Error estimates for iterative algorithms for minimizing regularized quadratic subproblems*

Nicholas I. M. Gould† and Valeria Simoncini‡

19th of March, 2019

Abstract

We derive bounds for the objective errors and gradient residuals when finding approximations to the solution of common regularized quadratic optimization problems within evolving Krylov spaces. These provide upper bounds on the number of iterations required to achieve a given stated accuracy. We illustrate the quality of our bounds on given test examples.

1 Introduction

In this paper, we derive upper bounds for the number of iterations required to reach a certain level of optimality by subspace methods for solving the trust-region subproblem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; q(x) := g^T x + \tfrac{1}{2} x^T H x \;\; \text{subject to} \;\; \|x\| \leq \delta \tag{1.1}$$

and its regularization variant

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; q^{\rm R}(x,\sigma,p) := q(x) + \tfrac{1}{p}\sigma\|x\|^p. \tag{1.2}$$

Here, we are given a gradient g, a symmetric, but possibly indefinite, Hessian H, a radius $\delta > 0$, a weight $\sigma > 0$ and a power $p > 2$, and use the Euclidean norm $\|\cdot\|$. Subproblems (1.1)–(1.2) lie at the heart of the step calculation in both trust-region and cubic-regularization methods for unconstrained optimization [6, 7, 19, 20].

A typical requirement in the trust-region case is that the computed x should decrease the objective function, i.e., $q(x) < q(0) \equiv 0$, and that the gradient of the Lagrangian for the problem, $g + Hx + \mu x$, should be smaller than a prescribed tolerance in norm, i.e.,

$$\|g + Hx + \mu x\| \leq \epsilon \tag{1.3}$$

*This work was supported by the EPSRC grant EP/M025179/1.

†Scientific Computing Department, STFC-Rutherford Appleton Laboratory, Chilton OX11 0QX, England. Email: nick.gould@stfc.ac.uk.

‡Dipartimento di Matematica, Università di Bologna, Piazza di Porta S.Donato 5, I-40127 Bologna & IMATI-CNR, Pavia, Italy. Email: valeria.simoncini@unibo.it.

for some given $\epsilon > 0$, whose precise value determines the rate of convergence of the trust-region algorithm, and a suitable Lagrange multiplier, $\mu \geq 0$, for the trust-region constraint $\|x\| \leq \delta$. For regularization problems, a similar requirement is that $q^{\rm R}(x,\sigma,p) < q^{\rm R}(0,\sigma,p) \equiv 0$ and that the norm of the gradient of $q^{\rm R}(x,\sigma,p)$ should be small. Since $\nabla_x q^{\rm R}(x,\sigma,p) = g + Hx + \mu x$ where $\mu = \sigma\|x\|^{p-2}$, the latter requirement is identical to (1.3) but for a different μ. As the subspace methods we consider automatically ensure that their relevant objectives decrease, our intention is to provide bounds on the number of steps (actually products with H) required by such methods to achieve (1.3) for the problems under consideration.

The subspaces of interest here are the nested Krylov spaces $\mathcal{K}_k := \mathcal{K}(H,g,k)$ for $k \geq 0$, where, for general A and b, we define $\mathcal{K}(A,b,k) := \mathrm{span}\{A^i b\}_{i=0}^{k-1}$. A sequence of estimates x_k are generated so that

$$x_k = \arg\min_{x \in \mathcal{K}_k} \; q(x) \;\; \text{subject to} \;\; \|x\| \leq \delta \tag{1.4}$$

for the trust-region subproblem, or

$$x_k = \arg\min_{x \in \mathcal{K}_k} \; q^{\rm R}(x,\sigma,p) \tag{1.5}$$

for the regularization case. This is useful as the well-known GLTR method [12] for (1.1) and the GLRT approach [6] for (1.2), which exploit the evolving Lanczos basis for $\mathcal{K}_k$, use precisely these formulations. However, we must be cautious as it is well known [12, Thm.5.8] that such methods may fail to solve the problem if the sequence of Krylov subspaces lies in an unpropitious non-trivial invariant subspace of $\mathbb{R}^n$, and in this case it may be necessary to enhance the search space with a specific eigenvector of H from outside the Krylov space. Fortunately, as we shall see, this is not necessary if our goal is merely to satisfy (1.3).

In §2 we examine the benefits and limitations of Krylov approximations to the solutions we wish to find. We follow this, in §3, by deriving bounds both on the decrease in the model objective functions and on the norm of the violation of the first-order criticality residuals from the Krylov space under consideration. We examine the latter on test examples that are designed to illustrate a variety of spectral distributions in §4. Finally, we make concluding remarks in §5.

2 Solutions from the Krylov space and beyond

Let $\lambda_1 \leq \ldots \leq \lambda_n$ be the eigenvalues of H, with the leftmost λ_1 having multiplicity n_1, and let u_i, $i \in \mathcal{N} := \{1,\ldots,n\}$ be the corresponding orthonormal eigenvectors. Crucially, there are well known characterisations of the global solutions of (1.1) and (1.2).

Theorem 2.1. [9, Thm.2.1; 18, Lem.2.1]. Any solution x_* to the trust-region subproblem (1.1) satisfies

$$(H + \mu_* I)x_* = -g, \tag{2.1}$$

where the Lagrange multiplier $\mu_* \geq \max(0, -\lambda_1)$ and $\mu_*(\|x_*\|^2 - \delta^2) = 0$. Moreover the solution is unique and $\mu_* > \max(0, -\lambda_1)$ whenever $g^T u_i \neq 0$ for some $1 \leq i \leq n_1$.

Theorem 2.2. [6, Thm.3.1; 19, Thm.10]. Any solution x_* to the regularization subproblem (1.2) satisfies (2.1), where the multiplier $\mu_* = \sigma\|x_*\|^{p-2} \geq -\lambda_1$. Moreover the solution is unique and $\mu_* > -\lambda_1$ whenever $g^T u_i \neq 0$ for some $1 \leq i \leq n_1$.

We consider the evolving Krylov spaces $\mathcal{K}_k$, $k \geq 0$, in more detail. Clearly we may decompose

$$g = \sum_{j=1}^{n} (g^T u_j) u_j$$

in terms of the basis of eigenvectors $\{u_j\}_{j\in\mathcal{N}}$ of H. Let $\mathcal{I}_+ := \{j \mid g^T u_j \neq 0\}$, $\mathcal{I}_0 := \mathcal{N} \setminus \mathcal{I}_+$ and $m := |\mathcal{I}_+|$.[1] Thus

$$g = \sum_{j\in\mathcal{I}_+} (g^T u_j) u_j \;\text{ and hence }\; H^i g = \sum_{j\in\mathcal{I}_+} \lambda_j^i (g^T u_j) u_j.$$

Therefore $\mathcal{K}_m = \cdots = \mathcal{K}_n$, since $\mathcal{K}_m = \text{span}\{u_j\}_{j\in\mathcal{I}_+}$ and the vectors $H^i g$ for $m < i \leq n$ are dependent on those in $\mathcal{K}_m$. Hence, our Krylov methods will make no further progress beyond the m-th iteration, and at that point provide estimates of their relevant solutions x_m and multipliers μ_m.

We now contrast x_m with the desired solution x_*. Let U_+ by the n by m matrix whose columns are the eigenvectors u_j, $j \in \mathcal{I}_+$, U_0 be the n by $n - m$ matrix whose columns are the remaining eigenvectors and $U = (U_+ : U_0)$. Likewise let Λ be the diagonal matrix of eigenvalues ordered as for U, and let Λ_+ and Λ_0 be its diagonal blocks. Thus

$$\Lambda = U^T H U = \begin{pmatrix} \Lambda_+ & 0 \\ 0 & \Lambda_0 \end{pmatrix}. \tag{2.2}$$

If we define $\bar{g} := U^T g$, and therefore $g = U\bar{g}$, this leads to

$$\begin{pmatrix} \bar{g}_+ \\ \bar{g}_0 \end{pmatrix} := \bar{g} = \begin{pmatrix} U_+^T g \\ U_0^T g \end{pmatrix} = \begin{pmatrix} U_+^T g \\ 0 \end{pmatrix} \;\text{ and }\; g = U_+ \bar{g}_+, \tag{2.3}$$

since $\bar{g}_0 = 0$ as $u_j^T g = 0$ for all $j \in \mathcal{I}_0$.

[1] This is equivalently the *grade* of H with respect to g

Consider the trust-region subproblem (1.1), and the change of variables $x = U\bar{x}$. In this case, $x_* = U\bar{x}_*$, where

$$\bar{x}_* = \underset{\bar{x}\in\mathbb{R}^n}{\arg\min}\ \bar{g}^T\bar{x} + \tfrac{1}{2}\bar{x}^T\Lambda\bar{x}\ \text{ subject to }\ \|\bar{x}\| \leq \delta. \tag{2.4}$$

The optimality conditions (2.1) for this are

$$\begin{pmatrix} \Lambda_+ + \mu_* I & 0 \\ 0 & \Lambda_0 + \mu_* I \end{pmatrix} \begin{pmatrix} \bar{x}_*^+ \\ \bar{x}_*^0 \end{pmatrix} = - \begin{pmatrix} \bar{g}_+ \\ 0 \end{pmatrix}, \tag{2.5}$$

where $\bar{x}_*$ and the Lagrange multiplier

$$\mu_* \geq \max(0, -\lambda_1) \tag{2.6}$$

satisfy

$$\|\bar{x}_*\| \leq \delta\ \text{ and }\ \mu_*(\|\bar{x}_*\|^2 - \delta^2) = 0, \tag{2.7}$$

and we have partitioned

$$\bar{x}_* = U^T x_* \equiv \begin{pmatrix} \bar{x}_*^+ \\ \bar{x}_*^0 \end{pmatrix}.$$

By contrast, if $x \in \mathcal{K}_m$, then $x = U_+\hat{x}_+$ for some vector $\hat{x}_+ \in \mathbb{R}^m$, in which case (1.4) gives $x_m = U_+\hat{x}_*^+$, where

$$\hat{x}_*^+ = \underset{\hat{x}_+\in\mathbb{R}^m}{\arg\min}\ \bar{g}_+^T\hat{x}^+ + \tfrac{1}{2}\hat{x}^{+T}\Lambda_+\hat{x}^+\ \text{ subject to }\ \|\hat{x}^+\| \leq \delta. \tag{2.8}$$

