

# Newton's method for concave operators with resolvent positive derivatives in ordered Banach spaces

T. Damm  
Zentrum Mathematik  
TU München  
80333 München, Germany  
damm@ma.tum.de

D. Hinrichsen  
Institut für Dynamische Systeme  
Universität Bremen  
28334 Bremen, Germany  
dh@math.uni-bremen.de

## Abstract

We prove a non-local convergence result for Newton's method applied to a class of nonlinear equations in ordered real Banach spaces. The key tools in our approach are special notions of concavity and the spectral theory of resolvent positive operators.

## 1 Introduction

A standard method to solve nonlinear equations is the Newton iteration. Its applicability to operator equations in normed spaces was first established by Kantorovich in [14]. In the operator-theoretic setup the method is therefore often referred to as the Newton-Kantorovich-procedure. Kantorovich specified boundedness conditions on the first and second Fréchet-derivatives of the nonlinear operator, that guarantee convergence of the Newton iteration starting in a small neighbourhood of the actual solution. These results can be simplified and generalized, if the underlying space is ordered and the sequence produced by the iteration can be shown to be monotonic and bounded; this can be the case, for instance, if the nonlinear operator satisfies certain convexity conditions (compare [23] and references therein).

In general, however, results on the convergence of the Newton iteration are of a local nature; they require a good initial guess. But in some special cases, e.g. for Riccati-operators ([15], [24], [13], [5], [11], [7]), it has been observed that the iteration converges from any point, where the derivative of the Riccati-operator has its spectrum in the left half plane. This is what we call a non-local convergence result, since the starting point may be far away from any solution of the Riccati equation.

In this paper we present a class of nonlinear operator equations in ordered Banach spaces that can be solved by a non-local Newton iteration. Our approach is motivated by our earlier investigations of Riccati type equations in the ordered real space of Hermitian matrices [7], and these turn out to be special cases of our main result.

We proceed as follows: In Section 2 we present basic facts about ordered linear spaces and positive linear operators. Some results from the spectral theory of resolvent positive operators which we need but did not find in the literature, are proved in an Appendix. In Section 3 we introduce the notions of  $D_+$ -concavity and directional concavity and prove a simple

but useful result on the relation between concavity and differentiability. The tools worked out in Sections 2 and 3 are then applied in Section 4 to prove a non-local convergence result for Newton's method applied to a class of concave operator equations with resolvent positive derivatives. We also discuss a modified fixed-point iteration. The paper is concluded with Section 5 where the results are illustrated by applications to a number of rational matrix equations arising in various control and realization problems.

## 2 Resolvent positive operators on ordered Banach spaces

In this section we summarize some basic concepts and results from the theories of ordered vector spaces and resolvent positive linear operators.

Let  $X$  be a real Banach space. Following the terminology in [3] we say that a nonempty subset  $C \subset X$  is a *convex cone* if  $C + C = C$ ,  $\alpha C \subset C$  for all real numbers  $\alpha \geq 0$ . If the cone is pointed (i.e.  $C \cap -C = \{0\}$ ) such a cone  $C$  induces an ordering on  $X$ . For  $x, y \in X$  we write  $x \geq y$ , if  $x - y \in C$ . If  $C$  has interior points and  $x - y \in \text{int } C$ , then we write  $x > y$ . If  $x \leq y$ , the set  $[x, y] = \{z \in X \mid x \leq z \leq y\}$  is called the order interval between  $x$  and  $y$ . We will need the following definitions [17].

**Definition 2.1** *Given a convex cone  $C$  in the real Banach space  $X$ , the dual cone  $C^*$  in the dual space  $X^*$  is given by  $C^* := \{y^* \in X^* \mid \forall x \in C : \langle x, y^* \rangle \geq 0\}$ .*

- (i)  $C$  is reproducing if  $C - C = X$ .
- (ii)  $C$  is solid if  $\text{int } C \neq \emptyset$ .
- (iii)  $C$  is normal if  $\exists b > 0 : \forall x, y \in C, x \leq y : \|x\| \leq b\|y\|$ .
- (iv)  $C$  is regular if every monotonically decreasing sequence  $x_1 \geq x_2 \geq \dots$ , which is bounded from below by some element  $\hat{x} \in X$ , converges in norm.

Our main results will be derived for real Banach spaces  $X$  ordered by a *closed solid, regular convex cone*  $C$ . It then follows that  $C$  is pointed, and both  $C$  and  $C^*$  are reproducing and normal (see Lemma A.1 in the Appendix).

**Example 2.2** Let  $\mathcal{H}^n \subset \mathbb{K}^{n \times n}$ ,  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{K} = \mathbb{C}$ , denote the real Hilbert space of real or complex  $n \times n$  Hermitian matrices, endowed with the Frobenius inner product  $\langle X, Y \rangle = \text{trace}(XY)$  and the corresponding (Frobenius) norm  $\|\cdot\|$ . By  $\mathcal{H}_+^n := \{X \in \mathcal{H}^n \mid X \geq 0\}$  we denote the subset of nonnegative definite matrices. The space  $\mathcal{H}^n$  is canonically ordered by this pointed convex cone which satisfies all the conditions (i)–(iv) from Definition 2.1. The interior of the cone  $\mathcal{H}_+^n$  consists of all the positive definite matrices in  $\mathcal{H}^n$  (see [3]).

**Definition 2.3** *Let  $X$  be a normed vector space ordered by the pointed convex cone  $C$ . A bounded linear operator  $T : X \rightarrow X$  is called positive if  $T(C) \subset C$ . It is called inverse positive if it has a bounded positive inverse and resolvent positive if there exists an  $\alpha_0 \in \mathbb{R}$ , such that for all  $\alpha > \alpha_0$  the operators  $\alpha I - T$  are inverse positive.*

**Example 2.4** Let  $A \in \mathbb{K}^{n \times n}$ , then the operator  $\Pi_A : \mathcal{H}^n \rightarrow \mathcal{H}^n$ ,  $\Pi_A(X) = A^*XA$  is positive with respect to the canonical ordering of  $\mathcal{H}^n$ , whereas both the *continuous-time Lyapunov operator*  $\mathcal{L}_A : \mathcal{H}^n \rightarrow \mathcal{H}^n$ ,  $\mathcal{L}_A(X) = A^*X + XA$ , and the *discrete-time Lyapunov operator*

(also *Stein operator*)  $\mathcal{S}_A : \mathcal{H}^n \rightarrow \mathcal{H}^n$ ,  $\mathcal{S}_A(X) = A^*XA - X$ , are resolvent positive but, in general, not positive (see [7]).

As a further illustration we mention the following equivalent characterizations of resolvent positive operators in finite dimensions [22], [8],[2].

**Proposition 2.5** *Let  $X$  be a finite-dimensional real vector space ordered by a solid, normal, pointed convex cone  $C$  with topological boundary  $\partial C$ . For  $T \in \mathcal{L}(X)$  the following are equivalent:*

- (i)  $T$  is resolvent positive.
- (ii)  $\exp(tT)$  is positive for all  $t \geq 0$ .
- (iii)  $\forall x \in \partial C : \exists v \in \partial C^* : \langle x, v \rangle = 0$  and  $\langle Tx, v \rangle \geq 0$ .
- (iv)  $x \in \partial C, v \in \partial C^*, \langle x, v \rangle = 0 \Rightarrow \langle Tx, v \rangle \geq 0$ .
- (v)  $T \in \text{cl}\{S - \alpha I \mid S \text{ positive}, \alpha \in \mathbb{R}\}$ .

*In particular, the set of resolvent positive operators is a closed convex cone in  $\mathcal{L}(X)$ . It contains all positive operators as well as all scalar multiples of the identity and hence is solid but not pointed.*

For a bounded linear operator  $T : X \rightarrow X$  we denote the spectrum by  $\sigma(T)$  and set

$$\begin{aligned} \beta(T) &= \max\{\text{Re}(\lambda); \lambda \in \sigma(T)\} \quad \text{for the spectral abscissa,} \\ \rho(T) &= \max\{|\lambda|; \lambda \in \sigma(T)\} \quad \text{for the spectral radius of } T. \end{aligned}$$

It is well-known that  $\sigma(T) = \sigma(T^*)$  and that the adjoint operator  $T^*$  is (resolvent) positive with respect to the positive cone  $C^*$  if and only if  $T$  is (resolvent) positive with respect to  $C$ .

