Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

On a Generalization of the Kalman-Yakubovich-Popov Lemma

P.F. Curran

School of Electrical, Electronic and Mechanical Engineering, University College, Belfield, Dublin 4, Ireland, (e-mail: paul.curran@ucd.ie)

Abstract: The purpose of this note is to develop a generalization of the K-Y-P lemma and to apply this result to the absolute stability of single-variable Lur'e systems with sector-restricted nonlinearities.

Keywords: Nonlinear systems; Switching control; Hybrid systems stability.

1. INTRODUCTION

The problem of the absolute stability of single-variable Lur'e systems with sector-restricted, time-varying nonlinearities has been extensively investigated (Boyd et al. [1994]). In light of the work of Pyatnitskii (Pyatnitskii [1970]) the problem may be viewed as a special case of the stability problem for hybrid dynamical systems (Liberzon & Morse [1999]). Although numerous results are available, few are independent of the co-ordinates or the order and the majority deploy some numerical procedures. The application of Lyapunov methods to this problem most commonly leads to an LMI in a certain standard form. Application of the K-Y-P lemma (Rantzer [1996]) transforms these standard LMIs to frequency-domain inequalities. The circle criterion (Boyd et al. [1994]) is, without doubt, the most useful and general absolute stability criterion which can be thus obtained. Moreover, this criterion comprises a necessary and sufficient condition for the existence of a common quadratic Lyapunov function for a single-variable Lur'e system with sector-restricted nonlinearity.

Johansson & Rantzer [1998] have considered the problem of determining the existence of common, piecewise-quadratic Lyapunov functions for Lur'e systems and present the solution in the form of LMIs which contain many free parameters. The sheer number of free parameters is certainly a problem in using this technique. Some work (Brockett [1965], Molchanov & Pyatnitskii [1986], Wulff et al. [2002], Duignan & Curran [2006]) has suggested that there may be merit in considering far more specialized common, piecewise-quadratic Lyapunov functions. In the present note necessary and sufficient conditions are sought for the existence of common Lyapunov functions of this specialized form. Conditions for the existence of such functions lead to two *simultaneous* LMIs, with each individual LMI in standard form.

In his proof of the K-Y-P lemma Rantzer [1996] introduced a novel approach to the treatment of a single, standard-form LMI. The purpose of the present paper is to show that this approach can be adapted to the study of two simultaneous, standard-form LMIs. Whereas in the case of a single, standard-form LMI Rantzer is able to present the necessary and sufficient conditions for the existence of a solution in the *simplest* possible form, this cannot be achieved in the case of two simultaneous, standard-form LMIs. Nonetheless we contest that in the latter case necessary and sufficient conditions for the existence of a solution can be presented in the *second simplest* possible form. The formal meaning of simplicity in this context is outlined below. The proof of this claim presented below is unfortunately subject to a simplifying assumption concerning the simultaneous LMIs. We feel that this assumption can be eliminated, but have yet to successfully do so.

Notation: Given any matrix M, M^* denotes the transpose conjugate, $tr(M)$ denotes the trace and $\lambda_i(M)$ denotes an eigenvalue of M.

2. A GENERALIZATION OF THE K-Y-P LEMMA

The LMI considered by Rantzer [1996] takes the following form:

$$\begin{bmatrix} A^T P + PA & PB \\ B^T P & 0 \end{bmatrix} + N < 0. \qquad (1)$$

We will call this a *standard-form LMI*. We seek necessary and sufficient conditions for the existence of a symmetric solution P to the two simultaneous, standard-form LMIs:

$$\begin{bmatrix} A^T P + PA & PB \\ B^T P & 0 \end{bmatrix} + N \pm M < 0. \qquad (2)$$

Although it is a constraint that we conjecture may be unnecessary and one which we will seek to eliminate in future work, for the present we must assume that $rank(M) \leq 2$. This is the unfortunate assumption referred to at the end of the introduction. The K-Y-P lemma provides necessary and sufficient conditions for the existence of a symmetric solution P to this LMI in the special case where $M = 0$ and under certain relatively mild assumptions concerning matrices A and B. Rantzer [1996] provides a proof of the K-Y-P lemma. Logically his method divides the proof into two parts. The first part comprises a translation of the problem into a question of whether the intersection of two convex sets is null. This same translation can be employed in a wide variety of alternative but related problems and in particular will be employed below in treating the simultaneous LMI (2) of the present paper. The second part of Rantzer's method comprises a useful insight into the intersection of these convex sets, namely that if there exist points in the intersection then there exist very special points (i.e. points having a special, simple structure) in this intersection. We will establish that a corresponding, if not quite so simple, property holds for the intersection points of the convex sets arising in the context of the simultaneous LMI (2) of the present paper. We state the main result as follows:

Theorem 1. Given $N = N^T$ and $M = M^T \in R^{(n+m)\times(n+m)}$, $A \in R^{n\times n}$, $B \in R^{n\times m}$, there exists $P = P^T \in R^{n\times n}$ such that

$$\begin{bmatrix} A^T P + PA & PB \\ B^T P & 0 \end{bmatrix} + N \pm M < 0 \tag{3}$$

if and only if

$$\sum_{k=1}^{2} (z_k^* N z_k + |z_k^* M z_k|) < 0$$

for all $z_1,\ z_2 \in\ C^{n+m}$ not both zero s.t.

$$\sum_{k=1}^{2} [\, I_n\ \ 0\,]\, z_k\, ([\, A\ \ B\,]\, z_k)^* + [\, A\ \ B\,]\, z_k\, ([\, I_n\ \ 0\,]\, z_k)^* = 0. \tag{4}$$

Proof. To commence we closely follow the proof of Rantzer [1996]. Let H_n denote the set of $n \times n$ Hermitian matrices, let $\|\ \|_2$ denote the 2-norm and let $conv(S)$ denote the convex hull of set S. Define sets

$$\begin{aligned} \Theta = \{ & (r, \mathcal{H}) \in R \times H_n : r = z^* N z + |z^* M z|\,, \\ & \mathcal{H} = ([\, I_n\ \ 0\,]\, z)\, ([\, A\ \ B\,]\, z)^* + ([\, A\ \ B\,]\, z)\, ([\, I_n\ \ 0\,]\, z)^*\,, \\ & z \in C^{n+m} \text{ s.t. } \|\, z \,\|_2 = 1 \} \end{aligned}$$

and

$$\mathcal{P} = \{(r, \mathcal{H}) \in R \times H_n : r \geq 0\ ,\quad \mathcal{H} = 0\}\,. \tag{5}$$

It is clear that $\mathcal{P}$ is convex. For any given real number p and real, symmetric $n \times n$ matrix P define an associated real linear functional (p, P) on $R \times H_n$ by

$$(p, P) : (r, \mathcal{H}) \mapsto pr + tr(P\mathcal{H}). \tag{6}$$

Following the reasoning of Rantzer it is clear that if P is a solution of the simultaneous LMI (2) then the functional $(1, P)$ is positive on $\mathcal{P} \setminus \{(0, 0)\}$ and negative on Θ. Accordingly there exists a symmetric solution P to LMI (2) *only if*

$$conv(\Theta) \cap \mathcal{P} = \emptyset. \tag{7}$$

We contend that necessary condition (7) is also sufficient for simultaneous LMI (2) to possess a symmetric solution P. In justifying this contention we depart somewhat from Rantzer's proof and, moreover, postpone the details to the appendix. Note that $\mathcal{P}$ is a closed convex cone with vertex $(0, 0)$ (Lay [1982]) satisfying $\mathcal{P} \cap (-\mathcal{P}) = \{(0, 0)\}$. Note also that since Θ is compact (being a continuous image of a compact set), $conv(\Theta)$ is convex and compact. It follows from lemma 2 of the appendix that if (7) holds then there exists a real linear functional on $R \times H_n$ which is negative on $conv(\Theta)$ and positive on $\mathcal{P} \setminus \{(0, 0)\}$. Accordingly there exists a symmetric solution P to LMI (2) if and only if condition (7) holds. This completes the first part of the proof of the theorem 1 by Rantzer's method where the question of the existence of solutions to an LMI problem has been translated into the question of whether the intersection of two convex sets is null.

Any element of $conv(\Theta)$ may be written as a convex combination of elements of Θ. Hence, after some manipulation, the condition (7) can be reduced to the more explicit form: for all $z_1, \ldots, z_K \in C^{n+m}$ for which

$$\|\, z_k \,\|_2 \leq 1 \quad \text{for all}\, k\ \ ,\ \ \sum_{k=1}^{K} \|\, z_k \,\|_2 = 1 \quad \text{and}$$

$$\sum_{k=1}^{K} ([\, I_n\ \ 0\,]\, z_k)\, ([\, A\ \ B\,]\, z_k)^* + ([\, A\ \ B\,]\, z_k)\, ([\, I_n\ \ 0\,]\, z_k)^* = 0$$

one has

$$\sum_{k=1}^{K} (z_k^* N z_k + |z_k^* M z_k|) < 0. \tag{8}$$

Caratheodary's theorem (Lay [1982]) assures that K need not exceed $n^2 + 2$. Although there are ways of viewing condition (8) which render it a little less intimidating there remains a fundamental problem that this condition is not simple. The number K is a good measure of the complexity of the condition, with smaller K yielding a simpler condition. K is the formal measure of simplicity referred to in the introduction. The second part of Rantzer's method in the special case where $M = 0$ justifies setting $K = 1$ in general. This renders condition (8) as simple as possible and in fact the condition reduces to a standard frequency domain inequality in this case. Unfortunately for the simultaneous LMI (2) one cannot in general set $K = 1$. However, we will go on to show that it is possible in certain cases (and in particular when $rank(M) \leq 2$) to do the *next best thing*, namely to set $K = 2$. It is in this sense that we can claim, as we did in the introduction, that the condition is reduced to the *second simplest* form. The completion of the proof therefore relies on establishing that if condition (8) holds for $K = 2$ then it holds for any K not exceeding $n^2 + 2$.