The optimality conditions (2.1) then imply that

$$(\Lambda_+ + \mu_*^+ I)\hat{x}_*^+ = -\bar{g}_+, \tag{2.9}$$

where $\hat{x}_*^+$ and the Lagrange multiplier

$$\mu_m \equiv \mu_*^+ > \max\left(0, -\min_{j\in\mathcal{I}_+}\lambda_j\right) \tag{2.10}$$

satisfy

$$\|\hat{x}^+\| \leq \delta\ \text{ and }\ \mu_*^+(\|\hat{x}_*^+\|^2 - \delta^2) = 0. \tag{2.11}$$

Given $\hat{x}_*^+$, let $\hat{x}^0 = 0$ and define

$$\hat{x}_* = \begin{pmatrix} \hat{x}_*^+ \\ \hat{x}_*^0 \end{pmatrix}\ \text{ so that }\ x_m = U\hat{x}_*.$$

In this case, (2.9) and (2.11) become

$$\begin{pmatrix} \Lambda_+ + \mu_*^+ I & 0 \\ 0 & \Lambda_0 + \mu_*^+ I \end{pmatrix} \begin{pmatrix} \hat{x}_*^+ \\ \hat{x}_*^0 \end{pmatrix} = - \begin{pmatrix} \bar{g}_+ \\ 0 \end{pmatrix}, \tag{2.12}$$

and

$$\|\hat{x}_*\| \leq \delta \;\text{ and }\; \mu_*^+(\|\hat{x}_*\|^2 - \delta^2) = 0. \tag{2.13}$$

Now compare $\bar{x}_*$ and μ_* from (2.5)–(2.7) with $\hat{x}_*$ and μ_*^+ from (2.10), (2.12) and (2.13). The only substantial difference is between (2.6) and (2.10). Indeed, if $\mu_*^+ \geq \max(0, -\lambda_1)$, the two sets of conditions are identical, and in this case $x_m = x_*$ and $\mu_m = \mu_*$, i.e., the solution from the subspace $\mathcal{K}_m$ solves the full-space trust-region problem (1.1). This must occur if $\min_{j\in\mathcal{I}_+} \lambda_j = \lambda_1$ or, equivalently $\mathcal{I}_+ \cap \{1,\ldots,n_1\} \neq \emptyset$, where we recall n_1 is the multiplicity of λ_1, but may also happen if $\min_{j\in\mathcal{I}_+} \lambda_j > \lambda_1$. If $\mu_*^+ < -\lambda_1$, $\mathcal{I}_+ \cap \{1,\ldots,n_1\} = \emptyset$, and x_m cannot solve (1.1), but it is nonetheless a critical point of the problem.[2] This possibility is often called the "hard case" [18] and $\mu_* = -\lambda_1$; the first block equation in (2.5) uniquely defines $\bar{x}_*^+$, and $\bar{x}_*^0$ is a multiple of any eigenvector of the second (singular) block, the precise combination ensuring that $\|\bar{x}_*\| = \delta$.

The main lesson here is that if we wish to solve (1.1) we shall have to look outside the Krylov space and may need to compute an eigenvector corresponding to λ_1. If we are content simply in finding a critical point of (1.1), the Krylov space suffices. An essentially identical argument may be used in the case of the regularization subproblem (1.2) with the same conclusions.

3 Error bounds

3.1 Bounds on the decrease of the objective functions

In essence Carmon and Duchi [4] provide the following bounds.[3]

Theorem 3.1. [4, Thm.1 & Cor.3]. Let λ_1 and λ_n be the leftmost and rightmost eigenvalues of H. Then, for all $k \geq 0$,
(i)

$$q(x_k) - q(x_*) \leq 36[q(0) - q(x_*)] \left(e^{-4\sqrt{\frac{\lambda_1+\mu_*}{\lambda_n+\mu_*}}} \right)^k, \tag{3.1}$$

where x_k is given by (1.4), x_* is a minimizer of (1.1), and μ_* is its corresponding Lagrange multiplier, and
(ii)

$$q^{\mathrm{R}}(x_k,\sigma,p) - q^{\mathrm{R}}(x_*,\sigma,p) \leq 36\,[q^{\mathrm{R}}(0,\sigma,p) - q^{\mathrm{R}}(x_*,\sigma,p)] \left(e^{-4\sqrt{\frac{\lambda_1+\mu_*}{\lambda_n+\mu_*}}} \right)^k, \tag{3.2}$$

where x_k is given by (1.5), x_* is a minimizer of (1.2) and $\mu_* = \sigma\|x_*\|$.

[2] It will only be a local minimizer of $\mu_*^+ > -\lambda_2$ [17].
[3] Strictly [4, Cor.3] only considers the case $p = 3$, but their method of proof holds in general.

Thus the error in the relevant objective function decreases at worst linearly as a function of the subspace dimension unless $\mu_* = -\lambda_1$, in which case Theorem 3.1 provides no useful bound. As we have already mentioned, the unlikely "hard case" $\mu_* = -\lambda_1$ only occurs if g is orthogonal to the space of eigenvectors corresponding to the eigenvalue λ_1 of H, and should this happen these eigenvectors will not occur in the Krylov spaces $\mathcal{K}_k$, except through numerical rounding. A simple expedient advocated by others [4] is to perturb g by a small random vector.

We note that Carmon and Duchi actually provide a second, sublinear decrease estimate that may be less pessimistic for small k, but we shall not use this here.

We now restrict our attention to the best estimate x_m available from the evolving Krylov space. We exclude the special case $g = 0$ since then $x = 0$ is a critical point of both of the subproblems under consideration.

Corollary 3.2. Suppose that $g \neq 0$. Let $\{\lambda_j, u_j\}_{j\in\mathcal{N}}$ be eigenpairs of H, $\mathcal{I}_+ = \{j \mid g^T u_j \neq 0\}$, $m = |\mathcal{I}_+|$, and

$$\lambda_+^{\min} = \min_{j\in\mathcal{I}_+} \lambda_i \;\text{ and }\; \lambda_+^{\max} = \max_{j\in\mathcal{I}_+} \lambda_i. \tag{3.3}$$

Then, for all $k \geq 0$,
(i)

$$q(x_k) - q(x_m) \leq 36\,[q(0) - q(x_m)] \left(e^{-\frac{4}{\sqrt{\boldsymbol{\kappa}_m}}}\right)^k, \tag{3.4}$$

where x_k and x_m are given by (1.4),

$$\boldsymbol{\kappa}_m := \frac{\lambda_+^{\max} + \mu_m}{\lambda_+^{\min} + \mu_m}, \tag{3.5}$$

and μ_m is the Lagrange multiplier corresponding to x_m, and
(ii)

$$q^{\mathrm{R}}(x_k, \sigma, p) - q^{\mathrm{R}}(x_m, \sigma, p) \leq 36\,[q^{\mathrm{R}}(0, \sigma, p) - q^{\mathrm{R}}(x_m, \sigma, p)] \left(e^{-\frac{4}{\sqrt{\boldsymbol{\kappa}_m}}}\right)^k, \tag{3.6}$$

where x_k and x_m are given by (1.5), $\boldsymbol{\kappa}_m$ is given by (3.5) but now $\mu_m = \sigma\|x_m\|$.

Proof. Since $g \neq 0$, $\mathcal{I}_+ \neq \emptyset$, $m > 0$ and both $\lambda_+^{\min}$ and $\lambda_+^{\max}$ are well defined. Let H_+ be the matrix with eigenpairs $\{\lambda_j, u_j\}$ for $j \in \mathcal{I}_+$ and $\{\lambda_+^{\max}, u_j\}$ for $j \in \mathcal{I}_0 = \mathcal{N}\backslash\mathcal{I}_+$. Then $\mathcal{K}_k = \mathrm{span}\{H^i g\}_{i=0}^{k-1} = \mathrm{span}\{H_+^i g\}_{i=0}^{k-1}$ for $k \geq 0$, and the iterates x_k generated from the Krylov spaces $\mathcal{K}_k$ for (1.4) and (1.5) for the problem with Hessian H_+ are identical to those with Hessian H. However the hard case cannot occur with the Hessian H_+ as none of the eigenvalues for $j \in \mathcal{I}_0$ is smaller than the smallest for $j \in \mathcal{I}_+$. Hence, as we saw in §2, for this Hessian $x_* = x_m$ and $\mu_* = \mu_m$. Thus we may apply Theorem 3.1 for

the problem with Hessian H_+ to deduce (3.4) and (3.6). □

As Carmon and Duchi mention, this then implies a worst-case estimate of

$$k \leq \min(m, O(\sqrt{\boldsymbol{\kappa}_m}\log(1/\epsilon))) \tag{3.7}$$

iterations in order to guarantee $q(x_k) - q(x_m) \leq \epsilon$ or $q^{\mathrm{R}}(x_k, \sigma, p) - q^{\mathrm{R}}(x_m, \sigma, p) \leq \epsilon$ as appropriate.

3.2 Bounds on the decrease of the gradients of the objective functions

Recall that the orthonormal Lanczos basis matrix $V_k \in \mathbb{R}^{n\times k}$ for $\mathcal{K}_k$ satisfies

$$HV_k = V_k T_k + \gamma_k v_{k+1} e_k^T, \tag{3.8}$$

where

$$T_k = \begin{pmatrix} \delta_1 & \gamma_1 & & & \\ \gamma_1 & \delta_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \delta_{k-1} & \gamma_{k-1} \\ & & & \gamma_{k-1} & \delta_k \end{pmatrix} \tag{3.9}$$

is tridiagonal and the γ_i, $i = 1, \ldots, k-1$ with $k \leq m$, are strictly positive [1,14,15,23]. As the off diagonals $\gamma_i > 0$, T_k is irreducible, and has distinct real eigenvalues (the so-called Ritz values) $\theta_{i,k}$, $i = 1, \ldots, k$, arranged in increasing order. It is well known [10, Cor.8.1.7] that the Ritz values satisfy the interlacing properties

$$\theta_{i,k} \leq \theta_{i,k+1} \leq \theta_{i+1,k} \tag{3.10}$$

for $i = 1, \ldots, k$, $k < m$, and

$$\lambda_+^{\min} = \theta_{1,m} \leq \theta_{1,k+1} \leq \theta_{1,k} \leq \theta_{1,1} \equiv \frac{g^T H g}{\|g\|^2} \leq \theta_{k,k} \leq \theta_{k+1,k+1} \leq \theta_{m,m} = \lambda_+^{\max} \tag{3.11}$$

for $k = 1, \ldots, m-1$, where $\lambda_+^{\min}$ and $\lambda_+^{\max}$ are defined in (3.3).

