Kybernetika

Uwe Helmke; Michael Prechtel; Mark A. Shayman

Riccati-like flows and matrix approximations

Kybernetika, Vol. 29 (1993), No. 6, 563--582

Persistent URL: http://dml.cz/dmlcz/125038

Terms of use:

© Institute of Information Theory and Automation AS CR, 1993

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

KYBERNETIKA — VOLUME 29 (1993), NUMBER 6, PAGES 563 – 582

RICCATI-LIKE FLOWS AND MATRIX APPROXIMATIONS[1]

UWE HELMKE, MICHAEL PRECHTEL AND MARK A. SHAYMAN

A classical problem in matrix analysis, total least squares estimation and model reduction theory is that of finding a best approximant of a given matrix by lower rank ones. It is common believe that behind every such least squares problem there is an algebraic Riccati equation. In this paper we consider the task of minimizing the distance function $f_A(X) = \|A - X\|^2$ on varieties of fixed rank symmetric matrices, using gradient-like flows for the distance function f_A. These flows turn out to have similar properties as the dynamic Riccati equation and are thus termed Riccati-like flows. A complete phase portrait analysis of these Riccati-like flows is presented, with special emphasis on positive semidefinite solutions. A variable step-size discretization of the flows is considered. The results may be viewed as a prototype for similar investigations one would like to pursue in model reduction theory of linear control systems.

1. INTRODUCTION

A classical result from matrix analysis, the Eckart–Young–Mirsky Theorem, states that the best approximation of a given $M \times N$-matrix A by matrices of smaller rank is given by a truncated singular value decomposition of A. Such results are of interest for Total Least Squares Approximation; see Golub and Van Loan [2]. It is common believe that behind every least squares problem there is an associated algebraic Riccati equation. This has been recently confirmed by De Moor and David [7] for the total linear least squares problem. Here we show that a similar connection also exists with the dynamic Riccati equation.

We consider the approximation problem of a symmetric matrix by lower rank symmetric matrices. In Helmke and Shayman [5] the critical points and their Morse indices of the distance function $f_A(X) = \|A - X\|^2$ on the variety of fixed rank symmetric matrices X have been determined. These results, which contain those of Higham [6] and others as special cases, are summarized in Section 2.

As the set of symmetric matrices of fixed rank is a smooth manifold, this opens the way to study the minimization task for the function $f_A(X)$ by steepest descent

[1] This research was supported by a Grant I-0184-078.06/91 from the G.I.F., the German-Israeli Foundation for Scientific Research and Development

methods. In this paper, a systematic study of gradient flows related to the minimization of the distance function $f_A(X)$ is presented. With respect to a certain Riemannian metric the gradient flow of $f_A(X)$ has a simple, appealing form, being a cubic matrix differential equation. It is termed a Riccati-like flow as its convergence properties are similar to those of the Riccati differential equation. A simplified version of the gradient flow is a certain Riccati equation. A complete phase portrait analysis of both flows is given. On the subset of positive semidefinite matrices the Riccati equation turns out to be a gradient-like flow. We thus establish a close link between stable equilibria of the Riccati equation and positive semidefinite best approximants of a given symmetric matrix A; i.e. with the polar decomposition.

In Section 4 a related rank preserving flow on rectangular matrices is considered. This flow seems to have certain advantages such as structural stability properties, compared with the Riccati equation. Recursive versions of this flow involving a variable step size are considered in the last section. This leads to a new recursive algorithm for finding the best approximant of A by positive semidefinite matrices of fixed rank.

For a general discussion of such optimization problems from a dynamical systems viewpoint we refer to the forthcoming book Helmke and Moore [4]. Related questions arise in control theory where one is interested in finding best approximations of linear systems by lower order ones. The results obtained in this paper may thus be viewed as a prototype for an investigation one would like to pursue in model reduction theory of linear systems.

2. APPROXIMATIONS BY SYMMETRIC MATRICES

Let $S(N)$ denote the set of all $N \times N$ real symmetric matrices. For integers $1 \leq n \leq N$ let

$$S(n, N) = \{X \in \mathbb{R}^{N\times N} \mid X^{\mathrm{T}} = X, \ \operatorname{rank} X = n\} \tag{2.1}$$

denote the set of real symmetric $N \times N$ matrices of rank n. Given a fixed real symmetric $N \times N$ matrix A we consider the distance function

$$f_A \colon S(n, N) \longrightarrow \mathbb{R}, \quad X \longmapsto \|A - X\|^2 \tag{2.2}$$

where $\|X\|^2 = \operatorname{tr}(XX^{\mathrm{T}})$ is the Frobenius norm. We are interested in finding the critical points and local and global minima of f_A, i.e. the best rank n symmetric approximants of A. The following result summarizes some basic geometric properties of the constraint set $S(n, N)$. For proofs of the subsequent results in this section we refer to Helmke and Shayman [5].

Proposition 2.1. $S(n, N)$ is a smooth manifold of dimension $\frac{1}{2}n(2N - n + 1)$ and has $n + 1$ connected components

$$S(p, q; N) = \{X \in S(n, N) \mid \text{signature } X = p - q\} \tag{2.3}$$

where p, $q \geq 0$, $p + q = n$. The tangent space of $S(n, N)$ at an element X is

$$T_X S(n, N) = \{\Delta X + X\Delta^{\mathrm{T}} \mid \Delta \in \mathbb{R}^{N\times N}\}. \tag{2.4}$$

Theorem 2.2. (i) Let $A \in \mathbb{R}^{N\times N}$ be symmetric and let

$$N_+ = \dim \operatorname{Eig}^+(A), \ N_- = \dim \operatorname{Eig}^-(A) \tag{2.5}$$

be the numbers of positive and negative eigenvalues of A, respectively. The critical points X of the distance function $f_A: S(n,N) \to \mathbb{R}$ are characterized by $AX = X^2$.

(ii) If A has N distinct eigenvalues $\lambda_1 > \ldots > \lambda_N$, and $A = \Theta \operatorname{diag}(\lambda_1,\ldots,\lambda_N)\Theta^{\mathrm{T}}$ for $\Theta \in O(N)$, then the restriction of the distance function $f_A: S(p,q;N) \to \mathbb{R}$ has exactly

$$\binom{N_+}{p}\binom{N_-}{q}$$

critical points. (Here $O(N)$ denotes the set of $N \times N$ real orthogonal matrices.) In particular, f_A has critical points in $S(p,q;N)$ if and only if $p \le N_+$ and $q \le N_-$. The critical points $X \in S(p,q;N)$ of f_A with $p \le N_+, q \le N_-$, are characterized by

$$X = \Theta \ \operatorname{diag}(x_1,\ldots,x_N)\,\Theta^{\mathrm{T}} \tag{2.6}$$

with

$$x_i = 0 \ \text{ or } \ x_i = \lambda_i, \quad i = 1,\ldots,N \tag{2.7}$$

and exactly p of the x_i are positive and q are negative.

Theorem 2.2 has the following immediate consequence.

Corollary 2.3. Let $A = \Theta \ \operatorname{diag}(\lambda_1,\ldots,\lambda_N)\,\Theta^{\mathrm{T}}$ with $\Theta \in O(N)$ a real orthogonal $N\times N$ matrix and $\lambda_1 \ge \ldots \ge \lambda_N$. A minimum $\hat{X} \in S(p,q;N)$ for $f_A: S(p,q;N) \to \mathbb{R}$ exists if and only if $p \le N_+$ and $q \le N_-$. One such minimizing $\hat{X} \in S(p,q;N)$ is given by

$$\hat{X} = \Theta \ \operatorname{diag}(\lambda_1,\ldots,\lambda_p,0,\ldots,0,\lambda_{N-q+1},\ldots,\lambda_N)\,\Theta^{\mathrm{T}} \tag{2.8}$$

and the minimum value of $f_A: S(p,q;N) \to \mathbb{R}$ is $\sum_{i=p+1}^{N-q} \lambda_i^2$. $\hat{X} \in S(p,q;N)$, given by (2.8) is the unique minimum of $f_A: S(p,q;N) \to \mathbb{R}$ if $\lambda_p > \lambda_{p+1}$ and $\lambda_{N-q} > \lambda_{N-q+1}$.

An important question in linear algebra is that of finding the best positive semidefinite symmetric approximant of a given $N \times N$ matrix A. By Corollary 2.3 we have

Corollary 2.4. Let $A = \Theta \operatorname{diag}(\lambda_1,\ldots,\lambda_N)\,\Theta^{\mathrm{T}}$ with $\Theta \in O(N)$ real orthogonal and $\lambda_1 \ge \ldots \ge \lambda_n > \lambda_{n+1} \ge \ldots \ge \lambda_N, \lambda_n > 0$. Then

$$\hat{X} = \Theta \operatorname{diag}(\lambda_1,\ldots,\lambda_n,0,\ldots,0)\,\Theta^{\mathrm{T}} \in S(n,N) \tag{2.9}$$

is the unique positive semidefinite symmetric best approximant of A of rank $\le n$. In particular

$$\hat{X} = \Theta \operatorname{diag}(\lambda_1,\ldots,\lambda_{N_+},0,\ldots,0)\,\Theta^{\mathrm{T}} \tag{2.10}$$

is the uniquely determined best approximant of A in the class of positive semidefinite symmetric matrices.

