Preconditioning for Generalized Jacobians with the ω-Condition Number [*†]

Woosuk L. Jung[‡] David Torregrosa-Belén[§] Henry Wolkowicz[‡]

August 24, 2023

Key words and phrases: ω-condition number, preconditioning, generalized Jacobian, iterative methods

AMS subject classifications: 15A12, 65F35, 65F08, 65F10, 65G50, 49J52

Contents

[*]Emails resp.: w2jung@uwaterloo.ca, david.torregrosa@ua.es, hwolkowicz@uwaterloo.ca

[†]This report is available at URL: www.math.uwaterloo.ca/~hwolkowi/henry/reports/ABSTRACTS.html

[‡]Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

[§]Department of Mathematics, University of Alicante, Carretera San Vicente del Raspeig, Alicante, 03690, Spain

List of Algorithms

List of Tables

List of Figures

Abstract

Preconditioning is essential in iterative methods for solving linear systems of equations. We study a nonclassic matrix condition number, the ω-condition number, in the context of optimal conditioning for low rank updating of positive definite matrices. For a positive definite matrix, this condition measure is the ratio of the arithmetic and geometric means of the eigenvalues. In particular, we concentrate on linear systems with low rank updates of positive definite matrices which are close to singular. These systems arise in the contexts of nonsmooth Newton methods using generalized Jacobians. We derive an explicit formula for the optimal ω-preconditioned update in this framework.

Evaluating or estimating the classical condition number κ can be expensive. We show that the ω-condition number can be evaluated exactly following a Cholesky or LU factorization and it estimates the actual condition of a linear system significantly better. Moreover, our empirical results show a significant decrease in the number of iterations required for a requested accuracy in the residual during an iterative method, i.e., these results confirm the efficacy of using the ω-condition number compared to the classical condition number.

1 Introduction

In this paper we study the ω-condition number, a nonclassic matrix condition number. In particular, we look at finding the optimal ω-conditioned low rank updates of the positive definite generalized Jacobian that arises in nonsmooth Newton methods e.g., [3]. We illustrate both the efficiency and effectiveness of using this condition number compared to the classic κ-condition number when solving positive definite linear systems. In addition, we show that the ω-condition number can be evaluated exactly following a Cholesky or LU factorization. Moreover, we illustrate that the ω-condition number is a better indication of the conditioning of a problem, i.e., how random perturbations in the data affect the solution.

In numerical analysis, a condition number of a matrix A is the main tool in the study of error propagation in the problem of solving the linear equation $Ax = b$. The classical condition number of A, denoted as $\kappa(A)$, is defined as the ratio of the largest and smallest singular values of A. The linear system $Ax = b$ is said to be well-conditioned when A has a low condition number. In particular, $\kappa(A)$ is a measure of how much a solution x will change with respect to changes in the right-hand side b, e.g., [25]. In general, iterative algorithms used to solve the system $Ax = b$ require a large number of iterations to achieve a solution with high accuracy if the problem is not well-conditioned, i.e., is ill-conditioned. In this paper, we restrict ourselves to A positive definite and so $\kappa = \lambda_1(A)/\lambda_n(A)$, the ratio of largest and smallest eigenvalues.

In order to improve the conditioning of a problem, preconditioners are employed for obtaining equivalent systems with better condition number. For example, in [6] a preconditioner that minimizes the classical condition number κ is obtained in the Broyden family of rank-two updates. Also, for applications to inexact Newton methods see [1, 2], where it is emphasized that the goal is to improve the *clustering of eigenvalues* around 1. The ω-condition number in particular uses *all* the eigenvalues, rather than just the largest and

smallest as in the classical κ.

In the current literature, reducing the κ-condition number is the main aim for preconditioners. Nevertheless, a nonstandard condition number was proposed in [8]. Interestingly enough, the authors show that the inverse-sized BFGS and sized DFP [21] are obtained as optimal quasi-Newton updates with respect to this measure. This nonstandard condition number is known as the *ω-condition number* and is defined as the ratio of the arithmetic and geometric means of the eigenvalues of a positive definite matrix A:

$$\omega(A) := \frac{\operatorname{tr}(A)/n}{\det(A)^{\frac{1}{n}}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\lambda_i(A)}{\left(\prod_{i=1}^{n}\lambda_i(A)\right)^{\frac{1}{n}}}. \tag{1.1}$$

Our goal in this paper is to establish the basic properties of the ω-condition number and to study whether it is a better indicator of whether the problem $Ax = b$ is well- or ill-conditioned as it uses all the eigenvalues rather than the extreme pair of eigenvalues.

In addition, the ω-condition number presents some advantages with respect to the classic condition number, since it is differentiable and pseudoconvex in the interior of the positive semidefinite cone. This facilitates obtaining optimal preconditioners, see e.g., [8]. Moreover, it is expensive to evaluate the classic condition number [15], i.e., one needs the largest and smallest eigenvalues or the norms of the matrix and its inverse. We show that we can find the exact value of the ω-condition number when a Cholesky or LU factorization is done. Finally, we show that the ω-condition number provides a significantly better estimate for the true conditioning of a linear system.

1.1 Outline

Preliminaries are presented in Section 1.2. Then Section 2.1 introduces the basic properties of the ω-condition number. In particular, in Section 2.2 we empirically motivate the use of the ω-condition number as a better indicator of the conditioning of the problem compared to the κ-condition number. In Section 3, we derive an optimal ω-preconditioning for low rank updates of positive definite matrices. These updates often arise in the construction of generalized Jacobians. In Section 4, we use the linear equations that involve the generalized Jacobians. We empirically illustrate that reducing the ω-condition number improves the performance of iterative methods for solving linear systems. Conclusions are provided in Section 5.

1.2 Preliminaries and Notation

We let $\mathbb{R}^n$ denote the real Euclidean space of dimension n, and $\mathbb{R}^{n\times m}$ the space of $n \times m$ matrices. The space of $n \times n$ symmetric matrices is denoted as $\mathbb{S}^n$ and I stands for the identity matrix. We use $\mathbb{S}^n_+$ and $\mathbb{S}^n_{++}$ for the cone of positive semidefinite and positive

definite $n \times n$ symmetric matrices, respectively. We use $A \succeq 0$ (resp., $\succ 0$) to denote A is in $\mathbb{S}^n_+$ (respectively, $\mathbb{S}^n_{++}$). The transpose of a matrix $M \in \mathbb{R}^{n\times m}$ is written as $M^T \in \mathbb{R}^{m\times n}$. Given $A \in \mathbb{S}^n$, $\mathrm{tr}(A)$ and $\det(A)$ denote the trace and determinant of A, respectively.

For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we use ∇f for the gradient. If $\mathbb{R}^n = \mathbb{R}$, we just write f' for the derivative of f. Given a nonempty open set $\Omega \subseteq \mathbb{R}^n$, a function $f : \Omega \to \mathbb{R}$ is said to be *pseudoconvex* on Ω if it is differentiable and

$$\nabla f(x)(y-x)^T \geq 0 \Longrightarrow f(y) \geq f(x), \quad \forall x, y \in \Omega.$$

This implies that for a convex set Ω and a *pseudoconvex function* $f : \Omega \to \mathbb{R}$, we have: $\nabla f(x) = 0$ is a necessary and sufficient condition for x to be a global minimizer of f in Ω, see e.g., [19].

2 Properties and Numerical Evaluation of ω

We now introduce the basic properties of the ω-condition number and study its numerical evaluation. In addition we empirically compare its effectiveness with the κ-condition number for estimating the actual conditioning of positive definite linear systems

2.1 Basic Properties

It is known that the simple scaling diagonal preconditioner is the optimal preconditioner with respect to the ω-condition number, see [8,22]. This result is extended for block diagonal preconditioners in [9]. We summarize these and other basic properties of the ω-condition number in the following Proposition 2.1.

Proposition 2.1 ([8,9]). *The following statements are true.*

1 *The measure ω is pseudoconvex on the set of symmetric positive definite matrices, and thus any stationary point is a global minimizer of ω.*

2 *Let A be a full rank $m \times n$ matrix, $n \leq m$. Then the optimal column scaling that minimizes the measure ω, i.e.,*

$$\min \ \omega((AD)^T(AD)),$$

over D positive, diagonal, is given by

$$D_{ii} = \frac{1}{\|A_{:,i}\|}, \quad i = 1, ..., n,$$

where $A_{:,i}$ is the i-th column of A.

3 Let A be a full rank $m \times n$ matrix, $n \leq m$ with block structure $A = \begin{bmatrix} A_1 & A_2 & \dots & A_k \end{bmatrix}$, $A_i \in \mathbb{R}^{m \times n_i}$. Then an *optimal corresponding block diagonal scaling*

$$D = \begin{bmatrix} D_1 & 0 & 0 & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & D_k \end{bmatrix}, \quad D_i \in \mathbb{R}^{n_i \times n_i},$$

that minimizes the measure ω, i.e.,

$$\min \ \omega((AD)^T(AD)),$$

over D block diagonal, is given by the factorization

$$D_i D_i^T = \{A_i^T \, A_i\}^{-1}, \quad i = 1, ..., k.$$

■

2.2 $\omega(A)$ in Error Analysis

Suppose that A, b for the linear (square) system $Ax = b$ are given. Let $\Delta A, \Delta b, \epsilon > 0$ be perturbations resulting in the perturbed linear system

$$(A + \epsilon \Delta A)(x + \Delta x) = b + \epsilon \Delta b.$$

Let

$$\rho_A = \epsilon \|\Delta A\| / \|A\| \ \text{ and } \ \rho_b = \epsilon \|\Delta b\| / \|b\|.$$

It is well known, e.g., [11, Sect. 2.5], that the relative error

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A)(\rho_A + \rho_b) + O(\epsilon^2),$$

i.e., the condition number $\text{cond}(A)$ is estimated by the ratio of the relative error in the solution to the sum of relative errors in the data and $\kappa(A)$ is an upper bound:

$$\kappa(A) \geq \text{cond}(A) :\approx \frac{\frac{\|\Delta x\|}{\|x\|}}{(\rho_A + \rho_b)}. \tag{2.1}$$

We have run many tests[1] for various size n problems with $A \succ 0$ with various condition numbers determined by the sprandsym command in MATLAB. For each test we generated 30 random perturbations to estimate $\text{cond}(A)$. From our tests, we conclude that the ω-condition number resulted in a better estimate of $\text{cond}(A)$. One instance of the tests is ilustrated in Figure 2.1.

