BIT manuscript No.
(will be inserted by the editor)

On monotonic estimates of the norm of the minimizers of regularized quadratic functions in Krylov spaces

Coralia Cartis · Nicholas I. M. Gould · Marius Lange

5th April 2019, revised 30th August 2019 and 10th November 2019

Abstract We show that the minimizers of regularized quadratic functions restricted to their natural Krylov spaces increase in Euclidean norm as the spaces expand.

Keywords regularized quadratic functions · solution estimates · Krylov subspaces

Mathematics Subject Classification (2000) MSC 90C20 · MSC 90C26

1 Introduction

Given a real symmetric, possibly indefinite, matrix H and vector g, we are concerned with Krylov methods for approximating the global solution of the possibly nonconvex regularization problem

$$\underset{x\in\Re^n}{\text{minimize}}\ Q(x;\sigma,p) := \tfrac{1}{2}x^THx + g^Tx + \tfrac{1}{p}\sigma\|x\|^p \tag{1.1}$$

where $\sigma > 0$, $p > 2$ and $\|\cdot\|$ is the Euclidean norm—note that Q is bounded below over $\Re^n$, and all global minimizers have the same norm [7, §3]. Such methods have been advocated by a number of authors, e.g., [1–3, 8]. Here we are interested in how the norms of the estimates of the solution evolve as the Krylov process proceeds. The main utility is that these estimates provide useful predictions for the "multipliers" $\sigma\|x\|^{p-2}$ as the Krylov subspace expands

This work was supported by the EPSRC grant EP/M025179/1 (NIMG), and both by a German research foundation (DFG) fellowship through the Graduate School of Quantitative Biosciences Munich (QBM) and by the Joachim Herz Stiftung (ML).

Coralia Cartis
Mathematical Institute, Oxford University, Oxford OX2 6GG, England.
E-mail: coralia.cartis@maths.ox.ac.uk

Nicholas I. M. Gould
Scientific Computing Department, STFC-Rutherford Appleton Laboratory, Chilton OX11 0QX, England.
E-mail: nick.gould@stfc.ac.uk

Marius Lange
Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, 85764 Neuherberg, Germany and Department of Mathematics, Technische Universität München, 85748 Munich, Germany.
E-mail: marius.lange@helmholtz-muenchen.de

[9]. Our result is an analogue of that obtained by Lukšan, Matonoha and Vlček [11] for the trust-region subproblem.

By way of motivation and explanation, the solution x_* to (1.1) necessarily satisfies the first-order criticality condition $\nabla_x Q(x_*;\sigma,p) = 0$, i.e.,

$$(H+\mu_* I)x_* = -g, \text{ where } \mu_* = \sigma\|x_*\|^{p-2}. \tag{1.2}$$

In addition, $H+\mu_* I$ is positive semi-definite at any global minimizer, and if $H+\mu_* I$ is positive definite, the minimizer is unique [3, Thm.3.1]. If μ_* was known, the minimizer might be found simply by solving the linear system (1.2), and the skill is then in finding convergent estimates of μ_* by iteration [5]. Briefly, this is achieved by seeking the rightmost root of the secular equation

$$\|x(\mu)\|^{p-2} - \frac{\mu}{\sigma} = 0,$$

where

$$(H+\mu I)x(\mu) = -g \tag{1.3}$$

while ensuring that $H+\mu I$ is positive semi-definite. This is always possible so long as g intersects the subspace $\mathscr{U}$ of eigenvectors of H corresponding to the leftmost eigenvalue $\lambda_{\min}(H)$ of H, and in this case $H+\mu_* I$ is positive definite. The rare possibility that the later does not happen is known colloquially as the "hard case" [12], and the solution to (1.1) in the hard case involves an addition component from $\mathscr{U}$. We also make the connection between (1.1) and the regularized quadratic

$$Q_\nu(x) := \tfrac{1}{2}x^T(H+\nu I)x + g^T x, \tag{1.4}$$

namely that if $\nu = \mu_*$ and the hard case does not occur then x_* minimizes $Q_\nu(x)$.

