Abstract

In this paper, we consider the matricial version of generalized moment problem with degree constraint. Specifically we focus on computing the solution that minimize the Kullback-Leibler criterion. Several strategies to find such optimum via descent methods are considered and their convergence studied. In particular a parameterization with better numerical properties is derived from a spectral factorization problem. Such parameterization, in addition to guaranteeing descent methods to be globally convergent, it appears to be very reliable in practice.

A.1 Introduction

Generalized moment matching can be viewed as a broad family of moment problems that includes covariance extension and Nevanlinna-Pick interpolation as notable special cases. They arise in spectral estimation and robust control and find application in signal processing, speech recognition and coding, $H^\infty$ control and bioengineering. In particular, the complexity of the interpolant is of great interest for practical applicability. From this point of view, Byrnes, Georgiu, Lindquist and various collaborators has been developing a formulation for generalized moment problems that takes into account complexity in the form of degree constraints. In such approach the interpolants are obtained by solving a convex optimization problem. Despite the favorable properties of the resulting optimization, it has been noted that descent algorithms struggle to converge in certain conditions. The main difficulty lies in the numerical properties of the objective function near the boundary. The objective of the present work is to develop a descent algorithm able to
overcome such issue both from a theoretical and a practical point of view. In order to do so several approaches are proposed and compared.

A.2 Notation

For a matrix $M$, $\overline{M}$ denotes component-wise complex conjugate, $M^T$ denotes the transpose, whereas $M^*$ denotes its hermitian conjugate.

Let $\langle \cdot, \cdot \rangle$ denote the standard $L_2[-\pi, \pi]$ inner product for matrix valued square-integrable functions on the unit circle $\mathbb{T}$, – i.e.

\[
\langle A, B \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \text{tr} \{ A(e^{i\theta}) B(e^{i\theta})^* \} \, d\theta.
\]  

This inner product induces the norm $\|A\| = \sqrt{\langle A, A \rangle}$. In particular, if $A$ and $B$ are constant matrices then $\langle A, B \rangle = \text{tr} \{ AB^* \}$ and $\|A\|$ is the Frobenius norm.

In the norm (A.1) the adjoint is defined by $A^*(z) = \overline{A(z^{-1})}^T$, and then

\[
\langle A, B \rangle = \int_{\mathbb{T}} \text{tr} \{ A(z) B^*(z) \} \frac{dz}{z}.
\]

Additionally let $\|\cdot\|_2$ denote the matrix norm induced by the Euclidean vector norm – i.e

\[
\|M\|_2 = \max_v \frac{\|Mv\|_2}{\|v\|_2},
\]

It can be shown that $\|M\|_2$ is equal the maximum singular value of $M$. It relates to the Frobenius norm as the following

\[
\|M\|_2 \leq \|M\| \leq \sqrt{r} \|M\|_2
\]  

where $r$ is the rank of $M$.

A.3 State-covariance interpolation

Consider an $m$-dimensional, zero-mean, Gaussian, stationary stochastic vector process $\{x(t)\}_{t \in \mathbb{Z}}$ whose matrix-valued spectral density is defined as

\[
\Phi(z) = \sum_{j=-\infty}^{\infty} R_j z^j,
\]

where

\[
R_j := E \{ x(j)x(0) \}.
\]  

and where we assume that $\Phi(z) > 0$ for all $z \in \mathbb{T}$. Let $\mathcal{S}^m_\mathbb{T}$ denote the class of such spectral densities that are integrable on $\mathbb{T}$. 

Input-to-state filters is a simple tool that can be used to formulate many interpolation problems in a uniform fashion. An input-to-state filter \( G \) is defined by a matrix \( A \in \mathbb{C}^{n \times n} \) with all its eigenvalues in the open unit disc and a full-column rank matrix \( B \in \mathbb{C}^{n \times m} \) such that \((A, B)\) is reachable by
\[
G(z) = (zI - A)^{-1}B. \tag{A.5}
\]

The state-covariance of a spectral density \( \Phi \in \mathcal{S}_m^+ \) with respect to the input-to-state filter \([A.5]\) is related to the functional \( \Gamma \)
\[
\Gamma(F) = \int_{\mathbb{T}} G(z)F(z)G^*(z)\frac{dz}{z}. \tag{A.6}
\]
defined over the set of Hermitian valued functions that are integrable on \( \mathbb{T} \). Specifically, the state-covariance of \( \Phi \in \mathcal{S}_m^+ \) is defined as \( \Sigma = \Gamma(\Phi) \). Let \( \mathcal{L} \) be the range of \( \Gamma \) and denote \( \mathcal{L}_+ \) the image of \( \mathcal{S}_m^{+} \) i.e, \( \mathcal{L}_+ = \Gamma(\mathcal{S}_m^{+}) \).

For a given \( \Sigma \in \mathbb{C}^{m \times m} \), the corresponding interpolation problem consist mainly in the following:

- Find if \( \Sigma \) belongs to \( \Gamma(\mathcal{S}_m^{+}) \)
- If so, find \( \Phi \in \mathcal{S}_m^{+} \) such that \( \Gamma(\Phi) = \Sigma \).

The range set \( \mathcal{L}_+ \) has been completely characterized in \([7]\). Specifically a positive semi-definite \( \Sigma \) belongs to \( \mathcal{L}_+ \) if and only if there exists a matrix \( H \in \mathbb{C}^{m \times n} \) such that
\[
\Sigma - A\Sigma A^* = BH + H^*B^*. \tag{A.7}
\]

In this context, the poles of \( G \) or equivalently, the eigenvalues of \( A \), are denoted as interpolation points. If such points are placed all in the origin, the interpolation above is equivalent to a covariance extension problem.

### A.3.1 An optimization approach

Optimization theory can be employed to study the state-covariance interpolation problem. From this point of view, state-covariance matching is a feasibility problem, that is
\[
\begin{bmatrix}
\text{find } \Phi \\
\text{s.t. } \left\{ \begin{array}{l}
\int G\Phi G^* = \Sigma, \\
\Phi \in \mathcal{S}_m^{+}.
\end{array} \right.
\end{bmatrix}
\]

Tools from duality theory may be used to study feasibility and evaluate a solution.

It is natural now to modify the above feasibility problem into a maximization
\[
\begin{bmatrix}
\text{max } \Phi \\
\text{s.t. } \left\{ \begin{array}{l}
\int G\Phi G^* = \Sigma, \\
\Phi \in \mathcal{S}_m^{+}.
\end{array} \right.
\end{bmatrix}
\]

\(^1\) Maximization is chosen so that the dual problem, the one that will actually be solved, will turn to be a minimization.
where \( \mathbb{D} \) is any concave functional. The choice of \( \mathbb{D} \) will determine the structure of the solution for \( \mathcal{P} \) and whether it can be found via duality theory. One such choice is the entropy
\[
\mathbb{D}(\Phi) = \int_T \log \det \Phi(z) \frac{dz}{z},
\]
whose corresponding interpolant is denoted as maximum entropy solution and has the form
\[
\Phi_{ME}(z) = [G^*(z) \Lambda_{ME} G(z)]^{-1},
\]
for some \( \Lambda_{ME} \in \mathcal{L}_+ \).

Next, we develop a multivariable version of the duality theory in [8]. The primal functional that will be considered here is the Kullback-Leibler pseudo distance [8]
\[
\mathbb{D}(\Phi) = - \langle \Psi, \log(\Psi^{-1}) \rangle,
\]
between \( \Phi \) and a given \( \Psi \in \mathcal{S}^1_+ \). The spectral density \( \Psi \) can be used to embed in the problem some prior knowledge about the solution, it is particularly useful to convey information about positions of transmission zeros – i.e. where the desired solution should lose rank.

The dual optimization problem corresponding to \( \mathcal{P} \) with (A.10) as a primal functional is the following
\[
\begin{array}{ll}
\min_{\Lambda} & J_\Psi(\Lambda) \\
\text{s.t.} & G^*(z) \Lambda G(z) \geq 0, \quad \forall z \in T, \quad \Lambda \in \mathcal{L} \\
\end{array}
\]
where
\[
J_\Psi(\Lambda) = \langle \Lambda, \Sigma \rangle - \langle \Psi(z), I + \log G^*(z) \Lambda G(z) \rangle.
\]

It can be shown that the dual function is twice continuously differentiable in \( \mathcal{S}^n_+ \), in fact infinitely many times so. In particular, its first and second order differentials are
\[
\delta J_\Psi(\Lambda; \delta \Lambda) = \langle \Sigma - \Gamma(\Psi Q^{-1}), \delta \Lambda \rangle
\]
and
\[
\delta^2 J_\Psi(\Lambda; \delta \Lambda_1, \delta \Lambda_2) = \langle \Psi, Q^{-1} \delta Q_1 Q^{-1} \delta Q_2 \rangle
\]
respectively, where \( Q(z) = G^*(z) \Lambda G(z) \) and \( \delta Q_i = G^*(z) \delta \Lambda_i G(z) \).

Additionally, the following holds

**Lemma A.3.1** For any feasible \( \mathcal{P} \), the dual function \( J_\Psi(\Lambda) \) is finite in the whole dual feasibility region save for the origin where it tends to infinity.

**Proof.** Since the primal is feasible \( \mathcal{P} \) will be bounded below. Additionally, if \( \Lambda \neq 0 \), we have that the scalar rational spectrum \( \det G^*(z) \Lambda G(z) \) is positive semidefinite and admits a stable spectral factor. From the Szegö condition it follows that
\[
\int_T \log \det G^*(z) \Lambda G(z) \frac{dz}{z} > -\infty
\]
and (A.11) is also bounded above since $\Psi$ itself is bounded.

The following property of (D) is particularly important for convergence

**Proposition A.3.1** Suppose the primal optimization problem (P) is feasible, then the dual problem (D) has compact sublevel sets – i.e. for any $d \in \mathbb{R}$, the inverse image $\mathcal{J}_\Psi^{-1}(-\infty, d]$ is compact.

