the definition of $\mathcal{N}(\mathbf{x}_*)$ and (53) we have

$$\frac{\mu}{2}\|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{\mu}{2}\left(\frac{\Delta}{1 + L_g/\mu}\right)^2 \implies \|\mathbf{x}_k - \mathbf{x}_*\| < \frac{\Delta}{1 + L_g/\mu}.$$

Applying (55) we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| < \Delta,$$

which implies that $\mathbf{x}_{k+1} \in B(\mathbf{x}_*, \Delta) \cap X_*$. In addition, α_k satisfies the line search criterion, which guarantees that $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ and so

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq f(\mathbf{x}_k) - f(\mathbf{x}_*) \leq \frac{\mu}{2}\left(\frac{\Delta}{1 + L_g/\mu}\right)^2,$$

which implies $\mathbf{x}_{k+1} \in \mathcal{N}(\mathbf{x}_*)$, as needed. In the above argument we are free to replace Δ with any $\Delta' \in (0, \Delta]$ which implies that we can always choose $\mathcal{N}(\mathbf{x}_*)$ sufficiently small. □

We our now ready to prove Theorem 3.13.

Proof of Theorem 3.13. Note that we are free to choose the neighbourhood $\mathcal{N}(\mathbf{x}_*)$ of $\mathbf{x}_*$ on X_* from Lemma E.4 small. We therefore select $\Delta_0 < \Delta_{\text{cvx}}$ and $\mathcal{N}(\mathbf{x}_*)$ to satisfy the following inclusions

$$B(\mathbf{x}_*, \Delta_0) \cap X_* \subseteq \mathcal{N}(\mathbf{x}_*) \subseteq B(\mathbf{x}_* \Delta_{\text{bnd}}) \cap X_*. \tag{56}$$

By the second part of Lemma E.3, there exists $\Delta_{\text{cls}} \leq \Delta_{\text{bnd}}$ such that the following inclusions hold

$$\mathbf{x}_k \in B(\mathbf{x}_*, \Delta_{\text{cls}}) \implies \mathbf{x}_{k+1} \in B(\mathbf{x}_*, \Delta_0). \tag{57}$$

Choose $\Delta_{\text{actv}} = \Delta_{\text{cls}}$ and suppose that $\mathbf{x}_{\bar{k}} \in B(\mathbf{x}_*, \Delta_{\text{actv}})$. The first inclusion of (57), implies $\mathcal{A}(\mathbf{x}_{\bar{k}}, \delta) = \mathcal{A}(\mathbf{x}_{\bar{k}+1}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$, i.e. $\mathbf{x}_{\bar{k}+1} \in X_*$, by $\Delta_{\text{cls}} \leq \Delta_{\text{bnd}}$ and the first part of Lemma E.3. This fact and the second inclusion of (57), implies $\mathbf{x}_{\bar{k}+1} \in \mathcal{N}(\mathbf{x}_*)$ and therefore, by the second inclusion of (56), $\mathcal{A}(\mathbf{x}_{\bar{k}+1}, \delta) = \mathcal{A}(\mathbf{x}_{\bar{k}+2}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$, Again by the first part of Lemma E.3. Combining what we have so far, we obtain $\mathcal{A}(\mathbf{x}_{\bar{k}+1}, \delta) = \mathcal{A}(\mathbf{x}_{\bar{k}+1}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$, which is the result for $\bar{k}+1$. Additionally, however, we can apply Lemma E.4 applied to the iterate $\bar{k}+1$ to obtain $\mathbf{x}_{\bar{k}+2} \in \mathcal{N}(\mathbf{x}_*)$. The argument for $\bar{k}+1$ may now be repeated for $k \geq \bar{k}+2$. For instance, (56) and $\mathbf{x}_{\bar{k}+2} \in \mathcal{N}(\mathbf{x}_*)$ implies $\mathbf{x}_{\bar{k}+2} \in B(\mathbf{x}_*, \Delta_{\text{bnd}})$ and so $\mathcal{A}(\mathbf{x}_{\bar{k}+2}, \delta) = \mathcal{A}(\mathbf{x}_{\bar{k}+3}, 0) = \mathcal{A}(\mathbf{x}_*, 0)$ by Lemma E.3 and $\mathbf{x}_{\bar{k}+3} \in \mathcal{N}(\mathbf{x}_*)$ by Lemma E.4, which yields the result for $\bar{k}+2$ and sets up the argument for $\mathbf{x}_{\bar{k}+3}$. Continuing in this fashion yields the result for the given Δ_{actv}. □

F. Further Details and Extended Numerical Results

In this section we provide some additional elements of our proposed methods, further details on our experimental setup, and also give a more complete description of various problems we consider for our numerical simulations.