This all presupposes that one can solve the linear system (1.3), and unfortunately in some applications matrix factorization is out of the question, indeed H may only be available indirectly via Hessian-vector products Hv for given v. An attractive alternative in such cases is to find an approximation to the solution of (1.1) by restricting the search domain to a subspace or sequence of subspaces. A particularly appealing set of nested subspaces is provided by the Krylov space defined by H and g that we will define formally in the next section. Crucially, the k-th Krylov subspace, $\mathscr{K}_k(H,g)$, may be generated recursively through Hessian-vector products, and has an orthogonal basis V_k, with the desirable property that if

$$x_k = \operatorname*{arg\,min}_{x\in\mathscr{K}_k(H,g)} Q(x;\sigma,p)$$

then $x_k = V_k\bar{x}_k$, where

$$\bar{x}_k = \operatorname*{arg\,min}_{\bar{x}\in\Re^k} \tfrac{1}{2}\bar{x}^T T_k\bar{x} + \|g\|e_1^T\bar{x} + \tfrac{1}{p}\sigma\|\bar{x}\|^p, \tag{1.5}$$

the vector $e_1 \in \Re^k$ is the first column of the identity matrix and $T_k = V_k^T H V_k \in \Re^{k\times k}$ is tridiagonal. The latter implies that solving (1.5) is feasible via its optimality equations

$$(T_k+\mu_k I)\bar{x}_k = -\|g\|e_1, \text{ where } \mu_k = \sigma\|x_k\|^{p-2},$$

even when the dimension is large, since factorizing shifted tridiagonal matrices and solving linear systems involving them may be achieved in a few multiples of k floating-point operations.

Of course, we need to judge when x_k is a meaningful approximation to x_* as the subspaces evolve, and furthermore to solve each successive subproblem (1.5) efficiently. The former is addressed in [2,6] and requires estimates of μ_k, while the latter appeals to the ideas in [5] and relies on a good starting "guess" for μ_k. Thus generating a good starting guess provides motivation for our short paper.

In the next section we provide a set of lemmas leading to our main result, namely that so long as the evolving Krylov subspaces are of full dimensionality, the norms of the solution estimates $\|s_k\|$ and the corresponding "multipliers" μ_k increase monotonically. We summarize, extend and discuss implications and limitations of our results in the concluding Section 3.

2 The main result

We start with four vital lemmas that we use to prove our main result. The first shows a simple property of the conjugate gradient method. We use the generic notation $H \succ 0$ (resp. $H \succeq 0$) to mean that the real, symmetric matrix H is positive definite (resp. positive semi-definite).

Lemma 1 Given a real symmetric matrix H and real vector g, let

$$\mathscr{K}_k(H,g) := \text{span}\left\{g, Hg, \ldots, H^{k-1}g\right\},$$

$k \geq 1$, be the k-th Krylov subspace generated by H and the vector g, and let the columns of V_k provide an orthonormal basis for $\mathscr{K}_k(H,g)$. Letting $\ell \geq k \geq 1$, suppose that

$$V_\ell^T H V_\ell \succ 0 \tag{2.1}$$

and define

$$x_k = \underset{x \in \mathscr{K}_k(H,g)}{\arg\min} \; Q(x) := \tfrac{1}{2} x^T H x + g^T x.$$

Then

$$\|x_k\| \leq \|x_\ell\|.$$

Proof. This follows from [4, Thm.7.5.1] as the requirement there, namely that $p_k^T H p_k > 0$ for specific vectors $p_k \in \mathscr{K}_k(H,g)$, is implied by the more general assumption (2.1). □

Note that this is a generalization of [13, Thm.2.1] that relaxes the requirement that H be everywhere positive definite to be so merely over the evolving Krylov subspaces of interest.

Our second lemma compares Krylov subspaces of the matrices H and $H + \mu I$ for some $\mu \in \Re$.

Lemma 2 [11, Lem.2.3]. Let H, g and $\mathscr{K}_k$ be as in Lemma 1, and $\mu \in \Re$. Then

$$\mathscr{K}_k(H + \mu I, g) = \mathscr{K}_k(H,g). \tag{2.2}$$

Next, we state a crucial relation between the parameter ν that defines $Q_\nu(x)$ in (1.4) and the norm of the minimizer of $Q_\nu(x)$ within the k-th Krylov space.

Lemma 3 [11, Lem.2.5]. Suppose that the columns of V_k provide an orthonormal basis for $\mathscr{K}_k(H,g)$ for given real symmetric H and real g. Let $V_k^T HV_k + \nu_i I$, $\nu_i \in \mathfrak{R}$, $i \in \{1,2\}$, be symmetric and positive definite. Let

$$x_k(\nu_i) = \arg\min_{x \in \mathscr{K}_k(H,g)} Q_{\nu_i}(x) := \tfrac{1}{2} x^T (H + \nu_i I)x + g^T x.$$

Then

$$\nu_2 \leq \nu_1 \text{ if and only if } \|x_k(\nu_2)\| \geq \|x_k(\nu_1)\|.$$

We define the grade of H and g, $\text{grade}(H,g) \leq n$, to be the maximum dimension of the evolving Krylov spaces $\mathscr{K}_k(H,g)$, $k = 1,\ldots,n$ [10]. Our final lemma indicates that the evolving minimizers are unique.

