SIAM J. MATRIX ANAL. APPL.
Vol. 40, No. 2, pp. 774–799

THE STRUCTURED CONDITION NUMBER OF A DIFFERENTIABLE MAP BETWEEN MATRIX MANIFOLDS, WITH APPLICATIONS[*]

BAHAR ARSLAN[†], VANNI NOFERINI[‡], AND FRANÇOISE TISSEUR[§]

Abstract. We study the structured condition number of differentiable maps between smooth matrix manifolds, extending previous results to maps that are only $\mathbb{R}$-differentiable for complex manifolds. We present algorithms to compute the structured condition number. As special cases of smooth manifolds, we analyze automorphism groups, and Lie and Jordan algebras associated with a scalar product. For such manifolds, we derive a lower bound on the structured condition number that is cheaper to compute than the structured condition number. We provide numerical comparisons between the structured and unstructured condition numbers for the principal matrix logarithm and principal matrix square root of matrices in automorphism groups as well as for the map between matrices in automorphism groups and their polar decomposition. We show that our lower bound can be used as a good estimate for the structured condition number when the matrix argument is well conditioned. We show that the structured and unstructured condition numbers can differ by many orders of magnitude, thus motivating the development of algorithms preserving structure.

Key words. matrix function, Fréchet derivative, condition number, bilinear form, sesquilinear form, structured matrices, structured condition number, automorphism group, Lie algebra, Jordan algebra, polar decomposition

AMS subject classifications. 15A63, 65F15, 65F30, 65F35, 65F60

DOI. 10.1137/17M1148943

1. Introduction.

Condition numbers measure the sensitivity of a problem to perturbation in the data. Let $f : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ be differentiable, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Our interest is in the sensitivity of f when perturbations are constrained to preserve structure. A general theory of conditioning was first developed by Rice [20]. The special case of matrix functions was considered by Kenney and Laub [15] and was also analyzed in detail by Higham in the monograph [10, Chap. 3]. For a matrix $X \in \mathbb{F}^{n\times n}$ such that $f(X)$ is defined in an open neighborhood $\mathcal{D} \subseteq \mathbb{F}^{n\times n}$ of X, the absolute condition number of $f(X)$ is

$$\operatorname{cond}(f, X) = \lim_{\epsilon\to 0} \sup_{\substack{\|Y-X\|\le\epsilon \\ Y\in\mathcal{D}}} \frac{\|f(Y) - f(X)\|}{\epsilon}, \tag{1.1}$$

where $\|\cdot\|$ is an arbitrary, but fixed, matrix norm. It follows from this definition that

$$\|f(Y) - f(X)\| \le \operatorname{cond}(f, X)\|Y - X\| + o(\|Y - X\|), \tag{1.2}$$

[*] Received by the editors September 25, 2017; accepted for publication (in revised form) April 8, 2019; published electronically June 25, 2019.
http://www.siam.org/journals/simax/40-2/M114894.html
Funding: The first author's research was supported by the Republic of Turkey Ministry of National Education. The second author's work was partially supported by European Research Council Advanced grant MATFUN (267526). The third author's research was supported by Engineering and Physical Sciences Research Council grant EP/I005293 and by a Royal Society-Wolfson Research Merit Award.

[†] Faculty of Engineering and Natural Sciences, Mathematics Department, Bursa Technical University, 16330 Yıldırım/Bursa, Turkey (bahar.arslan@btu.edu.tr).
[‡] Aalto University, Department of Mathematics and Systems Analysis, P.O. Box 11100, FI-00076 Aalto, Finland (vanni.noferini@aalto.fi).
[§] School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (francoise.tisseur@manchester.ac.uk).

which provides a perturbation bound for small $\|Y - X\|$.

The aim is to extend the theory of conditioning when the differentiable map f is restricted to some smooth square matrix manifold $\mathcal{M} \subset \mathbb{F}^{n\times n}$ defining the structure of $X \in \mathcal{M}$ [17] (e.g., $\mathcal{M}$ is the set of $n \times n$ unitary matrices) and to derive algorithms to compute the corresponding *structured condition numbers* defined by (see, for example, [2], [10, p. 315], or [1, Intermez. I] for a version in the relative sense)

$$\operatorname{cond}_{\mathcal{M}}(f, X) := \operatorname{cond}(f|_{\mathcal{M}}, X) = \lim_{\epsilon\to 0} \sup_{\substack{\|Y-X\|\le\epsilon \\ Y\in\mathcal{M}}} \frac{\|f(Y) - f(X)\|}{\epsilon}. \tag{1.3}$$

Now if $\mathcal{M} \subseteq \mathbb{F}^{n\times n}$ and f is also defined in an open neighborhood of X in $\mathbb{F}^{n\times n}$, then, by the definition of supremum, we have the obvious fact that

$$\operatorname{cond}_{\mathcal{M}}(f, X) \le \operatorname{cond}(f, X) \tag{1.4}$$

since the condition $Y \in \mathcal{M}$ in (1.3) restricts the perturbation $Y-X$. Whether equality holds in (1.4) is unclear a priori and depends on both f and X. If it does not hold, and particularly when the ratio between the structured and the unstructured condition numbers is much smaller than 1, we get a clear indication that using a structured numerical method to compute f would be advantageous. This argument motivates a study of structured condition numbers. This task was done by Davies in [2] for the special case where $\mathcal{M}$ is a Jordan or Lie algebra associated with a scalar product, which greatly simplifies the theory since in this case $\mathcal{M}$ is a vector subspace of $\mathbb{F}^{n\times n}$ so that the structure is linear and the perturbation $Y - X$ is in $\mathcal{M}$.

Note that for a differentiable map f with domain of definition $\mathcal{D} \subseteq \mathbb{F}^{n\times n}$, and for $X \in \mathcal{D}$ such that $f(X)$ is defined in an open neighborhood of X, Rice [20] uses the right-hand side of (1.3) with $\mathcal{D}$ in place of $\mathcal{M}$ as the definition of the (unstructured) condition number of $f(X)$, i.e., $\operatorname{cond}(f, X)$. In this paper, the wording *structured condition number* and the notation $\operatorname{cond}_{\mathcal{M}}(f, X)$ will be used when the map f is restricted to some smooth manifold $\mathcal{M}$ contained in the domain of definition of f.

The paper is organized as follows. In section 2, we extend previous results on the condition number of differentiable maps to maps that are only $\mathbb{R}$-differentiable for complex manifolds. We show that the structured condition number $\operatorname{cond}_{\mathcal{M}}(f, X)$ can be expressed as the norm of the differential of the restriction of f to $\mathcal{M}$ at X. We then present a technique and two algorithms to evaluate or estimate $\operatorname{cond}_{\mathcal{M}}(f, X)$. We show in section 3 how to apply the technique derived in section 2 when $\mathcal{M}$ is a Jordan algebra or a Lie algebra or an automorphism group associated with a scalar product—note that automorphism groups have a nonlinear structure. The structured condition number is in general expensive to compute. Hence when $\mathcal{M}$ is a Lie or Jordan algebra, or an automorphism group, we derive upper and lower bounds on $\operatorname{cond}_{\mathcal{M}}(f, X)$ that are less expensive to compute than $\operatorname{cond}_{\mathcal{M}}(f, X)$. We apply the results of section 3 to the matrix logarithm, the matrix square, and the map that associates to X the unitary polar factor of its polar decomposition, the latter map being not complex differentiable but real differentiable, a situation our theory can handle even in the presence of complex perturbations. We use the simple 2×2 matrix $X = \operatorname{diag}(\mathrm{e}^{a}, \mathrm{e}^{-a})$, $a > 0$, which is real symplectic but also complex symplectic and conjugate symplectic.[1] When f is the matrix logarithm [10], we show

[1] $X \in \mathbb{R}^{2m\times 2m}$ (or $X \in \mathbb{C}^{2m\times 2m}$) is real (or complex) symplectic if $X^T JX = J$, where $J = \left[\begin{smallmatrix} 0 & I_m \\ -I_m & 0 \end{smallmatrix}\right]$; $X \in \mathbb{C}^{2m\times 2m}$ is conjugate symplectic if $X^* JX = J$.

that $\mathrm{cond}(\log, X) = \mathrm{e}^a > 1$, but if we restrict f to $\mathcal{M}$, the real symplectic group, then $\mathrm{cond}_{\mathcal{M}}(\log, X) = a/\sinh(a) < 1$ (the same result holds when $\mathcal{M}$ is the complex symplectic group or $\mathcal{M}$ is the conjugate symplectic group). Hence we conclude that $\mathrm{cond}_{\mathcal{M}}(f, X) \ll \mathrm{cond}(f, X)$ is possible since for $X = \mathrm{diag}(\mathrm{e}^a, \mathrm{e}^{-a})$, the ratio $\mathrm{cond}_{\mathcal{M}}(\log, X)/\mathrm{cond}(\log, X) = a/(\mathrm{e}^a \sinh a)$ exponentially decays when $a \to \infty$. We show with this simple example that the lower bound on $\mathrm{cond}_{\mathcal{M}}(f, X)$ obtained in section 3 can be attained. In section 4, we illustrate through numerical experiments the quality of the lower and upper bounds derived in section 3 for the matrix logarithm and matrix square root of matrices in automorphism groups as well as for the unitary polar factor of polar decompositions. The experiments show that our lower bound on $\mathrm{cond}_{\mathcal{M}}(f, X)$ tends to be much sharper than our upper bound, and the former, when combined with a backward error, provides a good approximation to the forward error $\|f(Y) - f(X)\|$.

2. Structured and unstructured condition numbers of matrix functions.

We start with a brief summary of the theory and algorithms for the unstructured condition number of a matrix function before discussing the structured case. Because the relative condition number of f at an $n \times n$ matrix X, denoted by $\mathrm{rcond}(f, X)$, can be written in terms of the absolute condition number $\mathrm{cond}(f, X)$ as

$$\mathrm{rcond}(f, X) = \mathrm{cond}(f, X) \cdot \frac{\|X\|}{\|f(X)\|}$$

(see [10, Chap. 3]), we just concentrate on absolute condition numbers. Here and throughout, $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

2.1. Unstructured condition number.

Let f be a differentiable endofunction of $\mathbb{F}^{n\times n}$. The unstructured condition number $\mathrm{cond}(f, X)$ in (1.1) can be expressed in terms of the Fréchet derivative of f at X, which is an $\mathbb{F}$-linear map $L_f(X, \cdot) : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ such that

$$\|f(X + E) - f(X) - L_f(X, E)\| = o(\|E\|) \tag{2.1}$$

for all $E \in \mathbb{F}^{n\times n}$. When the Fréchet derivative of f at X exists, it is unique. In that case, we have [10, Thm. 3.1]

$$\mathrm{cond}(f, X) = \max_{E\neq 0} \frac{\|L_f(X, E)\|}{\|E\|} =: \|L_f(X)\|.$$

An explicit formula for the Fréchet derivative is not always available. So we assume that we have a numerical method to evaluate $L_f(X, E)$ for a given E. Since the Fréchet derivative is linear in E, applying the vec operator, which stacks the columns of a matrix into one long vector [16], to $L_f(X, E)$ gives

$$\mathrm{vec}(L_f(X, E)) = K_f(X)\mathrm{vec}(E), \tag{2.2}$$

where $K_f(X) \in \mathbb{F}^{n^2\times n^2}$ is called the *Kronecker form of the Fréchet derivative.* When the canonical bases on $\mathbb{F}^{n^2}$ and $\mathbb{F}^{n\times n}$ are chosen (as we will implicitly do throughout the paper), then when we identify $\mathbb{F}^{n\times n}$ with $\mathbb{F}^{n^2}$ through the mapping represented by vec, the matrix $K_f(X)$ represents the derivative at X of the function f. If we specialize to the Frobenius norm, then

$$\mathrm{cond}(f, X) = \max_{E\neq 0} \frac{\|L_f(X, E)\|_F}{\|E\|_F} = \max_{E\neq 0} \frac{\|\mathrm{vec}(L_f(X, E))\|_2}{\|\mathrm{vec}(E)\|_2} = \|K_f(X)\|_2, \tag{2.3}$$

where we use the fact that for $A \in \mathbb{F}^{n\times n}$, $\|A\|_F = \|\text{vec}(A)\|_2$. The problem of computing $\text{cond}(f, X)$ in the Frobenius norm then reduces to finding the 2-norm of $K_f(X)$. Note that the latter matrix can be constructed explicitly by forming one column at a time using (2.2), that is,

$$K_f(X)e_{i+(j-1)n} = \text{vec}(L_f(X, e_i e_j^T)), \qquad i = 1\colon n, \quad j = 1\colon n, \tag{2.4}$$

where e_k is a vector of appropriate dimension with the kth entry equal to one and zero everywhere else. Constructing $K_f(X)$ this way costs $O(n^5)$ operations, assuming that $L_f(X, E)$ can be computed in $O(n^3)$ operations. Note that, based on (2.1), $L_f(X, E)$ can be approximated by finite differences,

$$L_f(X, E) \approx \frac{f(X + tE) - f(X)}{t}$$

for a small value of t. We refer the reader to [10, Chap. 3] on how to choose t. When f is a matrix function in the linear algebra sense (see [10, Chap.1] for a precise definition), the formula

$$f\left(\begin{bmatrix} X & E \\ 0 & X \end{bmatrix}\right) = \begin{bmatrix} f(X) & L_f(X, E) \\ 0 & f(X) \end{bmatrix}$$

also holds [10] and yields a useful tool to compute Fréchet derivatives. However, in this paper we consider a much wider class of functions f, and hence this expression is not always true.

A lower bound for $\text{cond}(f, X)$ can be computed by the power method applied to $K_f(X)^* K_f(X)$, which, for a nonzero matrix $E_0 \in \mathbb{F}^{n\times n}$, constructs the iterates [15]

$$Z_{k+1} = L_f(X, E_k), \quad E_{k+1} = L_f^{\text{adj}}(X, Z_{k+1}), \quad \gamma_{k+1} = \frac{\|E_{k+1}\|_F}{\|Z_{k+1}\|_F}, \quad k > 0, \tag{2.5}$$

with γ_k such that $\gamma_k \leq \|L_f(X)\|_F$ and $\gamma_k \to \|L_f(X)\|_F$ as $k \to \infty$ [5, Chap. 7]. In (2.5), L_f^{adj} is the adjoint of L_f and is given by

$$L_f^{\text{adj}}(X, E) = \begin{cases} L_f(X^T, E) \text{ when } \mathbb{F} = \mathbb{R}, \\ L_{\bar{f}}(X^*, E) \text{ when } \mathbb{F} = \mathbb{C}, \end{cases}$$

where $\bar{f}(z) := \overline{f(\bar{z})}$. The computation of the kth iteration in (2.5) costs $O(n^3)$ operations, and just a few iterations are usually needed for an accurate bound.