This implies a result due to Higham [6].

Thus under a genericity assumption on A the best symmetric approximant of a symmetric matrix $A \in \mathbb{R}^{N\times N}$ in the Frobenius norm is uniquely determined. To what extent is this also true for other critical points such as, e. g. local minima, of $f_A: S(n,N) \to \mathbb{R}$? To answer this question we will now show that – under a suitable genericity condition on A – each critical point of $f_A: S(n,N) \to \mathbb{R}$ is nondegenerate and we will compute the *index* of each critical point, i. e. the dimension of the largest subspace on which the Hessian is negative definite.

Theorem 2.5. Let $A \in \mathbb{R}^{N\times N}$ be symmetric with distinct eigenvalues.

(i) Every critical point $f_A: S(n,N) \to \mathbb{R}$ is nondegenerate.

(ii) The index of the critical point $X \in S(p,q;N)$ associated with the permutations μ, ν ($\mu_1 < \ldots < \mu_p$ and $\mu_{p+1} < \cdots < \mu_{N_+}$, $\nu_1 < \ldots < \nu_q$ and $\nu_{q+1} < \ldots < \nu_{N_-}$) is given by

$$\begin{aligned} &\operatorname{Card}\{(i,j) \mid \mu_i < \mu_j, 1 \le i \le p,\ p+1 \le j \le N_+\} \\ +\ &\operatorname{Card}\ \{(i,j) \mid \nu_i < \nu_j, 1 \le i \le q,\ q+1 \le j \le N_-\}. \end{aligned}$$

(iii) The number of critical points in $S(p,q;N)$ which have index d is equal to the dth mod 2 Betti number of the product of Grassmann manifolds

$$Grass(p, \mathbb{R}^{N_+}) \times Grass(q, \mathbb{R}^{N_-}).$$

In particular, $f_A: S(n,N) \to \mathbb{R}$ has exactly $n+1$ local minima, exactly one in each component $S(p,q;N)$.

3. GRADIENT FLOWS

In this section we develop a gradient flow approach to find the critical points of the distance function $f_A: S(n,N) \to \mathbb{R}$. For $A, B \in \mathbb{R}^{N\times N}$ we define

$$\{A, B\} = AB + B^{\mathrm{T}} A^{\mathrm{T}}. \tag{3.1}$$

Thus the tangent space $T_X S(n,N)$ is the image of the linear map

$$\pi_X: \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}, \quad \Delta \mapsto \{\Delta, X\} \tag{3.2}$$

while the kernel of π_X is

$$\ker \pi_X = \{\Delta \in \mathbb{R}^{N\times N} \mid \Delta X + X\Delta^{\mathrm{T}} = 0\} \tag{3.3}$$

Taking the orthogonal complement $(\ker \pi_X)^\perp$ with respect to the standard inner product on $\mathbb{R}^{N\times N}$ yields the isomorphism of vector spaces

$$(\ker \pi_X)^\perp \cong \mathbb{R}^{N\times N} / \ker \pi_X \cong T_X S(n,N) \tag{3.4}$$

We have the orthogonal decomposition of $\mathbb{R}^{N\times N}$

$$\mathbb{R}^{N\times N} = \ker \pi_X \oplus (\ker \pi_X)^\perp$$

and hence every element $\Delta \in \mathbb{R}^{N\times N}$ has a unique decomposition

$$\Delta = \Delta_X + \Delta^X \tag{3.5}$$

where $\Delta_X \in \ker \pi_X$ and $\Delta^X \in (\ker \pi_X)^\perp$. Given any pair of tangent vectors $\{\Delta_1, X\}, \{\Delta_2, X\}$ of $T_X S(n, N)$ we define

$$\langle\langle \{\Delta_1, X\}, \{\Delta_2, X\}\rangle\rangle := 4\,\mathrm{tr}\,((\Delta_1^X)^{\mathrm{T}} \Delta_2^X). \tag{3.6}$$

It is easy to show that $\langle\langle\, , \rangle\rangle$ defines a nondegenerate symmetric bilinear form on $T_X S(n, N)$ for each $X \in S(n, N)$. In fact, $\langle\langle\, , \rangle\rangle$ defines a Riemannian metric of $S(n, N)$. We refer to $\langle\langle\, , \rangle\rangle$ as the *normal Riemannian metric* on $S(n, N)$.

A differential equation $\dot{X} = F(X)$ evolving on the matrix space $S(N)$ is said to be *rank preserving* if the rank $\mathrm{rk}X(t)$ of every solution $X(t)$ is constant as a function of t. The following lemma gives a simple characterization of rank preserving vector fields on $S(N)$.

Lemma 3.1. Let $I \subset \mathbb{R}$ be an interval and let $A(t) \in \mathbb{R}^{N\times N}$, $t \in I$, be a continuous family of matrices. Then

$$\dot{X}(t) = A(t)X(t) + X(t)A(t)^{\mathrm{T}}, \quad X(0) \in S(N), \tag{3.7}$$

is a rank preserving flow on $S(N)$. Conversely, any rank preserving vector field on $S(N)$ is of this form.

Proof. For any $X \in S(n, N)$, $0 \le n \le N$, $A(t)X + XA(t)^{\mathrm{T}} \in T_X S(n, N)$. Thus (3.7) defines a time varying vector field on $S(n, N)$. It follows that for any initial condition $X_0 \in S(n, N)$ the solution $X(t)$ of (3.7) satisfies $X(t) \in S(n, N)$ for $t \in I$. Therefore (3.7) is rank preserving. Conversely, suppose $\dot{X} = F(X)$ is rank preserving. Then $F(X) \in T_X S(n, N)$ for all $X \in S(n, N)$ and $0 \le n \le N$ arbitrary. By Proposition 2.1 therefore $F(X) = \Delta(X)X + X\Delta(X)^{\mathrm{T}}$. □

Theorem 3.2. Let $A \in \mathbb{R}^{N\times N}$ be symmetric.

(i) The gradient flow of $f_A: S(n, N) \to \mathbb{R}(\dot{X} = -\mathrm{grad} f_A(X))$ with respect to the normal Riemannian metric $\langle\langle\, , \rangle\rangle$ is

$$\dot{X} = \{(A - X)X, X\} = (A - X)X^2 + X^2(A - X) \tag{3.8}$$

(ii) For any $X(0) \in S(n, N)$ the solution $X(t)$ of (3.8) exists for all $t \ge 0$ and $X(t) \in S(n, N)$ for all $t \ge 0$.

(iii) Every solution $X(t) \in S(n, N)$ of (3.8) converges to an equilibrium point X_∞ characterized by $X_\infty(X_\infty - A) = 0$. Also X_∞ has rank $\le n$.

Proof. The gradient of f_A with respect to the normal metric is the uniquely determined vector field on $S(n, N)$ characterized by

$$\begin{aligned} Df_A(X)(\{\Delta, X\}) &= \langle\langle \mathrm{grad} f_A(X), \{\Delta, X\}\rangle\rangle \\ \mathrm{grad} f_A(X) &= \{\Omega, X\} \in T_X S(n, N) \end{aligned} \tag{3.9}$$

for all $\Delta \in \mathbb{R}^{N\times N}$ and some (unique) $\Omega \in (\ker \pi_X)^{\perp}$. A straightforward computation shows that the derivative of $f_A\colon S(n,N) \to \mathbb{R}$ at X is the linear map defined on $T_X S(n,N)$ by

$$\begin{aligned} Df_A(X)(\{\Delta, X\}) &= 2\,\mathrm{tr}\,(X\{\Delta, X\} - A\{\Delta, X\}) \\ &= 4\,\mathrm{tr}\,((X-A)X)^{\mathrm{T}}\Delta \end{aligned} \tag{3.10}$$

Thus (3.9) is equivalent to

$$\begin{aligned} 4\,\mathrm{tr}\,(((X-A)X)^{\mathrm{T}}\Delta) &= \langle\langle \mathrm{grad} f_A(X), \{\Delta, X\}\rangle\rangle \\ &= \langle\langle \{\Omega, X\}, \{\Delta, X\}\rangle\rangle \\ &= \mathrm{tr}((\Omega^X)^{\mathrm{T}}\Delta^X) \\ &= \mathrm{tr}(\Omega^{\mathrm{T}}\Delta^X) \end{aligned} \tag{3.11}$$

since $\Omega \in (\ker \pi_X)^{\perp}$ implies $\Omega = \Omega^X$. For all $\Delta \in \ker \pi_X$

$$(X(X-A)\Delta) = \frac{1}{2}\,\mathrm{tr}\,[(X-A)(\Delta X + X\Delta^{\mathrm{T}})] = 0$$

and therefore $(X-A)X \in (\ker \pi_X)^{\perp}$. Thus

$$\begin{aligned} 4\,\mathrm{tr}\,[((X-A)X)^{\mathrm{T}}\Delta] &= 4\,\mathrm{tr}\,[((X-A)X^{\mathrm{T}}(\Delta_X + \Delta^X)] \\ &= 4\,\mathrm{tr}\,[((X-A)X)^{\mathrm{T}}\Delta^X)] \end{aligned}$$

and (3.11) is equivalent to

$$\Omega = 4(X-A)X$$

Thus

$$\mathrm{grad} f_A(X) = 4\{(X-A)X, X\} \tag{3.12}$$

which proves (i).