**Proof.** Suppose $\{\Lambda_k\}$ is any sequence in $\mathcal{J}_\Psi^{-1}(-\infty, d]$, as suggested in [3] it suffice to show that $\{\Lambda_k\}$ has a convergent subsequence.

Let $\gamma_k = \|\Lambda_k\|$ and $\bar{\Lambda}_k = \Lambda_k/\|\Lambda_k\|$, then

$$
\mathcal{J}_\Psi(\Lambda_k) = \mathcal{J}_\Psi(\gamma_k \bar{\Lambda}_k) = (\gamma_k - 1) \langle \bar{\Lambda}_k, \Sigma \rangle - \langle \Psi(z), 1 \rangle \log \gamma_k + \mathcal{J}_\Psi(\bar{\Lambda}_k).
$$

Since we assume the primal to be feasible $\Sigma \geq \lambda_{\text{min}}(\Sigma) I$ where $\lambda_{\text{min}}(\Sigma) > 0$ is the smallest eigenvalue of $\Sigma$. Then, since $\bar{\Lambda}_k \geq 0$,

$$
\langle \bar{\Lambda}_k, \Sigma \rangle \geq \lambda_{\text{min}}(\Sigma) \langle \bar{\Lambda}_k, I \rangle = \lambda_{\text{min}}(\Sigma) \text{tr} \bar{\Lambda}_k \geq \lambda_{\text{min}}(\Sigma) \|\bar{\Lambda}_k\|^2 \geq \frac{\lambda_{\text{min}}(\Sigma)}{r_k} \|\bar{\Lambda}_k\|^2 \geq \frac{\lambda_{\text{min}}(\Sigma)}{m}
$$

where $r_k$, the rank of $\bar{\Lambda}_k$, is positive since $\bar{\Lambda}_k$ cannot be equal to zero.

Finally, due to duality, for any primal feasible $\Phi$ we have

$$
\mathcal{J}_\Psi(\Lambda_k) \geq \frac{\lambda_{\text{min}}(\Sigma)}{\sqrt{m}} \gamma_k - \langle \Psi(z), 1 \rangle \log \gamma_k + \mathcal{D}_\Psi(\Phi)
$$

where the right hand side goes to infinity as $\gamma_k$ goes to either zero or infinity which is a contradiction. Thus the sequence $\{\Lambda_k\}$ has a convergent subsequence.

The nonnegativity constraint can shown to be inactive at any optima.

**Lemma A.3.2** The dual optimization problem (D) never attains a minimum in the boundary of its feasibility region.

**Proof.** Consider any $\Lambda, \bar{\Lambda}$ in the interior and boundary of the feasibility region respectively. That is $\bar{Q} = G^*\bar{\Lambda}G > 0$, $\bar{Q} = G^*\bar{\Lambda}G \geq 0$ and $\det \bar{Q}(z) = 0$ for some
$z \in T$. Now let $\Lambda_\gamma = \bar{\Lambda} + \gamma(\Lambda - \bar{\Lambda})$, then $Q_\gamma = G^*\Lambda_\gamma G > 0$. Then the directional derivative at $\Lambda_\gamma$ towards $\bar{\Lambda}$ is

$$
\delta \|_\Psi (\Lambda_\gamma; \bar{\Lambda} - \Lambda) = \langle \Sigma - \Gamma(vQ^{-1}), \bar{\Lambda} - \Lambda \rangle = \langle \Sigma, \bar{\Lambda} - \Lambda \rangle + \langle \Psi Q^{-1}, Q - \bar{Q} \rangle = \langle \Sigma, \bar{\Lambda} - \Lambda \rangle + \int_T h_\gamma(z) \frac{dz}{z}
$$

where

$$
h_\gamma(z) = \text{tr}\{\Psi Q^{-1}(z)[Q(z) - \bar{Q}(z)]\}.
$$

Evaluating the derivative of $h_\gamma(z)$ with respect to $\gamma$ we obtain

$$
\frac{d}{d\gamma} h_\gamma(z) = \text{tr}\{\Psi Q^{-2}(z)[Q(z) - \bar{Q}(z)]^2\} > 0
$$

hence $h_\gamma(z)$, for any $z \in T$, is a monotonely increasing function of $\gamma$. The rest of the proof proceeds as in [3].

□

A.3.2 On the choice of optimization’s variables

Before discussing about algorithms, it is convenient to reparameterize (D) in terms of a vector $x \in \mathbb{R}^M$. This change of optimizing variables will align our problem formulation to the one mostly used both in optimization’s theory and algorithms implementation.

Let $\Lambda_1, \Lambda_1, ... \Lambda_M$ be a basis of the linear subspace $L$, then we can parameterize all $\Lambda \in L$ by $x \in \mathbb{R}^M$ as

$$
\Lambda(x) = x_1\Lambda_1 + x_2\Lambda_2 + \ldots x_M\Lambda_M. \quad (A.14)
$$

Where $N$ is the real dimension of $L$. The following has been shown in [6].

**Proposition A.3.2** The dimension of $L$ as a space over $\mathbb{R}$ is $m(2n - m)$.

□

Now,

$$
f(x) = J_\psi(x_1\Lambda_1 + x_2\Lambda_2 + \ldots x_N\Lambda_M) \quad (A.15)
$$

being the dual objective function parameterized in $x$, we can rewrite problem (D) as

$$
\begin{bmatrix}
\min & f(x) \\
\text{s.t.} & G^*(z)\Lambda(x)G(z) > 0, \quad \forall z \in T \\
\end{bmatrix} \quad (D_x)
$$

where the constraint $\Lambda(x) \in L$ is satisfied by construction. Only strict positivity needs to be enforced since, in view of Lemma A.3.2, any optimal solution is internal.
Furthermore, it is convenient to consider such positivity constraint implicit so that simple, unconstrained minimization algorithms can be used to solve \( \{D_x\} \). To such end we must make sure that the sequence of points generated are bounded away from the boundary of the feasible region, as it will be shown later.

Note that first and second order derivatives of \((A.15)\) can be obtained in a straightforward way from \((A.12)\) and \((A.13)\) respectively in the following way:

\[
\frac{\partial f(x)}{\partial x_i} = \delta \|\phi(\Lambda(x); \Lambda_i)\|,
\]

\[
\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \delta^2 \tilde{J}_\phi(\Lambda(x); \Lambda_i, \Lambda_j)
\]

for \(i, j = 1, 2,...N\).

### A.4 Descent algorithms

In general, descent algorithms can be described as the generation of a sequence of points \( \{x_k\} \) such that

\[
x_{k+1} = x_k + \alpha_k p_k
\]

where \(\alpha_k > 0\) is the step length and \(p_k\) is a descent direction – i.e.

\[
\nabla f(x_k)^T p_k < 0.
\]

Equivalently, the descent direction can be defined from a sequence of positive definite matrices \(\{B_k\}\) as

\[
B_k p_k = -\nabla f(x_k).
\]

#### A.4.1 Line search

The step length \(\alpha_k\) is usually assumed to satisfy certain conditions of sufficient decrease. For instance, the Wolfe conditions

\[
|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq \eta_w |\nabla f(x_k)^T p_k|
\]

and

\[
f(x_k + \alpha_k p_k) \leq f(x_k) + \eta_s \alpha_k \nabla f(x_k)^T p_k,
\]

where \(0 \leq \eta_w < \eta_s < 1\), can guarantee global convergence under certain assumptions. A step length satisfying such conditions can be found performing a line search on the ray \(x_k + \alpha p_k\) for \(\alpha > 0\).

Unfortunately the most common step length strategies in the literature fails when applied to \(\{D_x\}\). This is due mainly to the fact that such line search methods can not guarantee the sequence \(\{x_k\}\) to be bounded away from the boundary where the objective function is not differentiable.
In the following discussion a simple alternative method will be proposed such that \( \{x_k\} \) is contained in a compact set in the interior of the feasible region.

Consider the squared norm of the gradient

\[
g(x) = \|\nabla f(x)\|^2, \tag{A.22}
\]

whose corresponding gradient will then be

\[
\nabla g(x) = \nabla^2 f(x) \nabla f(x). \tag{A.23}
\]

**Proposition A.4.1** Let \( p_k \) be such that (A.19) is satisfied, then \( p_k \) is also a descent direction for \( g(x) \) at \( x_k \).

**Proof.** We have

\[
p_k^T \nabla g(x_k) = p_k^T \nabla^2 f(x_k) \nabla f(x_k) = -p_k^T \nabla^2 f(x_k) B_k p_k < 0
\]

since, both \( B_k \) and \( \nabla^2 f(x_k) \) being symmetric and positive definite, \( \nabla^2 f(x_k) B_k \) has all positive eigenvalues [4].

Thus \( p_k \) is a descent direction for both the objective function (A.15) and the gradient squared norm (A.22). It is therefore possible to find an \( \bar{\alpha} > 0 \) such that for all \( \alpha \in (0, \bar{\alpha}) \) both

\[
f(x_k + \alpha p_k) \leq f(x_k) + \eta_s \alpha \nabla f(x_k)^T p_k \tag{A.24}
\]

and

\[
g(x_k + \alpha p_k) \leq g(x_k) + \eta_n \alpha \nabla g(x_k)^T p_k \tag{A.25}
\]

are satisfied for some \( \eta_s, \eta_n \in (0,1) \).

It is easy to show the following

**Lemma A.4.1** For any feasible \( x \), the set

\[
f^{-1}(-\infty, f(x)] \cap g^{-1}(-\infty, g(x)]
\]

is compact and contained in the interior of the feasibility region of \( \mathcal{D} \).

**Proof.** By Proposition [A.3.1] \( f^{-1}(-\infty, f(x)] \) is bounded, so is its intersection with \( g^{-1}(-\infty, g(x)] \). Consider a converging sequence \( \{x_k\} \) in the intersection. \( \{x_k\} \) can not approach the boundary since otherwise \( g(x_k) \) would tend to infinity. Thus it must converge to some \( \bar{x} \) in the interior of the feasibility region. Since both \( f \) and \( g \) are continuous there, we have

\[
\bar{x} \in f^{-1}(-\infty, f(x)] \cap g^{-1}(-\infty, g(x)].
\]
Any sequence \((A.18)\) such that \((A.24)\) and \((A.25)\) hold for \(\alpha_k\) is such that both
\[
\{x_k\} \subset f^{-1}(-\infty, f(x_0)] \cap g^{-1}(-\infty, g(x_0]].
\]
\[(A.26)\]  
In particular \(\{x_k\}\) will be bounded away from the boundary where the objective function is not differentiable.