Having fixed a convenient map to express the isomorphism between $n \times n$ complex matrices and real vectors with $2n^2$ entries, the Fréchet derivative can then be represented in the canonical basis as a $2n^2 \times 2n^2$ real matrix $K_f^{(\mathbb{R})}(X)$ such that

$$\text{vec}(\rho(L_f(X, E))) = K_f^{(\mathbb{R})}(X)\text{vec}(\rho(E)), \tag{2.6}$$

where

$$\rho : \mathbb{C}^{n\times n} \to \mathbb{R}^{n\times 2n}, \quad \rho(E) = \begin{bmatrix} \text{Re}(E) & \text{Im}(E) \end{bmatrix}. \tag{2.7}$$

Now when $\mathbb{F} = \mathbb{C}$, it also makes sense to consider maps $f : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ that satisfy the less demanding assumption of real Fréchet differentiability, that is, $L_f(X, E)$ in (2.1) is real linear (i.e, $L_f(X, E+\alpha F) = L_f(X, E)+\alpha L_f(X, F)$ for all $\alpha \in \mathbb{R}$; see also

[19, sect. 2.3]) but not necessarily complex linear. In this case, (2.2) does not hold but (2.6) does, so that

$$\mathrm{cond}(f,X) = \max_{E\neq 0} \frac{\|L_f(X,E)\|_F}{\|E\|_F} = \max_{E\neq 0} \frac{\|\mathrm{vec}\big(\rho(L_f(X,E))\big)\|_2}{\|\mathrm{vec}(\rho(E))\|_2} = \|K_f^{(\mathbb{R})}(X)\|_2.$$

Again, $K_f^{(\mathbb{R})}(X)$ can be explicitly formed by computing its $2n^2$ columns as

$$\mathrm{vec}(\rho(L_f(X,e_ie_j^T))), \quad \mathrm{vec}(\rho(L_f(X,\mathrm{i}e_ie_j^T))), \quad i=1\colon n, \quad j=1\colon n,$$

where $\mathrm{i}=\sqrt{-1}$ denotes the imaginary unit.

If f is not only real but also complex Fréchet differentiable, then, by the Cauchy–Riemann equations, $K_f^{(\mathbb{R})}$ has the form (we henceforth occasionally omit the dependence on X for notational simplicity)

$$K_f^{(\mathbb{R})} = \begin{bmatrix} R_f & -P_f \\ P_f & R_f \end{bmatrix}, \quad R_f, P_f \in \mathbb{R}^{n^2\times n^2}. \tag{2.8}$$

Let unvec be the inverse of the vec operator, which in this case is defined from $\mathbb{F}^{2n^2}$ to $\mathbb{F}^{n\times 2n}$. After applying $\mathrm{vec}\circ\rho^{-1}\circ\mathrm{unvec}$ to (2.6), the Cauchy–Riemann equation (2.8) implies that

$$\mathrm{vec}(L_f(X,E)) = (R_f + \mathrm{i}P_f)\mathrm{vec}(E) =: K_f^{(\mathbb{C})}\mathrm{vec}(E),$$

where $K_f^{(\mathbb{C})} \equiv K_f$ as in (2.2). Observe that the unitary matrix $Q = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 & \mathrm{i} \\ \mathrm{i} & 1\end{smallmatrix}\right] \otimes I_{n^2}$, where $\otimes$ denotes the Kronecker product, block diagonalizes $K_f^{(\mathbb{R})}$, i.e.,

$$Q^* K_f^{(\mathbb{R})} Q = \begin{bmatrix} K_f^{(\mathbb{C})} & 0 \\ 0 & \overline{K}_f^{(\mathbb{C})} \end{bmatrix}, \tag{2.9}$$

so that $\|K_f^{(\mathbb{C})}\|_2 = \|K_f^{(\mathbb{R})}\|_2$. This shows that the theory is coherent in the sense that the computation of the unstructured condition number is independent of whether the real or complex Kronecker form of the Fréchet derivative is considered.

Example 2.1. Let $X \in \mathbb{C}^{n\times n}$ be nonsingular and let $f : X \mapsto U$ associate to X the unitary factor U of its polar decomposition $X = UH$. The map f is real differentiable but not complex differentiable [19]. Now for $n=1$ and $z = x + \mathrm{i}y \neq 0$, $f(z) = \frac{z}{|z|} = f_1(x,y) + \mathrm{i}f_2(x,y)$ with $f_1(x,y) = x/(x^2+y^2)^{1/2}$, $f_1(x,y) = y/(x^2+y^2)^{1/2}$, and f has for the real Fréchet derivative in Kronecker form the matrix

$$K_f^{(\mathbb{R})}(z) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix} = |z|^{-3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}.$$

$K_f^{(\mathbb{R})}$ is not of the form (2.8), revealing that f is not complex Fréchet differentiable.

2.2. Structured condition number. Suppose now that $g : \mathcal{M} \to \mathcal{N}$ is a differentiable map, where $\mathcal{M},\mathcal{N} \subseteq \mathbb{F}^{n\times n}$ are smooth square matrix manifolds. We consider in particular three classes of smooth manifolds:

1. real submanifolds of the n^2-dimensional real vector space $\mathbb{R}^{n\times n}$ (e.g., orthogonal matrices, real symplectic matrices),

2. complex submanifolds of the n^2-dimensional complex vector space $\mathbb{C}^{n\times n}$ (e.g., complex orthogonal matrices, complex symplectic matrices), and
3. real submanifolds of the $2n^2$-dimensional *real* vector space $\mathbb{C}^{n\times n}$ (e.g., Hermitian matrices, unitary matrices, symplectic matrices).

These submanifolds are often associated with, respectively, real bilinear forms, complex bilinear forms, and sesquilinear forms. We use the expression "square matrix manifold" to mean any of the three classes of submanifolds of square matrices described above. The choice of considering square matrices is for the sake of simplicity in the exposition; we emphasize, though, that our theory can be easily extended to rectangular matrices.

We need to distinguish between the field in which the entries of the matrices are allowed to lie (i.e., the *ambient field*), denoted by $\mathbb{F}$ as in the previous sections, and the field on which the ambient vector space is built (i.e., the *base field*), which we instead denote by $\mathbb{K}$. We will also assume that $\mathbb{K}$ is a (possibly not proper) subfield of $\mathbb{F}$. To be more concrete, $\mathbb{K} = \mathbb{F}$ ($= \mathbb{R}$ or $\mathbb{C}$, respectively) for either case 1 or case 2 above, while case 3 of a real submanifold of $\mathbb{C}^{n\times n}$ is special in the sense that $\mathbb{C} = \mathbb{F} \neq \mathbb{K} = \mathbb{R}$. We will need to deal with case 2 carefully when g is only $\mathbb{R}$-differentiable since in this case it will be necessary to identify $\mathcal{M}$ with $\rho(\mathcal{M}) \subseteq \mathbb{R}^{n\times 2n}$ (see Remark 2.2 below).

For a smooth manifold $\mathcal{M} \subseteq \mathbb{F}^{n\times n}$ with base field $\mathbb{K}$, the matrix $E \in \mathbb{F}^{n\times n}$ is called a *tangent vector* of $\mathcal{M}$ at $X \in \mathcal{M}$ if there is a smooth curve $\gamma : \mathbb{K} \to \mathcal{M}$ such that $\gamma(0) = X$, $\gamma'(0) = E$. The set

$$T_X\mathcal{M} := \{E \in \mathbb{F}^{n\times n} \mid \exists\, \gamma : \mathbb{K} \to \mathcal{M} \text{ smooth with } \gamma(0) = X,\ \gamma'(0) = E\} \tag{2.10}$$

of tangent vectors of $\mathcal{M}$ at X is called the *tangent space* of $\mathcal{M}$ at X. It is a $\mathbb{K}$-linear subspace of $\mathbb{F}^{n\times n}$, in which it can be embedded, thus inheriting any usual matrix norm.[2] Note that, when $\mathbb{C} = \mathbb{F} \neq \mathbb{K} = \mathbb{R}$, $T_X\mathcal{M}$ is a real subspace but generally not a complex subspace. For example, take the unit sphere $\mathcal{M} = \{X \in \mathbb{C}^{n\times n} \mid \|X\|_F = 1\}$; then $T_X\mathcal{M} = \{E \in \mathbb{C}^{n\times n} \mid \operatorname{trace}(E^*X + X^*E) = 0\}$ has real codimension 1, and hence it clearly is not a complex subspace.

Consider a smooth curve $\gamma : \mathbb{K} \to \mathcal{U}$ such that $\gamma(0) = X$, where $\mathcal{U} \subset \mathcal{M}$ is an open set and $\mathcal{M} \subseteq \mathbb{F}^{n\times n}$ has base field $\mathbb{K}$. If $g : \mathcal{M} \to \mathcal{N}$ is $\mathbb{K}$-differentiable, then the differential of g at the point X is the map

$$dg_X : T_X\mathcal{M} \to T_{g(X)}\mathcal{N}, \quad dg_X(\gamma'(0)) = (g \circ \gamma)'(0). \tag{2.11}$$

Clearly, dg_X in (2.11) is a $\mathbb{K}$-linear operator [17, Exerc. 3.7], and it comes equipped with an induced norm,

$$\|dg_X\| := \max_{\substack{E \in T_X\mathcal{M}\\ E \neq 0}} \frac{\|dg_X(E)\|}{\|E\|}.$$

Remark 2.2. When $\mathcal{M} \subseteq \mathbb{C}^{n\times n}$ is a complex submanifold, there is freedom in the choice of the base field. Indeed, any complex submanifold is also a real submanifold, although the converse does not always hold. If g is only real differentiable, then it is necessary to pick $\mathbb{K} = \mathbb{R}$, thus (implicitly) identifying the complex submanifold $\mathcal{M}$ with the real submanifold $\rho(\mathcal{M}) =: \mathcal{M}_{\mathbb{R}} \subseteq \mathbb{R}^{n\times 2n}$, which has base field $\mathbb{R}$ and is such that $X \in \mathcal{M}$ if and only if $\rho(X) = [\operatorname{Re}(X)\ \ \operatorname{Im}(X)] \in \mathcal{M}_{\mathbb{R}}$.

[2] We note that other choices are possible, such as, for example, the intrinsic norm that makes $\mathcal{M}$ a path metric space. Different choices for the notion of distance on the manifold would, of course, lead to a different value of the condition number and hence a slightly different theory. Which choice is in some sense the "best" depends on the context. Our goal is to compare structured and unstructured condition numbers, and hence the induced Euclidean distance appears to be the most natural.

Table 1

Explicit expressions for the structured condition number and its cheaply computable lower and upper bounds, depending on the ambient and base fields and the differentiability of the map f.

Ambient field ($\mathbb{F}$)	Base field ($\mathbb{K}$)	Type of differentiability	$\mathrm{cond}_{\mathcal{M}}(f,X)$	Cheap bounds
$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C}$	Eq. (2.16)	Eq. (2.15)
$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	Eq. (2.19)	Eq. (2.18)
$\mathbb{C}$	$\mathbb{R}$	$\mathbb{R}$	Eq. (2.19)	Eq. (2.18)
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	Eq. (2.16)	Eq. (2.15)

A version of the following theorem appears in [3, pp. 4–5] and [20, Thms. 3 and 4]. Note, however, that our setting is more general in the sense that real submanifolds of complex matrices (the case $\mathbb{K} \neq \mathbb{F}$) are not analyzed in [3] or [20].

Theorem 2.3. *Let $g : \mathcal{U} \to \mathcal{V}$ be a $\mathbb{K}$-differentiable map between two open subsets $\mathcal{U} \subset \mathcal{M}$ and $\mathcal{V} \subset \mathcal{N}$, where $\mathcal{M}$ and $\mathcal{N}$ are smooth square matrix manifolds of $\mathbb{F}^{n\times n}$ with base field $\mathbb{K}$. Then for $X \in \mathcal{U}$ it holds that*

$$\lim_{\epsilon\to 0} \sup_{\substack{\|Y-X\|\leq\epsilon \\ Y\in\mathcal{M}}} \frac{\|g(Y)-g(X)\|}{\epsilon} = \|dg_X\|.$$

Proof. Suppose $X, Y \in \mathcal{U}$ and let $\gamma : \mathbb{K} \to \mathcal{U}$ be any curve such that $\gamma(0) = X$, $\gamma(\epsilon) = Y$. Letting $E = \gamma'(0) \in T_X\mathcal{U}$, we get $Y = X + \epsilon E + o(\epsilon)$ by using the definition of derivative along a curve, and by using the definition of differential we get $g(Y) = g(X) + \epsilon dg_X(E) + o(\epsilon)$. Then

$$\begin{aligned}\lim_{\epsilon\to 0} \sup_{\substack{\|Y-X\|\leq\epsilon \\ Y\in\mathcal{M}}} \frac{\|g(Y)-g(X)\|}{\epsilon} &= \lim_{\|X-Y\|\to 0} \sup_{Y\in\mathcal{M}} \frac{\|g(Y)-g(X)\|}{\|Y-X\|} \\ &= \lim_{\epsilon\to 0} \sup_{\substack{E\in\ T_X\mathcal{U} \\ E\neq 0}} \frac{\epsilon\|dg_X(E)\|+o(\epsilon)}{\epsilon\|E\|+o(\epsilon)} = \|dg_X\|. \qquad \square\end{aligned}$$

Now if $f|_{\mathcal{M}}$ denotes the restriction to the manifold $\mathcal{M}$ of the map $f : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$, then, by Theorem 2.3 applied to $g = f|_{\mathcal{M}}$, it follows that for the structured condition number in (1.3), we have that

$$\mathrm{cond}_{\mathcal{M}}(f,X) = \|d(f|_{\mathcal{M}})_X\|. \tag{2.12}$$

We now give more details about how the structured condition number (2.12) can be computed, distinguishing between the possible situations depending on $\mathbb{F}$, $\mathbb{K}$, and the differentiability properties of f, as summarized in Table 1. (Note that, when f is only real differentiable, we always take $\mathbb{K} = \mathbb{R}$ as the base field.)