F.1. Line Search Algorithms

Here, we gather the line search algorithms used for the theoretical analysis as well as the empirical evaluations of our methods.

Algorithm 5 Backward Tracking Line Search.

```
input Initial step size $\alpha_0$, Line search criterion, Scaling parameter $0 < \zeta < 1$.
$\alpha \leftarrow \alpha_0$.
while Line search criterion is not satisfied do
    $\alpha \leftarrow \zeta\alpha$.
end while
return $\alpha$.
```

Algorithm 6 Forward/Backward Tracking Line Search

```
input Initial step size $\alpha_0$, Line search criterion, Scaling parameter $0 < \zeta < 1$.
$\alpha \leftarrow \alpha_0$.
if Line search criterion is not satisfied then
    Call Algorithm 5
else
    while Line search criterion is satisfied do
        $\alpha = \alpha/\zeta$
    end while
    return $\zeta\alpha$.
end if
```

F.2. Smooth Reformulation of Nonsmooth ℓ_1 Regression

Consider ℓ_1 regularisation of a smooth function, f, as given in (12). Unfortunately, even when f is smooth, the objective (12) is non-differentiable when $\mathbf{x}^i = 0$ for some $i = 1, \ldots, d$. However, it was shown in Schmidt et al. (2007) that one can reformulate (12) into a smooth problem by splitting $\mathbf{x}$ into positive and negative parts, i.e., $\mathbf{x}_+ = \max(0, \mathbf{x})$ and $\mathbf{x}_- = -\min(0, \mathbf{x})$, where "max" and "min" are taken elementwise. Indeed, we have the identities

$$\mathbf{x}^i = \mathbf{x}_+^i - \mathbf{x}_-^i,$$

and

$$|\mathbf{x}^i| = \mathbf{x}_+^i + \mathbf{x}_-^i,$$

which we can use to reformulate (58) as a constrained problem on $\mathbb{R}^{2d}$. In particular, the following auxiliary function is equivalent to the objective of (12)

$$F(\mathbf{x}_+, \mathbf{x}_-) \triangleq f(\mathbf{x}_+ - \mathbf{x}_-) + \lambda \sum_{i=1}^{d} (\mathbf{x}_+^i + \mathbf{x}_-^i).$$

If we make the identification $\mathbf{z} = [\mathbf{x}_+, \mathbf{x}_-] \in \mathbb{R}^{2d}$, we obtain the auxiliary minimisation problem defined by

$$\min_{\mathbf{z} \in \mathbb{R}^{2d}} F(\mathbf{z}) \quad \text{subject to} \quad \mathbf{z} \geq \mathbf{0}. \tag{58}$$

The nonpositivity condition in (58) ensures that $\mathbf{z}$ can be interpreted as the positive and negative part of the underlying variable, $\mathbf{x}$. The gradient and Hessian of the auxiliary function, F, are given by

$$\nabla F(\mathbf{x}_+, \mathbf{x}_-) = \begin{pmatrix} \nabla f(\mathbf{x}_+ - \mathbf{x}_-) + \lambda \mathbf{1}_{d \times 1} \\ -\nabla f(\mathbf{x}_+ - \mathbf{x}_-) + \lambda \mathbf{1}_{d \times 1} \end{pmatrix}, \quad \nabla^2 F(\mathbf{x}_+, \mathbf{x}_-) = \begin{pmatrix} \nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-) & -\nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-) \\ -\nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-) & \nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-) \end{pmatrix}.$$