Accordingly, assume condition (8) holds for $K = 2$ and consider the condition for general K. Since

$$\sum_{k=1}^{K} ([\, I_n\ \ 0\,]\, z_k)\, ([\, A\ \ B\,]\, z_k)^* + ([\, A\ \ B\,]\, z_k)\, ([\, I_n\ \ 0\,]\, z_k)^* = 0$$

one has

$$[\, I_n\ \ 0\,]\, W\, [\, A\ \ B\,]^* + [\, A\ \ B\,]\, W\, [\, I_n\ \ 0\,]^* = 0 \tag{9}$$

where

$$W = \sum_{k=1}^{K} z_k z_k^*.$$

Reorder $\{z_1, \ldots, z_K\}$ if necessary such that $z_k^* M Z_k \geq 0$ for $k = 1, \ldots, r$ and $z_k^* M Z_k < 0$ for $k = r + 1, \ldots, K$, allowing that r may be zero or K if one of the subsets is empty. Let

$$W_1 = \sum_{k=1}^{r} z_k z_k^* \quad ,\quad W_2 = \sum_{k=r+1}^{K} z_k z_k^*$$

allowing that one of the sums may be zero. Clearly W_i is positive, semi-definite, Hermitian for each i. Accordingly, there exist $Q_i \in C^{(n+m)\times(n+m)}$ such that

$$W_i = Q_i Q_i^* \;\; i = 1,2 \quad \text{and} \quad W = W_1 + W_2.$$

It follows from (9) that

$$\begin{aligned}&([\, I_n \;\, 0\,]\,[\, Q_1 \;\, Q_2\,])\,([\, A \;\, B\,]\,[\, Q_1 \;\, Q_2\,])^* \\ &+([\, A \;\, B\,]\,[\, Q_1 \;\, Q_2\,])\,([\, I_n \;\, 0\,]\,[\, Q_1 \;\, Q_2\,])^* = 0 \qquad (10)\end{aligned}$$

Employing a lemma of Rantzer [1996] it follows that (10) holds if and only if

$$\begin{aligned}&[\, I_n \;\, 0\,]\,[\, Q_1 \;\, Q_2\,]\,\big(I_{2(n+m)} + U\big) \\ &= [\, A \;\, B\,]\,[\, Q_1 \;\, Q_2\,]\,\big(I_{2(n+m)} - U\big), \qquad (11)\end{aligned}$$

$U \in C^{2(n+m)\times 2(n+m)}$, unitary. Lemma 3 of the appendix introduces a *second-order decomposition* for U. Employing such a decomposition it follows that for all k

$$\big(I_{2(n+m)} + U\big)\begin{bmatrix} F_k \\ G_k \end{bmatrix} = \begin{bmatrix} F_k \\ G_k \end{bmatrix}(I_2 - \Phi_k)\,.$$

Accordingly, by (11) for all k

$$\begin{aligned}&[\, I_n \;\, 0\,]\,[\, Q_1 \;\, Q_2\,]\begin{bmatrix} F_k \\ G_k \end{bmatrix}(I_2 + \Phi_k) \\ &= [\, A \;\, B\,]\,[\, Q_1 \;\, Q_2\,]\begin{bmatrix} F_k \\ G_k \end{bmatrix}(I_2 - \Phi_k)\,. \qquad (12)\end{aligned}$$

For all k let $Q_1 F_k + Q_2 G_k = [\hat{w}_k \;\, \tilde{w}_k] \in C^{(n+m)\times 2}$, then by (12)

$$[\, I_n \;\, 0\,]\,[\, \hat{w}_k \;\, \tilde{w}_k\,]\,(I_2 + \Phi_k) = [\, A \;\, B\,]\,[\, \hat{w}_k \;\, \tilde{w}_k\,]\,(I_2 - \Phi_k)\,. \qquad (13)$$

Employing the same lemma of Rantzer [1996] it follows that

$$\begin{aligned}&[\, I_n \;\, 0\,]\,(\hat{w}_k \hat{w}_k^* + \tilde{w}_k \tilde{w}_k^*)\,[\, A \;\, B\,]^* \\ &+[\, A \;\, B\,]\,(\hat{w}_k \hat{w}_k^* + \tilde{w}_k \tilde{w}_k^*)\,[\, I_n \;\, 0\,]^* = 0. \qquad (14)\end{aligned}$$