The line search strategy that will be considered here is similar to Armijo-Goldstein’s. Given \(\gamma \in (0, 1)\), it consist to find the smallest integer \(j_k\) such that both \((A.24)\) and \((A.25)\) are satisfied for \(\alpha = \gamma^j\).

**A.4.2 Convergence of Newton’s method**

One of the most used choices of descent direction is such that
\[
\nabla^2 f(x_k)p_k = -\nabla f(x_k)
\]
\[(A.27)\]  
where \(B_k = \nabla^2 f(x_k)\). The resulting descent algorithm is denoted Newton’s method and is widely used in the most disparate fields.

It is important to note that the line search with conditions \((A.24)\) and \((A.25)\) requires the evaluation of the hessian \(\nabla^2 f(x_k)\) already. Therefore, in this case, Newton’s method incur in a smaller performance overhead each iteration with respect to alternative algorithms such as Quasi-Newton methods.

**Theorem A.4.1** Let \(x_0\) be any feasible point for \((D_x)\), \(\{x_k\}\) generated by \((A.18)\) and \((A.27)\). Suppose \(\alpha_k = \gamma^j_k\), for \(\gamma \in (0, 1)\), where \(j_k\) is the smallest integer such that \((A.24)\) and \((A.25)\) are satisfied by \(\alpha_k\) for some \(\eta_s, \eta_n \in (0, 1)\). Then
\[
\lim_{k \to \infty} \|\nabla f(x_k)\| = 0.
\]

**Proof.** First note that the sequence of Newton’s descent direction \(\{p_k\}\) defined in \((A.27)\) is bounded. In fact, by Lemma A.4.1 \(\{x_k\}\) belongs to the compact set
\[
\{x_k\} \subset f^{-1}(-\infty, f(x_0)] \cap g^{-1}(-\infty, g(x_0])
\]
and is bounded away from the boundary of the feasibility region. Therefore, due to continuity, there exist \(\lambda_+ \geq \lambda_- > 0\) such that \(\lambda_- I \leq \nabla^2 f(x_k) \leq \lambda_+ I\). In particular
\[
\|\nabla f(x_k)\| = \|\nabla^2 f(x_k)p_k\| \geq \lambda_- \|p_k\|
\]
and
\[
\|p_k\| \leq \lambda_-^{-1}\|\nabla f(x_k)\| \leq \lambda_-^{-1}\sqrt{g(x_0)} < \infty.
\]

Let us proceed by contradiction. Assume \(\|\nabla f(x_k)\|\) does not converge to zero. Then there exists \(\epsilon > 0\) such that \(\|\nabla f(x_k)\|^2 > \epsilon\) for infinitely many \(k\). Let \(G\) be the set of such indices. Rearranging \((A.25)\), we obtain
\[
g(x_k) - g(x_k + \alpha_k p_k) \geq -\eta_n \alpha_k \nabla g(x_k)^T p_k = \eta_n \alpha_k \|\nabla f(x_k)\|^2.
\]
The sequence \( \{g(x_k)\} \) is decreasing due to \( \text{(A.25)} \) and bounded below by zero thus converges. Hence, we must have \( \alpha_k \to 0 \) for \( k \in \mathcal{G} \) and, since \( \{p_k\} \) is bounded, \( \alpha_k p_k \to 0 \) also.

Consider the set of indices

\[
\mathcal{G} = \{ k : j_k > 0, k \in \mathcal{G} \},
\]

such set is infinite since \( \alpha_k \to 0 \). For each \( k \in \mathcal{G} \), the step length \( \sigma_k = \alpha_k/\gamma \) is such that either \( \text{(A.24)} \) or \( \text{(A.25)} \) is not satisfied for \( \alpha = \sigma_k \). Therefore at least one of the conditions will not hold an infinite number of times.

Suppose first that this is the case for \( \text{(A.25)} \), then

\[
g(x_k + \sigma_k p_k) > g(x_k) + \eta_n \sigma_k \nabla g(x_k)^T p_k
\]

for \( k \) in an infinite subset of \( \mathcal{G} \). Subtracting \( \sigma_k \nabla g(x_k)^T p_k \) on both sides and rearranging we obtain

\[
g(x_k + \sigma_k p_k) - g(x_k) - \sigma_k \nabla g(x_k)^T p_k > -\sigma_k (1 - \eta_n) \nabla g(x_k)^T p_k
\]

\[
= \sigma_k (1 - \eta_n) \| \nabla g(x_k) \|^2
\]

\[
> \sigma_k (1 - \eta_n) \varepsilon.
\]

By taking the Taylor series expansion of \( g(x_k + \sigma_k p_k) \), the left hand side becomes

\[
g(x_k + \sigma_k p_k) - g(x_k) - \sigma_k \nabla g(x_k)^T p_k = \sigma_k \int_0^1 [ \nabla g(x_k + \xi \sigma_k p_k) - \nabla g(x_k) ]^T p_k d\xi
\]

thus

\[
(1 - \eta_n) \varepsilon < \int_0^1 [ \nabla g(x_k + \xi \sigma_k p_k) - \nabla g(x_k) ]^T p_k d\xi
\]

\[
\leq \max_{0 \leq \xi \leq 1} \| \nabla g(x_k + \xi \sigma_k p_k) - \nabla g(x_k) \| \| p_k \|.
\]

Then, by continuity of \( \nabla g \), there exists \( \theta_k \in (0, \alpha_k/\gamma] \) such that

\[
\max_{0 \leq \xi \leq 1} \| \nabla g(x_k + \xi \sigma_k p_k) - \nabla g(x_k) \| = \| g(x_k + \theta_k p_k) - \nabla g(x_k) \|
\]

hence

\[
0 < (1 - \eta_n) \varepsilon < \| \nabla g(x_k + \theta_k p_k) - \nabla g(x_k) \| \| p_k \|.
\]

Finally, since \( \| \theta_k p_k \| \leq \| \alpha_k p_k \|, \theta_k p_k \to 0 \) and, considering the boundedness of \( p_k \), the desired contradiction is obtained as the right hand side of \( \text{(A.29)} \) tends to zero as \( k \) tends to infinity.

The case where \( \text{(A.24)} \) is not satisfied an infinite number of times is similar,

\[
f(x_k + \alpha p_k) - f(x_k) \sigma_k - \nabla f(x_k)^T p_k > -\sigma_k (1 - \eta_n) \nabla f(x_k)^T p_k
\]

\[
= \sigma_k (1 - \eta_n) \| \nabla f(x_k)^T p_k \|.
\]
Then we have
\[ |\nabla f(x_k)^T p_k| = |\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)| \geq \lambda_+^{-1} \|\nabla f(x_k)\|^2 \]
where \( \lambda_+ \) is defined above. Therefore
\[ f(x_k + \sigma_k p_k) - f(x_k) - \sigma_k \nabla f(x_k)^T p_k > \sigma_k (1 - \eta_0) \lambda_+^{-1} \epsilon > 0 \]
and the rest of the proof can proceed \textit{mutatis mutandis} as the previous case. \( \square \)

\section*{A.4.3 Convergence of first order methods}

The computation of the objective function’s Hessian is an increasingly demanding task as the dimension of the problem increases. Hence it makes sense to consider a method that involves derivatives of the first order only.

However, as it has already been noted previously, one of the proposed conditions of sufficient decrease (A.25) involves the Hessian. Thus a choice of the matrices \( B_k \) in (A.18) other than the Newton directions (A.27) is not enough to avoid the computation of the Hessian.

A sensible adaptation of (A.25) is simply the following
\[ g(x_k + \alpha p_k) \leq g(x_k). \] (A.30)

It turns out that, with this weaker condition, global convergence still holds provided that \( B_k \) is bounded above and away from the set of rank deficient positive semidefinite matrices.

\begin{theorem}
Let \( x_0 \) be any feasible point for (D.e, 1) and let \( \{x_k\} \) be generated by (A.18) and (A.19). Here \( \alpha_k = \gamma^j_k \), for \( \gamma \in (0, 1) \) and \( j_k \) being the smallest integer such that (A.24) and (A.30) are satisfied by \( \alpha_k \) for some \( \eta_k \in (0, 1) \). Suppose there exist \( \beta_-, \beta_+ \in \mathbb{R} \), such that
\[ 0 < \beta_- I \leq B_k \leq \beta_+ I \]
for all \( k \), then
\[ \lim_{k \to \infty} \|\nabla f(x_k)\| = 0. \]
\end{theorem}

\begin{proof}
The proof is very similar to the one of Theorem A.4.1. By Lemma A.4.1 \( \{x_k\} \) belongs to the compact set
\[ f^{-1}(-\infty, f(x_0)] \cap g^{-1}(-\infty, g(x_0)] \]
and is bounded away from the boundary of the feasibility region. Therefore, due to continuity, there exist \( \lambda_+ \geq \lambda_- > 0 \) such that \( \lambda_- I \leq \nabla^2 f(x_k) \leq \lambda_+ I \). Additionally, we have that
\[ \|\nabla f(x_k)\| = \|B_k p_k\| \geq \beta_- \|p_k\| \]
and

\[ \|p_k\| \leq \beta^{-1}_- \|\nabla f(x_k)\| \leq \beta^{-1}_- \sqrt{g(x_0)} < \infty. \]

Thus the sequence of descent directions \( \{p_k\} \) defined in (A.19) is bounded.

Let us proceed by contradiction. Assume \( \|\nabla f(x_k)\| \) does not converge to zero. Then there exists \( \epsilon > 0 \) such that \( \|\nabla f(x_k)\| \) does not converge to zero.