The spectrum of positive operators on  $\mathbb{R}^n$  ordered by the cone  $\mathbb{R}_+^n$  was analyzed first by Perron and Frobenius. They showed, that the spectral radius of such an operator is an eigenvalue corresponding to a nonnegative eigenvector. This result was extended by Krein and Rutman in [18] to more general spaces and cones. For instance the following holds: If  $T$  is a positive linear operator in a real Banach space ordered by a normal and reproducing convex cone, then  $\rho(T) \in \sigma(T)$  (see [17], Thm. 8.1). In general, however, it is not true, that  $\rho(T)$  is an eigenvalue. But if one considers the adjoint operator  $T^*$  instead of  $T$ , the existence of an eigenvector in  $C^*$  for the eigenvalue  $\rho(T^*) = \rho(T)$  is guaranteed under fairly general conditions:

**Theorem 2.6** *Let  $X$  be a real Banach space, ordered by a normal, solid convex cone  $C$ , and  $T : X \rightarrow X$  a bounded linear operator.*

- (a)  $T$  is positive  $\Rightarrow \exists v \in C^*, v \neq 0 : T^*v = \rho(T)v$ .
- (b)  $T$  is resolvent positive  $\Rightarrow \exists v \in C^*, v \neq 0 : T^*v = \beta(T)v$ .

A proof of (a) can be found in [20] App. 2.6, or [17] Thm. 9.11, for a proof of (b) see the Appendix.

The following theorem is an infinite-dimensional variation of a result by H. Schneider in [21] and plays a central rôle.

**Theorem 2.7** *Let  $X$  be a real Banach space ordered by a solid, normal convex cone  $C$ . Suppose  $R : X \rightarrow X$  to be resolvent positive and  $P : X \rightarrow X$  to be positive, and set  $T = R + P$ . Then the following are equivalent:*

- (i)  $T$  is stable, i.e.  $\sigma(T) \subset \mathbb{C}_-$ .
- (ii)  $-T$  is inverse positive.
- (iii)  $\forall y \in \text{int } C : \exists x \in \text{int } C : -T(x) = y$
- (iv)  $\exists x \in \text{int } C : -T(x) \in \text{int } C$ .
- (v)  $\exists x \in C : -T(x) \in \text{int } C$ .
- (vi)  $\sigma(R) \subset \mathbb{C}_-$  and  $\rho(R^{-1}P) < 1$ .

A proof adapted from [21] can be found in the Appendix.

### 3 Concave maps

In the following let  $X$  be a real Banach space ordered by a pointed convex cone  $C$  and consider a (nonlinear) mapping  $f : X \supset \text{dom } f \rightarrow X$ .

**Definition 3.1** *Assume we are given subsets  $D_+ \subset D \subset \text{dom } f$ . Then we say that  $f$  is  $D_+$ -concave on  $D$  if we can attach a bounded linear operator  $T_x : X \rightarrow X$  at each point  $x \in D$ , such that*

$$\forall y \in D_+ : f(y) \leq f(x) + T_x(y - x). \quad (1)$$

**Remark 3.2** (i) Let  $X = \mathbb{R}$  with the canonical ordering. If e.g.  $D = \text{dom } f = X$  then  $f$  is  $D_+$ -concave on  $D$  if at each point of the graph of  $f$  we can attach a straight line, such that the whole graph of  $f$  above  $D_+$  lies below this line.

(ii) In the next chapter,  $D_+$  plays the role of a target set, in which we try to find a solution of  $f(x) = 0$ , whereas  $D$  contains possible initial values for a fixed-point iteration. The  $D_+$ -concavity on  $D$  is needed, to guarantee the transition from  $D$  to  $D_+$ .

We will also consider the following weaker version of  $D_+$ -concavity.

**Definition 3.3** *Assume the situation of Definition 3.1 and let  $K \subset X$ . We say that  $f$  is  $D_+$ -concave on  $D$  in direction  $K$  if (1) holds for all  $x \in D$ ,  $y \in D_+$ , such that  $y - x \in K$ . If  $D = D_+$  we just say that  $f$  is concave on  $D$  in direction  $K$ .*

Obviously  $f$  is  $D_+$ -concave on  $D$  if and only if it is  $D_+$ -concave on  $D$  in direction  $X$ . Finally we recall the definition of Gâteaux-differentiability.

**Definition 3.4** *Let  $f : X \supset \text{dom } f \rightarrow X$  and  $x \in \text{int dom } f$ . Then we say that  $f$  is Gâteaux-differentiable at  $x$  if there exists a mapping  $f'_x : X \rightarrow X$  such that for all  $h \in X$  and  $t \in \mathbb{R}$*

$$f(x + th) = f(x) + tf'_x(h) + t\phi_{x,h}(t) \quad \text{with } \lim_{t \rightarrow 0} \phi_{x,h}(t) = 0.$$

**Proposition 3.5** *Let  $D_+ \subset D \subset \text{dom } f$ ,  $K \subset X$ , and assume  $f : D \rightarrow X$  to be  $D_+$ -concave on  $D$  in direction  $K$ . Let further  $x \in D$  and  $y \in \text{int } D_+$  such that  $y - x \in \text{int } K$  and assume  $f$  to be Gâteaux differentiable at  $y$ .*

- (i) *If  $f(y) - f(x) = T_x(y - x)$ , then  $f'_y = T_x$ . In particular  $f'_y = T_y$  if  $0 \in \text{int } K$ .*
- (ii) *If  $\langle f(y) - f(x), v \rangle = \langle T_x(y - x), v \rangle$  for some  $v \in C^*$ , then  $(f'_y)^*(v) = T_x^*(v)$ .*

**Proof:** (i) If  $f(y) - f(x) = T_x(y - x)$ , we have for all  $z \in D_+$  with  $z - x \in K$

$$f(z) \leq f(x) + T_x(z - x) = f(x) + T_x(z - y) + T_x(y - x) = f(y) + T_x(z - y) .$$

In particular it follows for every  $h \in X$  and  $z = y \pm th$  with  $0 < t < 1$  sufficiently small that  $z \in D_+$ ,  $z - x \in K$  and

$$f(y \pm th) = f(y) + f'_y(\pm th) \pm t\phi_{x,h}(t) \leq f(y) + T_x(\pm th) .$$

As  $t \rightarrow 0$  we obtain  $\pm f'_y(h) \leq \pm T_x(h)$ , whence  $f'_y(h) = T_x(h)$  for all  $h \in X$ .

(ii) Applying the functional  $v$  to all the expressions in the proof of (i) we obtain  $\langle f'_y(h), v \rangle = \langle T_x(h), v \rangle$  for all  $h \in X$ , whence  $(f'_y)^*(v) = T_x^*(v)$ .  $\square$

## 4 Resolvent positive operators and Newton's method

In this section let  $X$  be a real Banach space, ordered by a closed, solid, regular convex cone  $C$  and  $f$  a continuous mapping from some subset  $\text{dom } f$  of  $X$  to  $X$ .

Moreover let there be given subsets  $D_+ \subset D \subset \text{dom } f$  and attached to each point  $x \in D$  a bounded linear mapping  $T_x : X \rightarrow X$ , such that the following holds:

**Assumption 4.1** (H1)  $D_+ = D_+ + C$ .

(H2)  $f$  is  $D_+$ -concave on  $D$  (with the given  $T_x$ ).

(H3) The  $T_x$  are resolvent positive for all  $x \in D$ .

(H4) The  $T_x$  are locally bounded on  $D_+$ , i.e.  $\forall x \in D_+ \exists \varepsilon : \sup_{y \in D_+, \|y-x\| < \varepsilon} \|T_y\| < \infty$ .