Setting $z_1 = \hat{w}_k$ and $z_2 = \tilde{w}_k$, then from (14) and the assumed $K = 2$ case of condition (8), it follows that for each $k = 1, \ldots, n+m$

$$\hat{w}_k^* N \hat{w}_k + |\hat{w}_k^* M \hat{w}_k| + \tilde{w}_k^* N \tilde{w}_k + |\tilde{w}_k^* M \tilde{w}_k| < 0$$

and therefore upon adding and applying the triangle inequality

$$tr(N\hat{W}_1) + \left|tr(M\hat{W}_1\right| + tr(N\hat{W}_2) + \left|tr(M\hat{W}_2\right| < 0. \qquad (15)$$

Let

$$\hat{W}_1 = \sum_{k=1}^{n+m} \hat{w}_k \hat{w}_k^* \quad , \quad \hat{W}_2 = \sum_{k=1}^{n+m} \tilde{w}_k \tilde{w}_k^*,$$

then

$$\begin{aligned}\hat{W}_1 + \hat{W}_2 &= \sum_{k=1}^{n+m} [\, \hat{w}_k \;\, \tilde{w}_k\,]\,[\, \hat{w}_k \;\, \tilde{w}_k\,]^* \\ &= \sum_{k=1}^{n+m} [\, Q_1 \;\, Q_2\,]\begin{bmatrix} F_k \\ G_k \end{bmatrix}[\, F_k^* \;\, G_k^*\,]\,[\, Q_1 \;\, Q_2\,]^* \\ &= [\, Q_1 \;\, Q_2\,]\,[\, Q_1 \;\, Q_2\,]^* = Q_1 Q_1^* + Q_2 Q_2^* = W_1 + W_2 = W.\end{aligned}$$

Accordingly (15) can be rewritten

$$tr(NW_1) + \left|tr(M\hat{W}_1\right| + tr(NW_2) + \left|tr(M\hat{W}_2\right| < 0. \qquad (16)$$

As (16) must hold for every unitary matrix U or alternatively for every matrix $\hat{T}_1$ related to matrix U after the manner of lemma 4 it follows that it must also hold for that special choice of $\hat{T}_1$ such that the upper bound in lemma 5 is attained. In this special case it is readily shown that (16) becomes:

$$tr(NW_1) + tr(NW_2) + \sum_i |\lambda_i(MW_1 + MW_2)| < 0. \qquad (17)$$

Finally, by lemma 6, the only lemma which requires the unfortunate assumption that $rankM \leq 2$, one obtains:

$$tr(NW_1) + tr(NW_2) + |tr(MW_1)| + |tr(MW_2)| < 0 \qquad (18)$$

which is readily seen to be equivalent to condition (8). Accordingly we obtain the required result, that the $K = 2$ case of condition (8) implies the general case.

3. ABSOLUTE STABILITY

Towards an application of theorem 1 consider the single-variable Lur'e system (Boyd et al. [1994])

$$\dot{x} = Ax - bf(c^T x, t) \qquad (19)$$

where $x \in R^n$, $A \in R^{n\times n}$, b, $c \in R^{n\times 1}$, such that $b^T b = 1$, $c^T b = 1$, $\{b, c\}$ linearly independent and f is a sector $[0, \infty)$, time-varying nonlinearity. We seek a common piecewise-quadratic Lyapunov function of a highly specialized form:

$$V(x) = \begin{cases} x^T P_1 x \text{ if } (b^T x)(c^T x) > 0 \\ x^T P_2 x \text{ if } (b^T x)(c^T x) < 0 \end{cases} \qquad (20)$$

where the *matching condition*

$$x^T P_1 x = x^T P_2 x \;\text{ if }\; (b^T x)(c^T x) = 0$$

is imposed. Let $P = (P_1 + P_2)/2$ and, for convenience, let $S = bc^T + cb^T$, then the matching condition is equivalent to

$$P_1 = P + p_0 S \quad , \quad P_2 = P - p_0 S \qquad (21)$$

for arbitrary scalar p_0. Applying the S-procedure yields the following conditions for the derivative of V to be negative-definite for *any* sector $[0, \infty)$ nonlinearity f:

$$\begin{aligned}&Pb = (\gamma - p_0)c + (\tau - p_0)b \qquad (22)\\ &\gamma \geq 2p_0 \geq 0, \quad 2p_0 \geq \tau \geq 0.\\ &A^T P + PA + p_0(A^T S + SA) + q_1 S < 0, \quad q_1 \geq 0\\ &A^T P + PA - p_0(A^T S + SA) - q_2 S < 0, \quad q_2 \geq 0\end{aligned}$$

The latter may be combined to yield:

$$\begin{aligned}&A^T P + PA + ((q_1 - q_2)/2)S \qquad (23)\\ &\pm \big(p_0(A^T S + SA) + ((q_1 + q_2)/2)S\big) < 0\end{aligned}$$