Then there exists \( \epsilon > 0 \) such that \( \|\nabla f(x_k)\|^2 > \epsilon \) for infinitely many \( k \). Let \( G \) be the set of such indices. Rearranging (A.24) we obtain

\[
\begin{align*}
    f(x_k) - f(x_k + \alpha_k p_k) &\geq -\eta_w \alpha_k \nabla f(x_k)^T p_k \\
    &= \eta_w \alpha_k \nabla f(x_k)^T B_k^{-1} \nabla f(x_k) \\
    &\geq \frac{\eta_w \alpha_k}{\lambda_+} \|\nabla f(x_k)\|^2.
\end{align*}
\]

The sequence \( \{f(x_k)\} \) is decreasing due to (A.24) and bounded below thus converges. Hence, we must have \( \alpha_k \to 0 \) for \( k \in G \) and, since \( \{p_k\} \) is bounded, \( \alpha_k p_k \to 0 \) also.

Consider the set of indices

\[ \mathcal{G} = \{k : j_k > 0, k \in G\}, \]

such set is infinite since \( \alpha_k \to 0 \). For each \( k \in G \), the step length \( \sigma_k = \alpha_k/\gamma \) is such that either (A.24) or (A.30) is not satisfied for \( \alpha = \sigma_k \). Therefore at least one of the conditions will not hold an infinite number of times.

The rest of the proof proceeds as an obvious adaptation of Theorem A.4.1 as once

\[
\begin{align*}
    f(x_k + \sigma_k p_k) - f(x_k) - \sigma_k \nabla f(x_k)^T p_k &> \sigma_k (1 - \eta_w) \|\nabla f(x_k)\|^2 \\
    &= \frac{\sigma_k (1 - \eta_w)}{\beta_+} \|\nabla f(x_k)\|^2
\end{align*}
\]

and

\[
\begin{align*}
    g(x_k + \sigma_k p_k) - g(x_k) - \sigma_k \nabla g(x_k)^T p_k &> \sigma_k \|\nabla g(x_k)\|^2 \\
    &= \frac{\sigma_k \lambda_-}{\beta_+} \|\nabla f(x_k)\|^2
\end{align*}
\]

are taken into account in case where (A.24) or (A.30) do not hold for an infinite number of times respectively.

\[ \square \]

A.4.4 About the choice of starting point

While Theorems A.4.1 and A.4.2 guarantee convergence in theory, severe numerical issues may be encountered in practice. In fact, while the sequence \( \{x_k\} \) is bounded
away from the boundary for each feasible problem instance, by varying the problem
parameters the corresponding optimal point can be arbitrarily close to the bound-
ary. This situation may result in extreme ill conditioning along the sequence \{x_k\}
and possibly inhibit convergence.

An appropriate choice for the starting point \(x_0\) can reduce numerical issues to
some extent. In fact, either (A.25) or (A.30) guarantees that the sequence \(\{x_k\}\)
is contained in the set

\[ g^{-1}(-\infty, g(x_0)] \]

which is bounded away from the boundary for each feasible \(x_0\). Suppose that \(\tilde{x}_0\)
such that \(g(\tilde{x}_0) < g(x_0)\) is chosen as a starting point instead. The resulting sequence \(\{\tilde{x}_k\}\) will be contained by

\[ g^{-1}(-\infty, g(\tilde{x}_0)] \subset g^{-1}(-\infty, g(x_0)] \]

thus, in the worst case scenario, \(\{\tilde{x}_k\}\) will be further away from the boundary than
\(\{x_k\}\).

Given an arbitrary starting point \(x_0\) – e.g such that \(\Lambda(x_0) = I\) – one could find
a strictly better starting point \(\tilde{x}_0\) in the sense described above by minimizing

\[ g(y^{-1}x_0) = \min_{y > 0} \| y \Gamma(\Psi Q_0^{-1}) - \Sigma \|^2, \tag{A.31} \]

where \(Q_0 = G^* \Lambda(x_0)G\), with respect to the scalar \(y\). The resulting solution \(\tilde{y}\) can
be easily obtained as

\[ \tilde{y} = \frac{\langle \Gamma(\Psi Q_0^{-1}), \Sigma \rangle}{\| \Gamma(\Psi Q_0^{-1}) \|^2}, \]

and the corresponding point \(\tilde{x}_0 = \tilde{y}^{-1}x_0\) will be such that either \(\tilde{x}_0 = x_0\) or
\(g(\tilde{x}_0) < g(x_0)\).

### A.5 Optimizing over the spectral factors

It has been noted – e.g in [5] for the scalar covariance extension problem – that
\(D\) is numerically badly conditioned when the solution has poles close to the unit
circle. In [5] a method has been proposed in order to attenuate these numerical
problems. Specifically the optimization is performed over the set of strictly sta-
ble spectral factors and is characterized by vastly better numerical properties. An
homotopy approach is then applied to overcome the nonconvexity of the reparam-
eterized problem and the resulting issues. In [1] such approach was applied to
matrix-valued Nevanlinna-Pick interpolation.

In the following discussion a similar, yet different approach will be developed
with respect to \(D\). The similarity lies in choosing to minimize over a set of spectral
factors of the spectrum \(G(z)^* \Lambda G(z)\). This, as in the scalar case, will lead to better
numerical conditioning. On the other hand, the main difference is in the approach
to handle the resulting nonconvexity.
First, consider the following equivalence relation regarding complex matrices. For hermitian matrices $\Lambda_1, \Lambda_2 \in \mathbb{C}^{n \times n}$, let $\Lambda_1 \sim \Lambda_2$ denote the equivalence

$$G(z)^*\Lambda_1G(z) = G(z)^*\Lambda_2G(z) \quad \forall z \in \mathbb{T}. \tag{A.32}$$

It can be characterized as per the following

**Proposition A.5.1** Consider an hermitian $\Lambda \in \mathbb{C}^{n \times n}$. Then

$$G^*(z)\Lambda G(z) = 0 \quad \forall z \in \mathbb{T}$$

if and only if $\Lambda \perp \mathcal{L}$.

**Proof.** Suppose $G^*(z)\Lambda G(z) = 0$ on the unit circle. Then for any function $F$ integrable on $\mathbb{T}$ we have

$$0 = \langle G^*\Lambda G, F \rangle = \langle \Lambda, GFG^* \rangle = \langle \Lambda, \Gamma(F) \rangle$$

thus $\Lambda \perp \mathcal{L}$. The converse follows similarly considering that for any $\Sigma \in \mathcal{L}$ there exist an integrable $F$ such that $\Sigma = \Gamma(F)$. \hfill $\square$

or, in other words,

**Corollary A.5.1** Consider hermitian matrices $\Lambda_1, \Lambda_2 \in \mathbb{C}^{n \times n}$. Then $\Lambda_1 \sim \Lambda_2$ if and only if $\Lambda_1 - \Lambda_2 \in \mathcal{L}^\perp$.

In particular, in view of (A.32) and Corollary A.5.1 if $\Lambda_1 \sim \Lambda_2$ then $\mathcal{J}_\Psi(\Lambda_1) = \mathcal{J}_\Psi(\Lambda_2)$, where $\mathcal{J}_\Psi$ is defined in (A.11).

Now, denote $\mathcal{Q}_+$ the subset of hermitian matrices such that the spectrum $G(z)^*\Lambda G(z)$ is positive definite on $\mathbb{T}$. Then the feasibility set of (D) can be written as $\mathcal{Q}_+ \cap \mathcal{L}$. For any feasible $\Lambda$ it will be shown that there exist a $C \in \mathbb{C}^{m \times n}$ such that

$$C^*C \sim \Lambda. \tag{A.33}$$

Then, in view of the properties of $\sim$, it follows that the dual problem (D) can indeed be parameterized in terms of the factors (A.33). The nonnegativity constraint becomes redundant since

$$G^*(z)C^*CG(z) \geq 0$$

for any $C$ and $z \in \mathbb{T}$. Therefore the optimization problem (D) simplify into

$$\min_{C \in \mathcal{C}} \mathcal{J}_\Psi(C^*C) \tag{A.34}$$

where $\mathcal{C}$ is a suitable set of factors (A.33).

We turn now to show that the factorization (A.33) indeed exists for every $\Lambda \in \mathcal{Q}_+ \cap \mathcal{L}$ and to study the resulting optimization problem.
A.5.1 Existence and characterization

Let \( \Lambda \) such that the spectrum \( G(z)^* \Lambda G(z) \) is positive definite on \( \mathbb{T} \). The following equality holds

\[
G^* \Lambda G = [G^* \mid I] \begin{bmatrix} A + A^* X A - X & A^* X B \\ B^* X A & B^* X B \end{bmatrix} [G \mid I]
\]

(A.35)

for any \( X \in \mathbb{C}^{n \times n} \).

Consider the algebraic Riccati equation

\[
X = A^* X A - A^* X B (B^* X B)^{-1} B^* X A + \Lambda.
\]

(A.36)

For any \( P \in \mathbb{C}^{n \times n} \) such that \( B^* P B \) is invertible and is a solution of (A.36), it follows that

\[
\begin{bmatrix} A + A^* P A - P & A^* P B \\ B^* P A & B^* P B \end{bmatrix} = \begin{bmatrix} A^* P B \\ B^* P B \end{bmatrix} (B^* P B)^{-1} [B^* P A \mid B^* P B]
\]

(A.37)

and thus

\[
G^* \Lambda G = \begin{bmatrix} (G^* A^* P B + B^* P B)(B^* P B)^{-1} (B^* P A G + B^* P B) \\ (AG + B)^* P B (B^* P B)^{-1} B^* P (AG + B) \\ G^* P B (B^* P B)^{-1} B^* P G. \end{bmatrix}
\]

Let \( L \in \mathbb{C}^{m \times m} \) be any matrix such that \( B^* P B = LL^* \). Then \( L \) must be invertible as \( B^* P B \) is and

\[
(B^* P B)^{-1} = (LL^*)^{-1} = (L^{-1})^* L^{-1}.
\]

Thus the factorization (A.33) holds for \( C \in \mathbb{C}^{m \times n} \) such that

\[
L C = B^* P.
\]

(A.38)

Clearly, (A.38) is not the only factor satisfying (A.33). In fact, both the solution of equation (A.36) and the decomposition \( B^* P B = LL^* \) are not unique. Among the many different factorizations (A.33), consider the ones for which \( P \) is a stabilizing solution of (A.36) – i.e such that all the eigenvalues of

\[
A - B(B^* P B)^{-1} B^* P A
\]

(A.39)

lies inside the open unit disc. Such solution, if it exists, is unique – cif. [11] – and its main advantage is that, as it will be shown in Proposition A.5.2, it leads to a minimum-phase spectral factor of \( G^*(z) \Lambda G(z) \).