For (ii) note that for any $X \in S(n,N)$ we have $\{X, X(X-A)\} \in T_X S(n,N)$ and thus (3.8) is a vector field on $S(n,N)$. Thus for any initial condition $X(0) \in S(n,N)$ the solution $X(t)$ of (3.8) satisfies $X(t) \in S(n,N)$ for all t for which $X(t)$ is defined. It suffices therefore to show the existence of solutions of (3.8) for all $t \geq 0$. To this end consider any solution $X(t)$ of (3.8). By

$$\begin{aligned} &\tfrac{\mathrm{d}}{\mathrm{d}t} f_A(X(t)) = 2\,\mathrm{tr}\,((X-A)\dot{X}) = 8\,\mathrm{tr}\,[(X-A)\{(A-X)X, X\}] \\ &= -16\,\mathrm{tr}\,((A-X)^2X^2) = -16\,\|(A-X)X\|^2 \end{aligned} \tag{3.13}$$

(since $\mathrm{tr}\,(A\{B,C\}) = 2\,\mathrm{tr}\,(BCA)$ for $A = A^{\mathrm{T}}$). Thus $f_A(X(t))$ decreases monotonically and the equilibria points of (3.8) are characterized by $(X-A)X = 0$. Also

$$\|A - X(t)\| \leq \|A - X(0)\|$$

and $X(t)$ stays in the compact set $\{X \in \mathbb{R}^{N\times N} \mid \|A - X\| \leq \|A - X(0)\|\}$. Thus $X(t)$ exists for all $t \geq 0$ and the result follows. □

Remark. An important consequence of Theorem 3.2 is that the differential equation (3.8) on the vector space of symmetric $N \times N$ matrices is rank preserving, i.e. $\operatorname{rank} X(t) = \operatorname{rank} X(0)$ for all $t \geq 0$. Also, $X(t)$ always converges in the spaces of symmetric matrices to some symmetric matrix $X(\infty)$ as $t \to \infty$ and hence $\operatorname{rk} X(\infty) \leq \operatorname{rk} X(0)$. Here $X(\infty)$ is a critical point of $f_A : S(n, N) \to \mathbb{R}$, $n \leq \operatorname{rk} X(0)$.

In order to investigate the local stability properties of the equilibrium points of (3.8) we determine the eigenvalues of the linearizations of (3.8). We consider both situations where the flow (3.8) is regarded as evolving on the constraint set $S(n, N)$ as well as the unconstrained case where the state space is $S(N)$.

Let $A := \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, $\lambda_1 \geq \ldots \geq \lambda_N$. The tangent space of $S(n, N)$ at an equilibrium point $X_\infty = \operatorname{diag}(\varepsilon_1 \lambda_1, \ldots, \varepsilon_N \lambda_N)$, $\varepsilon_i \in \{0, 1\}$, $\sum_{i=1}^N \varepsilon_i = n$ is given by

$$T_{X_\infty} S(n, N) = \{\xi \in \mathbb{R}^{N \times N} \text{ symmetric} \mid \xi_{ij} = 0 \text{ for } \varepsilon_i = \varepsilon_j = 0\}.$$

Let $X_\infty = \operatorname{diag}(\varepsilon_1 \lambda_1, \ldots, \varepsilon_N \lambda_N)$, $\varepsilon_i \in \{0, 1\}$, $\sum_{i=1}^N \varepsilon_i = n$ be an equilibrium point of $f_A : S(n, N) \to \mathbb{R}$. Let $I^- := \{i \mid \lambda_i < 0, \varepsilon_i = 1\}$, $I^+ := \{i \mid \lambda_i > 0, \varepsilon_i = 1\}$ and $I := I^- \cup I^+$.

Proposition 3.3. (i) The linearization of (3.8) at X_∞ is

$$\dot{\xi} = (A - X_\infty)\xi X_\infty + X_\infty \xi (A - X_\infty) - \xi X_\infty^2 - X_\infty^2 \xi, \quad \xi \in S(N), \tag{3.14}$$

or in matrix entries $\dot{\xi}_{ij} = \mu_{ij} \xi_{ij}$, $N \geq i \geq j \geq 1$, with

$$\mu_{ij} = -((\varepsilon_i \lambda_i + \varepsilon_j \lambda_j)^2 - (\varepsilon_i + \varepsilon_j)\lambda_i \lambda_j) = \\ = \begin{cases} -(\lambda_i^2 + \lambda_j^2) & \varepsilon_i = \varepsilon_j = 1 \\ \lambda_i(\lambda_j - \lambda_i) & \varepsilon_i = 1, \varepsilon_j = 0 \\ -\lambda_j(\lambda_j - \lambda_i) & \varepsilon_i = 0, \varepsilon_j = 1 \\ 0 & \varepsilon_i = \varepsilon_j = 0 \end{cases}$$

(ii) The linearization (3.14) restricted to the orthogonal complement of the tangent space $T_{X_\infty} S(n, N)$ in $S(n, N)$ is identically 0, and on the tangent space $T_{X_\infty} S(n, N)$ the eigenvalues satisfy

$\mu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \\ \text{(b)} & i \in I^-, \ j \notin I \text{ and } \lambda_i \neq \lambda_j \\ \text{(c)} & i \notin I, \ j \in I^+ \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

$\mu_{ij} = 0$ if and only if $\varepsilon_i + \varepsilon_j = 1$ and $\lambda_i = \lambda_j$

$\mu_{ij} > 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i \in I^+, \ j \notin I \text{ and } \lambda_i \neq \lambda_j \\ \text{(b)} & i \notin I, \ j \in I^- \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

Proof. The linearization is given by the derivative of the map

$$X \mapsto (A-X)X^2 + X^2(A-X)$$

at X_∞, i.e. by the linear map $S(N) \to S(N)$ defined by

$$\xi \mapsto (A-X_\infty)(\xi X_\infty + X_\infty \xi) - \xi X_\infty^2 + (\xi X_\infty + X_\infty \xi)(A - X_\infty) - X_\infty^2 \xi.$$

Together with $(A - X_\infty)X_\infty = 0$ this shows (3.14). For the (i,j)-component of (3.14) one computes

$$\dot{\xi}_{ij} = (\lambda_i + \varepsilon_i\lambda_i)\varepsilon_i\xi_{ij}\lambda_j + \varepsilon_i\xi_{ij}\lambda_i(\lambda_j + \varepsilon_j\lambda_j) - \varepsilon_j\xi_{ij}\lambda_j^2 - \varepsilon_i\xi_{ij}\lambda_i^2 = \mu_{ij}\xi_{ij}.$$

From this (ii) easily follows. □

Remark. For $X_\infty \geq 0$, i.e. $I^- = \emptyset$, $I = I^+$, the eigenvalues of the linearization (3.14) on the tangent space $T_{X_\infty}S(n,N)$ satisfy

$\mu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i,j \in I \\ \text{(b)} & i \notin I,\ j \in I \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

$\mu_{ij} = 0$ if and only if $\varepsilon_i + \varepsilon_j = 1$ and $\lambda_i = \lambda_j$

$\mu_{ij} > 0$ if and only if $i \in I$, $j \notin I$ and $\lambda_i \neq \lambda_j$.

Remark. Let $p = \#I^+$, $q = \#I^-$, i.e. $X_\infty \in S(p,q;N)$. Let $N_+ := \dim \mathrm{Eig}^+(A)$ and $N_- := \dim \mathrm{Eig}^-(A)$. If A is invertible with distinct eigenvalues, then $T_{X_\infty}S(n,N)$ is exactly the maximal subspace of $S(N)$ on which the restriction of (μ_{ij}) is invertible.