2.2.1. When the ambient and base fields are the same. Suppose that $\mathbb{F} = \mathbb{K}$ and that the map $f : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is $\mathbb{K}$-Fréchet differentiable (and so is the restriction of f to the manifold $\mathcal{M}$). The uniqueness of the differential and the uniqueness of the Fréchet derivative imply that for any $E \in T_X\mathcal{M}$,

$$d(f|_{\mathcal{M}})_X(E) = L_f(X,E). \tag{2.13}$$

Since $T_X\mathcal{M}$ is a $\mathbb{K}$-linear subspace of $\mathbb{F}^{n\times n}$, the linear nature of $E \in T_X\mathcal{M}$ is then encoded by

$$\mathrm{vec}(E) = By, \tag{2.14}$$

where $B \in \mathbb{F}^{n^2 \times p}$ is a matrix of full rank giving (in essence) a basis for $T_X\mathcal{M}$ and $y \in \mathbb{K}^p$ with $p = \dim_{\mathbb{K}} T_X\mathcal{M}$ being a vector of parameters; see section 3 for examples of how to construct B. Applying the vec operator to (2.13) and using (2.2), (2.14) yields (recalling that here and throughout we are implicitly picking the canonical bases on both $\mathbb{K}^{n\times n}$ and $\mathbb{K}^n$)

$$\mathrm{vec}(d(f|_{\mathcal{M}})_X(E)) = \mathrm{vec}(L_f(X,E)) = K_f(X)\mathrm{vec}(E) = K_f(X)By = K_f(X)BB^+By,$$

where B^+ denotes the Moore–Penrose pseudoinverse of B. Hence, from (2.12) and using the Frobenius norm, we find that

$$\mathrm{cond}_{\mathcal{M}}(f,X) = \max_{\substack{E\in T_X\mathcal{M}\\ E\neq 0}} \frac{\|L_f(X,E)\|_F}{\|E\|_F} = \max_{\substack{y\in\mathbb{K}^p\\ y\neq 0}} \frac{\|K_f(X)By\|_2}{\|By\|_2} = \|K_f(X)BB^+\|_2.$$

A similar equation holds for the structured condition number in any entrywise p-norm, $\|X\|_p := (\sum_{ij}|x_{ij}|^p)^{1/p}$. However, for concreteness, we only report results in the Frobenius norm ($p = 2$). Observe now that $\|K_f(X)B\|_2\|B\|_2^{-1} \le \|K_f(X)BB^+\|_2 \le \|K_f(X)B\|_2\|B^+\|_2$. Hence,

$$\|K_f(X)B\|_2\|B\|_2^{-1} \le \mathrm{cond}_{\mathcal{M}}(f,X) \le \|K_f(X)B\|_2\|B^+\|_2. \tag{2.15}$$

Moreover, if B can be chosen to have orthonormal columns, i.e., $B^*B = I_p$, then the lower and upper bounds in (2.15) are equal so that

$$\mathrm{cond}_{\mathcal{M}}(f,X) = \|K_f(X)B\|_2. \tag{2.16}$$

Remark 2.4. Since $\|B\|_2 \le \|K_f(X)^{-1}\|_2\|K_f(X)B\|_2$,

$$\mathrm{cond}(f,X) \le \frac{\kappa_2\big(K_f(X)\big)}{\kappa_2(B)}\|K_f(X)B\|_2\|B^+\|_2,$$

where $\kappa_2\big(K_f(X)\big) = \|K_f(X)\|_2\|K_f(X)^{-1}\|_2$ and $\kappa_2(B) = \|B\|_2\|B^+\|_2$ denote the matrix 2-norm condition numbers of $K_f(X)$ and B, respectively. So when $\kappa_2(B) \ge \kappa_2\big(K_f(X)\big)$, the upper bound in (2.15) is loose and we should use the unstructured condition number $\mathrm{cond}(f,X)$ as an upper bound for $\mathrm{cond}_{\mathcal{M}}(f,X)$.

To compute $\|K_f(X)B\|_2$, we need to characterize $T_X\mathcal{M}$, find its dimension $p := \dim_{\mathbb{K}} T_X\mathcal{M}$ over the base field $\mathbb{K}$, and then construct a matrix $B \in \mathbb{F}^{n^2\times p}$ such that for any $E \in T_X\mathcal{M}$, $\mathrm{vec}(E) = By$ for some $y \in \mathbb{K}^p$. If we assume that we have a numerical method to compute $L_f(X,E)$ for a given $X \in \mathcal{M}$ and $E \in T_X\mathcal{M}$, then we can compute the columns of $K_f(X)B$ using

$$K_f(X)Be_i = \mathrm{vec}\big(L_f(X, \mathrm{unvec}(Be_i))\big), \quad i = 1\colon p, \tag{2.17}$$

where the inverse vec operator unvec is in this case defined from $\mathbb{F}^{n^2}$ to $\mathbb{F}^{n\times n}$. Note that (2.17) reduces to (2.4) when $\mathbb{K} = \mathbb{F}$ and $X \in \mathcal{M} = \mathbb{F}^{n\times n} = T_X\mathcal{M}$ so that $B = I_{n^2}$ in (2.14).

2.2.2. When the ambient and base fields differ. Suppose now that $\mathbb{C} = \mathbb{F} \neq \mathbb{K} = \mathbb{R}$ and that the map f is $\mathbb{R}$-Fréchet differentiable. It follows from (2.14) that for $X \in \mathcal{M}$ and $E \in T_X\mathcal{M}$, we can write $\mathrm{vec}(E) = By$ for some $B \in \mathbb{C}^{n^2\times p}$ and

$y \in \mathbb{R}^p$, where $p = \dim_{\mathbb{R}} T_X\mathcal{M}$. Using the map ρ defined in (2.7), we have that, for $\rho(X) \in \mathcal{M}_{\mathbb{R}} \equiv \rho(\mathcal{M})$ and $\rho(E) \in T_{\rho(X)}\mathcal{M}_{\mathbb{R}}$, we can write $\operatorname{vec}(\rho(E)) = B_{\mathbb{R}}y$ with

$$B_{\mathbb{R}} = \begin{bmatrix} \operatorname{Re} B \\ \operatorname{Im} B \end{bmatrix} \in \mathbb{R}^{2n^2 \times p}.$$

Applying ρ and the vec operator to (2.13) yields

$$\operatorname{vec}\big(\rho(d(f|_{\mathcal{M}})_X(E))\big) = \operatorname{vec}\big(\rho(L_f(X,E))\big) = K_f^{(\mathbb{R})}(X)\operatorname{vec}\big(\rho(E)\big) = K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}y,$$

and since $B_{\mathbb{R}} = B_{\mathbb{R}}B_{\mathbb{R}}^+B_{\mathbb{R}}$, we have that by (2.12)

$$\begin{aligned}\operatorname{cond}_{\mathcal{M}}(f,X) &= \max_{\substack{E\in T_X\mathcal{M}\\ E\neq 0}} \frac{\|L_f(X,E)\|_F}{\|E\|_F} \\ &= \max_{\substack{E\in T_X\mathcal{M}\\ E\neq 0}} \frac{\|\rho(L_f(X,E))\|_F}{\|\operatorname{vec}(\rho(E))\|_2} \\ &= \max_{\substack{y\in\mathbb{R}^p\\ y\neq 0}} \frac{\|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}y\|_2}{\|B_{\mathbb{R}}y\|_2} \\ &= \|K_f^{(\mathbb{R})}B_{\mathbb{R}}B_{\mathbb{R}}^+\|_2.\end{aligned}$$

The lower and upper bounds on $\operatorname{cond}_{\mathcal{M}}(f,X)$ follow:

$$\|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}\|_2\|B_{\mathbb{R}}\|_2^{-1} \le \operatorname{cond}_{\mathcal{M}}(f,X) \le \|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}\|_2\|B_{\mathbb{R}}^+\|_2. \tag{2.18}$$

Now if $B_{\mathbb{R}}$ has orthonormal columns, then

$$\operatorname{cond}_{\mathcal{M}}(f,X) = \|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}\|_2. \tag{2.19}$$

Remark 2.5. If $\mathbb{C} = \mathbb{F} \neq \mathbb{K} = \mathbb{R}$, but f is also $\mathbb{C}$-differentiable, then both $K_f^{(\mathbb{R})}$ and $K_f^{(\mathbb{C})}$ exist. It is easy to show that

$$\operatorname{cond}_{\mathcal{M}}(f,X) = \|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}\|_2 \le \|K_f^{(\mathbb{C})}(X)B\|_2. \tag{2.20}$$

Depending on the manifold and the map, a strict inequality may hold (see Example 3.3). This shows that, when $\mathcal{M}$ is not a complex submanifold and even for holomorphic maps f, only the real Fréchet derivatives are linked to the structured condition number.

2.2.3. Computation of $\operatorname{cond}_{\mathcal{M}}(f,X)$.

The construction of $K_f(X)B$ in (2.17) extends almost trivially to the construction of $K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}$ and yields the following algorithm.

ALGORITHM 2.6. *Given*

(i) *any algorithm to compute the Fréchet derivative of* $f : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$,

(ii) $X \in \mathcal{M}$, *where* $\mathcal{M}$ *is a smooth manifold of* $\mathbb{F}^{n\times n}$ *with base field* $\mathbb{K}$, *and*

(iii) *either* $B \in \mathbb{F}^{n^2\times p}$ *such that for any* $E \in T_X\mathcal{M}$, $\operatorname{vec}(E) = By$ *for some* $y \in \mathbb{K}^p$ *if* $\mathbb{K} = \mathbb{F}$, *or* $B_{\mathbb{R}} \in \mathbb{F}^{2n^2\times p}$ *such that for any* $E \in T_X\mathcal{M}$, $\operatorname{vec}([\operatorname{Re}(E), \operatorname{Im}(E)]) = B_{\mathbb{R}}y$ *for some* $y \in \mathbb{K}^p$ *if* $\mathbb{K} \neq \mathbb{F}$, *this algorithm computes* $\kappa = \|K_f(X)B\|_2$ *if* $\mathbb{K} = \mathbb{F}$ *or* $\kappa = \|K_f^{(\mathbb{R})}(X)B_{\mathbb{R}}\|_2$ *otherwise. If* $B^*B = I_p$ *for* $\mathbb{K} = \mathbb{F}$ *or* $B_{\mathbb{R}}^T B_{\mathbb{R}} = I_p$ *otherwise, then* $\kappa = \mathrm{cond}_{\mathcal{M}}(f, X)$.

```
1  If $\mathbb{K} = \mathbb{F}$
2      $K = 0_{n^2 \times p}$
3      for $i = 1\colon p$
4          Compute $F = L_f(X, E)$, where $\mathrm{vec}(E) = Be_i$,
5          $Ke_i = \mathrm{vec}(F)$
6      end
7  else
8      $K = 0_{2n^2 \times p}$
9      for $i = 1\colon p$
10         Compute $F = L_f(X, E)$, where $\mathrm{vec}\big([\mathrm{Re}(E), \mathrm{Im}(E)]\big) = B_{\mathbb{R}} e_i$,
11         $Ke_i = \mathrm{vec}\big([\mathrm{Re}(F), \mathrm{Im}(F)]\big)$
12     end
13 $\kappa = \|K\|_2$
```

If $\mathbb{K} = \mathbb{F}$, the construction of $K = K_f(X)B$ in Algorithm 2.6 costs $O(pn^3)$ operations, assuming that $L_f(X, E)$ can be computed in $O(n^3)$ operations, and the cost of computing the 2-norm of K in step 6 is $O(p^2n^2)$, and so Algorithm 2.6 costs $O(pn^3 + p^2n^2)$ operations, which is high in particular when $p = O(n^2)$. As in the unstructured case, once B is known, we can use the power method to obtain a lower bound for $\|K_f(X)B\|_2$, which corresponds to a lower bound for $\mathrm{cond}_{\mathcal{M}}(f, X)$ when B has orthonormal columns.

Analogous remarks hold for $\mathbb{K} \neq \mathbb{F}$, except that the factor 2 in the sizes of $K_f^{(\mathbb{R})}$ and $B_{\mathbb{R}}$ leads to higher constants in front of the asymptotic complexities.

ALGORITHM 2.7. *Given the same input as Algorithm* 2.6, *this algorithm uses the power method to compute* γ *such that* $\gamma \le \|K_f(X)B\|_2$ *for* $\mathbb{K} = \mathbb{F}$ *or* $\gamma \le \|K_f^{(\mathbb{R})} B_{\mathbb{R}}\|_2$ *otherwise.*

```
1  Choose a nonzero starting vector $z_0 \in \mathbb{K}^p$.
2  for $k = 0\colon \infty$
3      if $\mathbb{K} = \mathbb{F}$
4          $\mathrm{vec}(E_k) = Bz_k$
5      else
6          $\mathrm{vec}\big([\mathrm{Re}(E_k), \mathrm{Im}(E_k)]\big) = B_{\mathbb{R}} z_k$
7      end
8      $W_{k+1} = L_f(X, E_k)$
9      $Y_{k+1} = L_f^{\mathrm{adj}}(X, W_{k+1})$
10     if $\mathbb{K} = \mathbb{F}$
11         $z_{k+1} = B^* \mathrm{vec}(Y_{k+1})$
12     else
13         $z_{k+1} = B_{\mathbb{R}}^T \mathrm{vec}\big(\,\mathrm{Re}(Y_{k+1}), \mathrm{Im}(Y_{k+1})]\big)$
14     end
15     $\gamma_{k+1} = \|z_{k+1}\|_2 / \|W_{k+1}\|_F$
16     if converged, $\gamma = \gamma_{k+1}$, quit, end
17 end
```

Unless B (or $B_{\mathbb{R}}$) has a special structure that can be exploited in steps 4 and 11 (or steps 6 and 13), the cost of Algorithm 2.7 is $O(kpn^2)$ operations, where k is the number of iterations and we assume that $p \ge n$ and $L_f(X, E_k)$ and $L_f^{\mathrm{adj}}(X, W_{k+1})$ can be computed in $O(n^3)$ operations.

3. Application: Matrix manifolds associated with scalar products.

We illustrate the technique presented in section 2.2 on smooth square matrix manifolds $\mathcal{M}$ associated with a scalar product on $\mathbb{F}^n$, that is, a nondegenerate bilinear or sesquilinear form defined by any nonsingular matrix M by, for $x, y \in \mathbb{F}^n$,

$$\langle x, y\rangle_{\mathrm{M}} = \begin{cases} x^T My & \text{for real or complex bilinear forms,} \\ x^*My & \text{for sesquilinear forms.} \end{cases}$$

For any matrix $A \in \mathbb{F}^{n\times n}$, there exists a unique $A^\star \in \mathbb{F}^{n\times n}$ called the adjoint of A with respect to $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ and given by

$$A^\star = \begin{cases} M^{-1}A^TM & \text{for real or complex bilinear forms,} \\ M^{-1}A^*M & \text{for sesquilinear forms.} \end{cases}$$

There are three classes of structured matrices associated with $\langle\cdot,\cdot\rangle_{\mathrm{M}}$: a Jordan algebra[3] $\mathbb{J}_M$, a Lie algebra $\mathbb{L}_M$, and an automorphism group $\mathbb{G}_M$ defined by

$$\mathbb{J}_M := \{A \in \mathbb{F}^{n\times n} \mid A^\star = A\}, \qquad \mathbb{L}_M := \{A \in \mathbb{F}^{n\times n} \mid A^\star = -A\},$$
$$\mathbb{G}_M := \{A \in \mathbb{F}^{n\times n} \mid A^\star = A^{-1}\}.$$

For several special choices of M, a specific nomenclature exists to indicate these sets. For example, for real bilinear forms ($\mathbb{F} = \mathbb{R}$) and

- $M = I$, $\mathbb{J}_I, \mathbb{L}_I$, and $\mathbb{G}_I$ are the set of symmetric, skew-symmetric, and orthogonal matrices, respectively;
- $M = S_{p,q} = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_q \end{smallmatrix}\right]$ with $p + q = n$, $\mathbb{J}_{S_{p,q}}, \mathbb{L}_{S_{p,q}}$, and $\mathbb{G}_{S_{p,q}}$ correspond to the class of pseudosymmetric, pseudoskew-symmetric, and pseudo-orthogonal matrices, respectively;
- $M = J := \left[\begin{smallmatrix} 0 & I_{n/2} \\ -I_{n/2} & 0 \end{smallmatrix}\right]$ with n even, $\mathbb{J}_J, \mathbb{L}_J$, and $\mathbb{G}_J$ correspond to the class of skew-Hamiltonian, Hamiltonian, and symplectic matrices, respectively;
- $M = R = \begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix}$, $\mathbb{J}_R, \mathbb{L}_R$, and $\mathbb{G}_R$ correspond to the class of persymmetric, perskew-symmetric, and perplectic matrices, respectively.