[1] We used a laptop with Intel Core i7-12700H 2.30 GHz, with 16GB RAM, running Windows 11 (64-bit).

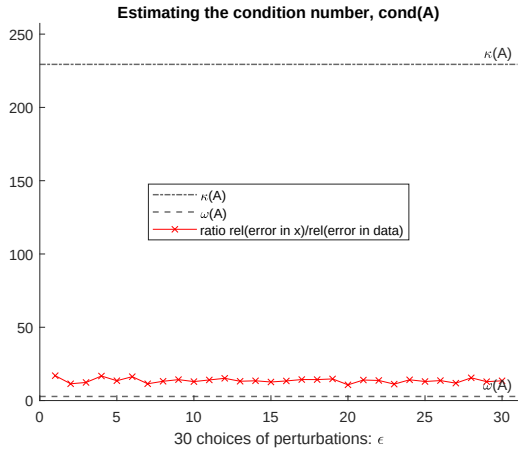

Figure 2.1: Estimating the condition number.

2.3 Efficiency and Accuracy of Evaluation of $\omega(A)$

Since eigenvalue decompositions can be expensive, one issue with $\kappa(A)$ is how to estimate it efficiently when the size of matrix A is large. A survey of estimates and, in particular, estimates using the ℓ_1-norm, is given in [15, 16]. Extensions to sparse matrices and block-oriented generalizations are given in [14, 17]. Results from these papers form the basis of the *condest* command in MATLAB; this illustrates the difficulty in accurately estimating $\kappa(A)$.

On the other hand, the measure $\omega(A)$ can be calculated using the trace and determinant function which do not require eigenvalue decompositions. However, for large n, the determinant is also numerically difficult to compute as it could easily result in an overflow $+\infty$ or 0 due to the limits of finite precision arithmetic, e.g., if the order of A is $n = 50$ and the eigenvalues $\lambda_i = .5, \forall i$, then the determinant $.5^n$ is zero to machine precision. A similar problem arises for e.g., $\lambda_i = 2, \forall i$ with overflow. In order to overcome this problem, we take the n-th root first and then the product, i.e., we define the value obtained from the spectral factorization as

$$\omega_{\text{eig}}(A) = \frac{\sum_{i=1}^{n} \lambda_i(A)/n}{\prod_{i=1}^{n} (\lambda_i(A)^{1/n})}.$$

We now let $A = R^T R = LUP$ denote the Cholesky and LU factorizations, respectively, with appropriate permutation matrix P. We assume that L is unit lower triangular. Therefore,

$$\det(A)^{1/n} = \det\left(R^T R\right)^{1/n} = \det(R)^{2/n} = \prod_{i=1}^{n} \left(R_{ii}^{2/n}\right). \tag{2.2}$$

Similarly,

$$\det(A)^{1/n} = \det(LUP)^{1/n} = \prod_{i=1}^{n} \left(|U_{ii}|^{1/n} \right). \tag{2.3}$$

Therefore, we find $\omega(A)$ with numerator $\mathrm{tr}(A)/n$ and denominator given in (2.2) and (2.3), respectively:

$$\omega_R(A) = \frac{\mathrm{tr}(A)/n}{\prod_{i=1}^{n} \left(R_{ii}^{2/n} \right)}, \quad \omega_{LU}(A) = \frac{\mathrm{tr}(A)/n}{\prod_{i=1}^{n} \left(|U_{ii}|^{1/n} \right)}.$$

Tables 2.1 and 2.2 provide comparisons on the time and precision from the three different factorization methods. Each column presents different order of κ-condition number, while each row corresponds to different decompositions with different size n of the problem. We form the random matrix using $A = QDQ^T$ for random orthogonal Q and positive definite diagonal D. We then symmetrize $A \leftarrow (A + A^T)/2$ to avoid roundoff error in the multiplications. Therefore, we consider the evaluation using D as the *exact value* of $\omega(A)$, i.e.,

$$\omega(A) = \frac{\sum_{i=1}^{n} (D_{ii})/n}{\prod_{i=1}^{n} \left(D_{ii}^{1/n} \right)}.$$

Table 2.2 shows the absolute value of the difference between the exact ω-condition number and the ω-condition numbers obtained by making use of each factorization, namely, ω_{eig}, ω_R and ω_{LU}. Surprisingly, we see that both the Cholesky and LU decompositions give better results than the eigenvalue decomposition.

n	Fact.	order κ 1e2	order κ 1e3	order κ 1e4	order κ 1e5	order κ 1e6	order κ 1e7	order κ 1e8	order κ 1e9
500	eig	5.5267e-02	5.7766e-02	5.2747e-02	5.9256e-02	6.0856e-02	6.2197e-02	5.5592e-02	5.7626e-02
	R	1.1218e-02	8.0907e-03	7.5172e-03	8.4705e-03	9.2774e-03	8.5553e-03	8.1462e-03	7.9027e-03
	LU	2.2893e-02	1.8159e-02	1.8910e-02	2.0902e-02	2.0057e-02	2.0308e-02	1.9060e-02	1.8879e-02
1000	eig	3.0664e-01	2.8968e-01	2.6095e-01	2.7796e-01	5.7083e-01	5.9007e-01	5.8351e-01	5.9630e-01
	R	2.9328e-02	2.8339e-02	2.7869e-02	3.1909e-02	5.8628e-02	6.0873e-02	6.2429e-02	6.1074e-02
	LU	7.5011e-02	7.2666e-02	7.0497e-02	7.6778e-02	1.6313e-01	1.7313e-01	1.7666e-01	1.7326e-01
2000	eig	3.4794e+00	3.4804e+00	3.1916e+00	3.4386e+00	3.4235e+00	3.4766e+00	3.2327e+00	3.3704e+00
	R	3.5644e-01	3.5989e-01	2.9556e-01	3.6375e-01	3.5847e-01	3.5972e-01	3.2629e-01	3.4227e-01
	LU	9.0136e-01	9.0537e-01	7.1161e-01	8.7445e-01	8.6420e-01	8.8027e-01	8.1990e-01	8.1383e-01

Table 2.1: CPU sec. for evaluating $\omega(A)$, averaged over the same 10 random instances; eig, R, LU are eigenvalue, Cholesky, LU decompositions, respectively.

3 Preconditioning for Generalized Jacobians

We now consider the problem of optimal preconditioning for low rank updates of very ill-conditioned (close to singular) positive definite matrices.

n	Fact.	order κ 1e2	order κ 1e3	order κ 1e4	order κ 1e5	order κ 1e6	order κ 1e7	order κ 1e8	order κ 1e9
500	eig	1.5632e-13	2.7853e-12	2.2618e-10	1.2695e-08	8.9169e-07	5.4109e-05	2.2610e-03	1.7349e-01
	R	1.7053e-13	2.5580e-12	1.0039e-10	1.1339e-08	4.9818e-07	2.6470e-05	1.3173e-03	1.6217e-01
	LU	1.5987e-13	2.4585e-12	1.0652e-10	1.1987e-08	5.1592e-07	2.1372e-05	1.3641e-03	1.4268e-01
1000	eig	2.1316e-13	2.1032e-12	8.7653e-11	4.6271e-09	3.1477e-07	1.9602e-05	9.9290e-04	7.6469e-02
	R	4.2633e-13	1.5632e-12	4.2235e-11	3.9297e-09	2.9562e-07	1.1498e-05	9.1506e-04	5.3287e-02
	LU	4.4054e-13	1.4850e-12	3.7858e-11	3.8287e-09	2.7390e-07	1.3820e-05	6.0492e-04	4.8568e-02
2000	eig	2.4336e-13	4.1780e-12	4.2019e-10	2.0080e-08	7.7358e-07	6.4819e-05	5.5339e-03	3.7527e-01
	R	4.3698e-13	2.0819e-12	5.0704e-11	2.3442e-09	1.8376e-07	8.9575e-06	5.5255e-04	4.8842e-02
	LU	4.3165e-13	2.2595e-12	2.3249e-11	2.5057e-09	1.5020e-07	6.0479e-06	5.4228e-04	4.4205e-02

Table 2.2: Precision of evaluation of $\omega(A)$ averaged over the same 10 random instances. eig, R, LU are eigenvalue, Cholesky, LU decompositions, respectively.

3.1 Preliminaries

More precisely, given a positive definite matrix $A \in \mathbb{S}^n_{++}$ and a matrix $U \in \mathbb{R}^{n\times t}$ with $t << n$, we aim to find $\gamma \in \mathbb{R}^t$ so as to minimize the condition number of the low rank update

$$A + U\,\mathrm{Diag}(\gamma)U^T. \tag{3.1}$$

Here $\mathrm{Diag}(v) \in \mathbb{S}^n_+$ is the diagonal matrix with diagonal given by the vector $v \in \mathbb{R}^n$.

This kind of updating arises when finding generalized Jacobians in nonsmooth optimization. We provide insight on the problem in the following Example 3.1.