Furthermore, the number of negative eigenvalues of the linearization (3.14) is given by

$$\begin{aligned} &\textstyle\sum_{i\in I^+}(N+1-i) + \sum_{i \in I^-}(i-p) = \\ &q(N_+ - p) + p(N_- - q) + pq + \textstyle\sum_{i\in I^+}(N_+ + 1 - i) + \sum_{i\in I^-}(i - N_+) \end{aligned} \tag{3.15}$$

and the number of positive eigenvalues is given by

$$\sum_{i\in I^+} i - \frac{p(p+1)}{2} + \sum_{i \in I^-}(N+1-i) - \frac{q(q+1)}{2} \tag{3.16}$$

This result coincides with the index formula appearing in Theorem 2.5 (ii). □

The linearization (3.14) at an equilibrium point $X_\infty \in S(p,q;N)$ is stable if

$$X_\infty = (\lambda_1, \dots, \lambda_p, 0, \dots, 0, \lambda_{N-q+1}, \dots, \lambda_N).$$

In particular, X_∞ is in this case a minimum of $f_A|S(p,q;N)$.

The linearization (3.14) at $X_\infty \in S(p,q;N)$ is asymptotically stable on the tangent space $T_{X_\infty}S(p,q;N)$ if and only if $X_\infty = (\lambda_1,\ldots,\lambda_p,0,\ldots,0,\lambda_{N-q+1},\ldots,\lambda_N)$ and $\lambda_p > \lambda_{p+1}$, $\lambda_{N-q} > \lambda_{N-q+1}$. In particular, X_∞ is in this case the unique minimum of $f_A|S(p,q;N)$. This shows that, if A has distinct eigenvalues, the distance function $f_A: S(n,N) \to \mathbb{R}$ has exactly $n+1$ local minima.

The following table illustrates the above results in the case $N = 2$. It lists the possible sign distributions of the eigenvalues.

Table 1. Signs of eigenvalues of the linearization (3.14) on $T_{X_\infty}S(n,2)$. Remember that the restriction of (3.14) on $\left(T_{X_\infty}S(n,2)\right)^\perp$ is zero.

	X_∞	diag$(\lambda_1,0)$	diag$(0,\lambda_2)$	diag(λ_1,λ_2)
(1) $\lambda_1 > \lambda_2 > 0$	$\mu_{ij} < 0$	μ_{11}, μ_{21}	μ_{22}	μ_{11}, μ_{21}, μ_{22}
	$\mu_{ij} > 0$	—	μ_{21}	—
(2) $\lambda_1 > 0 > \lambda_2$	$\mu_{ij} < 0$	μ_{11}, μ_{21}	μ_{21}, μ_{22}	μ_{11}, μ_{21} μ_{22}
	$\mu_{ij} > 0$	—	—	—
(3) $0 > \lambda_1 > \lambda_2$	$\mu_{ij} < 0$	μ_{11}	μ_{21}, μ_{22}	μ_{11}, μ_{21}, μ_{22}
	$\mu_{ij} > 0$	μ_{21}	—	—

4. A RICCATI FLOW

By Lemma 3.1, every rank preserving flow on $S(N)$ is of the form

$$\dot{X} = F(X)X + XF(X)^{\mathrm{T}}.$$

Thus the Riccati differential equation

$$\dot{X} = (A - X)X + X(A - X)$$

appears to be the simplest possible candidate for a rank preserving flow on $S(N)$ which has the same set of equilibria as the gradient flow (3.8). Moreover, the restriction of the gradient flow (3.8) on the subclass of projection operators, characterized by $X^2 = X$, is exactly the above Riccati equation. This motivates us to consider the above Riccati equation in more detail. As we will see, the situation is particularly transparent for positive definite matrices X.

Theorem 4.1. Let $A \in \mathbb{R}^{N\times N}$ be symmetric.

(i) The Riccati equation

$$\dot{X} = (A - X)X + X(A - X), \quad X(0) \in S(n,N), \tag{4.1}$$

defines a rank preserving flow on $S(n,N)$.

(ii) Assume A is invertible. Then the solutions $X(t)$ of (4.1) are given by

$$X(t) = e^{tA} X_0 [I_N + A^{-1}(e^{2At} - I_N)X_0]^{-1} e^{tA} \tag{4.2}$$

(iii) For any positive semidefinite initial condition $X(0) = X(0)^{\mathrm{T}} \geq 0$, the solution $X(t)$ of (4.1) exists for all $t \geq 0$ and is positive semidefinite.

(iv) Every positive semidefinite solution $X(t) \in S^+(n, N) = S(n, 0; N)$ of (4.1) converges to a connected component of the set of equilibrium points, characterized by $(A - X_\infty)X_\infty = 0$. Also X_∞ is positive semidefinite and has rank $\leq n$. If A has distinct eigenvalues then every positive semidefinite solution $X(t)$ converges to an equilibrium point.

Proof. (i) is an immediate consequence of Lemma 3.1. To prove (ii) it suffices to show that $X(t)$ defined by (4.2) satisfies the Riccati equation. By differentiation of (4.2) we obtain

$$\begin{aligned} \dot{X}(t) &= AX(t) + X(t)A - 2e^{tA} X_0 [I_n + A^{-1}(e^{2At} - I_n)X_0]^{-1} e^{2At} X_0 [\ldots]^{-1} e^{tA} \\ &= AX(t) + X(t)A - 2X(t)^2, \end{aligned}$$

which shows the claim.

For (iii) note that (i) implies that $X(t) \in S^+(n, N)$ for all $t \in [0, t_{\max}]$. Thus it suffices to show that $X(t)$ exists for all $t \geq 0$; i.e. $t_{\max} = \infty$. This follows from a simple Lyapunov argument. First, we note that the set $\overline{S^+(n, N)}$ of positive semidefinite matrices X of rank $\leq n$ is a closed subset of $S(N)$. Consider the distance function $f_A : S(N) \to \mathbb{R}_+$ defined by $f_A(X) = \|A - X\|^2$. Thus f_A is a proper function of $S(N)$ and hence also on $\overline{S^+(n, N)}$. For every positive semidefinite solution $X(t)$, $t \in [0, t_{\max}]$, let $X(t)^{1/2}$ denote the unique positive semidefinite symmetric square root. A simple computation shows

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f_A(X(t)) &= -4 \operatorname{tr}[(A - X(t))\dot{X}(t)] \\ &= -4\|(A - X(t))X(t)^{\frac{1}{2}}\|^2 \leq 0. \end{aligned}$$

This f_A is a Lyapunov function for (4.1), restricted to the class of positive semidefinite matrices, and equilibrium points $X_\infty \in \overline{S^+(n, N)}$ are characterized by $(A - X_\infty)X_\infty = 0$. In particular, $f_A(X(t))$ is a monotonically decreasing function of t and the solution $X(t)$ stays in the compact subset

$$\{X \in \overline{S^+(n, N)} \mid f_A(X) \leq f_A(X(0))\}.$$

Thus $X(t)$ is defined for all $t \geq 0$. By LaSalle's principle of invariance, the ω-limit set of $X(t)$ is a connected component of the set of positive definite equilibrium points. If A has distinct eigenvalues, then the set of positive definite equilibrium points is finite. Thus the result follows. □

Remark. The above proof shows that the least squares distance function $f_A(X) = \|A - X\|^2$ is a Lyapunov function for the Riccati equation, evolving on the subset of positive semidefinite matrices X. In particular, the Riccati equation exhibits gradient-like behaviour if restricted to positive semidefinite initial conditions. If X_0 is an indefinite matrix then also $X(t)$ is indefinite, and $f_A(X)$ is no longer a Lyapunov function. □

The local stability properties of the Riccati equation (4.1) around an equilibrium point X_∞ satisfying $(A - X_\infty)X_\infty = 0$ are analyzed in the next result.

Let $A := \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, $\lambda_1 \geq \ldots \geq \lambda_N$ and let $X_\infty = \operatorname{diag}(\varepsilon_1\lambda_1, \ldots, \varepsilon_N\lambda_N)$, $\varepsilon_i \in \{0, 1\}$, $\sum_{i=1}^N \varepsilon_i = n$ be an equilibrium point of $f_A\colon S(n, N) \to \mathbb{R}$.

Let $I^- := \{i \mid \lambda_i < 0, \varepsilon_i = 1\}$, $I^+ := \{i \mid \lambda_i > 0, \varepsilon_i = 1\}$ and $I := I^- \cup I^+$.