When M defines a real (respectively, complex) bilinear form, the three matrix classes defined above are real (respectively, complex) smooth manifolds; when M defines a sesquilinear form, they are real smooth submanifolds of the $2n$-dimensional real vector space $\mathbb{C}^{n\times n}$. This is immediate for $\mathbb{J}_M$ and $\mathbb{L}_M$, as they are $\mathbb{K}$-linear subspaces, and hence they are (flat) smooth manifolds. Automorphism groups are not vector subspaces, but they are known to be smooth manifolds, as we now show.

THEOREM 3.1. *The automorphism group $\mathbb{G}_M$ is a real submanifold of $\mathbb{F}^{n\times n}$. Furthermore, when M defines a complex bilinear form, $\mathbb{G}_M$ is also a complex submanifold of $\mathbb{C}^{n\times n}$.*

Proof. The first part of the statement is [21, Thm. 7.17]. For the second part, note that $\mathbb{G}_M$ is the set of solutions of the quadratic matrix equation $M = X^TMX$. Since the latter is equivalent to either n^2 complex polynomial equations or $2n^2$ real polynomial equations, it follows that $\mathbb{G}_M$ is both a complex algebraic variety and a real algebraic variety. Recall that a complex (respectively, real) algebraic variety is a complex (respectively, real) manifold if and only if it does not contain singular points,

[3] The name comes from the fact that $X, Y \in \mathbb{J}_M \Rightarrow XY + YX \in \mathbb{J}_M$, so that $\mathbb{J}_M$ is a commutative ring when endowed with the usual addition and this "symmetrized multiplication."

i.e., points such that the rank of the Jacobian is locally smaller than the dimension of the variety cut out by the aforementioned polynomial equations. It now suffices to observe that (in the canonical bases of $\mathbb{C}^{n\times n}$ as a complex and a real vector space, respectively), by the Cauchy–Riemann equations, $\mathcal{J} = \operatorname{Re}\mathcal{J} + \mathrm{i}\operatorname{Im}\mathcal{J}$ is the Jacobian of $\mathbb{G}_M$ as a complex algebraic variety if and only if $\mathcal{J}_{\mathbb{C}} = \left[\begin{smallmatrix} \operatorname{Re}\mathcal{J} & \operatorname{Im}\mathcal{J} \\ -\operatorname{Im}\mathcal{J} & \operatorname{Re}\mathcal{J}\end{smallmatrix}\right]$ is the Jacobian of $\mathbb{G}_M$ as a real algebraic variety. The statement then follows by observing that $\mathcal{J}_{\mathbb{C}}$ is unitarily similar to $\mathcal{J} \oplus \mathcal{J}^*$. ◻

Now suppose that X belongs to an automorphism group, Lie algebra, or Jordan algebra. Suppose, moreover, that $f(X)$ is a matrix function in the linear algebra sense (see [10, Chap. 1] for a precise definition). It is shown in [11, Thm. 3.1] that, for bilinear forms,

$$f(X)^{\star} = f(X^{\star}) \tag{3.1}$$

holds for all matrix functions f, assuming that f is defined at X and $X^{\star}$. For sesquilinear forms, (3.1) holds when, for example, the function f has a convergent power series representation $f(X) = \sum_{k=0}^{\infty} \alpha_k X^k$, with $\alpha_k \in \mathbb{R}$. Assuming that f is defined at the indicated arguments and that f satisfies (3.1), we have the following:

(i) if $X \in \mathbb{J}_M$, then $f(X) \in \mathbb{J}_M$ since $f(X)^{\star} = f(X)$;

(ii) if $X \in \mathbb{L}_M$, then $f(X)^{\star} = f(-X)$ so that the following hold:
- $f(X) \in \mathbb{J}_M$ if f is an even function [10, Prob. 1.20],
- $f(X) \in \mathbb{L}_M$ if f is an odd function [10, Prob. 1.20], and
- $f(X) \in \mathbb{G}_M$ if $f(-X) = f(X)^{-1}$ (e.g., the matrix exponential);

(iii) if $X \in \mathbb{G}_M$, then $f(X)^{\star} = f(X^{-1})$ so that the following hold:
- $f(X) \in \mathbb{L}_M$ if $f(X^{-1}) = -f(X)$ (e.g., the matrix logarithm), and
- $f(X) \in \mathbb{G}_M$ if $f(X^{-1}) = f(X)^{-1}$ for bilinear forms and $f(X^{-*}) = f(X)^{-*}$ for sesquilinear forms [11, Thm. 3.1] (e.g., the principal square root).

However, the map f does not necessarily have to be a matrix function in the sense of [10, Chap. 1]. For example, when the scalar product is unitary (i.e., $M = \beta\widetilde{M}$ for some unitary $\widetilde{M}$ and $\beta > 0$), if $X \in \mathbb{S}_M \in \{\mathbb{J}_M, \mathbb{L}_M, \mathbb{G}_M\}$ is nonsingular and has polar decomposition $X = UH$ with U unitary and H positive semidefinite, then $U \in \mathbb{S}_M$. Moreover, if $X \in \mathbb{G}_M$, then $H \in \mathbb{G}_M$ [18, sect. 5]. Note that the factor U is a "generalized matrix function" in the sense of [6]. The existence of structured singular value decompositions (see, e.g., [4], [22]) can be used to derive conditions on f such that other generalized matrix functions preserve structure. A precise statement is beyond our scope here and left for future research.

Recall from section 2 that to compute or approximate $\operatorname{cond}_{\mathbb{S}_M}(f, X)$ for $X \in \mathbb{S}_M \in \{\mathbb{J}_M, \mathbb{L}_M, \mathbb{G}_M\}$, we need to

(a) characterize the tangent space $T_X\mathbb{S}_M$ and its dimension p over $\mathbb{K}$, and

(b) construct a matrix B with orthonormal columns, if possible, such that for $E \in T_X\mathbb{S}_M$, $\operatorname{vec}(E) = By$ for some vector $y \in \mathbb{K}^p$.

Before we explain how to do so, we recall some useful properties of vec and the Kronecker product [16]. For all $A, C, Y \in \mathbb{F}^{n\times n}$,

$$\operatorname{vec}(AYC) = (C^T \otimes A)\operatorname{vec}(Y), \tag{3.2}$$

$$\operatorname{vec}(A^T) = P\operatorname{vec}(A), \tag{3.3}$$

where P is the $n^2 \times n^2$ vec-permutation matrix [7], and

$$P(A \otimes C) = (C \otimes A)P. \tag{3.4}$$

It follows from (3.3) that $\mathrm{vec}(e_i e_j^T) = P\mathrm{vec}(e_j e_i^T)$ so that $Pe_{(i-1)n+j} = e_{(j-1)n+i}$, $i, j = 1, \dots, n$. Hence

$$(P - I)e_{(i-1)n+i} = 0, \quad i = 1, \dots, n, \tag{3.5}$$

$$(P - \sigma I)(e_{(i-1)n+j} + \sigma e_{(j-1)n+i}) = 0, \quad \sigma = \pm 1, \quad 1 \le i < j \le n, \tag{3.6}$$

from which it follows that P has eigenvalue 1 with multiplicity $n(n+1)/2$ and eigenvalue -1 with multiplicity $n(n-1)/2$ [7].

3.1. Computation of $\mathrm{cond}_{\mathbb{S}_M}(f, X)$ when $\mathbb{S}_M$ is a Jordan or Lie algebra.

Davies [2] showed that for bilinear forms with $M = \pm M^T$ and $M^T M = I$, i.e., when the scalar product is orthosymmetric and unitary [18], $\mathrm{cond}(f, X) = \mathrm{cond}_{\mathbb{S}_M}(f, X)$ for

- all functions f and either $X \in \mathbb{J}_I$ or $X \in \mathbb{L}_I$ (i.e., for symmetric and skew-symmetric matrices X), and
- even functions f and $X \in \mathbb{L}_M$.

Davies also showed that equality between structured and unstructured condition numbers for $X \in \mathbb{J}_M$ does not always hold but the ratio $\mathrm{cond}(f, X)/\mathrm{cond}_{\mathcal{M}}(f, X)$ is bounded in terms of n for all functions f. On the other hand, the ratio is unbounded for odd functions and $X \in \mathbb{L}_M$ (see [2, Tab. 6.1]). For sesquilinear forms with $M = \pm M^T \in \mathbb{R}^{n \times n}$ and $M^T M = I$, Davies showed that $\mathrm{cond}(f, X) = \mathrm{cond}_{\mathcal{M}}(f, X)$ for Jordan algebras and that equality also holds for Lie algebras when f is an odd or even function.

In what follows, we show how to compute or approximate $\mathrm{cond}_{\mathcal{M}}(f, X)$ when it differs from $\mathrm{cond}(f, X)$. We make the simplifying assumption that $M = \mu M^T \in \mathbb{R}^{n \times n}$ with $\mu = \pm 1$ (which is true in essentially all the cases of practical interest), but we do not assume that M is orthogonal, as in [2]. The techniques we use are similar to that in [2].

It follows directly from (2.10) that the tangent space of any vector subspace is the vector subspace itself. Hence, for any $X \in \mathbb{S}_M$ with $\mathbb{S}_M \in \{\mathbb{J}_M, \mathbb{L}_M\}$,

$$T_X \mathbb{S}_M = \mathbb{S}_M.$$

To construct a basis for $T_X \mathbb{S}_M$, we start with bilinear forms first and then use the results to construct a basis when the scalar product is a sesquilinear form.

3.1.1. Bilinear forms on $\mathbb{F}^n$ and $\mathbb{F}$-differentiable maps.

Let $E \in \mathbb{S}_M$ and write $\mathrm{vec}(E) = B_{\mathbb{S}_M} y$ for some $y \in \mathbb{F}^p$ with $B \in \mathbb{F}^{n^2 \times p}$ of full column rank. Let $s = 1$ if $\mathbb{S}_M = \mathbb{J}_M$ and $s = -1$ if $\mathbb{S}_M = \mathbb{L}_M$. Then

$$E \in \mathbb{S}_M \iff E^\star = sE \iff E^T M - sME = 0. \tag{3.7}$$

If we assume $M = \mu M^T$ with $\mu \in \{+1, -1\}$, we can easily infer the value of p. Indeed, in this case, (3.7) shows that $ME = s\mu(ME)^T$, i.e., ME is either (complex) symmetric or (complex) skew-symmetric so that ME and therefore E, since M is nonsingular, depend on

$$p = n(n + s\mu)/2 = \mathrm{rank}(B_{\mathbb{S}_M}) = \dim_{\mathbb{F}}(T_X \mathbb{S}_M)$$

parameters that are real if $\mathbb{F} = \mathbb{R}$ and complex otherwise. Now applying the vec operator to (3.7) and using (3.2) and (3.3) we find that

$$\big((M^T \otimes I_n)P - s(I_n \otimes M)\big)\mathrm{vec}(E) = 0. \tag{3.8}$$

The property (3.4) combined with $M = \mu M^T$ implies that

$$(M^T \otimes I_n)P - s(I_n \otimes M) = (\mu P - sI_{n^2})(I_n \otimes M).$$

Since P has eigenvalues 1 with multiplicity $n(n+1)/2$ and -1 with multiplicity $n(n-1)/2$, and M is nonsingular, we have that

$$\operatorname{rank}\big((M^T \otimes I_n)P - s(I_n \otimes M)\big) = \operatorname{rank}\big((\mu P - sI_{n^2})(I_n \otimes M)\big) = n(n-\mu s)/2$$

so that the dimension of the null space of $(\mu P - sI_{n^2})(I_n \otimes M)$ is $n(n+\mu s)/2$. Hence, from this and on using (3.8) and $\operatorname{vec}(E) = B_{\mathbb{S}_M} y$, it follows that

$$\operatorname{range}(B_{\mathbb{S}_M}) = \operatorname{null}\big((P - \mu s I_{n^2})(I_n \otimes M)\big). \tag{3.9}$$

If we define $D_{\mathbb{S}_M} \in \mathbb{R}^{n^2 \times p}$ by

$$D_{\mathbb{S}_M} = \begin{cases} \begin{bmatrix} \widetilde{D}_{\mu s} & \check{I} \end{bmatrix} & \text{if } \mu s = 1, \\ \ \ \widetilde{D}_{\mu s} & \text{if } \mu s = -1, \end{cases} \tag{3.10}$$

where $\widetilde{D}_{\mu s} \in \mathbb{R}^{n^2 \times n(n-1)/2}$ has for columns the $n(n-1)/2$ unit vectors

$$(e_{(i-1)n+j} + \mu s e_{(j-1)n+i})/\sqrt{2}, \qquad 1 \le i < j \le n,$$

and $\check{I} \in \mathbb{R}^{n^2 \times n}$ has for columns the vectors $e_{(i-1)n+i}$, $i = 1, \dots n$, then $D_{\mathbb{S}_M}$ has full rank and orthonormal columns, and from (3.5)–(3.6) we have that $(P - \mu s I_{n^2})D_{\mathbb{S}_M} = 0$. From this and (3.9), it follows that

$$B_{\mathbb{S}_M} = (I_n \otimes M^{-1})D_{\mathbb{S}_M}, \tag{3.11}$$

and this construction can be done in at most $O(n^3)$ operations by exploiting (3.2) or by using the block diagonal structure of $(I \otimes M^{-1})$ and the special structure of $D_{\mathbb{S}_{\rm M}}$.