In this context the existence of the stabilizing solution of (A.36) is guaranteed by the following theorem whose proof is reported in the appendix.
Theorem A.5.1 Let \( G(z) = (zI - A)^{-1}B \) with \( B \) having full column rank and \((A, B)\) being a reachable pair. For any \( \Lambda \in \mathcal{Q}_+ \) exists \( P \in \mathbb{C}^{n \times n} \) such that \( B^*PB > 0 \) and is the unique stabilizing solution of the algebraic Riccati equation
\[
X = A^*X A - A^*X B (B^*X B)^{-1} B^*X A + \Lambda. \tag{A.40}
\]
\( \square \)

The resulting factorizations can be characterized in the following way. First, note that, by multiplying to the left \( A.38 \) by \( B \) we obtain that
\[
LCB = B^*PB = LL^*.
\]
As \( L \) is invertible, it follows that \( CB = L^* \) and thus also the matrix \( CB \) must be invertible. Consequently, in view of \( A.38 \) we have that
\[
B^*P = LC = B^*C^*C
\]
and
\[
A - B(B^*P B)^\dagger B^*PA = A - B(B^*C^*CB)^{-1}B^*C^*CA = A - B(CB)^{-1}CA \tag{A.41}
\]
that, since \( P \) is a stabilizing solution, must have all eigenvalues inside the unit disc.

Consider now the set \( C \subset \mathbb{C}^{m \times n} \) of all matrices satisfying these two conditions, namely such that

- \( CB \) is invertible,
- \( A - B(CB)^{-1}CA \) has eigenvalues inside the open unit disc.

The following can be shown

**Proposition A.5.2** Let \( C \in \mathcal{C} \). Then the transfer function
\[
W(z) = zCG(z) = CA(zI - A)^{-1}B + CB
\]
is minimum phase – i.e it is stable and admits a stable inverse.

**Proof.** Since \( CB \) is invertible \( W(z) \) is invertible too and its inverse has the following realization
\[
W^{-1}(z) = (CB)^{-1}CA \left[ zI - (A - B(CB)^{-1}CA) \right]^{-1} B(CB)^{-1} + (CB)^{-1}
\]
which is stable since \( A - B(CB)^{-1}CA \) has eigenvalues inside the unit disc.

Clearly, as shown above, for any \( \Lambda \in \mathcal{Q}_+ \cap \mathcal{L} \) there exist a \( C \in \mathcal{C} \) such that \( A.33 \) holds. It turns out that also the converse holds. In fact for any \( C \in \mathcal{C} \), by Proposition A.5.2 follows that \( C^*C \in \mathcal{Q}_+ \). Then, let \( \Lambda \in \mathcal{L} \) be the projection of \( C^*C \) onto \( \mathcal{L} \); by Corollary A.5.1 \( \Lambda \sim C^*C \) and thus \( \Lambda \in \mathcal{Q}_+ \) also.
A.5.2 Optimization’s properties

Consider now the optimization \( D(C) \) parameterized by \( C \in \mathcal{C} \)
\[
\left[ \min_{C \in \mathcal{C}} J_{\Psi}(C^*C) \right].
\]

Let \( J_{\Psi}^C : \mathcal{C} \to \mathbb{R} \) denote the objective function of \( D(C) \) – i.e
\[
J_{\Psi}^C(C) = J_{\Psi}(C^*C).
\]

In view of the above discussion it follows that \( D(C) \) is equivalent to \( D \) – i.e if \( \hat{C} \) is a global optimum of \( D(C) \) then the unique global optimum \( \hat{\Lambda} \) of \( D \) is such that \( \hat{\Lambda} \sim \hat{C}^* \hat{C} \) and can be retrieved by projecting \( \hat{C}^* \hat{C} \) onto \( \mathcal{L} \).

Such problems share some common properties. In particular, the sublevel sets of \( D(C) \) are still compact. In order to show this we need first to prove the following

**Lemma A.5.1** Let \( G(z) = (zI - A)^{-1}B \) with \( B \) having full column rank and \((A, B)\) being a reachable pair. Then there exist \( \epsilon > 0 \) such that
\[
\|\Lambda\| \geq \epsilon \|C\|^2
\]
for any \( \Lambda \) and any non zero \( C \) such that \((A.33)\) holds.

**Proof.** By \((A.2)\) we have
\[
\|G^*\Lambda G\|^2 = \int_T \|G^*(z)\Lambda G(z)\|^2 \frac{dz}{z} \geq \int_T \|G^*(z)\Lambda G(z)\|^2_2 \frac{dz}{z} = \int_T \|CG(z)\|^4_2 \frac{dz}{z}.
\]

Then, by Jensen inequality
\[
\|G^*\Lambda G\|^2 \geq \int_T \|CG(z)\|^4_2 \frac{dz}{z} \geq \left[ \int_T \|CG(z)\|^2 \frac{dz}{z} \right]^2.
\]

By taking square roots of both sides and applying \((A.2)\) again we obtain
\[
\|G^*\Lambda G\| \geq \int_T \|CG(z)\|^2 \frac{dz}{z} \geq \int_T \frac{1}{r(z)} \|CG(z)\|^2 \frac{dz}{z} \geq \frac{1}{m} \|CG\|^2
\]
where \( r(z) \) is the rank of \( CG(z) \) which is positive, since \( C \neq 0 \) and \((A,B)\) is reachable. Additionally,

\[
\|CG\|^2 = \int T \text{tr}\{CG(z)G^*(z)C^*\}\frac{dz}{z} \\
= \text{tr}\left\{C \left( \int T G(z)G^*(z)\frac{dz}{z} \right)C^*\right\} \\
= \text{tr}\{CP_C C^*\} \\
\geq \lambda_{\min}(P_G)\|C\|^2
\]

where \( \lambda_{\min}(P_G) \) is the minimum eigenvalue of \( P_G \) which is positive since \((A,B)\) is reachable.

The lemma follows from the fact that the linear operator \( Q : \mathcal{L} \rightarrow L_2[-\pi,\pi] \) defined as \( Q(\Lambda) = G^*(z)\Lambda G(z) \) is bounded.

\[\square\]

**Proposition A.5.3** Suppose the primal optimization problem \( \mathcal{P} \) is feasible, then the dual problem \( \mathcal{D}_C \) has compact sublevel set.

**Proof.** Straightforward consequence of Lemma A.5.1 and Proposition A.3.1. \(\square\)

Consider the differential of the objective function \(A.42\). In view of \(A.12\), it can be evaluated as

\[
\delta J_C^{\Psi}(C; \delta C) = 2\Re \left\langle \Sigma - \Gamma(\Psi Q^{-1}), C^* \delta C \right\rangle \tag{A.44}
\]

where \( \Re \) denotes the real part. Hence a stationary point for \(A.42\) is such that

\[
CT(\Psi Q^{-1}) = C\Sigma. \tag{A.45}
\]

It turns out that, despite not being a convex problem, stationarity implies global optimality for \( \mathcal{D}_C \). In order to show it we need the following

**Lemma A.5.2** Let \( C \in \mathcal{C} \), then for all \( \Lambda \in \mathcal{L} \) such that \( CA = 0 \) it follows that \( \Lambda = 0 \).

**Proof.** Let \( \Lambda \in \mathcal{L} \) be such that \( CA = 0 \) and consider the following

\[
\begin{align*}
\Lambda &= (A - B(CB)^{-1}CA)\Lambda(A - B(CB)^{-1}CA)^* = \\
&= \Lambda - \Lambda\Lambda^* + \\
&+ B(CB)^{-1}C\Lambda\Lambda^* + \Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* + \\
&- B(CB)^{-1}C\Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* \tag{A.46}
\end{align*}
\]

\[
\begin{align*}
&\quad + B(CB)^{-1}C\Lambda\Lambda^* + \Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* + \\
&- B(CB)^{-1}C\Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* \tag{A.47}
\end{align*}
\]

\[
\begin{align*}
&\quad + B(CB)^{-1}C\Lambda\Lambda^* + \Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* + \\
&- B(CB)^{-1}C\Lambda\Lambda^* C^*(B^*C^*)^{-1}B^* \tag{A.48}
\end{align*}
\]
Then, since $\Lambda$ belongs to $L$, there exist $H \in \mathbb{C}^{m \times n}$ such that

$$\Lambda - A\Lambda A^* = BH + H^* B^*.$$  

In particular, the terms $\text{(A.46)}$, $\text{(A.47)}$ and $\text{(A.48)}$ become

$$\text{(A.46)} = BH + H^* B^*,$$

$$\text{(A.47)} = B(CB)^{-1}CA + \Lambda C^*(B^*C^*)^{-1} B^* + B(CB)^{-1}C(BH + H^* B^*) - (BH + H^* B^*)C^*(B^*C^*)^{-1} B^* = -BH - B(CB)^{-1}CH^* B^* - BHC^*(B^*C^*)^{-1} B^* - H^* B^*$$

and

$$\text{(A.48)} = -B(CB)^{-1}CA^* C^*(B^*C^*)^{-1} B^* + B(CB)^{-1}C(BH + H^* B^*)C^*(B^*C^*)^{-1} B^* = BHC^*(B^*C^*)^{-1} B^* + B(CB)^{-1}CH^* B^*$$

respectively. Therefore we obtain the following Lyapunov equation

$$\Lambda - (A - B(CB)^{-1}CA) \Lambda (A - B(CB)^{-1}CA)^* = 0$$

and since $A - B(CB)^{-1}CA$ has no eigenvalues on the unit circle its unique solution is $\Lambda = 0$.

