SIAM J. Matrix Anal. Appl.
Vol. 31, No. 5, pp. 2580–2602

THE CONDITION METRIC IN THE SPACE OF RECTANGULAR FULL RANK MATRICES[*]

PAOLA BOITO[†] AND JEAN-PIERRE DEDIEU[†]

Abstract. The condition metric in spaces of polynomial systems has been introduced and studied in a series of papers by Beltrán, Dedieu, Malajovich, and Shub. The interest of this metric comes from the fact that the associated geodesics avoid ill-conditioned problems and are a useful tool to improve classical complexity bounds for Bézout's theorem. The linear case is examined here: using nonsmooth nonconvex analysis techniques, we study the properties of condition geodesics in the space of full rank, real, or complex rectangular matrices. Our main results include an existence theorem for the boundary problem, a differential inclusion for such geodesics based on Clarke's generalized gradients, regularity properties, and a detailed description of a few particular cases: diagonal and unitary matrices. Moreover, we study condition geodesics from a numerical viewpoint, and we develop an effective algorithm that allows us to compute geodesics with given endpoints and helps to illustrate theoretical results and formulate new conjectures.

Key words. matrices, condition metric, generalized gradient, BFGS method

AMS subject classifications. 15A15, 65F35, 15A12

DOI. 10.1137/08073874X

1. Introduction.

The interest of considering the condition metric in certain spaces of matrices or, more generally, in spaces of polynomial systems, comes from recent papers by Shub [19], Beltrán and Shub [3], [4], and Beltrán, Dedieu, Malajovich, and Shub [2], where these authors follow geodesics for the condition metric in certain solution varieties to improve classical complexity bounds for Bézout's theorem. These geodesics are designed to avoid ill-conditioned problems.

In this paper we investigate in more detail the linear case. Let us be more precise.

Let $\mathbb{GL}_{m,n}$ $(n \leq m)$ be the space of matrices $A \in \mathbb{K}^{m\times n}$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, with $\operatorname{rank} A = n$. The singular values of such matrices are denoted in decreasing order as follows:

$$\sigma_1(A) \geq \cdots \geq \sigma_{n-1}(A) \geq \sigma_n(A) > 0.$$

This set is equipped with the Frobenius inner product given by

$$\langle M, N\rangle_F = \text{trace }(N^* M) = \sum_{i,j} m_{ij}\overline{n_{ij}},$$

where N^* denotes the adjoint matrix of N. The length of an absolutely continuous curve $X(t)$, $a \leq t \leq b$, is given by the integral

$$L(X, a, b) = \int_a^b \left\| \dot{X}(s) \right\|_F ds,$$

where $\dot{X}$ denotes the derivative with respect to t.

[*] Received by the editors October 21, 2008; accepted for publication (in revised form) by A. S. Lewis June 23, 2010; published electronically September 14, 2010. This work was supported by the ANR Gecko project.

http://www.siam.org/journals/simax/31-5/73874.html

[†] Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse cedex 09, France (paola.boito@unilim.fr, jean-pierre.dedieu@math.univ-toulouse.fr).

The problem considered here is to connect two matrices A, $B \in \mathbb{GL}_{m,n}$ by a path $X(t) \in \mathbb{GL}_{m,n}$, $a \leq t \leq b$, which is as short as possible and also stays as far as possible from the boundary of $\mathbb{GL}_{m,n}$. This boundary consists of matrices with rank less than n, and the Frobenius distance of a matrix A from this boundary is equal to $\sigma_n(A)$, its smallest singular value (see the classical reference [11] for a discussion of matrix rank and distance from singularity via singular value decomposition). For this reason, instead of the length $L(X, a, b)$ we consider the *condition length* defined by

$$L_\kappa(X, a, b) = \int_a^b \left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} ds.$$

Such a path, if any, is called a *minimizing condition path* with given endpoints or a *minimizing condition geodesic* when $\|\dot{X}(s)\|_F \sigma_n(X(s))^{-1}$ is constant almost everywhere. It is *parametrized by arc length* when

$$\left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} = 1 \text{ a.e.}$$

A similar problem arises in hyperbolic geometry: instead of $\mathbb{GL}_{m,n}$ we consider the Poincaré model of hyperbolic space $\mathbb{H}_n = \{x \in \mathbb{R}^n \ : \ x_n > 0\}$, and the role of the condition length is played by the hyperbolic length $\int_a^b \|\dot{x}(s)\| \, x_n(s)^{-1} ds$. In this latter case the boundary to avoid is the hyperplane $x_n = 0$.

Given a matrix $A \in \mathbb{GL}_{m,n}$, define the *condition Riemannian structure* in $\mathbb{GL}_{m,n}$ as

$$\langle M, N \rangle_A = \langle M, N \rangle_F \, \sigma_n(A)^{-2}$$

for any $A \in \mathbb{GL}_{m,n}$ and $M, N \in \mathbb{K}^{m \times n}$. The corresponding norm is

$$\|M\|_A^2 = \langle M, M \rangle_A .$$

Our concept of condition length is related to the condition metric by

$$L_\kappa(X, a, b) = \int_a^b \left\| \dot{X}(s) \right\|_{X(s)} ds.$$

Unfortunately, we cannot use the usual tools of Riemannian geometry to study our problem because the condition metric defined above is not smooth. The maps $\sigma_n(A)$ and $\sigma_n(A)^{-1}$ are locally Lipschitz in $\mathbb{GL}_{m,n}$, but they are not necessarily smooth: they are not differentiable when $\sigma_{n-1}(A) = \sigma_n(A)$, that is, when $\sigma_n(A)$ has a multiplicity greater than 1. For this reason we qualify the condition metric as Lipschitz–Riemannian. Because of this lack of smoothness, the techniques used in this paper to prove our theorems are taken from linear algebra, calculus of variations, and nonsmooth, nonconvex analysis.

Our main results are the following.

1. Between two matrices A and B in the same connected component of $\mathbb{GL}_{m,n}$, there is always a minimizing condition path (Theorem 2.2).
2. These condition paths satisfy an Euler–Lagrange differential inclusion (Theorem 3.1).
3. Any condition geodesic is of class C^1 with a Lipschitz first derivative (Theorem 3.7).

4. When A and B are Stiefel matrices (i.e., when $A^*A = B^*B = I_n$), any geodesic in the Stiefel manifold for the Frobenius metric is also a geodesic in $\mathbb{GL}_{m,n}$ for the condition metric (Theorem 4.1).
5. In Theorem 5.1 we describe the condition geodesics contained in the space of positive diagonal matrices: they are union of segments of lines or arcs of circles, like in the case of hyperbolic geometry.
6. In the last section we investigate the numerical computation of the condition paths, and we give some examples. We carry out this computation by solving an optimization problem via an approximation process and the BFGS method.

2. Length in the condition metric. Let us denote as $W^{1,1}\left([a,b], \mathbb{K}^{m\times n}\right)$ the set of absolutely continuous paths $X : [a,b] \to \mathbb{K}^{m\times n}$. Every path $X(t)$ in $W^{1,1}$ is almost everywhere differentiable; its derivative $\dot{X}(t)$ is an integrable function, and

$$X(t) = X(a) + \int_a^t \dot{X}(s)ds$$

for every $t \in [a,b]$. The condition metric defined above is equal to $L_\kappa(X,a,b) = \infty$ when $\operatorname{rank} X(s) < n$ for some s because, in that case, the integral defining $L_\kappa(X,a,b)$ diverges.

Remark 1.

- The condition length of a path $X(t) \in \mathbb{GL}_{m,n}$ is invariant under any change of parametrization: it is a geometric concept.
- Any path $X(t) \in \mathbb{GL}_{m,n}$ may be parametrized by the (condition) arc length

$$s : [a,b] \to [0, L_\kappa(X,a,b)], \quad s(t) = \int_a^t \left\|\dot{X}(\tau)\right\|_F \sigma_n(X(\tau))^{-1}d\tau.$$

In that case we have

$$\left\|\dot{X}(s)\right\|_F \sigma_n(X(s))^{-1} = 1$$

for almost every s.

The *condition distance* between two full rank matrices X_0 and X_1 is defined as

$$d_\kappa(X_0, X_1) = \inf L_\kappa(X,a,b),$$

where the infimum is taken on the set of paths $X \in W^{1,1}\left([a,b], \mathbb{K}^{m\times n}\right)$ such that $X(a) = X_0$ and $X(b) = X_1$. The space $\mathbb{GL}_{m,n}$ equipped with the condition metric is a length space à la Gromov [12]. Its properties are summarized in Theorem 2.2 below. We begin by the following classical lemma (see, e.g., [1], [9]):

LEMMA 2.1. *$\mathbb{GL}_{m,n}(\mathbb{K})$ is connected except when $\mathbb{K} = \mathbb{R}$ and $m = n$. In that case, $\mathbb{GL}_n(\mathbb{R})$ has two connected components characterized by the sign of the determinant.*

THEOREM 2.2. *Let Ω be a connected component of $\mathbb{GL}_{m,n}$. The condition distance is a distance in Ω. This space is complete, locally compact, and the infimum defining d_κ is a minimum.*

The proof of this theorem is given by a series of 10 lemmas. The first one is classical.

LEMMA 2.3. *The generalized (or Moore–Penrose) inverse of $A \in \mathbb{GL}_{m,n}$ is given by $A^\dagger = (A^*A)^{-1}A^*$. Its spectral norm is equal to*

$$\left\| A^\dagger \right\|_2 = \sigma_n(A)^{-1}.$$

LEMMA 2.4. *Given A and $B \in \mathbb{GL}_{m,n}$ with $\sigma_n(A)^{-1} d_F(A,B) < 1$, the following inequality holds:*

$$\sigma_n(B)^{-1} \leq \frac{\sigma_n(A)^{-1}}{1 - \sigma_n(A)^{-1} d_F(A,B)}.$$

Proof. By Wedin's theorem (see [20, Theorem 3.9])

$$\left\| B^\dagger - A^\dagger \right\|_F \leq \left\| A^\dagger \right\|_2 \left\| B^\dagger \right\|_2 \left\| A - B \right\|_F$$

so that

$$\left\| B^\dagger \right\|_2 - \left\| A^\dagger \right\|_2 \leq \left\| B^\dagger - A^\dagger \right\|_2 \leq \left\| B^\dagger - A^\dagger \right\|_F \leq \left\| A^\dagger \right\|_2 \left\| B^\dagger \right\|_2 \left\| A - B \right\|_F .$$

Thus

$$\left\| B^\dagger \right\|_2 \left(1 - \left\| A^\dagger \right\|_2 \left\| A - B \right\|_F\right) \leq \left\| A^\dagger \right\|_2 ,$$

and, using Lemma 2.3 and the hypothesis, we are done. □

LEMMA 2.5. *For every $\varepsilon > 0$ there exists $C_2 > 0$ such that, for any A and $B \in \mathbb{GL}_{m,n}$,*

$$\sigma_n(A)^{-1} d_F(A,B) \leq C_2 \Longrightarrow \sigma_n(B)^{-1} \leq (1+\varepsilon)\sigma_n(A)^{-1}.$$

One may choose $C_2 = \varepsilon/(1+\varepsilon)$.