Example 3.1 (Generalized Jacobians)**.** *In many nonsmooth and semismooth Newton methods one aims to find a root of a function $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form*

$$F(y) := B(v + B^T y)_+ - c,$$

where $B \in \mathbb{R}^{n\times m}$, $v \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $(\,\cdot\,)_+$ denotes the projection onto the nonnegative orthant, e.g., [3, 18, 23]. At every iteration of these algorithms a generalized Jacobian of F is computed of the form

$$J := \sum_{i\in\mathcal{I}_+} B_i B_i^T + \sum_{j\in\mathcal{I}_0} \gamma_j B_j B_j^T, \;\; \textit{with } \gamma_j \in [0,1]$$

and where B_i and B_j denote columns of B over the set of indices $\mathcal{I}_+ := \{i \in \{1,\dots,m\} : (v + B^T y)_i > 0\}$ and

$$\mathcal{I}_0 := \{j \in \{1,\dots,m\} : (v + B^T y)_j = 0 \textit{ and } (B_j)_{j\in\mathcal{I}_0} \textit{ is a maximal linearly independent subset}\}.$$

The generalized Jacobian J, which is usually singular, is then used to obtain a Newton direction $d \in \mathbb{R}^n$ by solving a least-square problem for the system $(J + \epsilon I)\,d = -F(y)$, where ϵI, with $\epsilon > 0$, is analogous to the regularization term of the well-known Levenberg–Marquardt method*. Thus, this linear system is very ill-conditioned. This makes preconditioning by optimal updating appropriate.*

The optimal preconditioned update can be done in our framework as we start with

$$A := \sum_{i \in \mathcal{I}_+} B_i B_i^T + \epsilon I, \quad U = [B_j]_{j \in \mathcal{I}_0},$$

and then find an optimal low rank update as in (3.1); done with additional box constraints on γ, namely, $\gamma \in [0,1]^t$.

Similar conditioning questions also appear in the normal equations matrix, ADA^T, in interior point methods, e.g., modifying the weights in D appropriately to avoid ill-conditioning [5, 12]. For other related work on minimizing condition numbers for low rank updates see e.g., [4, 13].

Here, we propose obtaining an optimal conditioning of the update (3.1) by using the ω-condition number of [8], instead of the classic κ-condition number. The ω-condition number presents some advantages with respect to the classic condition number, since it is differentiable and pseudoconvex in the interior of the positive semidefinite cone, which facilitates addressing minimization problems involving it. Our empirical results show a significant decrease in the number of iterations required for a requested accuracy in the residual.

3.2 Optimal Conditioning for Rank One Updates

We first consider the special case where the update is rank one. Related eigenvalue results for rank one updates are well known in the quasi-Newton literature, e.g., [7, 24]. We include this special rank one case as it yields several interesting and surprising results. The general rank-t update will be studied in Section 3.3, below.

Theorem 3.2. *Suppose we have a given $A \in \mathbb{S}^n_{++}$ and $u \in \mathbb{R}^n$. Let $A = QDQ^T = LL^T$ be the (orthogonal) spectral and Cholesky decomposition of A, respectively. Let $U = uu^T$ and define the rank one update*

$$A(\gamma) = A + \gamma U, \; \gamma \in \mathbb{R}.$$

Set

$$w_s = D^{-1/2} Q^T u, \quad w_c = L^{-1} u, \tag{3.2}$$

and

$$\gamma_s = \frac{\operatorname{tr}(A)\|w_s\|^2 - n\|u\|^2}{(n-1)\|u\|^2\|w_s\|^2}, \quad \gamma_c = \frac{\operatorname{tr}(A)\|w_c\|^2 - n\|u\|^2}{(n-1)\|u\|^2\|w_c\|^2}. \tag{3.3}$$

Then

$$\gamma^* = \gamma_s = \gamma_c \in \left]-\|w_s\|^{-2}, +\infty\right[$$

provides the optimal ω-conditioning, i.e.,

$$\gamma^* = \operatorname{argmin}_{A(\gamma) \succ 0} \omega(\gamma). \tag{3.4}$$

Proof. Let

$$f(\gamma) := \operatorname{tr}(A(\gamma))/n \quad \text{and} \quad g(\gamma) := \det(A(\gamma))^{1/n}.$$

We want to find the optimal γ to minimize the condition number

$$\omega(\gamma) = f(\gamma)/g(\gamma)$$

subject to $A(\gamma)$ being positive definite. By Proposition 2.1 1, $\omega : \mathbb{R} \to \mathbb{R}; \gamma \to \omega(\gamma)$ is pseudoconvex as long as $A(\gamma) \succ 0$. We prove that the later is true for γ belonging to an open interval in the real line. Indeed, let $A = QDQ^T$ be the spectral decomposition of A and define

$$w = D^{-1/2}Q^T u \quad \text{and} \quad W = ww^T = D^{-1/2}Q^T uu^T QD^{-1/2}. \tag{3.5}$$

Then we can rewrite

$$A(\gamma) = QD^{1/2}(I + \gamma W)D^{1/2}Q^T, \tag{3.6}$$

which is positive definite if and only if the rank one update of I, $I + \gamma W$, belongs to the cone of positive definite matrices. Now, note that the eigenvalues of this term are $\lambda_1 = 1$, with multiplicity $n - 1$, and $\lambda_2 = 1 + \gamma\|w\|^2$ with multiplicity 1. We then conclude that

$$A(\gamma) \in \mathbb{S}^n_{++} \iff \gamma \in \left]-\frac{1}{\|w\|^2}, +\infty\right[,$$

in which case $\lambda_2 > 0$. Here $]a, b[$ denotes the open interval in $\mathbb{R}$ formed by a, b. Moreover, $\omega(\gamma)$ tends to ∞ as γ approaches the extreme of the above interval. Therefore ω possesses a minimizer in the open interval, $\gamma^* \in]-\|w\|^{-2}, +\infty[$, that satisfies $\omega'(\gamma^*) = 0$. Note that since ω is pseudoconvex the fact that its derivative is equal to zero is also a sufficient condition for global optimality (see Fact 3.5 below).

In the following we obtain an explicit expression for the (unique) minimizer of (3.4), γ^*, by studying the zeros of ω'. We accomplish this in two ways using the spectral and Cholesky decomposition. Since the proofs for the two decompositions are alike, we use similar notation.

- We first consider the spectral decomposition. Using the notation introduced in (3.5), f and its derivative are expressed as

  $$f(\gamma) = \left(\operatorname{tr}(A) + \gamma\|u\|^2\right)/n \quad \text{and} \quad f'(\gamma) = \|u\|^2/n,$$

  respectively. By making use of (3.6), g becomes

  $$g(\gamma) := \left(\det(A)\det(I + \gamma W)\right)^{1/n},$$

  since $\det(D) = \det(A)$. As explained above the eigenvalues of $I + \gamma W$, are $\lambda_1 = 1 + \gamma\|w\|^2$, and the others are all 1, which yields that

  $$g(\gamma) = \left(\det(A)(1 + \gamma\|w\|^2)\right)^{1/n} = \det(A)^{1/n}(1 + \gamma\|w\|^2)^{1/n}.$$

We get

$$g'(\gamma) = \frac{1}{n}\det(A)^{1/n}\|w\|^2(1+\gamma\|w\|^2)^{(1-n)/n}.$$

The derivative of ω is then obtained as follows

$$\begin{aligned}
\omega'(\gamma) &= \frac{f'(\gamma)g(\gamma) - f(\gamma)g'(\gamma)}{g(\gamma)} \\
&= \frac{1}{g(\gamma)^2}\left[\frac{\|u\|^2}{n}\det(A)^{1/n}\left(1+\gamma\|w\|^2\right)^{1/n}\right. \\
&\quad \left. -\frac{\|w\|^2}{n^2}\left(\mathrm{tr}(A)+\gamma\|u\|^2\right)\det(A)^{1/n}\left(1+\gamma\|w\|^2\right)^{(1-n)/n}\right] \\
&= \frac{\det(A)^{1/n}}{g(\gamma)^2 n^2}\left(1+\gamma\|w\|^2\right)^{(1-n)/n}\left[n\|u\|^2 + (n-1)\gamma\|u\|^2\|w\|^2 - \mathrm{tr}(A)\|w\|^2\right].
\end{aligned} \tag{3.7}$$

A simple computation shows that this derivative is 0 only when γ attains the value

$$\gamma^* = \frac{\mathrm{tr}(A)\|w\|^2 - n\|u\|^2}{(n-1)\|u\|^2\|w\|^2}, \tag{3.8}$$

which then has to be in the interval $]-\|w\|^{-2}, +\infty[$. Since ω is pseudoconvex, we conclude that γ^* is the optimal preconditioner which solves (3.4). Finally, we recover the expression for γ_s in (3.3) by substituting w_s for w.

- We now consider the Cholesky decomposition. By abuse of notation, for this part of the proof only, we set

$$w = L^{-1}u \quad \text{and} \quad W = ww^T.$$

We can rewrite

$$\begin{aligned}
A(\gamma) &= A + \gamma U \\
&= LL^T + \gamma U \\
&= L\left(I + \gamma L^{-1}uu^TL^{-T}\right)L^T \\
&= L(I+\gamma W)L^T.
\end{aligned}$$

Same than in the spectral case, f and its derivative are expressed as

$$f(\gamma) = \left(\mathrm{tr}(A) + \gamma\|u\|^2\right)/n \quad \text{and} \quad f'(\gamma) = \|u\|^2/n,$$

and g has the form

$$g(\gamma) = \left(\det(A)(1+\gamma\|w\|^2)\right)^{1/n}.$$

The derivative of g is then

$$g'(\gamma) = \frac{1}{n}\det(A)\|w\|^2(1+\gamma\|w\|^2)^{(1-n)/n}.$$

It straightforward follows that the expression for the derivative of ω is given by

$$\omega'(\gamma) = \frac{\det(A)}{g(\gamma)^2 n^2}(1+\gamma\|w\|^2)^{(1-n)/n}\left[n\|u\|^2 + (n-1)\gamma\|u\|^2\|w\|^2 - \mathrm{tr}\,(A)\,\|w\|^2\right],$$

which equals to zero at the point

$$\gamma^* = \frac{\mathrm{tr}(A)\|w\|^2 - n\|u\|^2}{(n-1)\|u\|^2\|w\|^2}.$$

Again, substituting w for w_c we obtain the expression for γ_c in (3.3).