Remark F.1 (Evaluating the gradients and Hessian-vector products). Clearly, evaluating the gradient of F requires only a single evaluation of the original gradient, ∇f. On the other hand, for computing a Hessian-vector product of F with a vector $\mathbf{v} = (\mathbf{v}_1^\intercal, \mathbf{v}_2^\intercal)^\intercal \in \mathbb{R}^{2d}$, we have

$$\nabla^2 F(\mathbf{x}_+, \mathbf{x}_-)\mathbf{v} = \begin{pmatrix} \nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-)\mathbf{v}_1 - \nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-)\mathbf{v}_2 \\ -\nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-)\mathbf{v}_1 + \nabla^2 f(\mathbf{x}_+ - \mathbf{x}_-)\mathbf{v}_2 \end{pmatrix},$$

which requires two Hessian-vector products of the original function f in the form $\nabla^2 f(\mathbf{x})\mathbf{w}$ where $\mathbf{w} \in \mathbb{R}^d$.

As a sanity check, we show that the first order stationary points of (58) and (12) coincide. The first order necessary conditions for (58) imply that for $j = 1, \ldots, d$ and $i = j$,

$$[\mathbf{x}_+^*]^i \geq \mathbf{0}, \quad \text{and} \quad \begin{cases} [\nabla f(\mathbf{x}_+^* - \mathbf{x}_-^*)]^i + \lambda = 0, & \text{if } [\mathbf{x}_+^*]^i > 0, \\ [\nabla f(\mathbf{x}_+^* - \mathbf{x}_-^*)]^i + \lambda \geq 0, & \text{if } [\mathbf{x}_+^*]^i = 0, \end{cases} \tag{59}$$

and for $j = d+1, \ldots, 2d$ and $i = j - d$,

$$[\mathbf{x}_-^*]^i \geq \mathbf{0}, \quad \text{and} \quad \begin{cases} -[\nabla f(\mathbf{x}_+^* - \mathbf{x}_-^*)]^i + \lambda = 0, & \text{if } [\mathbf{x}_-^*]^i > 0, \\ -[\nabla f(\mathbf{x}_+^* - \mathbf{x}_-^*)]^i + \lambda \geq 0, & \text{if } [\mathbf{x}_-^*]^i = 0. \end{cases} \tag{60}$$

On the other hand, the first order stationary points of the problem (12) can be expressed in terms of the Clarke subdifferential (Clarke, 1990, Chapter 2) as those points $\mathbf{x}^*$ for which $\mathbf{0} \in \nabla f(\mathbf{x}^*) + \partial \|\mathbf{x}^*\|_1$. That is, for $i = 1, \ldots, d$, we have

$$\begin{cases} [\nabla f(\mathbf{x}^*)]^i + \lambda = 0 & \text{if } [\mathbf{x}^*]^i > 0, \\ [\nabla f(\mathbf{x}^*)]^i - \lambda = 0 & \text{if } [\mathbf{x}^*]^i < 0, \\ \left|[\nabla f(\mathbf{x}^*)]^i\right| \leq \lambda & \text{if } [\mathbf{x}^*]^i = 0. \end{cases} \tag{61}$$

We first show that if $\mathbf{z}^* = [\mathbf{x}_+^*, \mathbf{x}_-^*]$ satisfies (59) and (60) then $\mathbf{x}^* = \mathbf{x}_+^* - \mathbf{x}_-^*$ satisfies (61). First, suppose $[\mathbf{x}^*]^i > 0$. In this case, we must have $[\mathbf{x}^*]^i = [\mathbf{x}_+^*]^i > 0 = [\mathbf{x}_-^*]^i$, which from the first case of (59) implies the first case of (61). When $[\mathbf{x}^*]^i < 0$, since $[\mathbf{x}^*]^i = [\mathbf{x}_-^*]^i > 0 = [\mathbf{x}_+^*]^i$, the first case of (60) implies the first case of (61). Finally, when $[\mathbf{x}^*]^i = 0$, we have $[\mathbf{x}_+^*]^i = [\mathbf{x}_-^*]^i = 0$, and we appeal to the second case of both (59) and (60) to obtain

$$[\nabla f(\mathbf{x}_*)]^i \geq -\lambda, \quad \text{and} \quad [\nabla f(\mathbf{x}_*)]^i \leq \lambda,$$

which implies $\left|[\nabla f(\mathbf{x}_*)]^i\right| \leq \lambda$, i.e., the third case of (61).