Proposition 4.2. (i) The linearization of the Riccati equation (4.1) at X_∞ is

$$\dot{\xi} = (A - 2X_\infty)\xi + \xi(A - 2X_\infty), \quad \xi \in S(N), \tag{4.3}$$

or in matrix entries $\dot{\xi}_{ij} = \nu_{ij}\xi_{ij}$, $N \geq i \geq j \geq 1$, with

$$\nu_{ij} = (1 - 2\varepsilon_i)\lambda_i + (1 - 2\varepsilon_j)\lambda_j =$$
$$= \begin{cases} -(\lambda_i + \lambda_j) & \varepsilon_i = \varepsilon_j = 1 \\ (\lambda_j - \lambda_i) & \varepsilon_i = 1, \varepsilon_j = 0 \\ -(\lambda_j - \lambda_i) & \varepsilon_i = 0, \varepsilon_j = 1 \\ \lambda_i + \lambda_j & \varepsilon_i = \varepsilon_j = 0 \end{cases}$$

(ii) The eigenvalues of the linearization on the tangent space $T_{X_\infty}S(n, N)$ satisfy

$\nu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \text{ and } \lambda_i > -\lambda_j \\ \text{(b)} & i \notin I,\ j \in I \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

$\nu_{ij} = 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \text{ and } \lambda_i = -\lambda_j \\ \text{(b)} & \varepsilon_i + \varepsilon_j = 1 \text{ and } \lambda_i = \lambda_j \end{cases}$$

$\nu_{ij} > 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \text{ and } \lambda_i < -\lambda_j \\ \text{(b)} & i \in I,\ j \notin I \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

(iii) The eigenvalues of the linearization on the whole space $S(N)$ satisfy

$\nu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \text{ and } \lambda_i > -\lambda_j \\ \text{(b)} & i \notin I,\ j \in I \text{ and } \lambda_i \neq \lambda_j \\ \text{(c)} & \lambda_i < -\lambda_j \text{ and } i, j \notin I \end{cases}$$

$\nu_{ij} = 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i, j \in I \text{ and } \lambda_i = -\lambda_j \\ \text{(b)} & \varepsilon_i + \varepsilon_j = 1 \text{ and } \lambda_i = \lambda_j \\ \text{(c)} & \varepsilon_i + \varepsilon_j = 1 \text{ and } \lambda_i = \lambda_j \end{cases}$$

$\nu_{ij} > 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i,j \in I \text{ and } \lambda_i < -\lambda_j \\ \text{(b)} & i \in I,\ j \notin I \text{ and } \lambda_i \neq \lambda_j \\ \text{(c)} & i,j \notin I \text{ and } \lambda_i > -\lambda_j \end{cases}$$

Proof. The linearization is given by the derivative of the map $X \mapsto (A-X)X + X(A-X)$ at X_∞, i.e. by the linear map

$$\xi \mapsto (A - X_\infty)\xi - \xi X_\infty + \xi(A - X_\infty) - X_\infty \xi.$$

This shows (4.3). For the (i,j)-component of (4.3) one computes

$$\dot{\xi}_{ij} = (\lambda_i - 2\varepsilon_i \lambda_i)\xi_{ij} + \xi_{ij}(\lambda_j - 2\varepsilon_j \lambda_j) = \nu_{ij}\xi_{ij}.$$

From this (ii) and (iii) easily follow. □

Remark. For $X_\infty \geq 0$, i.e. $I^- = \emptyset$, $I = I^+$, we have on the tangent space $T_{X_\infty} S(n, N)$

$\nu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i,j \in I \\ \text{(b)} & i \notin I,\ j \in I \text{ and } \lambda_i \neq \lambda_j \end{cases}$$

$\nu_{ij} = 0$ if and only if $\varepsilon_i + \varepsilon_j = 1$ and $\lambda_i = \lambda_j$

$\nu_{ij} > 0$ if and only if $i \in I$, $j \notin I$ and $\lambda_i \neq \lambda_j$.

On the whole space $S(N)$ we have for $X_\infty \geq 0$

$\nu_{ij} < 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i,j \in I \\ \text{(b)} & i \notin I,\ j \in I \text{ and } \lambda_i \neq \lambda_j \\ \text{(c)} & i,j \notin I \text{ and } \lambda_i < -\lambda_j \end{cases}$$

$\nu_{ij} = 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & \varepsilon_i + \varepsilon_j = 1 \text{ and } \lambda_i = \lambda_j \\ \text{(b)} & i,j \notin I \text{ and } \lambda_i = -\lambda_j \end{cases}$$

$\nu_{ij} > 0$ if and only if one of the following conditions hold:

$$\begin{cases} \text{(a)} & i \in I,\ j \notin I \text{ and } \lambda_i \neq \lambda_j \\ \text{(b)} & i,j \notin I \text{ and } \lambda_i > -\lambda_j. \end{cases}$$

Remark. Let $p = \#I^+$, $q = \#I^-$, i.e. $X_\infty \in S(p,q;N)$. Let $N_+ := \dim \operatorname{Eig}^+(A)$ and $N_- := \dim \operatorname{Eig}^-(A)$. Let A be invertible with distinct eigenvalues. The number of negative eigenvalues of the linearization (4.3) on the tangent space $T_{X_\infty} S(p,q;N)$ is given by

$$p(N_- - q) + \sum_{i \in I^+} (N_+ + 1 - i) + \sum_{i \in I^-} (N + 1 - i) - \frac{q(q+1)}{2} + \sum_{i \in I^-} \#\{j \in I^+ \mid \lambda_j > -\lambda_i\}, \quad (4.4)$$

the number of eigenvalues equal zero is given by

$$\sum_{i\in I^-} \#\{j \in I^+ \mid \lambda_j = -\lambda_i\} \tag{4.5}$$

and the number of positive eigenvalues is given by

$$q(N_+ - p) + \sum_{i\in I^+} i - \frac{p(p+1)}{2} + \sum_{i\in I^-}(i - N_+) + \sum_{i\in I^-} \#\{j \in I^+ \mid \lambda_j < -\lambda_i\} \tag{4.6}$$

Note that the numbers of negative, positive and zero eigenvalues for the linearizations (3.14) and (4.3) respectively are in general different. However, if X_∞ is positive semidefinite, then the numbers of eigenvalues coincide respectively. Comparing formulas (4.4), (4.5) and (4.6) with (3.15) and (3.16) shows that most summands do appear in both cases, except for $pq = \sum_{i\in I^-} \#\{j \in I^+ \mid \lambda_j > -\lambda_i\} + \sum_{i\in I^-} \#\{j \in I^+ \mid \lambda_j = -\lambda_i\} + \sum_{i\in I^-} \#\{j \in I^+ \mid \lambda_j < -\lambda_i\}$.

The linearization (4.3) for the Riccati flow restricted to $S(p,q;N)$ at an equilibrium point $X_\infty \in S(p,q;N)$ is stable on the tangent space $T_{X_\infty}S(p,q;N)$ if $q = 0$ and $X_\infty = (\lambda_1,\ldots,\lambda_p,0,\ldots,0)$. In particular, X_∞ is in this case a local minimum of $f_A|S(p,0;N)$. The linearization (4.3) at $X_\infty \in S(p,q;N)$ is asymptotically stable on the tangent space $T_{X_\infty}S(p,q;N)$ if and only if $q = 0$, $X_\infty = (\lambda_1,\ldots,\lambda_p,0,\ldots,0)$ and $\lambda_p > \lambda_{p+1}$. In particular, X_∞ is in this case the unique minimum of $f_A|S(p,0;N)$.

The linearization (4.3) of the Riccati flow on $S(N)$ at an equilibrium point $X_\infty \in S(p,q;N)$ is stable if $q = 0$, $p = N_+ = \dim \mathrm{Eig}^+(A)$ and $X_\infty = (\lambda_1,\ldots,\lambda_p,0,\ldots,0)$. In this case, the linearization is even asymptotically stable on $S(N)$. Thus the (standard) stability properties of the Riccati equation are different if they are considered on $S(N)$ or on the submanifold $S(n,N)$.

To illustrate the above results we consider the simplest nontrivial case, where $N = 2$. The eigenvalues of the linearization (4.3) on $S(2)$, and their signs, are given by Table 2.

For diagonal matrix initial data $X_0 = \mathrm{diag}(a_0, b_0)$ the solutions of (4.1) are given by $X = (a,b)$, where

$$\begin{aligned} a(t) &= \frac{\lambda_1 a_0}{a_0 + (\lambda_1 - a_0)e^{-2\lambda_1 t}} \\ b(t) &= \frac{\lambda_2 b_0}{b_0 + (\lambda_2 - b_0)e^{-2\lambda_2 t}} \end{aligned}$$

The associated phase portrait of (4.1) on the 2-dimensional subspace of diagonal matrices is given in Figure 1. The equilibrium points P_i there are:

$$\begin{array}{lllll} P_1 = (0,\lambda_2), & P_2 = (\lambda_1,\lambda_2), & P_3 = (0,0), & P_4 = (\lambda_1,0) & \text{if } \lambda_1 > \lambda_2 > 0 \\ P_1 = (0,0), & P_2 = (\lambda_1,0), & P_3 = (0,\lambda_2), & P_4 = (\lambda_1,\lambda_2) & \text{if } \lambda_1 > 0 > \lambda_2 \\ P_1 = (0,\lambda_2), & P_2 = (0,0), & P_3 = (\lambda_1,\lambda_2), & P_4 = (\lambda_1,0) & \text{if } 0 > \lambda_1 > \lambda_2 \end{array}$$

Table 2. Signs of eigenvalues of the linearization (4.3) on $S(2)$. Eigenvalues corresponding to the restriction on $T_{X_\infty}S(n,2)$ are underlined.