**Theorem A.5.2** Let $\hat{C} \in C$ be a stationary point for $\text{(A.42)}$ then $\hat{C}$ is a global optimum of $\mathcal{D}_C$.

**Proof.** By Lemma A.5.2 equation $\text{(A.45)}$ implies that $\Gamma(\Psi Q^{-1}) = \Sigma$ for $Q = G^* C^* CG$. Since there exist a $\Lambda \in Q_+ \cap L$ such that $\Lambda \sim \hat{C}^* \hat{C}$ then $\Lambda$ must be the unique optimum of $\mathcal{D}$ and the claim follows.

**A.5.3** Real parameterization

Next we turn to write $\mathcal{D}_C$ as an optimization over the real variable $y \in \mathbb{R}^N$ in a similar way as we did with $\mathcal{D}_x$. To this end, let $C_1, C_2, ..., C_N$ where $N = 2nm$ be a real basis of $C \in \mathbb{C}^{m \times n}$. Denote

$$C(y) = y_1 C_1 + y_2 C_2 + ... y_N C_N$$  \hspace{1cm} (A.49)

the linear combination of such basis with parameters given by the components of $y \in \mathbb{R}^N$. Now, consider the nonlinear map $q : \mathbb{R}^N \rightarrow \mathbb{R}^M$ such that

$$q_i(y) = \langle \Lambda_i, C^*(y) C(y) \rangle$$  \hspace{1cm} (A.50)
where \( \{ \Lambda_i \} \) is the basis of \( L \) considered in (A.14). Then \( D_C \) can be parameterized in the following way

\[
\begin{bmatrix}
\max_{y \in \mathbb{R}^N} f \circ q(y) \\
\text{s.t.}
\end{bmatrix}
\begin{align*}
B^*C^*(y)C(y)B &> 0, \\
A - B[C(y)B]^{-1}C(y)A &\text{ stable}
\end{align*}
\]  

(\( D_y \))

where \( f \) was defined in (A.15) and the constraints are such that \( C(y) \in C \). Clearly the feasibility set of \( D_y \) is in a one-to-one relation with \( C \) via (A.49) hence \( D_y \) is equivalent to \( D_C \) and thus to \( D \).

Let \( J_q(y) \in \mathbb{R}^{M \times N} \) denote the Jacobian of the map \( g \) at \( y \in \mathbb{R}^N \) – i.e

\[
[J_q(y)]_{i,j} = \frac{\partial q_i}{\partial y_j}(y) = \langle \Lambda_i, C^*(y)C_j \rangle + \langle \Lambda_i, C_j^*C(y) \rangle = 2\Re \langle \Lambda_i, C^*(y)C_j \rangle . \tag{A.51}
\]

Then the gradient of the objective function of \( D_y \) can be evaluated by

\[
\nabla (f \circ q)(y) = J_q^T(y)(\nabla f) \circ q(y).
\]

Specifically, in view of (A.12) and (A.51), it can be shown that

\[
\frac{\partial}{\partial y_j} (f \circ q)(y) = 2\Re \left( \Sigma - \Gamma(\Psi Q^{-1})^* \right), C^*(y)C_j \right) \tag{A.52}
\]

where \( Q(z) = G^*(z)C^*(y)C(z)G(z) \). In fact we have

\[
\frac{\partial}{\partial y_j} (f \circ q)(y) = \sum_{i=1}^N \frac{\partial f}{\partial x_i} \frac{\partial q_i}{\partial y_j}
\]

\[
= 2 \sum_{i=1}^N \langle \Sigma - \Gamma(\Psi Q^{-1})^* \Lambda_i \rangle \Re \langle \Lambda_i, C^*(y)C_j \rangle
\]

\[
= 2 \left\langle \left\langle \Sigma - \Gamma(\Psi Q^{-1})^* \right\rangle, C^*(y)C_j \right\rangle 
\]

where the last equality holds since

\[
\sum_{i=1}^N \Re \langle \Lambda_i, C^*(y)C_j \rangle \Lambda_i
\]

is the projection of \( C^*(y)C_j \) onto \( L \) and \( \Sigma - \Gamma(\Psi Q^{-1}) \in L \). The following property of the Jacobian is essential for convergence of descent algorithms applied to \( D_y \).
Proposition A.5.4 For any $y \in \mathbb{R}^N$ such that $C(y) \in C$ the Jacobian $J_q(y)$ defined in (A.51) has full row rank.

Proof. For any $x \in \mathbb{R}^M$ such that $x^T J_q(y) = 0$ we have that

$$0 = \sum_{j=1}^{M} x_j 2 \Re \langle \Lambda(x), C^*(y) C_j \rangle = 2 \Re \langle C(y) \Lambda(x), C_j \rangle$$

for any $j = 1, \ldots, N$. Therefore $C(y) \Lambda(x) = 0$ and, by Lemma A.5.2, $\Lambda(x) = 0$ hence $x = 0$. \qed

Before turning to study convergence, consider the hessian of the objective function. It can be evaluated similarly as its gradient starting from (A.12), (A.13) and amounts to

$$\frac{\partial^2}{\partial y_i \partial y_j} (f \circ q)(y) = \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{\partial^2 f}{\partial x_k \partial y_i} \frac{\partial q_k}{\partial y_l} \frac{\partial q_l}{\partial y_j} + \sum_{k=1}^{N} \frac{\partial f}{\partial x_k} \frac{\partial^2 q_k}{\partial y_i \partial y_j} + 4 \Re \langle \Psi, (CG)^{-1} C_i G(CG)^{-1} C_j G \rangle + 2 \Re \langle \Sigma - \Gamma \Psi Q^{-1}, C^*_i C_j \rangle. \tag{A.53}$$

Now we can compare with (A.13) by substituting $Q(z) = G^*(z) C^*(y) C(y) G(z)$. It can be noted that (A.13) contain a quartic term of $(CG)^{-1}$ while (A.53) contains only quadratic terms. Hence, it is to be expected a smaller growth of the condition number of (A.53) as $CG$ becomes rank deficient.

A.5.4 Convergence results

Now we turn to study the convergence properties of descent methods when applied to $\mathcal{P}_y$. As before we consider a sequence of feasible points $\{y_k\}$ such that

$$y_{k+1} = y_k + \alpha_k p_k \tag{A.54}$$

where the descent direction $p_k$ satisfies

$$B_k p_k = -\nabla (f \circ q)(y_k) \tag{A.55}$$

for a suitable sequence of positive definite matrices $\{B_k\}$. Similarly as before, the step length $\alpha_k$ will be chosen to satisfy the following conditions

$$f \circ q(y_k + \alpha p_k) \leq f \circ q(y_k) + \eta_k \alpha \nabla (f \circ q)(y_k)^T p_k, \tag{A.56}$$

$$g \circ q(y_k + \alpha p_k) \leq g \circ q(y_k) \tag{A.57}$$

where $g$ was defined in (A.22). The existence of an $\alpha$ such that (A.56) and (A.57) are satisfied is guaranteed by the following
Proposition A.5.5 Let $p_k$ be such that (A.55) is satisfied, then $p_k$ is also a descent direction for $g \circ q$ at $y_k$.

Proof. In view of (A.55) we have
\[0 \geq \nabla (f \circ q)^T(y_k)p_k = (\nabla f) \circ g^T(y_k)J_q(y_k)p_k\]
thus $J_q(y_k)p_k$ is a descent direction of $f$ at $q(y_k)$. It follows, by Proposition A.4.1, that $p_k$ is also a descent direction of $g$ at $q(y_k)$. Hence
\[0 \geq (\nabla g) \circ g^T(y_k)J_q(y_k)p_k = \nabla (g \circ q)^T(y_k)p_k.\]

□

The conditions (A.56) and (A.57) make sure that \(\{y_k\}\) does not approach the boundary as per the following

Lemma A.5.3 For any feasible $y_0$, the set
\[\{y : C(y) \in C, \ f \circ q(y) \leq f \circ q(y_0), \ g \circ q(y) \leq g \circ q(y_0)\}\]
is compact and contained in the interior of the feasibility region of $(D_y)$.

Proof. Follows directly by Lemmas A.4.1 and A.5.1

Finally convergence can be proven.

Theorem A.5.3 Let $y_0$ be any feasible point for $(D_y)$ and let $\{y_k\}$ generated by (A.54) and (A.55). Here $\alpha_k = \gamma^{j_k}$, for $\gamma \in (0, 1)$ and $j_k$ is the smallest integer such that (A.56) and (A.57) are satisfied by $\alpha_k$ for some $\eta_s \in (0, 1)$. Suppose there exist $\beta_-, \beta_+ \in \mathbb{R}$, such that
\[0 < \beta_- I \leq B_k \leq \beta_+ I\]
for all $k$, then
\[\lim_{k \to \infty} \|\nabla (f \circ g)(x_k)\| = 0.\]

Proof. By lemma A.5.3 the sequence $\{y_k\}$ belong to a compact set bounded away from the boundary of the feasibility region. Therefore, due to continuity, we have that
\[\infty > \|\nabla (f \circ q)(y_k)\| = \|B_k p_k\| \geq \beta_- \|p_k\|\]
thus the sequence of descent directions $\{p_k\}$ is bounded. The rest of the proof proceeds mutatis mutandis as Theorem A.4.1 and A.4.2 considering that
\[0 < \beta_- J_q \leq J_q B_k J_q^T \leq \beta_+ J_q J_q^T.\]
and, by Proposition A.5.4, $J_q J_q^T$ is positive definite. Hence, from the continuity of $J_q$, follows that there exists $\beta_-, \beta_+$ such that

$$0 < \beta_- \leq J_q B_k J_q^T \leq \beta_+$$

Now that convergence has been proven with respect to the formulation (D_y) it remains to verify that its favorable properties are indeed reflected in an increase of efficiency. In order to do so we compare the descent algorithms for (D_x) and (D_y) with a quasi-newton (BFGS) search direction.