Note that if M is orthogonal then $B_{\mathbb{S}_M} = \mu(I_n \otimes M)D_{\mathbb{S}_M}$ has orthonormal columns and its construction is essentially computation free. In this case,

$$\operatorname{cond}_{\mathbb{S}_M}(f, X) = \|K_f(X)(I_n \otimes M)D_{\mathbb{S}_M}\|_2 \tag{3.12}$$

and we refer the reader to section 2.2 for the computation or approximation of $\operatorname{cond}_{\mathbb{S}_M}(f, X)$. Davies showed in [2, Thm. 3.3] that (using our notation)

$$\operatorname{cond}_{\mathbb{S}_M}(f, X) = \frac{1}{2}\|K_f(X)P(I_n \otimes M)(P - \mu s I_{n^2})\|_2,$$

which is a less compact expression than in (3.12) since $D_{\mathbb{S}_M}$ has $p = n(n+\mu s)/2 < n^2$ columns for $n > 1$.

When M is not orthogonal, then, unless it has a special structure that somehow allows the use of computational tricks, orthogonalizing the columns of $B_{\mathbb{S}_M}$, or equivalently computing $B^+_{\mathbb{S}_M}$, costs $O(n^6)$ operations. However, instead of orthonormalizing B or computing its pseudoinverse we can use (2.15) to obtain lower and upper bounds for $\operatorname{cond}_{\mathcal{M}}(f, X)$ that are cheaper to compute than $\operatorname{cond}_{\mathcal{M}}(f, X)$. We note that $\|B_{\mathbb{S}_M}\|_2 \le \|M^{-1}\|_2$. For an upper bound on $\|B^+_{\mathbb{S}_M}\|_2$, we remark that the Moore–Penrose inverse of a full column rank matrix A is the minimal left inverse, i.e., for any left inverse A^L of A, $\|A^+\|_2 \le \|A^L\|_2$. This fact is a special case of [14, Thm. 1] but can also be shown directly using the definition of the Moore–Penrose pseudoinverse and the singular value decomposition. Now, $D_{\mathbb{S}_M}^T(I \otimes M)$ is a left inverse for $B_{\mathbb{S}_M}$, and hence $\|B_{\mathbb{S}_M}^+\|_2 \le \|D_{\mathbb{S}_M}^T(I \otimes M)\|_2 \le \|M\|_2$. Then (2.15) yields the following lower and upper bounds on the structured condition number:

$$\frac{\|K_f(X)B_{\mathbb{S}_M}\|_2}{\|M^{-1}\|_2} \le \operatorname{cond}_{\mathbb{S}_M}(f,X) \le \|K_f(X)B_{\mathbb{S}_M}\|_2\|M\|_2. \tag{3.13}$$

By exploiting the special structure of $B_{\mathbb{S}_M}$ in (3.11), approximating $\|K_f(X)B_{\mathbb{S}_M}\|_2$ using Algorithm 2.7 costs $O(kn^3)$ operations, where k is the number of iterations. So the lower bound in (3.13) and an estimate for the upper bound can be computed in $O(kn^3)$ operations.

3.1.2. Sesquilinear forms or complex bilinear forms with $\mathbb{R}$-differentiable map. When either the scalar product is a sesquilinear form or a complex bilinear form but we are dealing with a map that is only real differentiable, the ambient and base fields differ (see section 2.2.2). The latter case is simple to deal with since under our assumptions on M, the matrix in (3.11) is real, and on using the map ρ in (2.7), we have that

$$B_{\mathbb{S}_M,\mathbb{R}} = \begin{bmatrix} B_{\mathbb{S}_M} \\ 0 \end{bmatrix} = \begin{bmatrix} (I_n \otimes M^{-1})D_{\mathbb{S}_M} \\ 0 \end{bmatrix}, \tag{3.14}$$

where $\mathbb{S}_M \in \{\mathbb{J}_M, \mathbb{L}_M\}$. For sesquilinear forms, a little more handling is needed. We have that

$$E \in T_X\mathbb{S}_M = \mathbb{S}_M \quad \Longleftrightarrow \quad E^\star = sE \quad \Longleftrightarrow \quad \begin{cases} \big(\operatorname{Re}(E)\big)^\star = s\operatorname{Re}(E), \\ \big(\operatorname{Im}(E)\big)^\star = -s\operatorname{Im}(E). \end{cases}$$

Hence, if $E \in \mathbb{J}_M$, then $\operatorname{Re}(E) \in \mathbb{J}_M$ and $\operatorname{Im}(E) \in \mathbb{L}_M$ so that

$$\begin{aligned} \operatorname{vec}(E) &= \operatorname{vec}\big(\operatorname{Re}(E)\big) + \mathrm{i}\operatorname{vec}\big(\operatorname{Im}(E)\big) \\ &= B_{\mathbb{J}_M}x + \mathrm{i}B_{\mathbb{L}_M}y \\ &= [\, B_{\mathbb{J}_M} \quad \mathrm{i}\, B_{\mathbb{L}_M} \,] \begin{bmatrix} x \\ y \end{bmatrix} \\ &=: \widehat{B}_{\mathbb{J}_M} z \end{aligned}$$

for some $z \in \mathbb{R}^{n^2}$. The matrices $B_{\mathbb{J}_M}$ and $B_{\mathbb{L}_M}$ are as in (3.11) so that the $n^2 \times n^2$ matrix

$$\widehat{B}_{\mathbb{J}_M} = (I_n \otimes M^{-1})\,[\, D_{\mathbb{J}_M} \quad \mathrm{i}D_{\mathbb{L}_M} \,]$$

forms a basis over the base field $\mathbb{R}$ for $T_X\mathbb{J}_M$ (i.e., $p = n^2$). Similarly, we find that

$$\widehat{B}_{\mathbb{L}_M} = (I_n \otimes M^{-1})\,[\, D_{\mathbb{L}_M} \quad \mathrm{i}D_{\mathbb{J}_M} \,] \tag{3.15}$$

forms a basis over $\mathbb{R}$ for $T_X\mathbb{L}_M$. Using the map ρ in (2.7), we have that $\rho(X) \in \rho(\mathbb{S}_M)$ and $\rho(E) \in T_{\rho(X)}\rho(\mathbb{S}_M)$ with $\mathbb{S}_M \in \{\mathbb{J}_M, \mathbb{L}_M\}$. We can write $\operatorname{vec}\big(\rho(E)\big) = B_{\mathbb{S}_M,\mathbb{R}}z$ with $B_{\mathbb{S}_M,\mathbb{R}} = \left[\begin{smallmatrix} \operatorname{Re}(\widehat{B}_{\mathbb{S}_M}) \\ \operatorname{Im}(\widehat{B}_{\mathbb{S}_M}) \end{smallmatrix}\right] \in \mathbb{R}^{2n^2 \times n^2}$, and since M is real,

$$B_{\mathbb{J}_M,\mathbb{R}} = (I_n \otimes M^{-1})D_{\mathbb{J}_M} \oplus (I_n \otimes M^{-1})D_{\mathbb{L}_M}, \tag{3.16}$$

$$B_{\mathbb{L}_M,\mathbb{R}} = (I_n \otimes M^{-1})D_{\mathbb{L}_M} \oplus (I_n \otimes M^{-1})D_{\mathbb{J}_M}. \tag{3.17}$$

If M is orthogonal, then $B_{\mathbb{S}_M,\mathbb{R}}$ in (3.14) and (3.16)–(3.17) has orthonormal columns and

$$\operatorname{cond}_{\mathbb{S}_M}(f,X) = \|K_f^{(\mathbb{R})}(X)B_{\mathbb{S}_M,\mathbb{R}}\|_2. \tag{3.18}$$

Now if M is not orthogonal, we can use (2.15) to obtain lower and upper bounds for $\operatorname{cond}_{\mathbb{S}_M}(f,X)$. Note that $\|\widetilde{B}_{\mathbb{S}_M,\mathbb{R}}\|_2 \le \|M^{-1}\|_2$. We can use $B^L_{\mathbb{S}_M,\mathbb{R}} = D^T_{\mathbb{S}_M}(I_n \otimes M)$ as the left inverse for $B_{\mathbb{S}_M,\mathbb{R}}$ in (3.14), $B^L_{\mathbb{J}_M,\mathbb{R}} = D^T_{\mathbb{J}_M}(I_n\otimes M) \oplus D^T_{\mathbb{L}_M}(I_n \otimes M)$ as the left inverse for $B_{\mathbb{J}_M,\mathbb{R}}$ in (3.16), and $B^L_{\mathbb{L}_M,\mathbb{R}} = D^T_{\mathbb{L}_M}(I_n\otimes M) \oplus D^T_{\mathbb{J}_M}(I_n \otimes M)$ as the left inverse for $B_{\mathbb{L}_M,\mathbb{R}}$ in (3.17), yielding $\|B^+_{\mathbb{S}_M,\mathbb{R}}\|_2 \le \|B^L_{\mathbb{S}_M,\mathbb{R}}\|_2 \le \|M\|_2$. Therefore,

$$\frac{\|K_f^{(\mathbb{R})}(X)B_{\mathbb{S}_M,\mathbb{R}}\|_2}{\|M^{-1}\|_2} \le \operatorname{cond}_{\mathbb{S}_M}(f,X) \le \|K_f^{(\mathbb{R})}(X)B_{\mathbb{S}_M,\mathbb{R}}\|_2\|M\|_2, \tag{3.19}$$

which provides lower and upper bounds for $\operatorname{cond}_{\mathbb{S}_M}(f,X)$ that are cheap to evaluate compared to orthonormalizing $B_{\mathbb{S}_M,\mathbb{R}}$ into $\widetilde{B}_{\mathbb{S}_M,\mathbb{R}}$ and then computing $\operatorname{cond}_{\mathbb{S}_M}(f,X)$ in (3.18) with $\widetilde{B}_{\mathbb{S}_M,\mathbb{R}}$ in place of $B_{\mathbb{S}_M,\mathbb{R}}$.

3.2. Computation of $\operatorname{cond}_{\mathcal{M}}(f,X)$ when $\mathcal{M}$ is an automorphism group.

The next lemma provides an explicit characterization of $T_X\mathbb{G}_M$.

LEMMA 3.2. *Let $\mathbb{G}_M$ be the automorphism group of a scalar product $\langle\cdot,\cdot\rangle_M$ on $\mathbb{F}^n$ and let $X \in \mathbb{G}_M$. Then the tangent space at X to $\mathbb{G}_M$ is given by*

$$T_X\mathbb{G}_M = \{E \in \mathbb{F}^{n\times n} \mid E = XF,\ F \in \mathbb{L}_M\}, \tag{3.20}$$

where $\mathbb{L}_M$ is the Lie algebra associated with $\langle\cdot,\cdot\rangle_M$.

Proof. We only prove the lemma for bilinear forms, the proof for sesquilinear forms being analogous. We start by showing that

$$T_X\mathbb{G}_M = \{E \in \mathbb{F}^{n\times n} \mid E^TMX + X^TME = 0\}.$$

By definition, $E \in T_X\mathbb{G}_M$ is equivalent to the existence of a smooth curve $\gamma(t) \in \mathbb{G}_M$ satisfying $\gamma(0) = X$, $\gamma'(0) = E$. Now $\gamma(t) \in \mathbb{G}_M$ is equivalent to $\gamma(t)^TM\gamma(t) = M$. By differentiating the latter equation and evaluating at $t = 0$, we obtain

$$\gamma'(0)^TM\gamma(0) + \gamma(0)^TM\gamma'(0) = 0.$$

Substituting $\gamma(0) = X$ and $\gamma'(0) = E$ gives

$$E^TMX + X^TME = 0. \tag{3.21}$$

Defining $F := X^{-1}E$ (note that $X \in \mathbb{G}_M$ implies that X is nonsingular) we can rewrite (3.21) as $F^TM + MF = 0$, which shows that $-F = M^{-1}F^TM = F^\star$, i.e., $F \in \mathbb{L}_M$.

Conversely, suppose that E satisfies (3.21), and again let $F := X^{-1}E$. Then $tF \in L_M$ for all t. Hence, $e^{tF} \in \mathbb{G}_M$ for all t [10, sect. 14.1.1]. Since $\mathbb{G}_M$ is a group under matrix multiplication, this implies that the smooth curve $\gamma(t) = Xe^{tF} \in \mathbb{G}_M$. Manifestly, $\gamma(0) = X$ and $\gamma'(0) = E$. Thus, $E \in T_X\mathbb{G}_M$. □

Note that Lemma 3.2 implies that (a) $\dim_{\mathbb{K}}\mathbb{G}_M = \dim_{\mathbb{K}}\mathbb{L}_M$ so that $\dim_{\mathbb{K}}\mathbb{G}_M > 0$ (unless $n = 1$ and M represents a bilinear form, in which case $\mathbb{G}_M = \{1,-1\}$), and (b) $\mathbb{L}_M = T_I\mathbb{G}_M$ (observe that $I \in \mathbb{G}_M$). The latter property is, in fact, sometimes used as the *definition* of $\mathbb{L}_M$; see, e.g., [21, Def. 5.7]. It follows from Lemma 3.2 that any $E \in T_X\mathbb{G}_M$ can be written as $E = XF$ with $F \in \mathbb{L}_M$. If $M = \mu M^T$ with $\mu = \pm 1$, then

$$\mathrm{vec}(E) = \mathrm{vec}(XF) = (I_n \otimes X)\mathrm{vec}(F) = (I_n \otimes X)By,$$

where

- for bilinear forms, $B = B_{\mathbb{L}_M}$ is as in (3.11) with $\mathbb{S}_M = \mathbb{L}_M$ and $y \in \mathbb{F}^p$ with $p = n(n-\mu)/2$, and
- for sesquilinear forms, $B = \widehat{B}_{\mathbb{L}_M}$ is as in (3.15) and $y \in \mathbb{R}^p$ with $p = n^2$.