Hence

$$A^T P + PA + N \pm M \leq 0 \qquad (24)$$

where

$$N = \begin{bmatrix} ((q_1 - q_2)/2)S & -(\gamma - p_0)c - (\tau - p_0)b \\ -(\gamma - p_0)c^T - (\tau - p_0)b^T & 0 \end{bmatrix}$$

and

$$M = \begin{bmatrix} p_0(A^T S + SA) + ((q_1 + q_2)/2)S & 0 \\ 0^T & 0 \end{bmatrix}.$$

Accordingly the question of the existence of a common Lyapunov function of this specialized form reduces to two simultaneous, standard form, albeit non-strict LMIs of the form (2). We note that in this example matrix M has rank 4 in general. The non-strict nature of the LMIs and the rank of M exceeding 2 imply that the generalized K-Y-P lemma as presented cannot be applied. The purpose of the example, rather, is to indicate how simultaneous, standard-form LMIs can arise in a natural way in absolute stability problems. Furthermore if, as conjectured, the critical lemma 6 can be proven without the unfortunate rank assumption, then the generalized K-Y-P lemma can also be established without this assumption. In this event the excessive rank of M proves no hindrance. Rantzer's method certainly applies to non-strict, standard-form LMIs. Whether it can be adjusted to apply to the LMIs of the form (22)/(23) remains uncertain. The author could perhaps construct an example of a stability problem which leads to simultaneous, standard form LMIs satisfying the required rank condition, but this would surely be disingenuous and any such example would surely appear contrived. The author prefers to openly state the need to prove lemma 6 (and hence the generalized K-Y-P lemma) ideally for matrices M of any rank and at a minimum for matrices M of rank not exceeding 4.

4. CONCLUSION

It is possible to extend the K-Y-P lemma to two simultaneous LMIs. Whereas the extension no longer leads to criteria for the existence of a solution in the simplest possible form, according to a certain strict sense of simplicity, it does lead to criteria in the *second simplest* possible form.

REFERENCES

S.Boyd, L. El Ghaoui, E.Feron and V.Balakrishnan. *Linear Matrix Inequalities in System and Control theory.* Studies in Applied Mathematics, SIAM, Philadelphia, volume 15, 1994.

R.W. Brockett. Optimisation theory and the converse of the circle criterion. *Proceedings Nat. Electronic Conf.*, 697–701, 1965.

R. Duignan and P. F. Curran. A new absolute stability criterion for 2-D single variable Lur'e systems. *Proceedings IEEE Int. Symp. on Nonlinear Theory and Applications*, 2006.

M. Johansson and A. Rantzer. Computation of piecewise quadratic Lyapunov functions for hybrid systems. *IEEE Trans. Automat. Control*, 43: 555–559, 1998.

V.L. Klee. Separation properties of convex cones. *A.M.S. Proceedings*, 6: 313–318, 1955.

S.R. Lay. *Convex sets and their application* Pure & Applied Mathematics, Wiley, New York, 1982.

D. Liberzon and A.S. Morse. Basic problems in stability and design of switched systems. *IEEE Trans. Automat. Control*, 19: 59–70, 1999.

A.P. Molchanov and E.S. Pyatnitskii. Lyapunov functions that specify necessary and sufficient conditions of absolute stabilty of nonlinear control systems I. *Avtomatika i Telemekhanika*, 47: 63–73, 1986.

E.S. Pyatnitskii. Absolute stability of nonstationary nonlinear systems. *Avtomatika i Telemekhanika*, 31: 5–15, 1970.

A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems & Control Letters*, 28: 7–10, 1996.

K. Wulff, R. Shorten and P. Curran. On the 45^o criterion and the uniform asymptotic stability of classes of second order parameter-varying and switched systems. *Int. J. Control*, 75: 812–823, 2002.

Appendix A

The statements of some technical details are presented.

Lemma 2. Let X be a real vector space of finite dimension. Let C be a convex, compact set not containing the zero vector 0. Let D be a closed convex cone with vertex 0 such that $D \cap (-D) = \{0\}$. Assume $C \cap D = \emptyset$, then there exists a continuous real linear functional h on X such that $h < 0$ on C, $h = 0$ on $\{0\}$ and $h > 0$ on $D \setminus \{0\}$.

Proof. We shall base the proof on theorem 2.7 of Klee [1955]. To this end, let

$$\tilde{C} = \{x \in X \;:\; x = \lambda c \;,\; \lambda \geq 0 \;,\; c \in C\}$$

Since C is convex, it is elementary to verify that $\tilde{C}$ is a convex cone with vertex 0. As $0 \notin C$ and C is compact, there exists $\eta > 0$ such that $\| c \|_2 \geq \eta$ for all $c \in C$. Selecting $R > 0$ arbitrary, choose $\lambda \geq R/\eta$. It is elementary to show that

$$\tilde{C} \cap \overline{B(0,R)} = (conv\ [\{0\} \cup \lambda C]) \cap \overline{B(0,R)} \tag{A.1}$$

where $\overline{B(0,R)}$ denotes the closure of the ball of radius R centered at 0. From (A.1) and the compactness of C it follows that $\tilde{C} \cap \overline{B(0,R)}$ is closed [2.30, 5]. As R is arbitrary it follows that $\tilde{C}$ is closed, since every Cauchy sequence is bounded.