The simulations were conducted with the following parameters. Given a fixed $\Psi$, for $m = 5$ and $n = 10, 15, \ldots, 40$, the filter matrices $A, B$ were set so that (P) is equivalent to a covariance selection problem of compatible dimensions. On the other hand, for each $n$ the interpolation value $\Sigma$ was selected randomly in $L$ and then projected into $L_+$ by the addition of the identity times a small positive constant.

The results are depicted in Figure A.1 and represent the number of function evaluations necessary for the descent algorithm to bring the interpolation error $\|\Sigma - \Gamma(\psi Q^{-1})\|_2$ to $10^{-5}$. Here the dark and light gray correspond to formulations (D_x) and (D_y) respectively. It can be noted that the formulation (D_y) requires approximately 80% less function evaluation with respect to (D_x).

A.5.5 An homeomorphic factorization

The formulation (D_y) greatly improves numerical properties with respect to (D_x) while preserving global convergence. On the other hand, the optimization variables
dimension was increased by $m(m - 1)/2$ potentially degrading the algorithmic scalability with respect to $m$. This increase is due to the factorization (A.33) not being surjective with respect to the set $C$. It is therefore of practical interest to identify a set of factors such that the factorization (A.33) is a bijective map.

In order to do so, let $L$ be the Cholesky factor of $B^*PB$ — i.e a lower triangular matrix with positive entries on the diagonal such that $B^*PB = LL^*$. Since $B^*PB > 0$ the Cholesky factor exists and is unique. Such choice of $L$ let the system of equation (A.38) to be solved efficiently with respect to $C$ by forward-substitution methods.

Denote $H(A)$ the factorization (A.33) uniquely determined by the stabilizing solution $P$ of (A.36) and the Cholesky factor $L$. It is defined over $Q_+$ and its range $H(Q_+)$ can be characterized in the following way. Consider the set $C_+ \subset C$ of all factors (A.33) satisfying these two conditions, namely such that

- $CB$ upper triangular with positive diagonal entries,
- $A - B(CB)^{-1}CA$ has eigenvalues inside the open unit disc.

Clearly $C_+$ contains the range of $H$. It turns out that the two sets coincide and the factorization $H$ is surjective with respect to $C_+$. To show this we need the following

**Proposition A.5.6** Suppose $CB$ is invertible, then $C^*C$ is a solution of the discrete Riccati equation

$$
X = A^*XA - A^*XB(B^*XB)^{-1}B^*XA + C^*C. \tag{A.58}
$$

If, in addition the matrix

$$
A - B(CB)^{-1}CA \tag{A.59}
$$

has all eigenvalues inside the open unit disc, then $C^*C$ is the unique stabilizing solution.

**Proof.** We have

$$
A^*C^*CA - C^*C - A^*C^*CB(B^*CB)^{-1}B^*C^*CA + C^*C = A^*C^*CA - A^*C^*CB(CB)^{-1}(B^*C^*)^{-1}B^*C^*CA = 0.
$$

The second part of the proposition follows immediately since (A.41) still holds when $P = C^*C$.

Now, for any $C \in C_+$ consider $H(C^*C)$. Following the factorization procedure introduced previously we have that, by Proposition A.5.6 $P = C^*C$ and the Cholesky decomposition

$$
LL^* = B^*PB = B^*C^*CB
$$
gives \( L = B^*C^* \). Finally, by (A.38),

\[
B^*C^* H(C^*C) = LH(C^*C) = B^*P = B^*C^*C
\]

and \( H(C^*C) = C \). Hence \( H \) is indeed surjective onto \( C_+ \).

Clearly \( H \) defined over \( Q_+ \) is not injective. In fact consider some positive definite \( \Lambda \), hence \( \Lambda \in Q_+ \). Let \( C = H(\Lambda) \). On the other hand we have that \( C = H(C^*C) \) also and, since \( \Lambda \) is full rank while \( C^*C \) has rank \( m \) they can not be equal.

For injectivity to hold, we need to restrict the domain of \( H \) to include a single \( \Lambda \) for each equivalence class of \( \sim \). To this end, we need the following

**Lemma A.5.4** Let \( C, \hat{C} \in C_+ \) such that

\[
C^*C \sim \hat{C}^*\hat{C}, \tag{A.60}
\]

then \( C = \hat{C} \).

**Proof.**

Consider the observable canonical forms

\[
\begin{pmatrix}
A \\
C
\end{pmatrix}
\begin{pmatrix}
B \\
0
\end{pmatrix} \xrightarrow{T} \begin{pmatrix}
A_{11} & 0 & B_1 \\
A_{12} & A_{22} & B_2 \\
C_1 & 0 & 0
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
A \\
C
\end{pmatrix}
\begin{pmatrix}
B \\
0
\end{pmatrix} \xrightarrow{\tilde{T}} \begin{pmatrix}
\hat{A}_{11} & 0 & \hat{B}_1 \\
\hat{A}_{12} & \hat{A}_{22} & \hat{B}_2 \\
\hat{C}_1 & 0 & 0
\end{pmatrix}
\]

for \( CG(z) \) and \( \hat{C}G(z) \) respectively obtained by the orthogonal state transformations \( T, \hat{T} \). In particular \( CB = C_1B_1 \) and \( \hat{C}B = \hat{C}_1\hat{B}_1 \). It follows that \( W(z) = C_1(zI - A_{11})^{-1}B_1 \) and \( \hat{W}(z) = \hat{C}_1(zI - \hat{A}_{11})^{-1}\hat{B}_1 \) are both minimal minimum phase spectral factor of the spectrum \( G^*(z)C^*CG(z) \). Thus we have

\[
\begin{pmatrix}
A_{11} \\
C_1
\end{pmatrix}
\begin{pmatrix}
B_1 \\
0
\end{pmatrix} \xrightarrow{T} \begin{pmatrix}
\hat{A}_{11} \\
Y\hat{C}_1
\end{pmatrix}
\]

for some orthogonal matrices \( Y \) and \( \tilde{T} \). In view of this, we can assume without loss of generality that \( \tilde{T} = I \) hence \( A_{11} = \hat{A}_{11}, B_1 = \hat{B}_1 \) and \( C_1 = Y\hat{C}_1 \). Additionally, \( T \) and \( \tilde{T} \) must have the form

\[
T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}, \quad \tilde{T} = \begin{bmatrix} T_1 & \tilde{T}_2 \end{bmatrix}
\]

since the columns relative to the observable subspace are equal.

Now, the following holds

\[
B^*C^*CB = B_1^*C_1^*C_1B_1 = B_1^*\hat{C}_1^*\hat{C}_1B_1 = B^*\hat{C}^*\hat{C}B
\]
and since both $B^*_1C^*_1$ and $B^*_1\hat{C}^*_1$ are Cholesky factors of the same positive definite matrix they must be equal. Hence $CB = \hat{C}B$ and $C_1 = \hat{C}_1$. Furthermore we have

\[ CA = \begin{bmatrix} C_1 & 0 \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} T_1^* \\ T_2^* \end{bmatrix} = C_1A_{11}T_1^* = \begin{bmatrix} C_1 & 0 \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ A_{12} & \hat{A}_{22} \end{bmatrix} \begin{bmatrix} T_1^* \\ \hat{T}_2^* \end{bmatrix} = \hat{C}A. \]

Finally, since $CB = \hat{C}B$ and $CA = \hat{C}A$ it turns out that $CG(z) = \hat{C}G(z)$ and the Proposition follows from the controllability of $(A, B)$. □

Now bijectivity can be proven

**Theorem A.5.4** The factorization $H$ described above is a one-to-one map between $Q_+ \cap L$ and $C_+$.

**Proof.** If $H(\Lambda) = H(\hat{\Lambda}) = C$ for any $\Lambda, \hat{\Lambda} \in Q_+ \cap L$ then

\[ \Lambda \sim C^*C \sim \hat{\Lambda} \]

thus $\Lambda - \hat{\Lambda} \perp L$. Since both belong to $L$ then $\Lambda = \hat{\Lambda}$ and injectivity follows.

On the other hand for any $C \in C_+$ we have shown above that $C = H(C^*C)$. Consider now the decomposition

\[ C^*C = \Lambda + \Lambda_\perp \]

where $\Lambda \in L$ and $\Lambda_\perp \perp L$. Then

\[ C^*C \sim \Lambda \sim \hat{C}^*\hat{C} \]

where $\hat{C} = H(\Lambda)$. By Lemma A.5.4 $C = \hat{C} = H(\Lambda)$ and since $\Lambda \in Q_+ \cap L$ surjectivity follows.

Now that a bijective factorization has been established it remains to investigate how it interacts with our original problem $(D)$. To this end, consider the restriction of $(D_{C})$ to $C_+,$

\[ \left[ \min_{C \in C_+} \| \Psi(C^*C) \right]. \quad (D_{C_+}) \]

Unfortunately, at this point, a result for similar to Theorem A.5.2 is not know to hold with respect to $(D_{C_+})$. Thus, in general, stationarity in $C_+$ may not imply optimality. Nonetheless a weaker version can be proven, namely that any local optimum is also global. To this end, we need to show that $H : Q_+ \cap L \to C_+$ is continuous. In fact, more can be proven
**Theorem A.5.5** The factorization $H$ described above is a homeomorphism between $Q_+ \cap \mathcal{L}$ and $C_+$.

**Proof.** Consider the inverse map $H^{-1} : C_+ \to Q_+ \cap \mathcal{L}$. In order to prove the theorem we are going to apply Theorem 2.6 in [2] to $H^{-1}$. Specifically we need to show continuity, injectivity and properness. In addition we have to make sure that $C_+$ and $Q_+ \cap \mathcal{L}$ has the same real dimension.

First note that $H^{-1}(C)$ is equal to the projection of $C^*C$ onto $\mathcal{L}$. Being a composition of two continuous operations $H^{-1}$ is continuous too. Furthermore, from Theorem A.5.4 it follows that $H^{-1}$ is injective.