If we write $\mathrm{vec}(E) = B_{\mathbb{G}_M} y$, then

$$B_{\mathbb{G}_M} = \begin{cases} (I_n \otimes XM^{-1})D_{\mathbb{L}_M} & \text{for bilinear forms,} \\ (I_n \otimes XM^{-1}) \left[\, D_{\mathbb{L}_M} \quad \mathrm{i}D_{\mathbb{J}_M} \,\right] & \text{for sesquilinear forms.} \end{cases} \tag{3.22}$$

As a consequence, for complex bilinear forms ($\mathbb{F} = \mathbb{C}$) and a real differentiable, but not complex differentiable, f (so that we pick $\mathbb{K} = \mathbb{R}$) we have

$$B_{\mathbb{G}_M,\mathbb{R}} = \begin{bmatrix} \mathrm{Re}(B_{\mathbb{G}_M}) \\ \mathrm{Im}(B_{\mathbb{G}_M}) \end{bmatrix} = \begin{bmatrix} (I_n \otimes \mathrm{Re}(X)M^{-1})D_{\mathbb{L}_M} \\ (I_n \otimes \mathrm{Im}(X)M^{-1})D_{\mathbb{L}_M} \end{bmatrix}. \tag{3.23}$$

Similarly, for sesquilinear forms we also have $\mathbb{F} = \mathbb{C} \neq \mathbb{R} = \mathbb{K}$ and

$$B_{\mathbb{G}_M,\mathbb{R}} = \begin{bmatrix} \mathrm{Re}(B_{\mathbb{G}_M}) \\ \mathrm{Im}(B_{\mathbb{G}_M}) \end{bmatrix} = \begin{bmatrix} (I_n \otimes \mathrm{Re}(X)M^{-1})D_{\mathbb{L}_M} & -(I_n \otimes \mathrm{Im}(X)M^{-1})D_{\mathbb{J}_M} \\ (I_n \otimes \mathrm{Im}(X)M^{-1})D_{\mathbb{L}_M} & (I_n \otimes \mathrm{Re}(X)M^{-1})D_{\mathbb{J}_M} \end{bmatrix}. \tag{3.24}$$

Hence, if M and X are both orthogonal or unitary, then $B_{\mathbb{G}_M}$ in (3.22) and $B_{\mathbb{G}_M,\mathbb{R}}$ in (3.23)–(3.24) have orthonormal columns and

$$\mathrm{cond}_{\mathbb{G}_M}(f, X) = \begin{cases} \|K_f(X)B_{\mathbb{G}_M}\|_2 & \text{for } \mathbb{K} = \mathbb{F}, \\ \|K_f^{(\mathbb{R})}(X)B_{\mathbb{G}_M,\mathbb{R}}\|_2 & \text{for } \mathbb{K} \neq \mathbb{F}. \end{cases} \tag{3.25}$$

Now if M and X are not orthogonal or unitary, then $B_{\mathbb{G}_M}$ does not have orthonormal columns and orthonormalizing these columns, or computing the pseudoinverse, can cost as much as $O(n^6)$ operations. As for the Lie and Jordan algebra cases, we can use (2.15) to obtain lower and upper bounds for $\mathrm{cond}_{\mathbb{G}_M}(f, X)$ that are hopefully cheaper to compute than $\mathrm{cond}_{\mathbb{G}_M}(f, X)$. Note that $\|B_{\mathbb{G}_M}\|_2 \leq \|M^{-1}\|_2\|X\|_2$ and since $D_{\mathbb{L}_M}^T(I \otimes X^TM)$ is a left inverse for $B_{\mathbb{G}_M}$, $\|B_{\mathbb{G}_M}^+\|_2 \leq \|D_{\mathbb{L}_M}^T(I \otimes X^TM)\|_2 \leq \|X\|_2\,\|M\|_2$. Then, for the case of (real or complex) bilinear forms, (2.15) yields the following lower and upper bounds on the structured condition number:

$$\frac{\|K_f(X)B_{\mathbb{G}_M}\|_2}{\|M^{-1}\|_2\|X\|_2} \leq \mathrm{cond}_{\mathbb{G}_M}(f, X) \leq \|K_f(X)B_{\mathbb{G}_M}\|_2\|X\|_2\,\|M\|_2. \tag{3.26}$$

As for the bounds in (3.13), the cost of computing the lower and upper bounds in (3.26) is $O(kn^3)$ operations, where k is the number of iterations performed by Algorithm 2.7.

If M is orthogonal, then $\|X^{-1}\| = \|X\|$ for both the 2-norm and Frobenius norm since $X^{-1} = X^\star = M^{-1}X^TM$ so that $\|X\| = \kappa(X)^{1/2}$, where $\kappa(X) := \|X\|\|X^{-1}\|$. In particular, if X is well-conditioned, i.e., $\kappa_F(X) \approx 1$, then $\|K_f(X)B_{\mathbb{G}_M}\|_2/\|f(X)\|_F$ offers a good estimate of the relative structured condition number since

$$\frac{\|K_f(X)B_{\mathbb{G}_M}\|_2}{\|f(X)\|_F} \leq \mathrm{cond}_{\mathbb{G}_M}(f, X)\frac{\|X\|_F}{\|f(X)\|_F} \leq \kappa_F(X)\frac{\|K_f(X)B_{\mathbb{G}_M}\|_2}{\|f(X)\|_F}. \tag{3.27}$$

Hence, the quality of the bounds in (3.26) is likely to be influenced by the condition number of X.

Using Lemma A.1 and (2.18), and by an argument analogous to the one for bilinear forms, we obtain the following lower and upper bounds on the structured conditioned number for automorphism groups associated with sesquilinear forms and complex bilinear forms when f is only $\mathbb{R}$-differentiable:

$$\frac{\|K_f^{(\mathbb{R})}(X)B_{\mathbb{G}_M,\mathbb{R}}\|_2}{\|M^{-1}\|_2\|X\|_2} \le \mathrm{cond}_{\mathbb{G}_M}(f,X) \le \|K_f^{(\mathbb{R})}(X)B_{\mathbb{G}_M,\mathbb{R}}\|_2\|X\|_2\ \|M\|_2. \tag{3.28}$$

Example 3.3. We apply our theory to the 2×2 real diagonal symplectic matrix

$$X = \begin{bmatrix} \mathrm{e}^{a} & 0 \\ 0 & \mathrm{e}^{-a} \end{bmatrix}, \quad a > 0. \tag{3.29}$$

Note that X is in three different automorphism groups $\mathbb{G}_J$:

1. the real symplectic group associated with the real bilinear form defined by $J = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$ with ambient field $\mathbb{F} = \mathbb{R}$ and base field $\mathbb{K} = \mathbb{R}$,
2. the complex symplectic group associated with the complex bilinear form defined by J with $\mathbb{F} = \mathbb{K} = \mathbb{C}$, and
3. the conjugate symplectic group associated with the sesquilinear form defined by J with $\mathbb{F} = \mathbb{C}$ and $\mathbb{K} = \mathbb{R}$.

The matrix $B_{\mathbb{G}_M}$ in (3.22) with $M = J$ and $D_{\mathbb{L}_J} = [(e_2+e_3)/\sqrt{2}, e_1, e_4]$, $D_{\mathbb{J}_J} = [\frac{1}{\sqrt{2}}(e_2 - e_3)]$ is given by

$$B_{\mathbb{G}_J} = \begin{cases} [\frac{1}{\sqrt{2}}(\mathrm{e}^{-a}e_4 - \mathrm{e}^{a}e_1), \mathrm{e}^{-a}e_2, -\mathrm{e}^{a}e_3] \text{ for real/complex symplectic,} \\ [\frac{1}{\sqrt{2}}(\mathrm{e}^{-a}e_4 - \mathrm{e}^{a}e_1), \mathrm{e}^{-a}e_2, -\mathrm{e}^{a}e_3, -\frac{\mathrm{i}}{\sqrt{2}}(\mathrm{e}^{a}e_1 + \mathrm{e}^{-a}e_4)] \text{ for conj. symplectic.} \end{cases}$$

The orthonormalization of $B_{\mathbb{G}_J}$ for the real and complex symplectic groups and that of $B_{\mathbb{G}_J,\mathbb{R}}$ in (3.24) for the conjugate symplectic group take the form

$$\begin{aligned} \widetilde{B}_{\mathbb{G}_J} &= \left[(1+\mathrm{e}^{4a})^{-1/2}(\mathrm{e}^{2a}e_1 - e_4), e_2, e_3\right] \in \mathbb{R}^{4\times 3}, \\ \widetilde{B}_{\mathbb{G}_J,\mathbb{R}} &= \left[(1+\mathrm{e}^{4a})^{-1/2}(\mathrm{e}^{2a}e_1 - e_4), e_2, e_3, (1+\mathrm{e}^{4a})^{-1/2}(\mathrm{e}^{2a}e_5 + e_8)\right] \in \mathbb{R}^{8\times 4}. \end{aligned}$$

These bases will be needed in the computation of the structured condition numbers and their lower and upper bounds.

As differentiable map f, we consider the principal matrix logarithm [10]. Since X has no eigenvalues on $\mathbb{R}^-$, $\log X^{-1} = -\log X$, and since $X \in \mathbb{G}_J$, we have from section 3 that $\log X \in \mathbb{L}_J$. Indeed, $\log X = \left[\begin{smallmatrix} a & 0 \\ 0 & -a \end{smallmatrix}\right]$ is Hamiltonian. To compute the unstructured condition number, we construct $K_{\log}(X)$ one column at a time using $K_{\log}(X)e_{i+2j-2} = \mathrm{vec}(L_{\log}(X, e_ie_j^T))$, $i,j = 1,2$, as in (2.4). Because X and $e_ie_i^T$ commute, we have that $L_{\log}(X, e_ie_i^T) = X^{-1}e_ie_i^T$, $i = 1,2$ [10, Prob. 3.8]. Since $X + e_ie_j^T$ is triangular, we can use the explicit expression for the matrix function of a triangular matrix in [10, p. 84] and the definition of the Fréchet derivative in (2.1) to get an expression for $L_{\log}(X, e_ie_j^T)$. We find that

$$K_{\log}(X) = \mathrm{diag}\left(\mathrm{e}^{-a}, \frac{a}{\sinh a}, \frac{a}{\sinh a}, \mathrm{e}^{a}\right)$$

so that

$$\mathrm{cond}(\log, X) = \|K_{\log}(X)\|_2 = \mathrm{e}^{a}, \qquad \mathrm{rcond}(\log, X) = \frac{\mathrm{e}^{2a}}{a},$$

showing that the unstructured absolute and relative condition numbers increase rapidly with a. For the real and complex symplectic groups, we find that

$$\operatorname{cond}_{\mathbb{G}_J}(\log, X) = \|K_{\log}(X)\widetilde{B}_{\mathbb{G}_J}\|_2 = \frac{a}{\sinh a} < 1.$$

Similarly, for the conjugate symplectic group, we have that

$$\operatorname{cond}_{\mathbb{G}_J}(\log, X) = \|K_{\log}^{(\mathbb{R})}(X)\widetilde{B}_{\mathbb{G}_{J,\mathbb{R}}}\|_2 = \frac{a}{\sinh a},$$

where $K_{\log}^{(\mathbb{R})}(X) = K_{\log}(X) \oplus K_{\log}(X)$. Hence, for all three symplectic groups, the ratio $\operatorname{cond}_{\mathbb{G}_J}(\log, X)/\operatorname{cond}(\log, X) = a/(\mathrm{e}^a \sinh a)$ exponentially decays as $a \to \infty$. The lower and upper bounds in both (3.26) and (3.28) yield

$$\frac{a}{\sinh a} \le \operatorname{cond}_{\mathbb{G}_J}(\log, X) \le \frac{a\mathrm{e}^{2a}}{\sinh a},$$

showing that for this particular matrix and function f, the lower bound is attained, whereas the upper bound is larger than $\operatorname{cond}(\log, X)$. Also, for this particular choice of f and X, and the conjugate symplectic group, equality holds in (2.20); that is, $\operatorname{cond}_{\mathbb{G}_J}(\log, X) = \|K_f^{(\mathbb{R})}(X)\widetilde{B}_{\mathbb{G}_J,\mathbb{R}}\|_2 = \|K_f^{(\mathbb{C})}(X)\widetilde{B}_{\mathbb{G}_J}\|_2$ for some orthonormalization $\widetilde{B}_{\mathbb{G}_J}$ of the basis $B_{\mathbb{G}_J}$ for the conjugate symplectic group, but this is not always the case, as we next illustrate.

Example 3.4. Let us consider $f(X) = X^2$, X as in (3.29) with $a = \log 2$ and the conjugate symplectic group as the manifold. Since f is complex differentiable, we can compute $K_f^{(\mathbb{C})}(X) \equiv K_f(X)$ and find that $K_f^{(\mathbb{C})}(X) = \operatorname{diag}(4, 5/2, 5/2, 1)$ so that $K_f^{(\mathbb{R})}(X) = K_f^{(\mathbb{C})}(X) \oplus K_f^{(\mathbb{C})}(X)$. Hence,

$$\operatorname{cond}_{\mathbb{G}_J}(f, X) = \|K_f^{(\mathbb{R})}(X)\widetilde{B}_{\mathbb{G}_{J,\mathbb{R}}}\|_2 = \sqrt{\frac{257}{17}} \approx 3.89.$$

Now, for an orthonormalization $\widetilde{B}_{\mathbb{G}_J}$ of the 4×4 matrix $B_{\mathbb{G}_J}$ for the conjugate symplectic group, we find that $\|K_f^{(\mathbb{C})}(X)\widetilde{B}_{\mathbb{G}_J}\|_2 = 4 > \sqrt{\frac{257}{17}}$ and hence the inequality is strict in (2.20) for that particular case so that $\operatorname{cond}_{\mathbb{G}_J}(f, X) \ne \|K_f^{(\mathbb{C})}(X)\widetilde{B}_{\mathbb{G}_J}\|_2$.

Example 3.5. Finally, consider the map f that associates to X the unitary factor of its polar decomposition $X = UH$. This map is real differentiable but not complex differentiable. For the matrix X in (3.29), we have that $U = I_2$. Following [19, Cor. 3.12], we can compute the Kronecker form of the Fréchet derivative of the unitary factor analytically. Indeed, using [19, eq. (3.13)] with $U = V = I_2$ we have

$$L_f(X, E) = F \circ (E - E^*) + \mathrm{i}(H - F) \circ \operatorname{Im}(E),$$

where $\circ$ denotes the Schur product and, for our choice of X,

$$F = \frac{1}{2\cosh(a)}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad H = \begin{bmatrix} \mathrm{e}^{-a} & 0 \\ 0 & \mathrm{e}^{a} \end{bmatrix} + F.$$

For real perturbations only, $K_f(X)$ is the 4×4 matrix

$$K_f(X) = 0 \oplus \frac{1}{2\cosh(a)}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \oplus 0$$

so that the unstructured condition number is $\mathrm{cond}(f,X) = 1/\cosh(a)$. For the manifold of real symplectic matrices, we find that

$$\mathrm{cond}_{\mathbb{G}_J}(f,X) = \|K_f(X)\widetilde{B}_{\mathbb{G}_j}\|_2 = 1/\cosh(a).$$

Hence, $\mathrm{cond}_{\mathbb{G}_J}(f,X)/\mathrm{cond}(f,X) = 1$. For complex perturbations, the Kronecker form of the *real* Fréchet derivative is an 8×8 matrix $K_f^{(\mathbb{R})}(X)$. Using [19, eq. (3.13)], we obtain

$$K_f^{(\mathbb{R})}(X) = 0 \oplus \frac{1}{2\cosh(a)}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \oplus 0 \oplus \mathrm{e}^{-a} \oplus \frac{1}{2\cosh(a)}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \oplus \mathrm{e}^{a},$$

so that the unstructured conditioned number is $\mathrm{cond}(f,X) = \|K_f^{(\mathbb{R})}(X)\|_2 = \mathrm{e}^a$.