(H5) There exists an  $x_0 \in D$  such that  $\sigma(T_{x_0}) \subset \mathbb{C}_-$ .

(H6) There exists an  $\hat{x} \in \text{int } D_+$  such that  $f(\hat{x}) \geq 0$ .

(H7)  $f$  is Gâteaux-differentiable on  $\text{int } D_+$ .

Important examples for such mappings are provided by Riccati-type operators in the ordered vector space  $\mathcal{H}^n$  of Hermitian matrices, see Section 5.

We address the problem of approximating a solution  $x \in D_+$  to the equation  $f(x) = 0$  by the Newton-type iteration

$$x_{k+1} = x_k - (T_{x_k})^{-1}(f(x_k)) . \tag{2}$$

**Theorem 4.2** *Let Assumption 4.1 hold or assume, alternatively, the hypotheses (H1)–(H5) and*

(H8) *There exists an  $\hat{x} \in D_+$  such that  $f(\hat{x}) > 0$ .*

*Then the iteration scheme (2) starting at an arbitrary  $x_0 \in D$  such that  $\sigma(T_{x_0}) \subset \mathbb{C}_-$  defines a sequence  $x_1, x_2, \dots$  in  $D_+$  with the following properties:*

(i)  $\forall k = 1, 2, \dots : x_k \geq x_{k+1} \geq \hat{x}$ ,  $f(x_k) \leq 0$ , and  $\sigma(T_{x_k}) \subset \mathbb{C}_-$ .

(ii)  $x_+ := \lim_{k \rightarrow \infty} x_k \in D_+$  satisfies  $f(x_+) = 0$  and is the largest solution of the inequality  $f(x) \geq 0$  in  $D_+$  (i.e.  $f(x) \geq 0 \Rightarrow x \leq x_+$  for all  $x \in D_+$ ).

**Proof:** We prove (i) inductively and will only use conditions (H1) to (H5) in the first few steps.

Suppose that  $x_0, \dots, x_k$  have been constructed for some  $k \geq 0$  such that  $T_{x_i}$  is stable for  $i = 0, \dots, k$ ,  $x_1 \geq \dots \geq x_k$  and  $f(x_i) \leq 0$  for  $i = 1, \dots, k$ . Then  $T_{x_k}$  is inverse positive by Theorem 2.7 so that  $x_{k+1}$  is well defined by (2) and satisfies

$$T_{x_k}(x_k - x_{k+1}) = f(x_k). \quad (3)$$

We first prove  $x_k \geq x_{k+1} \geq \hat{x}$ . Since  $x_k \in D$  and  $\hat{x} \in D_+$  we obtain from the  $D_+$ -concavity of  $f$  on  $D$  that

$$T_{x_k}(\hat{x} - x_{k+1}) = T_{x_k}(\hat{x} - x_k) + T_{x_k}(x_k - x_{k+1}) = T_{x_k}(\hat{x} - x_k) + f(x_k) \geq f(\hat{x}) \geq 0.$$

But we know already that  $T_{x_k}$  is inverse positive and so we have  $\hat{x} \leq x_{k+1}$ . Hence  $x_{k+1} \in D_+$ , because  $\hat{x} \in D_+$  and  $D_+ = D_+ + C$ . By the same argument, if  $k \geq 1$  it follows from (3) and  $f(x_k) \leq 0$  that  $x_k - x_{k+1} \geq 0$ .

It remains to show that  $T_{x_{k+1}}$  is stable and  $f(x_{k+1}) \leq 0$ . For this we make use of the concavity condition for the pairs  $(x_{k+1}, x_k), (\hat{x}, x_{k+1}) \in D_+ \times D$ , to obtain

$$f(x_{k+1}) \leq f(x_k) + T_{x_k}(x_{k+1} - x_k), \quad (4)$$

$$f(\hat{x}) \leq f(x_{k+1}) + T_{x_{k+1}}(\hat{x} - x_{k+1}). \quad (5)$$

By (3) the right side in (4) vanishes, whence  $f(x_{k+1}) \leq 0$ . Therefore (5) yields

$$f(\hat{x}) \leq T_{x_{k+1}}(\hat{x} - x_{k+1}). \quad (6)$$

If now  $f(\hat{x}) > 0$  then  $\sigma(T_{x_{k+1}}) \subset \mathbb{C}_-$  by Theorem 2.7, which completes the proof of (i) under conditions (H1)–(H5) and (H8).

To complete the proof of (i) under Assumption 4.1 let us suppose that  $T_{x_{k+1}}$  is not stable. By Theorem 2.6 (ii) this is equivalent to the following condition:

$$\exists v \in C^* \setminus \{0\}, \beta \geq 0 : T_{x_{k+1}}^* v = \beta v. \quad (7)$$

Together with the four inequalities  $\hat{x} \leq x_{k+1}$ ,  $f(\hat{x}) \geq 0$ ,  $f(x_{k+1}) \leq 0$  and (5), this implies

$$0 \geq \langle \hat{x} - x_{k+1}, \beta v \rangle = \langle T_{x_{k+1}}(\hat{x} - x_{k+1}), v \rangle \geq \langle -f(x_{k+1}), v \rangle \geq 0. \quad (8)$$

Hence  $\langle f(x_{k+1}), v \rangle = 0$ , which by (3) means

$$\langle f(x_{k+1}) - f(x_k), v \rangle = \langle T_{x_k}(x_{k+1} - x_k), v \rangle.$$

But  $x_{k+1} \geq \hat{x} \in \text{int } D_+$  and  $D_+ = D_+ + C$  imply  $x_{k+1} \in \text{int } D_+$ , and so  $f$  is Gâteaux-differentiable at  $x_{k+1}$  by condition (H7). From Proposition 3.5 (with  $K = X$ ) we conclude that  $T_{x_{k+1}} = f'_{x_{k+1}}$ , and  $T_{x_k}^*(v) = (f'_{x_{k+1}})^*(v) = T_{x_{k+1}}^*(v) = \beta v$ , in contradiction with  $\sigma(T_{x_k}) \subset \mathbb{C}_-$ . Thus  $T_{x_{k+1}}$  must be stable, and this concludes our proof of (i) under Assumption 4.1. (ii) By (i) and the regularity of  $C$ , the  $x_k$  converge in norm to some  $x_+ \in \hat{x} + C \subset D_+$ . Since the  $T_x$  are locally bounded on  $D_+$ , we can pass to the limit in (3) to obtain  $f(x_+) = 0$ . By the first part of the proof the inequality  $x_{k+1} \geq \hat{x}$  holds true for all  $k \in \mathbb{N}$  and all solutions  $\hat{x} \in D_+$  of the inequality  $f(x) \geq 0$ . Therefore  $x_+$  is the largest solution of this inequality in  $D_+$ .  $\square$

We say that a solution  $\tilde{x} \in \text{dom } f$  of  $f(x) = 0$  is *stabilizing* if  $\sigma(T_{\tilde{x}}) \subset \mathbb{C}_-$ . The following lemma shows, that there can be at most one stabilizing solution in  $D_+$ .

**Lemma 4.3** *Suppose that  $f$  is  $D_+$ -concave on  $D$  with resolvent positive  $T_x$  and let  $y \in D_+$ ,  $z \in D$  and  $f(z) \leq f(y)$ . If  $\sigma(T_z) \subset \mathbb{C}_-$  then  $z \geq y$ . In particular, if  $z \in D_+$  is a stabilizing solution of  $f(x) = 0$  then  $z = x_+$ .*

**Proof:** By concavity  $T_z(y - z) \geq f(y) - f(z) \geq 0$ , whence by the stability and resolvent positivity of  $T_z$  we have  $y - z \leq 0$ . Now suppose  $z \in D_+$  is a stabilizing solution of  $f(x) = 0$ , then we choose  $y = x_+$  and get  $z \geq x_+$ . On the other hand we have  $z \leq x_+$  by Theorem 4.2.  $\square$

The following corollary of Theorem 4.2 gives a sufficient condition for the existence of a stabilizing solution.