From the convexity of C one readily obtains that $\tilde{C} \cap (-\tilde{C}) = \{0\}$. Finally, since C and D are disjoint, it follows that $D \cap \tilde{C} = \{0\}$.

As X is locally compact theorem 2.7 of Klee [1955] implies that there exists a continuous real linear functional h on X such that $h < 0$ on $\tilde{C} \setminus \{0\}$, $h = 0$ on $\{0\}$ and $h > 0$ on $D \setminus \{0\}$. Of course $0 \notin C$ implies $C \subset \tilde{C} \setminus \{0\}$, so that $h < 0$ on C.

Lemma 3. Let U be a $2(n+m) \times 2(n+m)$ unitary matrix then there exist 2×2 unitary matrices Φ_k and $(n+m) \times 2$ matrices F_k and G_k for $k = 1, \ldots, (n+m)$ such that

$$U = \sum_{k=1}^{n+m} \begin{bmatrix} F_k \\ G_k \end{bmatrix} \Phi_k \, [\, F_k^* \;\; G_k^* \,] \tag{A.2}$$

$$\sum_{k=1}^{n+m} \begin{bmatrix} F_k \\ G_k \end{bmatrix} \, [\, F_k^* \;\; G_k^* \,] = I_{2(n+m)}$$

$$[\,F_k^* \;\; G_k^*\,] \begin{bmatrix} F_l \\ G_l \end{bmatrix} = \delta_{kl} I_2$$

for all k, l.

Proof. As U is unitary $2(n+m) \times 2(n+m)$ there exist scalars θ_k and orthonormal vectors $u_k \in C^{2(n+m)}$ such that

$$U = \sum_{k=1}^{2(n+m)} e^{j\theta_k} u_k u_k^* \quad , \quad \sum_{k=1}^{2(n+m)} u_k u_k^* = I_{2(n+m)} \qquad \text{(A.3)}$$

Let

$$\Phi_k = e^{j\theta_{2k}} \hat{u}_k \hat{u}_k^* + e^{j\theta_{2k-1}} \tilde{u}_k \tilde{u}_k^* \qquad \text{(A.4)}$$

where $\hat{u}_k$ and $\tilde{u}_k$ are arbitrary 2×1 orthonormal vectors such that $\hat{u}_k \hat{u}_k^* + \tilde{u}_k \tilde{u}_k^* = I_2$. Let $F_k = [\, f_{1k} \;\; f_{2k}\,]$, $G_k = [\, g_{1k} \;\; g_{2k}\,]$ where f_{ik}, $g_{ik} \in C^{n+m}$. Let

$$u_k = \begin{bmatrix} u_{k1} \\ u_{k2} \end{bmatrix}$$

where $u_{ki} \in C^{(n+m)}$. With these notations equation (A.2) may be expanded as:

$$\begin{bmatrix} u_{2k,1} \\ u_{2k-1,1} \\ u_{2k,2} \\ u_{2k-1,2} \end{bmatrix} = \begin{bmatrix} \hat{u}_{k1} I & \hat{u}_{k2} I & 0 & 0 \\ \tilde{u}_{k1} I & \tilde{u}_{k2} I & 0 & 0 \\ 0 & 0 & \hat{u}_{k1} I & \hat{u}_{k2} I \\ 0 & 0 & \tilde{u}_{k1} I & \tilde{u}_{k2} I \end{bmatrix} \begin{bmatrix} f_{1k} \\ f_{2k} \\ g_{1k} \\ g_{2k} \end{bmatrix} \qquad \text{(A.5)}$$

where the identity matrices and the zero blocks are $(n+m)\times(n+m)$. Orthonormality of $\hat{u}_k$, $\tilde{u}_k$ implies invertibility, permitting one to express F_k, G_k in terms of u_k so that equation (A.2) holds. Moreover:

$$\sum_{k=1}^{n+m} = \begin{bmatrix} F_k \\ G_k \end{bmatrix} [\, F_k^* \;\; G_k^*\,]$$

$$= \sum_{k=1}^{n+m} \begin{bmatrix} f_{1k} & f_{2k} \\ g_{1k} & g_{2k} \end{bmatrix} (\hat{u}_k \hat{u}_k^* + \tilde{u}_k \tilde{u}_k^*) \begin{bmatrix} f_{1k}^* & g_{1k}^* \\ f_{2k}^* & g_{2k}^* \end{bmatrix}$$

$$= \sum_{k=1}^{n+m} \left(u_{2k} u_{2k}^* + u_{2k-1} u_{2k-1}^* \right) = \sum_{k=1}^{2(n+m)} u_k u_k^* = I_{2(n+m)}.$$