Proof. The proof is an easy consequence of Lemma 2.4. □

LEMMA 2.6. *For every $\varepsilon > 0$ there exists $C_3 > 0$ such that, for any A and $B \in \mathbb{GL}_n$,*

$$\sigma_n(A)^{-1} d_F(A,B) \leq C_3 \Longrightarrow \frac{\sigma_n(A)^{-1}}{1+\varepsilon} \leq \sigma_n(B)^{-1} \leq (1+\varepsilon)\sigma_n(A)^{-1}.$$

One may choose $C_3 = \varepsilon/(1+\varepsilon)^2$.

Proof. Observe that $C_3 = C_2/(1+\varepsilon)$ so that, by Lemma 2.5, the fact that $\sigma_n(A)^{-1} d_F(A,B) \leq C_3$ implies $\sigma_n(B)^{-1} \leq (1+\varepsilon)\sigma_n(A)^{-1}$. This proves the second inequality. Moreover, multiplying $\sigma_n(B)^{-1} \leq (1+\varepsilon)\sigma_n(A)^{-1}$ by $d_F(A,B)$ and applying again Lemma 2.5, we have $\sigma_n(B)^{-1} d_F(A,B) \leq (1+\varepsilon)\sigma_n(A)^{-1} d_F(A,B) \leq (1+\varepsilon)C_3 = C_2$. Applying Lemma 2.5 with the roles of A and B reversed yields $\sigma_n(A)^{-1} \leq (1+\varepsilon)\sigma_n(B)^{-1}$, which proves the first inequality. □

LEMMA 2.7. *Given $\varepsilon > 0$, C_3 as in Lemma 2.6, and $X \in W^{1,1}([a,b], \mathbb{GL}_{m,n})$, define a sequence $(s_i)_{1 \leq i \leq k}$ as follows:*

$$\begin{cases} s_0 = a, \\ s_i \text{ such that } \int_{s_{i-1}}^{s_i} \|\dot{X}(s)\|_F ds = C_3 \sigma_n(X(s_{i-1})) \\ \quad \text{when } \int_{s_{i-1}}^{b} \|\dot{X}(s)\|_F ds \geq C_3 \sigma_n(X(s_{i-1})), \\ \text{else } s_i = s_k = b. \end{cases}$$

Then $s_k = b$ *for*

$$k \leq \frac{1+\varepsilon}{C_3} L_\kappa(X,a,b) + 1.$$

Moreover,

$$\sigma_n(X(b))^{-1} \leq (1+\varepsilon)^k \sigma_n(X(a))^{-1}.$$

Proof. For any s with $s_{i-1} \leq s \leq s_i$ and $i \leq k-1$, we have

$$\sigma_n(X(s_{i-1}))^{-1} d_F(X(s), X(s_{i-1})) \leq \sigma_n(X(s_{i-1}))^{-1} \int_{s_{i-1}}^{s_i} \left\| \dot{X}(s) \right\|_F ds = C_3$$

so that, by Lemma 2.6, we have $\sigma_n(X(s))^{-1} \geq \sigma_n(X(s_{i-1}))^{-1}/(1+\varepsilon)$, and

$$\int_{s_{i-1}}^{s_i} \left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} ds \geq \int_{s_{i-1}}^{s_i} \left\| \dot{X}(s) \right\|_F \frac{\sigma_n(X(s_{i-1}))^{-1}}{1+\varepsilon} ds = \frac{C_3}{1+\varepsilon}.$$

Thus we obtain

$$\int_a^b \left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} ds \geq \sum_{i=1}^{k-1} \int_{s_{i-1}}^{s_i} \left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} ds \geq \frac{(k-1)C_3}{1+\varepsilon}$$

so that

$$k - 1 \leq \frac{1+\varepsilon}{C_3} \int_a^b \left\| \dot{X}(s) \right\|_F \sigma_n(X(s))^{-1} ds = \frac{1+\varepsilon}{C_3} L_\kappa(X,a,b).$$

In order to prove the last inequality $\sigma_n(X(b))^{-1} \leq (1+\varepsilon)^k \sigma_n(X(a))^{-1}$, we use a similar argument:

- when $i \leq k-1$, one has

$$\sigma_n(X(s_{i-1}))^{-1} d_F(X(s_{i-1}), X(s_i)) \leq \sigma_n(X(s_{i-1}))^{-1} \int_{s_{i-1}}^{s_i} \left\| \dot{X}(s) \right\|_F ds = C_3;$$

- whereas, when $i = k$, the definition of k gives

$$\sigma_n(X(s_{k-1}))^{-1} \int_{s_{k-1}}^{s_k} \left\| \dot{X}(s) \right\|_F ds < C_3.$$

In both cases, Lemma 2.6 yields the inequality

$$\sigma_n(X(s_i))^{-1} \leq (1+\varepsilon) \sigma_n(X(s_{i-1}))^{-1}.$$

The proof can then be completed by induction. □

LEMMA 2.8. *For any path* $X \in W^{1,1}([a,b], \mathbb{GL}_n)$ *with length* $L_\kappa(X,a,b)$ *in the condition metric, we have*

$$\sigma_n(X(b))^{-1} \leq e^{L_\kappa(X,a,b)} \sigma_n(X(a))^{-1}.$$

Proof. By Lemma 2.7 we have

$$\sigma_n(X(b))^{-1} \leq (1+\varepsilon)^{\frac{1+\varepsilon}{C_3} L_\kappa(X,a,b)+1} \sigma_n(X(a))^{-1},$$

and, according to Lemma 2.6, we can take $(1+\varepsilon)^{\frac{1+\varepsilon}{C_3}} = (1+\varepsilon)^{\frac{(1+\varepsilon)^3}{\varepsilon}}$. Since this last expression is increasing and tends to e when ε tends to 0, we obtain

$$\sigma_n(X(b))^{-1} \leq e^{L_\kappa(X,a,b)} \sigma_n(X(a))^{-1}. \qquad \square$$

LEMMA 2.9. *For any path $X \in W^{1,1}([a,b], \mathbb{GL}_n)$ with length $L_\kappa(X,a,b)$ in the condition metric, and for any $t \in [a,b]$ we have*

$$e^{-L_\kappa(X,t,b)} \sigma_n(X(b))^{-1} \leq \sigma_n(X(t))^{-1} \leq e^{L_\kappa(X,a,t)} \sigma_n(X(a))^{-1}.$$

Proof. The proof is an easy consequence of Lemma 2.8. $\square$

LEMMA 2.10. *d_κ is a distance in any connected component Ω of $\mathbb{GL}_{m,n}$.*

Proof. The only nontrivial fact to prove is

$$d_\kappa(X_0, X_1) = 0 \Longrightarrow X_0 = X_1.$$

For any $\varepsilon \in]0,1]$ let $X_\varepsilon \in W^{1,1}([a,b], \Omega)$ be such that $X_\varepsilon(a) = X_0$, $X_\varepsilon(b) = X_1$, and $L_\kappa(X_\varepsilon, a, b) \leq d_\kappa(X_0, X_1) + \varepsilon = \varepsilon$. Since all such paths have a condition length less than 1, we get, by Lemma 2.9,

$$e^{-1} \sigma_n(X_1)^{-1} \int_a^b \left\| \dot{X}_\varepsilon(s) \right\|_F ds \leq L_\kappa(X_\varepsilon, a, b) \leq \varepsilon$$

so that

$$e^{-1} \sigma_n(X_1)^{-1} \left\| X_1 - X_0 \right\|_F \leq e^{-1} \sigma_n(X_1)^{-1} \int_a^b \left\| \dot{X}_\varepsilon(s) \right\|_F ds \leq \varepsilon.$$

Thus $X_1 = X_0$ and we are done. $\square$

LEMMA 2.11. *Any connected component Ω of $\mathbb{GL}_{m,n}$ is complete and locally compact for the condition distance.*

Proof. Let (X_p) be a Cauchy sequence in Ω for the condition distance. Then $d_\kappa(X_p, X_0) \leq L$ for a certain $L > 0$, and, by Lemma 2.8,

$$\sigma_n(X_p)^{-1} \leq e^L \sigma_n(X_0)^{-1}$$

for any p. Since $\sigma_n(X_p)^{-1}$ is the inverse of the Frobenius distance of X_p from the set

$$\Sigma_{m,n} = \left\{ A \in \mathbb{K}^{m \times n} \ : \ \operatorname{rank} A < n \right\},$$

the sequence (X_p) stays inside a compact set in Ω; therefore, there exists a subsequence $(X_q) \subset (X_p)$ which converges in the usual Frobenius distance. Now, observe that Frobenius convergence in such a compact set implies condition convergence. Indeed, take X_a and X_b in this compact set, and let $X(t)$, $t \in [a,b]$, be the segment of line that joins X_a and X_b. Then, by Lemma 2.8, we have

$$\begin{aligned} \sigma_n(X(t))^{-1} &\leq e^L \sigma_n(X_a)^{-1} \\ \|\dot{X}(t)\|_F \sigma_n(X(t))^{-1} &\leq e^L \sigma_n(X_a)^{-1} \|\dot{X}(t)\|_F \\ d_\kappa(X_a, X_b) &\leq \int_a^b \|\dot{X}(t)\|_F \sigma_n(X(t))^{-1} dt \leq \\ &\leq e^L \sigma_n(X_a)^{-1} \int_a^b \|\dot{X}(t)\|_F dt = e^L \sigma_n(X_a)^{-1} d_F(X_a, X_b), \end{aligned}$$

which proves the convergence of (X_q) with respect to the condition distance. The convergence of (X_p) can be deduced from the fact that the sequence is Cauchy.