**Corollary 4.4** *If conditions (H1) - (H5) and (H8) hold then  $x_+$  in Theorem 4.2 is a stabilizing solution of  $f(x) = 0$  and satisfies  $x_+ > \hat{x}$  (whence  $x_+ \in \text{int } D_+$ ).*

**Proof:** We already know  $x_+ \geq \hat{x}$ , and by concavity

$$T_{x_+}(\hat{x} - x_+) \geq f(\hat{x}) - f(x_+) = f(\hat{x}) > 0;$$

hence  $\sigma(T_{x_+}) \subset \mathbb{C}_-$  and  $x_+ > \hat{x}$  follows from Theorem 2.7.  $\square$

If  $f$  is Fréchet-differentiable, condition (H8) is equivalent to the existence of a stabilizing solution of  $f(x) = 0$ .

**Corollary 4.5** *Assume conditions (H1) - (H5) and that  $f$  is Fréchet-differentiable on  $\text{int } D_+$ . Then*

$$(\exists \hat{x} \in D_+ : f(\hat{x}) > 0) \iff (\exists y \in \text{int } D_+ : f(y) = 0 \text{ and } \sigma(f'_y) \subset \mathbb{C}_-).$$

**Proof:** It only remains to prove ' $\Leftarrow$ '. By assumption  $0 \notin \sigma(f'_y)$  and so  $f'_y$  is an invertible bounded linear operator on  $X$ . By the implicit function theorem,  $f$  maps a small neighbourhood  $U \subset D_+$  of  $y$  onto a neighbourhood  $V$  of  $f(y) = 0$ . Choosing  $c \in V \cap \text{int } C$  we see that there exists  $\hat{x} \in U$  such that  $f(\hat{x}) = c > 0$ .  $\square$

The existence of stabilizing solutions implies quadratic convergence of the sequence  $(x_k)$ , provided  $f$  is sufficiently smooth. This is e.g. a consequence of the following well known result (compare [16] and also Remark 4.11):

**Proposition 4.6** *Assume the situation of Theorem 4.2, and let  $f$  be Fréchet-differentiable in a neighbourhood  $U$  of  $x_+$ . Moreover assume that the  $T_x$  satisfy a Lipschitz condition  $\|T_x - T_y\| \leq L\|x - y\|$  for all  $x, y \in U$ .*

*If  $x_+$  is stabilizing, then the sequence  $(x_k)$  converges quadratically, i.e. there exists a constant  $\kappa$  such that*

$$\|x_{k+1} - x_+\| \leq \kappa \|x_k - x_+\|^2, \quad k \in \mathbb{N}.$$

**Remark 4.7** (i) The sequence  $(x_k)$  is monotonically decreasing *after* the first step. The first step is, in general, not decreasing, and we have no control, how far it might lead away from the solution  $x_+$ . It is, however, needed to achieve  $f(x_1) \leq 0$  and  $x_1 \geq \hat{x}$ . Here the  $D_+$ -concavity of  $f$  comes into play. To start the iteration, it is sufficient to find an  $x_0 \in D$  and a resolvent positive, stable  $T_{x_0}$  satisfying (1) with  $x = x_0$ . We call Theorem 4.2 a *non-local convergence result*, since  $x_0$  does not have to be close to  $x_+$ .

(ii) If one considers the monotonic part of the sequence (after the first step), the concavity inequality is applied only at points  $x, y \in D_+$  with  $x \geq y$ . Under condition (H8) it is in fact sufficient to require only concavity on  $D_+$  in direction  $-C$  (see Thm. 4.9).

(iii) The iteration (2) requires the solution of a linear equation of the form  $T_{x_k}x = y$  in each step. It has been observed in the context of Riccati-equations (see e.g. [10]) that it can be advantageous to replace the operators  $T_{x_k}$  by other operators that are numerically easier to handle.

In the sequel we suggest a general framework to take advantage of these observations.

For all  $x \in D_+$  we consider a decomposition of  $T_x$  of the form  $T_x = R_x + P_x$ , where the  $R_x$  are resolvent positive and the  $P_x$  are positive. Replacing  $T_{x_k}$  by  $R_{x_k}$  in (2) we define the iteration

$$x_{k+1} = x_k - (R_{x_k})^{-1} (f(x_k)). \quad (9)$$

Let there be given  $\hat{x}, x_0 \in \text{dom } f$  such that  $\hat{x} \leq x_0$  and attached to each point  $x \in [\hat{x}, x_0]$  a bounded linear mapping  $T_x : X \rightarrow X$ , such that the following holds:

**Assumption 4.8** (H1')  $[\hat{x}, x_0] \subset \text{dom } f$ .

(H2')  $f$  is concave on  $[\hat{x}, x_0]$  in direction  $K$  where  $\text{int } K \supset -C \setminus \{0\}$  (with the given  $T_x$ ).

(H3')  $T_x = R_x + P_x$ , where the  $R_x$  are resolvent positive and the  $P_x$  are positive for all  $x \in [\hat{x}, x_0]$ .

(H4') The  $R_x$  are locally bounded on  $[\hat{x}, x_0]$ .

(H5')  $\sigma(T_{x_0}) \subset \mathbb{C}_-$  and  $f(x_0) \leq 0$ .

(H6')  $\hat{x} \in \text{int dom } f$  and  $f(\hat{x}) \geq 0$ .

(H7')  $f$  is Gâteaux-differentiable at  $\hat{x}$ .

The following result has two aspects corresponding to Remark 4.7 (ii) and (iii). Firstly, we weaken the concavity conditions in Theorem 4.2 at the price of strengthening the requirements imposed on the initial value  $x_0$ . Secondly we allow the operators  $T_x$  to be replaced by the  $R_x$ , at the price of possibly diminishing the rate of convergence.

**Theorem 4.9** *Let Assumption 4.8 hold or, alternatively, assume the hypotheses (H1') - (H6') with the weaker requirement  $K = -C$  and the stronger requirement  $f(\hat{x}) > 0$ . Then the iteration scheme (9) starting at  $x_0$  defines a monotonically decreasing sequence in  $[\hat{x}, x_0]$  with the following properties:*

- (i)  $\forall k = 0, 1, \dots : x_k \geq x_{k+1} \geq \hat{x}, f(x_k) \leq 0$ , and  $\sigma(T_{x_k}) \subset \mathbb{C}_-$ .
- (ii)  $x_+ := \lim_{k \rightarrow \infty} x_k \in D_+$  satisfies  $f(x_+) = 0$  and is the largest solution of the inequality  $f(x) \geq 0$  in  $[\hat{x}, x_0]$  (i.e.  $f(x) \geq 0 \Rightarrow x \leq x_+$  for all  $x \in [\hat{x}, x_0]$ ).

**Proof:** The proof of (i) proceeds by induction and follows the proof of Theorem 4.2. We only elaborate on those points where we must take into account the new assumptions. Let us assume that for some  $k \geq 0$  we have constructed  $x_0 \geq x_1 \geq \dots \geq x_k$  such that  $T_{x_i}$  is stable,  $x_i \geq \hat{x}$  and  $f(x_i) \leq 0$  for  $i \leq k$ . Then also  $R_{x_k}$  is stable by Theorem 2.7 and  $x_{k+1}$  is well defined by (9) and satisfies

$$R_{x_k}(x_k - x_{k+1}) = f(x_k) \leq 0. \quad (10)$$

We have to show that  $x_k \geq x_{k+1} \geq \hat{x}$ ,  $f(x_{k+1}) \leq 0$  and  $\sigma(T_{x_{k+1}}) \subset \mathbb{C}_-$ .