Since $v_1 = g/\|g\|$, it follows that

$$V_k^T g = \|g\| e_1 \ \text{ and } \ g = \|g\| V_k e_1 \tag{3.12}$$

as V_k has orthonormal columns. Furthermore pre-multiplying (3.8) by V_k^T yields

$$V_k^T H V_k = T_k,$$

and thus the definition (1.4) implies that

$$x_k = V_k y_k, \ \text{ where } \ (T_k + \mu_k I) y_k = V_k^T (H + \mu_k I) V_k y_k = -V_k^T g = -\|g\| e_1. \tag{3.13}$$

Moreover, applying Theorem 2.1 to (1.4) shows that $T_k + \mu_k I$ is positive definite.

Let

$$r_k := g + Hx_k + \mu_k x_k = g + (H + \mu_k I)x_k. \tag{3.14}$$

It then follows from (3.8), (3.12) and (3.13) that

$$\begin{aligned} Hx_k &= HV_k y_k = V_k T_k y_k + \gamma_k e_k^T y_k v_{k+1} = -V_k(\mu_k y_k + \|g\| e_1) + \gamma_k e_k^T y_k v_{k+1} \\ &= -\mu_k x_k - g + \gamma_k e_k^T y_k v_{k+1}. \end{aligned}$$

Hence $r_k = \gamma_k e_k^T y_k v_{k+1}$ and

$$\|r_k\| = \gamma_k \left| e_k^T y_k \right| = \gamma_k \|g\| \left| e_k^T (T_k + \mu_k I)^{-1} e_1 \right|. \tag{3.15}$$

Note that the definition of $\gamma_k > 0$ as the $(k, k+1)$-st entry of T_{k+1} and the Cauchy Schwarz inequality implies that

$$\gamma_k = e_{k+1}^T T_{k+1} e_k \leq \|T_{k+1}\| = \|V_{k+1}^T H V_{k+1}\| \leq \|H\|. \tag{3.16}$$

Thus, aside from the term $\gamma_k \|g\| > 0$, the residual norm decays with $\left| e_k^T (T_k + \mu_k I)^{-1} e_1 \right|$, and we now focus on this.

We recall a vital result by Demko, Moss and Smith [8] on the component-wise decay of the inverse of symmetric banded matrices. Here the bandwidth of a banded symmetric matrix M is the number of nonzero upper (or equivalently lower) super diagonals, and, if M is additionally positive definite, $\boldsymbol{\kappa}(M) := \lambda_{\max}(M)/\lambda_{\min}(M)$ is its spectral condition number, where $0 < \lambda_{\min}(M) \leq \lambda_{\max}(M)$ are the left- and right-most eigenvalues of M.

Lemma 3.3. [8, Thm.2.4]. Let $M \in \mathbb{R}^{n \times n}$ be a symmetric, positive definite matrix with bandwidth $\beta > 0$. Then

$$|(M^{-1})_{i,j}| \leq c t^{\frac{|i-j|}{\beta}}$$

for all $i, j = 1, \ldots, n$, where

$$c = \frac{1}{\lambda_{\min}(M)} \max\left(1, \frac{\left(\sqrt{\boldsymbol{\kappa}(M)} + 1\right)^2}{2\boldsymbol{\kappa}(M)}\right) \quad \text{and} \quad t = \frac{\sqrt{\boldsymbol{\kappa}(M)} - 1}{\sqrt{\boldsymbol{\kappa}(M)} + 1}.$$

Note that we shall prefer the slightly weaker, but simpler, bound

$$c \leq \frac{2}{\lambda_{\min}(M)}, \tag{3.17}$$

and indeed $c = 1/\lambda_{\min}(M)$ so long as $\boldsymbol{\kappa}(M) \geq \sqrt{1 + \sqrt{2}}$.

For any $k \leq m$, we have that T_k from (3.9) is symmetric and tridiagonal with left- and right-most eigenvalues (Ritz values) respectively $\theta_{1,k} < \theta_{k,k}$. As we have seem $T_k + \mu_k I$ is positive definite, and thus has distinct left- and right-most eigenvalues

$$\lambda_{\min}(T_k + \mu_k I) \equiv \theta_{1,k} + \mu_k < \lambda_{\max}(T_k + \mu_k I) \equiv \theta_{k,k} + \mu_k \tag{3.18}$$

as well as spectral condition number

$$\boldsymbol{\kappa}_k := \boldsymbol{\kappa}(T_k + \mu_k I) = \frac{\theta_{k,k} + \mu_k}{\theta_{1,k} + \mu_k}. \tag{3.19}$$

We may apply Lemma 3.3 to $T_k + \mu_k I$ to deduce our main result.

Theorem 3.4. The residual (3.14) for the k-th iterate, x_k^*, generated by either the trust-region subproblem (1.4) or the regularization subproblem (1.5) satisfies the bound

$$\|r_k\| \leq \|g\| \left(\frac{2\gamma_k \boldsymbol{\kappa}_k}{\theta_{k,k} + \mu_k} \right) \left(\frac{\sqrt{\boldsymbol{\kappa}_k} - 1}{\sqrt{\boldsymbol{\kappa}_k} + 1} \right)^{k-1}, \tag{3.20}$$

where $\boldsymbol{\kappa}_k$ is given by (3.19) and γ_k is the $(k, k+1)$-st entry of T_{k+1}.

Proof. Since $\left| e_k^T (T_k + \mu_k I)^{-1} e_1 \right| = \left| ((T_k + \mu_k I)^{-1})_{k,1} \right|$, we may apply Lemma 3.3 to $T_k + \mu_k I$, with $\beta = 1$, together with (3.17)–(3.19) to deduce the bound

$$\left| e_k^T (T_k + \mu_k I)^{-1} e_1 \right| \leq c_k t_k^{k-1} \tag{3.21}$$

for all $k \leq m$, where

$$c_k = \frac{2}{\theta_{1,k} + \mu_k} \equiv \frac{2\boldsymbol{\kappa}_k}{\theta_{k,k} + \mu_k} \quad \text{and} \quad t_k = \frac{\sqrt{\boldsymbol{\kappa}_k} - 1}{\sqrt{\boldsymbol{\kappa}_k} + 1}. \tag{3.22}$$

The desired bound (3.20) then follows directly from (3.15) and (3.22). □

In the trust-region case, this leads to the following residual bound[4].

Corollary 3.5. The residual (3.14) for the k-th iterate, x_k^*, generated by the trust-region subproblem (1.4) satisfies the bound

$$\|r_k\| \leq \|g\| \left(\frac{2\delta \|H\| \boldsymbol{\kappa}_k}{\|g\|} \right) \left(\frac{\sqrt{\boldsymbol{\kappa}_k} - 1}{\sqrt{\boldsymbol{\kappa}_k} + 1} \right)^{k-1}, \tag{3.23}$$

where $\boldsymbol{\kappa}_k$ is given by (3.19).

[4]Another thing we know but we have not used.: $0 \leq \mu_1 \leq \mu_k \leq \mu_m$ for $k = 1, \ldots, m$ [16].

Proof. It follows from (3.13), the Cauchy-Schwarz and Rayleigh-Ritz inequalities and the bound $\|y_k\| \leq \delta$ that

$$\|g\|^2 = \|(T_k + \mu_k I)y_k\|^2 \leq (\theta_{k,k} + \mu_k)^2 \|y_k\|^2 \leq (\theta_{k,k} + \mu_k)^2 \delta^2,$$

and hence

$$\frac{1}{\theta_{k,k} + \mu_k} \leq \frac{\delta}{\|g\|}. \tag{3.24}$$

Thus combining (3.16), (3.20) and (3.24), we find that (3.23) holds. □

Corollary 3.6. Let $\boldsymbol{\kappa}_* = \max_{1 \leq k \leq m} \boldsymbol{\kappa}_k$. Then the iteration defined by (1.4) satisfies

$$\|r_k\| \leq \epsilon \tag{3.25}$$

as soon as

$$k \leq \min \left[m, \left\lceil \log \left(\frac{2\delta \|H\| \boldsymbol{\kappa}_*}{\epsilon} \right) / \log \left(\frac{\sqrt{\boldsymbol{\kappa}_*} - 1}{\sqrt{\boldsymbol{\kappa}_*} + 1} \right) \right\rceil + 1 \right]. \tag{3.26}$$

Proof. Since the function

$$q(\boldsymbol{\kappa}) = \frac{\sqrt{\boldsymbol{\kappa}} - 1}{\sqrt{\boldsymbol{\kappa}} + 1}$$

is monotonically increasing for $\boldsymbol{\kappa} \geq 1$, we deduce from Corollary 3.5 that

$$\|r_k\| \leq \|g\| \left(\frac{2\delta \|H\| \boldsymbol{\kappa}_*}{\|g\|} \right) \left(\frac{\sqrt{\boldsymbol{\kappa}_*} - 1}{\sqrt{\boldsymbol{\kappa}_*} + 1} \right)^{k-1}. \tag{3.27}$$

Recalling that $r_m = 0$, (3.25) and (3.27) lead directly to (3.26). □

A similar result is possible for the regularization case.