For the manifold $\mathbb{G}_J$ of complex symplectic matrices that we view as a real manifold since f is not complex differentiable, it follows from (3.23) and (3.24) that $B_{\mathbb{G}_J,\mathbb{R}} = \left[\begin{smallmatrix} B_{\mathbb{G}_J} \\ 0 \end{smallmatrix}\right]$ and its orthonormalization takes the form $\widetilde{B}_{\mathbb{G}_J,\mathbb{R}} = \left[\begin{smallmatrix} \widetilde{B}_{\mathbb{G}_J} \\ 0 \end{smallmatrix}\right]$. Hence,

$$\mathrm{cond}_{\mathbb{G}_J}(f,X) = \|K_f^{(\mathbb{R})}(X)\widetilde{B}_{\mathbb{G}_J,\mathbb{R}}\|_2 = 1/\cosh(a)$$

and $\mathrm{cond}_{\mathbb{G}_J}(f,X)/\mathrm{cond}(f,X) = 1/(\mathrm{e}^a\cosh(a))$, which decays exponentially with a.

Finally, for the real manifold of conjugate symplectic matrices $\mathbb{G}_J$, it follows from (3.24) that $B_{\mathbb{G}_J,\mathbb{R}} = \left[\begin{smallmatrix} \mathrm{Re}(B_{\mathbb{G}_J}) \\ \mathrm{Im}(B_{\mathbb{G}_J}) \end{smallmatrix}\right]$. It can be readily proved that

$$\mathrm{cond}_{\mathbb{G}_J}(f,X) = \|K_f^{(\mathbb{R})}(X)\widetilde{B}_{\mathbb{G}_J,\mathbb{R}}\|_2 = 1/\cosh(a).$$

Hence, $\mathrm{cond}_{\mathbb{G}_J}(f,X)/\mathrm{cond}(f,X) = 1/(\mathrm{e}^a\cosh(a))$, which again decays exponentially with a.

4. Numerical experiments.

The purpose of this section is to compare the structured and unstructured condition numbers numerically and to illustrate the quality of the lower and upper bounds on the structured condition number for automorphism groups of real and complex bilinear forms and sesquilinear forms displayed in (3.26) and (3.28) since these bounds are cheaper to compute than $\mathrm{cond}_{\mathcal{M}}(f,X)$. All our experiments are performed with MATLAB R2017a, for which the unit roundoff is $u \approx 1.1 \times 10^{-16}$.

We consider the maps

$$\begin{aligned} f_1 &: \mathbb{G}_M \to \mathbb{L}_M, & f_2 &: \mathbb{G}_M \to \mathbb{G}_M, & f_3 &: \mathbb{G}_M \to \mathbb{G}_M \cap \mathbb{G}_I, \\ & X \mapsto \log X & & X \mapsto X^{1/2} & & X \mapsto U \end{aligned}$$

where $\log X$ is the principal logarithm of X, $X^{1/2}$ is the principal square root of X, and U is the unitary factor in the polar decomposition $X = UH$ of X. Both f_1 and f_2 are complex differentiable, but f_3 is only real differentiable. Algorithms 2.6 and 2.7 require an algorithm to compute $L_f(X,E)$ for a given E. For the logarithm, we use the MATLAB function `logm_frechet_pade` from Higham's Matrix Function Toolbox [8]. For the matrix square root, $L_{f_2} = L_{f_2}(X,E)$ is the solution to the Sylvester equation $XL_{f_2} + L_{f_2}X = E$ [10, p. 134], which can be computed with the function `sylvsol` from [8]. For the polar orthogonal factor, one possibility is to follow Higham, who showed in [9, Thm. 2.5] that $L_{f_3}(X,E) = (E - UL_H)H^{-1}$, where L_H is the solution to the Sylvester equation $HL_H + L_HH = A^TE + E^TA$. Alternatively, in [19, Thm. 3.8 and Cor. 3.10], Noferini obtained explicit formulae for the Fréchet derivatives of any generalized matrix function [6]. We prefer the latter approach for efficiency and numerical stability.

We use Jagger's MATLAB Toolbox for Classical Matrix Groups [13] to generate random matrices with specified condition numbers in the

- real perplectic group (bilinear form with $M = R$, ambient field $\mathbb{F} = \mathbb{R}$, and base field $\mathbb{K} = \mathbb{R}$),
- real pseudo-orthogonal group (bilinear form with $M = S_{p,q}$, $\mathbb{F} = \mathbb{K} = \mathbb{R}$),
- complex orthogonal group (bilinear form with $M = I$, $\mathbb{F} = \mathbb{K} = \mathbb{C}$),
- complex pseudo-orthogonal group (bilinear form with $M = S_{p,q}$, $\mathbb{F} = \mathbb{K} = \mathbb{C}$),
- pseudo-unitary group (sesquilinear form with $M = S_{p,q}$, $\mathbb{F} = \mathbb{C}$, $\mathbb{K} = \mathbb{R}$),
- real symplectic group (bilinear form with $M = J$, $\mathbb{F} = \mathbb{K} = \mathbb{R}$),
- complex symplectic group (bilinear form with $M = J$, $\mathbb{F} = \mathbb{K} = \mathbb{C}$),
- conjugate symplectic group (sesquilinear form with $M = J$, $\mathbb{F} = \mathbb{C}$, $\mathbb{K} = \mathbb{R}$),

with J, R, and $S_{p,q}$ as defined at the beginning of section 3. We check that the generated matrices X have no eigenvalues on the negative real line so that their principal logarithm $\log X$ and principal square root $X^{1/2}$ exist.

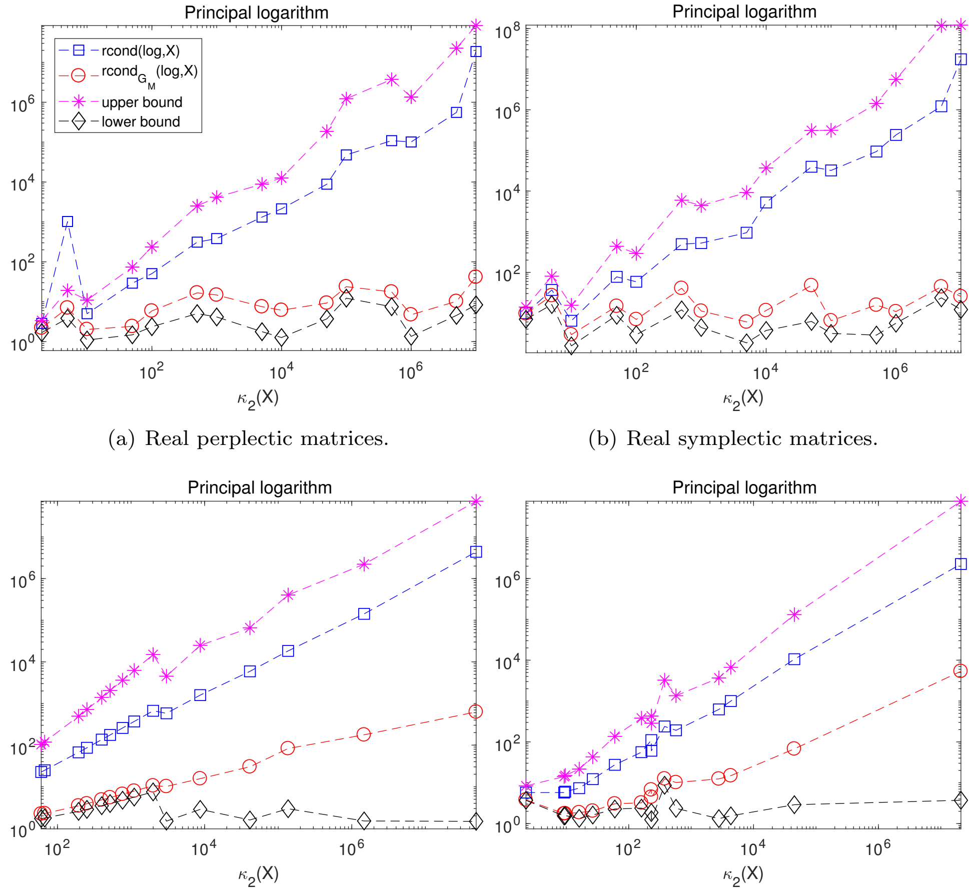
For our numerical experiments, we use Algorithm 2.6 and report the relative unstructured/structured condition numbers, i.e.,

$$\mathrm{rcond}(f, X) = \mathrm{cond}(f, X)\frac{\|X\|_F}{\|f(X)\|_F}, \qquad \mathrm{rcond}_{\mathbb{G}_M}(f, X) = \mathrm{cond}_{\mathbb{G}_M}(f, X)\frac{\|X\|_F}{\|f(X)\|_F},$$

rather than the absolute ones, as we will be varying the condition number of X. The upper and lower bounds in (3.26) and (3.28) multiplied by $\|X\|_F/\|f(X)\|_F$ provide upper and lower bounds for $\mathrm{rcond}_{\mathbb{G}_M}(f, X)$. These bounds and condition numbers are reported in Figure 1 for the principal logarithm of real perplectic, real symplectic, complex symplectic, and conjugate symplectic matrices of increasing condition numbers. In Figure 2, we report the same quantities for the principal square root of real pseudo-orthogonal matrices with $p = q = 5$, real pseudo-orthogonal matrices with $p = 1$, $q = 9$, complex pseudo-orthogonal matrices with $p = q = 5$, and pseudo-unitary matrices with $p = q = 5$. Figure 3 compares the relative unstructured/structured condition numbers for the map f_3 as well as the lower and upper bounds on $\mathrm{rcond}_{\mathbb{G}_M}(f, X)$. All plots show that the unstructured condition number can be much larger than the structured one, in particular when the argument X of the matrix function has a large condition number. When the latter is not too large, the lower bound in (3.26) or (3.28) offers a good approximation to $\mathrm{cond}_{\mathbb{G}_M}(f, X)$.

Our numerical experiments also suggest that for most of the real automorphism groups we consider (see Figure 1(a)–(b), Figure 2(a), and Figure 3(a)) the lower bound in (3.26) is a good approximation to $\mathrm{cond}_{\mathbb{G}_M}(f, X)$ even for badly conditioned matrices X. An explanation of why (3.26) is a good approximation for many, but not all, groups is left for future research; here, we note that more insight could be gained by studying the angle between the dominant singular vector of the matrix B and the vector y that achieves the maximum in the definition of a structured condition number. The plots show that the upper bound $\mathrm{ub\_cond}(f, X)$ is in general not sharp, in particular when $\kappa_2(X)$ is large, as expected from the analysis in section 3.2. Our experiments show that $\mathrm{cond}(f, X)$ is usually a better upper bound on $\mathrm{cond}_{\mathbb{G}_M}(f, X)$ than the upper bound in (3.26) or (3.28) but not always as plot (a) in Figure 1 indicates.

Let $A = X^{1/2}$ be the principal square root of X and let $\widehat{A}$ be the computed square root of X. The backward error of $\widehat{A}$ is $\|E\|_F$, where $E = \widehat{A}^2 - X$ is the unique matrix satisfying $\widehat{A} = (X + E)^{1/2}$. Then, on using (1.2), we get the following approximate



(a) Real perplectic matrices.

(b) Real symplectic matrices.

(c) Complex symplectic matrices.

(d) Conjugate symplectic matrices.

FIG. 1. *Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal logarithm of* 10×10 *randomly generated real perplectic matrices in plot* (a)*, real symplectic matrices in plot* (b)*, complex symplectic matrices in plot* (c)*, and conjugate symplectic matrices in plot* (d)*, with increasing condition number* $\kappa_2(X)$*.*

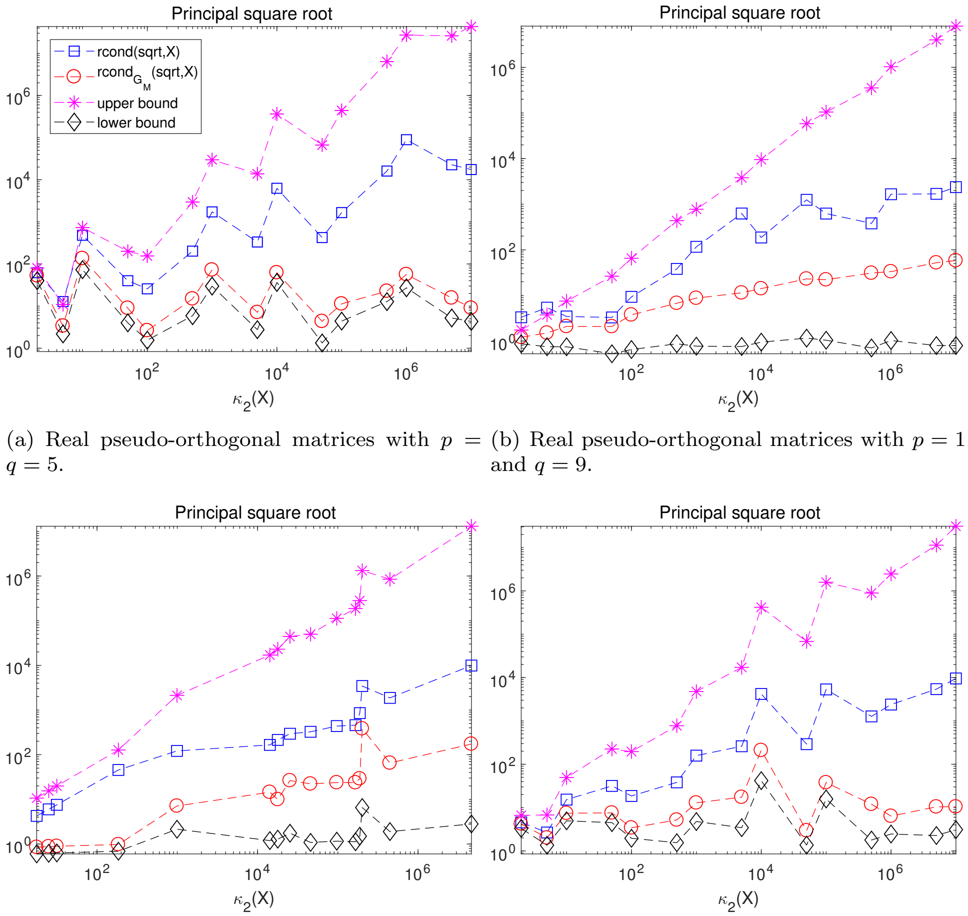
upper bound on the relative error:

$$\text{err}(\widehat{A}) := \frac{\|\widehat{A} - X^{1/2}\|_F}{\|X^{1/2}\|_F} \lesssim \text{cond}(\text{sqrt}, X)\frac{\|\widehat{A}^2 - X\|_F}{\|X^{1/2}\|_F}. \tag{4.1}$$

When $X \in \mathbb{G}_M$ and $X^{1/2}$ is computed with a structure preserving algorithm such as those derived in [11], $\text{cond}(\text{sqrt}, X)$ can be replaced by $\text{cond}_{\mathbb{G}_M}(\text{sqrt}, X)$ in (4.1), yielding a sharper upper bound. This is illustrated in Figure 4 for symplectic and pseudo-orthogonal matrices. To obtain the computed square root $\widehat{A}$, we use the structure preserving and cubically converging iteration

$$Y_{k+1} = \frac{1}{3}Y_k\big[I + 8\big(I + 3Z_kY_k\big)^{-1}\big], \qquad Y_0 = X, \tag{4.2a}$$

$$Z_{k+1} = \frac{1}{3}\big[I + 8\big(I + 3Z_kY_k\big)^{-1}\big]Z_k, \qquad Z_0 = I, \tag{4.2b}$$



(a) Real pseudo-orthogonal matrices with $p = q = 5$.