■

We now conclude a surprising relationship between the right singular vectors of the two factorizations in Corollary 3.4. First we need the following Lemma 3.3.

Lemma 3.3. *Let $X, Y \in \mathbb{S}^n$ be given and satisfy*

$$u^T X u = u^T Y u, \quad \forall u \in \mathbb{R}^n. \tag{3.9}$$

Then, $X = Y$.

Proof. We first start noting that the diagonal elements of X and Y coincide. Indeed, let $i \in \{1, \ldots, n\}$ and e_i be the ith unit vector of the standard basis of $\mathbb{R}^n$. Then, we get that $X_{ii} = Y_{ii}$ by just setting $u := e_i$ in (3.9).

Now, let $i, j \in \{1, \ldots, n\}$, with $i \neq j$. Setting $u := e_i + e_j$, equation (3.9) yields

$$X_{ii} + 2X_{ij} + X_{jj} = Y_{ii} + 2Y_{ij} + Y_{jj} \Longrightarrow X_{ij} = Y_{ij}.$$

Since this holds for any $i, j \in \{1, \ldots, n\}$, we conclude that $X = Y$.

■

Corollary 3.4. *Under the assumptions of Theorem 3.2, we get that the two matrices*

$$R_s := D^{-1/2}Q^T, \quad R_c := L^{-1}$$

are orthogonally equivalent[2]*, with the same right singular vectors. We conclude that*

$$A^{-1} = R_s^T R_s = R_c^T R_c.$$

Proof. Let γ_s and γ_c as in (3.3). By equating both expressions, we get after cancellation that

$$\|D^{-1/2}Q^T u\| = \|L^{-1}u\| \quad \forall u \in \mathbb{R}^n. \tag{3.10}$$

Squaring this equation we get,

$$u^T R_s^T R_s u = u^T R_c^T R_c u \quad \forall u \in \mathbb{R}^n.$$

By Lemma 3.3, this implies that $R_s^T R_s = R_c^T R_c$. Therefore, R_s and R_c share the same singular values and eigenvectors.

[2]The $s \times t$ matrices C, D are orthogonally equivalent if they have the same singular values.

■

As shown in Example 3.1, in some applications the preconditioner multiplier γ is required to take values in the interval $[0, 1]$. In the following, we analize the optimal ω-preconditioner for the rank 1 update subject to this interval constraint. This will require considering a constrained pseudoconvex minimization problem. In the following Fact 3.5, see e.g., [19, Chapter 10], we recall the sufficient optimality conditions for this class of optimization problems. We note that no constraint qualification is needed for *sufficiency*.

Fact 3.5 (Sufficient optimality conditions for pseudoconvex programming)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be nonempty open and convex. Let $f : \Omega \to \mathbb{R}$ be a pseudoconvex function and $(g_i)_{i=1}^m : \Omega \to \mathbb{R}$ a family of differentiable and quasiconvex functions. Consider the optimization problem*

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & x \in \Omega. \end{array} \tag{3.11}$$

Let $\bar{x} \in \Omega, \bar{\lambda} \in \mathbb{R}^m$, be a KKT primal-dual pair, i.e., the following KKT conditions hold:

$$\begin{aligned} \nabla f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla g_i(\bar{x}) &= 0 \\ \bar{\lambda}_i &\geq 0, \quad i = 1, \ldots, m \\ \bar{\lambda}_i g_i(\bar{x}) &= 0, \quad i = 1, \ldots, m \\ \bar{x} \in \Omega \text{ and } g_i(\bar{x}) &\leq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{3.12}$$

Then $\bar{x}$ solves (3.11).

Corollary 3.6. *Let the assumptions of Theorem 3.2 hold and let $\bar{\gamma}$ be the optimal ω-preconditioner in the interval $[0, 1]$, i.e.,*

$$\bar{\gamma} = \arg \min_{\substack{0 \leq \gamma \leq 1 \\ A(\gamma) \succ 0}} \omega(\gamma).$$

Then, if $\gamma^* \in]-\|w\|^2, +\infty[$ is the optimal "unconstrained" ω-preconditioner obtained in Theorem 3.2, the following hold:

(i) If $\gamma^* \in [0, 1] \implies \bar{\gamma} = \gamma^*$;

(ii) If $\gamma^* < 0 \implies \bar{\gamma} = 0$;

(iii) If $\gamma^* > 1 \implies \bar{\gamma} = 1$.

Proof. (i) In this case, since γ^* is the global optimum of ω in $]-\|w\|^2, +\infty[$, it would also be so in the interval $[0,1]$. For (ii) and (iii), we have a constrained pseudoconvex program in the form of (3.11) with $\Omega :=]-\|w\|^2, +\infty[$, $f := \omega$, $g_1(\gamma) = -\gamma$ and $g_2(\gamma) = \gamma - 1$. Therefore, it would be suffcient to check that the proposed minima satisfies the KKT conditions (3.12). We do this for the case (ii), and note that (iii) follows similarly.

Let $\bar{\gamma} = 0 \in \Omega$. Since g_1 is the unique active constraint, we just need to prove the existence of a Lagrange multiplier $\bar{\lambda}_1 \geq 0$ such that $(0, \bar{\lambda}_1)$ is a KKT point, i.e.,

$$\omega'(0) + \bar{\lambda}_1(-1) = 0.$$

For this, just note by (3.8), that

$$\mathrm{tr}(A)\|w\|^2 - n\|u\|^2 < 0,$$

since $\gamma^* < 0$. Then, making use of (3.7) we conclude that $\bar{\lambda}_1 := \omega'(0) \geq 0$ is a Lagrange multiplier, and thus $\bar{\gamma} = 0$ satisfies the sufficient condition of global optimality.

■

3.3 Optimal Conditioning with a Low Rank Update

We now consider the case where the update is low rank. We need the following notations. For a matrix $Z \in \mathbb{R}^{n\times t}$, we use MATLAB notation and define the function $\mathrm{norms}(Z) : R^{\mathrm{size}(Z)} \to \mathbb{R}^t$ as the (column) vector of column 2-norms of Z. We let $\mathrm{norms}^\alpha(Z)$ denote the vector of column norms with each norm to the power α.

Theorem 3.7 (Rank t-update)**.** *Let $A \in \mathbb{S}^n_{++}$, $U = [u_1, \ldots, u_t] \in \mathbb{R}^{n\times t}$, be given with $n > t \geq 2$, and $\mathrm{norms}(U) > 0$. Set*

$$A(\gamma) = A + U\,\mathrm{Diag}(\gamma)U^T, \text{ for } \gamma \in \mathbb{R}^t.$$

Let the spectral decomposition of A be given by $A = QDQ^T$, define $w_i = D^{-1/2}Q^Tu_i, i \in \{1, \ldots, t\}$, as in (3.2), with $W = \begin{bmatrix} w_1 \ldots w_t \end{bmatrix}$. Let

$$\begin{array}{rcl} K(U) & = & \left[n\,\mathrm{Diag}\left(\mathrm{norms}^2(U)\right) - e\,\mathrm{norms}^2(U)^T\right], \\ b(U) & = & \left(\mathrm{tr}(A)e - n\,\mathrm{norms}^2(U)./\,\mathrm{norms}^2(W)\right), \end{array} \tag{3.13}$$

where e denotes the vector of all ones and $./$ stands for the element-wise division of the two vectors. Then, the optimal ω-preconditioner,

$$\gamma^* = \mathrm{argmin}_{A(\gamma)\succ 0}\,\omega(\gamma), \tag{3.14}$$

is given component-wise by

$$\begin{aligned} (\gamma^*)_i &= (K(U)^{-1}b(U))_i \\ &= \frac{\mathrm{tr}(A)\|w_i\|^2 - (n-t+1)\|u_i\|^2}{(n-t)\|u_i\|^2\|w_i\|^2} - \frac{1}{(n-t)\|u_i\|^2}\sum_{j=1, j\neq i}^{t}\frac{\|u_j\|^2}{\|w_j\|^2}, \end{aligned} \tag{3.15}$$

for $i = 1, \ldots, t$.