A similar argument shows that the closed balls $B_K(X, r)$ are compact with respect to the condition distance. ∎

LEMMA 2.12. *For every X_0 and $X_1 \in \Omega$, the infimum defining the condition distance $d_\kappa(X_0, X_1)$ is a minimum: there exists a path $X \in W^{1,1}([a,b], \Omega)$ such that $X(a) = X_0$, $X(b) = X_1$, and $d_\kappa(X_0, X_1) = L_\kappa(X, a, b)$.*

Proof. Notice that, once it is proved that each connected component of $\mathbb{GL}_{m,n}$ is complete and locally compact for the condition distance (Lemma 2.11), Lemma 2.12 can be seen as a consequence of the Hopf–Rinow theorem (see, e.g., [12, section 1.10]). For the sake of completeness, however, we give a detailed proof.

For any $\varepsilon \in]0, 1[$ there exists a path $X_\varepsilon \in W^{1,1}([a,b], \Omega)$ such that $X_\varepsilon(a) = X_0$, $X_\varepsilon(b) = X_1$, and $d_\kappa(X_0, X_1) \le L_\kappa(X_\varepsilon, a, b) \le d_\kappa(X_0, X_1) + \varepsilon$. Let us denote $L_\varepsilon = L_\kappa(X_\varepsilon, a, b)$. We suppose that X_ε is nonconstant on any nonvoid open subinterval contained in $[a, b]$, and we parametrize X_ε by arc length

$$s : [a, b] \to [0, L_\varepsilon], \quad s(t) = \int_a^t \left\| \dot{X}_\varepsilon(\tau) \right\|_F \sigma_n(X_\varepsilon(\tau))^{-1} d\tau.$$

Set $d_\kappa(X_0, X_1) + 1 = B$. Since $L_\varepsilon \le d_\kappa(X_0, X_1) + \varepsilon \le B$, from Lemma 2.9 we obtain $e^{-B}\sigma_n(X_1)^{-1} \le \sigma_n(X_\varepsilon(s))^{-1}$ so that

$$e^{-B}\sigma_n(X_1)^{-1} \int_{s_1}^{s_2} \left\| \dot{X}_\varepsilon(s) \right\|_F ds \le \int_{s_1}^{s_2} \left\| \dot{X}_\varepsilon(s) \right\|_F \sigma_n(X_\varepsilon(s)) ds = s_2 - s_1$$

for every s_1 and s_2 such that $0 \le s_1 \le s_2 \le L_\varepsilon$. This gives

$$\int_{s_1}^{s_2} \left\| \dot{X}_\varepsilon(s) \right\|_F ds \le e^B \sigma_n(X_1)(s_2 - s_1).$$

We may now extend the definition of X_ε to the interval $[0, B]$ (independent of ε) by taking $X_\varepsilon(s) = X_1$ for every $L_\varepsilon \le s \le B$. The corresponding derivative $\dot{X}_\varepsilon(s)$ is extended by 0, and the previous inequality is still valid for any $0 \le s_1 \le s_2 \le B$. Now, Dunford–Pettis theorem (see, for example, [7, Corollary IV-8-11]) shows that the set of $\dot{X}_\varepsilon$, $\varepsilon \in]0, 1[$, is sequentially compact in $L^1([0, B], \mathbb{K}^{m\times n})$ for the topology $\sigma(L^1, L^\infty)$. This proves the existence of a limit point $Y \in L^1([0, B], \mathbb{K}^{m\times n})$ for $\dot{X}_\varepsilon$ as ε tends to 0. Let us define X as

$$X(t) = X_0 + \int_0^t Y(\tau) d\tau, \quad 0 \le t \le B.$$

This path is absolutely continuous (i.e., $X \in W^{1,1}([0, B], \mathbb{K}^{m\times n})$), its derivative $\dot{X}$ exists almost everywhere, and it is equal almost everywhere to Y. Thus we write $Y = \dot{X}$. Moreover, $X(t) = \lim_{\varepsilon\to 0} X_\varepsilon(t)$ for every $t \in [0, B]$ so that $X(0) = X_0$ and $X(B) = X_1$. The lower bound $e^{-B}\sigma_n(X_1)^{-1} \le \sigma_n(X_\varepsilon(s))^{-1}$ shows that X is in fact contained in $\mathbb{GL}_{m,n}$, so the condition distance $L_\kappa(X, 0, B)$ is well defined. Since the length functional $X \in L^1([0, B], \mathbb{GL}_{m,n}) \to L_\kappa(X, 0, B)$ is weakly lower semicontinuous as a consequence of Fatou's lemma (see, e.g., [8]), we have $L_\kappa(X, 0, B) \le \liminf_{\varepsilon\to 0} L_\kappa(X_\varepsilon, 0, B)$ so that $L_\kappa(X, 0, B) = d_\kappa(X_0, X_1)$, and we are done. ∎

3. The Euler–Lagrange differential inclusion for geodesics.

3.1. The smooth case.

The usual Euler–Lagrange equation for the solutions of the Bolza problem in the calculus of variations

$$\min_{\substack{X(a)=A\\ X(b)=B}} \int_a^b L(X(t), \dot{X}(t))dt \tag{3.1}$$

is given by

$$\begin{cases} -\dfrac{d}{dt}\dfrac{\partial L}{\partial \dot{X}} + \dfrac{\partial L}{\partial X} = 0, \\ X(a) = A, \ X(b) = B, \end{cases}$$

which is a second order differential equation with boundary conditions. In our case $L(X, \dot{X}) = \|\dot{X}\|_F \sigma_n(X)^{-1}$ is a smooth convex function in the variable $\dot{X}$, but it is nonsmooth in the variable X, and we cannot apply the usual Euler–Lagrange formalism. However, using nonsmooth analysis techniques, we obtain a differential inclusion that plays a similar role.

3.2. Generalized gradients.

Let $f : U \subset \mathbb{E} \to \mathbb{R}$ be a locally Lipschitz function defined on the open subset U of the Euclidean space $\mathbb{E}$. The generalized directional derivative in Clarke's sense of f at $x \in U$ in the direction $d \in \mathbb{E}$ is defined as

$$f^o(x, d) = \limsup_{\substack{y \to x\\ t \to 0_+}} \frac{f(y + td) - f(y)}{t},$$

and the generalized gradient of f at x is the nonempty compact subset of $\mathbb{E}$ given by

$$\partial f(x) = \{s \in \mathbb{E} \ : \ \langle s, d\rangle \le f^o(x, d) \text{ for all } d \in \mathbb{E}\}.$$

When $f \in C^1(U)$, the generalized gradient is just the usual one: $\partial f(x) = \{\nabla f(x)\}$. The directional derivative is related to the gradient via the equality

$$f^o(x, d) = \max_{s \in \partial f(x)} \langle s, d\rangle.$$

We say that f is *regular* at x when, for every $d \in \mathbb{E}$, the limit

$$f'(x, d) = \lim_{t \to 0_+} \frac{f(x + td) - f(x)}{t}$$

exists and is equal to $f^o(x, d)$. This is the case when $f \in C^1(U)$ and also when f is convex. In this latter case the map $t > 0 \to (f(x + td) - f(x))/t$ is decreasing, and

$$f'(x, d) = \inf_{t>0} \frac{f(x + td) - f(x)}{t}.$$

This concept of regularity is stable by composition; we will use this fact later. A good reference for this topic is Clarke's book [6].

3.3. The generalized gradient of $\boldsymbol{\sigma_n(X)^{-1}}$. It is given by the following.

Theorem 3.1. *The map $X \in \mathbb{GL}_{m,n} \to \sigma_n(X)^{-1}$ is locally Lipschitz, and*

$$\partial\sigma_n(X)^{-1} = -\sigma_n(X)^{-2}\mathrm{co}\,\{vu^* \ : \ \|u\| = \|v\| = 1, \ Xu = \sigma_n(X)v \text{ and } X^*v = \sigma_n(X)u\}, \tag{3.2}$$

or, equivalently,

$$\partial\sigma_n(X)^{-1} = -\sigma_n(X)^{-3}X\mathrm{co}\left\{uu^* \ : \ \|u\| = 1 \text{ and } (X^*X)u = \sigma_n(X)^2u\right\}, \tag{3.3}$$

where co *denotes the convex envelope.*

Proof. We first have to prove that the map $X \in \mathbb{GL}_{m,n} \to \sigma_n(X)^{-1}$ is locally Lipschitz. This will be done by showing that the map $X \in \mathbb{GL}_{m,n} \to \sigma_n(X)^2$ is locally Lipschitz. For a given Hermitian matrix H, let $\lambda_1(H) \geq \cdots \geq \lambda_n(H)$ denote the ordered eigenvalues of H. Recall that the equality $\sigma_n(X)^2 = \lambda_n(X^*X)$ holds. Moreover (see, [20, section IV.4]), we have $|\lambda_n(X^*X) - \lambda_n(Y^*Y)| \leq \|X^*X - Y^*Y\|_2$. There exist a constant K and a neighborhood $\mathcal{N} \subset \mathbb{GL}_{m,n}$ of X such that $\|X^*X - Y^*Y\|_2 \leq K\|X - Y\|_2$ for every matrix $Y \in \mathcal{N}$. Thus $\sigma_n(X)^2$ is Lipschitz in $\mathcal{N}$.

The existing literature provides formulas for generalized gradients of eigenvalues and singular values; see, e.g., [14] and [16]. Such results can be used to prove (3.2). For instance, observe that $\sigma_n(X)^{-1}$ can be seen as a composition of functions as follows:

$$X \in \mathbb{GL}_{m,n} \overset{\sigma_n}{\to} \sigma_n(X) \overset{(\cdot)^{-1}}{\to} \sigma_n(X)^{-1}.$$

Then (3.2) can be obtained from the formula for Clarke's subdifferential of singular values ([16, Corollary 6.4]) and from the chain rule given in [6, Theorems 2.3.9 and 2.3.10]. The equivalence of (3.2) and (3.3) is proved by setting $v = \sigma_n(X)^{-1}Xu$ so that $Xu = \sigma_n(X)v$ and $X^*v = \sigma_n(X)^{-1}X^*Xu = \sigma_n(X)u$. □

3.4. Generalized Euler–Lagrange equation. For the problem of Bolza described in (3.1), the counterpart of the Euler–Lagrange equation is given by the following result (see [6, Theorem 4.3.3] and [5]).