We now show that if $\mathbf{x}^* = \mathbf{x}_+^* - \mathbf{x}_-^*$ satisfies (61), then if $\mathbf{z}^* = [\mathbf{x}_+^*, \mathbf{x}_-^*]$ satisfies (59) and (60). Consider the first case of (61). Noting again that $[\mathbf{x}^*]^i = [\mathbf{x}_+^*]^i > 0 = [\mathbf{x}_-^*]^i$, it clearly implies the first and the second cases of (59) and (60), respectively (recall $\lambda > 0$). Similarly, the second case of (61) implies the second and the first cases of (59) and (60), respectively. Finally, it is clear that the third case of (61) implies the second case for both (59) and (60).

F.3. Additional Experimental Details

Oracle Calls as Complexity Measure Following the typical convection in the optimisation literature, in all our experiments, we plot the objective value against the total number of oracle calls for function, gradient, and Hessian-vector product evaluations. We adopt this approach because the measurement of “wall-clock” time can be heavily dependent on specific implementation details and computational platform. In contrast, counting the number of equivalent function evaluations, as an implementation and system independent unit of complexity is more appropriate and fair. More specifically, upon evaluating the function, computing its gradient is equivalent to one additional function evaluation, and computing a Hessian-vector product requires two additional function evaluations compared to a gradient evaluation (Pearlmutter, 1994). For example, in neural networks, for a given data at the input layer, evaluation of network’s output, i.e., function evaluation, involves one forward propagation. The corresponding gradient is computed by performing one additional backward propagation. After computing the gradient, an additional forward followed by a backward propagation give the corresponding Hessian-vector product (Goodfellow et al., 2016).

Parameter Settings In all experiments we set $\epsilon_k = \delta_k = \sqrt{\epsilon_g}$ as per Theorem 3.8. For the Newton-MR TMP we set the inexactness condition for MINRES, i.e., (10), to $\eta = 10^{-2}$ for convex problems and $\eta = 1$ for nonconvex problems. We apply a less stringent tolerance in the nonconvex case to maximise the chances of terminating early with a ”good enough”

SOL type solution. Indeed, running the solver too long increases the odds that spurious negative curvature direction will arise as part of iterations. Since such directions never occur in convex settings, one can afford to solve the subproblems more accurately.

For projected Newton-CG, we use the parameter settings from the experiments in Xie & Wright (2023). Specifically, in the notation of Xie & Wright (2023), we set the accuracy parameter and back tracking parameter to $\zeta = \theta = 0.5$ and the step acceptance parameter to $\eta = 0.2$. Furthermore, following the algorithmic description of Xie & Wright (2023), and to have equivalent termination conditions, we modify the gradient negativity check from $\mathbf{g}_k^i < -\epsilon_k^{3/2}$ to $\mathbf{g}_k^i < -\epsilon_k$ for this method.

For projected gradient and Newton-MR TMP, we set the scaling parameter in Algorithms 5 and 6 to $\zeta = 0.5$ and the sufficient decrease parameter to $\rho = 10^{-4}$. All line searches are initialised from $\alpha_0 = 1$. We note that, for both FISTA and PGM, we terminate the iterations when $|(f(\mathbf{x}_k) + \lambda\|\mathbf{x}_k\|_1 - (f(\mathbf{x}_{k-1}) + \lambda\|\mathbf{x}_{k-1}\|_1)| < 10^{-8}$ on the ℓ_1 problem and $|f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})| < 10^{-8}$ otherwise. We set the momentum term in PGM to $\beta = 0.9$ and select the fixed step size by starting from $\alpha = 1$ and successively shrinking the step size by a factor of 10 until the iterates are stable for the duration of the experiment, i.e., no divergence or large scale oscillations. This procedure resulted in a step size of $\alpha = 10^{-3}$ for the ℓ_1 MLP (Figure 3) and $\alpha = 1$ for the NNMF problems (Figures 4 and 5).