Corollary 3.7. The residual (3.14) for the k-th iterate, x_k^*, generated by the regularization subproblem (1.5) satisfies the bound

$$\|r_k\| \leq \|g\| \left(\frac{2\|H\| \boldsymbol{\kappa}_k}{\|g\|} \right) \left(\frac{\mu_k}{\sigma} \right)^{\frac{1}{p-2}} \left(\frac{\sqrt{\boldsymbol{\kappa}_k} - 1}{\sqrt{\boldsymbol{\kappa}_k} + 1} \right)^{k-1}, \tag{3.28}$$

where $\boldsymbol{\kappa}_k$ is given by (3.19).

Proof. It follows from (3.13), the Cauchy-Schwarz and Rayleigh-Ritz inequalities and the relationship $\mu_k = \sigma\|y_k\|^{p-2}$ that

$$\|g\|^2 = \|(T_k + \mu_k I)y_k\|^2 \leq (\theta_{k,k} + \mu_k)^2\|y_k\|^2 = (\theta_{k,k} + \mu_k)^2 \left(\frac{\mu_k}{\sigma}\right)^{\frac{2}{p-2}}$$

and hence

$$\frac{1}{\theta_{k,k} + \mu_k} \leq \frac{1}{\|g\|} \left(\frac{\mu_k}{\sigma}\right)^{\frac{1}{p-2}}. \tag{3.29}$$

Thus (3.28) follows by combining (3.16), (3.20) and (3.29). □

Corollary 3.8. Let $\boldsymbol{\kappa}_* = \max_{1\leq k\leq m} \boldsymbol{\kappa}_k$. Then the iteration defined by (1.4) satisfies (3.25) as soon as

$$k \leq \min\left[m, \left\lceil \log\left(\frac{2\|H\|\boldsymbol{\kappa}_*}{\epsilon}\left(\frac{\mu_m}{\sigma}\right)^{\frac{1}{p-2}}\right) / \log\left(\frac{\sqrt{\boldsymbol{\kappa}_*} - 1}{\sqrt{\boldsymbol{\kappa}_*} + 1}\right)\right\rceil + 1\right]. \tag{3.30}$$

Proof. Given Corollary 3.7, the proof is essentially identical to that of Corollary 3.6, except that we additionally make use of the bound $0 \leq \mu_k \leq \mu_m$ for $k = 1, \ldots, m$ [5, Thm.2.5]. □

3.3 Comments

We now comment on the bounds obtained in §3.1 and 3.2. Those on the (linear) rates of convergence given in Corollaries 3.2, 3.6 and 3.8 are very typical of the Chebyshev bounds that have been derived for conjugate-gradient (CG)-llke methods for solving symmetric, positive-definite systems of linear equations $Ax = b$ (see, e.g., [15, §5.6.2]). Briefly, in this case $\|r_k\|_{A^{-1}} = \|x_k - x_*\|_A$, where $r_k = Ax_k - b$ and $x_* = A^{-1}b$. Since $\|r_k\| \leq \sqrt{\lambda_{\max}(A)}\|r_k\|_{A^{-1}}$, the argument in the CG case focuses on finding an upper bound on $\|x_k - x_*\|_A$. In particular, x_k is chosen to minimize $\|x - x_*\|_A$ over all $x \in \mathcal{K}(A, b, k)$, and this is easily shown to lead to the bound

$$\|x_k - x_*\|_A \leq \|x_0 - x_*\|_A \min_{\psi\in\mathcal{P}_k} \max_{i\in\mathcal{N}} |\psi(\lambda_i(A))|, \tag{3.31}$$

where

$$\mathcal{P}_k = \{\text{polynomials } \psi \text{ of degree } k \text{ for which } \psi(0) = 1\}$$

and $\lambda_i(A)$, $i \in \mathcal{N}$, are a subset of the eigenvalues of A. Weakening the requirement that the maximum in (3.31) instead considers $\psi(\lambda)$ over all $\lambda \in [\lambda_{min}(A), \lambda_{max}(A)]$ and invoking a well-known bound from approximation theory relating to Chebyshev polynomials, it then

follows that

$$\|x_k - x_*\|_A \leq 2\left(\frac{\sqrt{\boldsymbol{\kappa}(A)}-1}{\sqrt{\boldsymbol{\kappa}(A)}+1}\right)^k \|x_0 - x_*\|_A.$$

Since $\|x_0 - x_*\|_A \leq \|r_0\|/\sqrt{\lambda_{\min}(A)}$, we thus obtain the bound

$$\|r_k\| \leq 2\sqrt{\boldsymbol{\kappa}(A)}\left(\frac{\sqrt{\boldsymbol{\kappa}(A)}-1}{\sqrt{\boldsymbol{\kappa}(A)}+1}\right)^k \|r_0\|;$$

if $x_0 = 0$, $\|r_0\| = \|b\|$ in the latter.

The presence of $\boldsymbol{\kappa}_m \leq \boldsymbol{\kappa}_*$ in the bounds in §3.1 and 3.2 is strongly reminiscent of the CG case, and indicates that rescaling (preconditioning) the problem so that $\boldsymbol{\kappa}_m$ or $\boldsymbol{\kappa}_* = O(1)$ would be beneficial. In the strictly convex case when H is positive definite, $\boldsymbol{\kappa}_m$ is no larger than the traditional condition number $\lambda_+^{\max}/\lambda_+^{\min}$ obtained from (3.11). Although we know that $\theta_{k,k}$ increases monotonically from (3.11), as does μ_k [5,16], we have not been able to prove that $\boldsymbol{\kappa}_k$ increases monotonically,[5] albeit in practice it appears to.

We need to be very cautious here as although such bounds accurately predict the worst possible case [4,14], they are often very pessimistic in general, a point stressed in [15]. We shall return to this in §4. Nevertheless, if one is interested in the worst-case, bounds such as (3.7), (3.26) and (3.30) are relevant.

We tried two other approaches to derive useful bounds on the norm of the residual, (3.14). The first aims to use the known decrease in the model objective given by Corollary 3.2 to deduce a similar bound on $\|r_k\|$. Since x_m from (1.4) is a critical point of (1.1), we have that

$$g + Hx_m + \mu_m x_m = 0, \tag{3.32}$$

and hence

$$\begin{aligned} r_k &:= g + Hx_k + \mu_k x_k = -Hx_m - \mu_m x_m + Hx_k + \mu_k x_k \\ &= (H + \mu_m I)(x_k - x_m) + (\mu_k - \mu_m)x_k \end{aligned}. \tag{3.33}$$

Elementary manipulation of these (see Appendix A) then leads to the bound

$$\begin{aligned} &(\lambda_+^{\max} + \mu_m)^{-1}\|r_k\|^2 \leq 2\left[q(x_k) - q(x_m)\right] + \rho_k, \quad \text{where} \\ &\rho_k := \mu_k(\|x_k\|^2 - \|x_m\|^2) - (\mu_m - \mu_k)\|x_m - x_k\|^2 + (\mu_m - \mu_k)^2 x_k^T (H + \mu_m I)^{-1} x_k, \end{aligned} \tag{3.34}$$

which exposes the dependence on the model objective decrease $q(x_k) - q(x_m)$. Unfortunately, aside from the case where $\mu_m = 0$ for which $\rho_k = 0$, we are not able to find a useful bound on ρ_k; ideally we would like to show that $\rho_k \leq 0$. Of course, even had we had succeeded in bounding ρ_k, the overall bound we would have obtained via Corollary 3.2 for the $q(x_k) - q(x_m)$ term would not have been substantially different from those given in Corollaries 3.6 and 3.8.

Our second approach tried to mimic that taken for the CG method for positive-definite linear systems. However, the argument relating the A-norm of the error to the A^{-1}-norm

[5] The result would follow if we could show that $\theta_{k,k} + \mu_k$ decreases monotonically with growing k.

of the residual and the subsequent min-max characterisation depends crucially on the definiteness of A, and thus this line of attack is not obvious for our case where H may be indefinite. Nevertheless it is easy to derive the bound

$$\|r_k\| \leq \max_{j \in \mathcal{I}_+} |\psi_k(\lambda_j + \mu_k)| \|g\|, \quad \text{where} \quad \psi_k(\lambda_j + \mu_k) = \prod_{i=1}^{k} \frac{(\theta_{i,k} - \lambda_j)}{(\theta_{i,k} + \mu_k)}. \tag{3.35}$$

on the residual (3.14) (see Appendix B). Although we do not know how to derive a useful bound on $\psi_k(\lambda_j + \mu_k)$, as we see in §4, to do so might provide a much closer match to the true residual than provided by Corollaries 3.6 and 3.8.

4 Experiments

We consider nine examples that aim to illustrate our analysis; all nine are available as part of the CUTEst [11] set of test problems. Each is of the form

$$q(x) = \tfrac{1}{2} \sum_{i=1}^{n} d_l x_i^2 + \sum_{i=1}^{n} x_i,$$

but vary according to the diagonal Hessian elements, d_i. Specifically we have examples

`DIAGPQT`: $d_i = -i^2/n + n + 1/n$,
`DIAGPQE`: $d_i = i$,
`DIAGPQB`: $d_i = i^2/n$,
`DIAGIQT`: $d_i = -i^2/n + n/2 + 1/n$,
`DIAGIQE`: $d_i = i - n/2$,
`DIAGIQB`: $d_i = i^2/n - n/2 + 1/n$,
`DIAGNQT`: $d_i = -i^2/n$,
`DIAGNQE`: $d_i = i - n - 1$, and
`DIAGNQB`: $d_i = i^2/n - n - 1/n$,

for $i = 1, \ldots, n$; in our tests we let $n = 1000$, and ignore the additional simple-bound constraints specified in the CUTEst examples. The first three are convex with Hessian spectra that bunch towards the top of the range, that are equispaced, and that coalesce towards the bottom of the range respectively. The second three shift the spectra of the first three downwards by $n/2$, leading to indefinite Hessians, while the last three concave examples shift downwards by $n + 1$.

In Figure 4.1 we compare the true residual against bounds derived in Section 3 when running GALAHAD's [13] `GLTR` package [12] to solve the trust-region subproblem (1.1) on the first three test examples. Almost identical plots have been obtained for the remaining examples for the trust-region case since the residual

$$r_k = Hx_k + \mu_k x_k + g = (H - \omega I)x_k + (\mu_k + \omega)x_k + g$$

shows that shifting the Hessian downwards by ω is compensated by shifting the multiplier upwards by the same amount once the trust-region constraint is active.