Theorem 3.2. *Let X solve* (3.1) *in the case in which $L(X, \dot{X})$ is a locally Lipschitz map, and suppose that $\dot{X}$ is essentially bounded. Then there is an absolutely continuous map P such that*

$$\dot{P}(t) \in \partial_X L(X(t), \dot{X}(t)) \text{ and } P(t) \in \partial_{\dot{X}} L(X(t), \dot{X}(t)) \text{ a.e.}$$

In our case $L(X, \dot{X}) = \|\dot{X}\|_F\sigma_n(X)^{-1}$ is smooth in the variable $\dot{X}$ and locally Lipschitz in the variable X. The generalized gradients in the variables X and $\dot{X}$ are given by (we write X and $\dot{X}$ instead of $X(t)$ and $\dot{X}(t)$)

$$\partial_X L(X, \dot{X}) = -\left\|\dot{X}\right\|_F \sigma_n(X)^{-3}X\mathrm{co}\left\{uu^* \ : \ \|u\| = 1 \text{ and } X^*Xu = \sigma_n(X)^2u\right\}$$

obtained from Theorem 3.1, and

$$\partial_{\dot{X}} L(X, \dot{X}) = \left\{ \frac{\dot{X}}{\left\|\dot{X}\right\|_F}\sigma_n(X)^{-1} \right\}$$

because L is smooth in $\dot{X}$. Here we implicitly assume $\dot{X} \neq 0$. Therefore, we have the following.

THEOREM 3.3. *Let $X \in W^{1,1}([a,b],\mathbb{K}^{m\times n})$ be a minimizing condition path in $\mathbb{GL}_{m,n}$ for the condition metric with endpoints A and $B \in \mathbb{GL}_{m,n}$, and suppose that $\dot{X}$ is essentially bounded. Then there exists $P \in W^{1,1}$ such that*

$$P(t) = \frac{\dot{X}(t)}{\left\|\dot{X}(t)\right\|_F}\sigma_n(X(t))^{-1} \tag{3.4}$$

and

$$\dot{P}(t) \in -\left\|\dot{X}(t)\right\|_F \sigma_n(X(t))^{-3}X(t)\,co\left\{u(t)u(t)^*\right\} \tag{3.5}$$

for almost all $a \le t \le b$ and where $u(t)$ is taken in the set of normalized eigenvectors of $(X(t)^*X(t))^{-1}$ for the eigenvalue $\sigma_n(X(t))^{-2}$.

In the case of a minimizing condition geodesic $X(t)$ parametrized by arc length, one has $\|\dot{X}(t)\|_F\sigma_n(X(t))^{-1} = 1$ for almost every t so that $\dot{X}$ is essentially bounded. We obtain the following.

COROLLARY 3.4. *Let $X \in W^{1,1}([0,L],\mathbb{K}^{m\times n})$ be a minimizing condition geodesic in $\mathbb{GL}_{m,n}$ parametrized by arc length, with endpoints A and $B \in \mathbb{GL}_{m,n}$. Then there exists $P \in W^{1,1}$ such that*

$$P(t) = \frac{\dot{X}(t)}{\left\|\dot{X}(t)\right\|_F^2} = \dot{X}(t)\sigma_n(X(t))^{-2} \tag{3.6}$$

and

$$\dot{P}(t) \in -\frac{X(t)}{\left\|\dot{X}(t)\right\|_F^2}co\left\{u(t)u(t)^*\right\} = -X(t)\,co\left\{u(t)u(t)^*\right\}\sigma_n(X(t))^{-2} \tag{3.7}$$

for almost all t.

DEFINITION 3.5. *We call* condition path *any curve $X \in W^{1,1}([a,b],\mathbb{GL}_{m,n})$ satisfying* (3.4) *and* (3.5)*. Such a path is called* condition geodesic *when $\|\dot{X}(t)\|_F\sigma_n(X(t))^{-1}$ is constant almost everywhere. A condition geodesic is parametrized by arc length when $\|\dot{X}(t)\|_F\sigma_n(X(t))^{-1} = 1$ for almost every t.*

Remark 2. In the introduction we have introduced the concept of *minimizing geodesic* as a shortest curve with given endpoints. The main interest of such a definition is to be transferable to a wide range of situations as soon as a reasonable concept of curve length is available; see Gromov [12]. However, such a definition is too restrictive: an arc of great circle on the unit sphere in $\mathbb{R}^3$ is not necessarily a minimizing geodesic! A natural extension is to define a *geodesic* as a curve which is locally a minimizing geodesic. With this definition any arc of great circle is a geodesic on the sphere. Another way to proceed is to define a geodesic, like in classical Riemannian geometry, via a second order differential equation or inclusion (see [10] or any other geometry textbook). This is what we did in our Definition 3.5. A locally minimizing geodesic satisfies such a differential inclusion. The converse is true in a smooth Riemannian manifold but not in a general Lipschitz–Riemannian structure. The case of $\mathbb{GL}_{m,n}$ equipped with the condition structure is still unclear.

COROLLARY 3.6. *Let $t \to X(t)$ be a condition path in $\mathbb{GL}_{m,n}$. Suppose that, for every t, $\sigma_n(X(t))^2$ is an eigenvalue with multiplicity* 1 *of $X(t)^*X(t)$. Then*

$t \to \sigma_n(X(t))^{-1}$ is a smooth function, and the Euler–Lagrange differential inclusion (Theorem 3.3) becomes the following:

$$\ddot{X}(t)\sigma_n(X(t))^2 \left\|\dot{X}(t)\right\|_F^2 - \dot{X}(t)\Re\left\langle \dot{X}(t), \ddot{X}(t)\right\rangle \sigma_n(X(t))^2$$

$$-\dot{X}(t)\Re\left\langle X(t)u(t)u(t)^*, \dot{X}(t)\right\rangle \left\|\dot{X}(t)\right\|_F^2 + X(t)u(t)u(t)^* \left\|\dot{X}(t)\right\|_F^4 = 0,$$

where $\Re$ denotes the real part of a complex number. When X is parametrized by arc length, this equation is equivalent to

$$\ddot{X}(t) - 2\dot{X}(t)\Re\left\langle X(t)u(t)u(t)^*, \dot{X}(t)\right\rangle \sigma_n(X(t))^{-2} + X(t)u(t)u(t)^* = 0.$$

Proof. If $\sigma_n(X(t))^{-1}$ is a smooth function, (3.4) and (3.5) become

$$\frac{d}{dt}\left(\frac{\dot{X}(t)}{\left\|\dot{X}(t)\right\|_F}\sigma_n(X(t))^{-1}\right) = \frac{\ddot{X}(t)}{\left\|\dot{X}(t)\right\|_F}\sigma_n(X(t))^{-1} - \frac{\dot{X}(t)\Re\left\langle \dot{X}(t), \ddot{X}(t)\right\rangle}{\left\|\dot{X}(t)\right\|_F^3}\sigma_n(X(t))^{-1}$$

$$-\frac{\dot{X}(t)}{\left\|\dot{X}(t)\right\|_F}\sigma_n(X(t))^{-3}\Re\left\langle X(t)u(t)u(t)^*, \dot{X}(t)\right\rangle = -\left\|\dot{X}(t)\right\|_F \sigma_n(X(t))^{-3}X(t)u(t)u(t)^*,$$

and this gives the first assertion. The second one is obtained from the first using the relation

$$\left\|\dot{X}(t)\right\|_F \sigma_n(X(t))^{-1} = 1$$

and its derivative. □

3.5. Regularity. What kind of regularity can we expect for a condition geodesic? For nonsmooth metrics, Pugh proved the following (see [18]).

THEOREM 3.7. *A Lipschitz–Riemannian structure on a smooth manifold has local length minimizing geodesics of class C^{1+Lip}, that is, C^1 with locally Lipschitz derivatives.*

Since the condition metric is Lipschitz, this theorem can be applied to our problem. It shows the existence of local length minimizing geodesics of class C^1 with Lipschitz derivatives. Pugh's argument is based on a smooth perturbation of the Lipschitz metric and a careful study of the corresponding smooth geodesics which become C^{1+Lip} geodesics for the Lipschitz metric as the perturbation tends to zero.

Using the generalized Euler–Lagrange equation previously established, we extend the regularity result of Theorem 3.7 to all condition geodesics.

THEOREM 3.8. *Let $X \in W^{1,1}([0,L], \mathbb{GL}_{m,n})$ be a condition geodesic. Then X belongs to $W^{2,\infty}([0,L], \mathbb{R}^{n\times n})$.*

Proof. We can, without any loss of generality, suppose that our condition geodesic is parametrized by arc length, that is, $\|\dot{X}(t)\|_F\sigma_n(X(t))^{-1} = 1$ a.e. Since $\sigma_n(X(t))$ is bounded away from 0, there exist two positive constants α and β such that

$$0 < \alpha \le \left\|\dot{X}(t)\right\|_F \le \beta$$

for almost every t. From Corollary 3.4 we have

$$\left\|\dot{X}(t)\right\|_F = \frac{1}{\|P(t)\|_F} \text{ and } \dot{X}(t) = \frac{P(t)}{\|P(t)\|_F^2}$$

so that

$$0 < \frac{1}{\beta} \leq \|P(t)\|_F \leq \frac{1}{\alpha},$$

and $\dot{X}(t)$ has almost everywhere a derivative given by

$$\ddot{X}(t) = \frac{\dot{P}(t)}{\|P(t)\|_F^2} - 2P(t)\frac{\Re\left\langle P(t), \dot{P}(t)\right\rangle}{\|P(t)\|_F^4}, \tag{3.8}$$

which is clearly in $L^1\left([0,L], \mathbb{K}^{m\times n}\right)$. Thus, the second derivative of X is in L^1; that is, $X \in W^{2,1}$. In order to prove that $\ddot{X}$ is bounded, we have to show that the expression in (3.8) is bounded; that is, $\dot{P}(t)$ is bounded. This comes easily from (3.7). □

Remark 3. According to Rademacher's theorem, $C^{1+Lip} = W^{2,\infty}$ (see [13, Theorem 4.1]), and we obtain the same regularity as in Theorem 3.7.