Lemma 4 Let H, g and V_k be as in Lemma 3 and let μ_k by the rightmost root of the secular equation

$$\|\bar{x}_k(\mu)\|^{p-2} - \frac{\mu}{\sigma} = 0, \text{ where } (V_k^T HV_k + \mu I)\bar{x}_k(\mu) = -V_k^T g \tag{2.3}$$

Then $V_k^T HV_k + \mu_k I \succ 0$ for all $1 \leq k \leq m := \text{grade}(H,g)$.

Proof. Using the Lanczos orthonormal basis, we have that $V_k^T HV_k = T_k$ for an irreducible tridiagonal matrix T_k for $k = 1,\ldots,m$. It then follows [4, Thm.7.5.12] that $\mathscr{K}_k(H,g)$ has a nontrivial intersection with the space of eigenvectors of T_k corresponding to the eigenvalue $\lambda_{\min}(T_k)$ (i.e., the "hard case" cannot occur), and thus that the only permitted root μ_k of the secular equation (2.3) for the problem satisfies $\mu_k > -\lambda_{\min}(T_k)$, where $\lambda_{\min}$ denotes the leftmost eigenvalue of its symmetric matrix argument [5, Sec.2.2]. □

We are now in a position to state and prove our main theorem.

Theorem 1 Given a real symmetric matrix H, vector g and scalars $\sigma > 0$ and $p > 2$, let $m = \text{grade}(H, g)$,

$$x_j = \underset{x \in \mathscr{K}_j(H,g)}{\arg\min}\, Q(x;\sigma,p) := \tfrac{1}{2}x^T Hx + g^T x + \tfrac{1}{p}\sigma\|x\|^p,$$

and

$$\mu_j = \sigma\|x_j\|^{p-2} \tag{2.4}$$

for $j \geq 1$. Then $\mu_k \leq \mu_\ell$ and $\|x_k\| \leq \|x_\ell\|$ for $1 \leq k \leq \ell \leq m$.

Proof. Let V_j be as in the statement of Lemma 3. The vector $x_j = V_j y_j$ is a minimizer of the j-th regularization subproblem if and only if

$$V_j^T(H + \mu_j I)V_j y_j = -V_j^T g,\ \ V_j^T(H + \mu_j I)V_j \succeq 0,\ \text{ and }\ \mu_j = \sigma\|y_j\|^{p-2}, \tag{2.5}$$

and the minimizer is unique since $V_j^T(H + \mu_j I)V_j \succ 0$ from Lemma 4 [5, Thm.2]. Consider two integers k and ℓ for which $1 \leq k \leq \ell \leq m$.
Since we have $V_k^T(H + \mu_k I)V_k \succ 0$ and $V_\ell^T(H + \mu_\ell I)V_\ell \succ 0$, and as $\mathscr{K}_k(H + \mu_k I, g) = \mathscr{K}_k(H, g)$ by Lemma 2, it follows from (2.5) that x_k is also the (unique) solution of the subspace minimization problem

$$x_k = \underset{x \in \mathscr{K}_k(H,g)}{\arg\min}\ \ Q_{\mu_k}(x),\ \text{ where }\ Q_\mu(x) = \tfrac{1}{2}x^T(H + \mu I)x + g^T x.$$

Assume that $\mu_k > \mu_\ell$, which implies that $V_\ell^T(H + \mu_k I)V_\ell \succ 0$. Let

$$x_\ell(\mu_k) = \underset{x \in \mathscr{K}_\ell(H,g)}{\arg\min}\ \ Q_{\mu_k}(x).$$

Then it follows from Lemma 1 that

$$\|x_k\| \leq \|x_\ell(\mu_k)\|. \tag{2.6}$$

But since $\mu_\ell < \mu_k$, Lemma 3 gives that

$$\|x_\ell(\mu_k)\| \leq \|x_\ell(\mu_\ell)\| = \|x_\ell\|. \tag{2.7}$$

Hence using the definition (2.4) and combining the inequalities (2.6) and (2.7)

$$\mu_k = \sigma\|x_k\|^{p-2} \leq \sigma\|x_\ell\|^{p-2} = \mu_\ell < \mu_k$$

which is a contradiction. Thus $\mu_k \leq \mu_\ell$ has to hold. It then follows from the definition (2.4) that $\|x_k\| \leq \|x_\ell\|$. □

The monotonic behaviour of the multipliers μ_k was predicted in [9, Lem.2.6] when $p = 3$, but the proof suggested there relied on [11, Thm.2.6], which appears to have a minor flaw—the proof depends on [13, Thm.2.1], but applies this at one point to an indefinite $H + \mu I$. Lemma 1 avoids this issue, and the same result fixes the proof of [11, Thm.2.6] that applies in the trust-region case.