We now give a more complete description of each of the objective functions.

Multinomial Regression We first consider is the problems of multinomial regression on C classes. Specifically, consider a set of data items $\{\mathbf{a}_i, b_i\}_{i=1}^n \subset \mathbb{R}^d \times \{1, \dots C\}$. Denote the weights of each class as $\mathbf{x}_1, \dots, \mathbf{x}_C$ and define $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_{C-1}]$. We are free to take $\mathbf{x}_C = \mathbf{0}$ as class C is identifiable from the weights of the other classes. The objective, f, is given by

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C-1} -\mathbf{1}(b_i = c)\log\left(\text{softmax}(\mathbf{x}_c, \mathbf{a}_i)\right), \tag{62}$$

where $\mathbf{1}(\cdot)$ is the indicator function and

$$\text{softmax}(\mathbf{x}_c, \mathbf{a}_i) = \frac{\exp\left(\langle \mathbf{x}_c, \mathbf{a}_i\rangle\right)}{\sum_{c=1}^{C}\exp\left(\langle \mathbf{x}_c, \mathbf{a}_i\rangle\right)}.$$

In this case, the objective is convex. We allow for a constant term in each set of weights, $\mathbf{x}_c$, which we do not apply the ℓ_1 penalisation to.

All methods for this example are initialised from $\mathbf{x}_0 = \mathbf{0}$.

Neural Network Again, suppose we have a set of data items $\{\mathbf{a}_i, b_i\}_{i=1}^n \subset \mathbb{R}^d \times \{1, \dots C\}$. We consider a small two layer network with a smooth activation function. Specifically, we consider the sigmoid weighted linear unit (SiLU) activation (Elfwing et al., 2017) defined by

$$\sigma(x) = \frac{x}{1 + e^{-x}}.$$

We note that the SiLU activation is similar to ReLU and is the product of a linear activation with a standard sigmoid activation. We define a network, $\mathbf{h}(\cdot; \mathbf{x})$ parameterised by the weights, $\mathbf{x}$, with the following architecture

$$\text{Input (d)} \to \text{Linear (100)} \to \text{SiLU} \to \text{Linear (100)} \to \text{SiLU} \to \text{Linear (10)},$$

where the number in brackets denotes the size of the output from the layer. Note that we allow for a bias term in each linear layer which we do not apply the ℓ_1 penalty to. The objective function, f, is given by cross entropy loss incurred by the network over the entire dataset

$$f(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\exp([\mathbf{h}(\mathbf{a}_i; \mathbf{x})]^{b_i})}{\sum_{c=1}^{C}\exp([\mathbf{h}(\mathbf{a}_i; \mathbf{x})]^{c})}\right). \tag{63}$$

The weights for layer i, denoted $\mathbf{x}_i$, are initialised with the default PyTorch initialisation, that is, via independent uniform draw

$$\mathbf{x}_i \sim U(-\sqrt{k}, \sqrt{k}),$$

where $k = 1/(\#\text{Inputs})$ with $(\#\text{Inputs})$ the number of input features into the layer.

NNMF Problem A common choice for (13) is a standard Euclidean distance function

$$D(\mathbf{Y}, \mathbf{WH}) = \frac{1}{nm} \|\mathbf{Y} - \mathbf{WH}\|_F^2 , \tag{64}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. In this case, (13) is nonconvex in both $\mathbf{W}$ and $\mathbf{H}$, when considered simultaneously, but convex so long as one of the variables is held fixed. This motivates the standard approach to solving (13) based on alternating updates to $\mathbf{W}$ and $\mathbf{H}$ (Gillis, 2014) where one variable is fixed while optimise over the other (e.g., alternating nonnegative least squares).