4. Condition geodesics in Stiefel manifolds.

Consider the *Stiefel manifold*

$$\mathbb{S}_{m,n}(\mathbb{K}) = \left\{U \in \mathbb{K}^{m\times n} \ : \ U^*U = I_n\right\}.$$

If $m = n$, $\mathbb{S}_{m,n}(\mathbb{K})$ is the unitary group $\mathbb{U}_m$ when $\mathbb{K} = \mathbb{C}$ and the orthogonal group $\mathbb{O}_m$ when $\mathbb{K} = \mathbb{R}$. If $n = 1$, then $\mathbb{S}_{m,n}(\mathbb{K})$ is the unit sphere in $\mathbb{K}^m$. The Stiefel manifold is a real compact submanifold in $\mathbb{K}^{m\times n}$. Its dimension is equal to $mn - n(n+1)/2$ when $\mathbb{K} = \mathbb{R}$ and $2mn - n^2$ when $\mathbb{K} = \mathbb{C}$. The Stiefel manifold is equipped with the Riemannian structure induced by the Frobenius metric. It becomes a smooth complete Riemannian manifold. The main result in this section is the following.

THEOREM 4.1. *A geodesic $U(t)$ in $\mathbb{S}_{m,n}(\mathbb{K})$ for the Frobenius metric is also a geodesic in $\mathbb{GL}_{m,n}$ for the condition metric.*

Proof. Let us first describe the tangent bundle $T\mathbb{S}_{m,n}(\mathbb{K})$ and the normal bundle $N\mathbb{S}_{m,n}(\mathbb{K})$. Let I_{mn} denote the $m \times n$ matrix with entries $(I_{mn})_{ij} = 1$ when $i = j$ and equal to 0 otherwise. Let $U \in \mathbb{S}_{m,n}(\mathbb{K})$ and $P \in \mathbb{U}_m$ be such that $U = PI_{mn}$. The tangent space at U is

$$T_U\mathbb{S}_{m,n}(\mathbb{K}) = \left\{P\begin{pmatrix} A \\ W \end{pmatrix} \ : \ A \in \mathbb{K}^{n\times n},\ A^* = -A,\ W \in \mathbb{K}^{(m-n)\times n}\right\},$$

and the normal space is

$$N_U\mathbb{S}_{m,n}(\mathbb{K}) = \left\{P\begin{pmatrix} B \\ 0 \end{pmatrix} \ : \ B \in \mathbb{K}^{n\times n},\ B^* = B\right\}.$$

Let $U : t \to U(t)$ be a geodesic curve in $\mathbb{S}_{m,n}(\mathbb{K})$ for the Frobenius metric parametrized by arc length so that $\dot{U}(t) \in T_{U(t)}\mathbb{S}_{m,n}(\mathbb{K})$ and $\ddot{U}(t) \in N_{U(t)}\mathbb{S}_{m,n}(\mathbb{K})$. For every t we have

$$U(t) = P(t)I_{mn}, \quad \dot{U}(t) = P(t)\begin{pmatrix} A(t) \\ W(t) \end{pmatrix}, \quad \ddot{U}(t) = P(t)\begin{pmatrix} B(t) \\ 0 \end{pmatrix},$$

with $A(t)$ skew-Hermitian, $B(t)$ Hermitian, and

$$\left\|\dot{U}(t)\right\|_F^2 = \text{trace } (A(t)^*A(t) + W(t)^*W(t)) = 1.$$

From $U(t)^*U(t) = I_n$, differentiating twice gives

$$U(t)^*\ddot{U}(t) + 2\dot{U}(t)^*\dot{U}(t) + \ddot{U}(t)^*U(t) = 0$$

so that

$$B(t) = -(A(t)^*A(t) + W(t)^*W(t)),$$

and, consequently, $B(t)$ is a negative semidefinite matrix and trace $B(t) = -1$.

Let us now show that $U(t)$ satisfies (3.4) and (3.5). Since $\sigma_n(U(t)) = 1$ and $\left\|\dot{U}(t)\right\|_F = 1$, we have to prove that

$$\frac{d}{dt}\dot{U}(t) \in -U(t)\text{co}\left\{u(t)u(t)^*\right\}$$

(the convex envelope in this formula is defined on vectors $u(t)$ with $\|u(t)\| = 1$ and $(U(t)^*U(t))^{-1}u(t) = u(t)$, that is, on all vectors $u(t)$ in the unit sphere); that is,

$$P(t)\begin{pmatrix} B(t) \\ 0 \end{pmatrix} \in -P(t)I_{mn}\text{co}\left\{uu^* \ : \ \|u\| = 1\right\}$$

or

$$B(t) \in -\text{co}\left\{uu^* \ : \ \|u\| = 1\right\}.$$

This last inclusion is true because $B(t)$ is a negative semidefinite matrix with trace $B(t) = -1$, and because co$\{uu^* \ : \ \|u\| = 1\}$ is equal to the set of positive semidefinite matrices with trace equal to 1 (a gentle exercise). □

5. Condition geodesics in the space of diagonal matrices. In this section we characterize condition geodesics in the positive, diagonal case, that is, when $X(t)$ satisfies

$$X(t)_{ij} = \begin{cases} \sigma_i(t), & \text{when } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

with

$$\sigma_1(t) \geq \sigma_2(t) \geq \cdots \geq \sigma_n(t) > 0.$$

Using the same techniques as for Theorem 2.2, it may be proved that there exists a path $X(t)$ in the space of absolutely continuous positive diagonal matrices with given endpoints which minimizes the condition length. Such a *diagonal condition geodesic* also satisfies the differential inclusion given in Theorem 3.3 so that it is a condition geodesic in $\mathbb{GL}_{m,n}$.

We also suppose that $X(t)$ is parametrized by arc length so that $\|\dot{X}(t)\|\sigma_n(t)^{-1} = 1$ for almost all t. In that case $X(t) \in W^{2,\infty}$; that is, $\sigma_i(t) \in W^{2,\infty}$ for every i.

From Corollary 3.4 it follows that $X(t)$ is characterized by the differential inclusion

$$\frac{d}{dt}\left(\dot{X}(t)\sigma_n(t)^{-2}\right) \in -X(t)\sigma_n(t)^{-2}\text{co}\left\{uu^* \ : \ \|u\| = 1 \text{ and } X(t)^*X(t)u = \sigma_n(t)^2u\right\}$$

for almost all t. Since

$$\frac{d}{dt}\left(\dot{X}(t)\sigma_n(t)^{-2}\right) = \ddot{X}(t)\sigma_n(t)^{-2} - 2\dot{X}(t)\sigma_n(t)^{-3}\dot{\sigma}_n(t),$$

we get

$$\ddot{X}(t)\sigma_n(t)^{-2} - 2\dot{X}(t)\sigma_n(t)^{-3}\dot{\sigma}_n(t) = -X(t)\sigma_n(t)^{-2}A(t) \tag{5.1}$$

for a matrix $A(t) \in \mathrm{co}\left\{uu^* \ : \ \|u\| = 1 \text{ and } X(t)^*X(t)u = \sigma_n(t)^2u\right\}$. To be more precise, when for every t the multiplicity of $\sigma_n(t)$ is $n - p \geq 1$, that is, when

$$\sigma_1(t) \geq \sigma_2(t) \geq \cdots \geq \sigma_p(t) > \sigma_{p+1}(t) = \cdots = \sigma_n(t) > 0$$

for every t, we have $u^T = (0, \ldots, 0, u_{p+1}, \ldots, u_n)$ so that

$$A(t) = \begin{pmatrix} 0 & 0 \\ 0 & U(t) \end{pmatrix},$$

where $U(t)$ is an $(n-p) \times (n-p)$ block. Since $A(t)$ is also a diagonal matrix, we obtain $A(t) = \mathrm{diag}(a_i(t),\ 1 \leq i \leq n)$, where the coefficients $a_i(t)$ satisfy $a_i(t) = 0$ for $1 \leq i \leq p$, $0 \leq a_i(t) \leq 1$, and $a_1(t) + \cdots + a_n(t) = 1$.

Equation (5.1) becomes

$$\ddot{\sigma}_i(t) - 2\dot{\sigma}_i(t)\dot{\sigma}_n(t)\sigma_n(t)^{-1} = 0, \quad 1 \leq i \leq p, \tag{5.2}$$

$$\ddot{\sigma}_i(t) - 2\dot{\sigma}_i(t)\dot{\sigma}_n(t)\sigma_n(t)^{-1} + a_i(t)\sigma_i(t) = 0, \quad p+1 \leq i \leq n, \tag{5.3}$$

$$\sum_{i=1}^{n} \dot{\sigma}_i(t)^2 = \sigma_n(t)^2; \tag{5.4}$$

the last equation comes from the fact that $X(t)$ is parametrized by arc length so that $\|\dot{X}(t)\|_F = \sigma_n(t)$. By adding the $n - p$ equations in (5.3), since $\sigma_i(t) = \sigma_n(t)$, $p + 1 \leq i \leq n$, and $\sum a_i(t) = 1$, we get

$$(n-p)\ddot{\sigma}_n(t) - 2(n-p)\dot{\sigma}_n(t)^2\sigma_n(t)^{-1} + \sigma_n(t) = 0.$$

Via the change of variable $z = \sigma_n^{-1}$, we obtain the linear equation

$$\ddot{z}(t) = \frac{1}{n-p}z(t)$$

so that

$$z(t) = a\exp\left(\frac{t}{\sqrt{n-p}}\right) + b\exp\left(-\frac{t}{\sqrt{n-p}}\right)$$

for suitable constants a and b. Thus

$$\sigma_n(t) = \frac{1}{a\exp\left(\frac{t}{\sqrt{n-p}}\right) + b\exp\left(-\frac{t}{\sqrt{n-p}}\right)},$$

and

$$\dot\sigma_n(t) = \frac{1}{\sqrt{n-p}}\left(b\exp\left(-\frac{t}{\sqrt{n-p}}\right) - a\exp\left(\frac{t}{\sqrt{n-p}}\right)\right)\sigma_n(t)^2.$$

A first integration of (5.2) gives

$$\dot\sigma_i(t) = c_i\sigma_n(t)^2,\ 1 \le i \le p,$$

and, by a second integration,

$$\sigma_i(t) = c_iS(t) + d_i,\ 1 \le i \le p,$$

where c_i and d_i are constants and $S(t)$ is an antiderivative of $\sigma_n(t)^2$. All these considerations prove that

$$\begin{pmatrix} \sigma_1(t) \\ \vdots \\ \sigma_p(t) \\ \sigma_{p+1}(t) \\ \vdots \\ \sigma_n(t) \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} S(t) + \begin{pmatrix} d_1 \\ \vdots \\ d_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \sigma_n(t)$$

so that $X(t)$ is a plane curve.