By Theorem 2.7, (10) implies  $x_k \geq x_{k+1}$ . To prove  $x_{k+1} \geq \hat{x}$  we utilize the concavity of  $f$  at  $x_k$  in direction  $x_{k+1} - x_k \leq 0$  and remember that  $T_{x_k}(y) \leq R_{x_k}(y)$  for  $y \leq 0$ :

$$R_{x_k}(\hat{x} - x_{k+1}) = R_{x_k}(\hat{x} - x_k) + R_{x_k}(x_k - x_{k+1}) \geq T_{x_k}(\hat{x} - x_k) + f(x_k) \geq f(\hat{x}) \geq 0.$$

Hence  $\hat{x} \leq x_{k+1}$  and so  $x_{k+1} \in [\hat{x}, x_0]$ . By the concavity of  $f$  at  $x_k$  and  $x_{k+1}$  in the directions  $(x_{k+1} - x_k) \leq 0$  and  $(\hat{x} - x_{k+1}) \leq 0$ , respectively, we have

$$f(x_{k+1}) \leq f(x_k) + T_{x_k}(x_{k+1} - x_k) \leq f(x_k) + R_{x_k}(x_{k+1} - x_k) = 0 \quad (11)$$

$$f(\hat{x}) \leq f(x_{k+1}) + T_{x_{k+1}}(\hat{x} - x_{k+1}) \quad (12)$$

and so  $f(x_{k+1}) \leq 0$ . Up till now we have only made use of the hypotheses (H1') - (H6') with  $K = -C$ . If we assume  $f(\hat{x}) > 0$  then  $\sigma(T_{x_{k+1}}) \subset \mathbb{C}_-$  follows as in Theorem 4.2. This concludes the proof of (i) for the alternative assumptions.

Now assume that Assumption 4.8 holds and suppose that  $T_{x_{k+1}}$  is not stable. By Theorem 2.6 this is equivalent to condition (7), which by the inequalities (8) implies  $\langle f(x_{k+1}), v \rangle = 0$ . Hence by (10) and (11)

$$\langle f(x_{k+1}) - f(x_k), v \rangle = \langle R_{x_k}(x_{k+1} - x_k), v \rangle \geq \langle T_{x_k}(x_{k+1} - x_k), v \rangle \geq \langle f(x_{k+1}) - f(x_k), v \rangle$$

where we obviously have equality everywhere. Since  $x_{k+1} - x_k \in \text{int } K$  we can apply Proposition 3.5 to obtain  $T_{x_k}^*(v) = T_{x_{k+1}}^*(v) = \beta v$ , in contradiction to  $\sigma(T_{x_k}) \subset \mathbb{C}_-$ . Thus  $T_{x_{k+1}}$  must be stable.

The proof of (ii) now is identical to the proof of (ii) from Theorem 4.2.  $\square$

It is clear that even under the conditions of Proposition 4.6 we cannot expect quadratic convergence of the sequence  $(x_k)$  in Theorem 4.9. But we can still expect at least linear convergence:

**Proposition 4.10** *Assume the situation of Theorem 4.9 and let  $f$  be Fréchet-differentiable in a small ball  $U$  around  $x_+$  such that  $T_x = f'_x$ . Moreover assume that the  $T_x$  satisfy a*

*Lipschitz condition*  $\|T_x - T_y\| \leq L\|x - y\|$  for all  $x, y \in U$ .

If the  $R_x$  depend continuously on  $x$  in  $U$  and  $x_+$  is stabilizing, then the sequence  $(x_k)$  converges linearly to  $x_+$ : There exists a constant  $0 \leq \theta < 1$  and an equivalent norm  $\|\cdot\|_+$  on  $X$ , such that for sufficiently large  $k$

$$\|x_{k+1} - x_+\|_+ \leq \theta \|x_k - x_+\|_+ . \quad (13)$$

**Proof:** By Theorem 2.7 and the stability of  $f'_{x_+} = R_{x_+} + P_{x_+}$ , we have  $\rho(R_{x_+}^{-1}P_{x_+}) < \theta$  for some  $\theta < 1$ . Hence (e.g. [16]) there exists an equivalent norm  $\|\cdot\|_+$  on  $X$  with the corresponding operator norm also denoted by  $\|\cdot\|_+$ , such that  $\|R_{x_+}^{-1}P_{x_+}\|_+ < \theta$ . By continuity also  $\|R_{x_k}^{-1}P_{x_k}\|_+ < \theta$  for sufficiently large  $k$ .

We denote the second order remainder term of the Taylor expansion of  $f$  at  $x$  by  $\phi_x$ :

$$\forall x, y \in U : f(y) = f(x) + f'_x(y - x) + \phi_x(y - x) \quad (14)$$

with  $\|\phi_x(y - x)\| \leq L\|y - x\|^2$ .

By the iteration scheme (9) we have

$$\begin{aligned} x_{k+1} - x_+ &= x_k - R_{x_k}^{-1} \left( f(x_+) + (R_{x_k} + P_{x_k})(x_k - x_+) + \phi_{x_k}(x_k - x_+) \right) - x_+ \\ &= -R_{x_k}^{-1}P_{x_k}(x_k - x_+) - R_{x_k}^{-1}\phi_{x_k}(x_k - x_+) , \end{aligned}$$

whence  $\|x_{k+1} - x_+\|_+ \leq \theta \|x_k - x_+\|_+$  for sufficiently large  $k$ . □

**Remark 4.11** Note that for  $\theta$  we could choose any number  $\rho(R_{x_+}^{-1}P_{x_+}) < \theta < 1$ . In particular if  $R_{x_+} = T_{x_+}$ ,  $P_{x_+} = 0$ , i.e. the iteration is just the Newton iteration in the limit, we have  $\rho(R_{x_+}^{-1}P_{x_+}) = 0$  and can make  $\theta$  arbitrarily small. This reflects the transition to a quadratically convergent sequence. In fact, if e.g.  $\|P_x\|_+ \leq L_+\|x - x_+\|_+$  for some  $L_+ \geq 0$  and all  $x \in U$ , we obviously regain quadratic convergence.

## 5 Application to generalized algebraic Riccati equations

In this section we will show, that operators satisfying Assumption 4.1 arise naturally in the theories of linear quadratic control (classical LQ control,  $H^\infty$ -control) of deterministic and stochastic linear systems and also in realization theory. We confine ourselves to the finite dimensional case. Throughout this section  $X$  will be the real Banach space  $\mathcal{H}^n$ , ordered by the closed, solid, regular convex cone  $\mathcal{H}_+^n$  (see Example 2.2,(ii)).

### 5.1 Resolvent positive operators and stability

The asymptotic stability of various classes of linear systems can be characterized by linear matrix inequalities of Lyapunov type:

1. The continuous-time deterministic system  $\dot{x} = Ax$  is asymptotically stable if and only if  $\exists X > 0 : T_1(X) = A^*X + XA < 0$ .

2. The discrete-time deterministic system  $x_{k+1} = Ax_k$  is asymptotically stable if and only if  $\exists X > 0 : T_2(X) = A^*XA - X < 0$ .
3. The continuous-time stochastic system  $dx = Ax dt + A_0x dw$  is mean-square stable if and only if  $\exists X > 0 : T_3(X) = A^*X + XA + A_0^*XA_0 < 0$ .
4. The discrete-time stochastic system  $x_{k+1} = Ax_k + A_0x_k w_k$  is mean-square stable if and only if  $\exists X > 0 : T_4(X) = A^*XA - X + A_0^*XA_0 < 0$ .
5. The deterministic delay system  $\dot{x}(t) = Ax(t) + A_1x(t-h)$  is asymptotically stable for all delays  $h > 0$  if  $\exists X > 0 : T_5(X) = A^*X + XA + X + A_1^*XA_1 < 0$ .

Note that the operators  $T_1, \dots, T_5$  are all of the general form

$$T(X) = A^*X + XA + \sum_{i=0}^N A_i^*XA_i, \quad (15)$$

and thus resolvent positive by Example 2.4 and Lemma A.2.

## 5.2 Stabilization and rational matrix equations

Let us now add a control input to the above systems:

1.  $\dot{x} = Ax + Bu$ ,
2.  $x_{k+1} = Ax_k + Bu_k$ ,
3.  $dx = Ax dt + A_0x dw + Bu dt + B_0u dw$ ,
4.  $x_{k+1} = Ax_k + A_0x_k w_k + Bu_k + B_0u_k w_k$ ,
5.  $\dot{x}(t) = Ax(t) + A_1x(t-h) + Bu(t)$ .