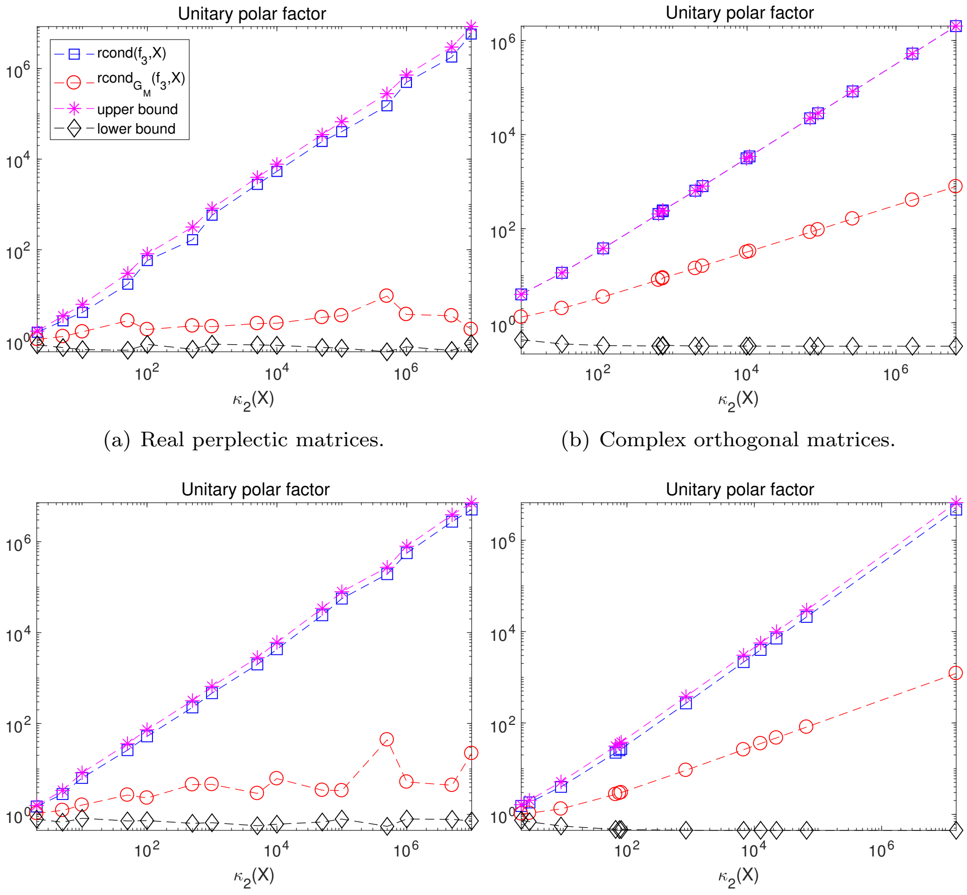
(b) Real pseudo-orthogonal matrices with $p = 1$ and $q = 9$.

(c) Complex pseudo-orthogonal matrices with $p = q = 5$.

(d) Pseudo-unitary matrices with $p = q = 5$.

FIG. 2. *Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of 10×10 randomly generated real pseudo-orthogonal matrices with $p = q = 5$ in plot* (a)*, real pseudo-orthogonal matrices with $p = 1$, $q = 9$ in plot* (b)*, complex pseudo-orthogonal matrices with $p = q = 5$ in plot* (c)*, and pseudo-unitary matrices with $p = q = 5$ in plot* (d)*, with increasing condition number $\kappa_2(X)$.*

where $Y_k, Z_k \in \mathbb{G}_M$ and $Y_k \to X^{1/2}$ [11, sect. 6]. For the relative error $\mathrm{err}(\widehat{A})$ in (4.1), we use `funm_x` from Higham's Matrix Function Toolbox [8] to compute $X^{1/2}$ in extended precision. The relative errors are plotted as "$*$" in Figure 4, and the test matrices are sorted with increasing values of $\mathrm{err}(\widehat{A})$. We compare these relative errors to the unstructured bound in (4.1). The structured bounds in Figure 4 correspond to (4.1) with $\mathrm{cond}(\mathrm{sqrt}, X)$ replaced with $\mathrm{cond}_{\mathbb{G}_M}(\mathrm{sqrt}, X)$, and the approximate bound corresponds to (4.1) with $\mathrm{cond}(\mathrm{sqrt}, X)$ replaced with the lower bound on $\mathrm{cond}_{\mathbb{G}_M}(\mathrm{sqrt}, X)$ displayed in (3.26) or (3.28) (so the approximate bound is not a strict bound). The plot on the right-hand side shows that even when $\mathrm{lb\_cond}(f, X)$ is an order of magnitude smaller than the relative structured condition number $\mathrm{rcond}_{\mathrm{struc}}(\mathrm{sqrt}, X)$ (which happens for pseudo-unitary matrices with $p \neq q$), the product of the lower bound on the relative structured condition number times the relative backward error, i.e.,



(a) Real perplectic matrices.

(b) Complex orthogonal matrices.

(c) Pseudo-unitary matrices with $p = q = 5$.

(d) Conjugate symplectic matrices.

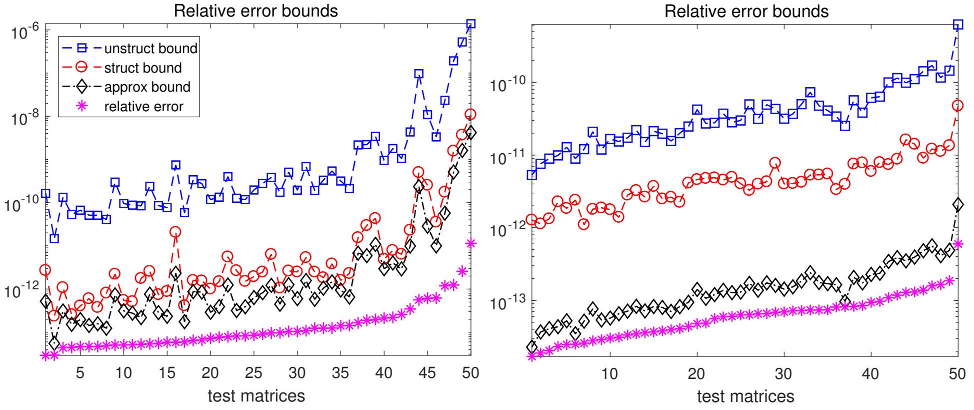
FIG. 3. *Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the unitary polar factor of* 10×10 *randomly generated real perplectic matrices* (a)*, complex orthogonal matrices* (b)*, pseudo-unitary matrices with* $p = q = 5$ (c)*, and conjugate symplectic matrices* (d)*, with increasing condition number* $\kappa_2(X)$*.*

$\mathrm{lb\_cond}(f, X) \times \|\widehat{A}^2 - X\|_F / \|X\|_F$, offers a good approximation of $\mathrm{err}(\widehat{A})$.

5. Concluding remark. We emphasize that in this work we have focused on manifolds embedded in a larger Euclidean space, which is the natural theoretical setting for algorithms working on the entries of a structured matrix. Several, but not all, algorithms in numerical linear algebra are in this category. A common alternative approach is to work via the atlas of a manifold, which is, for example, a standard paradigm to represent the manifold of semiseparable matrices. Extending the theory presented in this paper to include the study of computational problems based on parametrizations of manifolds, a setting where the condition number now also depends on the choice of an atlas, is an interesting line of research left for future work.

Appendix A. The following result is needed in section 3.2 and is of potential interest, per se.

LEMMA A.1. *For any* $A, B \in \mathbb{R}^{m\times n}$*, let* $C = A + \mathrm{i}B \in \mathbb{C}^{m\times n}$ *and* $C_{\mathbb{R}} = \left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]$*.*



(a) Real symplectic X with $\kappa_2(X) = 10^5$.

(b) Pseudo-unitary X with $p = 1$, $q = 9$, and $\kappa_2(X) = 10^5$.

FIG. 4. *Bounds on the relative error* $\mathrm{err}(\widehat{A}) = \|\widehat{A} - X^{1/2}\|_F/\|X^{1/2}\|_F$ *for* 50 *randomly generated* 10×10 *symplectic matrices X in plot* (a) *and for* 50 *randomly generated* 10×10 *pseudo-orthogonal matrices X in plot* (b)*, where $\widehat{A}$ is an approximate square root of X computed by the structure preserving iteration* (4.2).

Then $\|C_\mathbb{R}\|_2 \le \|C\|_2 \le \sqrt{2}\|C_\mathbb{R}\|_2$ *and the inequalities are tight.*

Moreover, suppose further that $\mathrm{rank}(C) = n \le m$. *Then* $\|C_\mathbb{R}^+\|_2 \le \|C^+\|_2$ *and the inequality is tight. However, for any* $M > 0$ *there exist* $A, B \in \mathbb{R}^{m\times n}$ *such that* $\mathrm{rank}(C) = n \le m$ *but* $\|C^+\| > M\|C_\mathbb{R}^+\|$.

Proof. Let us sort the singular values in nonincreasing order and let $\sigma_i(X)$ denote the ith singular value of X. By the same argument we used to derive (2.9), observe that the singular values of

$$C_{\mathbb{R}\mathbb{R}} := \begin{bmatrix} A & -B \\ B & A \end{bmatrix}$$

are the same as those of C, repeated twice (and reordered). Hence, by standard inequalities on singular values of submatrices that can be proved using Weyl's theorem on eigenvalues of sum of Hermitian matrices (see, e.g., [12, Chap. 3]), we conclude that, for any $i = 1, \dots, 2\min(m, n) - n$,

$$\sigma_{i+n}(C_{\mathbb{R}\mathbb{R}}) \le \sigma_i(C_\mathbb{R}) \le \sigma_i(C_{\mathbb{R}\mathbb{R}}).$$

This in particular implies that $\|C_\mathbb{R}\|_2 \le \|C\|_2$. Noting that $\mathrm{rank}(C_{\mathbb{R}\mathbb{R}}) = 2\,\mathrm{rank}(C) \le 2\,\mathrm{rank}(\mathbb{C}_\mathbb{R})$, if $m \ge n$ and $\mathrm{rank}(C) = n$, then $\mathrm{rank} C_{\mathbb{R}\mathbb{R}} = 2n$ and $\|C_\mathbb{R}^+\|_2 \le \|C^+\|_2$. To see that this equality can be achieved, it suffices to take scalars, i.e., $m = n = 1$.

Finally, we have $\|C\|_2 = \|\begin{bmatrix} I & \mathrm{i}I \end{bmatrix} C_\mathbb{R}\|_2 \le \sqrt{2}\|C_\mathbb{R}\|_2$ with equality achieved, for example, by picking $A = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$, $B = I_2$. The same choice of A, B shows that the statement is tight in two additional senses: (1) we cannot relax the assumption $\mathrm{rank}(C) = n$ in the pseudoinverse norm bound, and (2) as can be seen by considering a (generic) small perturbation of the example above, such that the smallest singular value of C becomes arbitrarily small, but nonzero, the ratio $\|C^+\|_2/\|C_\mathbb{R}^+\|_2$ is not bounded above, even if one requires $\mathrm{rank}(C) = n$. $\square$

Acknowledgments. We wish to thank Fabio Di Benedetto, Froilán Dopico, Endre Süli, and Nick Vannieuwenhoven for their insightful comments and questions after talks related to this research: they led to the addition of clarifying remarks to this paper. We are also grateful to Nick Higham for reading a preliminary draft and giving various suggestions. We thank the referees for carefully reading the paper and providing several suggestions to improve the presentation.

REFERENCES

[1] P. Bürgisser and F. Cucker, *Condition. The Geometry of Numerical Algorithms*, Springer-Verlag, Berlin, 2013.

[2] P. I. Davies, *Structured conditioning of matrix functions*, Electron. J. Linear Algebra, 11 (2004), pp. 132–161.

[3] J.-P. Dedieu, *Approximate solutions of numerical problems, condition number analysis and condition number theorems*, in The Mathematics of Numerical Analysis, Lectures in Appl. Math. 32, S. S. J. Renegar and M. Shub, eds., American Mathematical Society, Providence, RI, 1996, pp. 263–283.

[4] H. Fassbender and K. D. Ikramov, *Several observations on symplectic, Hamiltonian, and skew-Hamiltonian matrices*, Linear Algebra Appl., 400 (2005), pp. 15–29.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.

[6] J. B. Hawkins and A. Ben-Israel, *On generalized matrix functions*, Linear and Multilinear Algebra, 1 (1973), pp. 163–171.

[7] H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear and Multilinear Algebra, 9 (1981), pp. 271–288.

[8] N. J. Higham, *The Matrix Function Toolbox*, http://www.maths.manchester.ac.uk/~higham/mftoolbox.

[9] N. J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174, https://doi.org/10.1137/0907079.

[10] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717778.

[11] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Functions preserving matrix groups and iterations for the matrix square root*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 849–877, https://doi.org/10.1137/S0895479804442218.

[12] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[13] D. P. Jagger, *MATLAB Toolbox for Classical Matrix Groups*, M.Sc. Thesis, University of Manchester, Manchester, England, 2003.

[14] D. R. Jensen, *Minimal properties of Moore-Penrose inverses*, Linear Algebra Appl., 196 (1994), pp. 175–182.

[15] C. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209, https://doi.org/10.1137/0610014.

[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, London, 1985.

[17] J. M. Lee, *Introduction to Smooth Manifolds*, Springer-Verlag, New York, 2012.

[18] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured factorizations in scalar product spaces*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 821–850, https://doi.org/10.1137/040619363.

[19] V. Noferini, *A formula for the Fréchet derivative of a generalized matrix function*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 434–457, https://doi.org/10.1137/16M1072851.

[20] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310, https://doi.org/10.1137/0703023.

[21] K. Tapp, *Matrix Groups for Undergraduates*, American Mathematical Society, Providence, RI, 2005.

[22] H. Xu, *An SVD-like matrix decomposition and its applications*, Linear Algebra Appl., 368 (2003), pp. 1–24.