By contrast, to test our algorithm, we specifically consider solving (13) as a nonconvex problem in $\mathbf{W}$ and $\mathbf{H}$ simultaneously[3]. For our first experiment (Figure 4), we consider a text data application. When comparing text documents, we aim to have a similarity measure that is *independent* of document length. Indeed, we consider documents similar if they have similar word frequency *ratios*. This notion of similarity is naturally captured by measuring alignment between vectors, which motivates the use of a loss function based on cosine similarity as

$$D(\mathbf{Y}, \mathbf{WH}) = \frac{1}{n} \sum_{i=1}^{n} 1 - \cos\left(\theta(\mathbf{y}_i, (\mathbf{WH})_i)\right), \tag{65}$$

where $\theta(\mathbf{y}_i, (\mathbf{WH})_i)$ is the angle between the ith predicted and true document. This loss function only considers the alignment between documents. Indeed, we can write

$$\cos\left(\theta(\mathbf{y}_i, (\mathbf{WH})_i)\right) = \frac{\langle \mathbf{Y}_i, (\mathbf{WH})_i \rangle}{\|\mathbf{Y}_i\| \|(\mathbf{WH})_i\|}.$$

However, using this representation it is clear that, due to the nonnegativity of $\mathbf{Y}$ and $\mathbf{WH}$, (65) ranges between 0 and 1. It is also clear that (65) is equivalent to a Euclidean distance with normalisation

$$\frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\mathbf{Y}_i}{\|\mathbf{Y}_i\|} - \frac{(\mathbf{WH})_i}{\|(\mathbf{WH})_i\|} \right\|^2 .$$

In our second example (Figure 5), we consider (13) with a standard Euclidean distance function (64) and a nonconvex regularisation term R_λ. Specifically, we consider a version of the smooth clipped absolute deviation regularisation (SCAD) first proposed in Fan & Li (2001). SCAD uses a quadratic function to smoothly interpolate between a regular ℓ_1 penalty and a constant penalty

$$\text{SCAD}_{\lambda,a}(x) = \begin{cases} \lambda|x|, & |x| < \lambda, \\ \frac{a\lambda|x| - x^2 - \lambda^2}{a-1}, & \lambda \leq |x| < a\lambda, \\ \frac{\lambda^2(a+1)}{2}, & |x| \geq a\lambda. \end{cases}$$

The SCAD penalty reduces the downward bias on large parameters typical of the ℓ_1 penalty while still allowing for sparsification of small parameters. We consider a twice smooth clipped absolute deviation, which we call TSCAD. TSCAD replaces the quadratic interpolant with a quartic, $Q_{\lambda,a}(x)$, which allows for a twice continuously differentiable penalty

$$\text{TSCAD}_{\lambda,a}(x) = \begin{cases} \lambda|x|, & |x| < \lambda, \\ Q_{\lambda,a}(x), & \lambda \leq |x| < a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |x| \geq a\lambda. \end{cases}$$

The regularisation term is simply given by

$$R_\lambda(\mathbf{W}, \mathbf{H}) = \sum_{i,j} \text{TSCAD}_{\lambda,a}(\mathbf{W}_{ij}) + \sum_{i,j} \text{TSCAD}_{\lambda,a}(\mathbf{H}_{ij}).$$

[3] Our algorithm could be employed as a subproblem solver in alternating schemes on $\mathbf{W}$ and $\mathbf{H}$. Indeed, the original Bertsekas TMP has been applied for this purpose (Kuang et al., 2015).

Due to the inherent nonconvexity of the NNMF problem, initialisation is key to obtaining good results. We utilised a simple half normal initialisation. Indeed, because the data matrix for each NNMF example (Figures 4 and 5) satisfies $0 \leq \mathbf{Y} \leq 1$, we produced the initialisation by drawing $(\mathbf{W}_0')_{ij} \sim \mathcal{N}(0,1)$ and $(\mathbf{H}_0')_{ij} \sim \mathcal{N}(0,1)$ and normalising in the following manner

$$\mathbf{W}_0 \leftarrow \frac{|\mathbf{W}_0'|}{\sqrt{\max\left(|\mathbf{W}_0'||\mathbf{H}_0'|\right)}}, \quad \mathbf{H}_0 \leftarrow \frac{|\mathbf{H}_0'|}{\sqrt{\max\left(|\mathbf{W}_0'||\mathbf{H}_0'|\right)}},$$

where $|\cdot|$ is taken elementwise. This initialisation was found to result in nontrivial solutions (i.e., visually reasonable low rank representations $\mathbf{H}$) to (13).