Now consider the dimensions of $Q_+ \cap \mathcal{L}$ and $C_+$. The former is an open cone in $\mathcal{L}$ and thus has the same real dimension, namely $m(2n - m)$. Consider now the linear space of $C \in \mathbb{C}^{n \times m}$ such that $CB$ is upper triangular with real diagonal entries. The real dimension of such space is

$$2nm - m(m - 1) - m = m(2n - m)$$

since the triangular and the real diagonal conditions impose $m(m - 1)$ and $m$ real linear constraints respectively. Since $C$ is an open subset of such a space it has the same real dimension also equal to $m(2n - m)$.

It remains to show that $H^{-1}$ is proper – i.e for any compact $K \in Q_+ \cap \mathcal{L}$ $H(K)$ is compact also. Let $\{C_k\}$ be any converging sequence inside of $H(K)$ and $\Lambda_k = H^{-1}(C_k)$. Since $J_\Psi(C_k^*C_k) = J_\Psi(\Lambda_k)$, by continuity of $J_\Psi$ we have that

$$\|H(\Lambda) - \hat{C}\| \leq \max_{\Lambda \in K} J_\Psi(\Lambda) < \infty.$$ 

Thus, in view of Proposition A.5.3 the sequence $\{C_k\}$ is contained in a compact subset of the open set $C_+$. It follows that $H(K)$ is compact and thus $H^{-1}$ is proper. □

Now it can be shown the link between local and global optima for $\{D_{C_+}\}$.

**Proposition A.5.7** Suppose $\hat{C} \in C_+$ is a local minimum of $\{D_{C_+}\}$. Then $\hat{\Lambda} = H^{-1}(\hat{C})$ is a global optimum for $\{D\}$.

**Proof.** Since $\hat{C}$ is a local optimum there exist $\delta > 0$ such that for all $C$, $\|C - \hat{C}\| \leq \delta$ we have

$$J_\Psi(\hat{C}^*\hat{C}) \leq J_\Psi(C^*C).$$

On the other hand, by continuity of $H$, there exist $\epsilon > 0$ such that all $\Lambda$ such that $\|\Lambda - \hat{\Lambda}\| \leq \epsilon$ it holds $\|H(\Lambda) - \hat{C}\| \leq \delta$. Finally, for all such $\Lambda$

$$J_\Psi(\Lambda) = J_\Psi(H(\Lambda)^*H(\Lambda)) \geq J_\Psi(\hat{C}^*\hat{C}) = J_\Psi(\hat{\Lambda})$$

and $\hat{\Lambda}$ is a local, hence global, optimum of $\{D\}$. □

Note that Proposition A.5.7 together with the bijectivity of $H$ implies that any local minimum $\hat{C}$ is also global for $\{D_{C_+}\}$. 

37
A.6 Conclusions

In this paper we considered the application of descent methods in order to solve the generalized moment problem. Since classic convergence proofs can not be applied to the problem at hand, an alternative line search method was introduced. This allowed us to prove global convergence.

Additionally, the numerical issues occurring with solutions close to the boundary of the feasibility region were addressed. To overcome such difficulties, a formulation based on spectral factorization was proposed. Such an approach is not new and was applied, together with homotopy continuation, in [5] and [1] for covariance extension and Nevanlinna-Pick interpolation respectively. The novel result is that the homotopy continuation is not necessary for global convergence. Hence only one optimization problem needs to be solved rather than a sequence of problems.

The only apparent drawback of the proposed approach is an increase of the optimization variables that depend on \( m \). This issue can be solved by an homeomorphic factorization but, in this case, the link between stationarity and minimality remains to be proven. This will be considered in future works.

Appendix

Generalized eigenvalues problem

For some \( M, N \in \mathbb{C}^{n \times n} \), the first order matrix polynomial \( \lambda M - N \) is denoted matrix pencil. A matrix pencil is said to be regular if \( \det(\lambda M - N) \) is not identically zero.

Given a regular matrix pencil, any \( \lambda \in \mathbb{C} \) that is a root of the polynomial \( \det(\lambda M - N) \) is denoted finite generalized eigenvalue of the pencil. Since \( M \) may be rank deficient, the number \( n_f \), accounting of multiplicity, of finite generalized eigenvalues is generally less than \( n \).

Definition A.7.1 Given a regular matrix pencil \( \lambda M - N \) of dimension \( n \), a subspace \( V \subset \mathbb{C}^n \) such that

\[
\dim(MV + NV) = \dim(V)
\]

is called a deflating subspace of the pencil \( \lambda M - N \).

The following characterization can be shown, see e.g [9]

Theorem A.7.1 Given a regular matrix pencil \( \lambda M - N \) of dimension \( n \), a matrix \( V \in \mathbb{C}^{n \times l} \) spans a deflating subspace of dimension \( l \) if and only if there exist \( S \in \mathbb{C}^{l \times l} \) such that

\[
MVS = NV
\]

(A.61)

and \( MV \) is injective – i.e. has full column rank.
A deflating subspace is called \textit{stable} when the matrix \( S \) in (A.61) has all eigenvalues inside the unit disc.

\section*{Extended sympletic pencil and the Riccati equation}

Consider the discrete-time Riccati equation
\[
A^*XA - X - (L + A^*XB)(R + B^*XB)^{-1}(B^*XA + L^*) + Q = 0 \tag{A.62}
\]
where \( A \in \mathbb{C}^{n \times n}, \ B \in \mathbb{C}^{n \times m}, \ R = R^* \in \mathbb{C}^{m \times m}, \ L \in \mathbb{C}^{n \times m} \) and \( Q = Q^* \in \mathbb{C}^{n \times n} \). The matrix \( P = P^* \in \mathbb{C}^{n \times n} \) is a \textit{stabilizing solution} of the Riccati equation if \( R + B^*XB \) is invertible, \( X = P \) satisfies (A.62) and \( S = A + BF \) where
\[
F = -(R + B^*PB)^{-1}(B^*PA + L^*)
\]
has all eigenvalues inside the unit disc.

It turns out that the discrete-time Riccati equation is strictly linked to a matrix pencil \( \lambda M - N \) of dimension \( 2n + m \) where
\[
M = \begin{bmatrix} I & 0 & 0 \\ 0 & -A^* & 0 \\ 0 & -B^* & 0 \end{bmatrix}, \quad N = \begin{bmatrix} A & 0 & B \\ Q & -I & L \\ L^* & 0 & R \end{bmatrix} \tag{A.63}
\]
which will be denoted the \textit{extended sympletic pencil} associated with (A.62).

Now consider a deflating subspace \( V \) of dimension \( l \) relative to the pencil (A.63). Any matrix \( V \in \mathbb{C}^{(2n+m) \times l} \) generating \( V \) can be partitioned as
\[
V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}
\]
where \( V_1 \in \mathbb{C}^{n \times l}, \ V_2 \in \mathbb{C}^{n \times l} \) and \( V_3 \in \mathbb{C}^{m \times l} \).

The connection with the Riccati equation was proven in \[11\] Theorem A.7.2: \textit{The discrete-time Riccati equation (A.62) has a stabilizing solution if and only if the corresponding extended sympletic pencil (A.63) admits a stable deflating subspace \( V \) of dimension \( n \) such that \( V_1 \) is invertible.}

\section*{Proof of theorem A.5.1}

The characteristic polynomial of (A.63) with \( R, L = 0 \) simplifies to
\[
\det(\lambda M - N) = (-1)^m \lambda^n \det(\lambda I - A) \det(\lambda^{-1}I - A) \det\{G^*(\lambda)\Delta G(\lambda)\}.
\]
Since $G^*AG$ is positive definite on the unit circle, at least $n$ generalized eigenvalues are stable and no generalized eigenvalue lie on $T$. Therefore there exist a stable deflating subspace of dimension $n$ – cf. [10].

It remains to show that the corresponding $V_1$ is invertible. First it will be shown that $\ker V_1 \subset \ker V_1^* V_2 \subset \ker V_2$. Since a stable deflating exists equation \[A.61\] holds for some strictly stable matrix $S$, it can be rewritten as

\[AV_1 + BV_3 = V_1 S,\]
\[QV_1 - V_2 = -A^* V_2 S,\]
\[0 = -B^* V_2 S.\]

Adding the last two equations premultiplied by $V_1^*$ and $V_3^*$ respectively and using the first we obtain

\[S^* V_1^* V_2 S - V_1^* V_2 + V_1^* Q V_1 = 0\] \[A.65\]

where $V_1^* Q V_1$ is hermitian and positive semidefinite. Therefore $V_1^* V_2$ is a solution of a discrete-time Lyapunov equation and thus must be hermitian and positive semidefinite. It follows that $\ker V_1 \subset \ker V_2^* V_1 = \ker V_1^* V_2$.

Since both $Q$ and $V_1^* V_2$ are positive semidefinite, from \[A.65\] we obtain that

\[V_1^* V_2 S x = 0,\]
\[Q V_3 x = 0\]

for any $x \in \ker V_1^* V_2$. Hence $\ker V_1^* V_2$ is $S$-invariant and $\ker V_1^* V_2 \subset \ker Q V_3$.

Let $K$ be a basis matrix for $\ker V_1^* V_2$. By $S$-invariance there exist a matrix $J$ such that

\[SK = KJ\]

and the spectrum of $J$ is a subset of the spectrum of $S$. Postmultiplying the second of \[A.64\] we have

\[V_2 K = A^* V_2 S K = A^* V_2 K J\] \[A.66\]

and, since both $A$ and $J$ have eigenvalues inside the open unit circle, it follows that $V_2 K = 0$. Thus $\ker V_1 \subset \ker V_1^* V_2 \subset \ker V_2$.

Finally, consider the matrix

\[MV = \begin{bmatrix} V_1 \\ -A^* V_2 \\ -B^* V_2 \end{bmatrix}\]

that is injective by Theorem \[A.7.1\] On the other hand, for any $x \in \ker V_1$ we have that $MV x = 0$, so it follows that $\ker V_1 = \{0\}$ and $V_1$ is invertible.
A.8 Bibliography