To investigate more deeply its structure, we use (5.4) to get

$$\sum_{i=1}^n \dot\sigma_i(t)^2 = \sum_{i=1}^p \dot\sigma_i(t)^2 + (n-p)\dot\sigma_n(t)^2$$

$$= \sum_{i=1}^p c_i^2\sigma_n(t)^4 + \left(b\exp\left(-\frac{t}{\sqrt{n-p}}\right) - a\exp\left(\frac{t}{\sqrt{n-p}}\right)\right)^2 \sigma_n(t)^4 = \sigma_n(t)^2$$

so that

$$\sum_{i=1}^p c_i^2 = 4ab. \tag{5.5}$$

Either $\sum_{i=1}^p c_i^2 = 0$ so that

$$\begin{aligned} \sigma_i(t) &= d_i, && 1 \le i \le p, \\ \sigma_i(t) &= \sigma_n(t), && p+1 \le i \le n, \end{aligned}$$

and $X(t)$ is contained in a segment of line or $\sum_{i=1}^p c_i^2 \neq 0$ so that $a > 0$, $b > 0$, and

$$\begin{aligned} \sigma_i(t) &= -\tfrac{c_i\sqrt{n-p}}{2}\frac{1}{a\left(a\exp\left(2\frac{t}{\sqrt{n-p}}\right)+b\right)} + d_i, && 1 \le i \le p, \\ \sigma_i(t) &= \frac{1}{a\exp\left(\frac{t}{\sqrt{n-p}}\right)+b\exp\left(-\frac{t}{\sqrt{n-p}}\right)}, && p+1 \le i \le n. \end{aligned}$$

An easy computation shows that

$$\sum_{i\in Q}\left(\sigma_i(t) - d_i + \frac{\sqrt{n-p}}{qc_i}\right)^2 + \sum_{i=p+1}^n \sigma_i(t)^2 = \frac{n-p}{q^2}\sum_{i\in Q}\frac{1}{c_i^2},$$

where Q denotes the set of indices i such that $1 \leq i \leq p$, $c_i \neq 0$, and $q = \#Q$. This is the equation of a sphere in $\mathbb{R}^q \times \mathbb{R}^{n-p}$. Notice that $\sigma_i(t) = d_i$ when $1 \leq i \leq p$ and $i \notin Q$. Thus, in this second case, $X(t)$ is contained in an arc of circle.

We summarize these considerations in the following.

Theorem 5.1. *Let $X(t) \in \mathbb{GL}_{m,n}$ be a condition geodesic parametrized by arc length and contained in the set of diagonal positive matrices. Let us write $X(t) = \operatorname{diag}(\sigma_i(t))$, and suppose that*

$$\sigma_1(t) \geq \cdots \geq \sigma_p(t) > \sigma_{p+1}(t) = \cdots = \sigma_n(t)$$

for every t. Then $X(t)$ is contained either in a line segment or in an arc of circle (intersection of a plane with a sphere).

6. Numerical experiments. In this section we study condition paths from a numerical viewpoint. We consider here the particular case of paths in $\mathbb{GL}_n(\mathbb{R})$, and we discuss the numerical solution to the following task:

> Given matrices X_0 and X_1 belonging to the same connected component of $\mathbb{GL}_n(\mathbb{R})$, compute a minimizing condition path $X(t)$, with endpoints X_0 and X_1.

Notice that the computation of such paths cannot be derived from shooting methods because the corresponding initial value problem (IVP) is ill-posed; see, e.g., section 6.1.

A possible approach to the solution of our problem consists of

- applying standard tools of differential geometry or the classical Euler–Lagrange equation to compute equations for condition paths, assuming that the multiplicity of the smallest singular value is 1 in each point of the curve, and
- using a BVP solver (e.g., the function `bvp4c` in Matlab) to compute a solution of the equation, with boundary conditions defined by X_0 and X_1.

As it might be expected, this method works well as long as $\sigma_n(X(t))$ actually has multiplicity 1 for each t, but it tends to give unsatisfactory results otherwise, especially when a whole segment of the sought condition path belongs to the locus of matrices in $\mathbb{GL}_n(\mathbb{R})$ whose smallest singular value has multiplicity greater than 1.

For this reason, we prefer to follow an optimization approach and compute $X(t)$ by minimizing the condition length functional. In a theoretical framework, $X(t)$ is a minimizer among curves in $W^{2,\infty}$. In numerical applications, $W^{2,\infty}$ must be replaced by a finite dimensional space; in other words, we need to choose a discrete parametrization for the space of curves over which the minimization process is carried out. The general outline of the geodesic computation process goes as follows.

- Write the length functional (and, if necessary, its gradient) in discretized form, as a function of the chosen parameters.
- Apply an optimization method to compute a minimum of the functional.

Theoretical considerations and numerical tests show that a good choice for the discretization of the curve space is a Fourier parametrization; that is, we consider curves of the type

$$\begin{aligned} &Y(t) : [0,1] \longrightarrow \mathbb{GL}_n(\mathbb{R}),\\ &Y(t) = X_0 + (X_1 - X_0)t + A_1 \sin(\pi t) + A_2 \sin(2\pi t) + \cdots + A_N \sin(N\pi t), \end{aligned}$$

which are parametrized by the n^2N entries of the matrices $A_1, A_2, \ldots, A_N \in \mathbb{R}^{n\times n}$. This choice is theoretically motivated by the fact that, for $N = \infty$, curves of this type are dense in the set of curves on $W^{2,\infty}([0,1], \mathbb{GL}_n(\mathbb{R}))$ which have fixed endpoints X_0 and X_1. In other words, the curve $X(t) - (X_0 + (X_1 - X_0)t)$, whose value is 0 for

$t = 0$ and $t = 1$, can be approximated in the sense of L^2 or pointwise convergence using linear combinations of functions in $\{\sin(j\pi t)\}_{j=1,\ldots,\infty}$ with coefficients in $\mathbb{R}^{n\times n}$. Besides, sine functions are a very natural choice to approximate a function with zero boundary values.

This parametrization gives good results (better than, for instance, a piecewise linear parametrization) and requires few parameters: in the applications shown here we have always chosen $N = 9$.

The optimization method we employed is Overton's implementation of the BFGS method, written for Matlab; it can be downloaded from the website [17]. Though not originally conceived for nonsmooth problems, the method proves to be surprisingly robust, as discussed in [15].

This BFGS command requires in particular

- a user-defined function which computes the length functional and its gradient on a given curve,
- an initial guess of the solution.

The gradient of the length function has been computed analytically by using standard techniques in nonsmooth analysis, as the ones outlined in section 3. A straight line (i.e., the curve defined by $A_1 = \cdots = A_N = 0$) is usually a good choice as initial guess, provided, of course, that it does not intersect the set of singular matrices. In some cases, though, it is advisable to try an initial guess defined by some nonzero parameters.

We show now some interesting examples of condition paths. In the following, we will use the following notation:

$$\mathcal{S}_{n,k} = \{X \in \mathbb{GL}_n : \sigma_n(X) = \sigma_{n-1}(X) = \cdots = \sigma_{n-k+1}(X)\} \quad \text{for} \quad 2 \le k \le n;$$

this set will be called the *multiplicity locus.*

6.1. 2 × 2 diagonal matrices. We consider condition paths of the form

$$X(t) = \begin{bmatrix} x(t) & 0 \\ 0 & y(t) \end{bmatrix},$$

where we assume $x(t)$ and $y(t)$ are real and strictly positive, and we plot these curves on the (x, y) plane. The multiplicity locus $\mathcal{S}_{2,2}$ is the line defined by the equation $x = y$. By writing the metric explicitly, it can be seen that hyperbolic geometry holds in each of the semiquadrants $x > y$ and $y > x$. As a consequence, condition paths in this case are obtained as a C^1-gluing of the following "building blocks":

- arcs of circumference with center on the x axis (when $x > y$),
- segments,
- arcs of circumference with center on the y axis (when $y > x$).

The function $\sigma_2(X(t))^{-1}$ (inverse of the smallest singular value) is always convex along such geodesics; it is also C^1 when the geodesic meets $\mathcal{S}_{2,2}$ tangentially (or, of course, when it does not meet $\mathcal{S}_{2,2}$ at all).

Figures 6.1 and 6.2 show the condition path of endpoints $X_0 = \mathrm{diag}(4, 1)$ and $X_1 = \mathrm{diag}(1, 3)$, which meets $\mathcal{S}_{2,2}$ transversally, and the corresponding function $\sigma_2(X(t))^{-1}$, parametrized by arc length. Figures 6.3 and 6.4 show an example where $\mathcal{S}_{2,2}$ is met tangentially; here the endpoints are $X_0 = \mathrm{diag}(1, 3)$ and $X_1 = \mathrm{diag}(9, 8)$.