F.4. Simulations For Fast Local Convergence

In Figures 6 and 7, we consider an extended version of the results in Figures 1 and 2, respectively. Specifically, we plot the progress in each of the termination conditions (7). Part (a) of all figures depict the gradient norm on the inactive set. For Newton-MR TMP, this is the termination condition associated with the Newton-MR portion of the step. We see in both Figures 6 and 7 that, for our method, the inactive set termination condition is steadily reduced until a point is reached where the convergence becomes extremely rapid. This is consistent with the theoretical predictions in Theorem 3.13 and Corollary 3.14. We note that projected Newton-CG exhibits similar behaviour once it reaches Newton-CG step phase but to a lesser extent.

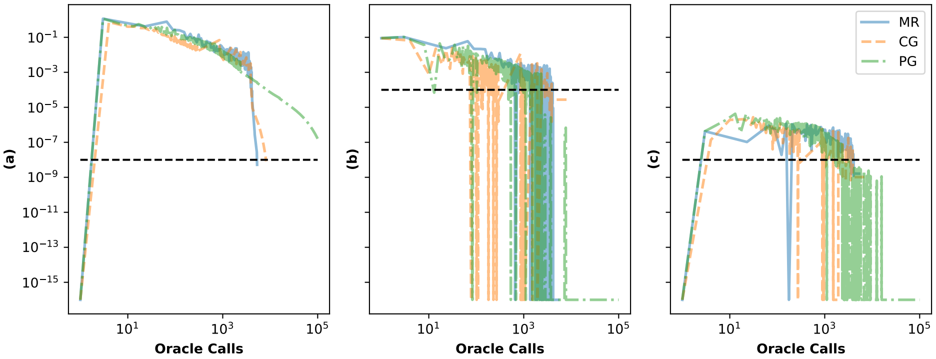

Figure 6. Termination conditions in (7) corresponding to experiment of Figure 1. (a) $\|\mathbf{g}_k^{\mathcal{I}}\|$, (b) $-\min(\mathbf{g}_k^{\mathcal{A}}, \mathbf{0})$ (min is taken elementwise) and (c) $\|\text{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\|$. The dashed line indicates the termination threshold for each of the respective conditions.

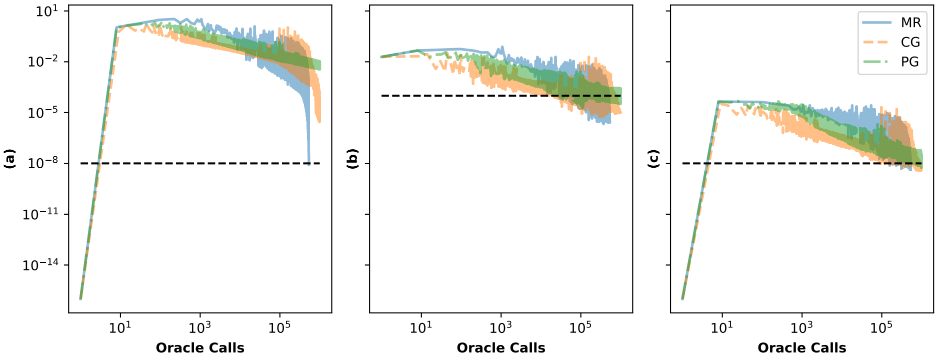

Figure 7. Termination conditions in (7) corresponding to experiment of Figure 2. (a) $\|\mathbf{g}_k^{\mathcal{I}}\|$, (b) $-\min(\mathbf{g}_k^{\mathcal{A}}, \mathbf{0})$ (min is taken elementwise) and (c) $\|\text{diag}(\mathbf{x}_k^{\mathcal{A}})\mathbf{g}_k^{\mathcal{A}}\|$. The dashed line indicates the termination threshold for each of the respective conditions.