We observe that although Theorem 3.4 and Corollary 3.5 provide bounds on the residual, they may be far from sharp, especially when the spectrum is equispaced or bunched towards the top end. In particular, the bounds do not capture the superlinear behaviour of the residuals in these cases; the slopes best mimic those from the earlier iterations. This largely agrees with the observations made and conclusions drawn in the linear-equation case [15]. The inferiority of the bound in Corollary 3.5 compared to that in Theorem 3.4 merely reflects the weakening that results when approximating unknown quantities (i.e., $\theta_{k,k} + \mu_k$) by known ones (i.e., H, g, δ). By contrast, the bound provided by (3.35) is quantitatively far better, but, of course, this requires full knowledge of the spectrum.

Figures 4.2–4.4 compare the estimates (3.20), (3.28) and (3.35) against the true residual when running GALAHAD's `GLRT` package [6] to solve the regularization subproblem (1.2) on all nine test examples; unlike for the trust-region case, a translation of the Hessian does not produce essentially identical plots when moving from the convex via the indefinite to the concave cases.

Once again, we observe that the bounds (3.20), (3.28) may be far from sharp, and can fail to reflect the later superlinear convergence of the residuals. The behaviour is most extreme for the concave examples, and for those whose spectra coalesce at the top of their ranges. As before, (3.35) provides a much more faithful bound.

Finally Figures 4.5 illustrate the effect of changing the regularization weight, σ, when solving (1.2) for the example `DIAGNQT`, on the residual estimates. The subproblems become increasingly hard as σ shrinks, and the estimates correspondingly poorer. Indeed, the decrease predicted by (3.20) when $\sigma = 100$ barely indicates convergence, while although the rates for the actual residual and the prediction (3.35) are initially slow, they later accelerate.

5 Conclusions

We have derived bounds for the objective errors and gradient residuals when finding approximations to the solution of common regularized quadratic optimization problems within evolving Krylov spaces. Those for the objective errors are trivial extensions of existing ones [4], while those for the gradient residuals generalize well-known ones for conjugate-gradient methods applied to definite linear systems. Quantitatively the bounds behave just as in the conjugate-gradient case, but reflect additional complication of the subproblems, particularly the potential indefiniteness of the matrices involved.

We express some caution since in exceptional cases Krylov methods may not find the global solutions to our problems. If this is the goal, then additional precautions [4, 12] that are not covered by our bounds may be necessary. If our goal is simply to find an approximation that yields a small gradient residual—and this is often the case when the subproblem occurs as component of a more general optimization calculation—then our bounds are appropriate, and provide upper bounds on the number of iterations required to achieve a given stated accuracy.

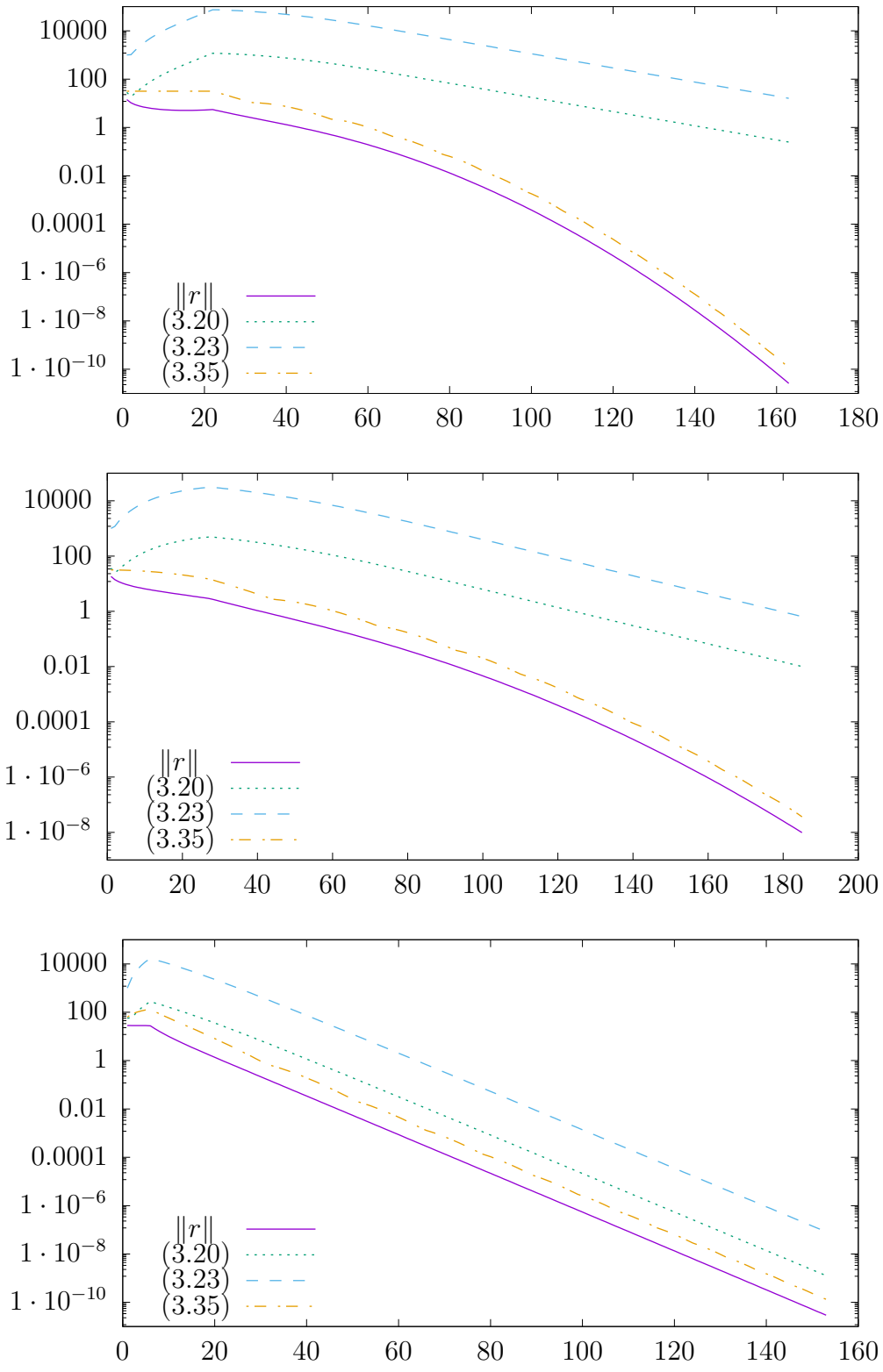

Figure 4.1: $\log_{10}$ of the residual (y-axis) as the iteration proceeds (x-axis) for GLTR as applied to the convex problems `DIAGPQT` (top plot), `DIAGPQE` (middle) and `DIAGPQB` (bottom) with a trust-region radius $\delta = 1$. Each figure shows the residual (3.15) (solid line), and the estimates (3.20) (dotted line), (3.23) (dashed line) and (3.35) (dash-dot line).

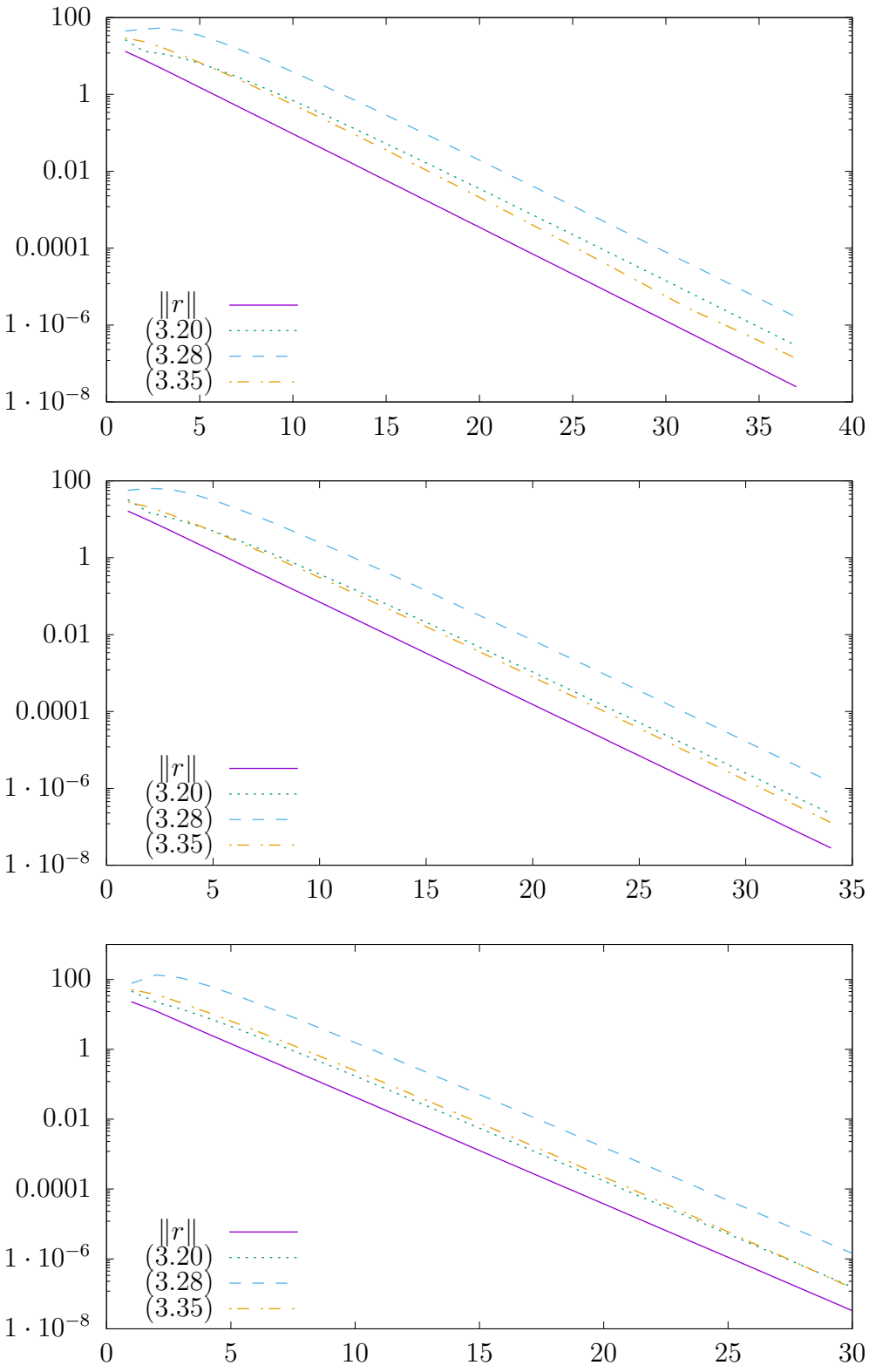

Figure 4.2: $\log_{10}$ of the residual (y-axis) as the iteration proceeds (x-axis) for GLRT as applied to the convex problems `DIAGPQT` (top plot), `DIAGPQE` (middle) and `DIAGPQB` (bottom) with a regularization weight $\sigma = 1000$ and $p = 3$. Each figure shows the residual (3.15) (solid line), and the estimates (3.20) (dotted line), (3.28) (dashed line) and (3.35) (dash-dot line).