The feedback stabilization problem consists in finding a matrix  $F$  such that the control input  $u(t) = Fx(t)$  or  $u_k = Fx_k$ , respectively, stabilizes the given system. For instance, in case 1.) one tries to find an  $F$ , such that  $\dot{x} = (A + BF)x$  is stable. By our stability criterion, the latter is equivalent to the existence of an  $X > 0$ , such that  $(A + BF)^*X + X(A + BF) < 0$ . In the general case with an operator  $T$  of the form (15) we have to find an  $F$ , such that there exists an  $X > 0$  satisfying

$$T_F(X) := (A + BF)^*X + X(A + BF) + \begin{bmatrix} I \\ F \end{bmatrix}^* \begin{bmatrix} A_0^*XA_0 & A_0^*XB_0 \\ B_0^*XA_0 & B_0^*XB_0 \end{bmatrix} \begin{bmatrix} I \\ F \end{bmatrix} \leq Y \quad (16)$$

for some  $Y < 0$ . (For simplicity of notation we have assumed  $N = 0$ .)

It follows from Theorem 2.7, that if (16) admits a positive definite solution for some  $Y < 0$  then it has a solution  $X > 0$  for arbitrary  $Y < 0$ .

In particular we choose  $Y = - \begin{bmatrix} I \\ F \end{bmatrix} M \begin{bmatrix} I \\ F \end{bmatrix}$ , where  $M = \begin{bmatrix} P_0 & S_0 \\ S_0^* & Q_0 \end{bmatrix}$  is a given weight

matrix. By an adequate choice of  $M$  one can express the performance criteria of different control problems (classical LQ-control,  $H^\infty$  control) for the above systems by equations of

the form (16), see [7], [6]. Here we restrict ourselves to the case  $M > 0$ . Then we can write (16) as

$$P(X) + F^*S(X)^* + S(X)F + F^*Q(X)F \leq 0, \quad (17)$$

where  $P$ ,  $S$  and  $Q$  are given by

$$\begin{aligned} P(X) &= A^*X + XA + A_0^*XA_0 + P_0, \\ S(X) &= A_0^*XB_0 + XB + S_0, \\ Q(X) &= B_0^*XB_0 + Q_0. \end{aligned}$$

By the simple inequality for  $Q > 0$

$$P + F^*S^* + SF + F^*QF = P + (F^* + SQ^{-1})Q(F + Q^{-1}S^*) - SQ^{-1}S^* \geq P - SQ^{-1}S^*,$$

we see that (17) implies  $P(X) - S(X)Q(X)^{-1}S(X)^* \leq 0$ , and, conversely, the latter implies (17) if we set  $F = -Q^{-1}S^*$ .

Thus we have reduced an optimal stabilization problem to the solution of the Riccati type matrix inequality

$$\mathcal{R}(X) = P(X) - S(X)Q(X)^{-1}S(X)^* \geq 0.$$

If we set  $D = \text{dom } \mathcal{R} = \{X \in \mathcal{H}^n : \det Q(X) \neq 0\}$  and  $D_+ = \{X \in \mathcal{H}^n : Q(X) > 0\}$ , then one can prove (see [7]) that  $\mathcal{R}$  satisfies Assumption 4.1, provided that the underlying control system is stabilizable. In particular one proves that the Fréchet derivative of  $\mathcal{R}$  at  $X$  is given by  $T_F$  in (16) with  $F = Q(X)^{-1}S(X)^*$ . From this it is immediate to see, that  $T_F$ , as the sum of a Lyapunov operator and a positive operator, is resolvent positive. Theorem 4.2 – as an existence theorem – yields, that there exists a stabilizing solution to the Riccati equation  $\mathcal{R}(X) = 0$  if and only if the inequality (16) has a positive definite solution  $X$  for some feedback matrix  $F$ , i.e. if and only if the underlying system can be stabilized. In this case  $F = Q(X)^{-1}S(X)^*$  provides us with a stabilizing feedback satisfying the performance requirement represented by  $M$ .

Note that the main ingredients of Assumption 4.1 – *resolvent positivity*, *concavity*, and *stabilizability* – are inherent in the optimal stabilization problem: The first comes from the Lyapunov approach, the second from the quadratic performance criterion and the third is a part of the problem itself.

### 5.3 $L^2$ -sensitivity optimization of realizations

In [25], and [12] problems of optimizing the realization of a transfer function were considered. Without going into the details we sketch the basic elements:

Given a strictly proper rational matrix  $G(s) \in \mathbb{R}^{p \times m}(s)$  of McMillan degree  $n$  and a minimal realization  $(A_0, B_0, C_0) \in L_{n,m,p}(\mathbb{R}) := \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m} \times \mathbb{R}^{p \times n}$

$$G(s) = C_0(sI - A_0)^{-1}B_0,$$

the set of all minimal realizations of  $G(s)$  is given as the orbit of  $(A_0, B_0, C_0)$  under the similarity action  $(S, (A, B, C)) \mapsto (SAS^{-1}, SB, CS^{-1})$  of  $\text{Gl}_n(\mathbb{R})$  on  $L_{n,m,p}(\mathbb{R})$ . Following

[12] we interpret the  $(A, B, C) \in L_{n,m,p}(\mathbb{R})$  as discrete time systems and define the  $L^2$ -sensitivity measure of a realization  $(A, B, C)$  by

$$\Gamma_2(A, B, C)^2 := \left\| \frac{\partial G}{\partial A} \right\|_2^2 + \left\| \frac{\partial G}{\partial B} \right\|_2^2 + \left\| \frac{\partial G}{\partial C} \right\|_2^2.$$

Now let

$$\hat{A} = \begin{bmatrix} A_0 & B_0 C_0 \\ 0 & A_0 \end{bmatrix}, \quad \hat{B} = \mathcal{S}_{A_0}^{-1}(B_0^* B_0), \quad \hat{C} = \mathcal{S}_{A_0}^{-1}(C_0^* C_0), \quad P_1 = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0 \\ I \end{bmatrix},$$

where  $\mathcal{S}_{A_0}$  denotes the Stein operator associated with  $A_0$  (Example 2.4), and define

$$\Pi_1(X) = P_1^* \mathcal{S}_{\hat{A}}^{-1}(P_2 X P_2^*) P_1^*, \quad \Pi_2(X) = P_2^* \mathcal{S}_{\hat{A}^*}^{-1}(P_1 X P_1^*) P_2.$$

Note that the operators  $\Pi_1$  and  $\Pi_2$  are completely positive, i.e. they have a representation of the form  $X \mapsto \sum A_i^* X A_i$ . Moreover,  $\hat{B}$  and  $\hat{C}$  are the controllability and the observability Gramians, respectively, of  $(A_0, B_0, C_0)$ , and thus positive definite by the minimality of the realization.