Finally:

$$\begin{bmatrix} \hat{u}_{k1} & \tilde{u}_{k1} \\ \hat{u}_{k2} & \tilde{u}_{k2} \end{bmatrix}^* [\, F_k^* \;\; G_k^*\,] \begin{bmatrix} F_l \\ G_l \end{bmatrix} \begin{bmatrix} \hat{u}_{l1} & \tilde{u}_{l1} \\ \hat{u}_{l2} & \tilde{u}_{l2} \end{bmatrix}$$

$$= \begin{bmatrix} u_{2k}^* \\ u_{2k-1}^* \end{bmatrix} [\, u_{2l} \;\; u_{2l-1}\,] = \delta_{kl} I_2.$$

From orthonormality of $\hat{u}_k$, $\tilde{u}_k$ one may deduce relatively easily that the last equation of the lemma follows.

In the statement of the lemma δ_{kl} denotes the Kronecker delta. We call decomposition (A.2) a *second order decomposition* of the unitary matrix U. The lemma therefore asserts that every unitary matrix of even order possesses a second order decomposition. Uniqueness of the decomposition is neither claimed nor required.

Lemma 4. Given any $Q_i \in C^{(n+m)\times(n+m)}$, $i = 1, 2$ and given a second order decomposition of a unitary matrix U, let $Q_1 F_k + Q_2 G_k = [\hat{w}_k \;\; \tilde{w}_k]$ for all $k = 1, \ldots, n+m$ then

$$\hat{W}_1 = \sum_{k=1}^{n+m} \hat{w}_k \hat{w}_k^* = [\, Q_1 \;\; Q_2\,]\, \hat{T}_1 \,[\, Q_1 \;\; Q_2\,]^*, \qquad \text{(A.6)}$$

for some Hermitian $\hat{T}_1 \in C^{2(n+m)\times 2(n+m)}$ such that $n+m$ eigenvalues of $\hat{T}_1$ are zero and $n+m$ eigenvalues of $\hat{T}_1$ are unity. Moreover, given any matrix $\hat{T}_1$ of this form, there exist a unitary matrix U such that the associated matrix $\hat{W}_1$ is given by (A.6).

Proof. With all of the same notations as in the proof of lemma 3 we obtain:

$$[\, u_{2k} \;\; u_{2k-1}\,] = \begin{bmatrix} f_{1k} & f_{2k} \\ g_{1k} & g_{2k} \end{bmatrix} \begin{bmatrix} \hat{u}_{k1} & \tilde{u}_{k1} \\ \hat{u}_{k2} & \tilde{u}_{k2} \end{bmatrix}$$

which upon inversion yields:

$$\begin{bmatrix} f_{1k} \\ g_{1k} \end{bmatrix} = \frac{\tilde{u}_{k2} u_{2k} - \hat{u}_{k2} u_{2k-1}}{\hat{u}_{k1} \tilde{u}_{k2} - \hat{u}_{k2} \tilde{u}_{k1}}.$$

Straightforward manipulation leads to equation (A.6) with:

$$\hat{T}_1 = \sum_{k=1}^{n+m} \frac{\left(\tilde{u}_{k2} u_{2k} - \hat{u}_{k2} u_{2k-1}\right) \left(\tilde{u}_{k2} u_{2k} - \hat{u}_{k2} u_{2k-1}\right)^*}{|\hat{u}_{k1} \tilde{u}_{k2} - \hat{u}_{k2} \tilde{u}_{k1}|^2}$$

Let

$$\hat{v}_k = \frac{\tilde{u}_{k2} u_{2k} - \hat{u}_{k2} u_{2k-1}}{\hat{u}_{k1} \tilde{u}_{k2} - \hat{u}_{k2} \tilde{u}_{k1}}$$

By orthonormality of $\{u_k\}$ it is possible to establish orthonormality of $\{\hat{v}_k\}$. Accordingly, as $\hat{T}_1 = \sum_{k=1}^{n+m} \hat{v}_k \hat{v}_k^*$ it follows that $(n+m)$ eigenvalues of $\hat{T}_1$ are unity and $(n+m)$ eigenvalues are zero. Moreover, given any $\hat{T}_1 \in C^{2(n+m)\times 2(n+m)}$, Hermitian with $(n+m)$ unity eigenvalues and $(n+m)$ zero eigenvalues, there exist $(n+m)$ orthonormal $\hat{v}_k \in C^{2(n+m)}$ such that $\hat{T}_1 = \sum_{k=1}^{n+m} \hat{v}_k \hat{v}_k^*$. Let $\{u_k\}$ be any orthonormal set in $C^{2(n+m)}$ such that $u_{2k} = \hat{v}_k$. Let $\tilde{u}_{k1} = 0$, $\tilde{u}_{k2} = 1$, $\hat{u}_{k1} = 1$ and $\hat{u}_{k2} = 0$, then

$$\hat{v}_k = \frac{\tilde{u}_{k2} u_{2k} - \hat{u}_{k2} u_{2k-1}}{\hat{u}_{k1} \tilde{u}_{k2} - \hat{u}_{k2} \tilde{u}_{k1}}$$

and $\{\hat{u}_k \;,\; \tilde{u}_k\}$ are orthonormal. The second part of the lemma follows from these observations.