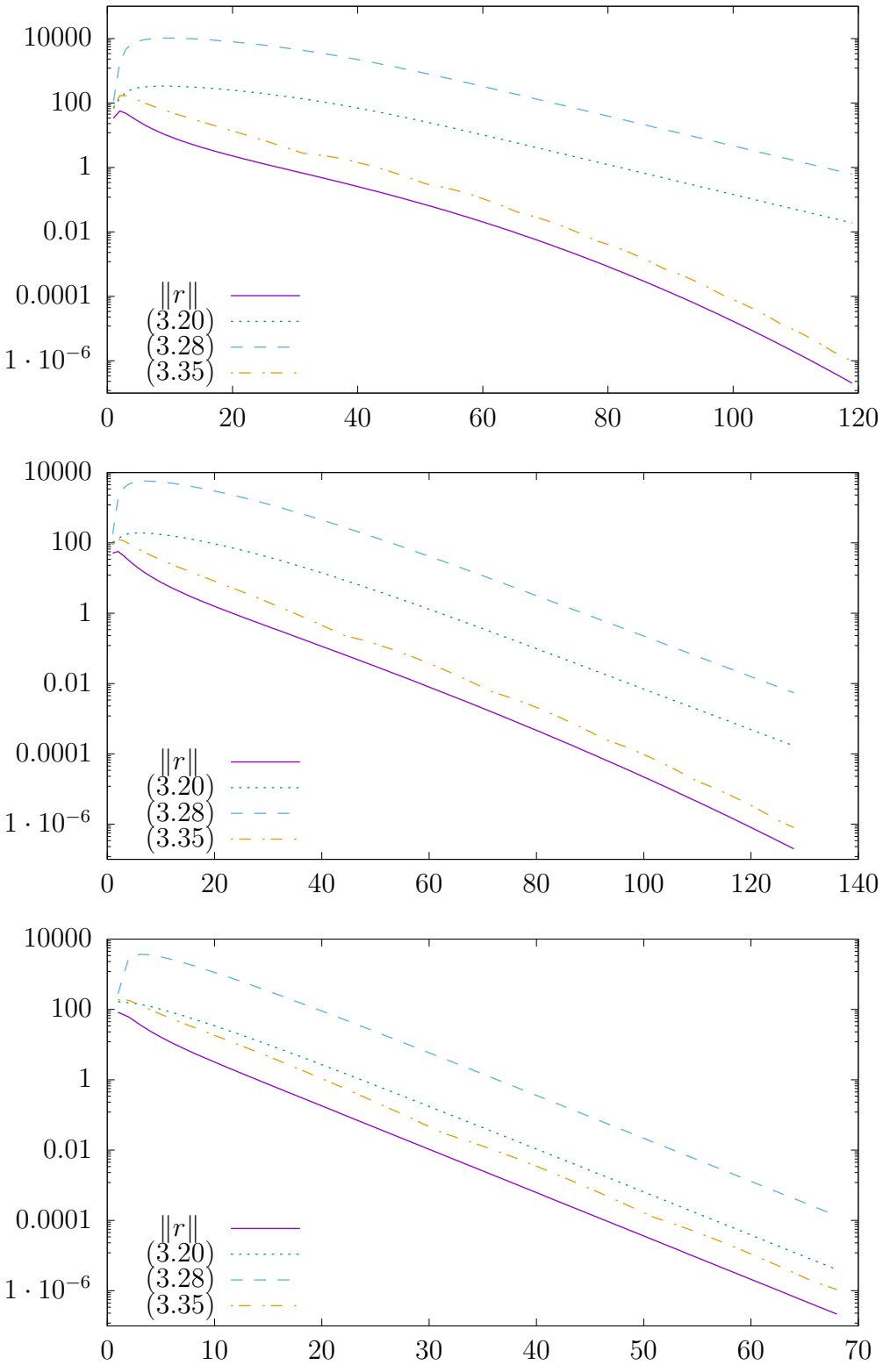

Figure 4.3: $\log_{10}$ of the residual (y-axis) as the iteration proceeds (x-axis) for GLRT as applied to the indefinite problems `DIAGIQT` (top plot), `DIAGIQE` (middle) and `DIAGIQB` (bottom) with a regularization weight $\sigma = 1000$ and $p = 3$. Each figure shows the residual (3.15) (solid line), and the estimates (3.20) (dotted line), (3.28) (dashed line) and (3.35) (dash-dot line).

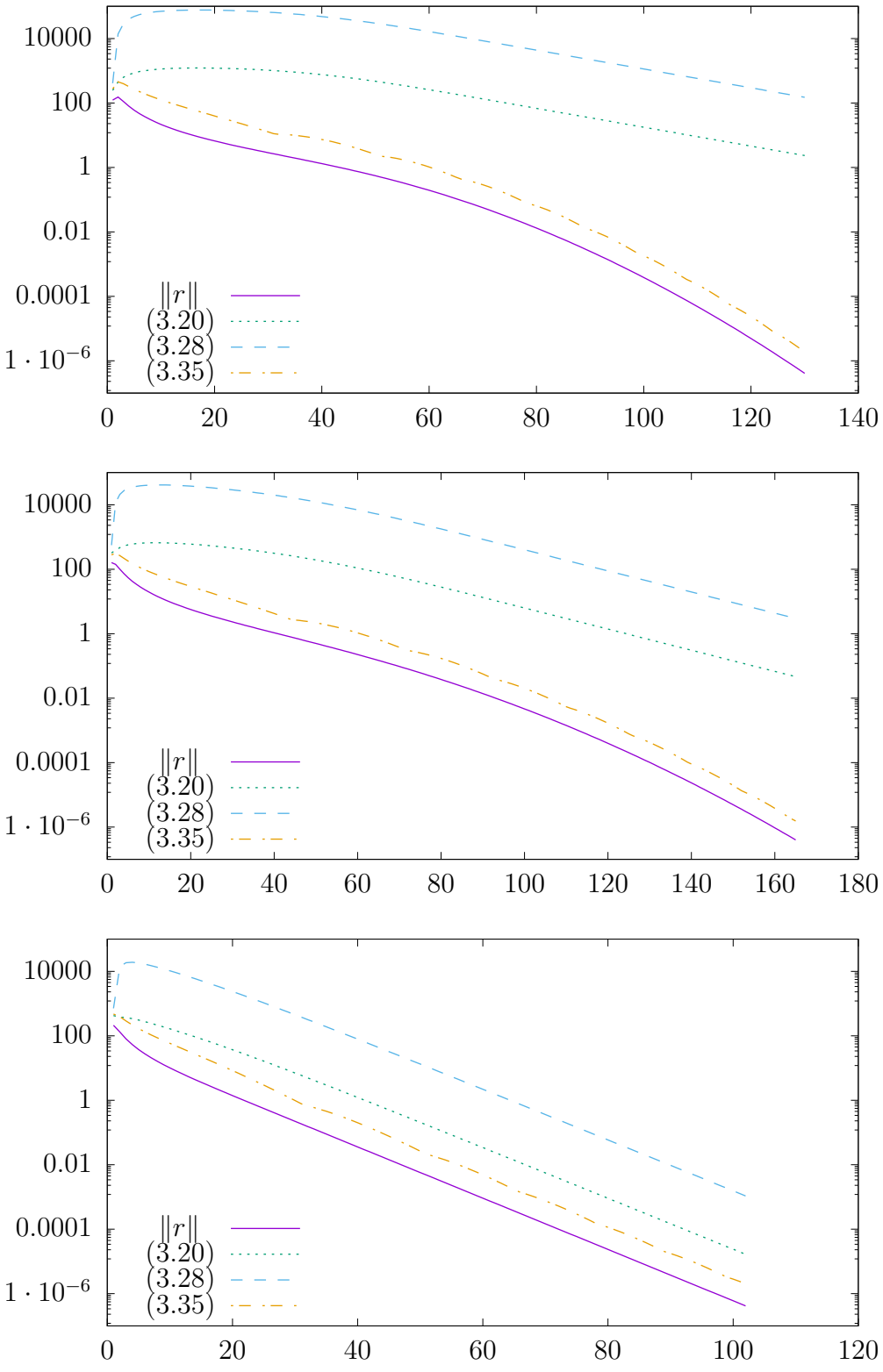

Figure 4.4: $\log_{10}$ of the residual (y-axis) as the iteration proceeds (x-axis) for GLRT as applied to the concave problems `DIAGNQT` (top plot), `DIAGNQE` (middle) and `DIAGNQB` (bottom) with a regularization weight $\sigma = 1000$ and $p = 3$. Each figure shows the residual (3.15) (solid line), and the estimates (3.20) (dotted line), (3.28) (dashed line) and (3.35) (dash-dot line).