It was shown in [12] that a realization  $(A, B, C) = (S A_0 S^{-1}, S B_0, C_0 S^{-1}) \in L_{n,m,p}(\mathbb{R})$  minimizes the  $L^2$ -sensitivity measure if and only if  $X = S^* S$  solves the matrix equation

$$\mathcal{R}(X) := \Pi_1(X) + \hat{C} - X \left( \Pi_2(X^{-1}) + \hat{B} \right) X = 0. \quad (18)$$

The operator  $\mathcal{R}$  is well defined on  $\text{dom } \mathcal{R} = \{X \in \mathcal{H}^n \mid \det X \neq 0\}$ . We wish to solve equation (18) in  $D_+ := \text{int } \mathcal{H}_+^n \subset \text{dom } \mathcal{R} =: D$ , and verify the conditions of Theorem 4.2. By a straightforward calculation we obtain the explicit form of the Fréchet derivative of  $\mathcal{R}$ :

**Lemma 5.1** *The Fréchet derivative  $\mathcal{R}'_X(H)$  of  $\mathcal{R}$  is given by*

$$\mathcal{R}'_X(H) = \mathcal{L}_{-X(\Pi_2(X^{-1}) + \hat{B}^*)}(H) + \Pi_1(H) + X \Pi_2(X^{-1} H X^{-1}) X. \quad (19)$$

*Being the sum of a Lyapunov operator and positive operators,  $\mathcal{R}'_X$  is resolvent positive.*

**Lemma 5.2** *The operator  $\mathcal{R}$  is  $D_+$ -concave on  $D$ .*

**Proof:**  $\Pi_1$  and the quadratic mapping  $X \mapsto -X \hat{B}^* X$  are obviously  $D_+$ -concave on  $D$ . Thus it remains to analyze the operator  $X \mapsto \mathcal{F}(X) := -X \Pi_2(X^{-1}) X$ . For  $Z \in D$  and  $Y \in D_+$  we have

$$\begin{aligned} \mathcal{F}(Z) - \mathcal{F}(Y) + \mathcal{F}'_Z(Y - Z) &= -Z \Pi_2(Z^{-1}) Z + Y \Pi_2(Y^{-1}) Y - Y \Pi_2(Z^{-1}) Z \\ &\quad + Z \Pi_2(Z^{-1}) Z - Z \Pi_2(Z^{-1}) Y + Z \Pi_2(Z^{-1}) Z \\ &\quad + Z \Pi_2(Z^{-1} Y Z^{-1}) Z - Z \Pi_2(Z^{-1}) Z \\ &= \begin{bmatrix} Y \\ Z \end{bmatrix}^* \begin{bmatrix} \Pi_2(Y^{-1}) & -\Pi_2(Z^{-1}) \\ -\Pi_2(Z^{-1}) & \Pi_2(Z^{-1} Y Z^{-1}) \end{bmatrix} \begin{bmatrix} Y \\ Z \end{bmatrix} \geq 0. \end{aligned}$$

The inequality holds because  $\begin{bmatrix} Y^{-1} & -Z^{-1} \\ -Z^{-1} & Z^{-1} Y Z^{-1} \end{bmatrix} = \begin{bmatrix} Y^{-1} \\ Z^{-1} \end{bmatrix} Y \begin{bmatrix} Y^{-1} \\ Z^{-1} \end{bmatrix}^* \geq 0$  for  $Y > 0$  and  $\Pi_2$  has a representation of the form  $X \mapsto \sum A_i^* X A_i$ . Thus  $\mathcal{F}$  is  $D_+$ -concave on  $D$ ,

which completes the proof.  $\square$

It follows from the positive definiteness of  $\hat{B}$  and  $\hat{C}$ , that  $\mathcal{R}'_{tI}$  is stable for sufficiently large  $t > 0$  and that  $\mathcal{R}(tI) > 0$  for sufficiently small  $t > 0$ . Thus Theorems 4.2 and 4.9 can be applied, to solve the equation  $\mathcal{R}(X) = 0$  by fixed-point iterations. As a starting point we can choose  $X_0 = tI$  for sufficiently large  $t > 0$ .

If we apply the Newton-iteration, we have to solve a linear equation of the form

$$\mathcal{L}_{-X_k(\Pi_2(X_k^{-1})+\hat{B}^*)}(H_k) + \Pi_1(H_k) + X_k\Pi_2(X_k^{-1}H_kX_k^{-1})X_k = \mathcal{R}(X_k) \quad (20)$$

for  $H_k$  in each step to obtain the next iterate  $X_{k+1} = X_k - H_k$ . Since the number of unknown scalar entries in  $H_k$  is  $n(n+1)$ , a direct solution of this equation would require  $O(n^6)$  steps. As has been pointed out in [10], and in view of Theorem 4.9 it might be advantageous to solve the following simplified equation for  $H_k$

$$\mathcal{L}_{-X_k(\Pi_2(X_k^{-1})+\hat{B}^*)}(H_k) = \mathcal{R}(X_k). \quad (21)$$

This is just a Lyapunov equation and can be solved efficiently by the Bartels-Stewart algorithm [1] in  $O(n^3)$  steps. But obviously there is a trade-off between the rate of convergence and the complexity of the linear equations to be solved.

To get a rough idea, we have tried the following mixed strategy, using the cheaper fixed-point iteration (21) as long as the step length (i.e. the norm  $h_1 := \|H_k\|$ ) decreases not too fast; that is, if  $\|H_k\|$  exceeds  $\|H_{k-1}\|$  divided by some level parameter  $\ell$ :

```

Generate an  $X_0 = t_0I > 0$ , such that  $\sigma(\mathcal{R}_{X_0}) \subset \mathbb{C}_-$ .
Set  $h_0 = 0$ ,  $h_1 = 1$ ,  $\ell = 2$ 
WHILE  $h_1 > \epsilon$ 
  IF  $h_1 < h_0/\ell$ : Solve (20) and set  $\ell = \ell + 1$ .
  ELSE Solve (21) and set  $\ell = \ell - 2/n$ .
  Set  $X_{k+1} = X_k - H_k$ ,  $h_0 = h_1$ ,  $h_1 = \|H_k\|$ 
END

```

If the step length  $\|H_k\|$  is smaller than the tolerance level  $\epsilon$ , the iteration stops.

For a few randomly generated examples of different dimensions  $n$  we compared the computing time (in seconds) for the pure Newton iteration and the mixed method:

dim $A = n$	5	10	15	20
pure Newton	4.0	23.4	101	352
mixed method	2.9	22.2	86	263

Though the numbers are not necessarily representative they indicate, that for large dimensions one might take advantage of a combination of the two methods. For the iteration (21) alone, the convergence is much slower.

We also applied both methods to the numerical example suggested in [25] with

$$A_0 = \begin{bmatrix} 0.5 & 0 & 1 \\ 0 & -0.25 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad C_0 = \begin{bmatrix} 1 \\ 5 \\ 10 \end{bmatrix}^\top, \quad X_+ = \begin{bmatrix} 0.2 & 0 & 0.5 \\ 0 & 5 & 0 \\ 0.5 & 0 & 5 \end{bmatrix}.$$

From the initial matrix  $X_0 = 40I$  the sequence converged to  $X_+$  in 16 (Newton) respectively 23 (mixed method) steps. Factorizing  $X_+ = S^*S$  one obtains a realization  $(A, B, C) = (SA_0S^{-1}, SB_0, C_0S^{-1})$  of minimal  $L^2$ -sensitivity.

# A Appendix

The elementary results summarized in the following lemma are taken from [17]:

**Lemma A.1** (i) *If  $C$  is solid, then it is reproducing. In finite-dimensional spaces the converse is also true.*

(ii) *If  $C$  is normal, then it is pointed. In finite-dimensional spaces the converse is also true.*

(iii)  *$C$  is normal if and only if  $C^*$  is reproducing ([17], Theorem 4.5).*

(iv)  *$C$  is reproducing if and only if  $C^*$  is normal ([17], Theorem 4.6).*

(v) *If  $C$  is regular, then it is normal ([17], Theorem 5.1).*

The following lemma summarizes some elementary properties of (resolvent) positive operators which we will use in the proofs of Theorems 2.6 and 2.7.

**Lemma A.2** *Let  $X$  be a real Banach space, ordered by a normal, solid pointed convex cone  $C$ , and let  $S, T : X \rightarrow X$  be bounded linear operators and  $S \geq T$ . Then*

(i) *If  $T$  is positive, then so is  $S$  and  $\rho(S) \geq \rho(T)$ .*

(ii) *If  $T$  is resolvent positive, then so is  $S$  and  $\beta(S) \geq \beta(T)$ .*

(iii) *If  $T$  is resolvent positive, then*

$$\alpha I - T \text{ is inverse positive} \iff \alpha > \beta(T) \iff \sigma(\alpha I - T) \subset \mathbb{C}_+.$$

Moreover,  $\beta(T) = \alpha - \frac{1}{\rho((\alpha I - T)^{-1})}$  for all  $\alpha > \beta(T)$ .