Notice that this last case provides a counterexample to the IVP formulation of the problem of computing condition paths. Indeed, once the path meets $\mathcal{S}_{2,2}$ tangentially (or if it already starts tangentially from $\mathcal{S}_{2,2}$), it may either continue along $\mathcal{S}_{2,2}$ or leave it at any moment; so the problem of determining a path from initial conditions does not have a unique solution. For instance, Figure 6.5 shows some condition paths

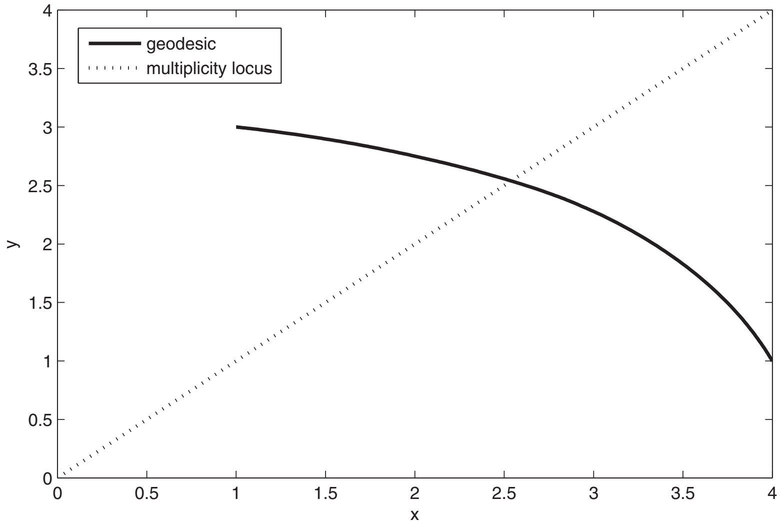

FIG. 6.1. *Condition path that meets the multiplicity locus transversally.*

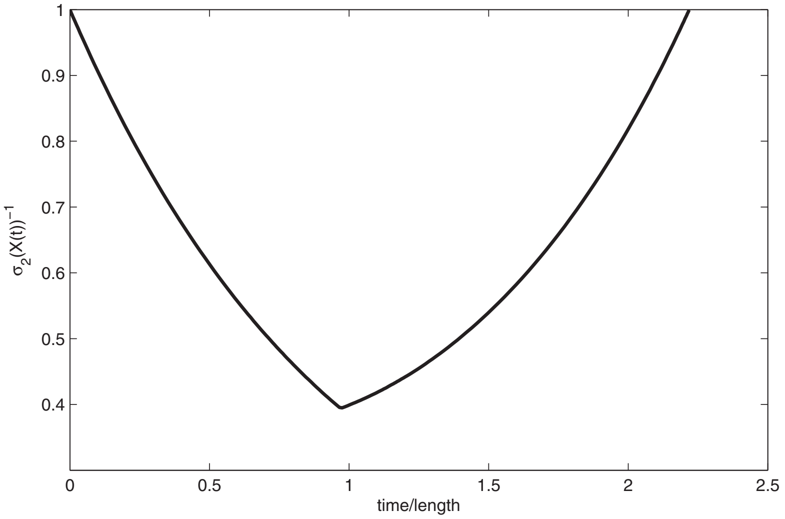

FIG. 6.2. *Behavior of* $\sigma_2(X(t))^{-1}$.

which have the same initial conditions as the one depicted in Figure 6.3 but a different global behavior.

6.2. General matrices. We consider here a general example in $\mathbb{GL}_5$. Let us take $X_0 = \mathrm{diag}(13, 7, 3, 9, 5)$ and $X_1 = USV^*$, where $S = \mathrm{diag}(8, 5, 2, 4, 6)$ and U, V are randomly chosen orthogonal matrices. Notice that choosing one of the endpoints as a positive diagonal matrix does not cause any loss of generality, since we can always apply orthogonal transformations that send a given matrix to a diagonal form. The behavior of singular values along the condition geodesic $X(t)$ that joins X_0 and X_1

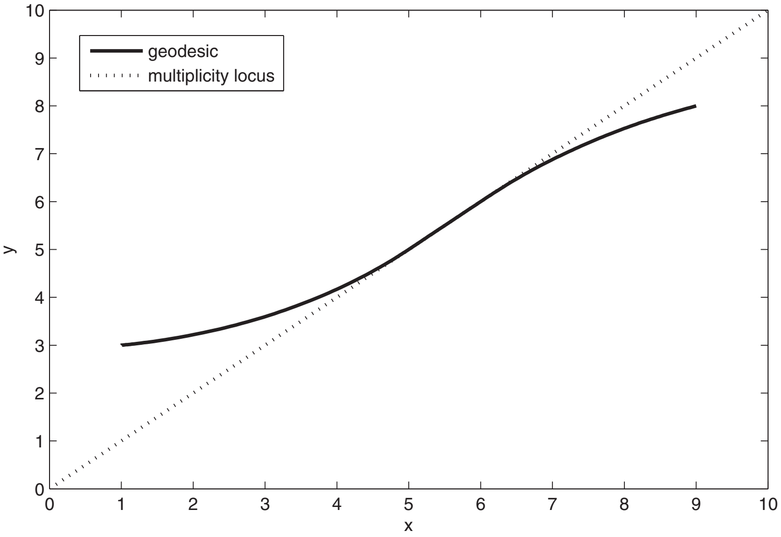

FIG. 6.3. *Condition path that meets the multiplicity locus tangentially.*

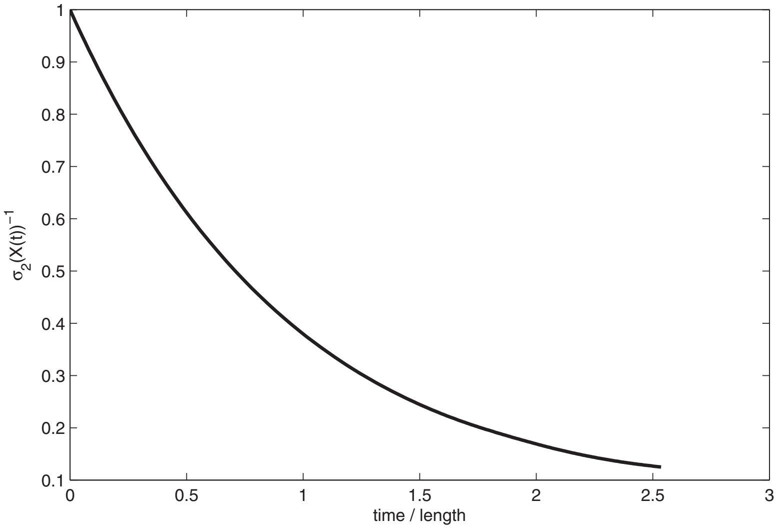

FIG. 6.4. *Behavior of* $\sigma_2(X(t))^{-1}$.

is plotted in Figure 6.6; the multiplicity of $\sigma_n(X(t))$ varies from 1 to 3. The function $\sigma_5(X(t))^{-1}$, parametrized by arc length, is shown in Figure 6.7, and it can be seen to be convex. This example shows that, generically, a condition path does not cut the multiplicity locus transversally.

6.3. Orthogonal endpoints. We study here the case examined in section 4, for the particular case where the considered Stiefel manifold is the orthogonal group $\mathbb{O}_n$. Theorem 4.1 suggests that, if we choose matrices X_0, X_1 in the same connected component of $\mathbb{O}_n$, then the condition geodesic in $\mathbb{GL}_n$ with endpoints X_0 and X_1

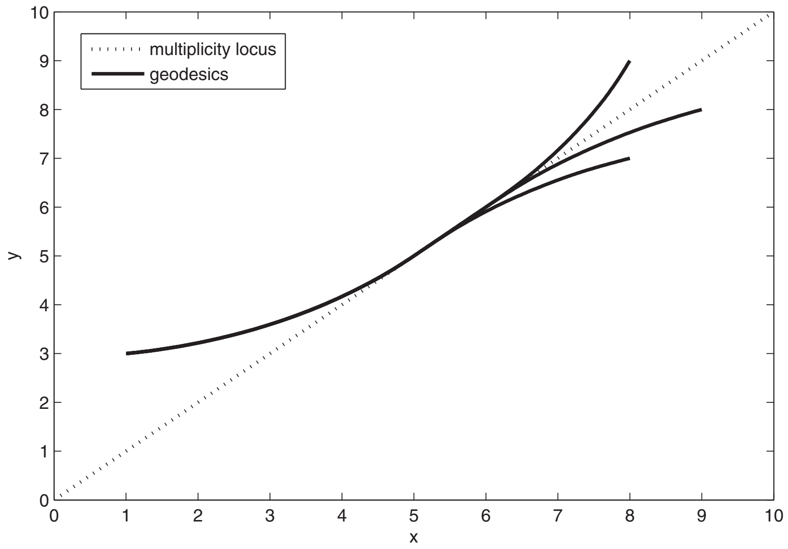

FIG. 6.5. *Condition paths that meet the multiplicity locus tangentially.*

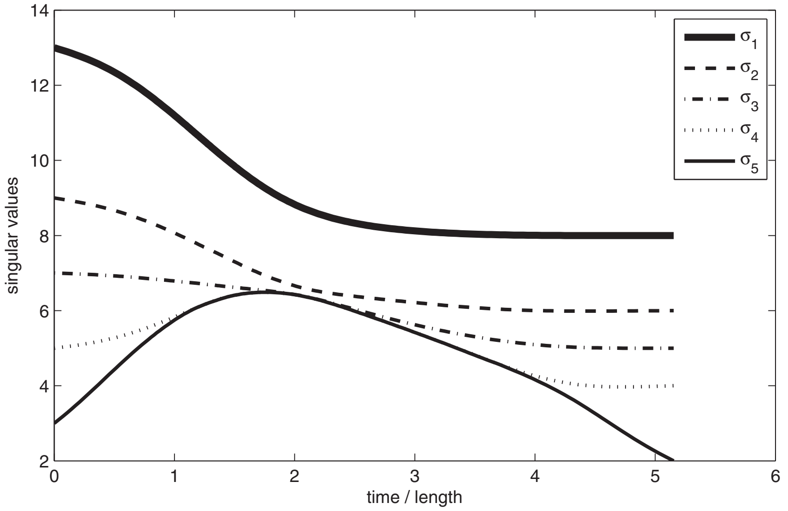

FIG. 6.6. *Singular values along a condition geodesic in* $\mathbb{GL}_5$.

belongs in fact to $\mathbb{O}_n$. Numerical experiments confirm this. In the example we report here, X_0 and X_1 are Householder matrices associated with vectors $v_0 = [1, 2, 3, 4]^T$ and $v_1 = [1, 1, 1, 1]^T$, respectively. Figure 6.8 shows the behavior of singular values along the computed geodesic; all the singular values are numerically close to 1.