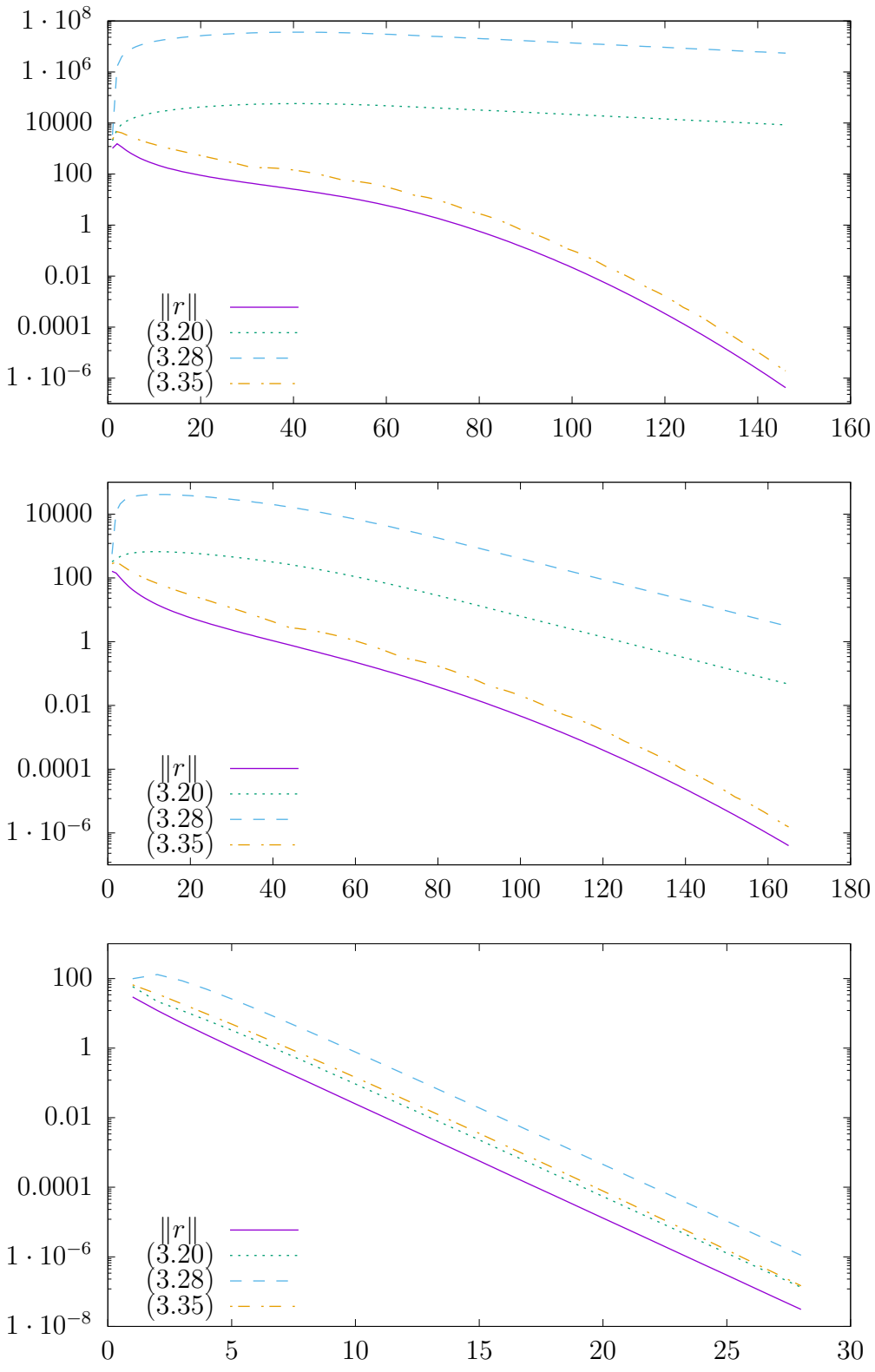

Figure 4.5: $\log_{10}$ of the residual (y-axis) as the iteration proceeds (x-axis) for GLRT as applied to the concave problem `DIAGNQT` with different values of σ when $p = 3$. Specifically $\sigma = 100$ (top plot), $\sigma = 1000$ (middle) and $\sigma = 10000$ (bottom). Each figure shows the residual (3.15) (solid line), and the estimates (3.20) (dotted line), (3.28) (dashed line) and (3.35) (dash-dot line).

Our bounds do not reflect the "superlinear" behaviour that is sometimes observed in practice that results from annihilation of extreme eigenvalues by the Krylov process. A more sophisticated analysis, akin to that by Axelsson, Kaporin and others [2, 3], might provide this, but we have not attempted it.

Acknowledgement

The authors are very grateful for fruitful discussions on aspects of this work with Coralia Cartis, Tyrone Rees and Philippe Toint.

References

[1] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, Cambridge, England, 1996.

[2] O. Axelsson and I. Kaporin. On the sublinear and superlinear rate of convergence of conjugate gradient methods. *Numerical Algorithms*, 25(1):1–22, 2000.

[3] O. Axelsson and J. Karátson. Reaching the superlinear convergence phase of the cg method. *Journal of Computational and Applied Mathematics*, 260:244–257, 2014.

[4] Y. Carmon and J. C. Duchi. Analysis of Krylov subspace solutions of regularized nonconvex quadratic problems. arXiv:1806.09222v1, 2018.

[5] C. Cartis, N. I. M. Gould, and M. Lange. On monotonic estimates of the norm of the minimizers of regularized quadratic functions in krylov spaces. Technical Report RAL-TR-2019, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, 2019.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, 127(2):245–295, 2011.

[7] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, 2000.

[8] S. Demko, W. F. Moss, and P. W. Smith. Decay rates for inverses of band matrices. *Mathematics of Computation*, 43(168):491–499, 1984.

[9] D. M. Gay. Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2(2):186–197, 1981.

[10] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, third edition, 1996.

[11] N. I. M. Gould, , D. Orban, and Ph. L. Toint. CUTEst : a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3):545–557, 2015.

[12] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999.

[13] N. I. M. Gould, D. Orban, and Ph. L. Toint. GALAHAD—a library of thread-safe fortran 90 packages for large-scale nonlinear optimization. *ACM Transactions on Mathematical Software*, 29(4):353–372, 2003.

[14] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia, USA, 1997.

[15] J. Liesen and Z. Strakoš. *Krylov subspace methods*. Oxford University Press, Oxford, 2013.

[16] L. Lukšan, C. Matonoha, and J. Vlček. On Lagrange multipliers of trust-region subproblems. *BIT*, 48(4):763–768, 2008.

[17] J. M. Martínez. Local minimizers of quadratic functions on Euclidean balls and spheres. *SIAM Journal on Optimization*, 4(1):159–176, 1994.

[18] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.

[19] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[20] J. Nocedal and S. J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, Heidelberg, Berlin, New York, second edition, 2006.

[21] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. arXiv:1706.03131v2, 2017.

[22] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.

[23] H. A. van der Vorst. *Iterative Krylov methods for large linear systems*. Cambridge University Press, Cambridge, 2003.

Appendix A

Using (3.32) to compare the model decrease from x_k to x_m, we deduce that

$$\begin{aligned} q(x_k) - q(x_m) &= \tfrac{1}{2}x_k^T H x_k + g^T x_k - \tfrac{1}{2}x_m^T H x_m - g^T x_m \\ &= \tfrac{1}{2}x_k^T H x_k - \tfrac{1}{2}x_m^T H x_m + g^T(x_k - x_m) \\ &= \tfrac{1}{2}x_k^T H x_k - \tfrac{1}{2}x_m^T H x_m + (x_m - x_k)^T H x_m + \mu_m (x_m - x_k)^T x_m \\ &= \tfrac{1}{2}(x_m - x_k)^T (H + \mu_m I)(x_m - x_k) + \tfrac{1}{2}\mu_m(\|x_m\|^2 - \|x_k\|^2). \end{aligned} \tag{A.1}$$

Since (2.10) shows that $H + \mu_m I$ is positive definite on $\mathcal{K}_m$, and as $r_k \in \mathcal{K}_m$, it follows that

$$(H + \mu_m I)^{-1} r_k = (x_k - x_m) + (\mu_k - \mu_m)(H + \mu_m I)^{-1} x_k,$$

and hence, taking the inner product with r_k from (3.33),

$$\begin{aligned} r_k^T (H + \mu_m I)^{-1} r_k &= (x_m - x_k)^T (H + \mu_m I)(x_m - x_k) \\ &\quad + 2(\mu_m - \mu_k)(x_m - x_k)^T x_k + (\mu_m - \mu_k)^2 x_k^T (H + \mu_m I)^{-1} x_k \\ &= 2[q(x_k) - q(x_m)] - \mu_m[\|x_m\|^2 - \|x_k\|^2] \\ &\quad + 2(\mu_m - \mu_k)(x_m - x_k)^T x_k + (\mu_m - \mu_k)^2 x_k^T (H + \mu_m I)^{-1} x_k \\ &= 2[q(x_k) - q(x_m)] + \mu_k(\|x_k\|^2 - \|x_m\|^2) \\ &\quad - (\mu_m - \mu_k)\|x_m - x_k\|^2 + (\mu_m - \mu_k)^2 x_k^T (H + \mu_m I)^{-1} x_k \end{aligned} \tag{A.2}$$

using (A.1). As $r_k \in \mathcal{K}_m$, referring back to (2.2), we have that

$$r_k = U_+ \hat{r}_k = U \begin{pmatrix} \hat{r}_k \\ 0 \end{pmatrix},$$

for some $\hat{r}_k$, and thus

$$r_k^T (H + \mu_m I)^{-1} r_k = \hat{r}_k^T (\Lambda_+ + \mu_m I)^{-1} \hat{r}_k \geq \frac{\|\hat{r}_k\|^2}{(\lambda_+^{\max} + \mu_m)} = \frac{\|r_k\|^2}{(\lambda_+^{\max} + \mu_m)}$$

where we recall the definition of $\lambda_+^{\max}$ from (3.3). Hence (A.2) leads directly to (3.34).