F.5. Timing Results

For completeness, in the following section we give results presented in Section 4 in terms of "wall-clock" time. As noted earlier, wall-clock timing results are implementation and platform dependent. In particular, results are unreliable for small time scales. However, we note that, over larger time scales, the wall-clock time results generally conform with the corresponding oracle call results.

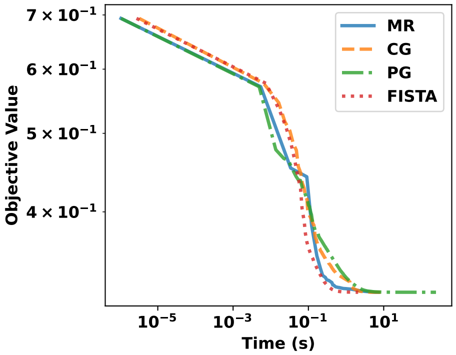

Figure 8. Wall-clock timing results for logistic regression ($C = 2$) on the binarised MNIST dataset (LeCun et al., 1998) ($d = 785$) with $\lambda = 10^{-3}$.

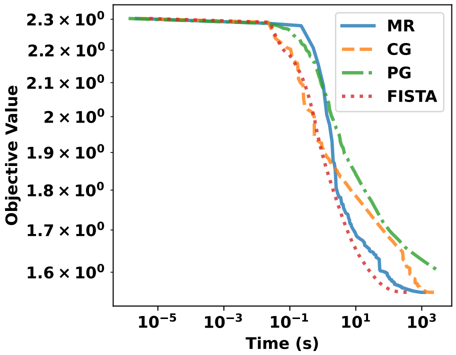

Figure 9. Wall-clock timing results for multinomial regression ($C = 10$) on CIFAR10 dataset (Krizhevsky, 2009) ($d = 27,657$) with $\lambda = 10^{-4}$.

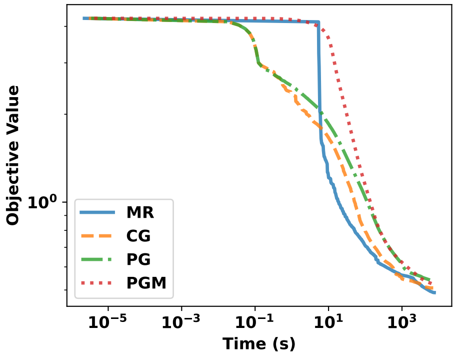

Figure 10. Wall-clock timing results for training a two-layer neural network on the `Fashion MNIST` dataset (Xiao et al., 2017) ($d = 89,610$) with $\lambda = 10^{-3}$.

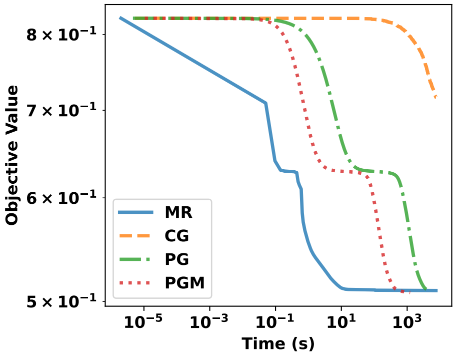

Figure 11. Wall-clock timing results for NNMF ($r = 20$) with cosine distance on top 1000 TF-IDF features of the `20 Newsgroup` dataset (Mitchell, 1999) ($d = 385,220$).

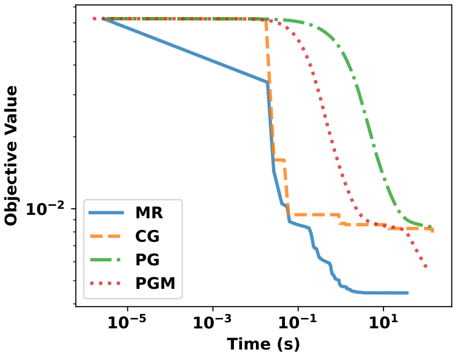

Figure 12. Wall-clock timing results for NNMF ($r = 10$) with nonconvex TSCAD regulariser on the `Olivetti faces` dataset (Pedregosa et al., 2011) ($d = 44,960$). We used $a = 3$ and $\lambda = 10^{-4}$ for the TSCAD regulariser.