3 Comments and conclusions

We have shown that the norms of the approximations generated by well-known Krylov methods for solving the regularization problem (1.1) increase monotonically as the dimension of the Krylov spaces expands. This implies that the corresponding "multipliers" μ_k also increase, and is useful as estimates of these multipliers are crucial when solving the Krylov subproblem; in particular, as the multiplier for the k-th problem is a lower bound for the $(k+1)$-st, Newton-like iterations for the required root of the secular equation

$$\|\bar{x}_{k+1}(\mu)\|^{p-2} - \frac{\mu}{\sigma} = 0, \text{ where } (V_{k+1}^T H V_{k+1} + \mu I)\bar{x}_{k+1}(\mu) = -V_{k+1}^T g,$$

will converge both globally and rapidly to μ_{k+1} when started from μ_k if additionally $\mu_k > \lambda_{\min}(T_{k+1})$ [5, §3]. In particular, Newton's method, the secant method or methods based upon certain higher (odd)-order Taylor approximations or nonlinear rescalings of the term $\|\bar{x}(\mu)\|^{p-2}$ all converge monotonically from such a starting μ. Knowledge of the monotonic nature of these quantities is also important when deriving convergence bounds [6] for such methods.

We warn readers that in exceptional circumstances, namely when g is orthogonal to the eigenspace corresponding to the leftmost eigenvalue of H and σ is not large enough, the global minimizer of (1.1) will not lie in $\mathscr{K}_m(H,g)$, and μ_m will underestimate the optimal multiplier. This (zero-probability) possibility is often referred to as the "hard case" [3, §6,1, 12], and, despite their popularity, might be viewed as an unavoidable defect of Krylov methods.

The main result here may trivially be extended for Krylov methods to

$$\underset{x\in\Re^n}{\text{minimize}}\ Q(x;\sigma,p,M) := \tfrac{1}{2}x^T H x + g^T x + \tfrac{1}{p}\sigma\|x\|_M^p,$$

for given symmetric $M \succ 0$, where $\|x\|_M^2 := x^T M x$, so long as we instead consider the Krylov spaces $\mathscr{K}(M^{-1}H, M^{-1}g)$. It is well known that this may be achieved using the M-preconditioned Lanczos method [3, Sec.6.3]. In particular, if

$$x_j = \underset{x\in\mathscr{K}_j(M^{-1}H,M^{-1}g)}{\arg\min} Q(x;\sigma,p,M) \text{ and } \mu_j = \sigma\|x_j\|_M^{p-2},$$

it follows (using the transformation $x \leftarrow M^{\frac{1}{2}}x$) that

$$\mu_k \le \mu_\ell \text{ and } \|x_k\|_M \le \|x_\ell\|_M$$

for $1 \le k \le \ell \le m$ just as in Theorem 1.

Acknowledgement

The authors would like to thank two referees and the associate editor for their very helpful comments on the original draft of this paper.

References

1. T. Bianconcini, G. Liuzzi, B. Morini, and M. Sciandrone. On the use of iterative methods in cubic regularization for unconstrained optimization. *Computational Optimization and Applications*, 60(1):35–57, 2015.
2. Y. Carmon and J. C. Duchi. Analysis of Krylov subspace solutions of regularized nonconvex quadratic problems. arXiv:1806.09222v1, 2018.
3. 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.
4. A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, 2000.
5. N. I. M. Gould, D. P. Robinson, and H. S. Thorne. On solving trust-region and other regularised subproblems in optimization. *Mathematical Programming Computation*, 2(1):21–57, 2010.
6. N. I. M. Gould and V. Simoncini. Error estimates for iterative algorithms for minimizing regularized quadratic subproblems. *Optimization Methods and Software*, DOI: 10.1080/10556788.2019.1670177, 2019.
7. A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England, 1981.
8. J. M. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1895–1904, 2017.
9. M. Lange. Subproblem solutions in cubic regularisation methods. M.Sc. thesis, University of Oxford, England, 2017.
10. J. Liesen and Z. Strakoš. *Krylov subspace methods*. Oxford University Press, Oxford, 2013.
11. L. Lukšan, C. Matonoha, and J. Vlček. On Lagrange multipliers of trust-region subproblems. *BIT*, 48(4):763–768, 2008.
12. J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.
13. T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.