We recall that [5, 22]

$$\|x_k\| \leq \|x_m\| \tag{A.3}$$

and [5, 16]

$$0 \leq \mu_k \leq \mu_m, \tag{A.4}$$

and hence $\mu_k(\|x_k\|^2 - \|x_m\|^2) \leq 0$.

Suppose that $\mu_m = 0$. Then (A.4) implies that $\mu_k = 0$, while (3.32) gives $Hx_m = -g$ and thus (1.1), (2.2), (2.3) and (3.3) combine to give

$$q(0) - q(x_m) = \tfrac{1}{2}\bar{g}_+^T \Lambda_+^{-1} \bar{g}_+ \leq \frac{\|\bar{g}_+\|^2}{(\lambda_+^{\min} + \mu_m)} = \frac{\|g\|^2}{(\lambda_+^{\min} + \mu_m)}. \tag{A.5}$$

Combining (3.4) (3.34) and (A.5) gives

$$\|r_k\|^2 \leq 2(\lambda_+^{\max} + \mu_m)[q(x_k) - q(x_m)] \leq 72\boldsymbol{\kappa}_m \left(e^{-\frac{4}{\sqrt{\boldsymbol{\kappa}_m}}}\right)^k \|g\|^2,$$

i.e.,

$$\|r_k\| \leq 6\sqrt{2}\sqrt{\boldsymbol{\kappa}_m}\left(e^{-\frac{2}{\sqrt{\boldsymbol{\kappa}_m}}}\right)^k \|g\|. \tag{A.6}$$

We note that obtaining an error bound via (3.32)–(A.2) is similar to the approach taken by [21, §3.2] in the absence of a trust region.

Unfortunately, it is unclear how to proceed when $\mu_m > 0$—there are two sub-cases $\mu_k = 0$ and $\mu_k > 0$, but we cannot see a way for either. The issue, of course, is the extra term

$$-(\mu_m - \mu_k)\|x_m - x_k\|^2 + (\mu_m - \mu_k)^2 x_k^T (H + \mu_m I)^{-1} x_k \tag{A.7}$$

in (3.34). Ideally, we would like to show that this is negative in which case a bound of the form (A.6) would follow. We also have a bound

$$\begin{aligned} x_k^T (H + \mu_m I)^{-1} x_k &= (x_k - x_m)^T (H + \mu_m I)^{-1} (x_k - x_m) \\ &\quad + 2x_k^T (H + \mu_m I)^{-1} x_m - x_m^T (H + \mu_m I)^{-1} x_m \\ &\leq (x_k - x_m)^T (H + \mu_m I)^{-1} (x_k - x_m) - 2g^T x_k \end{aligned}$$

on the second term in (A.7), but that doesn't seem to help. Another possibility is to show that the two terms in (A.7) decay exponentially, although we see no reason why in particular $(\mu_m - \mu_k)$ would—one might for example have $\mu_k = 0$ for all $k = 1, \ldots, m-1$ (i.e., the trust-region constraint is inactive), but $\mu_m > 0$ (the constraint becomes active).

Appendix B

Our second attempt to find a useful bound on the residual is based on the relationships (3.8)–(3.9), and uses the following identity.

Lemma 5.1. For any scalar λ, we have

$$V_k^T (H + \lambda I)^j V_k e_1 = (T_k + \lambda I)^j e_1 \tag{B.1}$$

for $j = 1, \ldots, k$.

Proof. We first show that

$$(H + \lambda I)^j V_k = V_k (T_k + \lambda I)^j + \gamma_k \sum_{i=0}^{j-1} (H + \lambda I)^{j-i-1} v_{k+1} e_k^T (T_k + \lambda I)^i \tag{B.2}$$

for all $k \geq 1$. This follows follow immediately when $j = 1$ as

$$(H + \lambda I) V_k = V_k (T_k + \lambda I) + \gamma_k v_{k+1} e_k^T \tag{B.3}$$

from (3.8). Suppose that (B.2) holds for some $j > 1$. Then multiplying by $H + \lambda I$ and using (B.3), we have

$$\begin{aligned}(H+\lambda I)^{j+1}V_k &= (H+\lambda I)V_k(T_k+\lambda I)^j + \gamma_k \sum_{i=0}^{j-1}(H+\lambda I)^{j-i}v_{k+1}e_k^T(T_k+\lambda I)^i \\ &= (V_k(T_k+\lambda I) + \gamma_k v_{k+1}e_k^T)(T_k+\lambda I)^j + \gamma_k \sum_{i=0}^{j-1}(H+\lambda I)^{j-i}v_{k+1}e_k^T(T_k+\lambda I)^i \\ &= V_k(T_k+\lambda I)^{j+1} + \gamma_k v_{k+1}e_k^T(T_k+\lambda I)^j + \gamma_k \sum_{i=0}^{j-1}(H+\lambda I)^{j-i}v_{k+1}e_k^T(T_k+\lambda I)^i \\ &= V_k(T_k+\lambda I)^{j+1} + \gamma_k \sum_{i=0}^{j}(H+\lambda I)^{j-i}v_{k+1}e_k^T(T_k+\lambda I)^i,\end{aligned}$$

and thus (B.2) holds for $j+1$. Hence (B.2) holds for all $j \geq 1$ by induction.

Now observe that $(T_k + \lambda I)e_1$ only has nonzeros in positions 1 and 2, $(T_k + \lambda I)^2 e_1 = (T_k + \lambda I)((T_k + \lambda I)e_1)$ only has nonzeros in positions 1 to 3, and in general $(T_k + \lambda I)^{j-1}e_1 = (T_k + \lambda I)((T_k + \lambda I)^{j-2}e_1)$ only has nonzeros in positions 1 to j. Thus from (B.2) and the orthogonality of the columns of V_k, we have

$$\begin{aligned}V_k^T(H+\lambda I)^j V_k e_1 &= V_k^T V_k(T_k+\lambda I)^j e_1 + \gamma_k \sum_{i=0}^{j-1} V_k^T(H+\lambda I)^{j-i-1}v_{k+1}e_k^T(T_k+\lambda I)^i e_1 \\ &= (T_k+\lambda I)^j e_1 + \gamma_k \sum_{i=0}^{j-2} V_k^T(H+\lambda I)^{j-i-1}v_{k+1}e_k^T(T_k+\lambda I)^i e_1 + \gamma_k V_k^T v_{k+1}e_k^T(T_k+\lambda I)^{j-1}e_1 \\ &= (T_k+\lambda I)^j e_1\end{aligned}$$

as required, since $e_k^T(T_k + \lambda I)^i e_1 = 0$ for $i = 0, \ldots, j-2$ and $j = 1, \ldots, k$, and $V_k^T v_{k+1} = 0$. □

Since $x_k \in \mathcal{K}_k = \text{span}\{H^i g\}_{i=0}^{k-1} \equiv \text{span}\{(H + \mu_k I)^i g\}_{i=0}^{k-1}$, we have

$$x_k = \sum_{j=0}^{k-1} \eta_j (H + \mu_k I)^j g$$

for coefficients η_j, $j = 0, \ldots, k-1$, and thus

$$\begin{aligned}r_k &= g + (H+\mu_k I)\sum_{j=0}^{k-1}\eta_j(H+\mu_k I)^j g = g + \sum_{j=0}^{k-1}\eta_j(H+\mu_k I)^{j+1}g \\ &= g + \sum_{j=1}^{k}\eta_{j-1}(H+\mu_k I)^j g \\ &= \psi_k(H+\mu_k I)g\end{aligned} \tag{B.4}$$

for some

$$\psi_k(\lambda) = \sum_{j=0}^{k}\omega_j\lambda^j \in \mathcal{P}_k = \{\text{polynomials } \psi \text{ of degree } k \text{ for which } \psi(0) = 1\}.$$

But then

$$\begin{aligned} V_k^T r_k &= V_k^T \psi_k(H + \mu_k I) g = \sum_{j=0}^{k} \omega_j V_k^T (H + \mu_k I)^j g \\ &= \|g\| \sum_{j=0}^{k} \omega_j V_k^T (H + \mu_k I)^j V_k e_1 = \|g\| \sum_{j=0}^{k} \omega_j (T_k + \mu_k I)^j e_1 \\ &= \|g\| \psi_k \left(T_k + \mu_k I\right) e_1 \end{aligned} \tag{B.5}$$

using (3.12) and (B.1). Since T_k is irreducible, it has distinct eigenvalues $\theta_{i,k}$, $i = 1, \ldots, k$, and as (3.13) indicates that $V_k^T r_k = 0$, (B.5) implies that ψ_k is a scalar multiple of the minimum polynomial

$$\phi_k(\lambda) = \prod_{i=1}^{k} (\lambda - \theta_{i,k} - \mu_k)$$

of the irreducible matrix $T_k + \mu_k I$. Indeed, $\psi_k(\lambda) = \phi_k(\lambda)/\phi_k(0)$ since we require that $\psi_k(0) = 1$.

Referring back to (2.2) and (2.3), we have that $H = U \Lambda U^T$ and $g = U\bar{g}$ for matrices U of eigenvectors and Λ of eigenvalues. Then (B.4) gives

$$r_k = \psi_k \left(U(\Lambda + \mu_k I) U^T \right) U\bar{g} = U \psi_k(\Lambda + \mu_k I)\bar{g} = U_+ \psi_k(\Lambda_+ + \mu_k I)\bar{g}_+,$$

and hence

$$\begin{aligned} \|r_k\|^2 &= \|\psi_k(\Lambda_+ + \mu_k I)\bar{g}_+\|^2 = \sum_{j \in \mathcal{I}_+} \psi_k^2(\lambda_j + \mu_k)\bar{g}_j^2 \\ &\leq \max_{j \in \mathcal{I}_+} \psi_k^2(\lambda_j + \mu_k) \sum_{j \in \mathcal{I}_+} \bar{g}_j^2 = \max_{j \in \mathcal{I}_+} \psi_k^2(\lambda_j + \mu_k) \|\bar{g}_*\|^2 \\ &= \max_{j \in \mathcal{I}_+} \psi_k^2(\lambda_j + \mu_k) \|g\|^2 \end{aligned} .$$

This then provides the estimate (3.35).