Proof. Let $A \succ 0$ and

$$U = \begin{bmatrix} u_1 & \dots & u_t \end{bmatrix} \in \mathbb{R}^{n \times t}, \qquad \text{with } n > t \geq 2.$$

We consider the update of the form

$$A(\gamma) = A + U \operatorname{Diag}(\gamma) U^T = A + \sum_{i=1}^{t} \gamma_i u_i u_i^t, \quad \gamma \in \mathbb{R}^t.$$

Same than in Theorem 3.2, we start characterizing an open subset of $\mathbb{R}^t$ where $A(\gamma)$ is positive definite. In order to do this, we again transform the problem using the spectral decomposition of A, $A = QDQ^T$, and setting

$$w_i = D^{-1/2} Q^T u_i \quad \text{and} \quad W_i = w_i w_i^T \quad \text{for } i = 1, \dots, t.$$

Then, we can express $A(\gamma)$ as

$$\begin{aligned} A(\gamma) &= A + U \operatorname{Diag}(\gamma) U^T \\ &= QD^{1/2} \left(I + D^{-1/2} Q^T U \operatorname{Diag}(\gamma) U^T Q D^{-1/2} \right) D^{1/2} Q^T \\ &= QD^{1/2} \left(I + \sum_{i=1}^{t} \gamma_i (D^{-1/2} Q^T u_i)(u_i^T Q D^{-1/2}) \right) D^{1/2} Q^T \\ &= QD^{1/2} \left(I + \sum_{i=1}^{t} \gamma_i W_i \right) D^{1/2} Q^T. \end{aligned}$$

By repeatedly making use of the formula for the determinant of the sum of an invertible matrix and a rank one matrix (see e.g., [20, Example 4]), we obtain the following expression for the determinant of $A(\gamma)$

$$\det(A(\gamma)) = \det(A) \left(\prod_{i=1}^{t} (1 + \gamma_i \|w_i\|^2) \right). \tag{3.16}$$

Consequently, $A(\gamma)$ is nonsingular and, by continuity of the eigenvalues, positive definite for γ belonging to the set

$$\Omega := \left] -\frac{1}{\|w_1\|^2}, +\infty \right[\times \left] -\frac{1}{\|w_2\|^2}, +\infty \right[\times \dots \times \left] -\frac{1}{\|w_t\|^2}, +\infty \right[. \tag{3.17}$$

Now, note that the constraint $A(\gamma) \succ 0$ is a positive definite constraint, so it is convex. Therefore, if there exists some γ outside of Ω such that $A(\gamma) \succ 0$, we would loose the convexity of the feasible set, since $A(\gamma)$ is singular on the boundary of Ω. This implies that

$$A(\gamma) \succ 0 \iff \gamma \in \Omega.$$

Moreover, since $\omega(\gamma) \to +\infty$ as γ tends to the border of Ω or to $+\infty$, we can ensure that γ has a minimizer in Ω. Since the function is pseudoconvex on this open set, the global

minimum is attained at a point γ^* such that $\nabla\omega(\gamma^*) = 0$. Next, we prove that γ^* is given by (3.15).

For this, note that $f(\gamma)$ can be expressed as

$$f(\gamma) = \frac{1}{n}\operatorname{tr}(A + U\operatorname{Diag}(\gamma)U^T) = \frac{1}{n}\left(\operatorname{tr}(A) + \sum_{i=1}^{t}\gamma_i\|u_i\|^2\right) = \frac{1}{n}\left(\operatorname{tr}(A) + \gamma^T \operatorname{norms}^2(U)\right),$$

and its gradient is $\nabla f(\gamma) = \frac{1}{n}\operatorname{norms}^2(U)$. On the other hand, by (3.16) $g(\gamma)$ can be expressed as

$$g(\gamma) = \det(A)^{1/n}\left(\prod_{i=1}^{t}(1+\gamma_i\|w_i\|^2)\right)^{1/n}.$$

The gradient of $g(\gamma)$ is then given component-wise by

$$\frac{\partial g(\gamma)}{\partial\gamma_j} = \frac{1}{n}\det(A)^{1/n}\left(\prod_{i=1}^{t}(1+\gamma_i\|w_i\|^2)\right)^{(1-n)/n}\left(\prod_{i=1,i\neq j}^{t}(1+\gamma_i\|w_i\|^2)\right)\|w_j\|^2,\ j=1,\ldots,t.$$

We make use of these expressions in order to compute the partial derivatives of ω. For every $j = 1,\ldots,t$, we have

$$\begin{aligned}\frac{\partial\omega(\gamma)}{\partial\gamma_j} &= \frac{\frac{\partial f(\gamma)}{\partial\gamma_j}g(\gamma) - f(\gamma)\frac{\partial g(\gamma)}{\partial\gamma_j}}{g(\gamma)^2} \\ &= \frac{1}{g(\gamma)^2}\left[\frac{\|u_j\|^2}{n}\det(A)^{1/n}\left(\prod_{i=1}^{t}(1+\gamma_i\|w_i\|^2)\right)^{1/n}\right. \\ &\quad \left. - \frac{\|w_j\|^2}{n^2}\left(\operatorname{tr}(A)+\gamma^T\operatorname{norms}^2(U)\right)\left(\prod_{i=1,i\neq j}^{t}(1+\gamma_i\|u_i\|^2)\right)^{(1-n)/n}\left(\prod_{i=1,i\neq j}^{t}(1+\gamma_i\|u_i\|^2)\right)\right].\end{aligned}$$

By defining the positive function $C(\gamma):\mathbb{R}^t\to\mathbb{R}_{++}$ as

$$C(\gamma) = \frac{\det(A)^{1/n}}{g(\gamma)^2 n^2}\left(\prod_{i=1,i\neq j}^{t}(1+\gamma_i\|u_i\|^2)\right)^{(1-n)/n}\left(\prod_{i=1,i\neq j}^{t}(1+\gamma_i\|u_i\|^2)\right),$$

we finally get that

$$\frac{\partial\omega(\gamma)}{\partial\gamma_j} = C(\gamma)\left[n\|u_j\|^2(1+\gamma_j\|w_j\|^2) - \left(\operatorname{tr}(A)+\gamma^T\operatorname{norms}^2(U)\right)\|w_j\|^2\right], \tag{3.18}$$

for all $j = 1,\ldots,t$. After setting the derivative (gradient) of ω to zero, and ignoring the positive factor given by C, we get that the minimum of the pseudoconvex function is obtained as the solution of the linear system defined by the t equations

$$(n-1)\|u_k\|^2\gamma_k - \sum_{i=1,i\neq k}^{t}\|u_i\|^2\gamma_i = \operatorname{tr}(A) - n\frac{\|u_k\|^2}{\|w_k\|^2},\quad k=1,\ldots,t.$$

Equivalently,

$$\begin{bmatrix} (n-1)\|u_1\|^2 & -\|u_2\|^2 & \dots & & -\|u_t\|^2 \\ -\|u_1\|^2 & (n-1)\|u_2\|^2 & -\|u_3\|^2 & \dots & -\|u_t\|^2 \\ \dots & & & & \\ -\|u_1\|^2 & \dots & \dots & -\|u_{t-1}\|^2 & (n-1)\|u_t\|^2 \end{bmatrix} \gamma = \begin{pmatrix} \mathrm{tr}(A) - n\|u_1\|^2/\|w_1\|^2 \\ \dots \\ \mathrm{tr}(A) - n\|u_t\|^2/\|w_t\|^2 \end{pmatrix}.$$

This is further equivalent to

$$\left[n\,\mathrm{Diag}(\mathrm{norms}^2(U)) - e\,\mathrm{norms}^2(U)^T\right]\gamma = \left(\mathrm{tr}(A)e - n\,\mathrm{norms}^2(U)./\,\mathrm{norms}^2(W)\right),$$

which is the system $K(U)\gamma = b(U)$ using the notation in (3.13).

Now we derive an explicit expression for the optimal γ. In order to do this, note that $K(U)$ is given as the sum of an invertible matrix, $n\,\mathrm{Diag}(\mathrm{norms}^2(U))$, and an outer product of vectors, $-e\,\mathrm{norms}^2(U)^T$. By the *Sherman-Morrison formula*, this sum is invertible if and only if

$$1 - \frac{1}{n}\,\mathrm{norms}^2(U)^T\,\mathrm{Diag}(\mathrm{norms}^2(U))^{-1}e \neq 0.$$

This is always true for $t < n$. Indeed, we have

$$1 - \frac{1}{n}\,\mathrm{norms}^2(U)^T\,\mathrm{Diag}(\mathrm{norms}^2(U))^{-1}e = 1 - \frac{1}{n}e^Te = 1 - \frac{t}{n} > 0.$$

Moreover, we obtain the following expression for the inverse

$$\begin{aligned} &\left(n\ \mathrm{Diag}(\mathrm{norms}^2(U)) - e\,\mathrm{norms}^2(U)^T\right)^{-1} \\ &\quad= \frac{1}{n}\,\mathrm{Diag}(\mathrm{norms}^2(U))^{-1} + \frac{1}{\left(1-\frac{t}{n}\right)n^2}\,\mathrm{Diag}(\mathrm{norms}^2(U))^{-1}e\,\mathrm{norms}^2(U)^T\,\mathrm{Diag}(\mathrm{norms}^2(U))^{-1} \\ &\quad= \frac{1}{n}\,\mathrm{Diag}(\mathrm{norms}^{-2}(U)) + \frac{1}{(n-t)n}\,\mathrm{Diag}(\mathrm{norms}^{-2}(U))ee^T. \end{aligned}$$

Therefore, the inverse of $K(U)$ in matrix form is given by

$$K(U)^{-1} = \frac{1}{n}\begin{bmatrix} \frac{1}{\|u_1\|^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\|u_2\|^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\|u_t\|^2} \end{bmatrix} + \frac{1}{(n-t)n}\begin{bmatrix} \frac{1}{\|u_1\|^2} & \frac{1}{\|u_1\|^2} & \dots & \frac{1}{\|u_1\|^2} \\ \frac{1}{\|u_2\|^2} & \frac{1}{\|u_2\|^2} & \dots & \frac{1}{\|u_2\|^2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\|u_t\|^2} & \frac{1}{\|u_t\|^2} & \dots & \frac{1}{\|u_t\|^2} \end{bmatrix}.$$

Finally, we obtain γ^* by calculating the product $\gamma^* = K(U)^{-1}b(U)$ which yields

$$\gamma_i^* = \frac{\mathrm{tr}(A)\|w_i\|^2 - (n-t+1)\|u_i\|^2}{(n-t)\|u_i\|^2\|w_i\|^2} - \frac{1}{(n-t)\|u_i\|^2}\sum_{j=1,j\neq i}^{t}\frac{\|u_j\|^2}{\|w_j\|^2}, \tag{3.19}$$

for all $i = 1, \dots, t$. Since γ^* is the unique zero of the gradient of ω, we conclude that it belongs to Ω and solves (3.14).