Lemma 5. Given any $W_i = Q_i Q_i^*$, $i = 1, 2$ and given $\hat{W}_1 = [\, Q_1 \;\; Q_2\,]\, \hat{T}_1 \,[\, Q_1 \;\; Q_2\,]^*$ for some Hermitian $\hat{T}_1$ of the form of lemma 4

$$\sum_{i:\lambda_i \le 0} \lambda_i(MW_1 + MW_2) \le tr(M\hat{W}_1)$$
$$\le \sum_{i:\lambda_i \ge 0} \lambda_i(MW_1 + MW_2). \qquad \text{(A.7)}$$

Moreover, there exist matrices $\hat{T}_1$ of the form of lemma 4 such that the upper and lower bounds are attained.

Proof. By definition of $\hat{W}_1$:

$$tr(M\hat{W}_1) = tr\left(\begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix} M \,[\, Q_1 \;\; Q_2\,]\, \hat{T}_1 \right).$$

Let

$$\hat{M} = \begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix} M \, [\, Q_1 \;\; Q_2 \,] \, .$$

As $\hat{T}_1$ and $\hat{M}$ are both Hermitian it follows that $\hat{M} = \hat{V}\Sigma_{\hat{M}}\hat{V}^*$ and $\hat{T}_1 = \hat{V}_1\Sigma_{\hat{T}_1}\hat{V}_1^*$, where $\hat{V}$ and $\hat{V}_1$ are unitary, $\Sigma_{\hat{M}}$ is real diagonal

$$\Sigma_{\hat{T}_1} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} .$$

Accordingly

$$tr(\hat{M}\hat{T}_1) = \sum_{i=1}^{2(n+m)} \lambda_i(\hat{M})(\tilde{V}\tilde{V}^*)_{ii}$$

where $\tilde{V} = \hat{V}^*\hat{V}_1\Sigma_{\hat{T}_1}$. It is possible to establish that $0 \le (\tilde{V}\tilde{V}^*)_{ii} \le 1$. Hence:

$$\sum_{i\,:\,\lambda_i \le 0} \lambda_i(\hat{M}) \le tr(\hat{M}\hat{T}_1) \le \sum_{i\,:\,\lambda_i \ge 0} \lambda_i(\hat{M}).$$

Moreover by selecting $\hat{T}_1 = \hat{V}\Sigma\hat{V}^*$, Σ diagonal with $\Sigma_{ii} = 1$ for each i such that $\lambda_i(\hat{M}) > 0$, $\Sigma_{ii} = 0$ for each i such that $\lambda_i(\hat{M}) < 0$ and $\Sigma_{ii} = 0$ or 1 for all other i, selected such that half of the eigenvalues of Σ are unity and half are zero, the upper bound is attained. Similarly the lower bound is attained for suitable $\hat{T}_1$. Hence:

$$\sum_{i\,:\,\lambda_i \le 0} \lambda_i(\hat{M}) \le tr(M\hat{W}_1) \le \sum_{i\,:\,\lambda_i \ge 0} \lambda_i(\hat{M})$$

with upper and lower bounds attained for suitable $\hat{T}_1$. It is elementary to establish that $n+m$ of the eigenvalues of $\hat{M}$ are zero and the remaining $n+m$ eigenvalues are the eigenvalues of $MQ_1Q_1^* + MQ_2Q_2^* = MW_1 + MW_2$. The statement of the lemma follows from this observation.

Lemma 6. Given any $q \times q$ Hermitian matrix M with $rank(M) \le 2$, let vectors v_k satisfy $v_k^*Mv_k \ge 0$ for $k = 1, \ldots, r$, $v_k^*Mv_k < 0$ for $k = r+1, \ldots, q$ and define Q_1 and Q_2 by:

$$Q_1Q_1^* = \sum_{k=1}^{r} v_k v_k^* \quad , \quad Q_2Q_2^* = \sum_{k=r+1}^{q} v_k v_k^*$$

then

$$\left| \sum_i \lambda_i \left(\begin{bmatrix} Q_1^* \\ 0 \end{bmatrix} M \, [\, Q_1 \;\; 0 \,] \right) \right| + \left| \sum_i \lambda_i \left(\begin{bmatrix} 0 \\ Q_2^* \end{bmatrix} M \, [\, 0 \;\; Q_2 \,] \right) \right|$$
$$\le \sum_i \left| \lambda_i \left(\begin{bmatrix} Q_1^* \\ Q_2^* \end{bmatrix} M \, [\, Q_1 \;\; Q_2 \,] \right) \right| . \tag{A.8}$$

The proof of this lemma is omitted at the present time since it is hoped to eliminate the rank condition in future.