	X_∞	$\operatorname{diag}(\lambda_1, 0)$	$\operatorname{diag}(0, \lambda_2)$	$\operatorname{diag}(\lambda_1, \lambda_2)$
(1) $\lambda_1 > \lambda_2 > 0$	$\nu_{ij} < 0$	$\underline{\nu_{11}}, \underline{\nu_{21}}$	$\underline{\nu_{22}}$	$\underline{\nu_{11}}, \underline{\nu_{21}}, \underline{\nu_{22}}$
	$\nu_{ij} > 0$	ν_{22}	$\nu_{11}, \underline{\nu_{21}}$	—
(2) $\lambda_1 > 0 > \lambda_2$				
(2a) $\lambda_1 > -\lambda_2$	$\nu_{ij} < 0$	$\underline{\nu_{11}}, \underline{\nu_{21}}, \nu_{22}$	—	$\underline{\nu_{11}}, \underline{\nu_{21}}$
	$\nu_{ij} > 0$	—	$\nu_{11}, \underline{\nu_{21}}, \underline{\nu_{22}}$	$\underline{\nu_{22}}$
(2b) $\lambda_1 = -\lambda_2$	$\nu_{ij} < 0$	$\underline{\nu_{11}}, \underline{\nu_{21}}, \nu_{22}$	—	$\underline{\nu_{11}}$
	$\nu_{ij} > 0$	—	$\nu_{11}, \underline{\nu_{21}}, \underline{\nu_{22}}$	$\underline{\nu_{22}}$
	$\nu_{ij} = 0$	—	—	$\underline{\nu_{21}}$
(2c) $\lambda_1 < -\lambda_2$	$\nu_{ij} < 0$	$\underline{\nu_{11}}, \underline{\nu_{21}}, \nu_{22}$	—	$\underline{\nu_{11}}$
	$\nu_{ij} > 0$	—	$\nu_{11}, \underline{\nu_{21}}, \underline{\nu_{22}}$	$\underline{\nu_{21}}, \underline{\nu_{22}}$
(3) $0 > \lambda_1 > \lambda_2$	$\nu_{ij} < 0$	$\underline{\nu_{21}}, \nu_{22}$	ν_{11}	—
	$\nu_{ij} > 0$	$\underline{\nu_{11}}$	$\underline{\nu_{21}}, \underline{\nu_{22}}$	$\underline{\nu_{11}}, \underline{\nu_{21}}, \underline{\nu_{22}}$

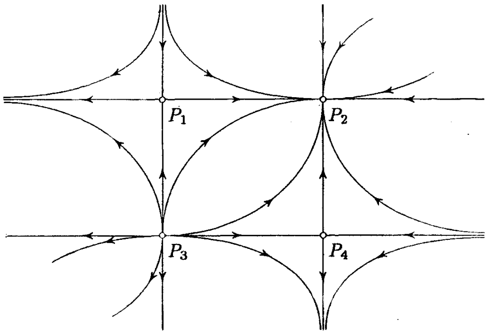

Fig. 1. Phase Portrait of the Riccati flow on diagonal matrices.

The following Figure 2 illustrates the phase portrait of (4.3) on $S(2)$ for $A = A^{\mathrm{T}}$ positive semidefinite. Only a part of the complete phase portrait is shown here, concentrating on the cone of positive semidefinite matrices in $S(2)$. There are two equilibrium points on $S(1,2)$. For the flow on $S(2)$, both equilibria are saddle points, one having 1 positive and 2 negative eigenvalues while the other one has 2 positive and 1 negative eigenvalues. The induced flow on $S(1,2)$ has one equilibrium as a local attractor while the other one is a saddle point.

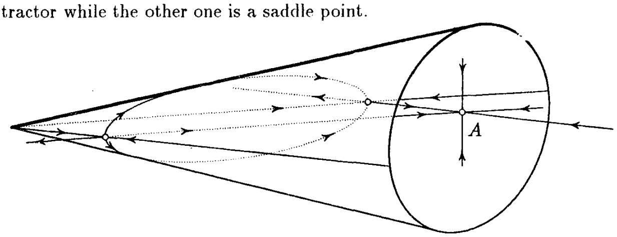

Fig. 2. Riccati flow on positive semidefinite matrices.

5. GRADIENT FLOWS FOR POSITIVE SEMIDEFINITE APPROXIMANTS

So far we have considered gradient like flows for the least squares distance function $f_A(X)$, evolving on spaces of fixed rank symmetric matrices. We now focus on the approximation problem by positive semidefinite matrices. As we have seen, the above flows may not always have particularly attractive stability properties when considered in the entire space $S(N)$ of symmetric matrices. This motivates us to consider the differential equation

$$\dot{X} = (A - XX^{\mathrm{T}})X, \quad X(0) \in \mathbb{R}^{N \times n} \tag{5.1}$$

for $A = A^{\mathrm{T}} \in \mathbb{R}^{N \times N}$ real symmetric. The motivation for studying the flow (5.1) on $\mathbb{R}^{N \times n}$ lies in the observation that it induces the Riccati equation (4.1) on $S^{+}(n, N)$. In fact, for $H = XX^{\mathrm{T}}$ we obtain

$$\dot{H} = \dot{X}X^{\mathrm{T}} + X\dot{X}^{\mathrm{T}} = (A - H)H + H(A - H)$$

for any solution $X(t)$ of (5.1). Thus the parametrization of positive semidefinite matrices H by $X \mapsto XX^{\mathrm{T}} = H$ maps solutions of (5.1) onto solutions of the Riccati equation. The basic convergence properties of the flow (5.1) are stated in the next result.

Theorem 5.1. Let $A \in \mathbb{R}^{N \times N}$ be symmetric.

(i) The differential equation

$$\dot{X} = (A - XX^{\mathrm{T}})X \quad , \quad X(0) \in \mathbb{R}^{N \times n}, \tag{5.1}$$

defines a rank preserving flow on $\mathbb{R}^{N \times n}$.

(ii) The solutions $X(t)$ of (5.1) exist for all $t \geq 0$ and converge for $t \rightarrow +\infty$ to a connected component of the set of equilibrium points X_∞ of rank $\mathrm{rk} X_\infty \leq n$. Equilibrium points X_∞ of (5.1) are characterized by $(A - X_\infty X_\infty^{\mathrm{T}})X_\infty = 0$ or, equivalently, by the condition that the columns of X_∞ generate an A-invariant subspace of $\mathbb{R}^N$. If $n = 1$ and A has distinct eigenvalues, then every solution $X(t)$ converges to an equilibrium point.

(iii) The function $V\colon \mathbb{R}^{N\times n} \to \mathbb{R}$, $V(X) := \|A - XX^{\mathrm{T}}\|^2$, is a Lyapunov function for (5.1).

Proof. (i) Let $M_k \subset \mathbb{R}^{N\times n}$ be the manifold of all rank k real $N \times n$ matrices. For any $X \in M_k$ the right hand side of (5.1) $AX - X(X^{\mathrm{T}}X)$ is a tangent vector of M_k at X. Thus the flow (5.1) on $\mathbb{R}^{N\times n}$ leaves M_k invariant for $0 \le k \le n$. This shows (i).

(ii) The arguments here are similar to those for Theorems 3.2, 4.1. Explicitly for $V(X) = \|A - XX^{\mathrm{T}}\|^2$, then

$$\begin{aligned} \tfrac{\mathrm{d}V(X(t))}{\mathrm{d}t} &= -4\,\mathrm{tr}\,[(A - XX^{\mathrm{T}})\dot{X}X^{\mathrm{T}}] \\ &= -4\,\mathrm{tr}\,\|(A - XX^{\mathrm{T}})X\|^2 \le 0. \end{aligned}$$

Thus $V(\cdot)$ is a proper Lyapunov function on $\mathbb{R}^{N\times n}$ and the result follows similarly as for Theorems 3.2, 4.1. □

In order to analyze the local stability properties of (5.1) we need the following characterization of equilibrium points.

Lemma 5.2. Let $A = \mathrm{diag}(\lambda_1, \dots, \lambda_N), \lambda_1 > \dots > \lambda_N$. An equilibrium point X_∞ of (5.1) is characterized by $X_\infty X_\infty^{\mathrm{T}} = \mathrm{diag}(\varepsilon_1\lambda_1, \dots, \varepsilon_N\lambda_N), \varepsilon_i \in \{0,1\}$. In particular, the rank $\mathrm{rk}X_\infty$ can be at most $\dim \mathrm{Eig}^+(A)$.