■

We note that the optimal ω-preconditioner for the rank one update in Theorem 3.2 is obtained from (3.15) when $t = 1$. On the other hand, we can also employ the Cholesky decomposition of A to derive the optimal ω-preconditioner in Theorem 3.7. We state this in the following corollary.

Corollary 3.8. *Given A and U as in Theorem 3.7. Let $A = LL^T$ be the Cholesky decomposition of A. Then, the formula for the optimal ω-preconditioner γ^* in (3.15) holds with the replacement*

$$w_i \leftarrow L^{-1}u_i, \quad i = 1, \ldots, t.$$

Proof. The result follows from (3.10) in Corollary 3.4.

■

With the same assumptions as in Theorem 3.7, we now consider the problem of finding the optimal ω-preconditioner in the box $[0, 1]^t$, i.e.,

$$\bar{\gamma} = \arg \min_{\substack{\gamma \in [0,1]^t \\ A(\gamma) \succ 0}} \omega(\gamma). \tag{3.20}$$

For the rank one update ($t = 1$), Corollary 3.6 shows that the solution to (3.20) can be obtained by first computing the minimum of the unconstrained problem, whose explicit expression was given in Theorem 3.2, and then projecting onto the box constraint, which in that case was the interval $[0, 1]$. However, this simple projection can fail in general for the low rank update, as we now show in Example 3.9.

Example 3.9 (Failure of the projection for solving the constrained problem (3.20))**.** *Let $n = 3$, $t = 2$ and consider the following initial data for the omega minimization problem*

$$A := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad U := \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix}.$$

Then, we get the following:

- *The optimal ω-preconditioner from Theorem 3.7 is*

$$\gamma^* = \frac{1}{3} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

  This follows from (3.19).

- *When projecting onto the set* $[0,1]^2$ *we get the point*

$$\gamma_p^* = \frac{1}{3}\begin{pmatrix}1\\0\end{pmatrix},$$

where ω *attains a value of* $\omega(\gamma_p^*) = 16/(9\sqrt[3]{5})$.

- *However, the value of* ω *is lower at the point*

$$\bar{\gamma} = \frac{1}{2}\begin{pmatrix}1\\0\end{pmatrix},$$

since $\omega(\bar{\gamma}) = 1/3\sqrt[3]{(2/11)^2} < 16/(9\sqrt[3]{5})$. *In fact,* $\bar{\gamma}$ *is the optimal* ω*-preconditioner in* $[0,1]^2$, *as we now show.*

To prove the last statement, note that (3.20) *can be written as the pseudoconvex program in* (3.11) *by setting* $f := \omega : \mathbb{R}^2 \to \mathbb{R}$, $g_1(\gamma) = -\gamma_1$, $g_2(\gamma) = -\gamma_2$, $g_3(\gamma) = \gamma_1 - 1$, $g_4(\gamma) = \gamma_2 - 2$ *and* Ω *defined as in* (3.17). *In particular, the only active constraint for* $\bar{\gamma} = (1/2, 0)^T$ *is* $g_2(\gamma) = 0$, *so the KKT conditions become*

$$0 = \frac{\partial\omega(\bar{\gamma})}{\partial\gamma_1},$$
$$0 = \frac{\partial\omega(\bar{\gamma})}{\partial\gamma_2} - \bar{\lambda}_2,$$

for some $\bar{\lambda}_2 \geq 0$. *This can be verified by simply substituting using the expressions of the partial derivatives of* ω *obtained in* (3.18). *By Fact 3.5, we conclude that for the given data,* $\bar{\gamma}$ *is the solution of* (3.20).

As done in the previous example, obtaining the optimal ω-preconditioner in the box $[0,1]^t$ would require obtaining a KKT point for the constrained pseudoconvex problem (3.20). This is not an easy task. To the author's knowledge, closed formulas for this kind of box constrained minimization problems are not known even when the objective is a quadratic. Nevertheless, using the projection of γ^* onto $[0,1]^t$ as an approximation to $\bar{\gamma}$ appears to give good results in practice. We see this in our numerical tests in Section 4.

4 Numerical Tests

We now present tests with different choices of γ for iteratively solving a linear system of the form $A(\gamma)x = b$. We us MATLAB's builtin functions **lsqr** and **cgs**. We focus our attention on the case where $A(\gamma) \in \mathbb{S}^n_{++}$ is a low rank update that appears in nonsmooth Newton methods, see Example 3.1.

4.1 Problem Generation

Specifically, we generate random instances of the linear system in the following way. We set

$$A(\gamma) := A + \epsilon I + U^T Diag(\gamma) U,$$

where ϵ is a random number in the interval $[10^{-7}, 10^{-9}]$, and A, U and b are chosen as follows:

- A is defined as $A = A_0 A_0^T$ with $A_0 \in \mathbb{R}^{r \times n}$ a normally distributed random sparse matrix of random density smaller than $0.5/\log(n)$ and r a random integer in the interval $[n/2 + 1, n - 1]$.
- The rank of the update, t, is randomly chosen in the interval $[2, r/2]$, and $U \in \mathbb{R}^{n \times t}$ is a normally distributed random sparse matrix of random density smaller than $1/\log(n)$.
- The right hand side, b, is chosen as the sum of two random vectors in the range of A and U, respectively. More precisely,

  $$c = A\, b^1 + U\, b^2,$$

  with $b^1 \in \mathbb{R}^n$ and $b^2 \in \mathbb{R}^t$ vectors randomly generated using the standard normal distribution.

As explained in Example 3.1, in this application the preconditioner γ is required to belong to the hypercube $[0, 1]^t$. Therefore, in our experiment we test the performance of four different choices of the preconditioner γ:

- The zero vector ($\gamma = 0$).
- The vector of ones ($\gamma = e$).
- Another common choice consists in setting the ith component of γ as

  $$\gamma_i = \min\{1, 1/\|u_i\|^2\},$$

  where we recall u_i denotes the ith column of U. In order to simplify notation, we use $\gamma = u^{-2}$ in the plots for this choice.
- The projection onto $[0, 1]^t$ of the optimal ω-preconditioner obtained in Theorem 3.7 ($\gamma = \gamma_p^*$). We recall that this is not necessarily the optimal ω-preconditioner in the set $[0, 1]^t$, i.e., it is not the solution of (3.20); but rather it is a heuristic approximation of it.

4.2 Description/Performance of Results

For each dimension choice $n \in \{100, 200, 500, 1000, 2000\}$, we generate 10 instances of random problems (50 problems in total) and solve each one of the system with the four different choices of preconditioner, and with both **lsqr** and **cgs**. We always employ the origin as initial point for the iterative algorithms.

Tables 4.1 and 4.2, are for **lsqr** and **cgs**, respectively. The tables show the average κ- and ω-condition numbers of every $A(\gamma)$ and the average relative residual, number of iterations and time employed by **lsqr** and **cgs** to solve the system for every choice of γ with a tolerance of 10^{-12} or until more than $50,000$ iterations are performed. Note that, both **lsqr** and **cgs** can stop before reaching the tolerance and the maximum number of iterations if two consecutive iterations are the same. This often happens for $\gamma = 0$. The two last columns of the tables indicate the time required for computing γ_p^* by making use of the spectral and the Cholesky decompositions, respectively. Regarding the difference in time, we want to mention that although obtaining the Cholesky decomposition of the positive definite matrix $A = LL^T$ is in general more efficient than computing its eigenvalue decomposition, the computation of the optimal ω-preconditioner in this case requires solving the system $LW = U$, see Corollary 3.8 and (3.2). This means that, for larger dimensions, employing the spectral decomposition for computing γ^* is more time efficient.

We also employ performance profiles, e.g., [10], to compare the different choices of preconditioners γ. These plots are constructed as follows. Let P denote a set of problems, and $\Gamma := \{0, e, u^{-2}, \gamma_p^*\}$ the set of preconditioners into comparison. For each $p \in P$ and $\gamma \in \Gamma$ we denote as $t_{p,\gamma}$ the measure we want to compare. In particular, we will separately consider the number of iterations and the time required for solving the system $A(\gamma)\, x = b$. In the case in which we are studying the time and $\gamma = \gamma_p^*$, the time for computing γ_p^*, this is computing the optimal ω-preconditioner γ^* and its projection onto $[0, 1]^t$ with the spectral decomposition, is also included in $t_{p,\gamma}$, i.e.,

$$t_{p,\gamma_p^*} = \{\text{time for solving the system } A(\gamma)\, x = b\} + \{\text{time for computing } \gamma_p^*\}.$$

Then, for every problem $p \in P$ and every $\gamma \in \Gamma$, we define the performance ratio as

$$r_{p,\gamma} := \begin{cases} \frac{t_{p,\gamma}}{\min\{t_{p,\gamma} : \gamma \in \Gamma\}} & \text{if convergence test passed,} \\ +\infty & \text{if convergence test failed.} \end{cases}$$

In our experiment, a convergence test passed if it succeded in solving the linear system with the required tolerance in less than 50 000 iterations, and failed otherwise. Note that the best performing preconditioner with respect to the measure under study (time or number of iterations), say $\tilde{\gamma}$, for problem p will have performance ratio $r_{p,\tilde{\gamma}} = 1$. In contrast, if the preconditioner γ underperforms in comparison with $\tilde{\gamma}$, but still manages to pass the test, then

$$r_{p,\gamma} = \frac{t_{p,\gamma}}{t_{p,\tilde{\gamma}}} > 1,$$

is the ratio between the overall time (number of iterations) required for solving the problem p for this particular choice and the time (number of iterations) employed by $\tilde{\gamma}$. Consequently, the larger the value of $r_{p,\gamma}$, the worse the preconditioner γ performed for problem p.