6.4. Endpoints in the multiplicity locus. Consider the following problem: given endpoints X_0, X_1 in the same connected component of $\mathbb{GL}_n$, with $\sigma_n(X_0) = \cdots = \sigma_{n-k+1}(X_0)$ and $\sigma_n(X_1) = \cdots = \sigma_{n-k+1}(X_1)$, does the associated condition geodesic belong to the multiplicity locus $\mathcal{S}_{n,k}$? In other words, is $\mathcal{S}_{n,k}$ geodesically complete?

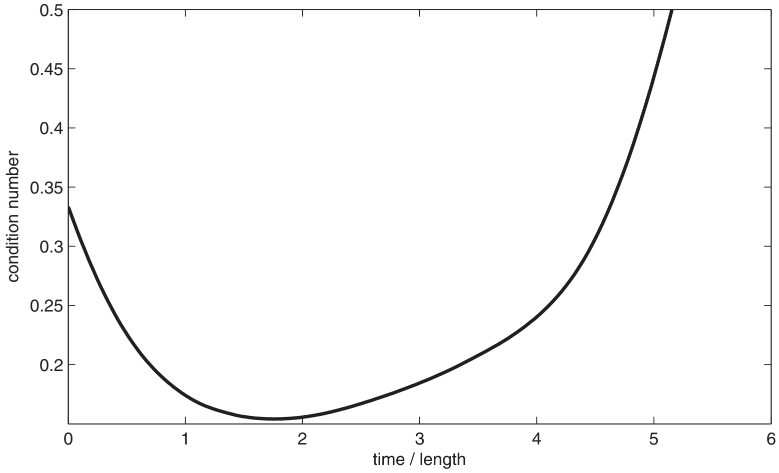

FIG. 6.7. *Behavior of the "normalized condition number"* $\sigma_5(X(t))^{-1}$.

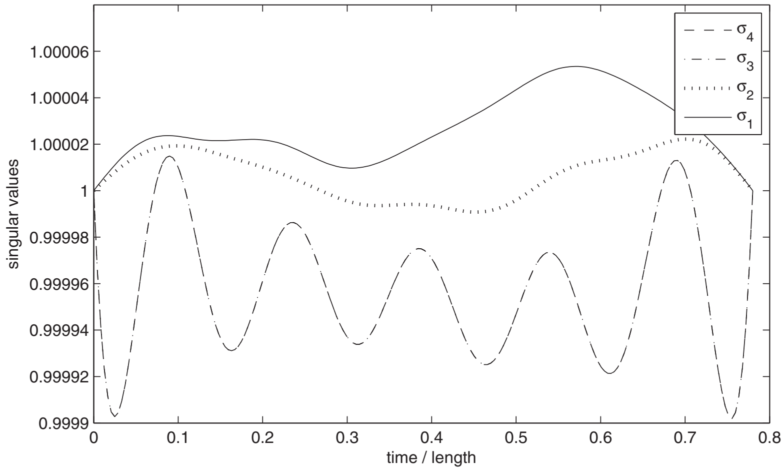

FIG. 6.8. *Singular values along a condition geodesic in the orthogonal group. Notice that the plots for* σ_3 *and* σ_4 *are nearly identical.*

We have theoretical proof that this is actually the case when working with diagonal matrices. Numerical experiments, however, seem to suggest that $\mathcal{S}_{n,2}$ is geodesically complete also in the general case. In the example shown here, $X_0 = \mathrm{diag}(7, 9, 3, 3)$ and $X_1 = USV^*$, where $S = \mathrm{diag}(8, 7, 1, 1)$ and U, V are randomly chosen orthogonal matrices. The singular values along the computed condition geodesic are plotted in Figure 6.9.

Notice, however, that there seem to be counterexamples to the conjecture that $\mathcal{S}_{n,k}$ may be geodesically complete for any k. Figure 6.10 shows the behavior of singular values along a numerically computed geodesic of endpoints $X_0 = \mathrm{diag}(7, 3, 3, 3)$ and $X_1 = USV^*$, where $S = \mathrm{diag}(8, 1, 1, 1)$ and U, V are again random orthogonal matrices.

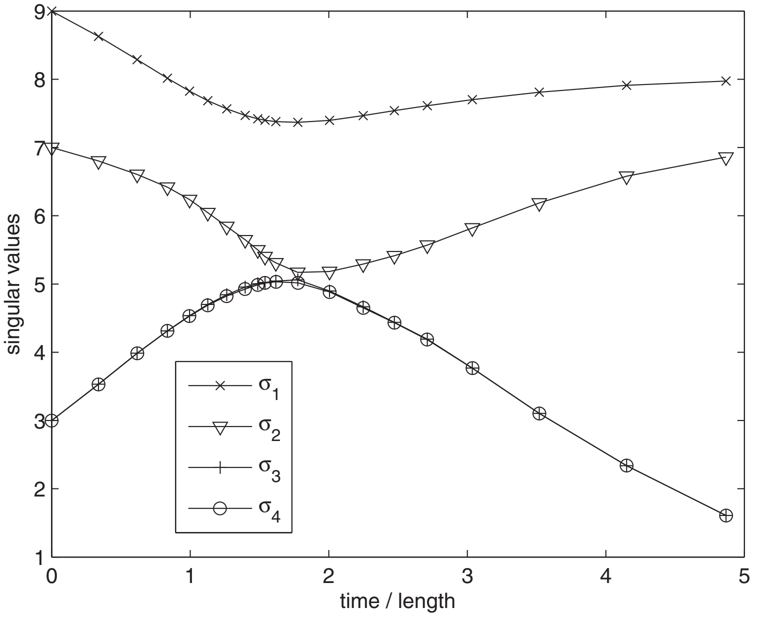

FIG. 6.9. *Is $\mathcal{S}_{4,2}$ geodesically complete?*

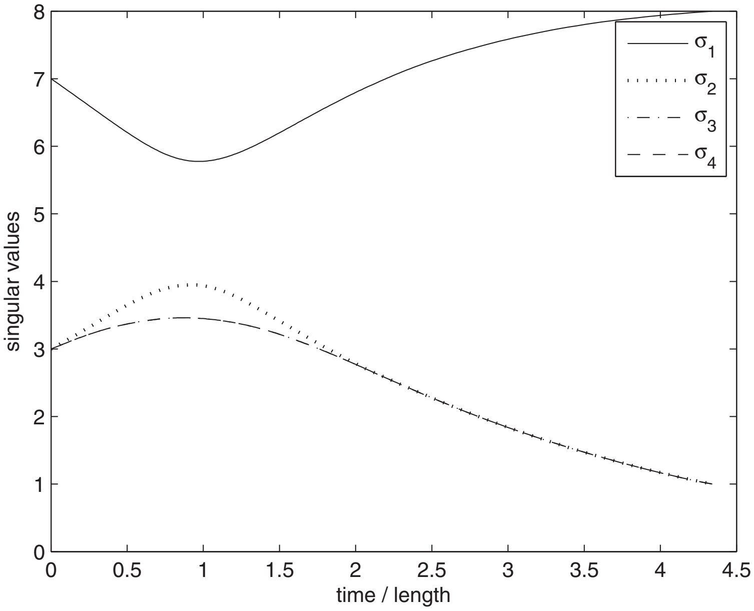

FIG. 6.10. *Numerical computations suggest that $\mathcal{S}_{4,3}$ is not geodesically complete.*

Acknowledgments. We want to thank Carlos Beltrán, Gregorio Malajovich, and Michael Shub for useful discussions about the condition metric and Michael Overton who lighted up our route to the BFGS method. Our thanks also go to SIMAX editor Adrian Lewis and to two anonymous referees for their helpful suggestions.

REFERENCES

[1] A. Baker, *Matrix Groups: An Introduction to Lie Group Theory*, Springer Undergrad. Math. Ser., Springer, London, 2002.

[2] C. Beltrán, J.-P. Dedieu, G. Malajovich, and M. Shub, *Convexity properties of the condition number*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1491–1506.

[3] C. Beltrán and M. Shub, *Complexity of Bézout's theorem* VII*: Distance estimates in the condition metric*, Found. Comput. Math., 9 (2009), pp. 179–195.

[4] C. Beltrán and M. Shub, *On the Geometry and Topology of the Solution Variety for Polynomial System Solving*, http://sites.google.com/site/beltranc/preprints (2008).

[5] F. Clarke, *The Erdmann condition and Hamiltonian inclusions in optimal control and the calculus of variations*, Canad. J. Math., 32 (1980), pp. 494–509.

[6] F. Clarke, *Optimization and Nonsmooth Analysis*, J. Wiley and Sons, New York, 1983.

[7] N. Dunford and J. Schwartz, *Linear Operators. Part* I: *General Theory*, Interscience, New York, 1957.

[8] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[9] J.-C. Evard and F. Jafari, *The set of all $m\times n$ rectangular real matrices of rank r is connected by analytic regular arcs*, Proc. Amer. Math. Soc., 120 (1994), pp. 413–419.

[10] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian Geometry*, Springer, New York, 2004.

[11] G. Golub and C. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[12] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces*, Birkhäuser, Boston, 2007.

[13] J. Heinonen, *Lectures on Lipschitz Analysis*, http://www.math.jyu.fi/research/reports/rep100.pdf (2005).

[14] J.-B. Hiriart-Urruty and D. Ye, *Sensitivity analysis of all the eigenvalues of a symmetric matrix*, Numer. Math., 70 (1995), pp. 45–72.

[15] A. Lewis and M. Overton, *Nonsmooth optimization via BFGS*, SIAM J. Optim., submitted.

[16] A. Lewis and H. Sendov, *Nonsmooth analysis of singular values. Part* II: *Applications.* Set-Valued Anal., 13 (2005), pp. 243–264.

[17] M. Overton, *HANSO package*, http://cs.nyu.edu/faculty/overton/software/hanso/index.html.

[18] C. Pugh, *private communication*, 2007.

[19] M. Shub, *Complexity of Bezout's theorem* VI: *Geodesics in the condition (number) metric*, Found. Comput. Math., 9 (2009), pp. 171–178.

[20] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.