Proof. We must show that $(A - X_\infty X_\infty^{\mathrm{T}})X_\infty = 0$ if and only if $X_\infty X_\infty^{\mathrm{T}} = \mathrm{diag}(\varepsilon_1\lambda_1, \dots, \varepsilon_N\lambda_N)$ for some $\varepsilon_i \in \{0,1\}$. Let $AX_\infty = X_\infty X_\infty^{\mathrm{T}} X_\infty$. Then $AX_\infty X_\infty^{\mathrm{T}} = X_\infty X_\infty^{\mathrm{T}} X_\infty X_\infty^{\mathrm{T}}$, which implies $X_\infty X_\infty^{\mathrm{T}} A = AX_\infty X_\infty^{\mathrm{T}}$. Since A has distinct eigenvalues, $X_\infty X_\infty^{\mathrm{T}}$ must be diagonal, say $X_\infty X_\infty^{\mathrm{T}} = \mathrm{diag}(\nu_1, \dots, \nu_N)$. From $AX_\infty = \mathrm{diag}(\nu_1, \dots, \nu_N)X_\infty$ we get $\nu_i = \varepsilon_i\lambda_i$. □

Proposition 5.3. Let $A = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$, $\lambda_1 \ge \dots \ge \lambda_N$, A invertible. Let $X_\infty \in \mathbb{R}^{N\times n}$ be an equilibrium point of (5.1), $X_\infty X_\infty^{\mathrm{T}} = \mathrm{diag}(\varepsilon_1\lambda_1, \dots, \varepsilon_N\lambda_N)$, $\varepsilon_i \in \{0,1\}$, $\sum \varepsilon_i = r = \mathrm{rk}X_\infty$.

(i) The linearization of (5.1) at X_∞ is given by

$$\dot{\xi} = F(\xi) := (A - X_\infty X_\infty^{\mathrm{T}})\xi - \xi X_\infty^{\mathrm{T}} X_\infty - X_\infty \xi^{\mathrm{T}} X_\infty \tag{5.2}$$

(ii) The eigenvalues of F are

$$\begin{array}{ll} 0 & \text{with multiplicity } \frac{r(2n-r-1)}{2} \\ \lambda_i & \text{with multiplicity } n-r \text{ for } 1 \le i \le N,\ \varepsilon_i = 0 \\ \lambda_i - \lambda_j & \text{for } 1 \le i,j \le N,\ \varepsilon_i = 0,\ \varepsilon_j = 1 \\ -\lambda_i - \lambda_j & \text{for } 1 \le i \le j \le N,\ \varepsilon_i = \varepsilon_j = 1. \end{array}$$

(iii) The number of negative eigenvalues of F is

$$\frac{r(r+1)}{2} + (n-r)\cdot \dim \mathrm{Eig}^-(A) + \#\{(i,j) \mid \lambda_i < \lambda_j, \varepsilon_i = 0, \varepsilon_j = 1\},$$

the number of eigenvalues equal to 0 is

$$\frac{r(2n-r-1)}{2} + \#\{(i,j) \mid \lambda_i = \lambda_j, \varepsilon_i = 0, \varepsilon_j = 1\}$$

and the number of positive eigenvalues is

$$(n-r)\#\{i \mid \lambda_i > 0, \varepsilon_i = 0\} + \#\{(i,j) \mid \lambda_i > \lambda_j, \varepsilon_i = 0, \varepsilon_j = 1\}$$

Proof. (i) is straightforward. (ii) Let $x_1, \dots, x_N$ denote the row vectors of X_∞. Let $X(i,j)$ denote the $N \times n$-matrix whose ith row vector is x_j and 0 otherwise. Let V denote the subvector space of $\mathbb{R}^{N\times n}$ generated by $\big(X(i,j) \mid 1 \le i,j \le N\big)$. First we consider $F|V$. To begin with, note that $X(i,j) = 0$ if $\varepsilon_j = 0$ and that $\big(X(i,j) \mid 1 \le i,j \le N, \varepsilon_j = 1\big)$ is a basis of V. It is easy to see that for $\varepsilon_j = 1$ we have $X(i,j)X_\infty^{\mathrm{T}} = \lambda_j E_{ij}$, where E_{ij} denotes the $N \times N$-matrix with (i,j)-entry 1 and 0 else. Therefore we get

$$\begin{aligned} F(X(i,j)) &= (A - X_\infty X_\infty^{\mathrm{T}})X(i,j) - \lambda_j E_{ij} X_\infty - \lambda_j E_{ji} X_\infty \\ &= (1-\varepsilon_i)\lambda_i X(i,j) - \lambda_j X(i,j) - \lambda_j X(j,i) \end{aligned}$$

and thus

$$F(E_{ij}X_\infty) = \begin{cases} (\lambda_i - \lambda_j)X(i,j) & \varepsilon_i = 0 \\ -\lambda_j(X(i,j) + X(j,i)) & \varepsilon_i = 1. \end{cases}$$

This shows that V is invariant under F and the eigenvalues of $F|V$ are

$$\begin{array}{cl} \lambda_i - \lambda_j & \text{for} \quad \varepsilon_i = 0,\ \varepsilon_j = 1,\ 1 \le i,j \le N \\ -\lambda_i - \lambda_j & \text{for} \quad \varepsilon_i = \varepsilon_j = 1,\ 1 \le i \le j \le N \\ 0 & \text{with multiplicity } \frac{r(r-1)}{2}. \end{array}$$

Now we analyze the restriction of F to the orthogonal complement $V^\perp$ of V. We choose an extension of $\{x_1, \dots, x_N\}$ to an orthogonal basis of $\mathbb{R}^n$, say by $y_1, \dots, y_{n-r}$. Let $Y(i,j)$ denote the $N \times n$-matrix whose ith row vector is y_j and 0 otherwise. Then $\big(Y(i,j) \mid 1 \le i \le N, 1 \le j \le n-r\big)$ is a basis of $V^\perp$. Furthermore $Y(i,j)X_\infty^{\mathrm{T}} = 0$ and therefore $F(Y(i,j)) = (\lambda_i - \varepsilon_i\lambda_i)Y(i,j)$.

This shows that $V^\perp$ is invariant under F and that the eigenvalues of $F|V^\perp$ are $\lambda_i - \varepsilon_i\lambda_i$ with multiplicity $n-r$, i.e. $F|V^\perp$ has eigenvalues 0 with multiplicity $r(n-r)$ and λ_i with multiplicity $n-r$, $1 \le i \le N, \varepsilon_i = 0$.

The formulas in (iii) are immediate consequences of (ii). □

Remark. The linearization (5.2) is stable if $X_\infty X_\infty^{\mathrm{T}} = \operatorname{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$ with $n = r$ or $\lambda_{r+1} < 0$. If A has distinct eigenvalues, then the converse is also true.

The linearization (5.2) is asymptotically stable if and only if $n = r = 1, X_\infty X_\infty^{\mathrm{T}} = \operatorname{diag}(\lambda_1, 0, \dots, 0)$ and $\lambda_1 > \lambda_2$.

6. RECURSIVE VERSIONS

Recursive algorithms for finding the best approximant of A by positive semidefinite matrices can be obtained by discretizing the continuous time gradient flows using suitable small step-size selections. Here we consider discretizations of (5.1). Our analysis parallels similar efforts in Moore, Mahony and Helmke [8], and Helmke and Moore [4].

We consider the rank preserving recursion

$$X_{k+1} = e^{\alpha_k (A - X_k X_k^{\mathrm{T}})} X_k, \quad k \in \mathbb{N}_0, \tag{6.1}$$

on $\mathbb{R}^{N \times n}$ with suitable small step-size selections $\alpha_k \geq 0$. The α_k are chosen such as to maximize the difference

$$\begin{aligned} \Delta V(X_k) &= V(X_{k+1}) - V(X_k) = \\ &= \left\| A - e^{\alpha_k (A - X_k X_k^{\mathrm{T}})} X_k X_k^{\mathrm{T}} e^{\alpha_k (A - X_k X_k^{\mathrm{T}})} \right\|^2 - \left\| A - X_k X_k^{\mathrm{T}} \right\|^2 . \end{aligned}$$

Lemma 6.1. Let $A, B \in \mathbb{R}^{N \times N}$, $F := A - B$, be symmetric matrices with $B \geq 0$, $B \neq 0$. For $\alpha \geq 0$ let $\Delta\phi(\alpha) := \phi(\alpha) - \phi(0)$ with

$$\phi(\alpha) := \left\| A - e^{\alpha F} B e^{\alpha F} \right\|^2 .$$

Then

$$\Delta\phi(\alpha) \leq 0$$

for

$$\alpha := \frac{1}{2\|F\|} \log \left[\frac{-\|A\| + \sqrt{\|A\|^2 + 4((\|A\| + \|B\|)\|B\| + \operatorname{tr}(F^2 B)\|F\|^{-1})}}{2\|B\|} \right] .$$

$\Delta\phi(\alpha) = 0$ if and only if $FB = 0$.