n	γ	$\kappa(A(\gamma))$	$\omega(A(\gamma))$	Rel. Residual	No. Iterations	Time	T. γ_p^* Spec	T. γ_p^* Chol
100	0	5.1176e+10	1.3955e+04	6.6846e-08	599.70	0.0072	-	-
	e	6.2576e+10	1.5794e+03	8.1311e-13	521.80	0.0054	-	-
	u^{-2}	5.1860e+10	1.5917e+03	5.8289e-13	659.60	0.0059	-	-
	γ_p^*	5.2773e+10	1.5055e+03	8.1212e-13	508.60	0.0040	0.0015	0.0007
200	0	2.8445e+10	9.4045e+03	2.5085e-08	1401.50	0.0270	-	-
	e	3.9524e+10	3.1707e+02	8.4896e-13	873.10	0.0192	-	-
	u^{-2}	2.6057e+10	3.6045e+02	8.7180e-13	1703.80	0.0194	-	-
	γ_p^*	2.9029e+10	2.9642e+02	8.2688e-13	832.80	0.0104	0.0018	0.0015
500	0	1.1976e+11	5.3664e+03	6.5646e-08	5637.60	0.6599	-	-
	e	1.0145e+11	2.4445e+02	9.8921e-13	4815.90	0.6816	-	-
	u^{-2}	5.6245e+10	2.8308e+02	9.9316e-13	18527.30	0.3633	-	-
	γ_p^*	6.5105e+10	2.1025e+02	9.7595e-13	3818.20	0.0763	0.0135	0.0192
1000	0	8.5673e+11	1.0529e+04	9.3333e-08	4546.70	2.2612	-	-
	e	7.4309e+11	2.2239e+03	9.8922e-13	7276.20	3.3259	-	-
	u^{-2}	7.1303e+11	2.4369e+03	7.7621e-06	23056.00	1.5090	-	-
	γ_p^*	7.5403e+11	2.2234e+03	9.8838e-13	7195.90	0.4960	0.0495	0.0838
2000	0	4.6865e+11	1.4188e+04	1.5624e-03	6044.70	8.4669	-	-
	e	2.0223e+12	2.6540e+03	9.9595e-13	3552.70	11.6022	-	-
	u^{-2}	4.7477e+11	2.2786e+03	9.9166e-13	10893.40	5.1524	-	-
	γ_p^*	5.7969e+11	1.9914e+03	9.8609e-13	1536.40	0.7630	0.2973	0.9005

Table 4.1: **LSQR**: For different dimensions n, and every choice of preconditioner γ, average κ- and ω-condition numbers of $A(\gamma)$, and average residual, number of iterations and time for solving the system $A(\gamma) = b$ with MATLAB's **lsqr** for the same 10 random instances of data. The last two columns gather the time for computing γ_p^*, which includes obtaining the optimal ω-preconditioner and subsequently projecting onto $[0, 1]^t$, with the spectral and Cholesky decomposition, respectively.

n	γ	$\kappa(A(\gamma))$	$\omega(A(\gamma))$	Rel. Residual	No. Iterations	Time	T. γ_p^* Spec	T. γ_p^* Chol
100	0	5.1176e+10	1.3955e+04	3.9280e-07	327.90	0.0036	-	-
	e	6.2576e+10	1.5794e+03	5.6786e-13	140.90	0.0012	-	-
	u^{-2}	5.1860e+10	1.5917e+03	5.8033e-13	165.10	0.0014	-	-
	γ_p^*	5.2773e+10	1.5055e+03	5.6108e-13	133.70	0.0008	0.0017	0.0007
200	0	2.8445e+10	9.4045e+03	6.1834e-07	533.00	0.0094	-	-
	e	3.9524e+10	3.1707e+02	7.2098e-13	114.20	0.0023	-	-
	u^{-2}	2.6057e+10	3.6045e+02	7.1703e-13	188.30	0.0023	-	-
	γ_p^*	2.9029e+10	2.9642e+02	7.4175e-13	112.20	0.0013	0.0018	0.0015
500	0	1.1976e+11	5.3664e+03	5.1672e-07	518.30	0.0683	-	-
	e	1.0145e+11	2.4445e+02	6.8841e-13	215.10	0.0369	-	-
	u^{-2}	5.6245e+10	2.8308e+02	7.2216e-13	508.90	0.0107	-	-
	γ_p^*	6.5105e+10	2.1025e+02	8.3652e-13	185.20	0.0038	0.0133	0.0189
1000	0	8.5673e+11	1.0529e+04	7.3646e-08	390.20	0.2322	-	-
	e	7.4309e+11	2.2239e+03	7.8024e-13	181.60	0.1056	-	-
	u^{-2}	7.1303e+11	2.4369e+03	7.6439e-13	487.20	0.0249	-	-
	γ_p^*	7.5403e+11	2.2234e+03	7.4485e-13	183.40	0.0091	0.0466	0.0818
2000	0	4.6865e+11	1.4188e+04	5.2164e-05	658.10	10.3990	-	-
	e	2.0223e+12	2.6540e+03	6.8001e-13	123.10	0.5321	-	-
	u^{-2}	4.7477e+11	2.2786e+03	6.8386e-13	230.10	0.1089	-	-
	γ_p^*	5.7969e+11	1.9914e+03	8.3978e-13	76.80	0.0404	0.2919	0.9060

Table 4.2: **CGS**: For different dimensions n, and every choice of preconditioner γ, average κ- and ω-condition numbers of $A(\gamma)$, and average residual, number of iterations and time for solving the system $A(\gamma) = b$ with MATLAB's **cgs** for the same 10 random instances of data. The last two columns gather the time for computing γ_p^*, which includes obtaining the optimal ω-preconditioner and subsequently projecting onto $[0, 1]^t$, with the spectral and Cholesky decomposition, respectively.

Finally, the performance profile of $\gamma \in \Gamma$ is defined as

$$\rho_\gamma(\tau) := \frac{1}{|P|} \text{size}\left\{p \in P \,:\, r_{p,\gamma} \leq \tau\right\},$$

where $|P|$ is the number of problems in P. This can be undestood as the relative portion of times that the performance ratio $r_{p,\gamma}$ is whithin a factor of $\tau \geq 1$ of the best possible performance ratio. In particular, $\rho_\gamma(1)$ represents the number of problems where γ is the best choice. Also, the existense of a $\tau \geq 1$ such that $\rho_\gamma(\tau) = 1$, indicates that γ passed the convergence test for every single problem in P.

For the same experiment described above, we display the preformance profiles, with $\log_2$ scale on τ, of the time and number of iterations required by **lsqr** for solving the linear systems $A(\gamma) = b$ with the differnet choices of preconditioners in Figure 4.1. The same is presented for **cgs** in Figure 4.2.

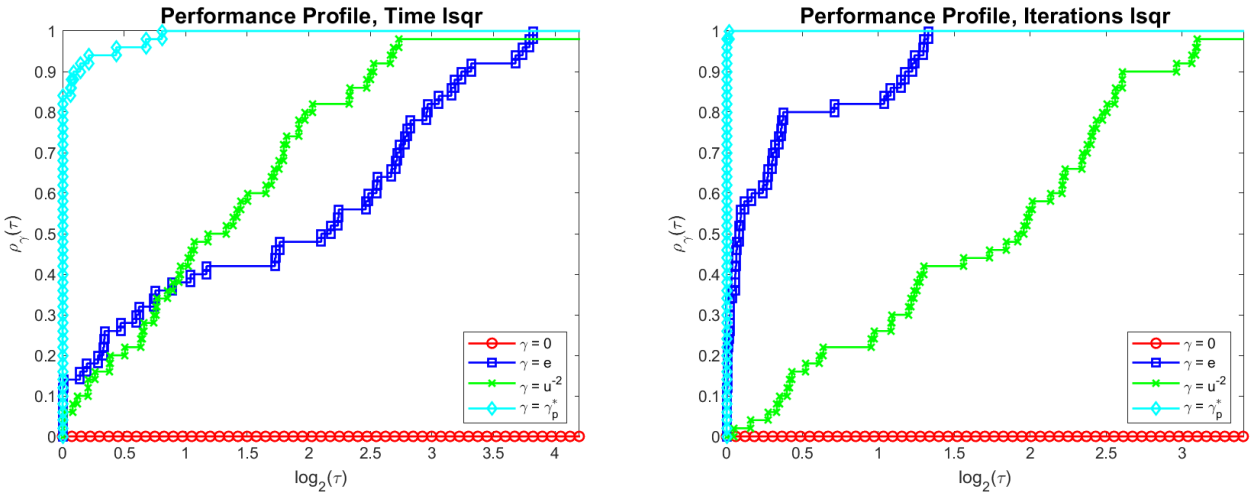

Figure 4.1: Performance profiles for the time (left) and number of iterations (right) required for solving the system $A(\gamma) = b$ with the different choices of preconditioner γ using MATLAB's **lsqr**.