**Proof:** See e.g. [19] or [4] for (i), [9] for (iii), and [7] for the first part of (ii) in the case  $X = \mathcal{H}^n$  (the proof carries over immediately); the second part of (ii) follows from (i) and (iii).  $\square$

**Lemma A.3** *Let  $X$  be an ordered Banach space. If  $S \in \mathcal{L}(X)$  is inverse positive and  $T$  is positive, such that  $\rho(S^{-1}T) < 1$ , then  $S - T$  is inverse positive.*

**Proof:** By assumption  $(S - T)^{-1} = (I - S^{-1}T)^{-1}S^{-1} = \sum_{k=0}^{\infty} (S^{-1}T)^k S^{-1}$  is a convergent series of positive operators and thus positive.  $\square$

**Proof of Theorem 2.6:** For  $\alpha > \beta(T) = \beta(T^*)$  the operator  $(\alpha I - T^*)^{-1}$  is positive. By (a) there exists a  $v \in C^*$ , such that  $(\alpha I - T^*)^{-1}v = \rho v$ , where  $\rho = \rho((\alpha I - T^*)^{-1})$ . Multiplying this equation from the left by  $\alpha I - T^*$  we obtain  $v = \rho \alpha v - \rho T^* v$ , that is  $T^* v = (\alpha - \frac{1}{\rho})v$ . It remains to show, that  $\beta(T^*) = \alpha - \frac{1}{\rho}$ . But this follows from Lemma A.2 (iii).  $\square$

**Proof of Theorem 2.7:** (i)  $\Leftrightarrow$  (ii): By Lemma A.2 (ii)  $T$  is resolvent positive and thus by

the last statement of the same lemma the first two conditions in Theorem 2.7 are equivalent. (ii) $\Rightarrow$ (iii): If  $-T$  is inverse positive then  $-T^{-1}$  maps  $\text{int } C$  into  $\text{int } C$ , which implies (iii). (iii) $\Rightarrow$ (iv) and (iv) $\Rightarrow$ (v) are trivial. (v) $\Rightarrow$ (i): Assume that  $x \in C$  satisfies  $-T(x) \in \text{int } C$ , but (i) fails. Then  $\beta := \beta(T) \geq 0$  and  $T^*$  has an eigenvector  $v \in C^*$  corresponding to  $\beta$  by Theorem 2.6. But this implies a contradiction:

$$0 > \langle Tx, v \rangle = \langle x, T^*v \rangle = \langle x, \beta v \rangle \geq 0,$$

hence (i) holds. It remains to prove that (vi) is equivalent to (i)–(v).

(vi) $\Rightarrow$ (ii): Suppose (vi), then  $-R$  is inverse positive by Lemma A.2 (iii) and  $\rho((-R)^{-1}P) < 1$ . Applying Lemma A.3 we obtain that  $-T = -R - P$  is inverse positive, i.e. (ii).

(i), (ii), (iv) $\Rightarrow$ (vi): Since  $\beta(R + P) \geq \beta(R)$  by Lemma A.2 (ii), condition (i) implies  $\sigma(R) \subset \mathbb{C}_-$ , whence  $-R$  is inverse positive and  $\Pi := -R^{-1}P$  is positive. By (iv) there exists a positive vector  $x \in \text{int } C$ , such that  $-Tx \in \text{int } C$  and since  $-R^{-1} \geq 0$  this implies  $R^{-1}Tx = (I - \Pi)x \in \text{int } C$ . But by Theorem 2.6 there exists a  $v \in C^*$ , such that  $\Pi^*v = \rho(\Pi)v$ . Therefore  $0 < \langle (I - \Pi)x, v \rangle = (1 - \rho(\Pi))\langle x, v \rangle$ , whence  $\rho(\Pi) < 1$  because  $\langle x, v \rangle > 0$ .  $\square$

## References

- [1] R. H. Bartels and G. W. Stewart. Solution of the matrix equation  $AX + XB = C$ : Algorithm 432. *Comm. ACM*, 15:820–826, 1972.
- [2] A. Berman, M. Neumann, and R. Stern. *Nonnegative Matrices in Dynamic Systems*. John Wiley & Sons, New York, 1989.
- [3] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Classics in Applied Mathematics. SIAM, 1994.
- [4] L. Burlando. Monotonicity of spectral radius for positive operators on ordered Banach spaces. *Arch. Math.*, 56:49–57, 1991.
- [5] W. A. Coppel. Matrix quadratic equations. *Bull. Austr. Math. Soc.*, 10:377–401, 1974.
- [6] T. Damm. State-feedback  $H^\infty$ -type control of linear systems with time-varying parameter uncertainty. *Lin. Alg. Appl.*, 2001. Submitted.
- [7] T. Damm and D. Hinrichsen. Newton’s method for a rational matrix equation occurring in stochastic control. Report 443, Institut für Dynamische Systeme, Universität Bremen, 1999. Accepted by *Lin. Alg. Appl.*
- [8] L. Elsner. Quasimonotonie und Ungleichungen in halbgeordneten Räumen. *Lin. Alg. Appl.*, 8:249–261, 1974.
- [9] A. Fischer, D. Hinrichsen, and N. K. Son. Stability radii of Metzler operators. *Vietnam J. of Mathematics*, 26:147–163, 1998.
- [10] C.-H. Guo. Iterative solution of a matrix Riccati equation arising in stochastic control. Preprint, Department of Mathematics and Statistics, University of Regina, 2000.

- [11] C.-H. Guo and P. Lancaster. Analysis and modification of Newton's method for algebraic Riccati equations. *Math. Comp.*, 67(223):1089–1105, 1998.
- [12] U. Helmke and J. B. Moore.  $L^2$  sensitivity minimization of linear system representations via gradient flows. *J. Math. Syst. Estim. Control*, 5(1):79–98, 1995.
- [13] G. A. Hewer. An iterative technique for the computation of steady state gains for the discrete optimal regulator. *IEEE Trans. Automat. Contr.*, AC-16:382–384, 1971.
- [14] L. V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Mat. Nauk*, 3(6):89–185, 1948.
- [15] D. L. Kleinman. On an iterative technique for Riccati equation computation. *IEEE Trans. Automat. Contr.*, AC-13:114–115, 1968.
- [16] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Y. B. Rutitskii, and V. Y. Stetsenko. *Approximate Solution of Operator Equations*. Wolters-Noordhoff, Groningen, 1972.
- [17] M. A. Krasnosel'skij, J. A. Lifshits, and A. V. Sobolev. *Positive Linear Systems - The Method of Positive Operators*, volume 5 of *Sigma Series in Applied Mathematics*. Heldermann Verlag, Berlin, 1989.
- [18] M. G. Krein and M. A. Rutman. Linear operators leaving invariant a cone in a Banach space. *Amer. Math. Soc. Transl.*, 26:199–325, 1950.
- [19] I. Marek. Frobenius theory of positive operators: Comparison theorems and applications. *SIAM Journal of Applied Mathematics*, 19:607–628, 1970.
- [20] H. H. Schaefer. *Topological Vector Spaces*. Springer-Verlag, Berlin, Heidelberg, New York, 1971.
- [21] H. Schneider. Positive operators and an inertia theorem. *Numerische Mathematik*, 7:11–17, 1965.
- [22] H. Schneider and M. Vidyasagar. Cross-positive matrices. *SIAM J. Numer. Anal.*, 7(4):508–519, December 1970.
- [23] J. S. Vandergraft. Newton's method for convex operators in partially ordered spaces. *SIAM J. Numer. Anal.*, 4(3):406–432, 1967.
- [24] W. M. Wonham. On a matrix Riccati equation of stochastic control. *SIAM J. Control Optim.*, 6:681–698, 1968.
- [25] W.-Y. Yan, J. B. Moore, and U. Helmke. Recursive algorithms for solving a class of non-linear matrix equations with applications to certain sensitivity optimization problems. *SIAM J. Cont.*, 32:1559–1576, 1994.