Proof. For an $N \times N$ matrix X let $\mathcal{A}_X : \mathbb{R}^{N \times N} \to \mathbb{R}^{N \times N}$ be the linear operator defined by

$$\mathcal{A}_X(Y) = XY + YX,$$

and let $\mathcal{A}_X^n(Y) = \mathcal{A}_X(\mathcal{A}_X^{n-1}(Y))$, $\mathcal{A}_X^0 = id$, be recursively defined for $n \in \mathbb{N}$. Thus for X, Y symmetric also $\mathcal{A}_X(Y)$ is symmetric and

$$\|\mathcal{A}_X(Y)\| \leq 2\|X\| \; \|Y\|.$$

The following identity is easily verified by differentiating both sides:

$$e^{tF} B e^{tF} = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathcal{A}_F^n(B)$$

Therefore for $t \geq 0$:

$$\|e^{tF} B e^{tF} - t\mathcal{A}_F(B) - B\| \leq (e^{2t\|F\|} - 2t\|F\| - 1)\|B\|$$

From this estimate we obtain:

$$\begin{aligned}\Delta\phi(\alpha) &= -2\ \mathrm{tr}\,[A(e^{\alpha F}Be^{\alpha F}-B)] + \mathrm{tr}\,[B(e^{2\alpha F}Be^{2\alpha F}-B)]\\ &= -2\alpha\,\mathrm{tr}\,(F\mathcal{A}_F(B)) - 2\,\mathrm{tr}\,[A(e^{\alpha F}Be^{\alpha F}-\alpha\mathcal{A}_F(B)-B)]\\ &+ \mathrm{tr}\,[B(e^{2\alpha F}Be^{2\alpha F}-2\alpha\mathcal{A}_F(B)-B)]\\ &\le -4\alpha\,\mathrm{tr}\,(F^2B)+2\|A\|\,\|B\|(e^{2\alpha\|F\|}-2\alpha\|F\|-1)+\|B\|^2(e^{4\alpha\|F\|}-4\alpha\|F\|-1)\\ &=: \psi(\alpha)\end{aligned}$$

Thus

$$\begin{aligned}\psi(\alpha) &= \|B\|^2e^{4\alpha\|F\|}+2\|A\|\,\|B\|e^{2\alpha\|F\|}\\ &\qquad -4\alpha\big(\mathrm{tr}\,(F^2B)+(\|A\|+\|B\|)\|B\|\,\|F\|\big)-\|B\|^2-2\|A\|\,\|B\|.\end{aligned}$$

By differentiating $\psi(\alpha)$:

$$\begin{aligned}\psi'(\alpha) = 4\Big[&\|B\|^2\|F\|e^{4\alpha\|F\|}+\|A\|\,\|B\|\,\|F\|e^{2\alpha\|F\|}-\mathrm{tr}\,(F^2B)\\ &-(\|A\|+\|B\|)\|B\|\,\|F\|\Big].\end{aligned}$$

and $\psi''(\alpha)>0$. Thus $\psi(\alpha)$ is a strictly convex function on $\mathbb{R}_+$ with the minimum given by $\psi'(\alpha_*)=0$. Equivalently, $x := e^{2\alpha_*\|F\|}$ is the positive root of

$$\|B\|^2\|F\|x^2+\|A\|\,\|B\|\,\|F\|x-(\mathrm{tr}\,(F^2B)+(\|A\|+\|B\|)\|B\|\,\|F\|)=0.$$

Thus

$$x=-\frac{1}{2}\frac{\|A\|}{\|B\|}+\sqrt{\frac{\|A\|^2}{4\|B\|^2}+\frac{\mathrm{tr}\,(F^2B)+(\|A\|+\|B\|)\|B\|\,\|F\|}{\|B\|^2\|F\|}}$$

$\iff$

$$x=\frac{-\|A\|+\sqrt{\|A\|^2+4((\|A\|+\|B\|)\|B\|+\mathrm{tr}\,(F^2B)\|F\|^{-1})}}{2\|B\|}$$

Hence

$$\alpha_*=\frac{1}{2\|F\|}\log\left[\frac{-\|A\|+\sqrt{\|A\|^2+4((\|A\|+\|B\|)\|B\|\ +\mathrm{tr}\,(F^2B)/\|F\|)}}{2\|B\|}\right]$$

Since $\psi(0)=0$, $\psi'(0)=-4\,\mathrm{tr}\,(F^2B)\le 0$, we have $\psi(\alpha_*)\le 0$. Moreover, $\psi(\alpha_*)=0$ if and only if $B=0$ or $FB=0$. This completes the proof. □

We now apply the lemma to the situation where $A=A^{\mathrm{T}}\in\mathbb{R}^{N\times N}$ and

$$B=X_kX_k^{\mathrm{T}},\ F:=A-X_kX_k^{\mathrm{T}}.$$

Then the lemma implies that

$$V(X_{k+1})\le V(X_k)$$

for

$$\alpha_k := \frac{1}{2\|A-X_kX_k^{\mathrm{T}}\|}\cdot$$

$$\log\left[\frac{-\|A\|+\sqrt{\|A\|^2\|A-X_kX_k^{\mathrm{T}}\|+4\big((\|A\|+\|X_kX_k^{\mathrm{T}}\|)\|A-X_kX_k^{\mathrm{T}}\|+\|(A-X_kX_k^{\mathrm{T}})X_k\|^2\big)}}{2\|X_kX_k^{\mathrm{T}}\|}\right] \tag{6.2}$$

with equality $V(X_{k+1}) = V(X_k)$ if and only if $(A - X_k X_k^{\mathrm{T}})X_k = 0$. Thus, unless $X_k \equiv X_0$ is an equilibrium point, the distance function $V(X_k)$ decreases strictly. In particular, $V(X) = \|A - XX^{\mathrm{T}}\|^2$ is a Lyapunov function for the discrete time system

$$X_{k+1} = e^{\alpha_k (A - X_k X_k^{\mathrm{T}})} X_k$$

with α_k as in (6.2).

This together with standard arguments from Lyapunov theory implies:

Theorem 6.2. Let $A \in \mathbb{R}^{N \times N}$ be symmetric and $1 \leq n \leq N$. The recursive system (6.1), (6.2) converges from any nonzero initial condition $X_0 \in \mathbb{R}^{N \times N}$ to the set of equilibria points. If A has distinct eigenvalues and $n = 1$, then every solution of (6.1), (6.2) converges to $\{X_\infty, -X_\infty\}$, where the columns of $X_\infty \in \mathbb{R}^{N \times n}$ generate an A-invariant subspace.

Standard linearization arguments similar to the above show that — for A having distinct eigenvalues with at least n positive eigenvalues — for almost every initial condition $X_0 \in \mathbb{R}^{N \times n}$ the column spaces of $(X_k \mid k \in \mathbb{N})$ converge to the maximal eigenspace of A. Thus the recursion can be used to find maximal eigenbasis vectors for A.

(Received March 2, 1993.)

REFERENCES

[1] G. Eckart and G. Young: The approximation of one matrix by another of lower rank. Psychometrika *1* (1936), 211–218.

[2] G. H. Golub and C. Van Loan: An analysis of the total least squares problem. SIAM J. Numer. Anal. *17* (1980), 883–843.

[3] G. H. Golub, A. Hoffmann and G. W. Stewart: A generalization of the Eckart-Young-Mirsky matrix approximation theorem. Linear Algebra Appl. *88/89* (1987), 317–327.

[4] U. Helmke and J. B. Moore: Optimization and Dynamical Systems. Springer-Verlag, Berlin 1993.

[5] U. Helmke and M. A. Shayman: Critical points of matrix least squares distance functions. Linear Algebra Appl., to appear.

[6] N. J. Higham: Computing a nearest symmetric positive semidefinite matrix. Linear Algebra Appl. *103* (1988), 103–118.

[7] B. De Moor and J. David: Total linear least squares and the algebraic Riccati equation. Systems Control Lett. *5* (1992), 329–337.

[8] J. B. Moore, R. E. Mahony and U. Helmke: Recursive gradient algorithms for eigenvalue and singular value decomposition. SIAM J. Matrix Anal. Appl., to appear.

Doc. Dr. Uwe Helmke and Dr. Michael Prechtel, Department of Mathematics, Regensburg University, D-8400 Regensburg. Germany.

Dr. Mark A. Shayman, Electrical Engineering Department, University of Maryland, College Park, Maryland 20742. U.S.A.