4.3 Conclusions of the Empirics

First, we observe that the preconditioner $\gamma = 0$ was notably underperforming with respect to the other choices. In fact, it never achieves the desired tolerance. So we omit it from the discussion below. The projection of the optimal ω-preconditioner onto $[0, 1]^t$, this is $\gamma = \gamma_p^*$, was the best performing preconditioner, in the sense that the average time (without counting the time for computing γ_p^*) and average number of iterations by both the **lsqr** and **cgs** methods is lower for this choice of γ as shown in Table 4.1 and Table 4.2, respectively. In particular, in the iterations performance plots we can observe that in almost every instance both solvers required less number of iterations when $\gamma = \gamma_p^*$. The significant improvement

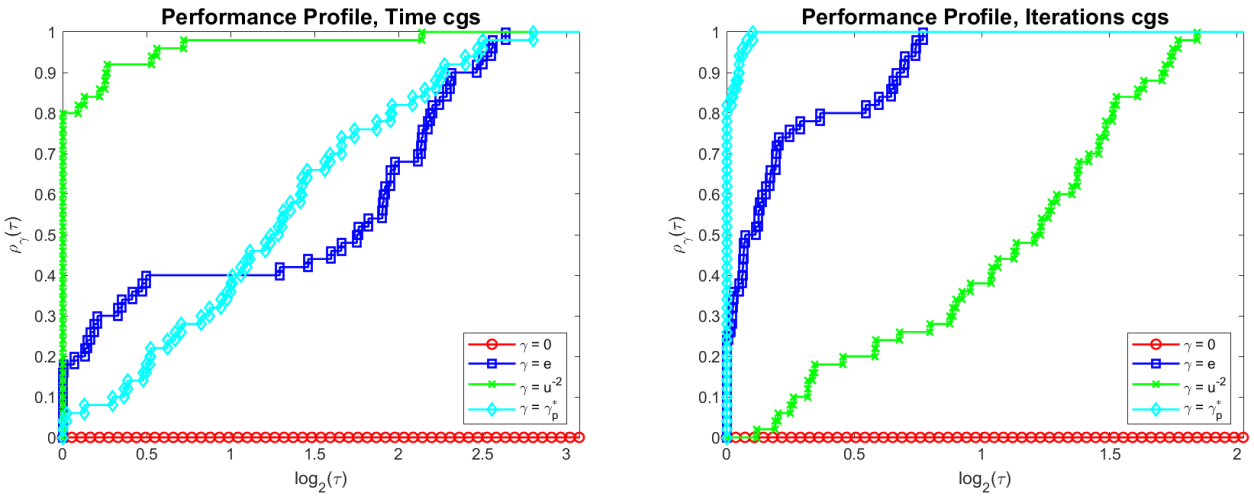

Figure 4.2: Performance profiles for the time (left) and number of iterations (right) required for solving the system $A(\gamma) = b$ with the different choices of preconditioner γ using MATLAB's **cgs**.

on time is also clear for the **lsqr** solver, where $\gamma = \gamma_p^*$ again beats both $\gamma = e$ and $\gamma = u^{-2}$ in a large number of instances, even after taking into account the time for computing γ_p^*, see Figure 4.1 (left). However, this does not happen for **cgs**. Although in this case the average time for this solver remains better for $\gamma = \gamma_p^*$ with respect to the other two options, once we add the time for computing γ_p^*, we observe on Figure 4.2 (left) and Table 4.2, that $\gamma = u^{-2}$ requires less average total time.

We therefore conclude that the projection of the optimal ω-preconditioner onto $[0, 1]^t$ is the best choice of preconditioner. However, evaluating γ_p^* requires some additional computational cost that seems to negate the solution success in the case of **cgs**. This is not the case for **lsqr**, where γ_p^* is definetly the best choice for γ.

We conclude that the ω-condition number is a good measure for determining well- or ill-conditioned linear systems as there is a direct correspondence between the value of $\omega(A(\gamma))$ and the efficiency of the iterative methods. Actually, looking at the third and forth column in Table 4.1 and Table 4.2, the results obtained are more correlated to the ω-condition number than to the κ-condition number. Specifically, a lower ω-condition number appears to be a better indication of performance of the solver (in both time and number of iterations) than a lower κ-condition number. In particular, γ_p^* which, as previously stated was significantly the best performing preconditioner, always makes $A(\gamma_p^*)$ have the smallest ω-condition number, but $\kappa(A(\gamma_p^*))$ was often larger than $\kappa(A(e))$ and $\kappa(A(u^{-2}))$.

5 Conclusion

In this paper we have studied the nonclassical matrix *ω-condition number*, i.e., the ratio of the arithmetic and geometric means of eigenvalues. We have shown that this condition

number has many properties that are advantageous over the classic κ-*condition number* that is the ratio of the largest to smallest eigenvalue. In particular, the differentiability of $\omega(A)$ facilitates finding optimal parameters for improving condition numbers. This was illustrated by characterizing the optimal parameters for low rank updates of positive definite matrices that arise in the context of nonsmooth Newton methods.

Moreover, the ω-condition number, when compared to the classical κ-condition number, is significantly more closely correlated to reducing the number of iterations and time for iterative methods for positive definite linear systems. This matches known results that show that preconditioning for clustering of eigenvalues helps in iterative methods, i.e., using all the eigenvalues rather than just the largest and smallest is better. This is further evidenced by the empirics that show that $\omega(A)$ is a significantly better estimate of the true conditioning of a linear system, i.e., how perturbations in the data A, b effect the solution x.

Finally, we have shown that an exact evaluation of $\omega(A)$ can be found using either the Cholesky or LU factorization. This is in contrast to the evaluation of $\kappa(A)$ that requires either a spectral decomposition or $\|A\|\|A^{-1}\|$ evaluation.

The results we presented here can be extended beyond A positive definite by replacing eigenvalues with singular values in the definition of $\omega(A)$.

Acknowledgements The authors would like to thank Haesol Im and Walaa M. Moursi for many useful and helpful conversations.

Funding The author D. Torregrosa-Belén was partially supported by the Ministry of Science, Innovation and Universities of Spain and the European Regional Development Fund (ERDF) of the European Commission, Grant PGC2018-097960-B-C22, the Generalitat Valenciana (AICO/2021/165) and by MINECO and European Social Fund (PRE2019-090751) under the program "Ayudas para contratos predoctorales para la formación de doctores" 2019.
All the authors were partially supported by the National Research Council of Canada.

Declarations

Conflict of interest The authors declare they have no conflict of interest.

Index

References

[1] L. Bergamaschi, *A survey of low-rank updates of preconditioners for sequences of symmetric linear systems*, Algorithms, 13 (2020). 3

[2] L. Bergamaschi, R. Bru, and A. Martínez, *Low-rank update of preconditioners for the inexact newton method with spd jacobian*, Mathematical and Computer Modelling, 54 (2011), pp. 1863–1873. Mathematical models of addictive behaviour, medicine & engineering. 3

[3] Y. Censor, , W. Moursi, T. Weames, and H. Wolkowicz, *Regularized nonsmooth newton algorithms for best approximation with applications*, tech. rep., University of Waterloo, Waterloo, Ontario, 2022 submitted. 37 pages, research report. 3, 9

[4] X. Chen, R. S. Womersley, and J. J. Ye, *Minimizing the condition number of a Gram matrix*, SIAM J. Optim., 21 (2011), pp. 127–148. 10

[5] S. Cipolla and J. Gondzio, *Proximal Stabilized Interior Point Methods and Low – Frequency – Update Preconditioning Techniques*, J. Optim. Theory Appl., 197 (2023), pp. 1061–1103. 10

[6] W. C. Davidon, *Optimally conditioned optimization algorithms without line searches*, Mathematical Programming, 9 (1975), pp. 1–30. 3

[7] J. Dennis Jr. and R. Schnabel, *Least change secant updates for quasi-Newton methods*, SIAM Review, 21 (1979), pp. 443–459. 10

[8] J. Dennis Jr. and H. Wolkowicz, *Sizing and least-change secant methods*, SIAM J. Numer. Anal., 30 (1993), pp. 1291–1314. 4, 5, 10

[9] X. Doan, S. Kruk, and H. Wolkowicz, *A robust algorithm for semidefinite programming*, Optim. Methods Softw., 27 (2012), pp. 667–693. 5

[10] E. Dolan and J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213. 22

[11] G. Golub and C. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013. 6

[12] J. Gondzio, S. Pougkakiotis, and J. Pearson, *General-purpose preconditioning for regularized interior point methods*, Comput. Optim. Appl., 83 (2022), pp. 727–757. 10

[13] C. Greif and J. M. Varah, *Minimizing the condition number for small rank modifications*, SIAM J. Matrix Anal. Appl., 29 (2006/07), pp. 82–97. 10

[14] R. Grimes and J. Lewis, *Condition number estimation for sparse matrices*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 384–388. 7

[15] W. W. Hager, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316. 4, 7

[16] N. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596. 7

[17] N. Higham and F. Tisseur, *A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1185–1201 (electronic). 7

[18] H. Hu, H. Im, X. Li, and H. Wolkowicz, *A semismooth Newton-type method for the nearest doubly stochastic matrix problem*, Math. Oper. Res., May (2023). arxiv.org/abs/2107.09631, 35 pages. 9

[19] O. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, NY, 1969. 5, 14

[20] K. S. Miller, *On the inverse of the sum of matrices*, Mathematics magazine, 54 (1981), pp. 67–72. 16

[21] S. S. Oren and D. G. Luenberger, *Self-scaling variable metric (ssvm) algorithms: Part i: Criteria and sufficient conditions for scaling a class of algorithms*, Management Science, 20 (1974), pp. 845–862. 4

[22] G. Pini and G. Gamboniati, *Is a simple diagonal scaling the best preconditioner for conjugate gradients on supercomputers?*, Advances in Water Resources, 13 (1990), pp. 147–153. 5

[23] H. Qi and D. Sun, *A quadratically convergent Newton method for computing the nearest correlation matrix*, SIAM journal on matrix analysis and applications, 28 (2006), pp. 360–385. 9

[24] R. Schnabel, *Analysing and improving quasi-Newton methods for unconstrained optimization*, PhD thesis, Department of Computer Science, Cornell University, Ithaca, NY, 1977. Also available as TR-77-320. 10

[25] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, vol. 1993, Springer, 1980. 3